Thailand Statistician
July 2017; 15(2): 149-156
http://statassoc.or.th
Contributed paper

Odds Ratio Estimation in Rare Data by Empirical Bayes Method

Kobkun Raweesawat [a], Yupaporn Areepong [a], Saowanit Sukparungsee [a] and Katechan Jampachaisri*[b]

[a] Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bang Sue, Bangkok 10800, Thailand.

[b] Department of Mathematics, Faculty of Science, Naresuan University, Phitsanuloke 65000, Thailand.

*Corresponding author; e-mail: katechanj@nu.ac.th

Received: 20 September 2016
Accepted: 26 October 2016

Abstract

The Empirical Bayes estimator (EB) of odds ratio in rare data is considered in this paper. The proposed estimate of odds ratio based on EB in Poisson distribution to approximate binomial distribution is then compared to conventional method, modified maximum likelihood estimator (MMLE), using the Estimated Relative Error (ERE) as a criterion of comparison. The result indicated that the EB estimator is a more efficient method than MMLE.

Keywords: Odds ratio, empirical Bayes, Poisson distribution, modified maximum likelihood estimator.

1. Introduction

A measure of association provides an index of how strongly the two factors under study were related. The odds ratio is a measure of association between two independent groups with binary outcome. Binary outcome usually refers to success or failure for example passing or failing an exam, good or bad conditions: two independent groups can be either treatment and control groups, or two treatment groups. The outcome in each group can be obtained by counting the number of successes in entire trials. The odds ratio is then used to compare the relative odds of the occurrence of the outcomes of interest (e.g. disease or disorder) in both groups, given exposure to the variable of interest (e.g. health characteristic, aspect of medical history).

The usual maximum likelihood estimator of odds ratio is defined as

$$\widehat{OR}_{MLE} = \frac{odds_1}{odds_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_{2)}}, \tag{1}$$

where $p_1 = x_1/n_1$ and $p_2 = x_2/n_2$ are probabilities of success in group 1 and group 2 and x_1 and x_2 are number of successes in group 1 and 2, respectively. The odds ratio can be 0 or ∞ if one or two of the observed data is 0 ($\widehat{OR}_{MLE} = 0$ if the numerator is 0, and $\widehat{OR}_{MLE} = \infty$ if the denominator is 0). If there is a 0 in both the numerator and denominator, then $\widehat{OR}_{MLE}$ is undefined. Haldane

(1955), and Gart and Zweifel (1967) preferred solutions by adding 0.5 to each cell, and calculate the odds ratio called the modified maximum likelihood estimate (MMLE)

$$\widehat{OR}_{MMLE} = \frac{(x_1 + 0.5)(n_2 - x_2 + 0.5)}{(x_2 + 0.5)(n_1 - x_1 + 0.5)}. \quad (2)$$

Even though MMLE solves the problem of having zeroes in the observed data, Bishop, Fienberg, and Holland (1975) and Agresti and Yang (1987) argued that original data is perturbed by adding 0.5 to each cell. Hitchcock (1962) suggested adding 0.25 to each cell instead of 0.5, similar to the study by Hauck, Anderson and Leahy (1982). Jewell (1984, 1986) recommended adding 1 in the cells x_2 and $n_1 - x_1$. As we can see, several attempts to search for an appropriate correction term have been mainly studied in order to make an odds ratio estimation with zero cell count plausible. However, there is no obvious conclusion and some researchers still disagree on the concept of data perturbation by adding a correction term to each cell.

This paper is concerned with estimating the odds ratio when the outcome under study is rare resulting in which a situation whereby the incidence of the outcome for both independent groups is less than 10% (a mathematical cutoff point) (Wu 2002) or small proportions. Wakeel and Aslam (2013) estimated mean of rare sensitive attribute using Bayes estimator, which Gamma distribution has been used as prior information. They compared Bayes estimator to the maximum likelihood estimator using mean square error, and found that Bayes estimator is more efficient than maximum likelihood estimator. The Poisson distribution is most often invoked for binomial approximation in a rare event, of which the rate of occurrence is small. Subsequently, EB is utilized to obtain the probability of success in each group. Our purposed estimation does not interfere with the original data, and the result of this study tends to outperform the conventional estimator, MMLE.

The remainder of this paper has been arranged in the following sequence. Section 2 discusses the odds ratio estimation using EB method. Section 3 illustrates simulated results, and the efficiency of EB is compared with MMLE. The application to real data set is presented in Section 4, and some concluding remarks are given in Section 5.

2. Approximation Solution of Odds Ratio: The EB Method

In this section, a new approximation method for rare data is proposed using EB. Data are assumed to be Poisson distribution. Let x_1 and x_2 be random variables, distributed as Binomial, but approximated as Poisson with equal and unequal sample sizes, $x_1 \sim poi(\lambda_1)$ and $x_2 \sim poi(\lambda_2)$, where λ_1 and λ_2 denote unknown mean of occurrence. The informative prior is adopted on $\lambda_i, \lambda_i \sim gamma(\eta_i, \tau_i), i = 1,2$ where η_i and τ_i denote hyper-parameters. The estimation of hyper-parameters can be obtained from the posterior marginal distribution function as follow,

$$\begin{aligned} m(\boldsymbol{x}|\eta,\tau) &= \int_0^\infty f(\boldsymbol{x}|\lambda)\pi(\lambda)d\lambda \\ &= \binom{x+\eta-1}{x}\left(\frac{1}{\tau+1}\right)^{\eta}\left(\frac{\tau}{\tau+1}\right)^{x}. \end{aligned} \quad (3)$$

Then, both hyper-parameters in each group can be estimated using maximum likelihood method. The likelihood function of posterior marginal distribution function is displayed as

$$l(\boldsymbol{x}|\eta,\tau) = \prod_{i=1}^{n}\binom{x_i+\eta-1}{x_i}\left(\frac{1}{\tau+1}\right)^{\eta}\left(\frac{\tau}{\tau+1}\right)^{x_i}$$

$$\ln l\left(\boldsymbol{x}\middle|\eta,\tau\right)=\sum_{i=1}^{n}\ln\Gamma\left(x_i+\eta\right)-n\eta\ln\left(\tau+1\right)+\sum_{i=1}^{n}x_i\ln\left(\tau\right)-\sum\ln\Gamma\left(x_i+1\right)-n\ln\Gamma\left(\eta\right)+\sum_{i=1}^{n}x_i\ln\left(\tau+1\right).$$

Applying Newton-Raphson method to solve a nonlinear equation, therefore the $\left(r+1\right)^{th}$ maximum likelihood estimator of hyper-parameters $\left(r=1,2,3,\ldots\right)$ can be obtained from

$$\begin{bmatrix}\hat{\eta}^{(r+1)}\\ \hat{\tau}^{(r+1)}\end{bmatrix}=\begin{bmatrix}\hat{\eta}^{(r)}\\ \hat{\tau}^{(r)}\end{bmatrix}-\begin{bmatrix}\frac{\partial^2 L^{(r)}}{\partial\eta^2} & \frac{\partial^2 L^{(r)}}{\partial\eta\partial\tau}\\ \frac{\partial^2 L^{(r)}}{\partial\tau\partial\eta} & \frac{\partial^2 L^{(r)}}{\partial\tau^2}\end{bmatrix}^{-1}\times\begin{bmatrix}\frac{\partial L^{(r)}}{\partial\eta}\\ \frac{\partial L^{(r)}}{\partial\tau}\end{bmatrix},$$

taking the partial derivatives of $\ln l\left(\boldsymbol{x}\middle|\eta,\tau\right)$, where $L(\bullet)$ denotes log likelihood function,

$$\frac{\partial^2 L^{(r)}}{\partial\eta^2}=\sum_{i=1}^{n}\psi'\left(x_i+\eta^{(r)}\right)-n\psi'\left(\eta^{(r)}\right),$$

$$\frac{\partial^2 L^{(r)}}{\partial\eta\partial\tau}=-\frac{n}{\tau^{(r)}+1},$$

$$\frac{\partial^2 L^{(r)}}{\partial\tau^2}=\frac{n\eta^{(r)}}{\left(\tau^{(r)}+1\right)^2}-\frac{\sum_{i=1}^{n}x_i}{\left(\tau^{(r)}\right)^2}-\frac{\sum_{i=1}^{n}x_i}{\left(\tau^{(r)}+1\right)^2},$$

$$\frac{\partial L^{(r)}}{\partial\eta}=\sum_{i=1}^{n}\psi\left(x_i+\eta^{(r)}\right)-n\ln\left(\tau^{(r)}+1\right)-n\psi\left(\eta^{(r)}\right),$$

$$\frac{\partial L^{(r)}}{\partial\tau}=-\frac{n\eta^{(r)}}{\tau^{(r)}+1}+\frac{\sum_{i=1}^{n}x_i}{\tau^{(r)}}-\frac{\sum_{i=1}^{n}x_i}{\tau^{(r)}+1}.$$

The parameters η and τ are initially estimated by the method of moments Fisher (1941), and Thom (1957) defined as

$$\eta^{*}=\frac{\overline{x}^2}{s^2-\overline{x}},\tag{4}$$

and

$$\tau^{*}=\frac{s^2-\overline{x}}{\overline{x}},\tag{5}$$

where $\overline{x}$ and s^2 are the sample mean and variance, respectively. The posterior distribution of λ is thus calculated, yielding

$$\pi\left(\lambda\middle|\boldsymbol{x},\eta,\tau\right)=\frac{1}{\Gamma\left(x+\eta\right)}\lambda^{x+\eta-1}e^{-\lambda\left(\frac{\tau+1}{\tau}\right)}\left(\frac{\tau+1}{\tau}\right)^{y+\eta}.$$

Substituting the estimators of η and τ, we obtain

$$\lambda\middle|\boldsymbol{x},\eta,\tau\sim ga\left(x+\hat{\eta},\frac{\hat{\tau}}{\hat{\tau}+1}\right).$$

Let λ_1' and λ_2' be estimators of λ_1 and λ_2 respectively, where

$$\lambda_1' = \frac{x_1 + \hat{\eta}_1}{\left(\frac{\hat{\tau}_1 + 1}{\hat{\tau}_1}\right)}, \tag{6}$$

and

$$\lambda_2' = \frac{x_2 + \hat{\eta}_2}{\left(\frac{\hat{\tau}_2 + 1}{\hat{\tau}_2}\right)}. \tag{7}$$

The EB of odds ratio can be obtained as following

$$\widehat{OR}_{EB} = \frac{p_1' / (1 - p_1')}{p_2' / (1 - p_2')}, \tag{8}$$

where $p_1' = \lambda_1' / n_1$ and $p_2' = \lambda_2' / n_2$ denote estimators for success probabilities in group 1 and 2, respectively.

3. Simulation Study and Result

In this section, performance of the proposed method in comparison with MMLE method is assessed. The approximate solutions are given in equation (2) and (8), respectively. Data in both groups are generated as independent binomial distributions with sample sizes $(n_1, n_2) = (10, 10)$ and $(10, 50)$, and success probabilities of 0.01, 0.03, 0.05 and 0.10. Each situation is repeated 5,000 times after 1,000 burn-ins, using R program (version 3.2.0) (2010). The efficiency of estimators is evaluated using the Estimated Relative Error (ERE) (%), defined as

$$ERE = \left[\frac{\left| OR - \widehat{OR}_i \right|}{OR} \right] \times 100, \tag{9}$$

where *OR* denotes the usual maximum likelihood estimator of odds ratio, and $\widehat{OR}_i$ denotes the estimate of odds ratio using EB and MMLE, respectively.

The estimated odds ratios resulted from simulation for sample sizes $(n_1, n_2) = (10, 10)$ and $(10, 50)$ are given in Tables 1-2. The numerical results of comparison using the performance indicator, ERE, are shown in Tables 3-4 and also illustrated in Figure 1 for the cases $(n_1, n_2) = (10, 50)$ and $p_2 = 0.01, 0.03, 0.05, 0.10,$ which are similar for the other cases not shown here. Based on the performance indicator, it can be seen that, for both equal and unequal sample sizes, the proposed estimator mostly outperforms the MMLE, except for the sample sizes $(n_1, n_2) = (10, 10)$ with $(p_1, p_2) = (0.01, 0.01)$. In addition, 31 out of 32 situations (96.875%) indicate smaller EREs for the EB method than MMLE method.

Table 1 The estimated values of odds ratio for $(n_1, n_2) = (10, 10)$

(p_1, p_2)	OR	$\widehat{OR}_{EB}$	$\widehat{OR}_{MMLE}$
(0.01, 0.01)	1.0000	1.1626	1.1486
(0.01, 0.03)	0.3266	0.3270	2.3733
(0.01, 0.05)	0.1919	0.0571	0.4472
(0.01, 0.10)	0.0909	0.0925	0.6227
(0.03, 0.01)	3.0619	3.5741	1.5989
(0.03, 0.03)	1.0000	1.0518	1.3940
(0.03, 0.05)	0.5876	0.6067	1.2145
(0.03, 0.10)	0.2784	0.2837	0.8676
(0.05, 0.01)	5.2105	6.1144	2.0736
(0.05, 0.03)	1.7018	1.7865	1.8013
(0.05, 0.05)	1.0000	1.0287	1.5718
(0.05, 0.10)	0.4737	0.4780	1.1174
(0.10, 0.01)	11.0000	13.0139	3.3491
(0.10, 0.03)	3.5926	3.7891	2.9181
(0.10, 0.05)	2.1111	2.1743	2.5410
(0.10, 0.10)	1.0000	1.0141	1.8097

Table 2 The estimated values of odds ratio for $(n_1, n_2) = (10, 50)$

(p_1, p_2)	OR	$\widehat{OR}_{EB}$	$\widehat{OR}_{MMLE}$
(0.01, 0.01)	1.0000	1.0219	4.2689
(0.01, 0.03)	0.3266	0.3270	2.3733
(0.01, 0.05)	0.1919	0.1935	1.4491
(0.01, 0.10)	0.0909	0.0915	0.6185
(0.03, 0.01)	3.0619	3.1678	5.9543
(0.03, 0.03)	1.0000	1.0144	3.3149
(0.03, 0.05)	0.5876	0.5917	2.0234
(0.03, 0.10)	0.2784	0.2781	0.8556
(0.05, 0.01)	5.2105	5.3877	7.7032
(0.05, 0.03)	1.7018	1.7246	4.2890
(0.05, 0.05)	1.0000	1.0069	2.6134
(0.05, 0.10)	0.4737	0.4748	1.1093
(0.10, 0.01)	11.0000	11.3334	12.4201
(0.10, 0.03)	3.5926	3.6340	6.9040
(0.10, 0.05)	2.1111	2.1198	4.2136
(0.10, 0.10)	1.0000	1.0048	1.7914

Table 3 The percentage of the estimated relative error of odds ratio estimation for $(n_1, n_2) = (10, 10)$.

(p_1, p_2)	ERE_{EB}	ERE_{MMLE}
(0.01, 0.01)	16.2590	14.8613
(0.01, 0.03)	3.7609	205.8135
(0.01, 0.05)	70.2297	133.0545
(0.01, 0.10)	1.7158	585.0082
(0.03, 0.01)	16.7271	47.7824
(0.03, 0.03)	5.1779	39.3969
(0.03, 0.05)	3.2457	106.6910
(0.03, 0.10)	1.8939	211.6315
(0.05, 0.01)	17.3471	60.2028
(0.05, 0.03)	4.9754	5.8475
(0.05, 0.05)	2.8739	57.1837
(0.05, 0.10)	0.9175	135.8892
(0.10, 0.01)	18.3080	69.5535
(0.10, 0.03)	5.4695	18.7745
(0.10, 0.05)	2.9936	20.3649
(0.10, 0.10)	1.4068	80.9732

Table 4 The percentage of the estimated relative error of odds ratio estimation for $(n_1, n_2) = (10, 50)$

(p_1, p_2)	ERE_{EB}	ERE_{MMLE}
(0.01, 0.01)	2.1866	326.8888
(0.01, 0.03)	0.1127	626.6820
(0.01, 0.05)	0.8355	655.1393
(0.01, 0.10)	0.6896	580.3816
(0.03, 0.01)	3.4588	94.4649
(0.03, 0.03)	1.4432	231.4878
(0.03, 0.05)	0.6918	244.3420
(0.03, 0.10)	0.0962	207.3378
(0.05, 0.01)	3.4013	47.8408
(0.05, 0.03)	1.3383	152.0266
(0.05, 0.05)	0.6860	161.3386
(0.05, 0.10)	0.2367	134.1791
(0.10, 0.01)	3.0306	12.9102
(0.10, 0.03)	1.1532	92.1720
(0.10, 0.05)	0.4110	99.5920
(0.10, 0.10)	0.4794	79.1387

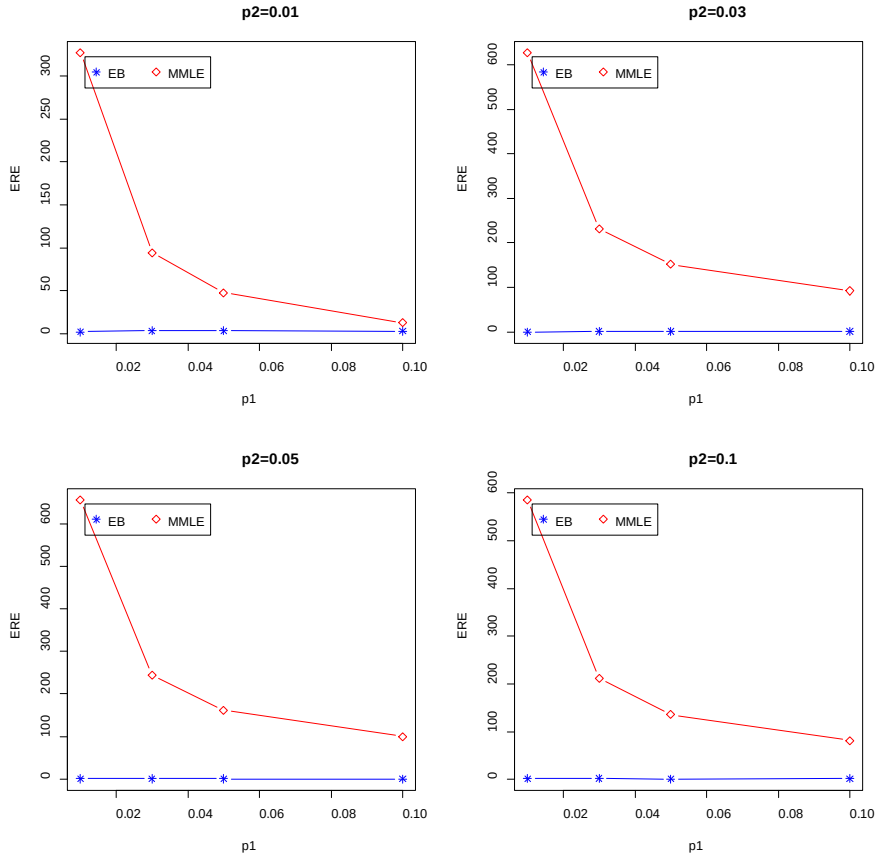

Figure 1 A comparison of the percentage of ERE for odds ratio estimation using EB and MMLE when $(n_1, n_2) = (10, 50)$

4. Application

Our example is taken from the A to Z trial by Blazing et al. (2004), which compared two treatments of enoxaparin and un-fractionated heparin in 3,905 patients with acute coronary syndrome. The count of patients with TIMI (The thrombolysis in Myocardial Infarction) major bleeding in each treatment group was considered as outcome measure, resulting in $(x_1, x_2) = (18, 8)$ out of $(n_1, n_2) = (1940, 1965)$ for enoxaparin and un-fractionated, respectively. The true odds ratio and their estimates using EB and MMLE methods are shown in Table 5. The results reveal that the estimate of the odds ratio using EB method yields the least ERE percentage with 0.3361 while that using MMLE method results in ERE percentage with 3.2693.

Table 5 True odds ratio and their estimates using EB and MMLE, with corresponding percentages of ERE

	Methods		
	True	EB	MMLE
OR	2.2910	2.2987	2.2161
ERE	-	0.3361	3.2693

5. Conclusions

This paper presents the odds ratio estimation in rare data with binomial distribution, and provides a Poisson approximate to the binomial distribution. The results obtained from simulated data indicate that the proposed method performs rather wells. The EB estimator of odds ratio is more efficient than the MMLE estimator. Hence our purposed estimator is an alternative to the MMLE method without interfering with initial data.

Acknowledgements

The authors would like to thank all referees and the editor for valuable comments and suggestions to improve this work.

References

Agresti A, Yang, M. An empirical investigation some effects of sparseness in contingency tables. Comput. Stat. Data Anal. 1987; 5: 9-21.

Bishop YMM, Fienberg SE, Holland PW. Discrete Multivariate Analysis: Theory and Practice. Cambridge: MA: MIT press; 1975.

Blazing MA, de Lemos JA, White HD, Fox KAA, Verheugt FWA, Ardissino D, DiBattiste PM, Palmisano J, Bilheimer DW, Snapinn SM, Ramsey KE, Gardner LH, Hasselblad V, Pfeffer MA, Lewis EF, Braunwald E, Califf RM. For the A to Z Investigators. Safety and efficacy of enoxaparin vs unfractionated heparinin patients with non–ST-segment elevationa cute coronary syndromes who receive tirofiban and aspirin: a randomized controlled trial. J. Am. Med. Assoc. 2004; 292:55-64.

Fisher RA. The negative binomial distribution. Ann. Eugen. 1941; 11:182-187.

Gart JJ, Zweifel JR. On the bias of various estimations of the logit and its variance with application to quantal bioassay. Biometrika. 1967; 54:181-187.

Jewell NP. Small-sample bias of point estimators of the odds ratio from matched sets. Biometrics. 1984; 40:421-435.

Jewell NP. On the bias of commonly used measures of association for 2×2 tables. Biometrics. 1986; 42: 351-358.

Haldane JBS. The estimation and significance of the logarithm of a ratio frequencies. Ann. Hum. Genet. 1955; 20: 309-311.

Hauck WW, Anderson S, Leahy FJ. Finite-sample properties of some new estimators of a common odds ratio from multiple 2×2 tables. J. Am. Stat. Assoc. 1982; 77:145-152.

Hitchcock SE. A note on the estimation of parameters of the logistic function using the minimum logit method. Biometrika. 1962; 49:250-252.

Thom HCS. The frequency of hail occurrence. Archiv für Meteorologie, Geophysik und Bioklimatologie. 1957; 8:185-194.

The R Development Core Team. An introduction to R, Vienna, http://R-project.org. 2010.

Wakeel A, Aslam M. Bayesian estimation of rare sensitive attribute. Thail. Stat. 2013; 11(1): 17-29.

Wu C. Estimating Adjusted Relative risk in studies of common outcomes: A comparative study and an example using vaccination coverage date. PHD thesis, USA, University at Albany; 2002.