The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.4.697
25*(4), 697–706*

The Role of Artificial Observations in Misclassified Binary Data with Common False-Positive Error

Seung-Chun Lee[1]

[1]Department of Applied Statistics, Hanshin University

(Received June 18, 2012; Revised July 13, 2012; Accepted August 3, 2012)

Abstract

An Agresti-Coull type test is considered for the difference of binomial proportions in two doubly sampled data subject to common false-positive error. The performance of the test is compared with likelihood-based tests. The Agresti-Coull test has many desirable properties in that it can approximate the nominal significance level well, and has comparable power performance with a computational advantage.

Keywords: Agresti-Coull test, likelihood-based tests, profile likelihood, double sampling.

1. Introduction

The test or the interval estimation for the difference of two proportions is often of prime interest in biology, medicine and other fields of scientific research. For instance, many experiments in clinical trials are designed to compare the difference in proportions of responses between a new treatment and an existing treatment. The test or the interval estimation plays a key role in these statistical problems.

Note that the Wald procedure employing the maximum likelihood estimate and its asymptotic variance has been considered as a standard method for these statistical problems; however, the erratic behavior of the Wald procedure has been recognized in recent literature. For example, Agresti and Coull (1998), Brown *et al.* (2001) and Lee (2006a) claimed that the coverage probability of the Wald interval for a binomial proportion is significantly smaller than the nominal level even if the sample size is moderately large. The literature also claimed that the performance of the Wald interval can be improved through the application of Agresti-Coull's approach of "adding two successes and two failures". The strategy worked quite well for the interval estimation of the difference between two proportions as well as the one-sample problem; see, Lee (2006b). We examined the performance of Agresti-Coull type test for a two-sample problem with misclassified binary data.

The misclassified binary data often occurs in clinical trials. For example, suppose that binary observations are obtained by classifying experimental or sampling units into two mutually exclusive

This work was supported by Hanshin University research grant.
[1]Professor, Department of Applied Statistics, Hanshin University, 411 Yangsan-Dong, Osan, Kyunggi-Do 447-791, Korea. E-mail: seung@hs.ac.kr

categories. Usually researcher uses an inerrant device for the classification. However, when the cost of the precise classification is expensive, the researcher often uses an inexpensive but fallible classifier with a supplementary inerrant classifier. The case-control study of Hildesheim *et al.* (1991) examined that invasive cervical cancer can influence exposure to Herpes Simplex Virus(HSV), is an example of the misclassified binary data. For the study, western blot procedure known to be relatively inaccurate but inexpensive in defecting the infection of HSV was applied to about two thousand women in case and control groups. Since the western blot procedure is fallible, it may classify an infected woman as normal (false-negative) and vice versa (false-positive). The observations were exposed to measurement error; in addition, the error rates as well as the true proportion of infection in each group are unestimable. An additional data was necessary. A small subsample from each group was further investigated by the refined western procedure, which is an inerrant but expensive classifier. The sampling scheme employed in the case-control study of double sampling. There are numerous examples taking the advantages of the double sampling scheme; see 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).

Some fallible devices may have only a single 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. 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. Boese *et al.* (2006) gave five likelihood-based confidence intervals in the false-positive misclassification model. Recently a two-sample problem was considered by Lee (2012). He investigated the Agresti-Coull type test for the difference of population proportions using two doubly sampled data. In this paper, we consider the same problem, but assume that false-positive error rates are common. Since the error rate is the characteristic of the fallible classifier, not the characteristic of groups, it would be more realistic to assume that the false-positive errors are common for both groups.

2. Doubly Sampling Model with Common False-Positive Error Rate

2.1. Model

In what follows, we will use the same notation of Lee (2012). That is, for each unit tested by the inerrant device, let $T_i = 1$, if i^{th} unit is recorded positive (or a success), and $T_i = 0$, otherwise. Likewise, for each unit tested by the fallible device, define $F_i = 1$, if i^{th} unit is classified as positive, and $F_i = 0$, otherwise. Then, the proportion of positive can be written as:

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

while the false-positive error rate incurred by the fallible device are defined to be

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

The false-negative error is assumed to be zero in this paper. Also we assumed that the misclassification errors are independent from sampling unit to sampling unit.

Suppose that a random sample of N units is drawn from the population of interest and a subsample of n units is drawn from the main sample. 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} denote the number of units in (t, f). Note that $N - n$ units only tested by a fallible device. Among these $N - n$ units, let x be the number of units tested positively and y be the number of units tested negatively. Then, the joint likelihood of p and ϕ is

$$L(p, \phi; \mathcal{Y}) = C(\mathcal{Y})\,(1-p)^{n_{0\cdot}+y}\,p^{n_{11}}(1-\phi)^{n_{00}+y}\phi^{n_{01}}\pi^{x},$$

where $C(\mathcal{Y}) = n!/(n_{00}!n_{01}!n_{11}!)\binom{N-n}{x}$, $n_{t\cdot} = n_{t0} + n_{t1}$, $\pi = \Pr[F_i = 1] = p + (1-p)\phi$ and $\mathcal{Y}$ represents $(n_{00}, n_{01}, n_{11}, x, y)$.

A two-sample double sampling model 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. Let $\lambda = p_1 - p_2$, then the joint likelihood of λ and $\Theta^* = (p_2, \phi_1, \phi_2)$ can be written as:

$$L(\lambda, \Theta^*; \mathcal{Y}_1, \mathcal{Y}_2) = L(\lambda + p_2, \phi_1; \mathcal{Y}_1)L(p_2, \phi_2; \mathcal{Y}_2). \tag{2.1}$$

Model (2.1) was considered by Lee (2012). He devised an Agresti-Coull type test for λ, compared the performance of the test with likelihood-based tests and concluded that the test is comparable with other computationally expensive likelihood-based tests in terms of power property and approximation of nominal significance level.

(2.1) is adequate for the inference of λ in general; however, once we notice that the false-positive error rate is a characteristic of fallible device, not a characteristic of population, there are cases in which it may be logical to assume $\phi_1 = \phi_2$. For instance, if the same fallible device is applied to obtain both $\mathcal{Y}_1$ and $\mathcal{Y}_2$ as the case-control study of Hildesheim *et al.* (1991), then the data sets probably have common error rate and the joint likelihood of λ and $\Theta = (p_2, \phi)$ would be

$$L(\lambda, \Theta; \mathcal{Y}_1, \mathcal{Y}_2) = L(\lambda + p_2, \phi; \mathcal{Y}_1)L(p_2, \phi; \mathcal{Y}_2). \tag{2.2}$$

The assumption of common false-positive error rate can reduce the dimension of parameter space. However, the reduction of dimension does not mean that model becomes more tractable. On the contrary, the reduction requires much more computational expense. For instance, (2.2) does not admit the closed form maximum likelihood estimators. Nonetheless, we will see that the computational expense could be compensated by the efficiency of an inferential method.

2.2. The observed information and the expected information

Taking logarithm on (2.1), and ignoring unnecessary constant terms we have

$$\begin{aligned}\ell(\lambda, \Theta) = {} & (n_{10\cdot} + y_1)\log(1-\lambda-p_2) + n_{111}\log(\lambda+p_2) + (n_{20\cdot} + y_2)\log(1-p_2) + n_{211}\log p_2 \\ & + (n_{100} + n_{200} + y_1 + y_2)\log(1-\phi) + (n_{101} + n_{201})\log\phi + x_1\log\pi_1 + x_2\log\pi_2,\end{aligned}$$

where $\pi_1 = (1-\lambda-p_2)\phi + (\lambda+p_2)$ and $\pi_2 = (1-p_2)\phi + p_2$. The maximum likelihood estimates are the solutions of following likelihood equations:

$$0 = -\frac{n_{10\cdot} + y_1}{1-\lambda-p_2} + \frac{n_{111}}{\lambda+p_2} + \frac{(1-\phi)x_1}{\pi_1}, \tag{2.3}$$

$$0 = -\frac{n_{10\cdot} + y_1}{1-\lambda-p_2} + \frac{n_{111}}{\lambda+p_2} - \frac{n_{20\cdot} + y_2}{1-p_2} + \frac{n_{211}}{p_2} + (1-\phi)\left(\frac{x_1}{\pi_1} + \frac{x_2}{\pi_2}\right), \tag{2.4}$$

$$0 = -\frac{n_{100} + n_{200} + y_1 + y_2}{1-\phi} + \frac{n_{101} + n_{201}}{\phi} + \frac{(1-\lambda-p_2)x_1}{\pi_1} + \frac{(1-p_2)x_2}{\pi_2}. \tag{2.5}$$

In addition, given the value of λ, the last two equations, (2.4) and (2.5) form the profile likelihood equations.

Note when $n_{itf} = 0$ for some (i, t, f), the maximum likelihood estimates cannot be defined; see Tenenbein (1970) for further details. Similarly, the profile log-likelihood does not admit a unique maximum. A customary remedy to prevent the undefined problem is to add a small number (say 0.005) to null observed counts. See Boese *et al.* (2006). Thus we will add a small number when necessary for the calculation of likelihood or profile likelihood equations.

One may use a solver of nonlinear system of equations such as "NEQNF" or "NEQNJ" of IMSL to obtain the maximum likelihood estimates or the profile likelihood estimates. However, those subroutines often fail to give solutions in our simulation study. Thus, it would be better to employ the algorithm given by Lee (2010) to obtain the maximum likelihood and the profile likelihood estimates. Let $(\hat{\lambda}, \hat{\Theta})$ and $\hat{\Theta}^{\lambda}$ be the solutions of likelihood equations and profile likelihood equations when λ is given, respectively.

The observed information matrix consists of minus the second-order derivatives of $\ell(\lambda, \Theta)$:

$$J_{\lambda\lambda} = J_{\lambda p_2} = \frac{n_{10\cdot} + y_1}{(1-p_1)^2} + \frac{n_{111}}{p_1^2} + \frac{(1-\phi)^2 x_1}{\pi_1^2}, \qquad J_{\lambda\phi} = \frac{x_1}{\pi_1^2},$$
$$J_{p_2p_2} = J_{\lambda\lambda} + \frac{n_{20\cdot} + y_2}{(1-p_2)^2} + \frac{n_{211}}{p_2^2} + \frac{(1-\phi)^2 x_2}{\pi_2^2}, \qquad J_{p_2\phi} = \frac{x_1}{\pi_1^2} + \frac{x_2}{\pi_2^2}$$

and

$$J_{\phi\phi} = \frac{n_{100} + n_{200} + y_1 + y_2}{(1-\phi)^2} + \frac{n_{101} + n_{201}}{\phi^2} + \frac{(1-p_1)^2 x_1}{\pi_1^2} + \frac{(1-p_2)^2 x_2}{\pi_2^2},$$

where $p_1 = \lambda + p_2$. Replacing observed counts by their expectations, we have

$$I_{\lambda\lambda} = I_{\lambda p_2} = \frac{n_1 + (N_1 - n_1)(1-\phi)}{(1-p_1)} + \frac{n_1}{p_1} + \frac{(1-\phi)^2 (N_1 - n_1)}{\pi_1}, \qquad I_{\lambda\phi} = \frac{N_1 - n_1}{\pi_1},$$
$$I_{p_2p_2} = I_{\lambda\lambda} + \frac{n_2 + (N_2 - n_2)(1-\phi)}{(1-p_2)} + \frac{n_2}{p_2} + \frac{(1-\phi)^2 (N_2 - n_2)}{\pi_2}, \qquad I_{p_2\phi} = \frac{N_1 - n_1}{\pi_1} + \frac{N_2 - n_2}{\pi_2}$$

and

$$I_{\phi\phi} = \frac{N_1(1-p_1) + N_2(1-p_2)}{1-\phi} + \frac{n_1(1-p_1) + n_2(1-p_2)}{\phi} + \frac{(1-p_1)^2 (N_1 - n_1)}{\pi_1} + \frac{(1-p_2)^2 (N_2 - n_2)}{\pi_2}.$$

Then the observed information for λ is obtained from

$$J^{\lambda\lambda}(\lambda, \Theta) = J_{\lambda\lambda} - (J_{\lambda p_2}, J_{\lambda\phi}) \begin{pmatrix} J_{p_2p_2} & J_{p_2\phi} \\ I_{p_2\phi} & J_{\phi\phi} \end{pmatrix}^{-1} \begin{pmatrix} J_{\lambda p_2} \\ J_{\lambda\phi} \end{pmatrix}. \tag{2.6}$$

Similarly, the expected information for λ, $I^{\lambda\lambda}(\lambda, \Theta)$ can be obtained through the replacement of J notations in (2.6) by I notations.

2.3. Likelihood-based tests

A large sample theory indicates that $\hat{\lambda}$ is asymptotically normally distributed with mean λ and inverse variance $I^{\lambda\lambda}(\lambda, \Theta)$. Thus, the asymptotic distribution of $(\hat{\lambda} - \lambda)^2 I^{\lambda\lambda}(\lambda, \Theta)$ is a χ^2-distribution

with 1 degree of freedom. However, the existence of nuisance parameter, Θ prevents us from directly using this result. Barndorff-Nielsen and Cox (1994) suggested that $I^{\lambda\lambda}(\lambda,\Theta)$ can be replaced by $I^{\lambda\lambda}(\lambda,\hat{\Theta}^{\lambda}), I^{\lambda\lambda}(\hat{\lambda},\hat{\Theta})$, $J^{\lambda\lambda}(\lambda,\hat{\Theta}^{\lambda})$ and $J^{\lambda\lambda}(\hat{\lambda},\hat{\Theta})$. Hence, for testing $H_0:\lambda=\lambda_0$ *vs.* $H_1:\lambda\neq\lambda_0$, we can setup four Wald-like test statistics,

$$W_{\mathrm{EP}}=\left(\hat{\lambda}-\lambda_0\right)^2 I^{\lambda\lambda}\left(\lambda_0,\hat{\Theta}^{\lambda_0}\right),\quad W_{\mathrm{EM}}=\left(\hat{\lambda}-\lambda_0\right)^2 I^{\lambda\lambda}\left(\hat{\lambda},\hat{\Theta}\right),\quad W_{\mathrm{OP}}=\left(\hat{\lambda}-\lambda_0\right)^2 J^{\lambda\lambda}\left(\lambda_0,\hat{\Theta}^{\lambda_0}\right)$$

and

$$W_{\mathrm{OM}}=\left(\hat{\lambda}-\lambda_0\right)^2 J^{\lambda\lambda}\left(\hat{\lambda},\hat{\Theta}\right),$$

which are known to have asymptotic χ^2-distribution with 1 degree of freedom.

Next four asymptotic tests are based on the score statistic obtained from (2.3)

$$U_{\lambda}\left(\hat{\Theta}^{\lambda_0}\right)=-\frac{n_{10\cdot}+y_1}{1-\hat{p}_1^{\lambda_0}}+\frac{n_{111}}{\hat{p}_1^{\lambda_0}}+\frac{\left(1-\hat{\phi}^{\lambda_0}\right)x_1}{\hat{\pi}_1^{\lambda_0}},$$

where $\hat{p}_1^{\lambda_0}=\lambda_0+\hat{p}_2^{\lambda_0}$, $\hat{\pi}_1^{\lambda_0}=(1-\hat{p}_1^{\lambda_0})\hat{\phi}^{\lambda_0}+\hat{p}_1^{\lambda_0}$, and $\hat{p}_2^{\lambda_0}$ and $\hat{\phi}^{\lambda_0}$ are the solutions of the profile likelihood equations when λ_0 is given. It is also known that the asymptotic distribution of $U_{\lambda}(\hat{\Theta}^{\lambda_0})$ is a normal distribution with mean 0 and variance $I^{\lambda\lambda}(\lambda,\Theta)$ under H_0. As before, four asymptotic tests can be setup as

$$S_{\mathrm{EP}}=\frac{U_{\lambda}^2\left(\hat{\Theta}^{\lambda_0}\right)}{I^{\lambda\lambda}\left(\lambda_0,\hat{\Theta}^{\lambda_0}\right)},\quad S_{\mathrm{EM}}=\frac{U_{\lambda}^2\left(\hat{\Theta}^{\lambda_0}\right)}{I^{\lambda\lambda}\left(\hat{\lambda},\hat{\Theta}\right)},\quad S_{\mathrm{OP}}=\frac{U_{\lambda}^2\left(\hat{\Theta}^{\lambda_0}\right)}{J^{\lambda\lambda}\left(\lambda_0,\hat{\Theta}^{\lambda_0}\right)},\quad S_{\mathrm{OM}}=\frac{U_{\lambda}^2\left(\hat{\Theta}^{\lambda_0}\right)}{J^{\lambda\lambda}\left(\hat{\lambda},\hat{\Theta}\right)}.$$

The last likelihood-based test is due to the well-known log-likelihood ratio statistic,

$$L_{\mathrm{R}}=2\left[\ell\left(\hat{\lambda},\hat{\Theta}\right)-\ell\left(\lambda_0,\hat{\Theta}^{\lambda_0}\right)\right].$$

All of these tests reject the null hypothesis at the significance level α when the observed value of test statistic is greater than the $(1-\alpha)\times 100$ percentile of a χ^2-distribution with 1 degree of freedom.

The Agresti-Coull test stem from W_{EM} which is the Wald test in original sense using the maximum likelihood estimate and its estimate of asymptotic variance. By adding artificial counts to observed counts, and then applying the Wald procedure, we can get an Agresti-Coull test. We will add 0.5 and 1 to each count classified by an inerrant and a fallible device, respectively. That is, let W_{A} be the W_{EM} using artificial observations $x_i^*=x_i+1$, $y_i^*=y_i+1$, $n_{i00}^*=n_{i00}+0.5$, $n_{i01}^*=n_{i01}+0.5$ and $n_{i11}^*=n_{i11}+0.5$ for $i=1,2$. We tried other values of artificial counts, but W_{A} was good in approximating the significance level at 5% significance test.

2.4. An example

The case-control study of Hildesheim *et al.* (1991) aimed to examine that invasive cervical cancer can affect the exposure to Herpes Simplex Virus(HSV). To explore the relationship, a 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 HSV. Since the western blot procedure is fallible, a sub-sample from each group was further investigated by a refined western blot procedure, which is known to be a

Table 2.1. Case-control data of Hildesheim *et al.* (absorbing false-negatives into true-positives)

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

Table 2.2. The observed values of test statistics and p-values for testing $H_0 : \lambda = 0$ *vs.* $H_1 : \lambda \neq 0$ (case-control data of Hildesheim *et al.*)

Test	Common error	No restriction	Test	Common error	No restriction
W_{EP}	22.41 (0.0000)	5.93 (0.0149)	S_{EP}	22.80670 (0.0000)	8.92 (0.0028)
W_{EM}	22.00 (0.0000)	8.48 (0.0036)	S_{EM}	23.22927 (0.0000)	6.23 (0.0125)
W_{OP}	22.25 (0.0000)	5.73 (0.0167)	S_{OP}	22.96908 (0.0000)	9.22 (0.0024)
W_{OM}	22.33 (0.0000)	10.35 (0.0013)	S_{OM}	22.89482 (0.0000)	5.11 (0.0238)
W_{A}	21.87 (0.0000)	7.70 (0.0055)	L_{R}	30.99226 (0.0000)	8.06 (0.0045)

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. The data are shown in Table 2.1.

This data was analyzed by Lee (2012). Presumably, he believed that there was no real restriction on the parameters. However, it may be logical to assume that the false-positive error rate is common, since the same fallible device was applied to the case and control groups. Under this assumption, the maximum likelihood estimate of λ is -0.1307; however, without the assumption it is -0.1566. Similarly the maximum likelihood estimate of ϕ is 0.1523, but it is 0.1633 and 0.1198 for control and case groups, respectively. The asymptotic variances of these estimates are 0.00039 and 0.00110. These estimates support the assumption of a common false-positive error rate. The p-value of a significance test for error rates was 0.260. For testing $H_0 : \lambda = 0$ against $H_1 : \lambda \neq 0$, the test statistics and p-values were calculated with and without the assumption (see Table 2.2). It can be seen that the tests under the assumption provide larger observed values and smaller p-values than corresponding tests with no restrictions.

3. Comparison of Tests

The tests considered in this paper are based on a large sample theory. The sizes of tests would eventually converge to the nominal level as the sample size increased. However, when the sample size is not large, the actual sizes of tests may not approximate to the nominal level well. To see this, under various configurations of parameter values, the sizes of tests were estimated with 1,000,000 random samples when $p\,(= p_1 = p_2) = 0.1, 0.3$ and 0.5. The results are shown in Table 3.1. Because of the duality between p and $q\,(= 1 - p)$, the results for large values p can be inferred from the table. Since the false-positive error rate is not large in general, we only considered small values of ϕ, say 0.1 or 0.2.

It can be observed that $W_{\mathrm{EP}}, W_{\mathrm{EM}}, W_{\mathrm{OM}}, S_{\mathrm{OP}}$ and L_{R} have a tendency to too often reject true null hypothesis too often. In other words, they are liberal. Since a philosophy of hypothesis testing is to control the maximum level of type I error, the liberality could be a defect of test. However, to be fair, we will focus on the approximation itself. Nonetheless, $W_{\mathrm{EM}}, W_{\mathrm{OM}}$ and S_{OP} seem to be undesirable, because the sizes of these tests are considerably larger than the nominal level when

Table 3.1. Estimated sizes of tests for testing $H_0 : \lambda = 0$ *vs.* $H_1 : \lambda \neq 0$ at 0.05 significance level

p	ϕ	Group 1		Group 2		Size of test									
		N_1	n_1	N_2	n_2	W_{A}	W_{EP}	W_{EM}	W_{OP}	W_{OM}	S_{EP}	S_{EM}	S_{OP}	S_{OM}	L_{R}
0.1	0.1	100	20	100	20	0.0413	0.0673	0.0768	0.0445	0.0731	0.0495	0.0417	0.0751	0.0449	0.0597
				200	40	0.0440	0.0597	0.0785	0.0444	0.0706	0.0493	0.0432	0.0665	0.0452	0.0580
				300	60	0.0446	0.0576	0.0802	0.0465	0.0696	0.0497	0.0443	0.0636	0.0455	0.0566
		200	40	200	40	0.0475	0.0571	0.0621	0.0462	0.0589	0.0502	0.0463	0.0602	0.0481	0.0554
				300	60	0.0477	0.0546	0.0591	0.0466	0.0561	0.0497	0.0460	0.0572	0.0478	0.0537
	0.2	100	20	100	20	0.0420	0.0652	0.0832	0.0367	0.0866	0.0482	0.0392	0.0947	0.0417	0.0649
				200	40	0.0451	0.0607	0.0933	0.0419	0.0890	0.0492	0.0428	0.0784	0.0435	0.0614
				300	60	0.0467	0.0597	0.1031	0.0464	0.0923	0.0496	0.0445	0.0710	0.0445	0.0600
		200	40	200	40	0.0495	0.0590	0.0671	0.0423	0.0657	0.0499	0.0443	0.0669	0.0454	0.0582
				300	60	0.0497	0.0569	0.0640	0.0445	0.0615	0.0498	0.0449	0.0616	0.0457	0.0562
0.3	0.1	100	20	100	20	0.0484	0.0516	0.0536	0.0497	0.0527	0.0504	0.0483	0.0523	0.0492	0.0518
				200	40	0.0488	0.0510	0.0535	0.0497	0.0530	0.0503	0.0481	0.0519	0.0486	0.0509
				300	60	0.0490	0.0507	0.0537	0.0496	0.0532	0.0502	0.0474	0.0514	0.0478	0.0507
		200	40	200	40	0.0491	0.0508	0.0518	0.0499	0.0515	0.0504	0.0493	0.0513	0.0496	0.0508
				300	60	0.0494	0.0506	0.0516	0.0499	0.0514	0.0502	0.0494	0.0511	0.0496	0.0507
	0.2	100	20	100	20	0.0497	0.0536	0.0561	0.0513	0.0547	0.0511	0.0483	0.0533	0.0495	0.0524
				200	40	0.0494	0.0514	0.0542	0.0499	0.0533	0.0501	0.0472	0.0515	0.0480	0.0520
				300	60	0.0498	0.0514	0.0549	0.0503	0.0540	0.0503	0.0466	0.0514	0.0475	0.0513
		200	40	200	40	0.0498	0.0515	0.0526	0.0504	0.0521	0.0505	0.0493	0.0515	0.0498	0.0514
				300	60	0.0499	0.0512	0.0523	0.0504	0.0519	0.0505	0.0493	0.0512	0.0496	0.0511
0.5	0.1	100	20	100	20	0.0481	0.0524	0.0535	0.0510	0.0535	0.0524	0.0510	0.0537	0.0511	0.0531
				200	40	0.0487	0.0511	0.0526	0.0499	0.0526	0.0511	0.0489	0.0516	0.0490	0.0507
				300	60	0.0491	0.0506	0.0527	0.0497	0.0526	0.0505	0.0481	0.0510	0.0483	0.0513
		200	40	200	40	0.0488	0.0500	0.0504	0.0495	0.0503	0.0499	0.0495	0.0504	0.0496	0.0506
				300	60	0.0493	0.0505	0.0512	0.0500	0.0511	0.0504	0.0498	0.0510	0.0499	0.0505
	0.2	100	20	100	20	0.0477	0.0516	0.0523	0.0495	0.0519	0.0510	0.0502	0.0529	0.0507	0.0515
				200	40	0.0480	0.0505	0.0517	0.0490	0.0516	0.0501	0.0487	0.0515	0.0490	0.0510
				300	60	0.0490	0.0508	0.0528	0.0497	0.0526	0.0506	0.0486	0.0517	0.0487	0.0506
		200	40	200	40	0.0488	0.0507	0.0510	0.0497	0.0509	0.0504	0.0501	0.0515	0.0503	0.0510
				300	60	0.0490	0.0506	0.0510	0.0497	0.0509	0.0504	0.0500	0.0512	0.0501	0.0504

the sample size is small. The approximation of the other tests seems to be acceptable in that the maximum difference is less than 0.01.

W_{EM} is too liberal when p is small; however, W_{A} is moderately conservative and has better approximations than W_{EM}. Thus, we may conclude that W_{EM} can be improved by adding small artificial counts to the observed counts. The sizes of W_{A} are quite close to the nominal level and comparable to other tests in the approximation.

Note that Efron and Hinkley (1978) claimed that the observed information is preferable form than the expected information in general. This may be true for the Wald-like tests. W_{OP} and W_{OM} which are scaled by the observed information have slightly better approximations than W_{EP} and W_{EM}, respectively. However, it may not be true for the score based tests. S_{EP} seems to be best among all likelihood-based tests in the approximation. Thus, we cannot conclude the preference between the observed information and the expected information. However, we may conclude that the information using the profile estimates is more suitable than the likelihood estimates for the testing problem.

The size of L_{R} is larger than those of $W_{\mathrm{A}}, W_{\mathrm{OP}}, S_{\mathrm{EP}}, S_{\mathrm{EM}}$ and S_{OM} for almost all cases examined in our study. It seems that L_{R} is too liberal, and we may exclude it from our consideration. However, we could not identify the preference among those tests, since none of these tests is uniformly better than other tests in the approximation, and their power properties are very similar (see Figure 3.1). We investigated the power of tests under various parameter values, and found that the tests have

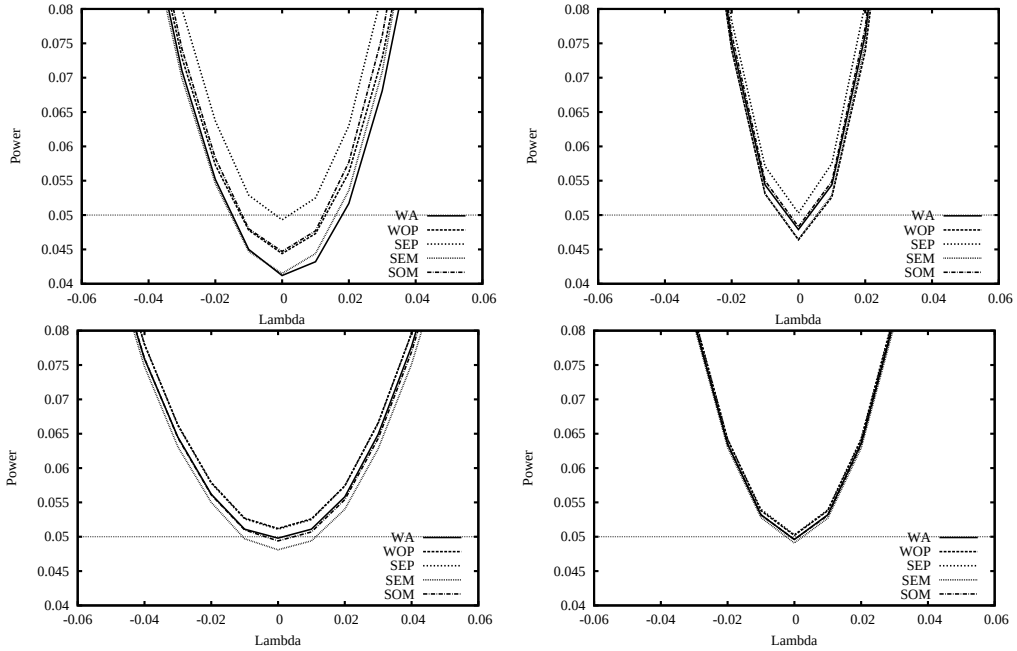

Figure 3.1. Power of $W_{\mathrm{A}}, W_{\mathrm{OP}}, S_{\mathrm{EP}}, S_{\mathrm{EM}}, S_{\mathrm{OP}}$ and L_{R} for testing $H_0 : \lambda = 0$ against $H_1 : \lambda \neq 0$ when $N_1 = N_2 = 100, n_1 = n_2 = 20$ (left) and $N_1 = N_2 = 200, n_1 = n_2 = 40$ (right) with $p = 0.1, \phi_1 = 0.1$ (top) and $p = 0.3, \phi_2 = 0.2$ (bottom).

similar powers if we take account of the actual size of the tests. Thus, we conclude that the tests are equally good for the testing problem.

The performance of W_{EM} is not good, but it has a computational advantage compared with other likelihood-based tests in that they do not require solving the profile likelihood equations that are more difficult to solve than likelihood equations. W_{A} can enjoy this advantage as well, because W_{A} has a computational advantage, and is comparable to other likelihood-based tests in view of approximation and power. These may justify the Agresti-Coull type test using artificial observations; subsequently, we can conclude that W_{A} is a preferable test to test the difference of the two proportions in two doubly sampled data that is subject to a common false-positive error.

4. Conclusion

The assumption of common error rates can simplify the model by reducing the number of nuisance parameters; however, it does not make a statistical problem more tractable. Rather, the assumption requires more computational expense. However, in the nature of double sampling design for binary data, it customarily happens that the error rates should be the same in two doubly sampled data. Researchers often do not notice this, and apply a standard method suitable to the model with no restriction on parameters. For instance, the error rates incurred by a fallible device should be the same in the case-control study of Hildesheim *et al.* (1991), but we believe Lee (2012) analyzed the data improperly in that the power of the tests employed is unsatisfying.

To demonstrate this, we estimate the power of Agresti-Coull type test under both the common error

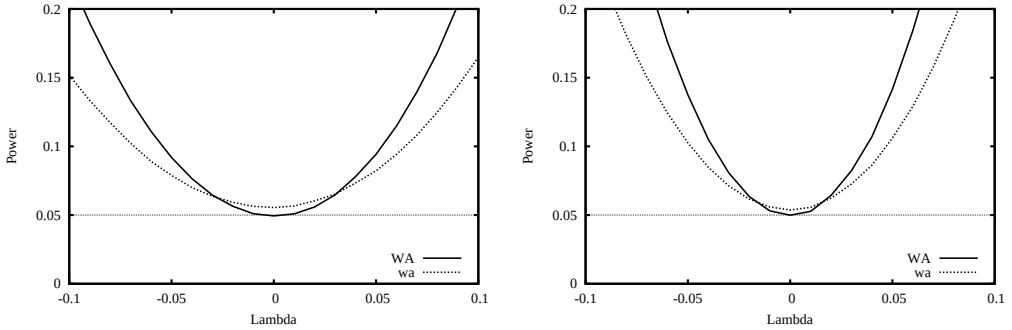

Figure 4.1. Power of W_{A} under the common error rate and the no restriction models for testing $H_0 : \lambda = 0$ against $H_1 : \lambda \neq 0$ when $p = 0.3, \phi = 0.2, N_1 = N_2 = 100, n_1 = n_2 = 20$ (left) and $p = 0.3, \phi = 0.2, N_1 = N_2 = 200, n_1 = n_2 = 40$ (right). Capital and small letters represent the common error and the no restriction models, respectively.

rate and no restriction models when the nuisance parameters are $p_2 = 0.3$ and $\phi_1 = \phi_2 = 0.2$. We consider two cases of sample size, $N_1 = N_2 = 100, n_1 = n_2 = 20$ and $N_1 = N_2 = 200, n_1 = n_2 = 40$. The results are shown in Figure 4.1. The power of W_{A} under the assumption of common error rate is represented by a solid line. It can be observed that the power of Agresti-Coull test designed for a general purpose (*i.e.*, no restriction on parameters), is significantly lower than that of W_{A} suitable to common error rate model. Since the Agresti-Coull type test and other likelihood-based tests have similar power pattern, we may apply this observation to other tests as well.

References

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

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*, Chapman and 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, Series A*, **164**, 407–422.

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. (2006a). Interval estimation of binomial proportions based on weighted Polya posterior, *Computational Statistics & Data Analysis*, **51**, 1012–1021.

Lee, S.-C. (2006b). The weighted Polya posterior confidence interval for the difference between two independent proportions, *The Korean Journal of Applied Statistics*, **19**, 171–181.

Lee, S.-C. (2010). Likelihood based confidence intervals for the difference of proportions in two doubly sampled data with a common false-positive error rate, *Communications of the Korean Statistical Society*, **17**, 679–688.

Lee, S.-C. (2012). The role of artificial observations in testing for the difference of proportions in misclassified binary data, *The Korean Journal of Applied Statistics*, **25**, 513–520.

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.

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.