The Bayes Factor for the Misclassified Categorical Data

Tze-San Lee[1]

[1] Western Illinois University, USA

Correspondence: Tze-San Lee, retired mathematical professor at Western Illinois University.

Received: May 2, 2018 Accepted: May 23, 2018 Online Published: June 28, 2018

doi:10.5539/ijsp.v7n4p91 URL: https://doi.org/10.5539/ijsp.v7n4p91

Abstract

This article addresses the issue of misclassification in a single categorical variable, that is, how to test whether the collected categorical data are misclassified. To tackle this issue, a pair of null and alternative hypotheses is proposed. A mixed Bayesian approach is taken to test these hypotheses. Specifically, a bias-adjusted cell proportion estimator is presented that accounts for the bias caused by classification errors in the observed categorical data. The chi-square test is then adjusted accordingly. To test the null hypothesis that the data are not misclassified under a specified multinomial distribution against the alternative hypothesis they are misclassified, the Bayes factor is calculated for the observed data and a comparison is made with the classical p-value.

Keywords: Bayes factor, classification errors, Dirichlet's distribution, Type II maximum likelihood

1. Introduction

The problem of misclassification is a major issue in observational epidemiologic studies. Not long after Bross (1954) pointed out that the non-differential misclassification would bias the corrected odds ratio toward the null hypothesis, Diamond and Lilienfeld (1962a-b) has extended the result to various types of epidemiologic studies. A 2 × 2 case-control studies with a single exposure variable being misclassified has been widely studied (Fleiss et al 2003, Chapter 17; Gustafson 2004, Chapter 5; Kleinbaum et al 1982, Chapter 12; Rothman et al 2008, Chapter 19). Yet, almost no authors pay attention to investigate the effect of misclassification in the analysis of a single categorical variable except Mote and Anderson (1965). Mote and Anderson primarily takes a deductive approach to account for the bias caused by the classification errors. Yet, the shortcoming with a deductive approach is that it does not take the sampling errors into consideration. As a result, the issue on how to deal with the misclassification in the analysis of categorical data still remains unsolved.

This article addresses another important issue, that is, whether the observed categorical data are misclassified. Instead of using a deductive method, an inductive approach is employed to account for the misclassification bias embedded in the collected data. First, the inverse way is taken by equating the expected value of the estimated sample cell proportion with its population parameter conditional on that the misclassification probabilities are given. Then the bias-adjusted estimator is presented for the population cell proportion parameter by inverting the misclassification matrix. Second, the appropriate misclassification probabilities are calculated depending on if the misclassification is possibly made either from one category to all other categories (scenario I) or merely to its neighboring categories (scenario II). Third, in order to test the null hypothesis that the data are not misclassified under a specified multinomial distribution, a mixed Bayesian approach is used to calculate the Bayes factor and compare it with the traditional p-value.

2. Methodology & Background

Given that X is a categorical variable with K (≥ 3) categories and the data are collected through a simple random sampling of size N, where $N = \sum_{i=1}^{K} n_i$ (table 1). The crude estimator, $\hat{p}_j$, for the population cell proportion p_j in the j^{th} category is then given by

$$\hat{p}_j = n_j / N \ . \tag{1}$$

Assume that $\hat{p}_j$ is distributed as a multinomial distribution with the population size N and the cell proportion of the j^{th} category p_j. It is well known that Eq. 1 is an unbiased estimator for the population cell proportion parameter, provided that the observed data are not misclassified (Agresti 2002). However, it is shown below by Eq. 4 that $\hat{p}_j$ of Eq. 1 is no longer unbiased for p_j, once the observed data are misclassified.

Table 1. Observed data for the categorical variable X

Variable	Categories			
X	1	2	………	K
Observation	n_1	n_2	………	n_K

Suppose that the observed data are misclassified. Let w_{jk} ($j \neq k$) be the misclassification probability of an observation belonging to the j^th category being incorrectly classified into the k^th category and w_{jj} the correct classification probability that an observation belonging to the j^th category being correctly classified into the j^th category. Then, it is easily shown that the expected value of $\hat{p}$ is

$$E(\hat{p}) = Wp\,, \tag{2}$$

where $p = (p_1, p_2, ..., p_K)$, $\hat{p} = (\hat{p}_1, \hat{p}_2, ..., \hat{p}_K)$, and $W = [w_{jk}]^T{}_{j,k=1,2,...,K}$ is the misclassification matrix, in which $\sum_{k=1}^{K} w_{jk} = 1$ for j = 1, 2, …, K. Eq. 2 shows that the crude estimator $\hat{p}_k$ is no longer unbiased for the population parameter p_k, provided that $W \neq I$,where I is the K × K identity matrix. A set of misclassification probabilities $\{w_{jk}\}$ is said to be feasible if the misclassification matrix W in Eq. 2 is invertible (or nonsingular) for $0 < w_{jk} < 1$.

Assume that W is invertible. Then bias-adjusted cell proportion (BACP) estimators ($\breve{p}_k$) are defined by

$$\breve{p} = W^{-1}\hat{p} = V\hat{p}\,, \tag{3}$$

where $\breve{p} = (\breve{p}_1, \breve{p}_2, ..., \breve{p}_K)^T$, V = [$v_{jk}$], j, k = 1, 2, …, K, denotes the inverse matrix of W, and $\breve{n} = Vn$, $\breve{n} = (\breve{n}_1,,,,\breve{n}_K)^T$, $n = (n_1,...,n_K)^T$. Note that by using Eqs. 2 and 3 it's easily shown: $E(\breve{p}) = p$, namely, $\breve{p}$ is an unbiased estimator for p, provided that W is known. The BACP estimators { $\breve{p}_k$ } are said to be admissible if for feasible w_{jk} we have $0 < \breve{p}_k < 1$ and $\sum_{j=1}^{K} \breve{p}_j = 1$. Similarly, a set of misclassification error probabilities $\{w_{jk}\}$ is said to be admissible if the corresponding BACP estimators { $\breve{p}_k$ } are admissible.

The misclassification matrix W has two possible forms depending on how the categorical variable X is misclassified. There are two possible scenarios that are given as follows:

Scenario I: The misclassification occurs after classifying one category incorrectly into all other categories. Also, because misclassification can occur equally likely from any one of the j^th correct category to the k^th (observed) wrong category, we thus have, for fixed j

$$\theta_j \equiv w_{jk} > 0,\ k \neq j, \text{ and } \ w_{jj} = 1 - \sum_{\substack{k=1 \\ k \neq j}}^{K} w_{jk}\,,\ \ j = 1, 2, \ldots, K, \tag{4}$$

Scenario II: The misclassification occurs after classifying one category incorrectly only into its neighboring categories. Therefore, we have, for fixed j

$$w_{jk} = 0 \text{ for } |k - j| > 1, \text{ and } \ w_{jj} = 1 - \sum_{\substack{k=1 \\ k \neq j}}^{K} w_{jk}\,,\ j = 1, 2, \ldots, K. \tag{5}$$

When K = 3, the associated misclassification matrix with its determinant and its inverse matrix for scenarios I and II are hereby obtained respectively. An explicit form of the misclassification matrix W_I and its inverse V_I for scenario I are given respectively by

$$W_I = \begin{bmatrix} 1-\theta_2-\theta_3 & \theta_2 & \theta_3 \\ \theta_1 & 1-\theta_1-\theta_3 & \theta_3 \\ \theta_1 & \theta_2 & 1-\theta_1-\theta_2 \end{bmatrix}, \tag{6a}$$

$$\Delta_I \equiv \det(W_I) = (1-\theta_1-\theta_2-\theta_3)^2 \neq 0, \tag{6b}$$

and

$$V_I \equiv [v_{jk(I)}] = \Delta_I^{-\frac{1}{2}} \cdot \begin{bmatrix} 1-\theta_1 & -\theta_1 & -\theta_1 \\ -\theta_2 & 1-\theta_2 & -\theta_2 \\ -\theta_3 & -\theta_3 & 1-\theta_3 \end{bmatrix}, \tag{6c}$$

where $\theta_1 \equiv w_{12} = w_{13}$, $\theta_2 \equiv w_{21} = w_{23}$, and $\theta_3 \equiv w_{31} = w_{32}$.

The BACP estimators for scenario I are given by

$$\breve{p}_{k(I)} = \sum_{j=1}^{K} v_{jk(I)} \cdot \hat{p}_j, \quad \text{k = 1, 2, …, K}, \tag{7}$$

By using Eqs. 6b and 7, the feasibility and admissibility constraints for the misclassification probability and BACP estimator are given respectively as follows:

$$\theta_1 + \theta_2 + \theta_3 < 1, \tag{8a}$$

and

$$\theta_1 < 1, \qquad \theta_2 < 1, \qquad \theta_3 < 1. \tag{8b}$$

For scenario II, an explicit form of the misclassification matrix W_{II} and its inverse V_{II} are given respectively by

$$W_{II} = \begin{bmatrix} 1-\gamma_2 & \gamma_2 & 0 \\ \gamma_1 & 1-\gamma_1-\gamma_3 & \gamma_3 \\ 0 & \gamma_2 & 1-\gamma_2 \end{bmatrix}, \tag{9a}$$

$$\Delta_{II} \equiv \det(W_{II}) = (1-\gamma_2)(1-\gamma_1-\gamma_2-\gamma_3) \neq 0, \tag{9b}$$

and

$$V_{II} \equiv [v_{jk(II)}] = \Delta_{II}^{-1} \cdot \begin{bmatrix} (1-\gamma_1)(1-\gamma_2) & -\gamma_1(1-\gamma_2) & \gamma_1\gamma_2 \\ -\gamma_2(1-\gamma_2) & (1-\gamma_2)^2 & -\gamma_2(1-\gamma_2) \\ \gamma_2\gamma_3 & -\gamma_3(1-\gamma_2) & (1-\gamma_2)(1-\gamma_3)-\gamma_1 \end{bmatrix}, \tag{9c}$$

where $\gamma_1 \equiv w_{12}$, $\gamma_2 \equiv w_{21} = w_{23}$, $\gamma_3 \equiv w_{32}$, and $w_{13} = w_{31} \equiv 0$.

The BACP estimator for scenario II is thus given by

$$\breve{p}_{j(II)} = \sum_{k=1}^{K} v_{jk(II)} \cdot \hat{p}_k, \quad \text{j = 1, 2, …, K.} \tag{10}$$

By using Eqs. 9b and 10, the feasibility and admissibility constraints for the misclassification probability and BACP are given respectively as follows:

$$\gamma_1 + \gamma_2 + \gamma_3 < 1, \tag{11a}$$

and

$$\gamma_2 < \hat{p}_2. \tag{11b}$$

To test whether the data in table 1 are misclassified, we need to test the following (sharp) null hypothesis that the data has no misclassification under p = p^0 versus the alternative hypothesis that the data are misclassified (Berger and Selleke

1987)

$$H_0: p = p^0, \omega = 0 \text{ versus } H_1: p \neq p^0, \omega > 0, \tag{12}$$

where $p = (p_1,..., p_K)^T$, $p^0 = (p_1^0,..., p_K^0)^T$, $\omega = (w_{11},..., w_{1K}, w_{21},..., w_{2K},..., w_{K1},...w_{KK})^T$, $\{w_{jk}\}$ are the entries of the misclassification matrix W given by Eq. 2.

To test Eq. 12 the bias-adjusted chi-square test (BACST) is given by

$$\breve{\Psi}_K = \sum_{k=1}^{K} N[(\breve{p}_k - p_k^0)^2 / p_k^0] = \sum_{k=1}^{K} (\breve{n}_k^2 / n_k^0) - N , \tag{13}$$

where $\breve{n}_k = \sum_{j=1}^{K} v_{jk} n_j$, v_{jk} denotes the entry of the j^th^ row and the k^th^ column of the inverse matrix V of the misclassification matrix W in Eq. 2 and $n_k^0 = Np_k^0$, k = 1,…, K.

For large samples, Eq. 13 is distributed under H_0 asymptotically as the central chi-square distribution with K – 1 degrees of freedom (df). Yet Eq. 13 is distributed asymptotically under H_1 as the noncentral chi-square distribution with K – 1 degrees of freedom and the non-centrality parameter given by (Lancaster 1969)

$$\breve{\lambda}_K = \sum_{j=1}^{K} (p_j - p_j^0)^2 = \sum_{j=1}^{K} (p_j^2 - 2p_j^0 p_j + p_j^{02}) . \tag{14}$$

When $w_{jk} = 0$ for all j and k, Eq. 13 reduces to

$$\hat{\Psi}_K = \sum_{j=1}^{K} (n_j^2 / n_j^0) - N . \tag{15}$$

Reject the null hypothesis H_0 if $\hat{\Psi}_K \geq C_0$, where $\hat{\Psi}_K$ is given by Eq. 15 and C_0 is the critical value of the central chi-square distribution with K – 1 df at the significance level α

As is well known from the Bayesian viewpoint, the p-value is not an adequate measure for the evidence to support the null hypothesis (Goodman 1999a-b). Hence the Bayes factor is calculated as a comparison with the p-value. To formulate the hypothesis-testing problem in a Bayesian setting we begin with the data $n = (n_1, n_2,..., n_K)$ and assume that its probability distribution follows in a family of distributions which are parameterized by $(p, \omega) \in \Sigma \times \Omega$, where $\Sigma = \{p \mid \sum_{k=1}^{K} p_k = 1, p_k > 0\}$ is the K-dimensional simplex. To test the hypotheses of $H_0 : p = p^0, \omega = 0$ vs $H_1 : p \neq p^0, \omega > 0$ (Eq.12), it is assumed that there exist a prior probability density function (PDF) $h_0(\omega)$ and another joint density $h(p, \omega)$ under H_1. Since p and ω are a priori independent under H_1, we have

$$h(p, \omega) = h_0(\omega) g(p) , \tag{16}$$

where g is a prior PDF on p ϵ Σ which assigns mass π_0 to $\{p = p^0\}$ and $1 - \pi_0$ to $\{p \neq p^0\}$. Define $g(p^0) = 0$ and writing the PDF of $\breve{\Psi}_K$ given p and ω as $f(\breve{\Psi}_K \mid p, \omega)$, the Bayes factor is given by (Kass and Raftery 1995)

$$B^g(\breve{\Psi}_K) = \frac{f(\breve{\Psi}_K \mid p^0, \omega = 0)}{m_g(\breve{\Psi}_K)} , \tag{17a}$$

where m_g is given by

$$m_g(\breve{\Psi}_K) = \iint\limits_{\Sigma\times\Omega} f(\breve{\Psi}_K \mid p, \omega) h_0(\omega) g(p) d\omega dp\,. \tag{17b}$$

In Eq. 17a, $f(\breve{\Psi}_K \mid p^0, \omega = 0)$ is the PDF of the central chi-square distribution with K – 1 df, while $f(\breve{\Psi}_K \mid p, \omega)$ in Eq. 17b is the PDF of the noncentral chi-square distribution with K – 1 degrees of freedom and the non-centrality parameter $\breve{\lambda}_K$ given by Eq. 14.

When K = 3, $m_g(\breve{\Psi}_{3(I)})$ of Eq. 17b is calculated for Scenario I with the assumption of $\theta_1 = \theta_2 = \theta_3 \equiv \theta$ and $h_0(\theta) = c^{-1}$, the PDF of uniform distribution over [0, c], where c is the upper bound on the admissible BACP for scenario I and obtain

$$m_g(\breve{\Psi}_{3(I)}) = \int\limits_{\Sigma}\int\limits_{0}^{c} \frac{1}{c} \cdot \frac{1}{2+\breve{\lambda}_3} \cdot \exp(-\frac{t}{2+\breve{\lambda}_3}) d\theta \cdot g(p) dp\,, \tag{18}$$

where an approximation to the noncentral chi-square distribution is provided by using the central chi-square distribution (Cox and Reid 1987). The lower bound for the Bayes factor after using a symmetric Dirichlet's prior for g(p) are obtained under scenario I and II:

$$\underline{B}_i^g = \frac{\frac{1}{2}\exp\{-\frac{1}{2}[\sum_{j=1}^{3}(n_j^2 / n_j^0) - N]\}}{m_g(\tau_{\max(i)} \mid \breve{\Psi}_{3(i)})} \quad \text{, i = I or II.} \tag{19}$$

The details for obtaining the value of $\tau_{\max(i)}$, i = I or II, are given in the appendix.

3. Example

The data in table 2 are taken from table C.1 in Woodward's book, pp. 756-760 (Woodward 2005). It represents the lung cancer data collected by the Bombay Cancer Registry from all cancer patients registered in the 168 government and private hospitals and nursing homes in Bombay, Australia, and from death records maintained by the Bombay Municipal Corporation. The survival times of each subject with lung cancer from time of first diagnosis to death (or censoring) were recorded over the period 1st January 1989 to 31st December 1991. Here we are only concerned with type of tumor of 682 subjects grouped by gender.

Table 2. 682 cancer patients are classified by sex and type-of-tumor

	Type of tumor			
Gender	Local	Regional	Advanced	Total
Male	165	169	229	563
Female	37	39	43	119

The issue of concern here is whether the data are misclassified separately for males and females. Because we do not have any prior belief on the values of p^0 in Eq. 12, they are thereby determined empirically from the observed data. As a result, the values of p^0 are chosen differently for males and females. For females the values of p^0 in the null hypothesis are chosen to be that of equiprobability, $H_{0(F)}: p_1 = p_2 = p_3 = \frac{1}{3}$ and $w_{jk} = 0$ vs $H_{1(F)}: p_1 \neq p_2 \neq p_3 \neq \frac{1}{3}$ and $w_{jk} > 0$, while that of p^0 in the null hypothesis for males are set up as follows: $H_{0(M)}: p_1 = 0.3, p_2 = 0.3, p_3 = 0.4$ and $w_{jk} = 0$

vs $H_{1(M)}: p_1 \neq 0.3, p_2 \neq 0.3, p_3 \neq 0.4$ and $w_{jk} > 0$. Because the misclassification probabilities of $\{w_{jk}\}$, j, k = 1, 2, 3 are zero under the null hypothesis, the BACST values of Eq. 15 are then given respectively by $\hat{\Psi}_M$ = 0.15 (p-value = 0.93) and $\hat{\Psi}_F$ = 0.47 (p-value = 0.79) for males and females. Therefore, the null hypothesis H_0 is not rejected at the significance level of 0.05 for both males and females. Yet, we would like to test the above hypotheses from the Bayesian perspective by calculating the Bayes factor as a comparison with the p-value.

For both males and females under scenarios I or II, Eq. A10 in the appendix has three negative and one positive real, and a pair of conjugate complex roots. Due to the constraint that $\tau > 0$, only the positive root is a stationary point for Eq. A9. Eq. A9 for males has only under scenario II a unique positive local maximum (Figure 1), while Eq. A9 has a unique positive local maximum at its stationary point for females only under scenario I (Figure 2).

Table 3. A comparison of the lower bound for Bayes factor (Eq. 19) with the p-value for admissible CF models

Scenario II					
Males	c_2	$\tau_{\max(II)}$	$m_g(\tau_{\max(II)})$	$\underline{B}_{II}^{g}$	p-value
Table 2	0.293073	0.0553	61	0.053	0.93
Scenario I					
Females	c_1	$\tau_{\max(I)}$	$m_g(\tau_{\max(I)})$	$\underline{B}_{I}^{g}$	p-value
Table 2	0.310924	0.0540	1.8	0.22	0.79

By taking the reciprocal of the lower bound of the Bayes factor (table 3, column 5) we are able to assess the evidence whether the cancer data in table 2 are misclassified. The collected data for males were in favor of supporting H_1 against H_0 by at most a factor of “19 to 1”, whereas for females by at most a factor of “5 to 1”.

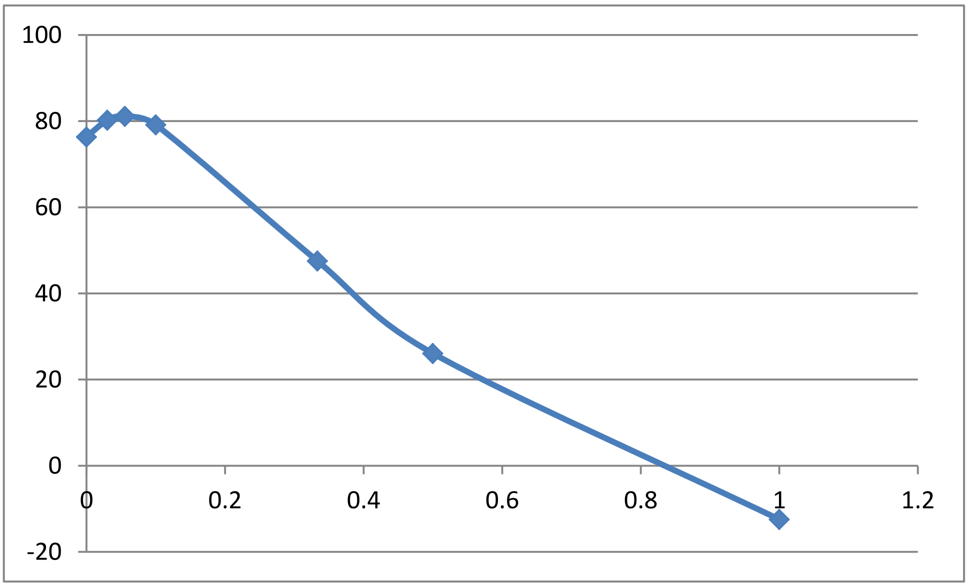

Figure 1. A plot of $m_g(\tau \mid \dot{\Psi}_{3(II)})$ given by Eq. A9 is for CF model 10 under scenario II for males

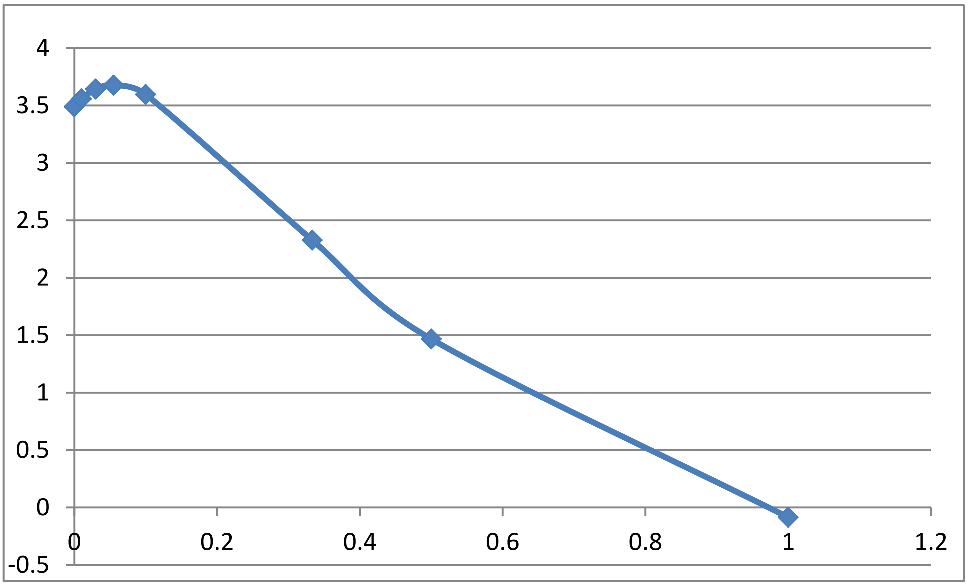

Figure 2. A plot of $m_g(\tau \mid \dot{\Psi}_{3(I)})$ given by Eq. A9 is for CF model 12 under scenario I for females

4. Discussion

Some interesting observations are worthy to be mentioned below:

1. So far, this author is not aware of any guideline available in the literature on deciding how large the lower bound for the Bayes factor should be so that we're confident the evidence provided by the data surely supporting H_1 rather than H_0. Yet, since the lower bounds for the Bayes factor from the cancer data for both genders were not large enough, a tentative conclusion was that the cancer data in table 2 seemed unlikely to be misclassified. Although H_0 was not rejected for both gender in table 2 either according to their p-values (table 3, column 6), the p-value is, strictly speaking, not an appropriate measure for assessing the evidence provided by the data due to its inherent fallacy (Goodman 1999a-b).
2. From the analysis of the Bombay cancer data, the existence of Bayes factor seems to depend not only on the scenario (I or II) (the misclassification pattern), but also the multinomial distribution of p^0 (table 3). To clarify this issue, another data set related to the degree of severity for the clinical condition of myocardial infarction patients was studied (Snow 1965), where the distribution of p^0 for the treated and control groups are respectively specified as (0.4, 0.4, 0.2) and (0.3, 0.4, 0.3). It was found that the Bayes factor existed for the treated group under scenario I, but not under scenario II, whereas for the control group it exists under both scenarios. It seems that a crucial condition for the existence of Bayes factor is whether the BACST value (Eq. 13) is positive. As far as the existence of the Bayes factor is concerned, I'd like to make a conjecture which is given as follows:

"For any data set under either scenario I or II the lower bound of $\underline{B}_i^g$, i = I or II, exists if the associated $\breve{\Psi}_{K(.)}$ of Eq. 13 is positive for K ≥ 3."

5. Conclusion

This paper addresses an issue: "how to test whether the collected categorical data are misclassified." A mixed Bayesian approach is used to test the null hypothesis that the collected data are not misclassified under a specified multinomial distribution for the studied categorical variable. The Bayes factor is employed as the main instrument to assess the evidence provided by the data. The lung cancer from all hospitals in the city of Bombay, Australia was used as an example for illustration. Based on the result of the Bayes factor in this study, the p-value was shown again not an appropriate measure to assess the evidence provided by the data.

References

Agresti, A. (2002). Categorical Data Analysis, 2nd edition. Wiley, New York.

Berger, J. O., & Selleke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and Evidence. J. Am. *Stat. Assoc.*, *82***,** 112-122.

Bross, I. (1954). Misclassification in 2 ×2 tables. *Biometrics, 10***,** 478-486.

Cox, D. R., & Reid, N. (1987). Approximations to noncentral distributions. Canad. J. Stat., 15**,** 105-114.

Diamond, E. L., & Lilienfeld, A. M. (1962a). Effects of errors in classification and diagnosis in various type of epidemiological studies. *Am. J. Public Health*, *52***,** 1137-1144.

Diamond, E. L., & Lilienfeld, A. M. (1962b). Misclassification errors in 2 ×2 tables with one margin fixed: some further comments. *Am. J. Public Health, 52*, 2106-2110.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical Methods for Rates and Proportions, 3rd edition. Wiley, New York.

Good, I. J. (1975). The Bayes factor against equiprobability of a multinomial population assuming a symmetric Dirichlet prior. *Ann. Stat., 3*, 246-250.

Good, I. J., & Crook, J. F. (1974). The Bayes/Non-Bayes compromise and the multinomial distribution. *J. Am. Stat. Assoc., 69*, 711-720.

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The p value fallacy. *Ann. Intern. Med., 130*, 995-1004.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med., 130,* 1005-1013.

Gustafson, P. (2004). Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments. Chapman & Hall, Boca Raton, FL.

Kass, R. E., & Raftery, A. E. (1995). Bayes factor. *J. Am. Stat. Assoc., 90*, 773-795.

Lancaster, H. O. (1969). The Chi-squared Distribution. Wiley, New York.

Mote, V. L., & Anderson, R. L. (1965). An investigation of the effect of misclassification on the properties of χ2-tests in the analysis of categorical data. *Biometrika, 52*, 95-109.

Redfern, D., & Campbell, C. (1998). The MATLAB 5 Handbook. Springer, New York.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). Modern Epidemiology, 3rd edition. Lippincott Williams & Wilkins, Philadelphia, PA.

Snow, P. J. D. (1965). Effects of propranolol in myocardial infarction. *Lancet, 286*(Sept 18)*,* 551-553.

Woodward, M. (2005). Epidemiology: Study Design and Data Analysis, 2nd ed. Chapman & Hall/CRC Press, Boca Raton, Florida.

Appendix A

With an assumption of $\theta_1 = \theta_2 = \theta_3 \equiv \theta$ and $\breve{n} = N\breve{p}$, we have under scenario I

$$\breve{n}_{(I)} \equiv \begin{bmatrix} \breve{n}_{1(I)} \\ \breve{n}_{2(I)} \\ \breve{n}_{3(I)} \end{bmatrix} = (1-3\theta)^{-1} \begin{bmatrix} (1-\theta)n_1 - \theta n_2 - \theta n_3 \\ -\theta n_1 + (1-\theta)n_2 - \theta n_3 \\ -\theta n_1 - \theta n_2 + (1-\theta)n_3 \end{bmatrix}, 0 < \theta < \mathrm{c}_1, \qquad \text{(A1)}$$

where $c_1 = \min_{j=1,2,3} \{\frac{1}{3}, \hat{p}_j\}$.

By substituting Eq. A1 into Eq. 13, we have

$$\breve{\Psi}_{3(I)} = \sum_{j=1}^{3} a_{j(I)} n_j^2 - 2\sum_{j \neq k} a_{jk(I)} n_j n_k - N , \qquad \text{(A2)}$$

where

$$a_{1(I)} = (1-3\theta)^{-2}[(1-\theta)^2 / n_1^0 + \theta^2 (1/n_2^0 + 1/n_3^0)] ,$$

$$a_{2(I)} = (1-3\theta)^{-2}[(1-\theta)^2 / n_2^0 + \theta^2 (1/n_1^0 + 1/n_3^0)] ,$$

$$a_{3(I)} = (1-3\theta)^{-2}[(1-\theta)^2 / n_3^0 + \theta^2 (1/n_1^0 + 1/n_2^0)] ,$$

$$a_{12(I)} = (1-3\theta)^{-2}[\theta(1-\theta)(1/n_1^0 + 1/n_2^0) - \theta^2 / n_3^0] ,$$

$$a_{13(I)} = (1-3\theta)^{-2}[\theta(1-\theta)(1/n_1^0 + 1/n_3^0) - \theta^2 / n_2^0] ,$$

$$a_{23(I)} = (1-3\theta)^{-2}[\theta(1-\theta)(1/n_2^0 + 1/n_3^0) - \theta^2 / n_1^0] .$$

By Eq. 14, we have

$$\breve{\lambda}_3 = \sum_{j=1}^{3} (p_j^2 - 2p_j^0 p_j + p_j^{02}) . \qquad \text{(A3)}$$

Note that $m_g(\breve{\Psi}_K)$ of Eq. 18 with a choice of $h_0(\theta)$ which equals to the pdf of uniform distribution over [0, c$_1$] is reduced to

$$m_g(\breve{\Psi}_3) = \int_\Sigma \int_0^{c_1} \frac{1}{c_1(2+\breve{\lambda}_3)} \cdot \exp(-\frac{\breve{\Psi}_3}{2+\breve{\lambda}_3}) d\theta \cdot g(p) dp , \qquad \text{(A4)}$$

where $\breve{\Psi}_{3(I)}$ and $\breve{\lambda}_3$ are given respectively by Eqs. A2 and A3. By using a linear approximation from the Taylor series expansion of $\exp(-\breve{\Psi}_3 /(2+\breve{\lambda}_3))$ and another linear approximation to $(2+\breve{\lambda}_3)^{-1}$, Eq. A4 simplifies under scenario I to

$$m_g(\breve{\Psi}_{3(I)}) \approx \int_\Sigma \int_0^{c_1} [\tfrac{1}{4}\breve{\lambda}_3^3 - \tfrac{1}{4}(2+\breve{\Psi}_{3(I)})\breve{\lambda}_3^2 + (\breve{\Psi}_{3(I)} - 1)\breve{\lambda}_3 + 2 - \breve{\Psi}_{3(I)}] \cdot c_1^{-1} d\theta \cdot g(p) dp .$$

By substituting Eqs. A2 and A3 into the above equation and integrating $\breve{\Psi}_{3(I)}$ with respect to θ, we have after algebraic simplification

$$m_g(\dot{\Psi}_{3(I)}) = \tfrac{1}{c_1} \int_\Sigma g(p) \{ \tfrac{1}{4} [\sum_{j=1}^{3} p_j^6 + 3 \sum_{j \neq k \neq \ell} p_j^4 (p_k^2 + p_\ell^2) + 6\prod_{j=1}^{3} p_j^2 - 6(\sum_{j=1}^{3} p_j^0 p_j^5 + \sum_{j \neq k \neq \ell} p_j^4 (p_k^0 p_k + p_\ell^0 p_\ell))$$

$$-12(\sum_{j\neq k\neq\ell} p_j^0 p_j^3 (p_k^2+p_\ell^2)+\sum_{j\neq k\neq\ell} p_j^0 p_j p_k^2 p_\ell^2)+12[\sum_{j=1}^{3} p_j^{02} p_j^4+\sum_{j\neq k}(p_j^{02}+p_k^{02})p_j^2 p_k^2+2(\sum_{j\neq k} p_j^0 p_k^0 (p_j^3 p_k+p_j p_k^3)$$

$$+\sum_{j\neq k\neq\ell} p_j^0 p_k^0 p_j p_k p_\ell^2)]+\tfrac{1}{4}(3\rho_0-2-\dot{\Psi}_{3(I)})\sum_{j=1}^{3} p_j^4+\tfrac{1}{2}(3\rho_0-2-\dot{\Psi}_{3(I)})\sum_{j\neq k} p_j^2 p_k^2-2[\sum_{j=1}^{3} p_j^{03} p_j^3$$

$$+3\sum_{j\neq k\neq\ell} p_j^{02} p_j^2 (p_k^0 p_k+p_\ell^0 p_\ell)+6\prod_{j=1}^{3} p_j^0 p_j]-(3\rho_0-2-\dot{\Psi}_{3(I)})[\sum_{j=1}^{3} p_j^0 p_j^3+\sum_{j\neq k\neq\ell} p_j^0 p_j (p_k^2+p_\ell^2)]+3\rho_0\sum_{j=1}^{3} p_j^{02} p_j^2$$

$$-(2+\dot{\Psi}_{3(I)})\sum_{j=1}^{3} p_j^0 p_j^2+\tfrac{1}{4}[3\rho_0^2-4\rho_0-4+2(2-\rho_0)\dot{\Psi}_{3(I)}]\sum_{j=1}^{3} p_j^2+[6\rho_0-2(2+\dot{\Psi}_{3(I)})]\sum_{j\neq k} p_j^0 p_k^0 p_j p_k$$

$$+[2(\rho_0+1)+(\rho_0-2)\dot{\Psi}_{3(I)}]\sum_{j=1}^{n} p_j^0 p_j+\tfrac{1}{4}\rho_0[\rho_0^2-2\rho_0-4+(4-\rho_0)\dot{\Psi}_{3(I)}]\}dp\,, \tag{A5}$$

where

$$\dot{\Psi}_{3(I)}=\int_0^{c_1}\breve{\Psi}_{3(I)}d\theta=\sum_{j=1}^{3}\dot{a}_{j(I)}n_j^2-2\sum_{j\neq k}\dot{a}_{jk(I)}n_j n_k-Nc_1\,,$$

$$\dot{a}_{1(I)}=b_{1(I)}/n_1^0+b_{2(I)}(1/n_2^0+1/n_3^0)\,,$$

$$\dot{a}_{2(I)}=b_{1(I)}/n_2^0+b_{2(I)}(1/n_1^0+1/n_3^0)\,,$$

$$\dot{a}_{3(I)}=b_{1(I)}/n_3^0+b_{2(I)}(1/n_1^0+1/n_2^0)\,,$$

$$\dot{a}_{12(I)}=b_{3(I)}(1/n_1^0+1/n_2^0)-b_{2(I)}/n_3^0\,,$$

$$\dot{a}_{13(I)}=b_{3(I)}(1/n_1^0+1/n_3^0)-b_{2(I)}/n_2^0\,,$$

$$\dot{a}_{23(I)}=b_{3(I)}(1/n_2^0+1/n_3^0)-b_{2(I)}/n_1^0\,,$$

$$b_{1(I)}=\tfrac{1}{9}[c_1(5-3c_1)(1-3c_1)^{-1}-\tfrac{4}{3}\ln(1-3c_1)]\,,$$

$$b_{2(I)}=\tfrac{1}{9}[\tfrac{2}{3}\ln(1-3c_1)+c_1(2-3c_1)(1-3c_1)^{-1}]\,,$$

$$b_{3(I)}=\tfrac{1}{9}[\tfrac{1}{3}\ln(1-3c_1)+c_1(1+3c_1)(1-3c_1)^{-1}]\,,$$

$$\rho_0=\sum_{j=1}^{3} p_j^{02}\,.$$

With an assumption of $\gamma_1=\gamma_2=\gamma_3\equiv\gamma$, we have under scenario II

$$\breve{n}_{(II)}\equiv\begin{bmatrix}\breve{n}_{1(II)}\\ \breve{n}_{2(II)}\\ \breve{n}_{3(II)}\end{bmatrix}=[(1-\gamma)(1-3\gamma)]^{-1}\begin{bmatrix}(1-\gamma)^2 n_1-\gamma(1-\gamma)n_2+\gamma^2 n_3\\ -\gamma(1-\gamma)n_1+(1-\gamma)^2 n_2-\gamma(1-\gamma)n_3\\ \gamma^2 n_1-\gamma(1-\gamma)n_2+(1-3\gamma+\gamma^2)n_3\end{bmatrix},\ 0<\gamma<\mathrm{c}_2, \tag{A6}$$

where $c_2\equiv c_1=\min_{j=1,2,3}\{\tfrac{1}{3},\hat{p}_j\}$.

By using Eq. A6, we have

$$\sum_{j=1}^{3}\breve{n}_{j(II)}^{2} = \sum_{j=1}^{3} a_{j(II)} n_j^2 - 2[a_{12(II)}n_1n_2 + a_{23(II)}n_2n_3 - a_{13(II)}n_1n_3], \qquad \text{(A7)}$$

where

$$a_{1(II)} = [(1-\gamma)^4 / n_1^0 + \gamma^2(1-\gamma)^2 / n_2^0 + \gamma^4 / n_3^0]/[(1-\gamma)(1-3\gamma)]^2,$$

$$a_{2(II)} = [\gamma^2(1-\gamma)^2(1/n_1^0 + 1/n_3^0) + (1-\gamma)^4 / n_2^0]/[(1-\gamma)(1-3\gamma)]^2,$$

$$a_{3(II)} = [(\gamma^4 / n_1^0 + \gamma^2(1-\gamma)^2 / n_2^0 + (\gamma^2 - 3\gamma + 1)^2 / n_3^0]/[(1-\gamma)(1-3\gamma)]^2,$$

$$a_{12(II)} = [\gamma(1-\gamma)^3(1/n_1^0 + 1/n_2^0) + \gamma^3(1-\gamma)/n_3^0]/[(1-\gamma)(1-3\gamma)]^2,$$

$$a_{23(II)} = [\gamma^3(1-\gamma)/n_1^0 + \gamma(1-\gamma)^3 / n_2^0 + \gamma(1-\gamma)(\gamma^2 - 3\gamma + 1)/n_3^0]/[(1-\gamma)(1-3\gamma)]^2,$$

$$a_{13(II)} = [\gamma^2(1-\gamma)^2(1/n_1^0 + 1/n_2^0) + \gamma^2(\gamma^2 - 3\gamma + 1)/n_3^0]/[(1-\gamma)(1-3\gamma)]^2.$$

By substituting Eq. A7 into Eq. 13 and integrating $\breve{\Psi}_{3(II)}$ with respect to γ over [0, c$_2$], we have

$$\dot{\Psi}_{3(II)} = \int_0^{c_2} \breve{\Psi}_{3(II)} d\gamma = \sum_{j=1}^{3} \dot{a}_{j(II)} n_j^2 - 2[\dot{a}_{12(II)}n_1n_2 + \dot{a}_{23(II)}n_2n_3 - \dot{a}_{13(II)}n_1n_3] - Nc_2, \text{ (A8)}$$

where

$$\dot{a}_{1(II)} = b_{1(II)}/n_1^0 + b_{2(II)}/n_2^0 + b_{3(II)}/n_3^0,$$

$$\dot{a}_{2(II)} = b_{1(II)}/n_2^0 + b_{2(II)}(1/n_1^0 + 1/n_3^0),$$

$$\dot{a}_{3(II)} = b_{3(II)}/n_1^0 + b_{2(II)}/n_2^0 + b_{4(II)}/n_3^0,$$

$$\dot{a}_{12(II)} = b_{5(II)}(1/n_1^0 + 1/n_2^0) + b_{6(II)}/n_3^0,$$

$$\dot{a}_{23(II)} = b_{6(II)}/n_1^0 + b_{5(II)}/n_2^0 + b_{7(II)}/n_3^0,$$

$$\dot{a}_{13(II)} = b_{2(II)}(1/n_1^0 + 1/n_2^0) + b_{8(II)}/n_3^0,$$

$$b_{1(II)} = \int_0^{c_2} \frac{(1-\gamma)^2}{(1-3\gamma)^2} d\gamma = \tfrac{1}{27}[\frac{3c_2(5-3c_2)}{1-3c_2} - 4\ln(1-3c_2)],$$

$$b_{2(II)} = \int_0^{c_2} \frac{\gamma^2}{(1-3\gamma)^2} d\gamma = \tfrac{1}{27}[\frac{3c_2(2-3c_2)}{1-3c_2} + 2\ln(1-3c_2)],$$

$$b_{3(II)} = \int_0^{c_2} \frac{\gamma^4}{[(1-\gamma)(1-3\gamma)]^2} d\gamma = \tfrac{1}{108}[\frac{3c_2(12c_2^2 - 44c_2 + 14)}{(1-c_2)(1-3c_2)} + 27\ln(1-c_2) + 5\ln(1-3c_2)],$$

$$b_{4(II)} = \int_0^{c_2} \frac{(\gamma^2 - 3\gamma + 1)^2}{[(1-\gamma)(1-3\gamma)]^2} d\gamma = \tfrac{1}{108}[\frac{3c_2(12c_2^2 - 44c_2 + 14)}{(1-c_2)(1-3c_2)} - 27\ln(1-c_2) - 23\ln(1-3c_2)]\,,$$

$$b_{5(II)} = \int_0^{c_2} \frac{\gamma(1-\gamma)}{(1-3\gamma)^2} d\gamma = \tfrac{1}{27}[\frac{3c_2(1+3c_2)}{1-3c_2} + \ln(1-3c_2)]\,,$$

$$b_{6(II)} = \int_0^{c_2} \frac{\gamma^3}{(1-\gamma)(1-3\gamma)^2} d\gamma = \tfrac{1}{108}[\frac{3c_2(6c_2+1)}{1-3c_2} - 27\ln(1-c_2) + 7\ln(1-3c_2)]\,,$$

$$b_{7(II)} = \int_0^{c_2} \frac{\gamma(\gamma^2 - 3\gamma + 1)}{(1-\gamma)(1-3\gamma)^2} d\gamma = \tfrac{1}{108}[\frac{6c_2(3c_2-2)}{1-3c_2} + 63\ln(1-c_2) - 31\ln(1-3c_2)]$$
,

$$b_{8(II)} = \int_0^{c_2} \frac{\gamma^2(\gamma^2 - 3\gamma + 1)}{[(1-\gamma)(1-3\gamma)]^2} d\gamma = \tfrac{1}{54}[\frac{3c_2(6c_2^2 + 5c_2 - 2)}{(1-c_2)(1-3c_2)} - 2\ln(1-3c_2)]\,.$$

If the prior distribution function for g(p) is taken to be a symmetric Dirichlet's distribution with the flattening constant (or hyper-parameter) τ (τ > 0) (Good 1975), then Eq. A5 is reduced to

$$m_g(\tau \mid \dot{\Psi}_{3(I)}) = \tfrac{1}{c_1} \frac{d_4\tau^4 + d_3\tau^3 + d_2\tau^2 + d_1\tau + d_0}{12(3\tau+1)(3\tau+2)(3\tau+4)(3\tau+5)}\,, \tag{A9}$$

where

$$d_4 = 243\rho_0^3 - 1377\rho_0^2 + 891\rho_0 + (648\rho_0 - 432)\rho_1 - 216\rho_2 - 72\rho_3 - (243\rho_0^2 - 1782\rho_0 + 432\rho_1 + 1647)\dot{\Psi}_{3(I)} + 1059\,,$$

$$d_3 = 972\rho_0^3 - 5130\rho_0^2 + 2568\rho_0 + (1176\rho_0 + 464)\rho_1 - 864\rho_2 - 432\rho_3 - (972\rho_0^2 - 7020\rho_0 + 392\rho_1 + 6470)\dot{\Psi}_{3(I)} + 4382\,,$$

$$d_2 = 1323\rho_0^3 - 6101\rho_0^2 + 5307\rho_0 + (2736\rho_0 + 1728)\rho_1 - 1128\rho_2 - 1240\rho_3 - (1323\rho_0^2 - 9306\rho_0 + 912\rho_1 + 8425)\dot{\Psi}_{3(I)} + 5943\,,$$

$$d_1 = 702\rho_0^3 + 2450\rho_0^2 + 138\rho_0 + (960\rho_0 + 1920)\rho_1 - 1440\rho_2 - 2736\rho_3 - (702\rho_0^2 - 4692\rho_0 + 320\rho_1 - 5274)\dot{\Psi}_{3(I)} + 2722\,,$$

$$d_0 = 40[3\rho_0^3 - 15\rho_0^2 + 48\rho_0 - 24\rho_3 - (3\rho_0^2 - 18\rho_0 - 14)\dot{\Psi}_{3(I)} + 3]\,,$$

$$\rho_1 = \sum_{j \neq k} p_j^0 p_k^0\,,$$

$$\rho_2 = \sum_{j \neq k \neq \ell} (p_j^0 + p_k^0) p_\ell^{02}\,,$$

$$\rho_3 = \sum_{j=1}^{3} p_j^{03}\,.$$

Similarly, $m_g(\tau \mid \dot{\Psi}_{3(II)})$ has exactly the same expression like Eq. A9 except that $\dot{\Psi}_{3(I)}$ and c_1 are replaced respectively by $\dot{\Psi}_{3(II)}$ and c_2.

To avoid the use of hyper-prior distribution on τ (Good and Crook 1974), the non-Bayesian approach is used to find the stationary point $\tau_{\max(.)}$ for $m_g(\tau \mid \breve{\Psi}_{3(.)})$. By using an elementary technique in calculus to calculate the first derivative of $m_g(\tau \mid \breve{\Psi}_{3(I)})$ and set it equal to zero, we have after simplification

$$(324d_4 - 81d_3)\tau^6 + (882d_4 - 162d_2)\tau^5 + (702d_4 + 441d_3 - 324d_2 - 243d_1)\tau^4 + (160d_4 + 468d_3 - 648d_1 - 324d_0)\tau^3$$

$$+(120d_3 + 234d_2 - 441d_1 - 972d_0)\tau^2 + (80d_2 - 882d_0)\tau + 40d_1 - 234d_0 = 0\,. \qquad \text{(A10)}$$

To solve Eq. A10 for the stationary points, I employed the “ROOTS” subroutine in the MATLAB (Redfern and Campbell 1998).

According to the terms of Good (1975), the way to estimate $\tau_{\max}$ is called by the type II maximum likelihood or the maximum hyper-prior likelihood method. This kind of approach to estimate the Bayes factor is called the Bayesian/Fisherian criterion which is a compromise from taking a full Bayesian approach.

Copyrights