


5-1-2011

# Matched-Pair Studies with Misclassified Ordinal Data

Tze-San Lee

Western Illinois University, leetzesan@gmail.com

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

## Recommended Citation

Lee, Tze-San (2011) "Matched-Pair Studies with Misclassified Ordinal Data," *Journal of Modern Applied Statistical Methods*: Vol. 10 : Iss. 1 , Article 8.

DOI: 10.22237/jmasm/1304222820

Available at: <http://digitalcommons.wayne.edu/jmasm/vol10/iss1/8>

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

## Matched-Pair Studies with Misclassified Ordinal Data

Tze-San Lee  
Western Illinois University  
Macomb, IL USA

The problem of matched-pair studies with misclassified ordinal data is considered. Misclassification is assumed to occur only between the adjacent columns/rows. Bias-adjusted generalized odds ratio and a test for marginal homogeneity are presented to account for misclassification bias. Data from lambing records of 227 Merino ewes are used to illustrate how to calculate these bias-adjusted estimators and – because validation data are not available – a sensitivity analysis is conducted.

Key words: Matched-pair, misclassification, ordinal scale.

### Introduction

Matched-pair studies with ordered categorical variables have received much attention in the literature (see Agresti, 1983, 1984; Clayton, 1974; McCullagh, 1977; Stuart, 1953, 1955). A few published studies investigated matched-pair with misclassified data (Greenland, 1982, 1989; Greenland & Kleinbaum, 1983; Lee, 2010); however, these studies consider only  $2 \times 2$  contingency tables with misclassified data. To date, matched-pair studies with misclassified data have not been investigated when the number of exposure categories is greater than two. A matched-pair misclassification problem is considered here with an exposure variable that has  $K (\geq 3)$  ordered levels. The generalized odds ratio is used for measuring the association in contingency tables with misclassified ordinal data and a test for marginal homogeneity proposed by Stuart (1955) is modified to manage the misclassified data.

### Methodology

Consider a 1:1 matched-pair study where X represents the case and Y represents the control population and the same exposure variable with

$K (\geq 3)$  ordered levels is used. Assume that a  $K \times K$  contingency table is realized with the following frequency counts:

$$A = [n_{ij}]_{i,j=1,\dots,K} \quad (1)$$

where  $\{n_{ij}\}$  are assumed to follow a multinomial distribution with parameters  $n (= \sum_{i,j} n_{ij})$  and the cell probability  $\{p_{ij} > 0\}$ . A naïve estimator ( $\hat{p}_{ij}$ ) for  $p_{ij}$  is given by

$$\hat{p}_{ij} = n_{ij} / n. \quad (2)$$

### Generalized Odds Ratio

As a measure for the association between X and Y, a generalized odds ratio (GOR),  $\zeta$ , is defined by Agresti (1980) as:

$$\zeta = \frac{p_C}{p_D} \equiv \frac{\sum_{i=1}^{K-1} \sum_{j=i+1}^K p_{ij}}{\sum_{i=2}^K \sum_{j=1}^{i-1} p_{ij}}, \quad (3)$$

where  $p_C$  (or  $p_D$ ) denotes the probability of a randomly selected matched-pair in which a case has a higher (or lower) level of exposure than his/her matched control. A naïve estimator, denoted by  $\hat{\zeta}$ , for (3) is obtained by replacing

Tze-San Lee is presently working at the Centers for Disease Control and Prevention, Mail Stop F-58, Chamblee, GA 34301, USA. Email: tj13@cdc.gov.

the unknown parameters  $p_{ij}$ ,  $i, j = 1, \dots, K$  by the sample estimator  $\hat{p}_{ij}$  shown in (2). Note that this naïve estimator of equation 3 could have substantial bias if the observed data in (1) are misclassified. Due to the faster convergence of  $\ln(\hat{\zeta})$ , a natural logarithm of  $\hat{\zeta}$ , to normality, is preferred to find a large sample Wald's  $100(1 - p)\%$  confidence interval:

$$\begin{aligned} & [\hat{\zeta} \cdot \exp(-z_{\alpha/2} \sqrt{\hat{\sigma}^2(\ln(\hat{\zeta}))}), \\ & \hat{\zeta} \cdot \exp(z_{\alpha/2} \sqrt{\hat{\sigma}^2(\ln(\hat{\zeta}))})] \end{aligned} \quad (4a)$$

for  $\zeta$ , where  $z_{\alpha/2}$  is the  $(\alpha/2)^{\text{th}}$  upper-tail percentile of the standard normal distribution. The asymptotic variance of  $\ln(\hat{\zeta})$  is given by Agresti (1980) as

$$\hat{\sigma}^2(\ln(\hat{\zeta})) = \sqrt{n^{-1}(\hat{p}_C^{-1} + \hat{p}_D^{-1})}, \quad (4b)$$

where  $\hat{p}_C$  (or  $\hat{p}_D$ ) is obtained by substituting equation 2 for  $p_{ij}$  in  $p_C$  (or  $p_D$ ).

#### A Test for Marginal Homogeneity

A global test for marginal homogeneity was proposed by Stuart (1955). A drawback of this global test is its failure to account for an ordinal nature in the categorical level of the exposure variable. Assume that the ordinal nature of the exposure variable is quantified by a variable  $U$  taking the score values  $u_k$  ( $u_1 < u_2 < \dots < u_K$ ) at the  $k^{\text{th}}$  level. Thus, the score test for the significance of the  $\beta$  coefficient in a linear trend model  $H_0: \beta_0 = \beta_1$ , where  $p_{i+} = \alpha_0 + \beta_0 u_i$  and  $p_{+j} = \alpha_1 + \beta_1 u_j$  is defined by

$$\hat{S} = \frac{n \cdot \left[ \sum_{j=1}^{K-1} \sum_{i=j+1}^K (\hat{p}_{ji} - \hat{p}_{ij})(u_j - u_i) \right]^2}{\sum_{j=1}^{K-1} \sum_{i=j+1}^K \hat{p}_{ij} (u_j - u_i)^2}, \quad (5)$$

where  $\hat{p}_{ij}$  is defined by equation 2,  $\hat{P}_{ij} = \hat{p}_{ij} + \hat{p}_{ji}$ , and  $u_i = i - 1$ ,  $i = 1, \dots, K$ . By a large sample theory (5) is distributed as a Chi-square distribution with 1 degree of freedom (Breslow, 1982).

#### Misclassification Probability

Suppose that the observed  $K \times K$  contingency table shown in (1) were misclassified with respect to both  $X$  and  $Y$ . Let  $X^*$  and  $Y^*$  be the classified surrogate variables for  $X$  and  $Y$ , respectively. Furthermore, assume that only adjacent rows in  $X^*$  or adjacent columns in  $Y^*$  are misclassified. For  $Z = X, Y$ , the misclassification probabilities (MPs) for the row or column variable are defined as follows:

$$\begin{aligned} \phi_{Z[k;j]} &= \Pr(Z^* = k + 1 | Z = k; \bar{Z} = j), \\ \overline{\phi_{Z[k;j]}} &= 1 - \phi_{Z[k;j]}; \end{aligned} \quad (6a)$$

and

$$\begin{aligned} \psi_{Z[j;k]} &= \Pr(Z^* = j - 1 | Z = j; \bar{Z} = k), \\ \overline{\psi_{Z[j;k]}} &= 1 - \psi_{Z[j;k]}; \end{aligned} \quad (6b)$$

where  $\bar{Z} = Y$  if  $Z = X$ ; and vice versa. Note that, due to symmetry,  $\phi_{Z[k;j+1]} = \psi_{Z[k+1;j]}$ . If

$$p = [p_{11}, \dots, p_{1K}, p_{21}, \dots, p_{2K}, \dots, p_{K1}, \dots, p_{KK}]^T, \quad (7a)$$

and

$$\hat{p} = [\hat{p}_{11}, \dots, \hat{p}_{1K}, \hat{p}_{21}, \dots, \hat{p}_{2K}, \dots, \hat{p}_{K1}, \dots, \hat{p}_{KK}]^T, \quad (7b)$$

then the expected value of a naïve estimator for equation 2 is given by

$$E(\hat{p}) = Wp \quad (8)$$

where the misclassification matrix  $W$  is a  $K^2 \times K^2$  matrix defined, respectively, by

$$W = \begin{bmatrix} W_{11} & W_{12} & 0 & 0 & \dots & \dots \\ W_{21} & W_{22} & W_{23} & 0 & 0 & \dots \\ 0 & W_{32} & W_{33} & W_{34} & \dots & \dots \\ 0 & 0 & W_{43} & W_{44} & \dots & \dots \\ \vdots & \vdots & 0 & W_{54} & \dots & \dots \\ \vdots & \vdots & \vdots & 0 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \vdots & \vdots \end{bmatrix} \quad (9)$$

where

$$\begin{aligned} W_{kl} &= [w_{ij}^{[kl]}], i, j, k, l = 1, 2, \dots, K; w_{11}^{[11]} = \overline{\varphi_{X[1;1]}} - \varphi_{Y[1;1]}, \\ w_{KK}^{[11]} &= \overline{\psi_{Y[K;1]}} - \varphi_{X[1;K]}, w_{22}^{[11]} = \overline{\psi_{Y[2;1]}} - \varphi_{X[1;1]} - \varphi_{Y[2;1]}, \\ w_{kk}^{[11]} &= \overline{\psi_{Y[k-1;1]}} - \varphi_{X[1;k-1]} - \varphi_{X[2;k-1]} - \dots - \psi_{Y[k-1;k-1]}, \\ k &= 3, \dots, K-1; w_{k,k+1}^{[11]} = \psi_{Y[k;k-1]}, k = 1, \dots, K-1; \\ w_{k+1,k}^{[11]} &= \varphi_{Y[k-1;k]}, k = 2, \dots, K, 0 \text{ elsewhere}; \\ W_{12} &= \text{diag}[\psi_{Y[2;1]}, \psi_{Y[3;1]}, \dots, \psi_{Y[K;1]}]; \\ W_{21} &= \text{diag}[\varphi_{Y[1;1]}, \varphi_{Y[2;1]}, \dots, \varphi_{Y[K-1;1]}]; w_{K+i,K+i}^{[22]} = \varphi_{X[1;1]}, \\ i &= 1, \dots, K-1; w_{2K,2K}^{[22]} = \overline{\psi_{Y[2K;2K-1]}} - \psi_{X[2K-1;2K]} - \varphi_{X[2K-1;2K]}, \\ w_{kk}^{[22]} &= \overline{\psi_{X[k-K;k-K]}} - \varphi_{X[k-K;k-K]} - \psi_{Y[k-K;k]} - \varphi_{Y[k-K;k-K]}, \\ k &= K+2, \dots, 2K-1, \dots, w_{k,k}^{[kk]} = \varphi_{X[k-K;k]}, \\ k &= (K-1)K+1, \dots, K^2-1, \\ w_{k,k}^{[kk]} &= \overline{\psi_{X[K-1;K(K-1)+1]}} - \varphi_{Y[1;K(K-1)+1]} - \psi_{X[K-1;k]}, \\ k &= K(K-1)+1, \dots, K^2-1, w_{K^2,K^2}^{[KK]} = \overline{\psi_{X[K-1;K^2]}} - \psi_{Y[K-1;K^2]}. \end{aligned}$$

Note that W is a block tri-diagonal matrix.

### Bias-Adjusted Cell Proportion

Using equation 8, a bias-adjusted cell proportion (BACP) estimator for p is given by

$$\tilde{p} = [\tilde{p}_{11}, \dots, \tilde{p}_{1K}, \dots, \tilde{p}_{K1}, \dots, \tilde{p}_{KK}]^T = W^{-1} \hat{p} = V \hat{p}, \quad (10a)$$

where  $\hat{p}$  is defined by (2) and V is defined by

$$W^{-1} \equiv V = [v_{ij}]_{i,j=1,\dots,K^2}. \quad (10b)$$

The appendix shows how to calculate its inverse V of the misclassification matrix W for K = 3, which was used to analyze the data for Table 1. When K = 3, then for i, j = 1, 2, 3

$$\tilde{p}_{ij} = \sum_{k=1}^3 (v_{ik} \hat{p}_{ij} + v_{i,k+3} \hat{p}_{i+1,j} + v_{i,k+6} \hat{p}_{i+2,j}), \quad (11)$$

where  $\{v_{ij}\}$ , i, j = 1, 2, ..., 9 are given respectively by equation A5 in the appendix with

$$\begin{aligned} \det(W_{11}) &= 1 - 2\varphi_{Y[1;1]} - 3\varphi_{Y[2;1]} - 2\psi_{Y[3;1]} \\ &\quad + 5\varphi_{Y[2;1]} (\varphi_{Y[1;1]} + \psi_{Y[3;1]}) \\ &\quad + 4\varphi_{Y[1;1]} \psi_{Y[3;1]} (1 - 2\overline{\varphi_{Y[2;1]}}) \end{aligned} \quad (12)$$

$$\begin{aligned} \det(\Delta_i) &= \delta_{11}^{[i]} (\delta_{22}^{[i]} \delta_{33}^{[i]} - \delta_{23}^{[i]} \delta_{32}^{[i]}) \\ &\quad - \delta_{12}^{[i]} (\delta_{21}^{[i]} \delta_{33}^{[i]} - \delta_{31}^{[i]} \delta_{23}^{[i]}) \\ &\quad + \delta_{13}^{[i]} (\delta_{21}^{[i]} \delta_{32}^{[i]} - \delta_{31}^{[i]} \delta_{22}^{[i]}), i = 1, 2 \end{aligned} \quad (13 \& 14)$$

Where  $\{\delta_{jk}^{[i]}\}$ , i = 1, 2 are given, respectively, by equations A5 and A6 in the appendix.

A set of MPs is said to be feasible if the values of all three determinants,  $\det(W_{11})$ ,  $\det(\Delta_1)$  and  $\det(\Delta_2)$ , from (12), (13) and (14), are nonzero for the given set of equation 6. Furthermore, a set of MPs is said to be admissible if - for all feasible  $\varphi_{Z[i;j]}$  and  $\psi_{Z[j;i]}$  - where Z = X, Y, the constraint of the sum of total probability  $\{\tilde{p}_{ij}\}$ , i, j = 1, 2, 3, that is,

$$\sum_{i=1}^3 \sum_{j=1}^3 \tilde{p}_{ij} = 1, \text{ is satisfied where } 0 < \tilde{p}_{ij} < 1.$$

### Bias-Adjusted Generalized Odds Ratio

By substituting (11) into (3), a bias-adjusted generalized odds ratio (BAGOR) is thus defined by

## MATCHED-PAIR STUDIES WITH MISCLASSIFIED ORDINAL DATA

$$\zeta = \frac{\bar{p}_C}{\bar{p}_D} \equiv \frac{\bar{p}_{12} + \bar{p}_{13} + \bar{p}_{23}}{\bar{p}_{21} + \bar{p}_{31} + \bar{p}_{32}} = \frac{\sum_{i,j=1}^3 v_{ij}^* \hat{p}_{ij}}{\sum_{i,j=1}^3 v_{ij}^{**} \hat{p}_{ij}}, \quad (15a)$$

where  $\{\hat{p}_{ij}\}$  are given by (2),  $\{v_{ij}^*\}$  and  $\{v_{ij}^{**}\}$  are given respectively for  $j = 1, 2, 3$ , by

$$\begin{aligned} v_{1j}^* &= v_{2j} + v_{3j} + v_{6j}, \\ v_{2j}^* &= v_{2,j+3} + v_{3,j+3} + v_{6,j+3}, \\ v_{3j}^* &= v_{2,j+6} + v_{3,j+6} + v_{6,j+6}, \end{aligned} \quad (15b)$$

and

$$\begin{aligned} v_{1j}^{**} &= v_{4j} + v_{7j} + v_{8j}, \\ v_{2j}^{**} &= v_{4,j+3} + v_{7,j+3} + v_{8,j+3}, \\ v_{3j}^{**} &= v_{4,j+6} + v_{7,j+6} + v_{8,j+6}, \end{aligned} \quad (15c)$$

and  $\{v_{ij}\}$  are given by equation A7 in the appendix. Using the delta method (Goodman & Kruskal, 1972), the asymptotic variance of  $\ln(\zeta)$  is given by

$$\sigma^2(\ln(\zeta)) = \frac{\sum_{i,j=1}^3 v_{ij}^{**} \sum_{i,j=1}^3 v_{ij}^* p_{ij} - \sum_{i,j=1}^3 v_{ij}^* \sum_{i,j=1}^3 v_{ij}^{**} p_{ij}}{n(\zeta \sum_{i,j=1}^3 v_{ij}^{**} p_{ij})^2}, \quad (15d)$$

where  $\{v_{ij}^*\}$  and  $\{v_{ij}^{**}\}$  are given, respectively, by equations 15(b-c). A large sample Wald's  $100(1-\alpha)\%$  confidence interval is given by

$$\left[ \zeta \exp\left(-z_{\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2(\ln(\zeta))}\right), \zeta \exp\left(-z_{\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2(\ln(\zeta))}\right) \right], \quad (15e)$$

where

$$\hat{\sigma}^2(\ln(\zeta)) = \sigma^2(\ln(\zeta)) \Big|_{p_{ij} = \bar{p}_{ij}}.$$

Bias-Adjusted Test for Marginal Homogeneity  
Substituting equation (11) into (5) for  $\hat{p}_{ij}$ , a bias-adjusted test for marginal homogeneity (BATMH) is given by

$$\tilde{S} = \frac{n \cdot \left[ \sum_{j=1}^{K-1} \sum_{i=j+1}^K (\bar{p}_{ji} - \bar{p}_{ij})(z_j - z_i) \right]^2}{\sum_{j=1}^{K-1} \sum_{i=j+1}^K \bar{P}_{ij} (z_j - z_i)^2}, \quad (16)$$

where  $\{\bar{p}_{ij}\}$  are given by equation (11), and  $\bar{P}_i = \bar{p}_{i+} + \bar{p}_{+i}$ .

### Results

Table 1 shows the first and second lambing records of a flock of 227 Merino ewes from 1952-1953 (Tallis, 1962). If the data in Table 1 are not misclassified, then the naïve GOR can be calculated as 1.22 (95% CI: 1.12–1.32) using equations 3 and 4. This implies that a significant association exists between the number of lambing records in 1952 and 1953. Also, the test value of equation 5 is obtained as  $\hat{S} = 70.0$  with  $p < 0.0001$  which indicates that the marginal distribution of the lambing records in 1952 is significantly different from that of 1953.

Table 1: Cross-Classification of Ewes According to Number of Lambs Born in Consecutive Years

Number of Lambs (1953)	Number of Lambs (1952)			Total
	0	1	2	
0	58	52	1	111
1	26	58	3	87
2	8	12	9	29
Total	92	122	13	227

Suppose that errors are present in the classification of the lambing records in Table 1; in that case, the bias-adjusting method would be applied. In order to use the formula of equation 11, the true MPs must be calculated. In order to accomplish this task, it is necessary to know the true cell counts; Through the use of theory of counterfactuals it is intuitively clear that the issue of getting a true table is simply a counterfactual of the observed [misclassified] table which is thought the factual one (Lewis, 1973). Hence, the above idea may be applied to obtain the hypothetically true cell counts by reshuffling the number of misclassified subjects in the observed table.

Because the row/column marginal totals in case-control studies have to be kept as being fixed, four out of nine cells can be chosen as free parameters to construct the true (counterfactual) table. By noting that there are two cells (1,3) and (2,3) with small observed counts, these two cells and two other cells (2,1) and (3,2) are selected as free parameters to construct ten true tables (column 2, Table 2). With 1 in the (1,3) cell to be kept unchanged, the assumed number of under- or over-misclassified subjects starts with the (2,3) cell and then increases one by one up to seven in that cell, while the assumed number of under- or over-misclassified subjects the other two cells (2,1) and (3,2) are chosen

discreetly we ended up with eight true cell count tables (#1 to #8, column 2 of Table 2); True cell counts in #9 and #10 of Table 2 are similarly constructed.

With the true cell counts as given, it is a matter of straightforward calculation to obtain true MPs; the MPs are calculated as the ratio of difference between true and observed marginal totals divided by their sum. These corresponding MPs were calculated (column 3, Table 2): the details are similar to that of Lee (2009a, 2010) and are hence omitted here. In order to check the feasibility of the MPs, three determinants (equations 12-14),  $\det(W_{11})$ ,  $\det(\Delta_1)$  and  $\det(\Delta_2)$ , were calculated. After examining their values, they are all feasible because all the determinant values are positive (columns 2-4, Table 3).

Although all MPs are feasible, it is interesting to note that only five out of ten (#1 to #5) are admissible because (1) they are positive real numbers between 0 and 1, and (2) they satisfy the constraint on the total probability sum:

$$\sum_{i=1}^3 \sum_{j=1}^3 \tilde{p}_{ij} = 1 \text{ (column 5, Table 3). As a result,}$$

BAGORs and BATMHs were calculated for models #1 to #5 (columns 2 and 3, Table 4). number of under- or over-misclassified subjects

Table 2: Ten Assumed True Cell Counts and their Corresponding MPs for Table 1

#	$(n_{11}, n_{12}, n_{13}; n_{21}, n_{22}, n_{23}; n_{31}, n_{32}, n_{33})$	$(a_{x[1]}, b_{x[1]}, d_{x[1]}; a_{x[2]}, b_{x[2]}, d_{x[2]}; a_{x[3]}, b_{x[3]}, d_{x[3]})^*$
1	(57,53,1;27,57,3;8,12,9)	(0.8,3,0;6,2,0;0,0,0)
2	(57,53,1;27,56,4;8,13,8)	(0.8,3,0;6,4,50;0,1,30)
3	(56,54,1;27,55,5;9,13,7)	(9,6,0;6,7,80;30,10,60)
4	(55,55,1;28,53,6;9,14,6)	(10,9,0;10,10,110;30,30,10)
5	(54,56,1;29,51,7;9,15,5)	(20,10,0;20,20,130;30,40,140)
6	(48,62,1;30,49,8;14,11,4)	(50,30,0;20,20,150;140,10,190)
7	(45,65,1;31,47,9;16,10,3)	(60,40,0;30,30,170;170,30,250)
8	(49,61,1;25,52,10;18,9,2)	(40,30,0;7,20,180;190,50,320)
9	(55,54,2;24,60,3;13,8,8)	(10,6,170;10,4,0;120,70,30)
10	(55,54,2;22,61,4;15,7,7)	(10,6,170;30,6,50;150,90,60)

\*All entries inside the parenthesis defined by equations A1 and A2 in the appendix need to multiply by  $10^{-3}$ .

## MATCHED-PAIR STUDIES WITH MISCLASSIFIED ORDINAL DATA

The value of BAGOR/BATMH was not computed if the corresponding BACPs were inadmissible.

Table 4 shows that admissible BACPs the BAGOR ( $\check{\zeta} = 2.08$ ) is biased further away

from the value of the null hypothesis than the classical estimator ( $\hat{\zeta} = 1.22$ ), but the BATMHs ( $\check{S} = 12.0$  is biased toward the null value than the classical estimator ( $\hat{S} = 70.0$ ).

Table 3: Feasibility and Admissibility of MPs and/or BACP in Table 2

#	$\det(W_{11})$	$\det(\Delta_1)$	$\det(\Delta_2)$	$(\check{P}_{11}, \check{P}_{12}, \check{P}_{13}, \check{P}_{21}, \check{P}_{22}, \check{P}_{23}, \check{P}_{31}, \check{P}_{32}, \check{P}_{33})$
1	0.99	0.96	0.96	(0.260, 0.0008, 0.004, 0.11, 0.26, 0.22, 0.03, 0.05, 0.04)
2	0.99	0.82	0.74	(0.26, 0.007, 0.005, 0.11, 0.26, 0.22, 0.03, 0.05, 0.03)
3	0.96	0.69	0.54	(0.27, 0.02, 0.008, 0.11, 0.26, 0.21, 0.03, 0.04, 0.03)
4	0.95	0.58	0.39	(0.27, 0.04, 0.02, 0.11, 0.26, 0.22, 0.02, 0.04, 0.03)
5	0.93	0.49	0.28	(0.27, 0.07, 0.04, 0.11, 0.26, 0.22, 0.02, 0.03, 0.02)
6	0.82	0.38	0.15	(0.32, 0.13, 0.06, 0.10, 0.25, 0.21, 0.02, -0.01, 0.02)
7	0.77	0.32	0.08	(0.34, 0.22, 0.14, 0.09, 0.22, 0.21, 0.01, -0.04, 0.006)
8	0.84	0.36	0.05	(0.37, 0.59, 0.54, 0.10, 0.11, 0.20, 0.008, -0.06, -0.002)
9	0.64	0.60	0.34	(0.32, -0.09, 0.02, 0.10, 0.28, 0.19, 0.01, -0.008, 0.04)
10	0.64	0.48	0.21	(0.34, -0.11, 0.006, 0.10, 0.28, 0.19, -0.003, -0.04, 0.03)

Table 4: Estimated BAGORs with 95% Confidence Interval (CI) and BATMHs with p-value for Table 3

#	$\check{\zeta}$ (95% CI)	$\check{S}$ (p - value)
1	0.74 (0.26 – 2.12)	0.49 (0.31)
2	1.17 (0.66 – 2.08)	2.05 (0.02)
3	1.34 (0.72 – 2.48)	4.05 (< 0.0001)
4	1.60 (0.86 – 3.00)	6.80 (< 0.0001)
5	2.08 (1.06 – 4.07)	12.0 (< 0.0001)
6-10	-*	-*

\*MPs are not admissible; thus values of  $\check{\zeta}/\check{S}$  are not calculated.

### Conclusion

A new method is presented here to study the misclassification problem associated with matched-pair case-control studies for the polytomous exposure variable. Based on results from this study, the following conclusions are put forth:

1. Determining whether there are classification errors in the collected data is a difficult issue. Strictly, this requires the principal investigator using personal expertise to exercise subjective judgment on the collected data. However, from the sensitivity analysis of this data set of lambing records, the method presented herein can vindicate itself empirically. Note that this example indicates that, at most, one record in the (1,3) cell can be under- or over-misclassified. It is impossible to have more than one record misclassified in that cell due to the occurrence of inadmissible MPs.
2. This method does not require non-differential misclassification as an assumption. In fact, differential misclassification is inclined to be the norm rather than exception in practical applications. Indeed, the example provided shows that, even if both the column and the row marginal totals misclassify, just the same number of records to their corresponding MPs are not the same because they have different marginal totals for the column and row respectively.
3. The direction of the bias is not the same for two measures of association - it depends on which measure is used.
4. The close-form formula for this method are derived only for  $K = 3$ . For  $K = 4, 5, 6$  it is workable to obtain the closed-form formula by hand. For much bigger values of  $K$ , it is a formidable task to work out all the details by hand. Fortunately, there is an alternative way to bypass the necessity of getting closed-form formula. Taking a closer look at two criteria for MA: feasibility and admissibility, it is found that feasibility is not essential, but admissibility is critical,

meaning that it is not necessary to pay much attention to feasibility, the main focus is only on admissibility. Hence, instead of getting closed-form formula, equation 10 can be solved numerically for BACP and the admissibility of MP checked by examining whether all components of BACP are positive real numbers between 0 and 1.

5. The confidence interval given by equations 4 or 15(e) is large sample asymptotic. If the sample sizes are small, an exact confidence interval should instead be used (Lui, 2002).

Although the traditional naïve estimator can be viewed as a special case of bias-adjusted estimator when all misclassification probabilities are zero, a huge difference exists between these two estimators. Note that a bias-adjusted generalized odds ratio as shown in equation (11) uses both the concordant and discordant data in the observed table, while the naïve estimator shown in (3) uses only the discordant data. As a result, a bias-adjusted generalized odds ratio will be more efficient than the naïve one.

Finally, a limitation of this study is that the results presented do not apply to a situation in which more than two adjacent columns/rows are misclassified in the contingency table. Clearly, the question remains open regarding how to adjust the naïve estimator for the misclassification bias if the assumption of only two adjacent columns/rows being misclassified is not satisfied.

### Acknowledgements

All numerical calculations obtained in Tables 2-4 are facilitated by using the Microsoft EXCEL spreadsheet. This article is an edited version of Lee (2009b), except that the data used for illustration has been changed from the unaided eyesight data of employees at Royal Ordnance factories in 1943-1946 to the current lambing records of Merino ewes in 1952-1953. Also, I was grateful for the editorial help which greatly improved the presentation of this paper.

### References

Agresti, A. (1980). Generalized odds ratios for ordinal data. *Biometrics*, 36, 59-67.



## MATCHED-PAIR STUDIES WITH MISCLASSIFIED ORDINAL DATA

Agresti, A. (1983). Testing marginal homogeneity for ordinal categorical variables. *Biometrics*, 39, 505-510.

Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: Wiley.

Breslow, N. (1982). Covariance adjustment of relative-risk estimates in matched studies. *Biometrics*, 38, 661-672.

Clayton, D. G. (1974). Some odds ratio statistics for the analysis of ordered categorical data. *Biometrika*, 61, 525-531.

Goodman, L. A., & Kruskal, W. H. (1972). Measures of association for cross classifications IV: Simplification of asymptotic variances. *Journal of the American Statistical Association*, 67, 415-421.

Greenland, S. (1982). The effect of misclassification in matched-pair case-control studies. *American Journal of Epidemiology*, 116, 402-406.

Greenland, S. (1989). On correcting misclassification in twin studies and other matched pair studies. *Statistics in Medicine*, 8, 825-829.

Greenland, S. and Kleinbaum, D. G. (1983). Correcting for misclassification in two-way tables and matched-pair studies. *International Journal of Epidemiology*, 12, 93-97.

Lee, T-S. (2009a). Bias-adjusted exposure odds ratio for misclassified data. *The Internet Journal of Epidemiology*, 6(2). Accessed from <http://www.ispub.com/journal/the-internet-journal-of-epidemiology/volume-6-number-2/bias-adjusted-exposure-odds-ratio-for-misclassification-data-1.html>.

Lee, T-S. (2009b). Misclassified matched studies with ordinal data. Proceedings of the 6<sup>th</sup> Sino-International Symposium on Probability, Statistics, and Quantitative Management, pp. 48-67, Taipei, Taiwan.

Lee, T-S. (2010). Bias-adjusted McNemar's test for misclassified data. *Journal of Probability & Statistical Science*, 8, 81-95.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Lui, K.-J. (2002). Notes on estimation of the general odds ratio and the general risk difference for paired-sample data. *Biometrical Journal*, 44, 957-968.

McCullagh, P. (1977). A logistic model for paired comparisons with ordered categorical data. *Biometrika*, 64, 449-453.

Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables. *Biometrika*, 40, 105-110.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42, 412-416.

Tallis, G. M. (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics*, 18, 342-353.

### Appendix

For  $j = 1, 2, 3$ , let

$$\begin{aligned} a_{Z[j]} &= \varphi_{z[1;j]}, b_{Z[j]} = \varphi_{z[2;j]}, c_{\bar{Z}[j]} = \\ \psi_{Z[2;j]}, d_{Z[j]} &= \psi_{z[3;j]}, \end{aligned} \quad (A1)$$

Where  $\bar{Z} = Y$  if  $Z = X$ , and vice versa.

Because of the symmetry in matched-pair studies, it is reasonable to assume that

$$\begin{aligned} a_{Z[1]} &= a_{\bar{Z}[1]}, a_{Z[2]} = b_{Z[1]} = c_{Z[1]}, b_{Z[2]} = \\ b_{\bar{Z}[2]} &= c_{Z[2]} = c_{\bar{Z}[2]}, b_{Z[3]} = c_{Z[3]} = \\ d_{\bar{Z}[2]}, a_{Z[3]} &= d_{\bar{Z}[1]}, d_{Z[3]} = d_{\bar{Z}[3]}. \end{aligned} \quad (A2)$$

For  $K = 3$ , the matrix  $W$  in equation 9 was given by

$$W_{9 \times 9} = \begin{bmatrix} W_{11} & W_{12} & 0 \\ W_{21} & W_{22} & W_{23} \\ 0 & W_{32} & W_{33} \end{bmatrix}, \quad (A3)$$

where

$$W_{11} = \begin{bmatrix} 1 - 2a_{X[1]} & b_{X[1]} & 0 \\ a_{X[1]} & 1 - 3b_{X[1]} & d_{X[1]} \\ 0 & b_{X[1]} & 1 - 2d_{X[1]} \end{bmatrix},$$

$$W_{12} = \text{diag}[c_{Y[1]}, c_{Y[2]}, c_{Y[3]}], W_{13} = W_{31} = [0], W_{21} = \text{diag}[a_{X[1]}, b_{X[1]}, d_{X[1]}];$$

$$W_{22} = \begin{bmatrix} 1 - 3a_{X[2]} & b_{X[2]} & 0 \\ a_{X[2]} & 1 - 4b_{X[2]} & d_{X[2]} \\ 0 & b_{X[2]} & 1 - 3d_{X[2]} \end{bmatrix},$$

$$\begin{aligned}
 W_{23} &= \\
 \text{diag}[d_{Y[1]}, d_{Y[2]}, d_{Y[3]} &= \\
 \text{diag}[a_{X[3]}, b_{X[3]}, d_{X[3]}], W_{32} &= \\
 \text{diag}[b_{Y[1]}, b_{Y[2]}, b_{Y[3]}] &= \\
 \text{diag}[a_{X[2]}, b_{X[2]}, d_{X[2]}]; & \\
 W_{33} &= \begin{bmatrix} 1 - 2a_{X[3]} & b_{X[3]} & 0 \\ a_{X[3]} & 1 - 3b_{X[3]} & d_{X[3]} \\ 0 & b_{X[3]} & 1 - 2d_{X[3]} \end{bmatrix},
 \end{aligned}$$

where  $\text{diag}[d_{11}, d_{22}, d_{33}]$  denotes a  $3 \times 3$  diagonal matrix.

Solving the matrix equation of  $W_{9 \times 9} \cdot V_{9 \times 9} = I_{9 \times 9}$ , where  $V_{9 \times 9}$  was given by equation 10(b) and  $I_{9 \times 9}$  was a  $9 \times 9$  identity matrix, results in

$$\begin{aligned}
 V_{31} &= \Delta_2^{-1} W_{32} W_{21}, \\
 V_{21} &= -\Delta_1^{-1} (W_{11} W_{23} V_{31} + W_{21}), \\
 V_{11} &= W_{11}^{-1} (I_{3 \times 3} - W_{12} V_{21}), \\
 V_{32} &= -\Delta_2^{-1} W_{32} W_{11}, \\
 V_{22} &= \Delta_1^{-1} W_{11} (I_{3 \times 3} - W_{23} V_{32}), \quad (\text{A4}) \\
 V_{12} &= -W_{11}^{-1} W_{12} V_{22}, \\
 V_{33} &= \Delta_2^{-1} \Delta_1, \\
 V_{23} &= -\Delta_1^{-1} W_{11} W_{23} V_{33}, \\
 V_{13} &= -W_{11}^{-1} W_{12} V_{23}.
 \end{aligned}$$

where  $\Delta_1$  and  $\Delta_2$  were defined, respectively, by

$$\Delta_1 = W_{11} W_{22} - W_{21} W_{12},$$

and

$$\Delta_2 = \Delta_1 W_{33} - W_{32} W_{11} W_{23}.$$

If  $\{\delta_{ij}^{[1]}\}$  and  $\{\delta_{ij}^{[2]}\}$  denote the  $(i, j)^{\text{th}}$  entry of  $\Delta_1$  and  $\Delta_2$ , then

$$\delta_{11}^{[1]} = 1 - 2a_{X[1]} - 3a_{X[2]} + 5a_{X[1]}a_{X[2]} + a_{X[2]}b_{X[1]},$$

$$\delta_{12}^{[1]} = b_{X[2]}(1 - 2a_{X[1]}) + b_{X[1]}(1 - 4b_{X[2]}),$$

$$\delta_{13}^{[1]} = b_{X[1]}d_{X[2]},$$

$$\delta_{21}^{[1]} = a_{X[1]}(1 - 3a_{X[2]}) + a_{X[2]}(1 - 3b_{X[1]}),$$

$$\delta_{22}^{[1]} = a_{X[1]}b_{X[2]} + 1 - 3b_{X[1]} - 4b_{X[2]} + 11b_{X[1]}b_{X[2]} + d_{X[1]}b_{X[2]}, \quad (\text{A5})$$

$$\delta_{23}^{[1]} = b_{X[1]}(1 - 4b_{X[2]}) + b_{X[2]}(1 - 2d_{X[1]});$$

$$\delta_{31}^{[1]} = b_{X[1]}a_{X[2]},$$

$$\delta_{32}^{[1]} = b_{X[1]}(1 - 4b_{X[2]}) + b_{X[2]}(1 - 2d_{X[1]}),$$

$$\delta_{33}^{[1]} = b_{X[1]}d_{X[2]} + 1 - 2d_{X[1]} - 3d_{X[2]} + 5d_{X[1]}d_{X[2]};$$

$$\delta_{11}^{[2]} = (1 - 2a_{X[3]})\delta_{11}^{[1]} + a_{X[3]}\delta_{12}^{[1]} - a_{X[2]}a_{X[3]}(1 - 2a_{X[1]}),$$

$$\delta_{12}^{[2]} = b_{X[3]}\delta_{11}^{[1]} + (1 - 3b_{X[3]})\delta_{12}^{[1]} + b_{X[3]}\delta_{13}^{[1]} - b_{X[1]},$$

$$\delta_{13}^{[2]} = d_{X[3]}\delta_{12}^{[1]} + (1 - 2d_{X[3]})\delta_{13}^{[1]},$$

$$\delta_{21}^{[2]} = (1 - 2a_{X[3]})\delta_{21}^{[1]} + a_{X[3]}\delta_{22}^{[1]} - a_{X[1]},$$

$$\delta_{22}^{[2]} = b_{X[3]}\delta_{21}^{[1]} + (1 - 3b_{X[3]})\delta_{22}^{[1]} + b_{X[3]}\delta_{23}^{[1]} - b_{X[2]}b_{X[3]}(1 - 3b_{X[1]}), \quad (\text{A.6})$$

$$\delta_{23}^{[2]} = d_{X[3]}\delta_{22}^{[1]} + (1 - 2d_{X[3]})\delta_{23}^{[1]} - d_{X[1]};$$

$$\delta_{31}^{[2]} = (1 - 2a_{X[3]})\delta_{31}^{[1]} + a_{X[3]}\delta_{32}^{[1]},$$

$$\delta_{32}^{[2]} = b_{X[3]}\delta_{31}^{[1]} + (1 - 3b_{X[3]})\delta_{32}^{[1]} + b_{X[3]}\delta_{33}^{[1]} - b_{X[1]},$$

$$\delta_{33}^{[2]} = d_{X[3]}\delta_{32}^{[1]} + (1 - 2d_{X[3]})\delta_{33}^{[1]} - d_{X[2]}d_{X[3]}(1 - 2d_{X[1]}).$$

## MATCHED-PAIR STUDIES WITH MISCLASSIFIED ORDINAL DATA

The nine equations in A4 were solved by grouping them into three sets. First, the first three equations were solved together in A4 letting the entries for  $V_{31}$ ,  $V_{21}$ , and  $V_{11}$  be denoted, respectively, by  $\{a_{ij}\}$ ,  $\{s_{ij}\}$ , and  $\{x_{ij}\}$ , namely,  $V_{31} = [a_{ij}]$ ,  $V_{21} = [s_{ij}]$ , and  $V_{11} = [x_{ij}]$ . Next the second set of three equations in A4 were solved for  $V_{32}$ ,  $V_{22}$ , and  $V_{12}$ ; the entries of these matrices are given, respectively, by  $V_{32} = [b_{ij}]$ ,  $V_{22} = [t_{ij}]$  and  $V_{12} = [y_{ij}]$ . Finally, after solving equations A4 for  $V_{33}$ ,  $V_{23}$ , and  $V_{13}$ , the entries of these matrices are given, respectively, by  $V_{33} = [c_{ij}]$ ,  $V_{23} = [u_{ij}]$  and  $V_{13} = [z_{ij}]$ .

Putting together the above result, the inverse of the misclassification matrix  $W$  was thus obtained as follows:

$$V_{9 \times 9} = [v_{ij}] = \begin{bmatrix} x_{11} & x_{12} & x_{13} & y_{11} & y_{12} & y_{13} & z_{11} & z_{12} & z_{13} \\ x_{21} & x_{22} & x_{23} & y_{21} & y_{22} & y_{23} & z_{21} & z_{22} & z_{23} \\ x_{31} & x_{32} & x_{33} & y_{31} & y_{32} & y_{33} & z_{31} & z_{32} & z_{33} \\ s_{11} & s_{12} & s_{13} & t_{11} & t_{12} & t_{13} & u_{11} & u_{12} & u_{13} \\ s_{21} & s_{22} & s_{23} & t_{21} & t_{22} & t_{23} & u_{21} & u_{22} & u_{23} \\ s_{31} & s_{32} & s_{33} & t_{31} & t_{32} & t_{33} & u_{31} & u_{32} & u_{33} \\ a_{11} & a_{12} & a_{13} & b_{11} & b_{12} & b_{13} & c_{11} & c_{12} & c_{13} \\ a_{21} & a_{22} & a_{23} & b_{21} & b_{22} & b_{23} & c_{21} & c_{22} & c_{23} \\ a_{31} & a_{32} & a_{33} & b_{31} & b_{32} & b_{33} & c_{31} & c_{32} & c_{33} \end{bmatrix}.$$

where the closed-form solutions for all the entries  $\{v_{ij}\}$  can be found in the appendix of Lee (2009b).