The Korean Communications
in Statistics Vol. 7, No. 2, 2000
pp. 481-487

On the Interval Estimation of the Difference between Independent Proportions with Rare Events

Yongdai Kim[1], Daewoo Choi[2]

Abstract

When we construct an interval estimate of two independent proportions with rare events, the standard approach based on the normal approximation behaves badly in many cases. The problem becomes more severe when no success observations are observed on both groups. In this paper, we compare two alternative methods of constructing a confidence interval of the difference of two independent proportions by use of simulation. One is based on the profile likelihood and the other is the Bayesian probability interval. It is shown in this paper that the Bayesian interval estimator is easy to be implemented and performs almost identical to the best frequentist's method - the profile likelihood approach.

Keywords : Profile likelihood interval, Bayesian probability interval, rare events

1. Introduction

Let X_1 and X_2 be independent random variables from Binomial (m, p_1) and Binomial (n, p_2) respectively. The standard method of constructing a confidence interval of $p_1 - p_2$ is to use the normal approximation, which provides an interval estimation as

$$(\widehat{p}_1 - \widehat{p}_2) \pm 1.96\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}} \tag{1}$$

where $\widehat{p}_1 = X_1/m$ and $\widehat{p}_2 = X_2/n$. However, this conventional interval estimation behaves badly when X_1 and X_2 are very small. In particular, if $X_1 = X_2 = 0$, this interval fails to cover the true difference unless $p_1 = p_2$ since the interval length is 0. So better interval estimators are needed for rare event cases.

An exact confidence interval of a single proportion whose coverage probability does not

[1] (449-791) Dept. of Statistics, Hankuk University of Foreign Studies, Yong-in, Kyunggi
E-mail : kimy@stat.hufs.ac.kr

[2] (449-791) Dept. of Statistics, Hankuk University of Foreign Studies, Yong-in, Kyunggi

depend on the true proportion is suggested by Blyth and Still (1983) and the Bayesian probability interval of a proportion with a rare event is studied by Louis (1981). Several confidence intervals of the difference of two independent proportions are studied extensively by Newcombe (1998).

In this paper, we compare two interval estimators - the profile likelihood interval and the Bayesian probability interval for the difference of two independent proportions by use of simulation, especially focusing on rare event cases. Our work can be considered to be an extension of Louis's (1991) result.

The paper is organized as follows. In section 2, the two intervals to be compared are described in details. Section 3 presents simulation results and discussions are followed in section 4.

2. Description of methods

The basic idea of the profile likelihood confidence interval is as follows. Let $\theta = p_1 - p_2$ and $\psi = p_1 + p_2$. Consider a test $H_0(x): p_1 - p_2 = x$ versus $H_1(x): p_1 - p_2 \neq x$. We reject $H_0(x)$ when $|\widehat{p}_1 - \widehat{p}_2 - x|$ is larger than the critical value C_x. If we know the true value of ψ, then the critical value C_x is calculated for a given level α by

$$\Pr\{|\widehat{p}_1 - \widehat{p}_2 - x| > C_x | \theta = x, \psi\} = \alpha/2. \tag{2}$$

Now, we can construct the exact $100(1-\alpha)\%$ confidence interval (L, U) where

$$L = \inf\{x: H_0(x) \text{ is not rejected}\}$$

and

$$U = \sup\{x: H_0(x) \text{ is not rejected}\}.$$

See Bickel and Doksum (1977, p.155). However, we cannot construct such an interval estimator since we don't know ψ. One simple remedy of this problem is to replace ψ with the maximum likelihood estimator ψ_x constrained on $\theta = x$. This idea is essentially the same as the profile likelihood approach. In the profile likelihood approach, the maximum likelihood estimator of θ is obtained by maximizing $l(\theta, \psi_\theta)$ where l is a likelihood function and ψ_x is the maximum likelihood estimator of ψ constrained on $\theta = x$. Newcombe (1998) performed extensive simulation studies of the profile confidence interval as well as another 10 confidence intervals and he concluded that the profile likelihood confidence interval is the best among them.

The Bayesian probability interval of the difference of two independent proportions is constructed as follows. Let $\pi(p_1, p_2)$ be the prior distribution of (p_1, p_2) . Then the posterior distribution is given by

$$\pi(p_1, p_2 | X_1, X_2) \propto p_1^{X_1}(1-p_1)^{m-X_1} p_2^{X_2}(1-p_2)^{n-X_2} \pi(p_1, p_2).$$

Let $\pi(\theta, \psi | X_1, X_2)$ be the corresponding posterior distribution of θ and ψ , which can be obtained by use of variable transformation technique. Now, the equal tail $(1-\alpha) \times 100\%$ probability interval has the form of (L, U) which satisfies

$$\int_U^1 \int_0^2 \pi(\theta, \psi | X_1, X_2) d\psi d\theta = \alpha/2$$

and

$$\int_{-1}^L \int_0^2 \pi(\theta, \psi | X_1, X_2) d\psi d\theta = \alpha/2$$

See Gelman et al. (1995).

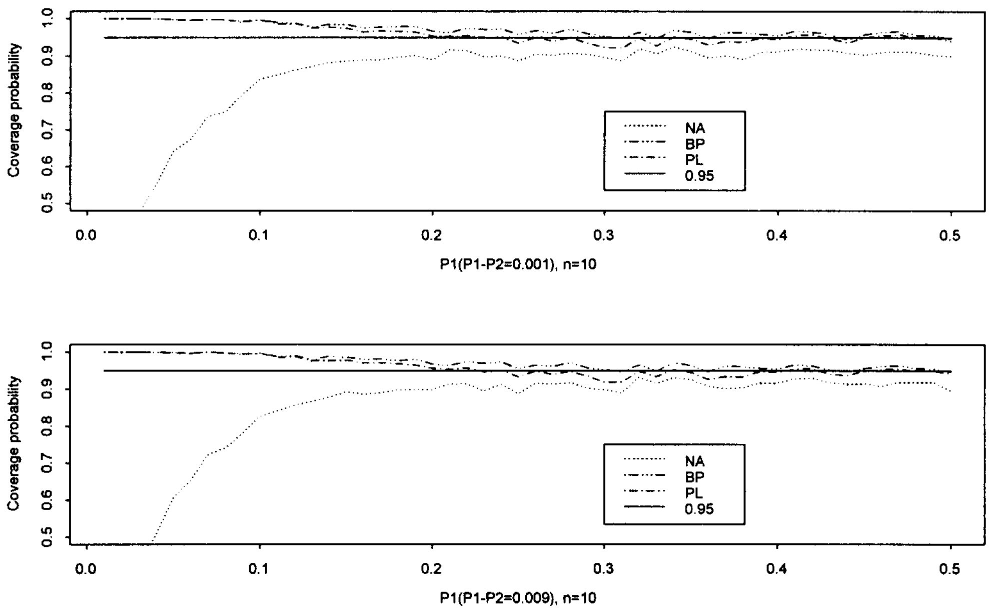
3. Simulation results

In this section, simulation results of comparing the profile confidence interval and the Bayesian probability interval of the difference of two independent proportions are presented. For the prior distribution $\pi(p_1, p_2)$, we use the flat prior, that is

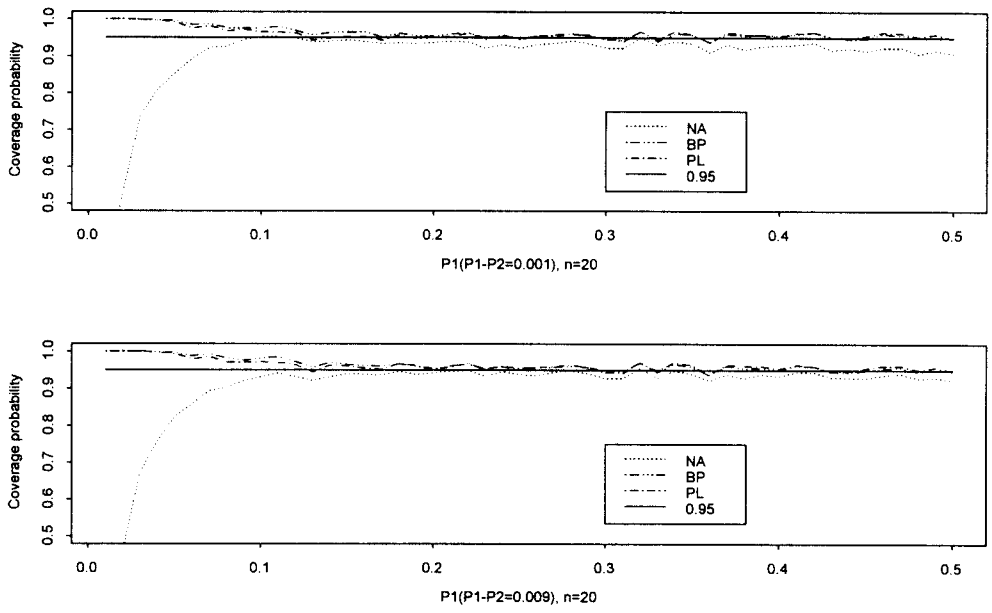
$$\pi(p_1, p_2) = I((p_1, p_2) \in [0,1] \times [0,1]).$$

This prior is known to be a noninformative prior and widely used in practice. With this prior, the posterior distribution of p_1 and p_2 are independent beta distributions with parameters $(X_1+1, m-X_1+1)$ and $(X_2+1, n-X_2+1)$ respectively.

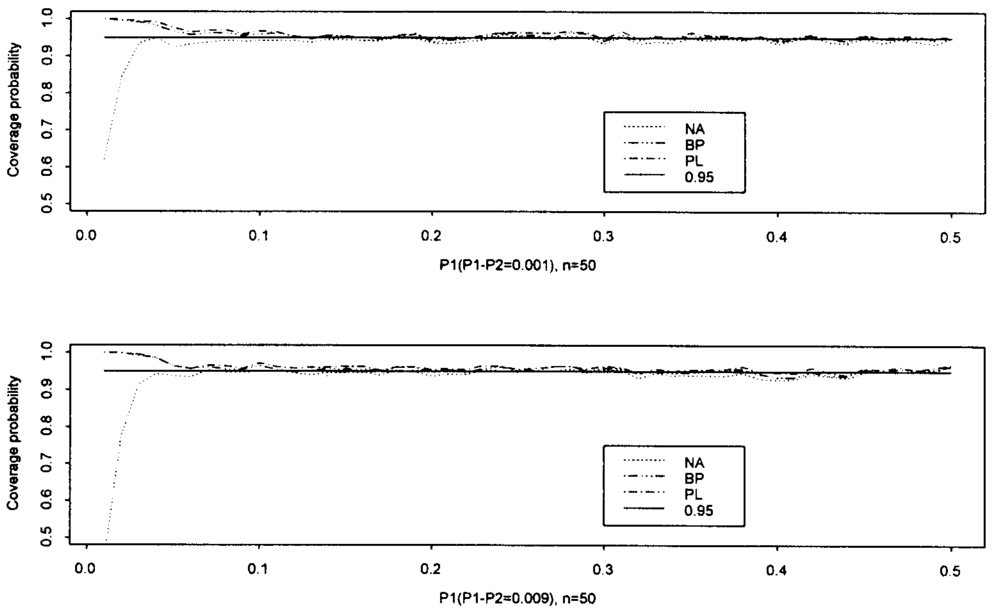
Figures 1,2,3 and 4 present the coverage probabilities of those two intervals as well as the conventional interval based on the normal approximation. The horizontal line is the values of p_1 and the vertical line is the coverage probabilities when $p_1 - p_2$ is given. "NA", "BP", "PL" and "0.95" in the legends of the figures mean the confidence interval based on the normal approximation, the Bayesian probability interval, the profile likelihood interval and the reference line of the given significant level respectively. It can be clearly seen that the conventional interval has very low coverage probability for small values of p_1 even with relatively large sample sizes while the profile confidence interval and the Bayesian probability interval cover the true value approximately 95% for all values of p_1 and p_2. Also, it should be noted that the profile likelihood and the Bayesian probability interval work well similarly.



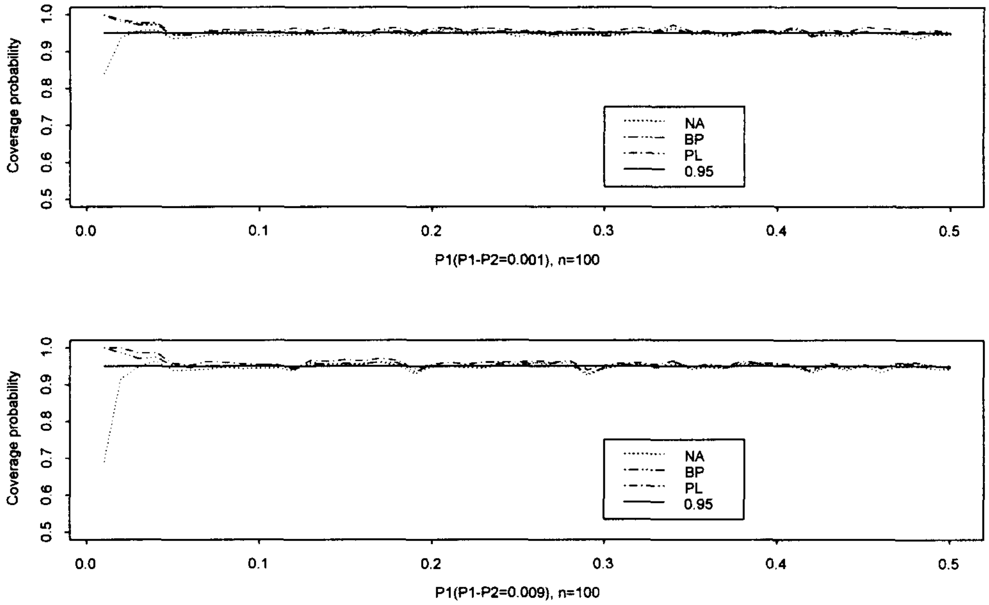
<Fig. 1> The coverage probabilities of several interval estimators when the sample size is 10.



<Fig. 2> The coverage probabilities of several interval estimators when the sample size is 20.



<Fig. 3> The coverage probabilities of several interval estimators when the sample size is 50.



<Fig. 4> The coverage probabilities of several interval estimators when the sample size is 100.

4. Discussion

We observed that the profile likelihood confidence interval and the Bayesian probability interval work well even when the true value of p_1 and p_2 are very small while the conventional interval fails many times. Another comment of the simulation result is that the profile likelihood confidence interval and the Bayesian probability interval cover the true value more than necessary when the true values are very small. This phenomenum is acceptable in view of conservation. That is, if we use these intervals for testing $p_1 - p_2 = 0$, the size of the test is smaller than α . In contrast with this, the conventional confidence interval is too short when the true value is small, which should be avoided. Conclusively, when the observations are close to 0, the conventional method should be abolished and either the profile likelihood confidence interval or the Bayesian probability interval is recommended.

From computational points of view, the Bayesian probability interval is preferred to the profile likelihood confidence interval. In the profile likelihood confidence interval, computation of ψ_x is very computational demanding. In our simulation work, we use grid search. On the other hand, the lower limit and the upper limit of the Bayesian probability interval can be easily obtained by use of a simple Monte Carlo method.

Continuity-corrected profile likelihood confidence interval can be obtained by employing usual continuity-correcting methods. Also in the Bayesian probability interval, the Jeffrey's prior (Gelman et al., 1995) can be used instead of the flat prior since Jeffrey's prior has many desirable frequentist's properties (Nicolaou, 1993). These alternative confidence intervals will be studied later on.

References

[1] Bickel, P.J. and Doksum, K.A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*, San Francisco: Holden-Day.

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

[3] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*, Chapman & Hall.

[4] Louis, T.A. (1981). Confidence intervals for a binomial parameter after observing no successes, *The American Statistician*, Vol. 35, 154-154.

[5] Newcombe, R.G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods, *Statistics in Medicine*, Vol. 17, 873-890.

[6] Nicolaou, A. (1993). Bayesian intervals with good frequentist behavior in the presence of nuisance parameters, *Journal of the Royal Statistical Society, Series B*, Vol. 55, 377-390.