Revista Colombiana de Estadística
Junio 2010, volumen 33, no. 1, pp. 63 a 88

Confidence and Credibility Intervals for the Difference of Two Proportions

Intervalos de confianza y de credibilidad para la diferencia de dos proporciones

Hanwen Zhang[1,a], Hugo Andrés Gutiérrez Rojas[1,b], Edilberto Cepeda Cuervo[2,c]

[1]Centro de Investigaciones y Estudios Estadísticos (CIEES), Facultad de Estadística, Universidad Santo Tomás, Bogotá, Colombia

[2]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

This paper presents a frequentist comparison of the performance of confidence and credibility intervals for the difference of two proportions from two independent samples. The comparison is carried out considering three frequentist criteria. It was found that the intervals with the best performance, in terms of coverage probability, are Bayesians; in terms of expected length and variance of the length, the Newcombe interval shows the best performance. As a final remark, it was found that traditional intervals such as the Wald and adjusted Wald have a poor performance.

***Key words*:** Confidence intervals, Credibility intervals, Difference of two proportions..

Resumen

Este artículo presenta una comparación del comportamiento de intervalos de confianza frecuentistas y de credibilidad bayesianos para la diferencia de dos proporciones provenientes de muestras aleatorias independientes. La comparación se lleva cabo considerando tres criterios frecuentistas con los cuales se concluyó que el mejor comportamiento, en términos de la probabilidad de cobertura, lo tienen los intervalos bayesianos, y en términos de la longitud esperada y varianza de la longitud el mejor comportamiento está dado por el intervalo frecuentista de Newcombe. Como resultado de esta investigación se encontró que los intervalos frecuentistas más populares como Wald y Wald ajustado tienen un comportamiento deficiente.

***Palabras clave*:** intervalos de confianza, intervalos de credibilidad, diferencia de dos proporciones.

[a]Docente investigadora. E-mail: hanwenzhang@usantotomas.edu.co
[b]Director. E-mail: hugogutierrez@usantotomas.edu.co
[c]Profesor asociado. E-mail: ecepedac@unal.edu.co

1. Background

A common problem in practical statistics is estimatig the difference of two proportions by means of interval estimation. This topic is especially important in clinical trials where it is necessary to investigate cure rates of two drugs or treatments. The theoretical background of this research is as follows: suppose that $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$ are two independent samples such that $X_i \sim Bernoulli(p_1)$ and $Y_j \sim Bernoulli(p_2)$, with $i = 1, \dots, n_1$ and $j = 1, \dots, n_2$. It is necessary to construct a confidence interval or a credibility interval for the difference of the proportions $p_1 - p_2$.

The most popular method for estimating $p_1 - p_2$ by means of frequentist confidence interval is the Wald interval, which is presented in most introductory statistics textbooks in spite of its poor performance. Many modifications have been made to the Wald interval in order to improve it. One of them is the adjusted Wald interval obtained by widening the Wald interval to increase the coverage probability. This improvement is especially meaningful when the sample sizes are small. Another important interval is the score interval (Wilson 1927), obtained by inverting the score test statistics. This interval was first obtained for one proportion, and thereafter was to be extended to deal with the difference of two proportions. However, in that case, the interval lacks a closed form (Pan 2002) and must be computed by numerical approximations. Agresti & Caffo (2000) analyzed the score interval, and derived the Adding-4 method: add 2 successes and 2 failures to sample observation. A considerable number of authors agree that Agresti and Caffo method has a very good performance (Pan 2002, Correa & Sierra 2003, Agresti et al. 2008). Another interval obtained by modifying the score method is the Newcombe interval (Newcombe 1998*a*, 1998*b*), and it seems to have a similar performance to the Agresti and Caffo interval (Correa & Sierra 2003).

In the Bayesian approach, Pham-Gia & Turkkan (1993) used the hypergeometric Appell function and derived the posterior distribution of $p_1 - p_2$ when beta priors are used for each proportion. Given the exact posterior distribution, an exact Bayesian credibility interval for $p_1 - p_2$ can be found. However the computational procedures are somewhat tedious, therefore new computational methods such as the Markov Chain Monte Carlo (MCMC), can be used to make it easier to evaluate posterior distributions for $p_1 - p_2$, as Agresti & Min (2005) argued.

In the literature, many comparisons between confidence intervals have been done (Newcombe 1998*a*, Newcombe 1998*b*, Agresti & Caffo 2000, Pan 2002, Correa & Sierra 2003). The aim of this research is to take into account Bayesian credibility intervals jointly with frequentist confidence intervals. After a brief introduction, Section 2 presents some frequentist and Bayesian intervals for $p_1 - p_2$. Traditional confidence intervals such as the Wald and adjusted Wald are considered, as well as Bayesian credibility intervals with two noninformative priors. Section 3 deals with the comparison criteria for the considered intervals: the coverage probability, the expected length, and the variance of the length are used in order to evaluate the performance of the intervals. Section 4 presents results for the performance of the intervals with varying sample sizes, varying values of a single proportion and, finally, the difference of the two proportions. Other scenarios were analyzed, but all of them yield similar conclusions. Section 5 provides a survey of other intervals and their performance, and finally Section 6 gives some conclusions and recommendations.

2. Some intervals

In this section we introduce some confidence and credibility intervals that are considered and lead the research through out this paper. We denote $\widehat{p}_1$ as the maximum likelihood estimator of p_1 defined as $\sum_{i=1}^{n_1} \dfrac{X_i}{n_1}$ and analogously for $\widehat{p}_2$.

2.1. Frequentist intervals

The Wald interval is based on the normal approximation to the distribution of $\widehat{p}_1 - \widehat{p}_2$, when the sample sizes are large, by considering that

$$E(\widehat{p}_1 - \widehat{p}_2) = p_1 - p_2$$
$$Var(\widehat{p}_1 - \widehat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

By the the central limit theorem a $(1-\alpha)100\%$ interval for $p_1 - p_2$ is clearly defined by (L_{low}, L_{upp}), where

$$L_{low} = \widehat{p}_1 - \widehat{p}_2 - z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \tag{1}$$

and

$$L_{upp} = \widehat{p}_1 - \widehat{p}_2 + z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \tag{2}$$

The computation of this interval is very simple, and it is presented in most of the statistical inference textbooks. Despite the fact of its popularity, many authors have shown that the performance of this interval is quite poor (Ghosh 1979, Vollset 1993, Newcombe 1998*a*, Newcombe 1998*b*). Moreover, when the sample sizes are large, the Wald interval still performs poorly (Brown et al. 2001).

Considering that the Wald interval uses a continuous distribution to approximate a discrete distribution, an alternative to for improving the performance of the Wald interval is to incorporate the continuity correction factor by adding a constant term to both the lower and upper limits. The resulting limits of the adjusted Wald interval are:

$$L_{low} = \widehat{p}_1 - \widehat{p}_2 - z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} - \frac{n_1+n_2}{2n_1n_2} \tag{3}$$

and

$$L_{upp} = \widehat{p}_1 - \widehat{p}_2 + z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} + \frac{n_1+n_2}{2n_1n_2} \tag{4}$$

The adjusted Wald interval, by definition, has a wider length than the Wald interval. This leads to an increasing the coverage probability, but at the same time, widening the interval leads to a loss of precision.

Agresti & Caffo (2000) proposed to combine the Wald interval and the score interval, due to Wilson (1927), by adding pseudo observations in order to increase the coverage probability. They found that the optimum number of pseudo observations to add is four: two successes and two failures, and they showed that the performance of the resulting Agresti-Caffo interval is surprisingly high even for small sample sizes. The limits of Agresti-Caffo interval are:

$$L_{low} = \widetilde{p}_1 - \widetilde{p}_2 - z_{1-\alpha/2}\sqrt{V(\widetilde{p}_1, \widetilde{n}_1) + V(\widetilde{p}_2, \widetilde{n}_2)} \tag{5}$$

and

$$L_{upp} = \widetilde{p}_1 - \widetilde{p}_2 + z_{1-\alpha/2}\sqrt{V(\widetilde{p}_1, \widetilde{n}_1) + V(\widetilde{p}_2, \widetilde{n}_2)} \tag{6}$$

with

$$V(\widetilde{p}_i, \widetilde{n}_i) = \frac{1}{\widetilde{n}_i}\left[\widetilde{p}_i - \widetilde{p}_i\frac{n_i}{\widetilde{n}_i} + \frac{1}{2\widetilde{n}_i}\right]$$

where $\widetilde{n}_i = n_i + 2$ for $i = 1, 2$, $\widetilde{p}_1 = \frac{\sum_{j=1}^{n_1} X_j + 1}{\widetilde{n}_1}$ and $\widetilde{p}_2 = \frac{\sum_{j=1}^{n_2} Y_j + 1}{\widetilde{n}_2}$.

Another confidence interval obtained by combining the Wald and the score interval is the Newcombe interval. To compute this interval, the following equation for each p_i should first be solved

$$|\widehat{p}_i - p_i| = z_{1-\alpha/2}\sqrt{\frac{p_i(1-p_i)}{n_i}}$$

Let's denote the solutions by l_i and u_i with $l_i < u_i$, $i = 1, 2$. The limits of the Newcombe interval are

$$L_{low} = \widehat{p}_1 - \widehat{p}_2 - z_{1-\alpha/2}\sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}} \tag{7}$$

and

$$L_{upp} = \widehat{p}_1 - \widehat{p}_2 + z_{1-\alpha/2}\sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{n_2}} \tag{8}$$

Newcombe found that this interval has good coverage and average length properties.

2.2. Bayesian intervals

Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities, such as predictions for new observations (Gelman et al. 2004). This process can be carried out by using Markov Chain Monte Carlo methods that simulate values from the posterior distribution of the parameter of interest[1]. Thus, we appeal to the Gibbs sampling algorithm to simulate values from the posterior distribution.

In order to implement a Gibbs sampling algorithm for the problem of finding a credibility interval for $p_1 - p_2$, we chose the prior distributions of p_1 and p_2 to be $Beta(a_1, b_1)$ and $Beta(a_2, b_2)$, respectively. Once the samples are drawn, the observed information is given by $x_1, \dots, x_{n_1}$ and $y_1, \dots, y_{n_2}$ or equivalently by $S_x = \sum_{j=1}^{n_1} x_j$ and $S_y = \sum_{j=1}^{n_2} y_j$. The posterior marginal distributions of p_1 and p_2 are obtained by Bayes theorem and are given by $Beta(a_1+S_x, b_1+n_1-S_x)$ and $Beta(a_2+S_y, b_2+n_2-S_y)$, respectively (Gelman et al. 2004, p. 34). Since the samples come from two independent populations, the posterior joint distribution of (p_1, p_2) is a product of its marginal distributions and, for this reason, one can get samples from the posterior distribution of $p_1 - p_2$ by simulating N values from the posterior distribution of p_1 and p_2, say $p_1^{(1)}, \dots, p_1^{(N)}$ and $p_2^{(1)}, \dots, p_2^{(N)}$, respectively. Then, by computing $p_1^{(1)} - p_2^{(1)}, \dots, p_1^{(N)} - p_2^{(N)}$, we obtain simulated values from the posterior distribution of $p_1 - p_2$. Note that the algorithm presented here generates independent samples from the posterior, so it is fair to name it as just a Monte Carlo algorithm, rather than a Markov Chain Monte Carlo algorithm.

After that, it is possible to compute the credibility interval[2] of $100 \times (1-\alpha)\%$ for $p_1 - p_2$ using the percentiles of the values simulated that induce the shortest credible intervals. In this research, we consider two noninformative priors for p_1 and p_2: $Beta(1,1)$ and $Beta(0.5, 0.5)$ priors. $Beta(1,1)$ corresponds to the uniform distribution, which provides the same weight along all values in the range $(0,1)$ for each p_i with $i = 1, 2$. When both priors of p_1 and p_2 are uniform priors, the prior distribution for the difference $p_1 - p_2$ is a triangular distribution with vertices $(-1, 0)$, $(1, 0)$ and $(0, 1)$. That is to say the prior distribution provides greater weight to values of $p_1 - p_2$ close to 0, and small weights to values close to the extremes -1 and 1.

The $Beta(0.5, 0.5)$ is known as the Jeffreys prior, which, according to Carlin & Louis (1998, p. 51), is noninformative in a transformation-invariate sense. However, it provides extra weight to extreme values of p_i, that is, values close to 0 and 1. When both priors of p_1 and p_2 are the Jeffreys prior, the prior distribution of $p_1 - p_2$ is symmetric at the value 0 where it is not defined, increasing for values

[1] In the case of estimating the difference of two proportions, the exact posterior distribution of $p_1 - p_2$ is given by Pham-Gia & Turkkan (1993). However, this exact distribution is somewhat complicated and computationally expensive to obtain.

[2] There are many ways to construct a Bayesian credible interval from the posterior distribution. A naive way to construct it is by using the upper and lower $\alpha/2$ quantiles. However, as the intervals are to be judged by expected length and its variance, it would make more sense to use the highest posterior density intervals which are, by definition, the shortest credible intervals with the given coverage (Carlin & Louis 1998, p. 43).

in $(0,1)$ and decreasing for values in $(-1,0)$. The explicit density function of the priori distribution of $p_1 - p_2$ when both priors of p_1 and p_2 are beta is studied in Pham-Gia & Turkkan (1993).

3. Comparison criteria

In this section, we establish some criteria in order to measure the performance of the intervals in a frequentist sense. A good confidence or credibility interval should have the true coverage probability close to or larger than the nominal value. Of course, in most cases, a way to increase the coverage probability is by widening the interval, obtaining intervals with little precision. The comparison of different methods for obtaining confidence intervals for one parameter must take into account their lengths. To accomplish this, mean and variance of those lengths are analyzed in this paper. In conclusion, we use the following criteria:

1. The true coverage probability defined by:

$$CP = E(I(X,Y,p_1,p_2)) \tag{9}$$

where X and Y denote the number of successes in n_1 and n_2 trails, respectively. $I(x,y,p_1,p_2)$ defines an indicator function that is equal to one if the interval contains $p_1 - p_2$ when $X = x$ and $Y = y$, and equal to zero if the interval does not contain $p_1 - p_2$. The coverage probability is given by:

$$CP = \sum_{x=0}^{n_1}\sum_{y=0}^{n_2} I(x,y,p_1,p_2)\binom{n_1}{x}p_1^x(1-p_1)^{n_1-x}\binom{n_2}{y}p_2^y(1-p_2)^{n_2-y} \tag{10}$$

2. The expected length defined by:

$$l = E(U(X,Y) - L(X,Y)) \tag{11}$$

where $U(X,Y)$ and $L(X,Y)$ are the upper and lower limit of the confidence or credibility interval for $p_1 - p_2$. Note that they are functions of the variables X and Y. The expected length is given by:

$$l = \sum_{x=0}^{n_1}\sum_{y=0}^{n_2}(U(x,y) - L(x,y))\binom{n_1}{x}p_1^x(1-p_1)^{n_1-x}\binom{n_2}{y}p_2^y(1-p_2)^{n_2-y} \tag{12}$$

3. Analogously, we define the variance of length by:

$$V = Var(U(X,Y) - L(X,Y)) \tag{13}$$

and it is easy to show that

$$V = \sum_{x=0}^{n_1}\sum_{y=0}^{n_2}(U(x,y)-L(x,y))^2\binom{n_1}{x}p_1^x(1-p_1)^{n_1-x}\binom{n_2}{y}p_2^y(1-p_2)^{n_2-y}$$
$$-\left(\sum_{x=0}^{n_1}\sum_{y=0}^{n_2}(U(x,y)-L(x,y))\binom{n_1}{x}p_1^x(1-p_1)^{n_1-x}\binom{n_2}{y}p_2^y(1-p_2)^{n_2-y}\right)^2 \tag{14}$$

Notice that these criteria are frequentist, in the sense that in (10), (12) and (14), the proportions p_1 and p_2 are assumed to be fixed values, rather than random variables.

4. Comparison among intervals

In this section, we compare several confidence and Bayesian credibility intervals with respect to coverage probability and mean and variance of their lengths. For confidence intervals (Wald, adjusted Wald, Agresti-Caffo and Newcombe), those values were exactly computed for several combinations of p_1, p_2 and different sample sizes. For Bayesian intervals, the computation was done by means of the simulation of samples of the posterior distributions of p_1 and p_2. These distributions were obtained through the Markov Chain approach, and prior distributions used for p_1 and p_2 were the same: $Beta(1,1)$ and $Beta(0.5,0.5)$. In subsection 4.1, the true coverage probability of 0.95 confidence level or credibility level of intervals are obtained for $p_2 = 0.5$, $p_1 \in (0,1)$ and $n_j \in \{10,50,100\}$, with $j = 1,2$, for the two priors described above. Subsequently, the mean and variance of the intervals were computed. Subsection 4.2 shows the same kind of study, with the same chosen values as in 4.1, except that n_2 is fixed at $n_2 = 30$. In 4.3, $n_1 \in \{1,2,\dots,500\}$, $n_2 = 30$ and $(p1-p2) \in \{0,0.1,0.5,0.8\}$.

4.1. Performance of intervals by varying sample sizes

We compare the performance of the confidence and credibility intervals for different sample sizes n_1 and n_2. First, we calculate the true confidence level for the confidence intervals as a function of p_1. The value p_2 is fixed as 0.5, the samples sizes of X and Y are assumed to be the same, and we consider the values $n_1 = n_2 = 10, 50, 100$. The resulting coverage probabilities for the Wald and adjusted Wald intervals are presented in Figure 1. It is seen that the coverage probability of the adjusted Wald interval is always larger than the Wald interval; this fact is intuitive since the adjusted Wald interval is obtained by widening the Wald interval. Additionally, the coverage probability is not affected by the different values of p_1. Also, the poor performance of the Wald interval is noted, especially in small samples.

On the other hand, the coverage probabilities of the Agresti-Caffo and Newcombe intervals are presented in Figure 2. It can be seen that both intervals have

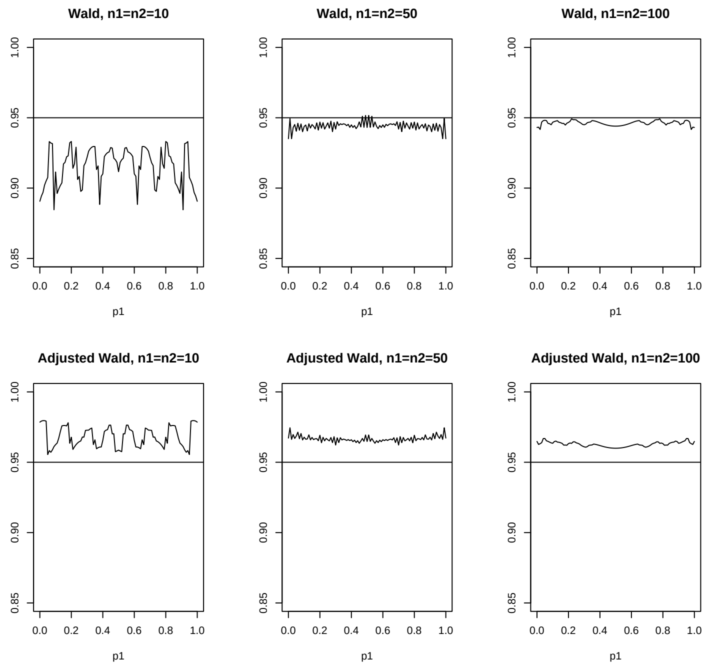

Figure 1: True coverage probability of the Wald and Adjusted Wald intervals varying p_1 with $n_1 = n_2 = 10, 50, 100$ with a nominal coverage probability of 0.95

coverage probability quite close to the nominal coverage 0.95, a desirable property that the Wald and adjusted Wald do not have. Although the adjusted Wald interval has coverage probability larger than 0.95, we will see later that its length is the largest. Also, the coverage probability of the Newcombe interval is sean to be affected by different values of p_1, especially when the samples are small.

The coverage probability for Bayesian intervals is presented in Figure 3, where it is seen that the performance of these two intervals are similar, and are quite good in the sense that the coverage probability is stable with respect to p_1, and is close to the nominal 0.95 even when the samples are small. So we can conclude that, in terms of true coverage probability, the Bayesian intervals are better than the frequentist intervals, without ignoring the notable performance of the Agresti-Caffo and Newcombe intervals. As a final remark, the true coverage probabilities of all the intervals considered become more stable with respect to p_1 as the sample sizes increases.

We now compare the intervals in terms of the expected length. The expected lengths of the considered intervals with different sample sizes are presented in Figure 4. It is be seen that the interval with largest length is the adjusted Wald interval. This shows that the high coverage probability is due to the length of the interval, but is not due to its good performance. The shape of the curve for the Wald interval is similar to the adjusted Wald; this is intuitive since the adjusted

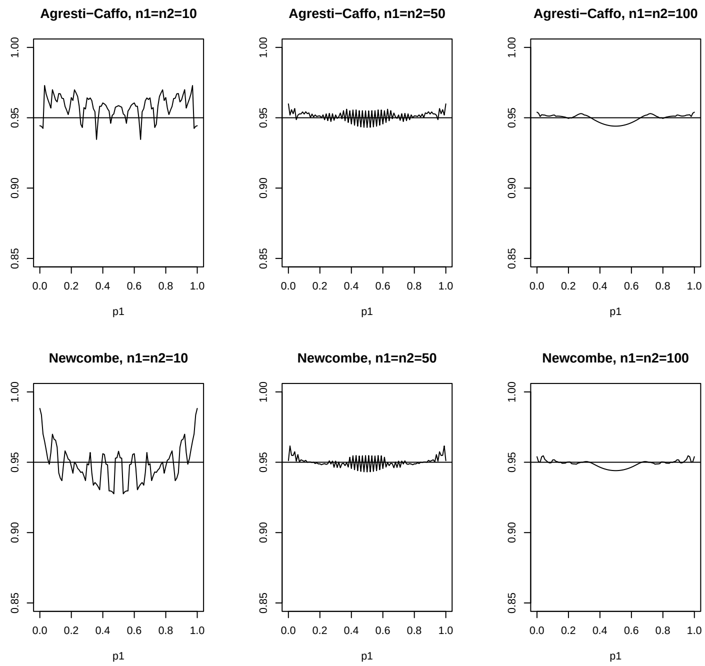

FIGURE 2: True coverage probability of the Agresti-Caffo and Newcombe intervals varying p_1 with $n_1 = n_2 = 10, 50, 100$ with a nominal coverage probability of 0.95

Wald interval is obtained by subtracting and adding a constant to the lower and upper limit of the Wald interval, respectively. As a result then, the following relationship between the lengths of these intervals remains:

$$l_{A.Wald} = l_{Wald} + \frac{n_1 + n_2}{n_1 n_2} \tag{15}$$

The Agresti-Caffo and Newcombe intervals have a more stable expected length with respect to p_1 than the Wald and adjusted Wald intervals. The improvement is noted especially in small samples. In samples with $n_1 = n_2 = 50, 100$, the length of the Agresti-Caffo and Newcombe intervals are smaller than the Wald and adjusted Wald intervals.

The expected lengths of the Bayesian intervals are also presented also in Figure 4, where it is seen that the performance of the intervals with the uniform and Jeffreys prior are similar. However, their expected lengths are larger than the Agresti-Caffo and Newcombe intervals when $n_1 = n_2 = 100$; when $n_1 = n_2 = 50$, the lengths are similar; when $n_1 = n_2 = 10$, the Bayesian intervals show a similar performance to the Newcombe interval while the Agresti-Caffo interval has a slightly larger expected length. In conclusion, the Newcombe interval has the smallest expected length in all sample sizes.

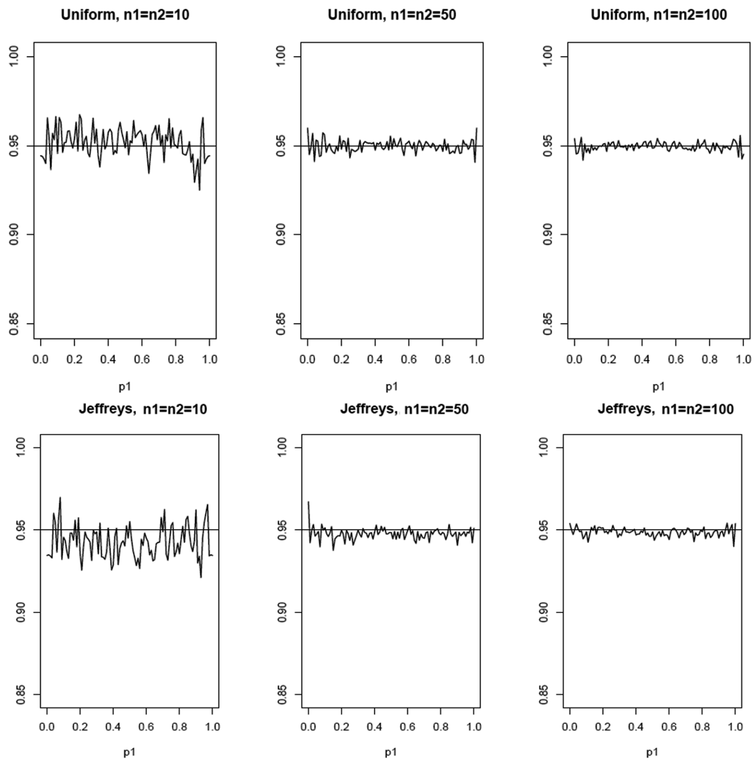

FIGURE 3: True coverage probability of the Bayesian intervals varying p_1 with $n_1 = n_2 = 10, 50, 100$ with a nominal coverage probability of 0.95

Finally, we compare the intervals in terms of variance of the length. Notice that the variance of the length of the adjusted Wald interval is equal to the Wald interval. Recalling (15) and using the property of variance, we have that $Var(l_{A.Wald}) = Var(l_{Wald})$. So in the figures related of the variance of the length, we only plot the variance of length for the Wald interval.

The variances of length for Wald/adjusted Wald, Agresti-Caffo and Newcombe intervals are presented in Figure 5. It is seen that the Newcombe interval has the smallest variance, although very close to the variance of the Agresti-Caffo interval. The huge variance of the Wald and adjusted Wald intervals in small samples is also seen. On the other hand, the variances of the Bayesian intervals are presented in Figure 6, and, the performance of the intervals with the uniform prior and the Jeffreys prior are similar. However, their variance is larger than both the Agresti-Caffo and Newcombe intervals.

In conclusion, in terms of true coverage probability, the best intervals are the Bayesian; in terms of the expected length, the best interval is the Newcombe interval, as well as in terms of variance of the length.

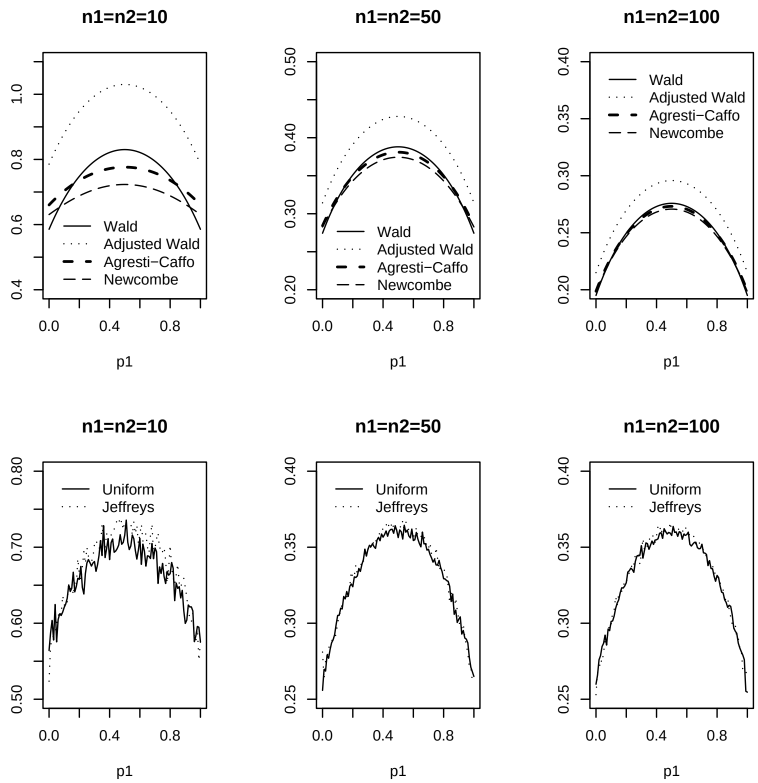

FIGURE 4: Expected length of confidence and Bayesian intervals varying p_1 with $n_1 = n_2 = 10, 50, 100$.

4.2. Performance of intervals varying values of p_1

In this section, we compare the performance of the intervals when different values of p_1 are considered.

First, we compute the true coverage probability as a function of n_1, the value of n_2 is fixed to be 30, the value of p_2 is 0.5, and we consider the values $p_1 = 0.01, 0.1, 0.3, 0.5$. The true coverage probability for the Wald and adjusted Wald intervals are presented in Figure 7. It is seen that, as in the previous section, the coverage probability of the adjusted Wald interval is always larger than the Wald interval. Additionally for the adjusted Wald interval, as the sample size n_1 increases, the coverage probability becomes more stable, while for the Wald interval, the increasing sample size does not improve the coverage probability when $p_1 = 0.01, 0.1, 0.3$.

In Figure 8, the coverage probabilities of the Agresti-Caffo and Newcombe intervals are presented. We see that the performance of the Newcombe interval is better than Agresti-Caffo interval as its coverage probability is more stable; when $p = 0.01, 0.1, 0.3$, it is always larger than the nominal 0.95, and when $p_1 = 0.5$, i

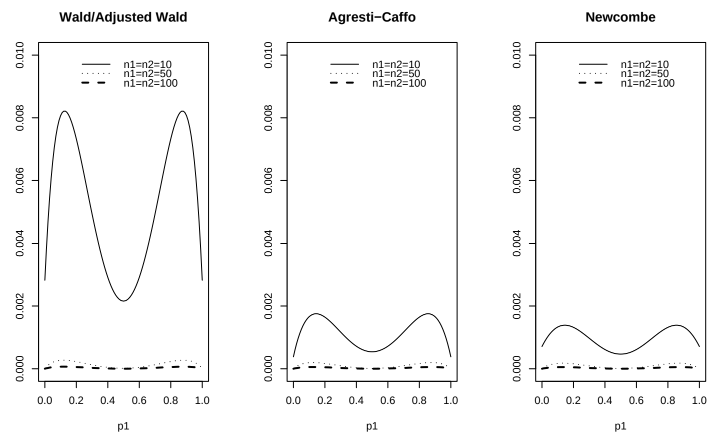

FIGURE 5: Variance of the length of the confidence intervals varying p_1 with $n_1 = n_2 = 10, 50, 100$.

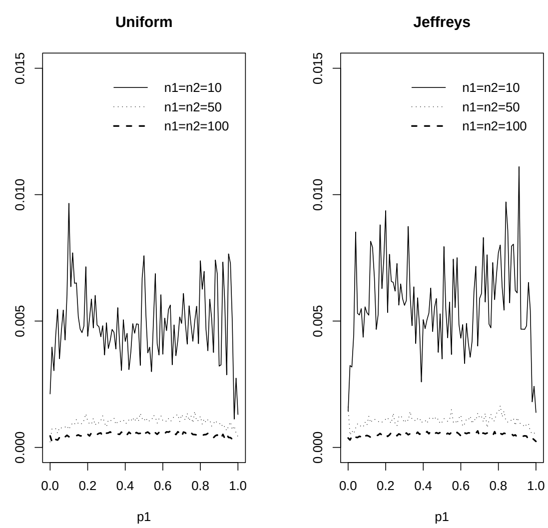

FIGURE 6: Variance of the length of the Bayesian intervals varying p_1 and $n_1 = n_2 = 10, 50, 100$.

tis very close to 0.95. Although the adjusted Wald interval has a larger coverage probability than the Newcombe interval, we will see later that this interval also has a larger expected length.

The results for the Bayesian intervals are those presented in Figure 9, where it is seen that for both intervals, the coverage probability is close to the nominal probability 0.95, and is not affected by different values of p_1; however, it is smaller than the adjusted Wald and Newcombe interval. In conclusion, the best intervals

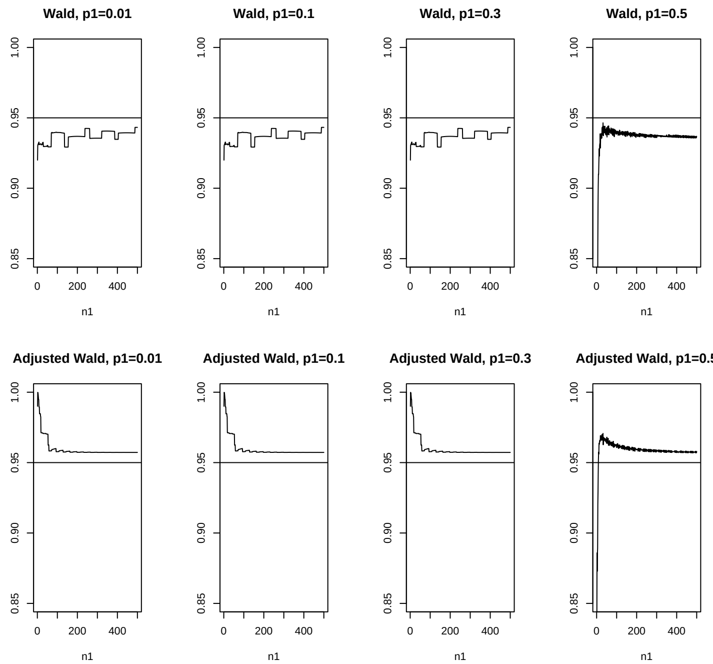

FIGURE 7: True coverage probability of the Wald and Adjusted Wald intervals varying n_1 and p_1 with a nominal coverage probability of 0.95.

in terms of the true coverage probability, are the adjusted Wald and Newcombe intervals.

We compare the intervals in terms of the expected length for different values of p_1. In Figure 10, the expected lengths of the Wald and adjusted Wald intervals are presented. Note that, as in the previous section, the expected length of the adjusted Wald interval is always larger. Thus we do not recommend this interval in spite of its large coverage probability. It is also noted that the lengths get smaller as the value of p_1 decreases and n_1 increases. The expected lengths of theAgresti-Caffo and Newcombe intervals are presented in Figure 11. It can be seen that their performance are very similar, although the length of the Newcombe interval is slightly smaller. In addition it is seen that their lengths are similar to the length of the Wald interval.

In Figure 12, the expected lengths of the Bayesian intervals are presented. It is seen that their performances are almost the same as the Agresti-Caffo and Newcombe intervals. In conclusion, except for the adjusted Wald interval, the performance of the other intervals in terms of the expected length is very similar.

We also compare the intervals considering the variance of the length. The performance of Wald and adjusted Wald intervals is presented in Figure 13. It is seen that for large sample sizes, the variance is almost zero. The variance of the Agresti-Caffo and Newcombe intervals is presented in Figure 14, where it is seen

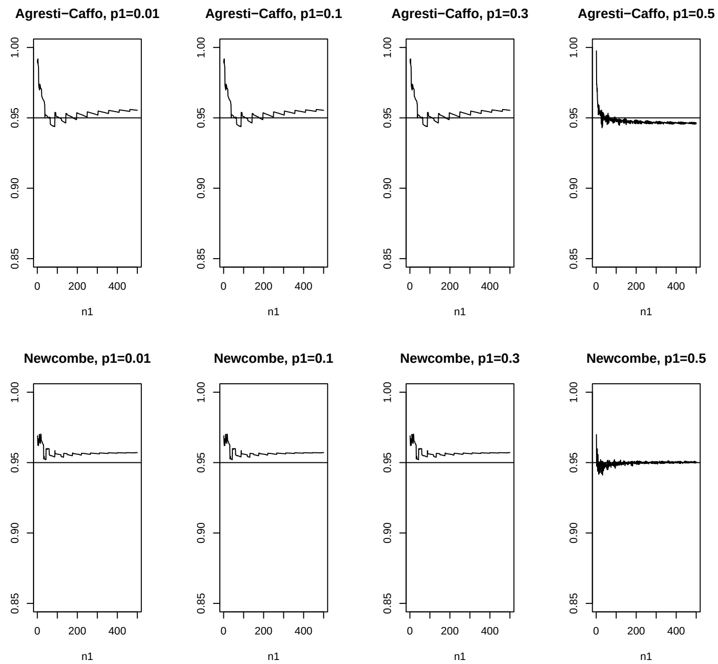

Figure 8: True coverage probability of the Agresti-Caffo and Newcombe intervals varying n_1 and p_1 with a nominal coverage probability of 0.95.

that when the sample size n_1 is small, the Newcombe interval always has a smaller variance than the Agresti-Caffo interval; while the difference is negligible when n_1 is large. At any rate, the variance of the Agresti-Caffo and Newcombe intervals is smaller than the Wald and adjusted Wald intervals.

In Figure 15, the variances for the Bayesian intervals are presented. Notice that there is no significant difference between the uniform and Jeffreys prior. However, their variances are smaller than the Wald and adjusted Wald intervals and larger than the Agresti-Caffo and Newcombe intervals. In conclusion, the interval with the smallest variance in length is the Newcombe interval.

4.3. Performance of intervals by varying values of p1 - p2

Since the parameter of interest is the difference between the proportions $p = p_1 - p_2$, it is natural to check the performance of the intervals when this parameter changes. Therefore, we calculate the true coverage probability of the intervals in the case that $p_1 - p_2$=0,0.1,0.5,0.8, the value of n_2 is fixed to be 30, and n_1 takes values $1, 2, \ldots, 500$.

The performance of the Wald and adjusted Wald intervals are presented in Figure 16, where we see that when the difference between p_1 and p_2 is large, the coverage probability of the Wald interval is really small. Further more, in

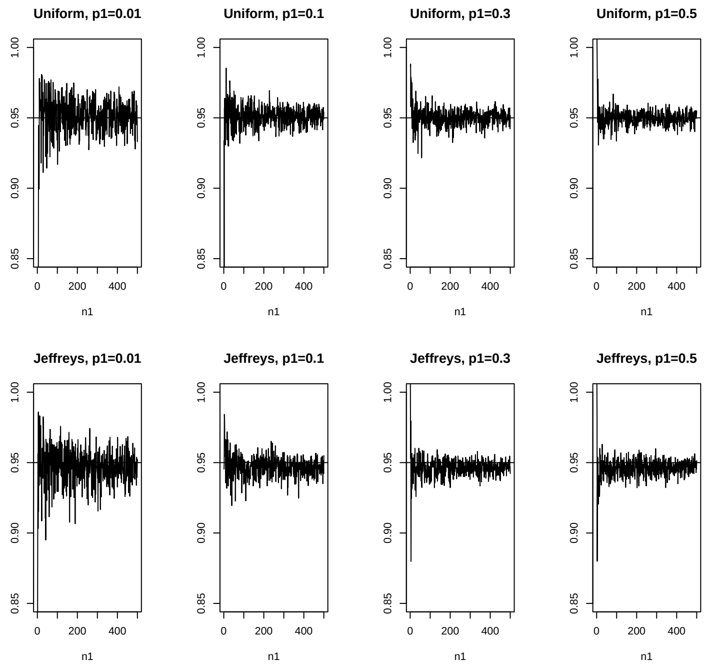

FIGURE 9: True coverage probability of the Bayesian intervals varying n_1 and p_1 with a nominal coverage probability of 0.95.

previous sections, the adjusted Wald always has larger coverage probability than the nominal 0.95, but in the case that $p_1 - p_2 = 0.8$, its coverage probability decreases considerably.

The coverage probabilities of the Agresti-Caffo and Newcombe intervals are presented in Figure 17, where we note that, contrary to the Wald and adjusted Wald intervals, the Agresti-Caffo and Newcombe intervals have larger coverage probability when $p_1 - p_2$ takes larger values. Regarding the Bayesian intervals, whose coverage probabilities are presented in Figure 18, we note that their performance is not affected by the values of $p_1 - p_2$, and that this is an advantage over the confidence intervals.

5. Other intervals

There are many other confidence intervals in statistical literature. Some of them will be briefly presented. Pan (2002) modified the Agresti-Caffo interval using the t distribution instead of the normal distribution to take of the uncertainty in estimating the variance of the observed pseudo proportion into account. It was found that in some situations the proposed method can have a higher coverage probability than the Agresti-Caffo interval. However, the price payed for the Pan interval is the resulting wider length of the intervals. The limits of this interval

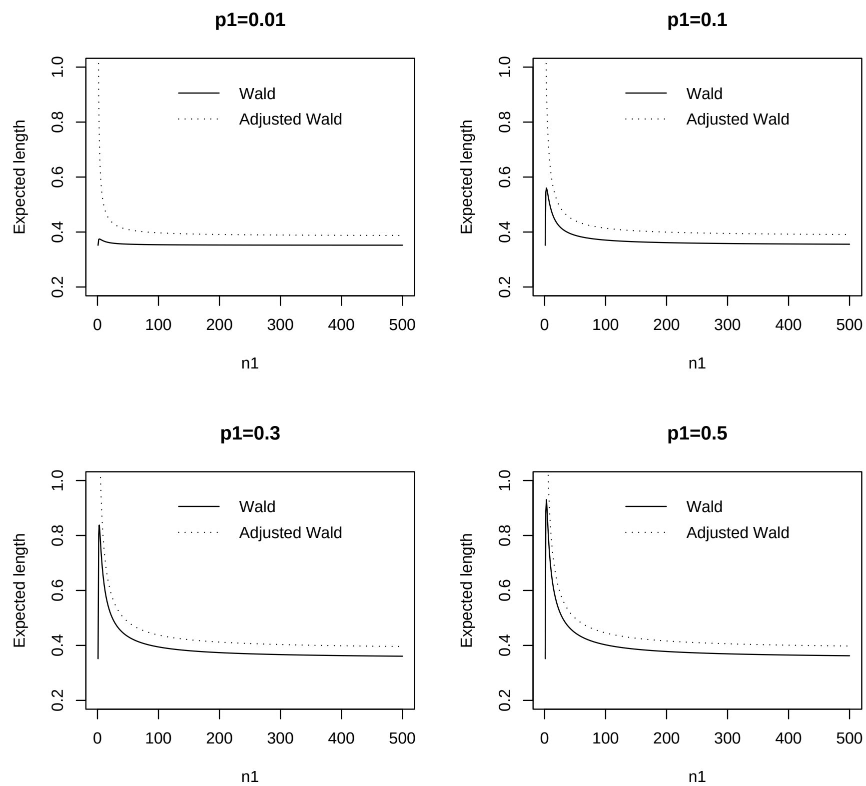

FIGURE 10: Expected length of the Wald and adjusted Wald intervals varying n_1 and p_1.

are:

$$L_{low} = \widetilde{p}_1 - \widetilde{p}_2 - t_{d,1-\alpha/2}\sqrt{V(\widetilde{p}_1, \widetilde{n}_1) + V(\widetilde{p}_2 + \widetilde{n}_2)} \tag{16}$$

and

$$L_{upp} = \widehat{p}_1 - \widehat{p}_2 + t_{d,1-\alpha/2}\sqrt{V(\widetilde{p}_1, \widetilde{n}_1) + V(\widetilde{p}_2 + \widetilde{n}_2)} \tag{17}$$

where

$$d \approx \frac{2[V(\widetilde{p}_1, \widetilde{n}_1) + V(\widetilde{p}_2 + \widetilde{n}_2)]}{\Omega(\widetilde{p}_1, \widetilde{n}_1) + \Omega(\widetilde{p}_2 + \widetilde{n}_2)}$$

and

$$\Omega(\widetilde{p}_i, \widetilde{n}_i) = \frac{\widetilde{p}_i - \widetilde{p}_i^2}{\widetilde{n}_i^3} + \widetilde{p}_i + (6\widetilde{n}_i - 7)\widetilde{p}_i^2 + 4(\widetilde{n}_i - 1)(\widetilde{n}_i - 3)\widetilde{p}_i^2 - \\ 2(\widetilde{n}_i - 1)\frac{(2\widetilde{n}_i - 3)\widetilde{p}_i^3}{\widetilde{n}_i^5} - \frac{2\widetilde{p}_i + (2\widetilde{p}_i - 3)\widetilde{p}_i^2 - 2(\widetilde{n}_i - 1)\widetilde{p}_i^3}{\widetilde{n}_i^4}$$

where $\widetilde{p}_i$ and $\widetilde{n}_i$ are similarly defined as in the Agresti-Caffo interval for $i = 1, 2$.

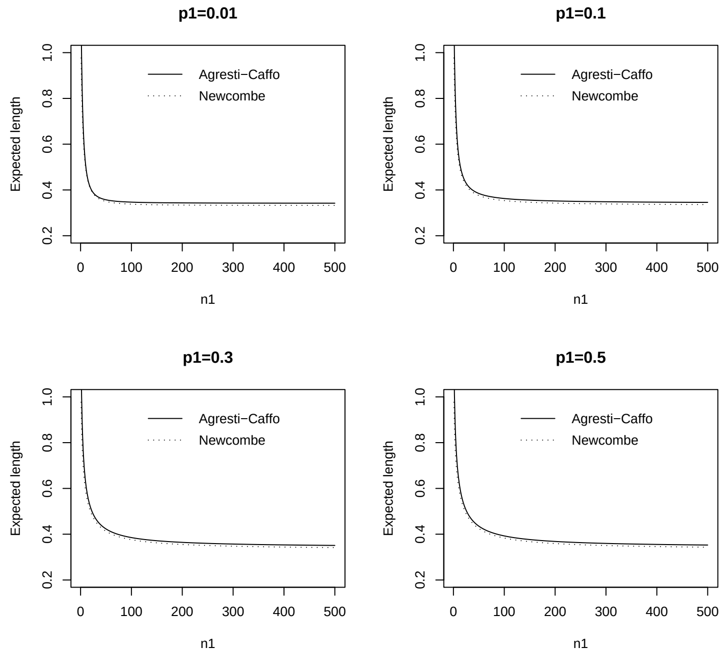

FIGURE 11: Expected length of the Agresti-Caffo and Newcombe intervals varying n_1 and p_1.

Miettinen & Nurminen (1985) proposed an asymptotic method based on the score test statistic, where the following system is considered:

$$H_0 : p_1 - p_2 = p^* \quad versus \quad H_1 : p_1 - p_2 \neq p^*$$

the score test statistic for testing this system is given by

$$S = \frac{\widehat{p}_1 - \widehat{p}_2 - p^*}{\sqrt{\widetilde{p}_1(1-\widetilde{p}_1)/n_1 + \widetilde{p}_2(1-\widetilde{p}_2)/n_2}} \tag{18}$$

where $\widetilde{p}_1$ and $\widetilde{p}_2$ are the maximum likelihood estimates of p_1 and p_2, respectively, under the restriction that $p_1 - p_2 = p^*$. The limits of the score interval L_{low} and L_{upp} are defined to satisfy:

$$1 - \Phi\left(\frac{\widehat{p}_1 - \widehat{p}_2 - L_{low}}{\sqrt{\widetilde{p}_1(1-\widetilde{p}_1)/n_1 + \widetilde{p}_2(1-\widetilde{p}_2)/n_2}}\right) = \Phi\left(\frac{\widehat{p}_1 - \widehat{p}_2 - L_{upp}}{\sqrt{\widetilde{p}_1(1-\widetilde{p}_1)/n_1 + \widetilde{p}_2(1-\widetilde{p}_2)/n_2}}\right) = \frac{\alpha}{2} \tag{19}$$

and the solution of L_{low} and L_{upp} must be found using numerical methods.

In addition, there is the Clopper-Pearson interval, which is strongly associated with the Clopper-Pearson test. However, many authors have criticized this interval for being too conservative in its coverage probability (Vos & Hudson 2008).

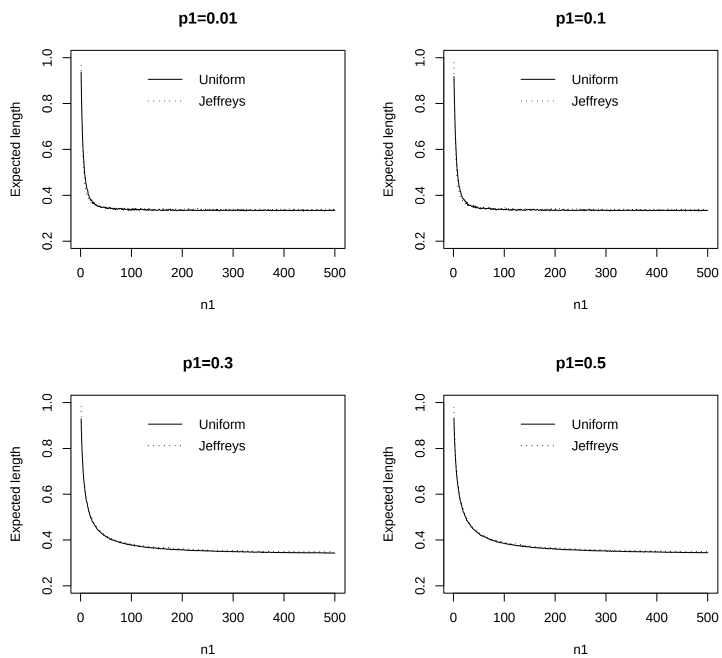

FIGURE 12: Expected length of the Bayesian intervals varying n_1 and p_1.

Another well-known interval is the Blaker interval. This interval has a smaller length than the Clopper-Pearson interval, i.e. it is always contained within the Clopper-Pearson intervals (Blaker 2000).

As we mentioned in Section 2, the exact Bayesian interval for $p_1 - p_2$ can be obtained using the exact posterior distribution. Pham-Gia & Turkkan (1993) established that when the prior distribution for p_i is $Beta(a_i, b_i)$ for $i = 1, 2$, the posterior distribution for $p = p_1 - p_2$ is given by

$$p(p \mid \mathbf{x}, \mathbf{y}) = \begin{cases} \frac{1}{k} B(\alpha_2, \beta_1) p^{\beta_1+\beta_2-1}(1-p)^{\alpha_2+\beta_1-1} \\ \qquad F_1(\beta_1, \alpha_1+\beta_1+\alpha_2+\beta_2-2, 1-\alpha_1, \beta_1+\alpha_2, 1-p1-p^2) & \text{for } \ 0 < p \leq 1 \\ \frac{1}{k} B(\alpha_1+\alpha_2-1, \beta_1+\beta_2-1) & \text{for } \ p = 0 \\ \frac{1}{k} B(\alpha_1, \beta_2)(-p)^{\beta_1+\beta_2-1}(1+p)^{\alpha_1+\beta_2-1} \\ \qquad F_1(\beta_2, 1-\alpha_1, \alpha_1+\beta_1+\alpha_2+\beta_2-2, \beta_2+\alpha_1, 1-p^2, 1+p) & \text{for } \ -1 \leq p < 0 \end{cases} \tag{20}$$

where $\mathbf{x} = (X_1, \ldots, X_{n_1})$ and $\mathbf{y} = (Y_1, \ldots, Y_{n_2})$.

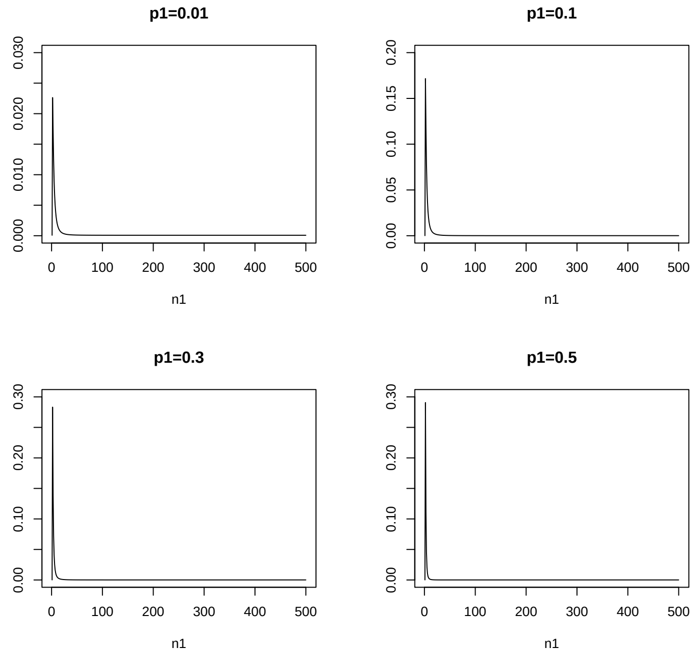

Figure 13: Variance of the length of the Wald and Adjusted Wald intervals varying n_1 and p_1.

$k = B(a_1, b_1)B(a_2, b_2)$, with $B(a, b)$ the beta function evaluated in a y b, that is,

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt \tag{21}$$

And $F_1(\varphi, \eta_1, \eta_2, \psi, w_1, w_2)$ is the fourth hypergeometric Appell's function, given by

$$\frac{\Gamma(\psi)}{\Gamma(\varphi)\Gamma(\psi-\varphi)} \int_0^1 u^{\varphi-1}(1-u)^{\psi-\varphi-1}(1-uw_1)^{-\eta_1}(1-uw_2)^{-\eta_2}du \tag{22}$$

when the real part of φ y $\psi - \varphi$ are all positive, for more details, see Bailey (1934).

Given the exact posterior distribution of $p = p_1 - p_2$, a Bayesian interval is defined by the lower limit l and upper limit u such that:

$$Pr(l \leq p \leq u \mid \mathbf{x}, \mathbf{y}) = 1 - \frac{\alpha}{2}$$

l and u are chosen to satisfy $Pr(p < l \mid \mathbf{x}, \mathbf{y}) = Pr(p > u \mid \mathbf{x}, \mathbf{y}) = \alpha/2$. Pham-Gia & Turkkan (1993) considered a numerical example where the prior distribution of p_1 and p_2 are $Beta(3, 5)$ and $Beta(2, 8)$, respectively, and the sampling results are $n_1 = 10$, $s_x = 4$, $n_2 = 6$ and $s_y = 2$. The resulting posterior distribution of $p_1 - p_2$ is bell-shaped, symmetric at the value 0.17, and an exact 90% credibility interval is $(-0.11, 0.39)$.

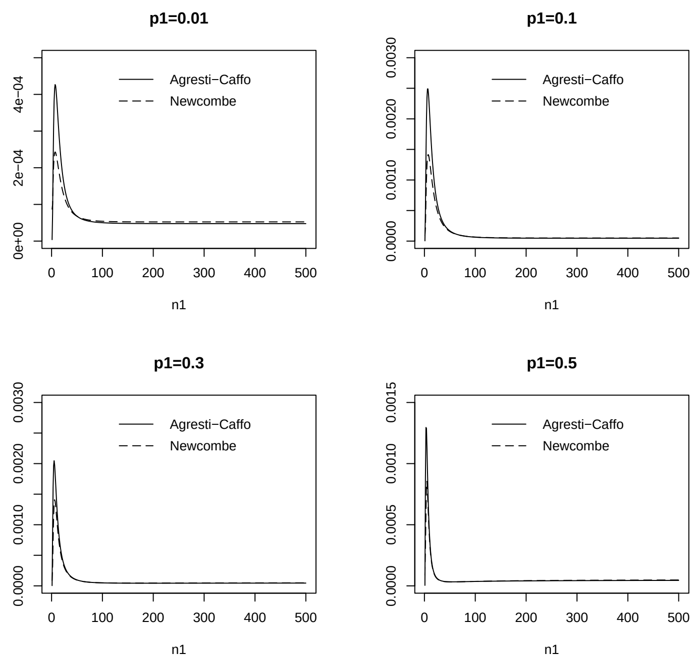

FIGURE 14: Variance of the length of the Agresti-Caffo and Newcombe intervals varying n_1 and p_1.

6. Conclusions

As a first conclusion, we point out that the performance of the Bayesian intervals is not greatly affected by the sample sizes nor by different values of p_1, p_2 or $p_1 - p_2$. In terms of true coverage probability, the best interval is the Bayesian interval, since its coverage probability is always close to the nominal coverage probability and is always stable with respect to different samples sizes. They are followed by the Newcombe and Agresti-Caffo intervals. We discard the use of adjusted Wald interval since its large coverage probability is obtained at the expense of a large length. The Wald interval performs poorly although this poor performance in small samples is a result that is well-known empirically and theoretically (Cepeda et al. 2008). In terms of expected length, the best interval is the Newcombe interval followed by the Agresti-Caffo interval, Bayesian intervals, and the Wald interval. The adjusted Wald interval always has largest length. In terms of the variance of length, the best interval is again the Newcombe interval, followed by the Agresti-Caffo interval, the Wald and adjusted Wald intervals. The intervals with the largest length variance are the Bayesian intervals, there fore the Newcombe interval is strongly recommend. The Wald and adjusted Wald intervals are not recommended.

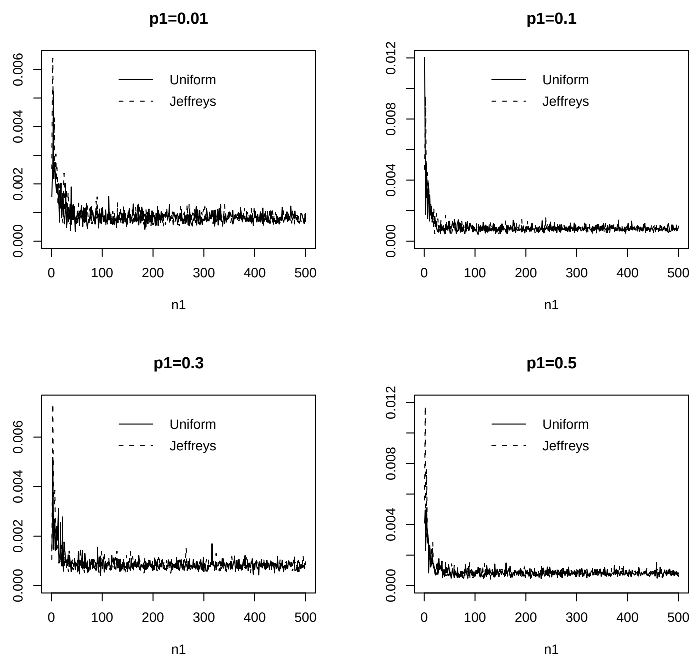

FIGURE 15: Variance of the length of the Bayesian intervals varying n_1 and p_1.

Acknowledgements

The authors are very grateful to professor Turkkan who kindly answered our inquires and to the anonymous referees for valuable suggestions.

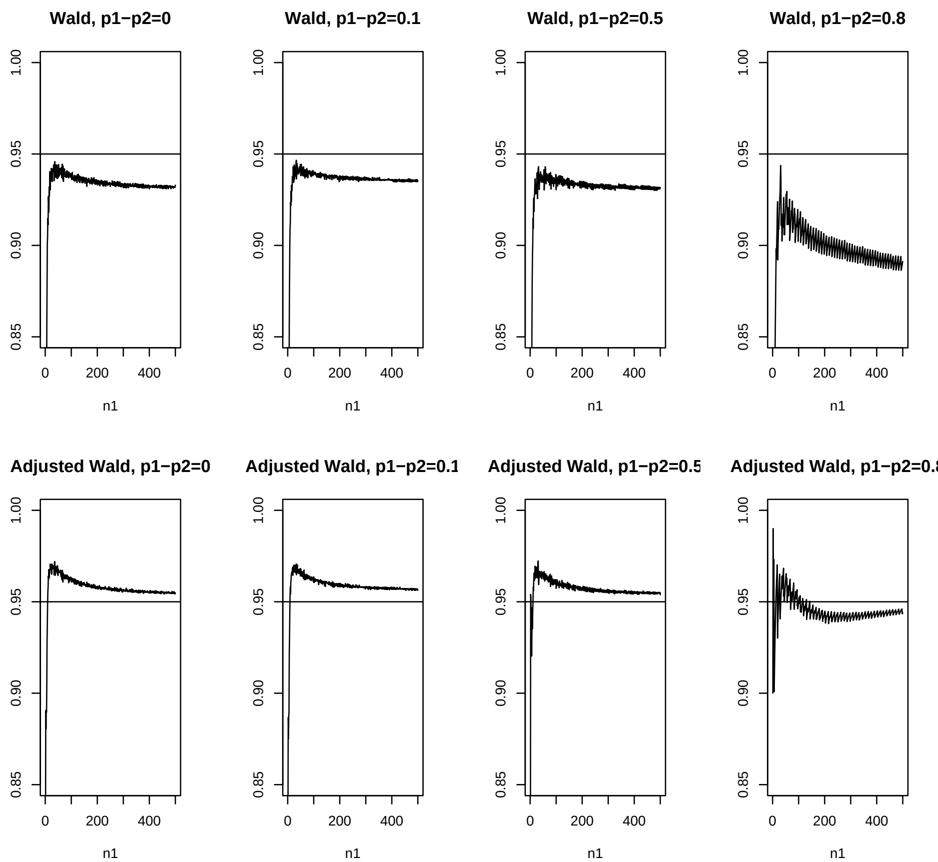

FIGURE 16: True coverage probability of the Wald and Adjusted Wald intervals varying n_1 and $p_1 - p_2$=0,0.1,0.5,0.8 with a nominal coverage probability of 0.95.

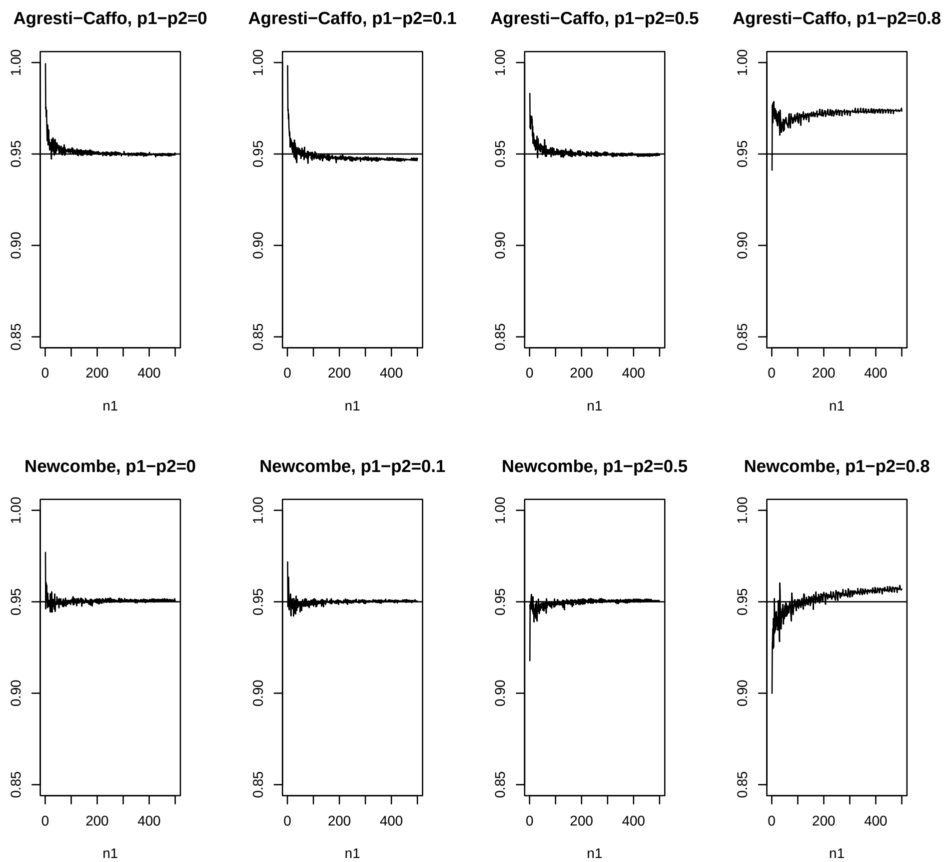

FIGURE 17: True coverage probability of the Agresti-Caffo and Newcombe intervals varying n_1 and $p_1 - p_2$=0,0.1,0.5,0.8 with a nominal coverage probability of 0.95.

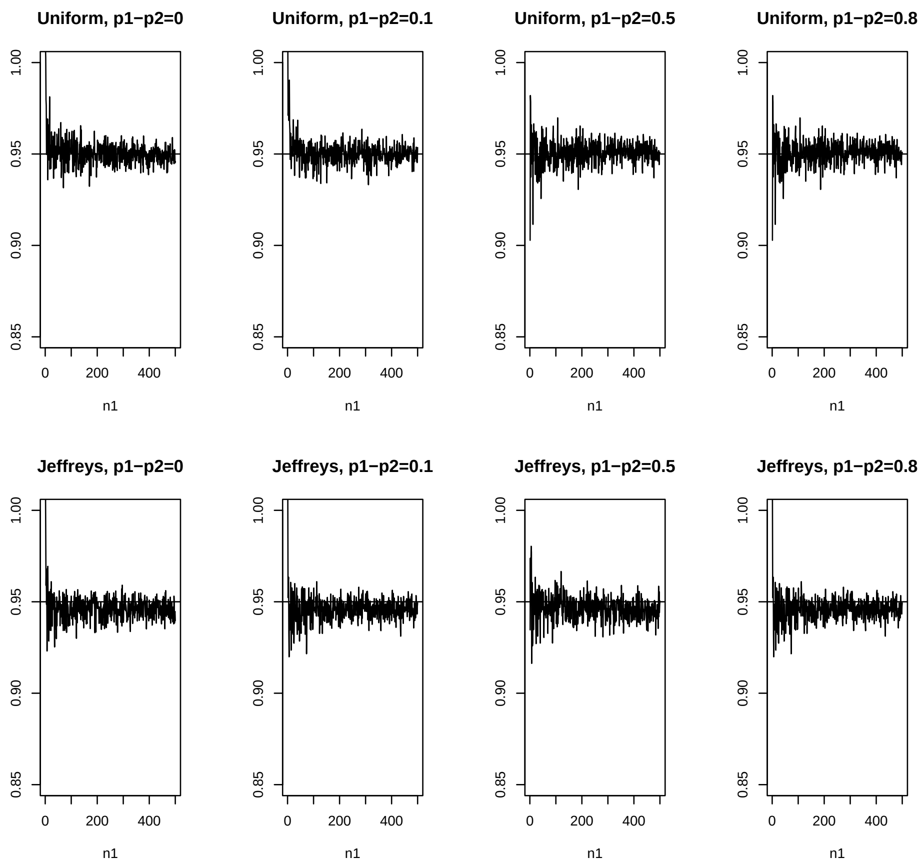

Figure 18: True coverage probability of the Bayesian intervals varying n_1 and $p_1 - p_2$=0,0.1,0.5,0.8 with a nominal coverage probability of 0.95.

[Recibido: julio de 2009 — Aceptado: marzo de 2010]

References

Agresti, A., Bini, M., Bertaccini, B. & Ryu, E. (2008), 'Simultaneous Confidence Intervals for Comparing Binomial Parameters', *Biometrics* **64**, 1270–1275.

Agresti, A. & Caffo, B. (2000), 'Simple and Effective Confidence Intervals for Proportions and Differences of Proportions', *American Statistician* **54**(4), 280–288.

Agresti, A. & Min, Y. (2005), 'Frequentist Performance of Bayesian Confidence Intervals for Comparing Proportions in 2×2 Contingency Tables', *Biometrics* **61**, 515–523.

Bailey, W. N. (1934), 'On the reducibility of Appell's Function F4', *The Quarterly Journal of Mathematics* **5**, 291–292.

Blaker, H. (2000), 'Confidence Curves and Improved Exact Confidence Intervals for Discrete Distributions', *The Canadian Journal of Statistics* **28**(4), 783–798.

Brown, L. D., Cai, T. T. & DasGupta, A. (2001), 'Interval estimation of a binomial proportion', *Statistical Science* **16**, 101–133.

Carlin, B. P. & Louis, T. A. (1998), *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall.

Cepeda, E., Aguilar, W., Cervantes, V., Corrales, M., Díaz, I. & Rodríguez, D. (2008), 'Intervalos de confianza e intervalos de credibilidad para una proporción', *Revista Colombiana de Estadística* **31**(2), 211–228.

Correa, J. C. & Sierra, E. (2003), 'Intervalos de confianza para la comparación de dos proporciones', *Revista Colombiana de Estadística* **26**(1), 61–75.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004), *Bayesian Data Analysis*, second edn, Chapman & Hall.

Ghosh, B. K. (1979), 'A Comparison of Some Approximate Confidence Intervals for the Binomial Parameter', *Journal of the American Statistical Association* **74**, 894–900.

Miettinen, O. S. & Nurminen, M. (1985), 'Comparative analysis of two rates', *Statistics in Medicine* **4**, 213–226.

Newcombe, R. (1998*a*), 'Interval Estimation for the Difference between Independent Proportions: Comparison of Eleven Methods', *Statistics in Medicine* **17**, 873–890.

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

Pan, W. (2002), 'Approximate Confidence Intervals for One Proportion and Difference of Two Proportions', *Computational Statistics and Data Analysis* **40**, 143–157.

Pham-Gia, T. & Turkkan, N. (1993), 'Bayesian Analysis of the Difference of Two Proportions', *Communications in Statistics. Theory and Methods* **22**(6), 1755–1771.

Vollset, S. E. (1993), 'Confidence intervals for a binomial proportion', *Statistics in Medicine* **12**, 809–824.

Vos, P. W. & Hudson, S. (2008), 'Problems with Binomial Two-Sided Tests and the Associated Confidence Intervals', *Australian & New Zealand Journal of Statistics* **50**(1), 81–89.

Wilson, E. B. (1927), 'Probable Inference, the Law of Succession, and Statistical Inference', *Journal of the American Statistical Association* **22**, 209–212.