Odds Ratio Estimation for Small Proportion in Binomial Distribution

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

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

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

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

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

[*] *Email address: kobkun.ma [AT] gmail.com*

ABSTRACT---- ***In this study, we introduce the new estimator of odds ratio using Empirical Bayes (EB) for small proportions of success in a 2x2 table. The proposed estimate of odds ratio based on EB 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 more efficient than MMLE.***

Keyword--- Odds Ratio, Empirical Bayes, Modified Maximum Likelihood Estimator

1. INTRODUCTION

The odds ratio, defined as a ratio of two odds, is a measure of association between two independent groups with binary outcome. Binary outcome is referred to as success or failure; such as dead or alive, good or bad conditions, and two independent groups can be treatment and control groups or two treatment groups. The data can be arranged in a (2X2) table as in Table 1, there are n_1 subjects in group 1 with y_1 successes, and n_2 subjects in group 2 with y_2 successes. Total number of subjects in each group is assumed to be fixed. Thus, y_1 and y_2 are considered as independent binomial random variables.

Table 1: The 2X2 contingency table

Group	Outcome		Total
	Success	Failure	
1	y_1	$n_1 - y_1$	n_1
2	y_2	$n_2 - y_2$	n_2
Total	$y_1 + y_2$	$(n_1 + n_2) - (y_1 + y_2)$	$n_1 + n_2$

The usual maximum likelihood estimator of odds ratio from a (2X2) table,

$$\hat{O}R_r = \frac{odd_1}{odd_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{y_1(n_2 - y_2)}{y_2(n_1 - y_1)}. \quad (1)$$

Agresti [1] explained property of the odds ratio is nonnegative real value. It equals to 1 when groups are independent of response. The value greater than 1 indicates that success is more likely to occur in group 1 than group 2, and vice versa for the value less than 1. Consequently, the farther value from 1 in either direction represents strength of association. In the usual method, odds ratio lead to zero (if the numerator is 0) or infinity (if the denominator is 0) or undefined (if there is 0's in both numerator and denominator). Haldane [2] and Gart and Zweifel [3] suggested to add a correction term 0.5 to each cell, when having zero cell count, which gives the modified maximum likelihood estimator (MMLE) as

$$\hat{OR}_{mmle} = \frac{(y_1 + 0.5)(n_2 - y_2 + 0.5)}{(y_2 + 0.5)(n_1 - y_1 + 0.5)}. \quad (2)$$

Even though $\hat{OR}_{mmle}$ still lies between 0 and infinity, some researchers, Bishop, Fienberg, and Holland [4] and Agresti and Yang [5], discouraged adding 0.5 to each cell, because of the appearance of adding "fake data".

As mentioned, small cell counts in clinical trials involving rare event can cause difficulty in the odds ratio estimation since it may lead to zeros or small observed counts in numerator, or denominator, or both, yielding large standard error and thus less precise confidence interval. As a result, only a rough idea of the value of true odds ratio is obtained. In this study, we purpose new estimation method of the odds ratio in a 2 x 2 table with small proportions of success based on Empirical Bayes (EB). Our purposed estimation does not interfere with the original data and tends to outperform the conventional estimator, MMLE. Both simulated and actual data are utilized to evaluate the purposed estimator.

2. APPROXIMATE SOLUTION OF ODDS RATIO

2.1 The Empirical Bayes Method

In this section, a new approximation method for small proportion is proposed using EB. Data are assumed to be binomial distribution. Let y_1 and y_2 be random variables, distributed as binomial with equal and unequal sample sizes and unknown probability, $y_1 \sim Bin(n_1, p_1)$ and $y_2 \sim Bin(n_2, p_2)$ where n_1, n_2 and p_1, p_2 denote sample sizes and unknown probability. The informative prior is adopted on p_i, $p_i \sim beta(\alpha_i, \beta_i)$, $i = 1, 2$ where α_i and β_i denote hyper-parameters. The estimation of hyper-parameters can be obtained from the posterior marginal distribution function as follow,

$$\begin{aligned} m(y|\alpha,\beta) &= \int_{-\infty}^{\infty} f(y|p)\pi(p)\,dp \\ &= \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}. \end{aligned} \quad (3)$$

Then, we estimate both hyper-parameters using maximum likelihood method. The likelihood function of posterior marginal distribution function is displayed as

$$L(y|\alpha,\beta) = \prod_{i=1}^{n}\binom{n}{y_i}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y_i+\alpha)\Gamma(n-y_i+\beta)}{\Gamma(n+\alpha+\beta)}$$

$$= \prod_{i=1}^{n}\binom{n}{y_i}\frac{(\alpha+y_i-1)(\alpha+y_i-2)\cdots\alpha(\beta+n-y_i-1)(\beta+n-y_i-2)\cdots\beta}{(\alpha+\beta+n-1)(\alpha+\beta+n-2)\cdots(\alpha+\beta)}.$$

Applying Newton-Raphson method to solve a nonlinear equation, therefore the maximum likelihood estimator of hyper-parameters can be obtained from

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

where,

$$\frac{\partial^2 L^{(r)}}{\partial\alpha^2} = -\sum_{i=1}^{n}\left[\frac{1}{\left(\alpha^{(r)}+y_i-1\right)^2} + \frac{1}{\left(\alpha^{(r)}+y_i-2\right)^2} + \cdots + \frac{1}{\left(\alpha^{(r)}\right)^2}\right]$$

$$+\sum_{i=1}^{n}\left[\frac{1}{\left(\alpha^{(r)}+\beta^{(r)}+n-1\right)^2} + \frac{1}{\left(\alpha^{(r)}+\beta^{(r)}+n-2\right)^2} + \cdots + \frac{1}{\left(\alpha^{(r)}+\beta^{(r)}\right)^2}\right]$$

$$\frac{\partial^2 L^{(r)}}{\partial\alpha\partial\beta} = \sum_{i=1}^{n}\left[\frac{1}{\left(\alpha^{(r)}+\beta^{(r)}+n-1\right)^2}+\frac{1}{\left(\alpha^{(r)}+\beta^{(r)}+n-2\right)^2}+\ldots+\frac{1}{\left(\alpha+\beta\right)^2}\right]$$

$$\frac{\partial^2 L^{(r)}}{\partial\beta^2} = \sum_{i=1}^{n}\left[\frac{1}{\left(\beta^{(r)}+n+y_i-1\right)^2}+\frac{1}{\left(\beta^{(r)}+n+y_i-2\right)^2}+\cdots+\frac{1}{\left(\beta^{(r)}\right)^2}\right]$$

$$+\sum_{i=1}^{n}\left[\frac{1}{\left(\alpha^{(r)}+\beta^{(r)}+n-1\right)^2}+\frac{1}{\left(\alpha^{(r)}+\beta^{(r)}+n-2\right)^2}+\cdots+\frac{1}{\left(\alpha^{(r)}+\beta^{(r)}\right)^2}\right]$$

$$\frac{\partial L^{(r)}}{\partial\alpha} = \sum_{i=1}^{n}\left[\frac{1}{\alpha^{(r)}+y_i-1}+\frac{1}{\alpha^{(r)}+y_i-2}+\cdots+\frac{1}{\alpha^{(r)}}\right]$$

$$-\sum_{i=1}^{n}\left[\frac{1}{\alpha^{(r)}+\beta^{(r)}+n-1}+\frac{1}{\alpha^{(r)}+\beta^{(r)}+n-2}+\cdots\frac{1}{\alpha^{(r)}+\beta^{(r)}}\right]$$

$$\frac{\partial L^{(r)}}{\partial\beta} = \sum_{i=1}^{n}\left[\frac{1}{\beta^{(r)}+n+y_i-1}+\frac{1}{\beta^{(r)}+n+y_i-2}+\cdots+\frac{1}{\beta^{(r)}}\right]$$

$$-\sum_{i=1}^{n}\left[\frac{1}{\alpha^{(r)}+\beta^{(r)}+n-1}+\frac{1}{\alpha^{(r)}+\beta^{(r)}+n-2}+\cdots\frac{1}{\alpha^{(r)}+\beta^{(r)}}\right]$$

and r represents the iteration number (r = 1,2,3,…).
In this study, the moment estimator of hyper-parameters in Beta Binomial distribution is used as initial values [6] as follow

$$\hat{\alpha} = \frac{nm_1 - m_2}{n\left(\frac{m_2}{m_1}-m_1-1\right)+m_1} \quad (4)$$

$$\hat{\beta} = \frac{\left(n-m_1\right)\left(n-\frac{m_2}{m_1}\right)}{n\left(\frac{m_2}{m_1}-m_1-1\right)+m_1} \quad (5)$$

where m_1 and m_2 denote first and second raw sample moment respectively.
The posterior distribution of p is thus calculated, yielding

$$\pi\left(p\middle|\underline{y},\alpha,\beta\right) = p^{(y+\alpha-1)}\left(1-p\right)^{(n-y+\beta-1)}\frac{\Gamma\left(n+\alpha+\beta\right)}{\Gamma\left(y+\alpha\right)\Gamma\left(n-y+\beta\right)}$$

Consequently, the posterior marginal distribution of y is the beta-binomial distribution (BBD).

$$p\middle|y \square\ Beta\left(y+\hat{\alpha},n-y+\hat{\beta}\right)$$

Thus, the approximation of p_1 as p_1' is

$$p_1' = \frac{y_1 + \hat{\alpha}_1}{n_1 + \hat{\alpha}_1 + \hat{\beta}_1},$$

(6)

and the approximation of p_2 as p_2' is

$$p_2' = \frac{y_2 + \hat{\alpha}_2}{n_2 + \hat{\alpha}_2 + \hat{\beta}_2}.$$

(7)

The EB of odds ratio can be obtained as follow.

$$\hat{OR}_{eb} = \frac{p_1'/(1-p_1')}{p_2'/(1-p_2')}.$$

(8)

2.2 Simulation study

Simulation studies have been carried out using R program (version 3.2.0) to assess performance of the proposed method in comparison with MMLE method. The approximate solutions are given in equation (2) and (8) respectively. Data in both groups are generated as independent binomial distributions with equal $(n_1 = n_2 = 10)$ and unequal $(n_1 = 14, n_2 = 11)$ sample sizes. Let probabilities of success of each size are 0.01, 0.05 and 0.10. Each situation is repeated 5,000 iterations after 1,000 burn-ins. The efficiency of estimators is measured using the Estimated Relative Error (ERE), defined as

$$ERE = \left[\frac{OR_r - \overline{\hat{OR}_i}}{OR_r}\right] \times 100\%$$

, where OR_r denotes the usual maximum likelihood estimator of odds ratio

, and $\overline{\hat{OR}_i}$ denotes the means of the approximate of odds ratio for EB method and MMLE method respectively. The simulation results are given in Table 2-5.

Table 2: Odds ratio estimation with equal sample sizes $(n_1, n_2) = (10,10)$

p_1	p_2	OR_r	$\hat{OR}_{eb}$	$\hat{OR}_{mmle}$
0.01	0.01	1.0000	1.3665	1.1514
0.01	0.05	0.1919	0.2248	0.8723
0.01	0.10	0.0909	0.1040	0.6219
0.05	0.01	5.2105	6.5657	2.0724
0.05	0.05	1.0000	1.0787	1.5693
0.05	0.10	0.4737	0.4989	1.1181
0.10	0.01	11.0000	13.7434	3.3472
0.10	0.05	2.1111	2.2585	2.5352
0.10	0.10	1.0000	1.0445	1.8068

Table 3: Odds ratio estimation with unequal sample sizes $(n_1, n_2) = (14,11)$

p_1	p_2	OR_r	$\hat{OR}_{eb}$	$\hat{OR}_{mmle}$
0.01	0.01	1.0000	1.2901	0.9652
0.01	0.05	0.1919	0.2168	0.7131
0.01	0.10	0.0909	0.1001	0.4944
0.05	0.01	5.2105	6.4376	1.9403
0.05	0.05	1.0000	1.0823	1.4326
0.05	0.10	0.4737	0.4997	0.9926
0.10	0.01	11.0000	13.5656	3.2847
0.10	0.05	2.1111	2.2797	2.4257
0.10	0.10	1.0000	1.0529	1.6811

Table 4: ERE for equal sample sizes $(n_1, n_2) = (10,10)$

p_1	p_2	$\hat{OR}_{eb}$	$\hat{OR}_{mmle}$
0.01	0.01	36.6535	15.1385
0.01	0.05	17.1281	354.5204
0.01	0.10	14.3630	584.0771
0.05	0.01	26.0092	60.2273
0.05	0.05	7.8661	56.9330
0.05	0.10	5.3167	136.0498
0.10	0.01	24.9404	69.5705
0.10	0.05	6.9818	20.0893
0.10	0.10	4.4467	80.6791

Table 5: ERE for unequal sample sizes $(n_1, n_2) = (14,11)$

p_1	p_2	$\hat{OR}_{eb}$	$\hat{OR}_{mmle}$
0.01	0.01	29.0090	3.4829
0.01	0.05	12.9665	271.5570
0.01	0.10	10.1454	443.9483
0.05	0.01	23.5505	62.7620
0.05	0.05	8.2274	43.2629
0.05	0.10	5.5025	109.5594
0.10	0.01	23.3239	70.1889
0.10	0.05	7.9863	14.9028
0.10	0.10	5.2871	68.1098

Based on the performance indicator, it can be seen that both equal and unequal sample sizes, the proposed estimator mostly outperform the MMLE, except for the case $(p_1, p_2) = (0.01, 0.01)$.

3. ILLUSTRATIVE EXAMPLES USING APPLICATION DATA SET

Our example is taken from the study of Parzen et.al [7], which was a randomized phase II clinical trial developed by the United States Eastern Cooperative Oncology group (ECOG) and opened for patient accrual from 1987 through 1990 to evaluate two new chemotherapy treatments in patients with advanced large bowel cancer. Two treatments referred to Homoharringtonine and Caracemide. In this clinical trial, the investigators were interested in toxicity or side effect of treatment, defined as life-threatening toxicity. Twenty-five patients entered the study, $n_1 = 14$ on the

Homoharringtonine and $n_2 = 11$ on the Caracemide. The accrual goal was set to 30 patients in each arm. However, the study was terminated early after being open for 43 months. The outcome found $y_1 = 2$ subjects with life-threatening toxicities in harringtonine and the $y_2 = 1$ subject with life-threatening toxicities in Caracemide.

For outcome, in which $(y_1, y_2) = (2,1)$, our proposed estimate of the odds ratio is $\hat{O}R_{eb} = 0.1517$, the estimate of the odds ratio after adding 0.5 to every cell is $\hat{O}R_{mmle} = 1.40$, and the usual estimate of odds ratio is $\hat{O}R_r = 0.1667$.

4. CONCLUSION

Based on simulated and real data, it can be shown that the EB Estimator of odds ratio is more efficient than the conventional estimator, MMLE. In addition, our purposed estimator is an alternative to the MMLE without disturbing the original data.

5. REFERENCES

[1] Agresti, A., Categorical Data Analysis, 3rd Etd., John Willey and Sons, USA, 2013.
[2] Haldane, J.B.S., The estimation and Significance of the Logarithm of a ratio frequencies. *Annals of Human Genetics*, 1955; 20: 309-311.
[3] Gart, J.J., and Zweifel, J.R., On the bias of various estimations of the logit and its variance with application
[4] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W., Discrete multivariate analysis. Theory and Practice, Cambridge, MA:MIT press.
[5] Agresti, A., and Yang, M., An empirical investigation some effects of sparseness in contingency tables. Computational Statistics and Data Analysis, 1987; 5: 9-21.
[6] Minka, T.P, Estimating a Dirichlet Distribution. Technical report. MIT, 2000.
[7] Parzen, M., Lipsitz, S., Ibrahim, J., and Klar, N., An estimation of the odds ratio that always exists. *Journal of Computation and Graphical Statistics*, 2002; 11: 420-436.