Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Statistical Planning and Inference 130 (2005) 359–375

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Confidence intervals for two sample binomial distribution

Lawrence Brown[a,*], Xuefeng Li[b]

[a] *Statistics Department, Wharton School, University of Pennsylvania, 3730 Walnut Street., Philadelphia, PA 19104-6340, USA*

[b] *Division of Biostatistics, Center for Devices and Radiological Health, Food and Drug Administration, 1350 Piccard drive, Rockville, MD 20850, USA*

Received 22 August 2003; accepted 5 September 2003
Available online 29 July 2004

Abstract

This paper considers confidence intervals for the difference of two binomial proportions. Some currently used approaches are discussed. A new approach is proposed. Under several generally used criteria, these approaches are thoroughly compared. The widely used Wald confidence interval (CI) is far from satisfactory, while the Newcombe's CI, new recentered CI and score CI have very good performance. Recommendations for which approach is applicable under different situations are given.

Keywords: Confidence intervals; Binomial distribution; Two-sample problem; Wald interval; Newcombe's interval; Jeffrey's prior

1. Introduction

Suppose that X and Y are two independent random variables drawn from two different populations that both have binomial distributions. The first is of size m and has success probability p_1. The second is of size n and has success probability p_2. We are interested in comparing the difference of the success probability between these two populations.

We let $X \sim \text{binomial}(m, p_1)$ and $Y \sim \text{binomial}(n, p_2)$ and let $\Delta = p_1 - p_2$. We want to find the confidence interval with approximate level $1 - \alpha$ for Δ. When $\Delta = 0$, the related

[*] Corresponding author.
E-mail address: lbrown@wharton.upenn.edu (L. Brown).

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2003.09.039

testing problem is equivalent to the classical problem of testing independence in a 2 × 2 contingency table.

Because of its wide application in practice many approaches have been provided for this problem. However, most of them are concentrated on testing the independence hypothesis. Pearson (1947) proposed the χ^2 goodness-of-fit test, which is still widely used today. To improve its performance, Yates (1934) and Pearson (1947) gave different corrections or modifications to the χ^2 test. Fisher (1935) proposed the exact test. It is well known but less convenient for large sample sizes. See also Boschloo (1970) and Haber (1986). The likelihood ratio test was discussed by Wilks (1935). Freeman and Tukey (1950), Cox (1953) and Gart (1966) gave some other test statistics that have approximately χ^2 distribution with one degree of freedom. Barnard (1947) and Liddell (1976) suggested some tests in the spirit of ordering the sample space. Goodman (1964) explicitly gave a test statistic for the hypothesis $\mathrm{H}_0 : p_1 - p_2 = \Delta$. Some other tests, including Bayesian tests, are discussed in the literature. For instance, see Howard (1998), Tango (1998) and Agresti and Caffo (2000). Chernoff (2002) provides another interesting procedure on testing $p_1 = p_2$.

The reason we mention the above tests is because of the dual relationship between statistical tests and confidence sets. We can always obtain a confidence set for the parameter we are interested by inverting the family of tests . But we cannot always get a clear and convenient form for the confidence interval of Δ from these tests.

The well-known Wald confidence interval (CI) can be derived from Goodman's test, in which the standard errors are evaluated at the maximum likelihood estimates. Because of its simplicity and convenience, it has gained nearly universal application in practice and in textbooks. An alternative procedure is provided by the score test approach. This is based on inverting the test with the standard errors evaluated at the null hypothesis. Wilson (1927) gave the score CI for the proportion of one binomial population. For the difference of the proportions of two populations, we need an estimate of the standard error to avoid the nuisance parameter. A heuristic idea is to use the constrained maximum likelihood estimate of the standard error in the test. For simplicity we refer to the resulting CI as the score CI in this paper. It does not have an easily explained form. When $m = n$, Yule and Kendall (1950) obtained a CI by inverting the χ^2 test. For the case $m \neq n$ we propose a modified Yule's CI by changing the variance estimate. Newcombe (1998) gave a hybrid score interval by using information from the single score intervals for p_1 and p_2. Combining informative Bayesian estimates and the general procedure of inverting tests yields what we refer to as the Jeffrey's CI as a pseudo Bayesian CI. Inspired by the score test, Agresti and Caffo (2000) gave another pseudo Bayesian CI. Real Bayesian CIs are not fully explored here because of their computing difficulty. Finally, we propose a CI that is similar to the Wald CI but has a recentered coefficient. We call it the recentered CI.

After some exploration, we chose six representatives for comparison, Wald CI, Newcombe's CI, Jeffrey's CI, Agresti's CI, score CI and recentered CI. We compare their performance under some plausible criteria in order to give a broad picture for this problem. Recommendation is given for practical application.

We first give a brief summary of existing CIs in Section 2. Since the Bayesian CI is different from others, we consider it separately in Section 3. In Section 4, we propose the recentered CI. All the criteria used to compare the performance of these CIs are listed in Section 5. Section 6 gives empirical results for the comparisons and describes the various figures and tables that support the conclusions of the papers. Recommendation is given in Section 7. Some related problems are discussed in Section 8.

2. Some existing confidence intervals

Newcombe (1998) made a good summary of the problem and compared eleven methods, including most of the existing confidence intervals. We are not intending to do the same kind of work. Instead, we pick up some of them which either are widely used or have superior performance and compare them with some new methods, including our recentered CI and Agresti and Caffo's Bayesian CI. Following are the existing intervals which are used for comparison.

First we introduce some notation. Let $q_i = 1 - p_i,\ i = 1, 2,\ \hat{p}_1 = X/m,\ \hat{p}_2 = Y/n$, and let $\hat{q}_i = 1 - \hat{p}_i,\ i = 1, 2$. $\hat{p}_1$ and $\hat{p}_2$ are the MLEs of p_1 and p_2, respectively. Let $z_{\alpha/2}$ denote the upper $\alpha/2$ quantile of the standard normal distribution.

1. *The standard* (*Wald*) *CI*. Let $\hat{\Delta} = \hat{p}_1 - \hat{p}_2$, then $T = (\hat{\Delta} - \Delta)/\sigma_{\hat{\Delta}}$ asymptotically has standard normal distribution. Here $\sigma^2_{\hat{\Delta}}$ is some consistent estimate of $\mathrm{Var}(\hat{\Delta}) = p_1q_1/m + p_2q_2/n$. Substituting the MLE $\hat{p}_1\hat{q}_1/m + \hat{p}_2\hat{q}_2/n$ in T as the estimate of $\mathrm{Var}(\hat{\Delta})$, we get the Wald CI of Δ:
$$(\hat{\Delta} \pm z_{\alpha/2}\sqrt{\hat{p}_1\hat{q}_1/m + \hat{p}_2\hat{q}_2/n}). \tag{1}$$
The Wald CI has the most intuitive motivation. It also has a very simple form and is widely used in textbooks. Its performance can be somewhat improved in its coverage by use of some of the other CIs (see Figs. 1–8).
2. *Yule's CI*. In the above CI, we made no assumption about a target value Δ_0 for $p_1 - p_2$ when we were constructing the estimate of $\mathrm{Var}(\Delta)$. That is, we did not impose a constraint such as $\hat{p}_1 - \hat{p}_2 = \Delta$ on $\hat{p}_1$ and $\hat{p}_2$ in defining the estimation of $\mathrm{Var}(\Delta)$. Assuming $\Delta = 0$, $\bar{p} = (X+Y)/(m+n)$ is a better estimator of $p = p_1 = p_2$ and hence $(1/m + 1/n)\bar{p}\bar{q}$ is a more accurate estimate of $\mathrm{Var}(\Delta) = (1/m + 1/n)pq$. This yields Yule's CI
$$(\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{(1/m + 1/n)\bar{p}\bar{q}}). \tag{2}$$
We derived Yule's CI here under the assumption $\Delta = 0$. But it also performs reasonably well when $|\Delta|$ is not too big, especially when $m \approx n$.
3. *The modified Yule's CI*. We found Yule's CI performs well when $m = n$. If $m \neq n$, significant deviation of coverage probability from the nominal level appears when p_1 or p_2 is close to 0 or 1. So we need to make some modification to it. Using the weighted estimate $\check{p} = (nX/m + mY/n)/(m + n)$ instead of $\bar{p}$ we get the modified Yule's CI:
$$(\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{(1/m + 1/n)\check{p}\check{q}}). \tag{3}$$
This procedure has smaller bias when $m \neq n$. Note that when $m = n$ Yule's CI is actually a special case of modified Yule's CI. Similar to the situation for Yule's CI, the standard deviation in the Modified Yule's CI converges to the true values only if $\Delta = 0$ as $m \vee n \to \infty$.

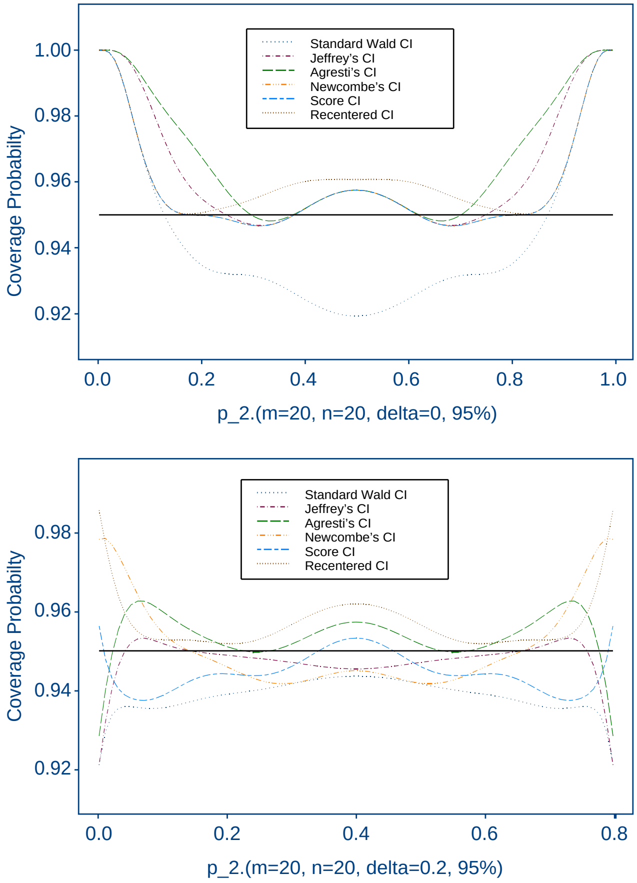

Fig. 1. Compare coverage when p is varying for $d = 0, 0.2$ and $m = n = 20$.

4. *The Newcombe's interval*. Using information from the single sample score intervals for p_1 and p_2, Newcombe (1998) gave a hybrid interval. Let (l_i, u_i) be the score confidence interval for p_i, that is, (l_i, u_i) are the roots for p_i in the quadratic equation $z_{\alpha/2} = (\hat{p}_i - p_i)/\sqrt{p_i(1-p_i)/n_i}$, $i = 1, 2,\ \ n_1 = m,\ \ n_2 = n$. The CI has the form

$$\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{l_1(1-l_1)}{m} + \frac{u_2(1-u_2)}{n}},\right.$$
$$\left.\hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{u_1(1-u_1)}{m} + \frac{l_2(1-l_2)}{n}}\right). \tag{4}$$

In some sense, the estimate of the Var$(\varDelta)$ in Newcombe CI is the average value of two boundary variances.

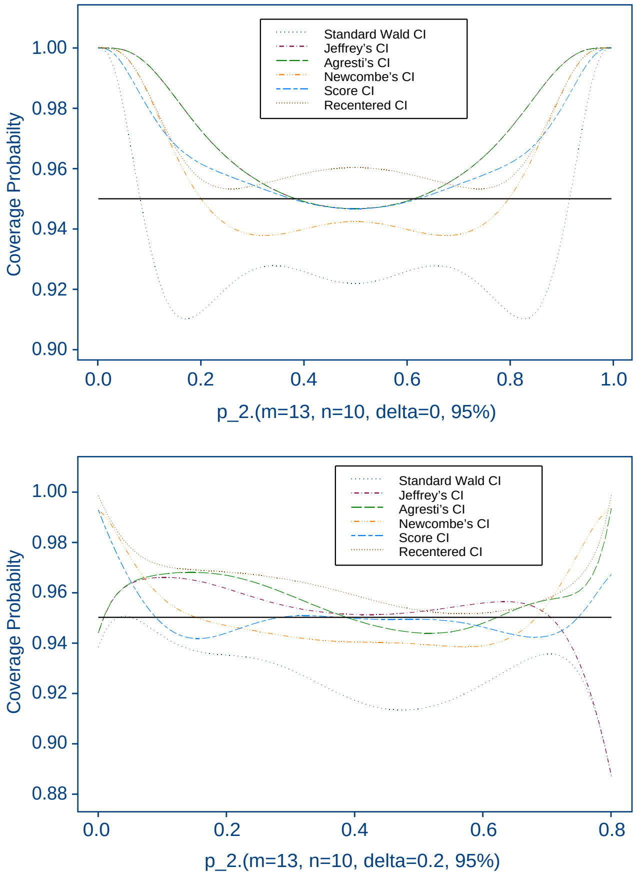

Fig. 2. Compare coverage when p is varying for $d = 0, 0.2$ and $m = 13$, $n = 10$.

5. *The score-test interval*. This interval is based on inverting the test

$$\frac{\hat{p}_1 - \hat{p}_2 - \Delta}{\hat{\sigma}} = z_\alpha. \tag{5}$$

But here $\hat{\sigma}$ is the estimate of the standard deviation of $\hat{p}_1 - \hat{p}_2$ under the constraint $\hat{p}_1 - \hat{p}_2 = \Delta$. More specifically,

$$\hat{\sigma}^2 = \frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{(\hat{p}_1 - \Delta)(1 - \hat{p}_1 + \Delta)}{n},$$

where $\hat{p}_1$ is the maximum likelihood estimate of p_1 under the constraint $\hat{p}_1 - \hat{p}_2 = \Delta$. It does not have a conveniently expressed form. We use the method of bisection to find $\hat{p}_1$.

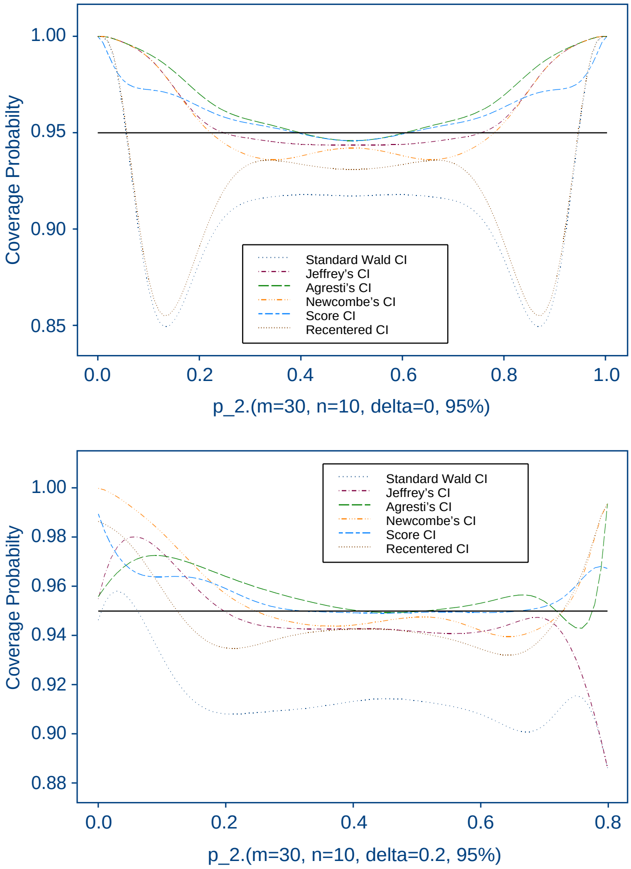

Fig. 3. Compare coverage when p is varying for $d = 0, 0.2$ and $m = 30, n = 10$.

In the case of just one binomial proportion the analogous procedure is due to Wilson (1927). There it has a more convenient, clearly expressed formula and performs quite well. A score-test interval for a different but related multinomial problem appears in Tango (1998).

3. Bayesian confidence intervals

3.1. General idea

Given a prior density $f(p_1, p_2)$ for p_1, p_2, we can construct the Bayesian Confidence intervals (HPD) for $\Delta = p_1 - p_2$ through a posterior distribution. Now the posterior density of p_1, p_2 is

$$l(p_1, p_2) = p_1^X (1 - p_1)^{m-X} p_2^Y (1 - p_2)^{n-Y} f(p_1, p_2).$$

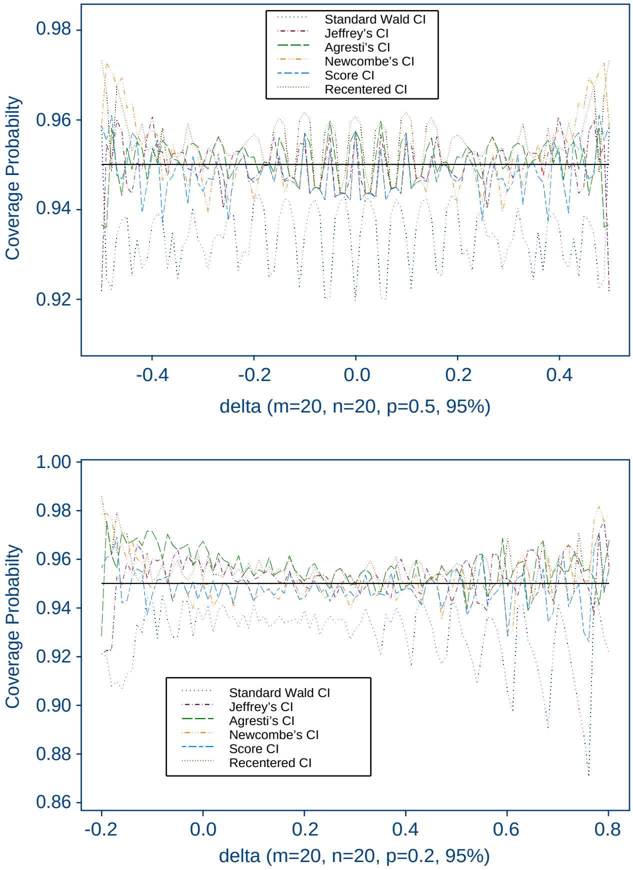

Fig. 4. Compare coverage when d is varying for $p = 0.5, 0.2$ and $m = n = 20$.

The lower bound can be obtained by solving the equations in u

$$\int_0^{1+u} l(p_1, p_1 - u)\,\mathrm{d}p_1 = \alpha/2 \quad \text{if} \quad \int_0^1 l(p_1, p_1)\,\mathrm{d}p_1 > \alpha/2,$$
$$\int_u^{1} l(p_1, p_1 - u)\,\mathrm{d}p_1 = \alpha/2 \quad \text{if} \quad \int_0^1 l(p_1, p_1)\,\mathrm{d}p_1 < \alpha/2.$$

Similarly, we can solve for the upper bound. Generally we cannot get a simple formula for these Bayesian CIs.

A widely used prior is the independent beta prior (see e.g. Howard (1998))

$$f(p_1, p_2) = Cp_1^{\alpha_1}(1 - p_1)^{\alpha_2} p_2^{\beta_1}(1 - p_2)^{\beta_2},$$

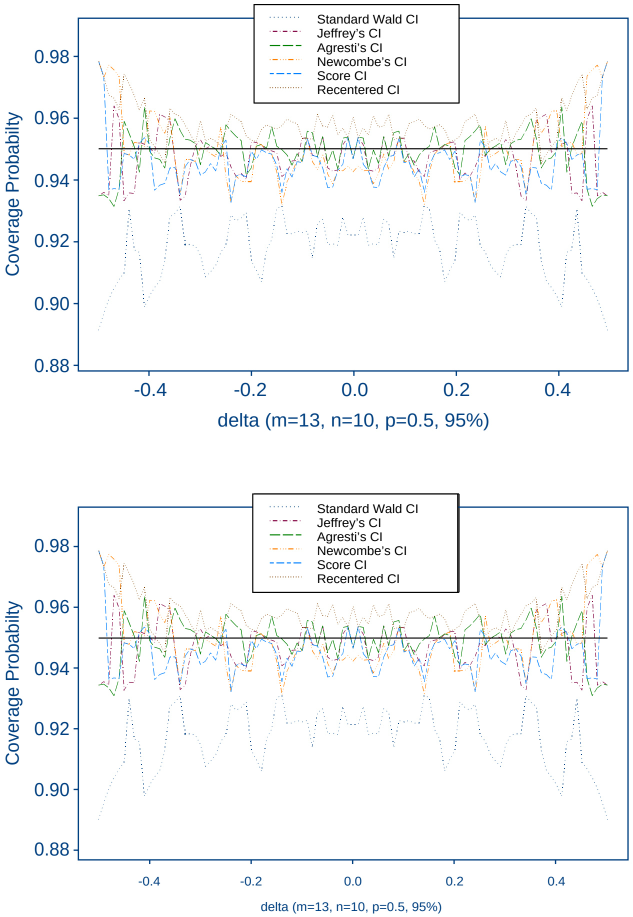

Fig. 5. Compare coverage when d is varying for $p = 0.5, 0.2$ and $m = 13, n = 10$.

where C is some constant. Howard also proposed a dependent prior

$$f(p_1, p_2) = C\, \mathrm{e}^{-(1/2)u^2} p_1^{\alpha-1}(1-p_1)^{\beta-1} p_2^{\gamma-1}(1-p_2)^{\delta-1},$$

where

$$u = \frac{1}{\sigma} \ln\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right).$$

3.2. Real and pseudo Bayesian CIs

- Jeffrey's CI
  In the one sample situation, the Bayes estimator $\tilde{p} = (X + 1/2)/(m+1)$ derived from Jeffrey's prior Beta$(1/2, 1/2)$ performs very well in constructing the CI of p, even when p is close to 0 or 1 (see Brown et al., 2001). Inspired by this fact, we use $\tilde{p}_1 = (X +$

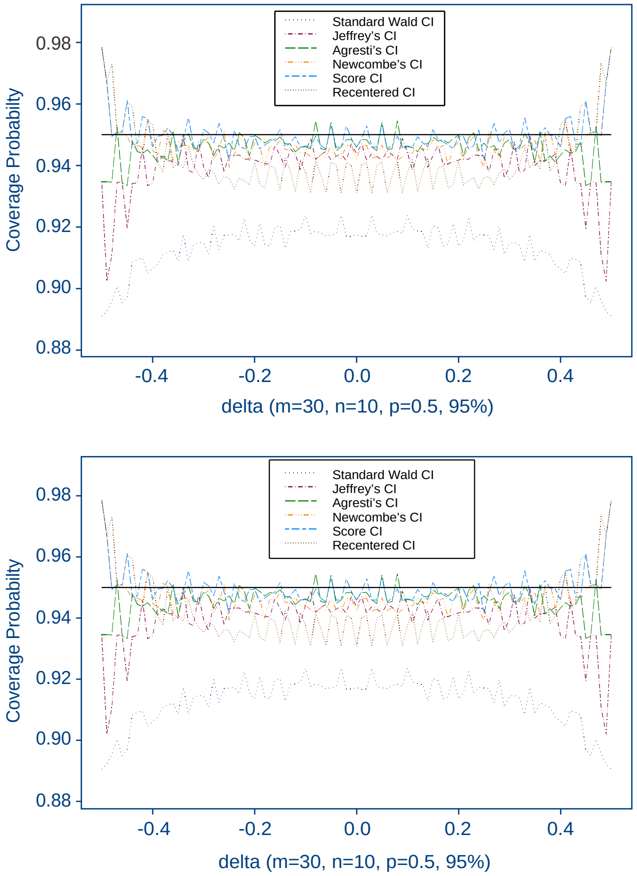

Fig. 6. Compare coverage when d is varying for $p = 0.5, 0.2$ and $m = 30$, $n = 10$.

$1/2)/(m+1)$ and $\tilde{p}_2 = (Y+1/2)/(n+1)$ instead of $\hat{p}_1$ and $\hat{p}_2$ as the estimates of p_1 and p_2 in the previous T statistic. We call this the Jeffrey's estimate CI. It has the form

$$(\tilde{p}_1 - \tilde{p}_2 \pm z_{\alpha/2}\sqrt{\tilde{p}_1\tilde{q}_1/m + \tilde{p}_2\tilde{q}_2/n}). \tag{6}$$

It has very good coverage performance, almost uniformly better than the Wald CI.

- Agresti's CI

  Leaning in the conservative direction, Agresti and Caffo (2000) propose $\acute{p}_1 = (X+1)/(m+2)$ and $\acute{p}_2 = (Y+1)/(n+2)$ in the above CI, i.e. adding one success and one failure for each sample. We call this procedure Agresti's CI.
- Approximate Jeffrey's CI

  Given an independent Jeffrey's prior, the posterior distributions of p_1, p_2 are

$$p_1|X \sim \text{Beta}(X+1/2, m-X+1/2), \quad p_2|Y \sim \text{Beta}(Y+1/2, n-Y+1/2).$$

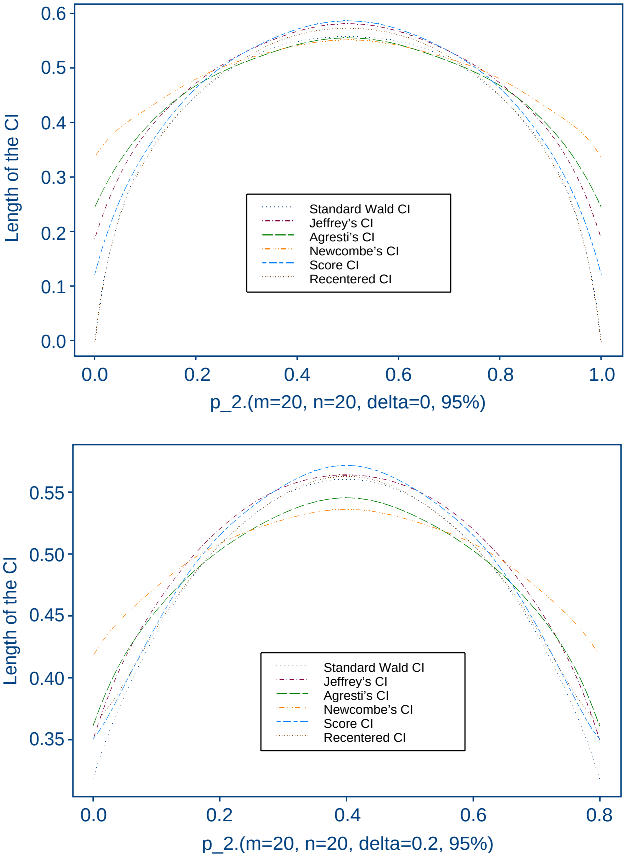

Fig. 7. Compare width when p is varying for $d = 0, 0.2$ and $m = n = 20$.

By normal approximation

$$p_1|X \sim \mathrm{N}\left(\tilde{p}_1, \frac{\tilde{p}_1(1-\tilde{p}_1)}{m+2}\right), \quad p_2|Y \sim \mathrm{N}\left(\tilde{p}_2, \frac{\tilde{p}_2(1-\tilde{p}_2)}{n+2}\right),$$

where $\tilde{p}_1 = (X + 1/2)/(m + 1)$ and $\tilde{p}_2 = (Y + 1/2)/(n + 1)$. Hence we obtain an approximate Jeffrey's CI which has the form

$$(\tilde{p}_1 - \tilde{p}_2 \pm z_{\alpha/2}\sqrt{\tilde{p}_1\tilde{q}_1/(m+2) + \tilde{p}_2\tilde{q}_2/(n+2)}). \tag{7}$$

It is uniformly shorter than Jeffrey's CI and does not perform well when the sample size is small.

- Real Jeffrey's interval
  All above three CIs are not real Bayesian CIs. They only use Bayesian estimates in constructing the CIs. We may call them pseudo Bayesian CI. Difficulty of calculation

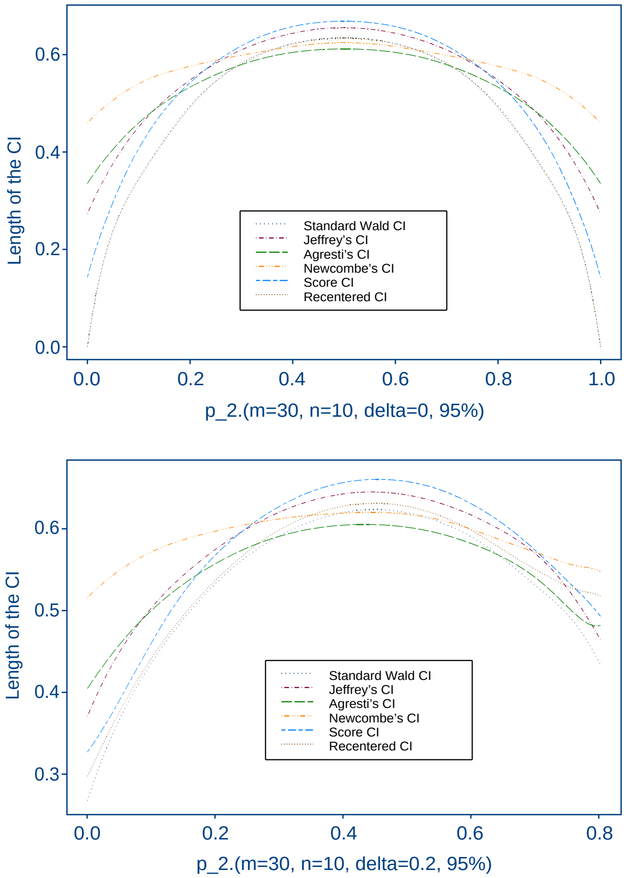

Fig. 8. Compare width when p is varying for $d = 0, 0.2$ and $m = 30$, $n = 10$.

hinders the convenient application of a real Bayesian CI. However, using numerical integration and the power of the computer, we tried the real Jeffrey's CI. The prior is

$$f(p_1, p_2) = Cp_1^{1/2}(1 - p_1)^{1/2}p_2^{1/2}(1 - p_2)^{1/2}$$

and the posterior is

$$l(p_1, p_2) = p_1^{X+1/2}(1 - p_1)^{m-X+1/2}p_2^{Y+1/2}(1 - p_2)^{n-Y+1/2}.$$

Results show that this CI does not perform better in terms of coverage than the pseudo Jeffrey's CI described above.

4. Recentered CI

In constructing the Wald confidence interval of p for a one sample binomial proportion, we simply invert the test by solving the equation $(\hat{p}-p)/\sqrt{p(1-p)/n}=z_{\alpha/2}$ in p. But for two binomial population, the same procedure is no longer applicable, since there is a nuisance parameter p_1 (or equivalently p_2). However, with reparametrization and some reasonable approximation, we can still construct tests and invert them to form confidence intervals.

Without loss of generality, we assume $\varDelta \geqslant 0$. Introduce a new parameter p defined as $p=(np_1+mp_2)/(m+n)$. Then $p_1=p+\Delta m/(m+n)$, $p_2=p-\Delta n/(m+n)$ and let $q=1-p$. A natural estimate of p is

$$\hat{p}=\frac{\hat{p}_1/m+\hat{p}_2/n}{1/m+1/n}=\frac{n\hat{p}_1+m\hat{p}_2}{m+n}.$$

By simple calculation, the variance of $\hat{\varDelta}$ is

$$\mathrm{Var}(\hat{\varDelta})=\left(\frac{1}{m}+\frac{1}{n}\right)pq-\frac{\varDelta^2}{m+n}.$$

Hence the related estimate of $\mathrm{Var}(\hat{\varDelta})$ is

$$\hat{\sigma}^2_{\varDelta}=\left(\frac{1}{m}+\frac{1}{n}\right)\hat{p}\hat{q}-\frac{\hat{\varDelta}^2}{m+n}.$$

It is easy to show that $\hat{\varDelta}$ is asymptotically orthogonal to $\hat{p}$, i.e. the following two statistics

$$Z_1=\frac{\hat{\varDelta}-\varDelta}{\sqrt{\mathrm{Var}(\hat{\varDelta})}},\qquad Z_2=\frac{\hat{p}-p}{\sqrt{\mathrm{Var}(\hat{p})}}$$

are asymptotically independent when $\min(n,m)$ goes to infinity. We reject H$_0$: $p_1-p_2=\varDelta$ if $|\hat{\varDelta}-\varDelta|\geqslant\kappa\hat{\sigma}_{\varDelta}$, where κ is upper $\alpha/2$ quantile of t distribution with degree of freedom $m+n-2$. Actually we do not have theoretical justification for using the t quantile instead of z quantile. The empirical results show that using t does much better than using z when n and m are small. Inverting this test, i.e. by solving

$$|\hat{\varDelta}-\varDelta|=\kappa\hat{\sigma}_{\varDelta}$$

we can get the corresponding CI.

But considering the definition, p must be less than $1-\Delta m/(m+n)$ and bigger than $\Delta n/(m+n)$. Consequently, $\hat{p}$ should satisfy the same conditions (these conditions can ensure $\hat{\sigma}_{\varDelta}>0$, but the inverse is not true). This consideration leads to the truncated estimate

$$\tilde{p}=\begin{cases}\Delta n/(m+n) & \text{if } \hat{p}<\Delta n/(m+n)\\ \hat{p} & \text{if } \Delta n/(m+n)\leqslant\hat{p}\leqslant 1-\Delta m/(m+n)\\ 1-\Delta m/(m+n) & \text{if } \hat{p}>1-\Delta m/(m+n)\end{cases}$$

instead of $\hat{p}$. The final version of this CI has the form

$$\left(\frac{\hat{\Delta}}{1+\kappa^2/(m+n)} \pm \frac{\kappa\sqrt{(1+\kappa^2/(m+n))(1/m+1/n)\tilde{p}\tilde{q} - \hat{\Delta}^2/(m+n)}}{1+\kappa^2/(m+n)} \right). \quad (8)$$

We call it the recentered CI, since it is centered by a value $(1+\kappa^2/(m+n))^{-1}$. Though it seems complex, it does have an explicit form, and it performs pretty well.

5. Criteria of comparison

There are several commonly used criteria for evaluating the performance of CIs.

1. Coverage probability

   - The average coverage should be close to the nominal level (generally we use 95%; 90% and 99% are also considered).
   - The region of poor coverage (e.g., less than 0.93) should be small.
   - The absolute deviance or the square root of the L_2 distance of coverage probability

     $$D_{\mathrm{A}} = \int |\text{coverage} - \alpha| \quad \text{or} \quad D_{\mathrm{S}} = \left(\int (\text{coverage} - \alpha)^2 \right)^{1/2}$$

     should be small.

2. The expected length of the CI should be as small as possible so long as the coverage probability generally is greater than or near to the nominal level.
3. The coverage converges to the nominal level uniformly and quickly with the increasing of the sample size, especially when p_1 or p_2 is close to 0 or 1.
4. Simplicity, i.e. easy to remember, easy to calculate, easy to understand and easy to present.
5. The CI is reasonable, i.e. it is within the domain of Δ.

In the next section, we mainly focus on the first two standards to describe the performance of those CIs.

6. Empirical results

Our main objective is to compare the CIs under different situations and provides some advice in application. A lot of comparison work is needed for these CIs, since there are four parameters in this problem. Here we precisely calculate the coverage probability and expected widths of the CIs on a grid of parameter values.

We compare the performance of all the CIs described above under various situations and three different confidence levels $\alpha = 0.01, 0.05, 0.1$. Yule’s CI is dominated by the

Table 1
Performance under different circumstances

	Wald	Jeffrey	Agresti	Newcombe	Score	Recenter
$m=n\leqslant 25$ and $\varDelta\approx 0$	−	+	o	+	+	+
$m=n\leqslant 25$ and $\varDelta\neq 0$	−	o	+	+	o	+
$m=n>25$ and $\varDelta\approx 0$	o	o	o	+	+	+
$m=n>25$ and $\varDelta\neq 0$	−	+	+	+	+	+
$m\neq n\leqslant 25$ and $\varDelta\approx 0$	−	o	o	+	+	−
$m\neq n\leqslant 25$ and $\varDelta\neq 0$	−	o	+	+	+	+
$m\neq n>25$ and $\varDelta\approx 0$	−	+	+	+	+	−
$m\neq n>25$ and $\varDelta\neq 0$	−	+	+	+	+	+

Remark: Here $\varDelta\approx 0$ means $|\varDelta|<0.1$. “+” ,“o” and “−” stand for good, acceptable and poor, respectively. The comparison is based mainly on the coverage probability for $p\in(0,1)$.

Table 2
Coverage: $m=n=10$, p is changing, $\varDelta=0, 0.2$, $1-\alpha=0.95$

CI	Average error		Cov.prob. <0.93		D_{A}		D_{S}	
	$\varDelta=0$	$\varDelta=0.2$	$\varDelta=0$	$\varDelta=0.2$	$\varDelta=0$	$\varDelta=0.2$	$\varDelta=0$	$\varDelta=0.2$
Wald	−0.023	−0.025	0.722	0.679	0.037	0.025	0.039	0.027
Jeffrey	0.023	−0.003	0.000	0.148	0.023	0.011	0.027	0.018
Agresti	0.023	0.018	0.000	0.000	0.023	0.018	0.027	0.020
Newcombe	0.011	0.004	0.000	0.000	0.018	0.013	0.025	0.017
Score	0.023	0.009	0.000	0.000	0.023	0.009	0.027	0.012
Recentered	0.023	0.018	0.000	0.000	0.023	0.018	0.027	0.020

Remark: Cov.prob. <0.93 gives the percentage of the grid points for which coverage probability are less than 0.93. D_{A} is the average of absolute deviance and D_{S} is the square root of the average of square error.

modified Yule’s CI. Furthermore, the modified Yule’s CI does not do well enough when $m\neq n$ compared to some other CIs. For simplicity, we ignore these CIs and only present the results of comparing the performance of the following six CIs: Wald CI, Jeffrey’s CI, Agresti’s CI, Newcombe’s CI, score CI and recentered CI. And in the figures we present, we always take the confidence level to be 95%.

First we examined the coverage probability. Eight subgroups are considered. The division is based on whether $m=n$ or $m\neq n$, m and n are small or m and n are large, $\varDelta\approx 0$ or $\varDelta\neq 0$. More specifically, we say m is small if $m\leqslant 25$, $\varDelta\approx 0$ if $|\varDelta|<0.05$. It is of course not a complete collection. The dividing point 25 and 0.05 are empirical. We drew a lot of plots within each group and the visual results are summarized in Table 1. By simply counting the number of “+” and “o”, we can see that Newcombe appears best, with the score, Agresti, recentered and Jeffrey are generally acceptable. In Figs. 1–6 we provide a few representative plots to display typical results supporting the summary in Table 1.

To verify the results numerically, we list some statistics based on coverage probability in Table 2. They are average coverage deviance, the percentage that the coverage probability is less than 0.93, the average of absolute deviance and the average of square error. Heuristically,

Table 3
Width: p is changing, $\varDelta = 0, 0.2$, $1-\alpha = 0.95$

CI	$m = m = 10$		$m = n = 30$		$m = 15, n = 10$	
	$\varDelta = 0$	$\varDelta = 0.2$	$\varDelta = 0$	$\varDelta = 0.2$	$\varDelta = 0$	$\varDelta = 0.2$
Wald	0.57	0.66	0.36	0.40	0.54	0.60
Jeffrey	0.68	0.71	0.38	0.41	0.61	0.65
Agresti	0.65	0.68	0.38	0.40	0.60	0.62
Newcombe	0.72	0.73	0.39	0.41	0.64	0.66
Score	0.64	0.70	0.37	0.41	0.57	0.64
Recentered	0.59	0.68	0.36	0.40	0.54	0.62

the bigger these statistics are, the poorer the CI is. We can see that all the statistics of the Wald CI are unacceptable. The rest of the CIs are doing well, especially the Newcombe CI.

Second we compare the expected widths of these CIs. Table 3 gives the average widths of these CIs under four special cases. See Figs. 7–8 for details. Wald and recentered CI have the smallest width while Newcombe CI is the widest. The other four are in the middle.

Finally, we compare the conservativity and boundary property. Regarding conservativity, we can see that in many cases, Wald CI is well below the nominal level and the Jeffrey and Agresti CI are above the nominal level, while all other CIs are fine. Surprisingly, the coverage of all CIs go to one when p is close to 0 and 1 except for some $m \neq n$ cases.

In summary, we have the following observation from those figures and tables:

- When m and n are small, the Wald CI is below the nominal level. But it has small width.
- In some cases where n and m are small, Jeffrey and Agresti's CI are a little bit conservative. And their coverage is generally above the nominal level.
- In most of the cases, the score, Newcombe and recentered CI are similar with each other and do a good job.
- When m and n are large, say, $m \wedge n \geqslant 50$, all the CIs are doing well. However, for the case that m and n have large common factor and $\varDelta$ is small, the recentered CI is not good.
- Generally for fixed $\varDelta$, most of the CIs are very conservative when p is close to 0 or 1 except for very few cases, see, Wald CI.
- The coverage of all the CIs are symmetric in p about $p = 0.5$ when $m = n$ and $\varDelta = 0$.
- For fixed $\varDelta$ and small m, n, Wald, score and recentered have small width for all $p \in (0, 1)$. Jeffrey and Agresti CI are much wider when p is close to 0 and 1. Newcombe CI is the shortest when p is close to 0.5 but the widest when p is close to 0 and 1.
- For fixed p and small m, n, Wald and recentered CI have small width. Modified Yule CI much wider than others when $p \neq 0.5$ and d is big.
- There is some oscillation when d is varying and m and n are small. But for the Wald CI, it can be objectionably large. For most of the procedures, it is negligible.

7. Recommendations

Under most of the circumstances the score, Newcombe and recentered CI perform very well. We strongly recommend these three intervals (see Table 1 for details). Roughly speaking, when m and n are small and $\Delta \approx 0$, we should use any of these three. When m and n are small and $\Delta \neq 0$, we should use Newcombe or recentered CI. When $m \neq n$ and $\Delta \approx 0$, we should use score CI or Newcombe CI. Do not use recentered CI if m is a common factor of n or vice visa. Under other circumstances we can use any of them. If simplicity and conservativity in coverage are the most important issue, Agresti and our Jeffrey CI are the best choice.

8. Remarks

1. *Boundary problem and Poisson modification*. In most of the figures we see that the coverage probability is either too conservative or too low when p is near 0 or $1 - \Delta$. For example, the coverage probability of Jeffrey's CI is always 1 at $p = 0$ or 1 when $m = n$ and $\Delta = 0$. When $\Delta \neq 0$, it is always below 0.95 but converges to 0.95 quickly with increasing of $m = n$. Recentered CI behaves similarly but is always conservative at the boundaries whether $\Delta = 0$ or not. We should be careful and it is better to choose conservative CIs when p is expected to be close to 0 or 1. It also suggests the desirability of making some modifications of those CIs at the boundary. One method is to apply a Poisson modification to the boundary (see Brown et al., 2001). We applied such a modification to those six CIs. Only the boundary performance of Jeffrey's and Agresti's CIs are improved a little bit while those of all other CIs remain virtually the same.
2. *Bias when $m \neq n$*. From the figures the coverage probability is not only asymmetric but also biased when $m \neq n$ and $\Delta \neq 0$. When $\min(m, n)$ is large, this phenomenon becomes insignificant. When $\max(m, n)$ is small, it is really a problem. However, for a conservative objective, one can use Agresti's CI under this situation.
3. *Limiting behavior of coverage prob*. when $m, n \to \infty$ *or* $p \to 0, 1$. In most circumstances the poorest point of coverage probability is in the neighborhood of p or $q = 0$ or 1. For example, the coverage probability of all those intervals is always 1 when $\Delta = 0$ and $p = 0$.
4. *Oscillation and Edgeworth expansion*. In the one sample situation, oscillation of the coverage probability is a serious problem for Wald CI for the proportion, see Brown et al. (2002). But for the two sample situation, it is not so serious a problem. In plots of the coverage as a function of p for fixed d the coverage function is continuous and smooth. This is not surprising since p is only a nuisance parameter for such plots. In plots of coverage as a function of d for fixed p (and fixed m and n) there is noticeable oscillation, but it is qualitatively less pronounced than that observed in the one parameter situation. We found that there may exist significant oscillation in Δ when $m \approx n$ are small. Even when it exists, it seems always acceptable. Actually, according to empirical results the magnitude of oscillation of Jeffrey's and recentered CIs are very small when Δ is not so big, say, less than 0.4. In practice it is rarely seen that Δ is bigger than 0.4. So we need not worry about oscillation in our problem.
5. *Other exponential distributions*. Like Brown et al. (2003), one could consider extending the current results to the problem of comparing the means of two samples with other distributions in simple exponential families.

References

Agresti, A., Caffo, B., 2000. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. Amer. Statist. 54 (4), 280–288.

Barnard, G.A., 1947. Significance test for 2 × 2 tables. Biometrica 34, 123–138.

Boschloo, R.D., 1970. Raised conditional level of significance for the 2 × 2 table when testing for the equality of two probabilities. Statist. Neerlandica 21 (1), 1–35.

Brown, L.D., Cai, T., DasGupta, A., 2001. Interval estimation for a binomial proportion (with discussion). Statist. Sci. 16, 101–133.

Brown, L.D., Cai, T., DasGupta, A., 2002. Confidence intervals for a binomial proportion and asymptotic expansions. Ann. Statist. 30, 160–201.

Brown, L.D., Cai, T., DasGupta, A., 2003. Interval estimation in exponential families. Stat. Sinica 13, 19–50.

Chernoff, H., 2002. Another view of the classical problem of comparing two probabilities. J. Iranian Statist. Soc. 1 (1–2), 35–53.

Cox, D.R., 1953. Some simple approximate tests for Poisson variates. Biometrika 40, 354–360.

Fisher, R.A., 1935. The Design of Experiments, Oliver & Boyd, Edinburgh.

Freeman, M.F., Tukey, J.W., 1950. Translations related to the angular and square root. Ann. Math. Statist. 21, 607–611.

Gart, J.J., 1966. Alternative analysis of contingency tables. J. Roy. Statist. Soc. B 28, 164–179.

Goodman, L.A., 1964. Simultaneous confidence intervals for contrast among multinomial populations. Ann. Math. Statist. 35, 716–725.

Haber, M., 1986. A modified exact test for 2 × 2 contingency tables. Biometrical 28, 455–463.

Howard, J.V., 1998. The 2 × 2 table: a discussion from a Bayesian viewpoint. Statist. Sci. 13, 351–367.

Liddell, D., 1976. Practical tests of 2 × 2 contingency tables. The Statistician 25, 295–304.

Newcombe, R., 1998. Interval estimation for the difference between independent proportions: comparison of eleven methods. Statist. Med. 17, 873–890.

Pearson, E.S., 1947. The choice of statistical tests illustrated on the interpretation of data classed in a 2 × 2 table. Biometrika 34, 139–167.

Tango, T., 1998. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. Statist. Med. 17, 891–908.

Wilks, S.S., 1935. The likelihood test of independence in contingency tables. Ann. Math. Statist. 6, 190–196.

Wilson, E.B., 1927. Probable inference, the law of succession, and statistical inference. J. Amer. Statist. Assoc. 22, 209–212.

Yates, F., 1934. Contingency tables involving small numbers and the χ^2-test. J. Roy. Statist. Soc. Suppl. 1, 217–235.

Yule, G.U., Kendall, M.G., 1950. An Introduction to the Theory of Statistics, 14th ed. Hafner Publishing Co., New York, NY.