The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.3.513
25*(3), 513–520*

The Role of Artificial Observations in Testing for the Difference of Proportions in Misclassified Binary Data

Seung-Chun Lee[1]

[1]Department of Applied Statistics, Hanshin University

(Received May 1, 2012; Revised May 31, 2012; Accepted June 12, 2012)

Abstract

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

Keywords: Agrestt-Coull interval, double sampling, profile likelihood, Rao score.

1. Introduction

The Wald interval using the maximum likelihood estimate of the binomial parameter is considered the standard method for the interval estimations of binomial proportions; however, the erratic behavior of the coverage probability of the Wald interval has been recognized in various literature, see for example, Blyth and Still (1983), Agresti and Coull (1998), and Brown *et al.* (2001). In particular, Brown et al. investigated the unsatisfactory coverage properties of the Wald interval in detail, and Agresti and Coull showed that an improved interval for the parameter of a binomial distribution could be obtained by "adding two successes and two failures" to the observed counts and then using the standard method.

This strategy works quiet 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

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

sampling design subject to false-positive misclassification. He compared the performance of the Agresti-Coull type confidence interval with the original Wald interval and the confidence interval given by Boese *et al.* (2006). Again, the Agresti-Coull type interval is comparable to or even better than the Wald interval in terms of the closeness of coverage probability to nominal level.

The interval estimation is closely related to the hypothesis test. In this paper, we will provide a simple but effective test for the difference of population proportions with double sampled data subject to false-positive classification relying on the Agrest-Coull's argument.

A double sampling scheme on binary observations occurs when the cost of the precise test is expensive. To reduce the cost, a large sample is classified by an inexpensive but fallible device and a subsample is classified by a supplementary inerrant device.

A significant amount of literature 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 Moor (2003), Boese *et al.* (2006) and Lee (2011). In particular, Lee (2011) showed that the Agresti-Coull's approach can be justified by the Bayesian paradigm.

Note that among the false-positive and false-negative errors, only one type error model occurs frequently in real world. For example, Moors *et al.* (2000) and Perry *et al.* (2000) analyzed an auditing data and blood testing data, respectively, where each data has only one type of error. the false-positive error model can represent the one type of error model, since the role of false-positive error is switchable to the false-negative errors. In this paper, we will apply the Agresti-Coull's approach to test the difference of two proportions with double sampled data subject to false-positive error. The power performance of the proposed test is compared with well-known likelihood based tests.

In Section 2 and Section 3, we will briefly describe the model and tests considered in this paper. We also give an example for the tests. The comparison of tests is shown in Section 4 with some conclusions.

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

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

and the false-positive error rate is

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

The false-negative error rate, $\Pr\left[F_i = 0 | T_i = 1\right]$, 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

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

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:

$$\hat{p} = \frac{n_{11}}{n_{01}+n_{11}} \frac{x+n_{01}+n_{11}}{N} \quad \text{and} \quad \hat{\phi} = \frac{n_{01}}{n_{01}+n_{11}} \frac{x+n_{01}+n_{11}}{N(1-\hat{p})} \tag{2.1}$$

with

$$\widehat{\mathrm{Var}}(\hat{p}) = \frac{\hat{p}\,\hat{q}}{n} - \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n_{11}}{n_{11}+n_{01}} \hat{p}(1-\hat{\pi}) \tag{2.2}$$

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. Let $\lambda = p_1 - p_2$. Then, the joint likelihood of λ and $\Theta = (p_2, \phi_1, \phi_2)$ can be written as:

$$L(\lambda, \Theta; \mathcal{Y}_1, \mathcal{Y}_2) = L(\lambda + p_2, \phi_1; \mathcal{Y}_1) L(p_2, \phi_2; \mathcal{Y}_2). \tag{2.3}$$

3. Tests

In this section, we will define an Agresti-Coull type test for testing $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda \neq \lambda_0$ in the two sample false-positive misclassification model. In addition, we will review likelihood-based tests. For this purpose the profile likelihood and the information are important in what follows.

3.1. Profile likelihood and information

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

Taking logarithm of (2.3), we have the full log-likelihood,

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

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 $\ell_P(\Theta; \lambda)$ 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:

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

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

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

Note that when all observed counts are greater than zero, $\hat{p}_2^\lambda$ lies in interval $(\max\{-\lambda, 0\}, \min\{1 - \lambda, 1\})$, which in turn results in $\hat{\phi}_1^\lambda \in (0, 1)$ and $\hat{\phi}_2^\lambda \in (0, 1)$. For this case, one may refer Lee (2010) for solving the nontrivial profile equations. 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 $n_{201} = 0$, $\hat{p}_2$ or $\hat{\phi}_2$ 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

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

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 $I^{\lambda\lambda}(\lambda, \Theta)$ is obtained by replacing observed counts by their expectations. However, Efron and Hinkley (1978) claimed that the observed information is preferable form to the expected information in general. In fact, tests using the adjusted restricted information did not have good features in our simulation study. Thus we do not consider tests based the adjusted restricted information.

3.2. Asymptotic tests

The first likelihood-based test considered in this paper is the Wald test which rejects the null hypothesis when

$$W = \frac{(\hat{\lambda} - \lambda_0)^2}{\widehat{\mathrm{Var}}(\hat{\lambda})} \tag{3.4}$$

is greater than $\chi^2_{1,\alpha}$, where $\hat{\lambda} = \hat{p}_1 - \hat{p}_2$ and $\widehat{\mathrm{Var}}(\hat{\lambda}) = \widehat{\mathrm{Var}}(\hat{p}_1) + \widehat{\mathrm{Var}}(\hat{p}_2)$ which are obtained using (2.1) and (2.2), and $\chi^2_{1,\alpha}$ represents the $1 - \alpha$ quantile of a χ^2-distribution with 1 degree of freedom. The Wald test may be the most popular, but it does not approximate the nominal level well in many sampling designs. Thus, we are doubtful of the usability of (3.4). However, some adjustments

Table 3.1. 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

of the Wald test may have good statistical properties as shown in Agresti and Coull (1998) and Brown *et al.* (2001). We may expect better performance by adding some artificial observations to observed counts and applying the Wald procedure. The theoretical justification of these artificial observations in the double sampling model can be found in Lee and Byun (2008). We add 0.5 to each n_{ijk}, but add 1 to x_i and $y_i, i = 1, 2$. Thus, the total sample size in each group is increased by 3.5. This test will be denoted by W_{A}.

The log-likelihood ratio test reject the null hypothesis when

$$L_R = 2\left[\ell\left(\hat{\lambda}, \hat{p}_2, \hat{\phi}_1, \hat{\phi}_2\right) - \ell_P\left(\hat{p}_2^{\lambda_0}, \hat{\phi}_1^{\lambda_0}, \hat{\phi}_2^{\lambda_0}; \lambda_0\right)\right]$$

is greater than $\chi^2_{1,\alpha}$. It is well-known that the test has many nice statistical properties.

A large sample theory also indicates that $\hat{\lambda}$ is asymptotically normally distributed with mean λ and inverse variance $I^{\lambda\lambda}(\lambda, \Theta)$. Because of nuisance parameter Θ, we cannot use this result directly. Barndorff-Nielsen and Cox (1994) suggested that $J^{\lambda\lambda}(\lambda, \Theta^{\lambda})$ or $J^{\lambda\lambda}(\hat{\lambda}, \Theta^{\hat{\lambda}})$ can replace $I^{\lambda\lambda}(\lambda, \Theta)$. This suggestion gives two Wald-like test statistics

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

Next two tests are based on the Rao's 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

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

Then, weighting by $J^{\lambda\lambda}(\lambda_0, \Theta^{\lambda_0})$ and $J^{\lambda\lambda}(\hat{\lambda}, \Theta^{\hat{\lambda}})$, we have two tests

$$S_{\mathrm{OP}} = s\left(\hat{\Theta}^{\lambda_0}; \lambda_0\right)\left(J^{\lambda\lambda}\left(\lambda_0, \Theta^{\lambda_0}\right)\right)^{-1} \quad \text{and} \quad S_{\mathrm{OM}} = s\left(\hat{\Theta}^{\lambda_0}; \lambda_0\right)\left(J^{\lambda\lambda}\left(\hat{\lambda}, \Theta^{\hat{\lambda}}\right)\right)^{-1}.$$

Both tests also reject the null hypothesis when their observed values are greater than $\chi^2_{1,\alpha}$.

3.3. An example

The case-control study of Hildesheim *et al.* (1991) aimed to examine if invasive cervical cancer can influence exposure to the 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 infections. 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

Table 4.1. Simulated significance level when $p_1 = 0.3, p_2 = 0.3$ and $\alpha = 0.05$

Group 1		Group 2			Simulated significance level						
N_1 n_1	ϕ_1	N_2	n_2	ϕ_2	W	W_A	W_OP	W_OM	S_OP	S_OM	L_R
100 20	0.1	100	20	0.1	0.0760	0.0460	0.0312	0.0723	0.1149	0.0424	0.0650
				0.2	0.0796	0.0521	0.0384	0.0757	0.0914	0.0384	0.0646
			30	0.1	0.0716	0.0473	0.0399	0.0682	0.0919	0.0448	0.0615
				0.2	0.0719	0.0497	0.0415	0.0685	0.0832	0.0427	0.0606
		200	40	0.1	0.0737	0.0476	0.0378	0.0701	0.1071	0.0407	0.0617
				0.2	0.0728	0.0501	0.0395	0.0694	0.0962	0.0413	0.0611
			60	0.1	0.0743	0.0480	0.0392	0.0712	0.1105	0.0382	0.0604
				0.2	0.0727	0.0488	0.0396	0.0697	0.1026	0.0391	0.0598
	0.2	100	20	0.1	0.0798	0.0521	0.0384	0.0757	0.0915	0.0385	0.0647
				0.2	0.0806	0.0558	0.0426	0.0766	0.0802	0.0375	0.0643
			30	0.1	0.0764	0.0541	0.0456	0.0731	0.0735	0.0385	0.0615
				0.2	0.0754	0.0553	0.0459	0.0719	0.0705	0.0383	0.0608
		200	40	0.1	0.0805	0.0565	0.0441	0.0771	0.0774	0.0329	0.0622
				0.2	0.0778	0.0566	0.0450	0.0745	0.0743	0.0350	0.0612
			60	0.1	0.0811	0.0573	0.0435	0.0781	0.0760	0.0292	0.0608
				0.2	0.0789	0.0571	0.0443	0.0759	0.0745	0.0314	0.0608
200 40	0.1	200	40	0.1	0.0622	0.0484	0.0415	0.0597	0.0646	0.0513	0.0576
				0.2	0.0637	0.0520	0.0466	0.0612	0.0596	0.0469	0.0563
			60	0.1	0.0599	0.0489	0.0443	0.0580	0.0606	0.0495	0.0555
				0.2	0.0604	0.0505	0.0465	0.0584	0.0586	0.0488	0.0552
		300	60	0.1	0.0610	0.0492	0.0422	0.0588	0.0637	0.0496	0.0566
				0.2	0.0611	0.0508	0.0453	0.0590	0.0599	0.0487	0.0557
			90	0.1	0.0610	0.0497	0.0424	0.0592	0.0636	0.0483	0.0562
				0.2	0.0601	0.0499	0.0438	0.0583	0.0606	0.0485	0.0554
	0.2	200	40	0.1	0.0633	0.0518	0.0464	0.0607	0.0590	0.0467	0.0558
				0.2	0.0639	0.0539	0.0483	0.0615	0.0572	0.0444	0.0552
			60	0.1	0.0618	0.0527	0.0478	0.0597	0.0558	0.0442	0.0539
				0.2	0.0617	0.0534	0.0488	0.0598	0.0555	0.0448	0.0542
		300	60	0.1	0.0631	0.0531	0.0473	0.0608	0.0572	0.0439	0.0547
				0.2	0.0623	0.0537	0.0480	0.0601	0.0557	0.0441	0.0541
			90	0.1	0.0637	0.0547	0.0476	0.0617	0.0563	0.0426	0.0544
				0.2	0.0626	0.0542	0.0480	0.0606	0.0557	0.0433	0.0541

relatively accurate procedure. Originally the fallible procedure is exposed to the two types of error, however 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 3.1.

We found the maxmum likelihood estimates of λ is $\hat{\lambda} = -0.1566$ with standard error 0.0538, while the artificial observations adjusted it to $\tilde{\lambda} = -0.1511$ with standard error 0.0545. Thus, W and W_A were calculated as 8.480 and 7.700. Since $\chi^2_{1,0.05} = 3.8416$, the null hypothesis is rejected by both tests with p-values 0.0036 and 0.0055, respectively. The other tests were $W_\mathrm{OP} = 5.732, W_\mathrm{OM} = 10.346, S_\mathrm{OP} = 9.223, S_\mathrm{OM} = 5.1097$ and $L_\mathrm{R} = 8.061$. These give p-values as 0.0167, 0.0013, 0.0024 and 0.0045, respectively. Thus, we can reject the null hypothesis at the 5% significance level. That is, we may conclude that invasive cervical cancer would affect exposure to HSV. However, the tests reported different p-values. Since they are all asymptotic tests, actual levels of test are not the nominal level. In fact, actual level of some tests is quite different from the nominal level 0.05 in our simulation study.

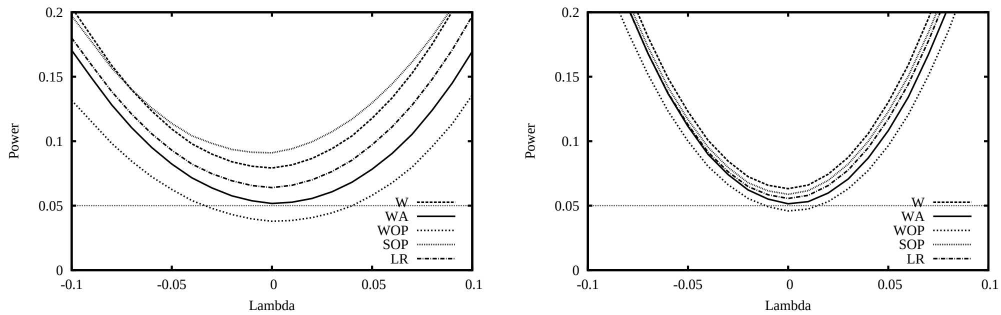

Figure 4.1. Power of $W, W_{\mathrm{A}}, W_{\mathrm{OP}}, S_{\mathrm{OP}}$ and L_{R} for testing $H_0 : \lambda = 0$ against $H_1 : \lambda \neq 0$ when $N_1 = N_2 = 100, n_1 = n_2 = 20, \phi_1 = 0.1, \phi_2 = 0.2$ and $p_1 = 0.3$ (left), and $N_1 = N_2 = 200, n_1 = n_2 = 40, \phi_1 = 0.1, \phi_2 = 0.2$ and $p_1 = 0.3$ (right).

4. Comparison of Tests and Conclusions

Note that $W_{\mathrm{OP}}, S_{\mathrm{OP}}$ and L_{R} require to solve the profile likelihood equations, which are computationally expensive. Thus, one may prefer relatively simple W, W_{A} or W_{OM}. In particular W and W_{A} can be obtained simply by a calculator and hence computationally most preferable. However, the difference between the actual level of W and the nominal level is quite big in our simulation study. For instance, when $N_1 = N_2 = 100, n_1 = n_2 = 20, \phi_1 = \phi_2 = 0.1$ and $p_1 = p_2 = 0.3$, the actual level of test based on W was estimated to 0.076. We estimated the actual levels of tests under various configurations of parameter values with each 1,000,000 random samples. These results are shown in Table 4.1. If we increase sample size, then the size of tests considered in this paper would eventually converge to the nominal level. Thus, we compared the actual size of tests in relatively small or moderately large samples.

Some messages of Table 4.1 are quite clear. For instance, W cannot approximate the nominal level well, even if sample size is moderately large, while W_{A} has the ability in approximating the nominal level compared with other likelihood-based tests. L_{R} also gives good approximations.

The power property of $W_{\mathrm{A}}, W_{\mathrm{OM}}, S_{\mathrm{OM}}$ and L_{R} is similar in that they have near same shape of power curve as shown Figure 4.1. For small sample size, the power of W_{A} is slightly lower than W, S_{OP} and $L\mathrm{R}$, which is because they make more type I errors than nominal level achieving more power (4.1, left). We also examined the power of tests under various configurations of parameter values, but the power patterns were not changed dramatically. We may conclude that W_{A} is a desirable test in approximation and power property with computational simplicity.

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 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.

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. (2010). Confidence intervals for the difference of binomial proportion in two double sampled data, *Communications of the Korean Statistical Society*, **17**, 309–318.

Lee, S.-C. (2011). Theoretical considerations for the Agresti-Coull type confidence intervals in misclassified binary data, *Communications of the Korean Statistical Society*, **18**, 445–455.

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.

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

Perry, M., Vakil, N. and Cutler, A. (2000). Admixture with whole blood does not explain false-negative urease tests, *Journal of Clinical Gastroenterology*, **30**, 64–65.

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.