Communications of the Korean Statistical Society
2010, Vol. 17, No. 3, 309–318

Confidence Intervals for the Difference of Binomial Proportions in Two Doubly Sampled Data

Seung-Chun Lee[1,a]

[a]Department of Statistics, Hanshin University

Abstract

The construction of asymptotic confidence intervals is considered for the difference of binomial proportions in two doubly sampled data subject to false-positive error. The coverage behaviors of several likelihood based confidence intervals and a Bayesian confidence interval are examined. It is shown that a hierarchical Bayesian approach gives a confidence interval with good frequentist properties. Confidence interval based on the Rao score is also shown to have good performance in terms of coverage probability. However, the Wald confidence interval covers true value less often than nominal level.

Keywords: Profile likelihood, Rao score, hierarchical Bayesian approach, coverage probability, expected width, double sampling.

1. Introduction

A double sampling scheme on binary observations occurs when the cost of precise test is expensive. To reduce the cost, most samples are classified by an inexpensive but fallible device, and a small subsample is classified by a supplementary inerrant device. Numerous literatures are concerned with the inference on the population proportion in the double sampling scheme; see Tenenbein (1970), Geng and Asano (1989), York *et al.* (1995), Moors *et al.* (2000), Barnett *et al.* (2001), Raats and Moors (2003) and Boese *et al.* (2006). For instance, York *et al.* (1995) illustrated the advantage of the double sampling scheme in estimating the proportion of infants born with Down's syndrome nationwide. For every birth during a certain period of time, the midwife or obstetrician classified the child with Down's syndrome based on a visual inspection, and for a small subsample of births, expensive but accurate cytogenetic tests were applied for the classification.

The classification by a visual inspection could not be expected to be accurate; the visually inspected sample might be exposed to measurement error. It is well known that usual estimators can be extremely biased when data is subjected to misclassification. It might be dangerous to use the data alone. On the other hand, the use of only the cytogenetic data would give a large variance because of the small sample size. It would have clear advantages to use all the data in estimating the accuracy of the visual test as well as the population proportion.

The sample classified by only the visual inspection might contain two types of error done by the midwife or obstetrician. He or she might classify erroneously a normal child as Down's syndrome (false-positive) and vice versa (false-negative). Tenenbein (1970) presented the maximum likelihood(ML) estimator for the population proportion as well as for false-positive and false-negative error rates.

This work was supported by Hanshin University research grant.
[1] Professor, Department of Statistics, Hanshin University, 411 Yangsan-Dong, Osan, Kyunggi-Do 449-791, Korea.
E-mail: seung@hs.ac.kr

Some fallible devices may have only one type of misclassification. For example, Lie *et al.* (1994) considered the case that the false-negative counts were corrected using multiple fallible classifiers and gave the ML estimators. The same model was considered by York *et al.* (1995). They estimated the proportion of Down's syndrome in Norway from the Bayesian perspective. Moors *et al.* (2000) analyzed an auditing data with no observed false-negative count. They put the corresponding error rate equal to zero a priori, and gave one-sided confidence intervals for the population proportion.

For interval estimation problem, Boese *et al.* (2006) gave five likelihood-based confidence intervals in the false-positive misclassification model. Among them, they recommended the interval based on the Rao score function. The recommended interval seems to have a good frequentist property in that it has the coverage probability close to nominal level. Under the same model, Lee and Byun (2008) gave a Bayesian confidence interval which outperforms slightly the confidence interval recommended by Boese *et al.* (2006).

In this paper, we consider the interval estimation of the difference of binomial proportions in two data sets obtained by a double sampling scheme subjected to false-positive misclassification. A naive approach is employing the Wald method. However, many recent works showed that the Wald method does not provide proper interval in various sampling situations; see for example, Blyth and Still (1983), Agresti and Coull (1998) and Brown *et al.* (2001). In particular, Brown *et al.* (2001) investigated the unsatisfactory coverage properties of the Wald interval in detail. We will demonstrate that the Wald method is not adequate as well for the sampling design considered in this paper.

Agresti and Coull (1998) showed that an improved interval for the parameter of a binomial distribution could be obtained by so-called "adding two successes and two failures" to the observed counts and then using the Wald method. Agresti-Coull's strategy works well in various sampling designs as well as in the 1-group design. For instance, Agresti and Caffo (2000) examined the interval estimation for the difference of two binomial proportions, and concluded that the strategy performs about as well as the best available methods in this 2-group design. See also Agresti and Min (2005). The more general problem of interval estimation for a linear function of binomial proportions was considered by Price and Bonett (2004). Unlike the 1-group and the 2-group cases for which competitive alternatives exist, they also concluded that the Agresti-Coull's method would provide effective confidence intervals. In addition, Lee (2007) investigated the performance of the Agresti-Coull type confidence interval in a double sampling design subject to false-positive misclassification and concluded that the Agresti-Coull type interval is comparable to or even better than likelihood-based confidence intervals.

It might be peculiar why "adding two successes and two failures" would work well in those designs. Agresti and Coull (1998) justified their approach by the Bayesian perspective. Lee and Byun (2008) showed that this justification is also applicable to the double sampling design. In other words, the Agresti-Coull type interval is essentially a Bayesian confidence interval. Thus, it is natural to consider a Bayesian approach for the difference of binomial proportions. Likelihood-based confidence intervals described in Barndorff-Nielsen and Cox (1994) also could be good candidates for the problem.

2. Two Sample False-Positive Misclassification Model

A double sampling scheme consists of two stages of sampling. A sample of size N is selected at random from the population of interest and a fallible device classifies each unit in the sample, and then a subset of size n is selected from the initial sample. Each unit in the subsample is tested by an inerrant device. Thus, a unit in the subsample is tested by both the inerrant and the fallible device.

For each unit tested by the inerrant device, let $T_i = 1$, if i^{th} unit is recorded positive (or a success), and $T_i = 0$, if otherwise. Likewise, for each unit tested by the fallible device, define $F_i = 1$, if i^{th} unit is classified as positive, and $F_i = 0$, if otherwise. The proportion of success p can be written as

$$p = \Pr[T_i = 1],$$

and the false-positive error rate is

$$\phi = \Pr[F_i = 1 | T_i = 0].$$

The false-negative error rate, $\Pr[F_i = 0|T_i = 1]$, is assumed to be zero in this model. Thus, each unit in the subsample belongs to one of three mutually disjoint categories $\{(t, f) \,|\, (0, 0), (0, 1), (1, 1)\}$ with probabilities $(1 - p)(1 - \phi), (1 - p)\phi$ and p, respectively. Let n_{tf} be the observed count in (t, f). $N - n$ units are tested by only fallible device. Among these units, let x be the number of units tested positively, and $y = N - n - x$. Define $\pi = \Pr[F_i = 1] = p + (1 - p)\phi$.

Assuming each unit is tested independently, the joint likelihood of p and ϕ is given by

$$L(p, \phi; \mathcal{Y}) = C(\mathcal{Y}) \left[(1 - p)\phi\right]^{n_{01}} p^{n_{11}} \pi^{x} (1 - \pi)^{n_{00}+y},$$

where $C(\mathcal{Y}) = n!/(n_{00}!n_{01}!n_{11}!)\binom{N-n}{x}$ and $\mathcal{Y}$ represents $(n_{00}, n_{01}, n_{11}, x, y)$.

The maximum likelihood estimate of p and ϕ were obtained by Tenenbein (1970) as:

$$\hat{p} = \frac{n_{11}}{n_{01} + n_{11}} \frac{x + n_{01} + n_{11}}{N} \tag{2.1}$$

and

$$\hat{\phi} = \frac{n_{01}}{n_{01} + n_{11}} \frac{x + n_{01} + n_{11}}{N(1 - \hat{p})}. \tag{2.2}$$

See also Barnett *et al.* (2001). While the estimate of asymptotic variance of $\hat{p}$ is given as:

$$\widehat{\mathrm{Var}}(\hat{p}) = \frac{\hat{p}\,\hat{q}}{n} - \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n_{11}}{n_{11} + n_{01}} \hat{p}(1 - \hat{\pi}), \tag{2.3}$$

where $\hat{q} = 1 - \hat{p}$ and $\hat{\pi} = (x + n_{01} + n_{11})/N$.

A two-sample false-positive misclassified data consists of two data sets $\mathcal{Y}_1 = (n_{100}, n_{101}, n_{111}, x_1, y_1)$ and $\mathcal{Y}_2 = (n_{200}, n_{201}, n_{211}, x_2, y_2)$, where each $\mathcal{Y}_i$ is sampled from $L(p_i, \phi_i; \mathcal{Y}_i)$ independently. Thus, the joint likelihood of p_1, p_2, ϕ_1 and ϕ_2 can be written as:

$$L(p_1, p_2, \phi_1, \phi_2; \mathcal{Y}_1, \mathcal{Y}_2) = L(p_1, \phi_1; \mathcal{Y}_1)\, L(p_2, \phi_2; \mathcal{Y}_2). \tag{2.4}$$

3. Confidence Intervals

3.1. Frequentist confidence intervals

3.1.1. Profile likelihood and information

The profile likelihood for λ and the restricted information are the keys of likelihood-based confidence intervals. Thus, the calculation of them is essential in what follows.

Substituting p_1 by $\lambda + p_2$ and taking logarithm of (2.4), we have the full log-likelihood,

$$\begin{aligned}\ell(\lambda, p_2, \phi_1, \phi_2) = {} & (n_{100} + n_{101} + y_1)\log(1-\lambda-p_2) + n_{111}\log(\lambda+p_2) + (n_{100}+y_1)\log(1-\phi_1) + \\ & n_{101}\log\phi_1 + x_1\log\pi_1 + (n_{200}+n_{201}+y_2)\log(1-p_2) + n_{211}\log p_2 + \\ & (n_{200}+y_2)\log(1-\phi_2) + n_{201}\log\phi_2 + x_2\log\pi_2,\end{aligned}$$

where $\pi_1 = (1-\lambda-p_2)\phi_1 + (\lambda+p_2)$ and $\pi_2 = (1-p_2)\phi_2 + p_2$. Profile log-likelihood $\ell_P(p_2, \phi_1, \phi_2; \lambda)$ is the full log-likelihood regarding λ as a given value.

Note that, given $\lambda \in (-1, 1)$, the maximum of the log-profile likelihood is $\ell_P(\hat{p}_2^\lambda, \hat{\phi}_1^\lambda, \hat{\phi}_2^\lambda; \lambda)$ where $\hat{p}_2^\lambda, \hat{\phi}_1^\lambda$ and $\hat{\phi}_2^\lambda$ are the solutions of following profile likelihood equations:

$$0 = -\frac{n_{100}+n_{101}+y_1}{1-\lambda-p_2} + \frac{n_{111}}{\lambda+p_2} + \frac{(1-\phi_1)x_1}{\pi_1} - \frac{n_{200}+n_{201}+y_2}{1-p_2} + \frac{n_{211}}{p_2} + \frac{(1-\phi_2)x_2}{\pi_2} \tag{3.1}$$

$$0 = -\frac{n_{100}+y_1}{1-\phi_1} + \frac{n_{101}}{\phi_1} + \frac{(1-\lambda-p_2)x_1}{\pi_1} \tag{3.2}$$

$$0 = -\frac{n_{200}+y_2}{1-\phi_2} + \frac{n_{201}}{\phi_2} + \frac{(1-p_2)x_2}{\pi_2}. \tag{3.3}$$

Let

$$\phi_1(p) = \frac{B_1(p) + \sqrt{B_1^2(p) + 4(N_1 - n_{111})n_{101}(p+\lambda)(1-\lambda-p)}}{2(N_1 - n_{111})(1-\lambda-p)} \tag{3.4}$$

and

$$\phi_2(p) = \frac{B_2(p) + \sqrt{B_2^2(p) + 4(N_2 - n_{211})n_{201}p(1-p)}}{2(N_2 - n_{211})(1-p)}, \tag{3.5}$$

where $B_i(p) = n_{i01} + x_i - (N_i - n_{i11} + n_{i01})(p+\lambda)$, $i = 1, 2$. Then it can be shown that $\hat{p}_2^\lambda$ is the solution of a nonlinear equation

$$g(p) = (1-p)\left(n_{111} + \frac{n_{101}}{\phi_1(p)} - \frac{n_{111}}{\lambda+p}\right) + (1-\lambda-p)\left(n_{211} + \frac{n_{201}}{\phi_2(p)} - \frac{n_{211}}{p}\right) = 0.$$

Substituting p in (3.4) and (3.5) by $\hat{p}_2^\lambda$, we have $\hat{\phi}_1^\lambda = \phi_1(\hat{p}_2^\lambda)$ and $\hat{\phi}_2^\lambda = \phi_2(\hat{p}_2^\lambda)$.

When all observed counts are greater than zero, $\hat{p}_2^\lambda$ lies in the interval $(\max\{-\lambda, 0\}, \min\{1-\lambda, 1\})$, which in turn results in $\hat{\phi}_1^\lambda \in (0, 1)$ and $\hat{\phi}_2^\lambda \in (0, 1)$. Thus, a simple numerical algorithm such as the bisection method or the Newton-Raphson method can be applicable. However, when some observed counts are zero, then the full likelihood or the profile likelihood does not admit unique maximum. For instance, when $n_{211} = 0$ or $n_{201} = 0$, $\log(\hat{p}_2)$ or $\log(\hat{\phi}_2)$ is undefined. A customary remedy to prevent the undefined problem is to add a small number, say 1.e-5, to null observed counts; see for example Boese *et al.* (2006). Thus we will add a small number when necessary for the calculation of likelihood-based confidence intervals.

Let $\hat{p}_1^\lambda = \lambda + \hat{p}_2^\lambda, \hat{\pi}_1^\lambda = (1-\hat{p}_1^\lambda)\hat{\phi}_1^\lambda + \hat{p}_1^\lambda$ and $\hat{\pi}_2^\lambda = (1-\hat{p}_2^\lambda)\hat{\phi}_2^\lambda + \hat{p}_2^\lambda$. Then the adjusted observed information for λ is

$$J^{\lambda\lambda} = J_{\lambda\lambda} - (J_{\lambda p_2}, J_{\lambda\phi_1}, J_{\lambda\phi_2})\begin{pmatrix} J_{p_2p_2} & J_{p_2\phi_1} & J_{p_2\phi_2} \\ J_{p_2\phi_1} & J_{\phi_1\phi_1} & J_{\phi_1\phi_2} \\ J_{p_2\phi_2} & J_{\phi_1\phi_2} & J_{phi_2\phi_2}\end{pmatrix}^{-1}\begin{pmatrix} J_{p_2\lambda} \\ J_{\phi_1\lambda} \\ J_{\phi_2\lambda}\end{pmatrix},$$

where

$$J_{\lambda\lambda} = J_{\lambda p_2} = \frac{n_{100}+n_{101}+y_1}{(1-\hat{p}_1^\lambda)^2} + \frac{n_{111}}{(\hat{p}_1^\lambda)^2} + \frac{(1-\hat{\phi}_1^\lambda)^2 x_1}{(\hat{\pi}_1^\lambda)^2}, \quad J_{\phi_1\phi_1} = \frac{n_{100}+y_1}{(1-\hat{\phi}_1^\lambda)^2} + \frac{n_{101}}{(\hat{\phi}_1^\lambda)^2} + \frac{(1-\hat{p}_1^\lambda)^2 x_1}{(\hat{\pi}_1^\lambda)^2},$$

$$J_{p_2p_2} = J_{\lambda\lambda} + \frac{n_{200}+n_{201}+y_2}{(1-\hat{p}_2^\lambda)^2} + \frac{n_{211}}{(\hat{p}_2^\lambda)^2} + \frac{(1-\hat{\phi}_2^\lambda)^2 x_2}{(\hat{\pi}_1^\lambda)^2}, \quad J_{\phi_2\phi_2} = \frac{n_{200}+y_2}{(1-\hat{\phi}_2^\lambda)^2} + \frac{n_{201}}{(\hat{\phi}_2^\lambda)^2} + \frac{(1-\hat{p}_2^\lambda)^2 x_2}{(\hat{\pi}_2^\lambda)^2},$$

$$J_{\lambda\phi_1} = J_{p_2\phi_1} = \frac{x_1}{(\hat{\pi}_1^\lambda)^2}, \quad J_{p_2\phi_2} = \frac{x_2}{(\hat{\pi}_2^\lambda)^2}, \qquad J_{\lambda\phi_2} = J_{\phi_1\phi_2} = 0.$$

The adjusted restricted information is obtained by replacing observed counts by their expectations. That is,

$$I^{\lambda\lambda} = I_{\lambda\lambda} - (I_{\lambda p_2}, I_{\lambda\phi_1}, I_{\lambda\phi_2}) \begin{pmatrix} I_{p_2p_2} & I_{p_2\phi_1} & I_{p_2\phi_2} \\ I_{p_2\phi_1} & I_{\phi_1\phi_1} & I_{\phi_1\phi_2} \\ I_{p_2\phi_2} & I_{\phi_1\phi_2} & I_{phi_2\phi_2} \end{pmatrix}^{-1} \begin{pmatrix} I_{p_2\lambda} \\ I_{\phi_1\lambda} \\ I_{\phi_2\lambda} \end{pmatrix},$$

where

$$I_{\lambda\lambda} = I_{\lambda p_2} = \frac{(1-\hat{\phi}_1^\lambda)N_1 - n_1\hat{\phi}_1^\lambda}{1-\hat{p}_1^\lambda} + \frac{n_1}{\hat{p}_1^\lambda} + \frac{(N_1-n_1)(1-\hat{\phi}_1^\lambda)^2}{\hat{\pi}_1^\lambda}, \quad I_{\phi_1\phi_1} = \frac{1-\hat{p}_1^\lambda}{1-\hat{\phi}_1^\lambda}\left[\frac{n_1}{\hat{\phi}_1^\lambda} + \frac{N_1-n_1}{\hat{\pi}_1^\lambda}\right],$$

$$I_{p_2p_2} = I_{\lambda\lambda} + \frac{(1-\hat{\phi}_2^\lambda)N_2 - n_2\hat{\phi}_2^\lambda}{1-\hat{p}_2^\lambda} + \frac{n_2}{\hat{p}_2^\lambda} + \frac{(N_2-n_2)(1-\hat{\phi}_2^\lambda)^2}{\hat{\pi}_1^\lambda}, \quad I_{\phi_2\phi_2} = \frac{1-\hat{p}_2^\lambda}{\hat{\phi}_2^\lambda}\left[\frac{n_2}{\hat{\phi}_2^\lambda} + \frac{N_2-n_2}{\hat{\pi}_2^\lambda}\right],$$

$$I_{\lambda\phi_1} = I_{p_2\phi_1} = \frac{N_1-n_1}{\hat{\pi}_1^\lambda}, \quad I_{p_2\phi_2} = \frac{N_2-n_2}{\hat{\pi}_2^\lambda}, \qquad I_{\lambda\phi_2} = I_{\phi_1\phi_2} = 0.$$

3.1.2. Likelihood-based confidence intervals

The first likelihood-based confidence interval considered in this paper is the Wald interval which can be constructed by using (2.1) and (2.3) as

$$\hat{\lambda} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{p}_1) + \widehat{\mathrm{Var}}(\hat{p}_2)}, \tag{3.6}$$

where $\hat{\lambda} = \hat{p}_1 - \hat{p}_2$ and z_α represents the $1-\alpha$ quantile of a standard normal distribution. We will denote (3.6) as $\mathrm{CI_W}$.

Efron and Hinkley (1978) claimed that the observed information is preferable form than the expected information in general. However, under the double sampling model, the restricted observed information yields nonsensical results in some cases as shown in Boese *et al.* (2006). Thus we do not consider the confidence intervals based the restricted observed information.

Using the adjusted restrictive information, an Wald related confidence intervals can be setup as:

$$W_n = \left\{\lambda : \left(\lambda - \hat{\lambda}\right)^2 I^{\lambda\lambda} \le z_{\alpha/2}^2\right\}.$$

Next interval is based on the Rao score which is obtained from the partial derivative of $\ell(\lambda, p_2, \phi_1, \phi_2)$ with respective to λ. Substituting nuisance parameters by the corresponding solutions of profile likelihood equations, we have

$$s\left(\hat{p}_2^\lambda, \hat{\phi}_1^\lambda, \hat{\phi}_2^\lambda; \lambda\right) = -\frac{n_{100}+n_{101}+y_1}{1-\lambda-\hat{p}_2^\lambda} + \frac{n_{111}}{\lambda+\hat{p}_2^\lambda} + \frac{(1-\hat{\pi}_1^\lambda)x_1}{\pi_1}.$$

Then, weighting by the adjusted restrictive information, we get an approximate confidence interval,

$$S_n = \left\{\lambda : s\left(\hat{p}_2^\lambda, \hat{\phi}_1^\lambda, \hat{\phi}_2^\lambda; \lambda\right)^2 \left(I^{\lambda\lambda}\right)^{-1} \le z_{\alpha/2}^2\right\}.$$

The last likelihood-based confidence interval is due to the well-known log-likelihood ratio statistic, *i.e.*

$$Q_n = \left\{\lambda : 2\left[\ell\left(\hat{\lambda}, \hat{p}_2, \hat{\phi}_1, \hat{\phi}_2\right) - \ell_P\left(\hat{p}_2^\lambda, \hat{\phi}_1^\lambda, \hat{\phi}_2^\lambda; \lambda\right)\right] \le z_{\alpha/2}^2\right\}.$$

W_n, S_n and Q_n do not admit closed-form intervals. Although the wide array method can give confidence limits, one may have difficulty in calculating the limits. They are computationally too expensive.

3.2. Confidence interval based on Bayesian approach

Various researchers have previously discussed the Bayesian approach for the double sampling scheme. Geng and Asano (1989) used Dirichlet priors for the joint probabilities rather than for the natural model parameters p and ϕ. However, as the conjugate prior for the binomial sample results, the beta distribution might be appropriate for the marginal prior distributions of p and ϕ. In addition, it is logical to assume that p and ϕ are independent.

Raats and Moors (2003) used the beta prior distribution. However, the beta distribution leads to complex posterior distributions. In particular, the marginal posterior distribution of p is a nontrivial linear combination of beta distributions, and it requires a heavy computation to calculate density. Even, we can only obtain the mean of the posterior distribution numerically. Thus, their approach is not appropriate for the interval estimation problem.

Recently Lee and Byun (2008) gave a confidence interval for population proportion in a doubly sampled data. Applying a hierarchical Bayesian approach, a relatively simple and effective confidence interval was derived. This result can be applicable to the problem considered in this paper.

They applied priors hierarchically as

$$g(p_i|\pi_i) = \frac{p_i^{\alpha_i-1}(\pi_i - p_i)^{\beta_i-1}}{\mathrm{B}(\alpha_i, \beta_i)\pi_i^{\alpha_i+\beta_i-1}}, \quad 0 < p_i < \pi_i$$

and

$$g(\pi_i) = \frac{1}{\mathrm{B}(\gamma_i, \delta_i)}\pi_i^{\gamma_i-1}(1 - \pi_i)^{\delta_i-1}, \quad 0 < \pi_i < 1,$$

where $\mathrm{B}(m, n) = \int_0^1 x^{m-1}(1 - x)^{n-1}dx$. Then, the posterior mean and variance of p_i for $i = 1, 2$ are

$$\hat{p}_i^B = \hat{\pi}_i^B \frac{n_{i11} + \alpha_i}{n_{i11} + n_{i01} + \alpha_i + \beta_i}$$

and

$$\begin{aligned}\mathrm{Var}\left(\hat{p}_i^B\right) = &\frac{(n_{i11} + \alpha_i)(n_{i01} + \beta_i)}{(n_{i11} + n_{i01} + \alpha_i + \beta_i)^2} \frac{(\hat{\pi}_i^B)^2 + \hat{\pi}_i^B(1 - \hat{\pi}_i^B)/(\tilde{N}_i + 1)}{n_{i11} + n_{i01} + \alpha_i + \beta_i + 1} \\ &+ \left(\frac{n_{i11} + \alpha_i}{n_{i11} + n_{i01} + \alpha_i + \beta_i}\right)^2 \frac{\hat{\pi}_i^B(1 - \hat{\pi}_i^B)}{(\tilde{N}_i + 1)},\end{aligned}$$

Table 1: Case-control data of Hildesheim *et al.* (absorbing false-negatives into true-positives)

		Fallible device			
		Control group		Case group	
	Inerrant device	0	1	0	1
Subsample	0	33	11	13	3
	1	na	32	na	23
		701	535	318	375

respectively, where $\tilde{N}_i = N_i + \gamma_i + \delta_i$ and $\hat{\pi}_i^B = (x_i + n_{i01} + n_{i11} + \gamma_i)/\tilde{N}_i$. Thus, a Bayesian confidence interval due to Lee and Byun (2008) can be established as:

$$\hat{p}_1^B - \hat{p}_2^B \pm z_{\alpha/2}\sqrt{\mathrm{Var}\left(\hat{p}_1^B\right) + \mathrm{Var}\left(\hat{p}_2^B\right)}. \tag{3.7}$$

To utilize (3.7), we must specify the parameters of prior distributions. Usually the parameters reflect the prior knowledge about p_i's and π_i. However, noninformative priors were adequate for interval estimation problem, and it is logical that the prior parameters of the first group are the same as those of the second group. We had considered three choices of noninformative priors, uniform prior ($\alpha_i = \beta_i = \gamma_i = \delta_i = 1$), Jeffrey's prior ($\alpha_i = \beta_i = \gamma_i = \delta_i = 1/2$) and $\alpha_i = \gamma_i = z_{\alpha/2}^2/4$, $\beta_i = z_{\alpha/2}^2/8$ and $\delta = 3z_{\alpha/2}^2/4$. The last one is closely related to "adding two successes and two failures". See Lee and Byun (2008) for further detail.

It turned out that the last set of priors outperforms the others in terms of the coverage probability and the expected width. Thus, we only consider the last one in what follows and denoted it by $\mathrm{CI_B}$.

3.3. An example

The case-control study of Hildesheim *et al.* (1991) aimed to examine that invasive cervical cancer can affect exposure to Herpes Simplex Virus(HSV). To explore the relationship, western blot procedure was applied to 693 women in the case group and for 1236 women in the control group to detect the infection of HIV. Since the western blot procedure is fallible, a sub-sample from each group was further investigated by refined western blot procedure, which is known to be a relatively accurate procedure. Originally the fallible procedure is exposed to the two types of error, but we assume the false-negative error rate is zero. The false-negative cases are absorbed into the true-positive. This artificial data is shown in Table 1.

Using the values in Table 1, we find the Bayesian and the maximum likelihood estimates of p are −0.151 and −0.157, respectively with standard errors 0.0484 and 0.0539. Figure 1 shows the values of likelihood-based statistics against λ. The limits of 95% confidence intervals are the points crossing with $z_{0.025}^2 = 3.8416$, which are $(-0.238, -0.058)$, $(-0.247, -0.052)$ and $(-0.254, 0.034)$ for each S_n, Q_n and W_n, respectively. Thus, depending upon the statistic, one can have different decisions for the null value of hypothesis at 5% significance level. Note that W_n has the widest width in this case. 95% Bayesian and Wald confidence limits are $(-0.245, -0.056)$ and $(-0.262, -0.051)$.

4. Comparison of Confidence Intervals and Conclusions

As we noted before, the likelihood-based confidence intervals are computationally expensive. For instance, when $N_1 = N_2 = 100$ and $n_1 = n_2 = 20$, it requires 18,711 × 18,711 = 350,101,521 iterations to calculate actual coverage probability for each parameter point (p_1, p_2). In particular, the wide array method for calculating confidence limits requires a huge computation work. Thus, it is practically impossible to compare actual coverage probability or expected width. We abandon

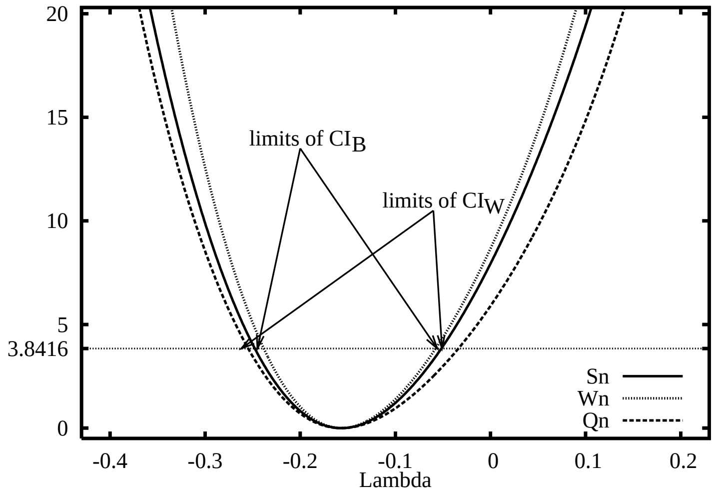

Figure 1: *Values of likelihood-based statistics against λ for the data in Table 1. The crossing points with 3.8416 are 95% confidence limits.*

to compare actual values. The comparison is done through simulations. That is, we estimated the coverage probability and the expected width of 95% confidence intervals at every 81 grid points of (p_1, p_2) where $p_i = 0.1$ up to 0.9 with 10,000 random samples, and then calculated the averages of these 81 estimated coverage probabilities and expected widths. We also computed the averages of mean absolute deviation of coverage probabilities from nominal level(AMADN). The AMADN is multiplied by 10,000. The results are shown in Table 2. The random samples were generated by the well-known IMSL Fortran Library.

Some messages of Table 2 are quite clear. For instance, CI_W, Q_n and W_n do not approximate the nominal level well enough. The AMADN's of these confidence intervals are always larger than those of S_n and CI_B. In particular, CI_W has significantly smaller averages of coverage probabilities than the nominal level. As a result, it has the largest AMADN's for almost always. It seems that the Wald procedure should be used with a great care for the interval estimation of the difference of binomial proportions in two doubly sampled data as well. Q_n and W_n also are not interesting. These two intervals are dominated by S_n and CI_B in the approximation. Since the averages of coverage of these two are always lower than nominal level, we may conclude that they tend to be narrower than necessary under the configurations consider in Table 2.

However, the preference between S_n and CI_B is not clear. S_n gives better approximation than CI_B in most cases, but CI_B also dominates S_n in some cases. Although it seems that S_n would be better than CI_B in the approximation, it is hard to conclude that S_n is better than CI_B. Note that the coverage of S_n fluctuate around the nominal level, but those of CI_B are slightly larger than the nominal level. In other words, CI_B is conservative. If we notes that CI_B has almost always smaller averages of expected widths than S_n, one may choose CI_B for the interval estimation problem.

Another important factor for judging confidence intervals may be the simplicity emphasized by Agresti and Coull (1998), Brown *et al.* (2001) and many other researchers. It is possible to argue that the computation is not a problem for modern computational techniques; however, we believe that the simplicity is important as it is a matter of practice and not of computation. From this point of view, we prefer CI_B to S_n in that CI_B is much simpler with competitive powers.

Table 2: Averages of 81 estimated coverage probabilities and expected widths, MADN

Group 1		Group 2			Average coverage probability and expected width, AMADN×10, 000														
N_1 n_1	ϕ_1	N_2	n_2	ϕ_2	$\mathrm{CI_W}$			Q_n			W_n			S_n			$\mathrm{CI_B}$		
100 20	0.1	100	20	0.1	.923	.306	268	.936	.309	160	.947	.326	165	.958	.339	112	.958	.312	108
				0.2	.922	.330	282	.934	.331	172	.944	.343	150	.954	.354	94	.953	.330	105
			30	0.1	.928	.294	221	.936	.293	144	.944	.303	132	.952	.311	77	.955	.297	95
				0.2	.929	.309	210	.937	.308	139	.944	.316	123	.951	.324	73	.953	.310	86
		200	40	0.1	.926	.267	242	.938	.267	134	.947	.278	125	.954	.287	83	.955	.271	84
				0.2	.929	.282	216	.939	.281	125	.946	.291	121	.953	.299	74	.952	.284	80
			60	0.1	.925	.258	255	.940	.257	129	.948	.268	119	.954	.275	74	.954	.262	72
				0.2	.926	.268	236	.940	.266	128	.948	.277	124	.954	.284	79	.952	.271	63
		300	60	0.1	.923	.251	273	.940	.251	137	.949	.262	126	.954	.270	85	.954	.255	66
				0.2	.925	.262	246	.940	.261	130	.948	.272	128	.954	.279	83	.951	.265	66
			90	0.1	.921	.244	293	.941	.244	138	.949	.255	132	.955	.262	87	.953	.249	66
				0.2	.923	.251	267	.941	.250	129	.950	.262	129	.955	.268	84	.952	.255	58
	0.2	100	20	0.1	.922	.330	281	.934	.331	170	.944	.343	151	.954	.354	97	.953	.330	108
				0.2	.922	.353	283	.932	.352	186	.942	.360	146	.951	.370	88	.950	.347	100
			30	0.1	.924	.319	259	.935	.316	158	.943	.321	122	.949	.329	71	.951	.316	101
				0.2	.926	.334	245	.934	.330	161	.941	.335	123	.948	.342	72	.949	.328	89
		200	40	0.1	.920	.294	298	.935	.291	156	.945	.298	111	.951	.304	70	.952	.291	89
				0.2	.923	.308	274	.935	.304	156	.943	.310	115	.949	.317	69	.949	.303	79
			60	0.1	.918	.286	323	.936	.282	153	.946	.288	107	.951	.294	73	.950	.283	90
				0.2	.920	.295	300	.936	.290	150	.946	.296	106	.951	.302	66	.950	.291	77
		300	60	0.1	.914	.279	355	.936	.275	164	.946	.282	112	.951	.288	71	.951	.277	87
				0.2	.919	.289	314	.936	.285	157	.945	.292	112	.951	.298	69	.949	.286	74
			90	0.1	.912	.273	377	.936	.269	164	.947	.275	107	.952	.281	72	.950	.271	84
				0.2	.915	.280	352	.937	.275	157	.947	.281	103	.951	.287	66	.949	.276	72
200 40	0.1	200	40	0.1	.918	.244	318	.939	.249	140	.949	.265	140	.955	.274	98	.959	.252	131
				0.2	.912	.277	385	.934	.277	183	.944	.286	123	.951	.293	75	.951	.277	143
			60	0.1	.937	.222	126	.935	.217	152	.937	.220	144	.942	.225	102	.953	.224	73
				0.2	.937	.240	134	.936	.234	138	.939	.236	126	.942	.240	90	.951	.239	63
		300	60	0.1	.930	.222	199	.932	.221	182	.942	.230	140	.949	.237	88	.956	.226	98
				0.2	.930	.248	204	.934	.244	165	.941	.249	119	.946	.255	79	.950	.247	101
			90	0.1	.939	.203	113	.936	.197	143	.939	.200	126	.943	.204	92	.953	.204	61
				0.2	.939	.217	110	.937	.210	132	.939	.212	119	.942	.216	86	.950	.216	54
	0.2	200	40	0.1	.922	.260	286	.938	.264	137	.948	.279	134	.954	.288	92	.957	.266	102
				0.2	.917	.292	333	.934	.291	168	.945	.299	117	.951	.307	70	.950	.290	113
			60	0.1	.937	.240	128	.937	.234	135	.939	.236	122	.942	.240	89	.951	.239	66
				0.2	.937	.256	134	.937	.250	130	.939	.251	119	.942	.255	87	.949	.253	56
		300	60	0.1	.932	.240	183	.934	.237	162	.941	.245	131	.948	.251	86	.954	.241	79
				0.2	.931	.264	188	.935	.260	152	.940	.264	118	.945	.269	74	.949	.261	80
			90	0.1	.937	.222	130	.937	.215	127	.940	.217	106	.944	.220	81	.951	.221	61
				0.2	.938	.234	119	.938	.227	118	.940	.229	101	.943	.232	75	.949	.232	51

References

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

Agresti, A. and Coull, B. A. (1998). Approximation is better than “exact” for interval estimation of binomial proportions, *The American Statistician,* **52**, 119–126.

Agresti, A. and Min, Y. (2005). Simple improved confidence intervals for comparing matched proportions, *Statistics in Medicine,* **24**, 729–740.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and asymptotics,* Chapman & Hall, London.

Barnett, V., Haworth, J. and Smith, T. M. F. (2001). A two-phase sampling scheme with applications to auditing or sed quis custodiet ipsos custodes? *Journal of Royal Statistical Society, Serie A,* **164**, 407–422.

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

Boese, D. H., Young, D. M. and Stamey, J. D. (2006). Confidence intervals for a binomial parameter based on binary data subject to false-positive misclassification, *Computational Statistics and Data Analysis,* **50**, 3369–3385.

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

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information, *Biometrika,* **65**, 457–482.

Geng, Z. and Asano, C. (1989). Bayesian estimation methods for categorical data with misclassifications, *Communications in Statistics, Theory and Methods,* **18**, 2935–2954.

Hildesheim, A., Mann, V., Brinton, L. A., Szklo, M., Reeves, W. C. and Rawls, W. E. (1991). Herpes simplex virus type 2: A possible interaction with human papillomavirus types 16/18 in the development of invasion cervical cancer, *International Journal of Cancer,* **49**, 335–340.

Lee, S.-C. (2007). An improved confidence interval for the population proportion in a double sampling scheme subject to false-positive misclassification, *Journal of the Korean Statistical Society,* **36**, 275–284.

Lee, S.-C. and Byun, J.-S. (2008). A Bayesian approach to obtain confidence intervals for binomial proportion in a double sampling scheme subject to false-positive misclassification, *Journal of the Korean Statistical Society,* **37**, 393–403.

Lie, R. T., Heuch, I. and Irgens, L. M. (1994). Maximum likelihood estimation of proportion of congenital malformations using double registration systems, *Biometrics,* **50**, 433–444.

Moors, J. J. A., van der Genugten, B. B. and Strijbosch, L. W. G. (2000). Repeated audit controls, *Statistica Neerlandica,* **54**, 3–13.

Price, R. M. and Bonett, D. G. (2004). An improved confidence interval for a linear function of binomial proportions, *Computational Statistics and Data Analysis,* **45**, 449–456.

Raats, V. M. and Moors, J. J. A. (2003). Double-checking auditors: A Bayesian approach, *The Statistician.* **52**, 351–365.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications, *Journal of the American Statistical Association,* **65**, 1350–1361.

York, J., Madigan, D., Heuch, I. and Lie, R. T. (1995). Birth defects registered by double sampling: A Bayesian approach incorporating covariates and model uncertainty, *Applied Statistics,* **44**, 227–242.

Received March 2010; Accepted April 2010