Journal of Data Science **9**(2011), 487-500

Tests of Independence with Incomplete Contingency Tables Using Likelihood Functions

Shin-Soo Kang[1] and Michael D. Larsen[2]*
[1]*KwanDong University and* [2]*George Washington University*

Abstract: Kang (2006) used the log-likelihood function with Lagrangian multipliers for estimation of cell probabilities in two-way incomplete contingency tables. The constraints on cell probabilities can be incorporated through Lagrangian multipliers for the likelihood function. The method can be readily extended to multidimensional tables. Variances of the MLEs are derived from the matrix of second derivatives of the log likelihood with respect to cell probabilities and the Lagrange multiplier. Wald and likelihood ratio tests of independence are derived using the estimates and estimated variances. Simulation results, when data are missing at random, reveal that maximum likelihood estimation (MLE) produces more efficient estimates of population proportions than either multiple imputation (MI) based on data augmentation or complete case (CC) analysis. Neither MLE nor MI, however, leads to an improvement over CC analysis with respect to power of tests for independence in 2×2 tables. Thus, the partially classified marginal information increases precision about proportions, but is not helpful for judging independence.

Key words: Lagrangian multiplier, likelihood ratio test, missing at random, missing value analysis, Wald statistic.

1. Introduction

It may happen that some observations are not fully cross-classified when forming contingency tables from two or more categorical variables. Complete-case (CC) analysis discards cases with missing data, thereby restricting analysis to only counts with fully observed variables used for cross classification into a contingency table.

An alternative approach involves constructing a complete table, in which all cases are completely classified, by imputing information for the missing classification dimensions. Multiple imputation, proposed by Rubin (1978; see also

*Corresponding author.

Rubin 1987, 1996), provides a way to take advantage of commonly used tests of independence for completely classified tables. Li *et al.* (1991) proposed a Wald test statistic and Meng and Rubin (1992) proposed likelihood ratio tests with F reference distributions. In addition to such tests of independence, one can estimate joint probabilities and their standard errors using sets of, say, five imputed contingency tables obtained from data augmentation. Finch (2010) has studied some methods of imputation for categorical data. In this paper, we consider multiple imputation of categorical values using a Jeffrey's prior distribution on the unknown cell probabilities.

Instead of imputing information for missing classifications, maximum likelihood estimates of population proportions can be obtained from the observed information, including both the completely and partially classified cases. Little (1982) used a simple EM algorithm to estimate cell probabilities. Lipsitz *et al.* (1998) show how to use generalized linear model software to evaluate maximum likelihood estimates of cell probabilities using the connection between the multinomial and Poisson likelihoods. Molenberghs and Goetghebeur (1997) present estimation methods using the observe data log likelihood directly. In their approach, they use Fisher scoring, Newton-Raphson, and other algorithms. They cite McCullagh and Nelder (1989) for their algorithms, and this source does not utilize Lagrangian multipliers. Instead, they incorporate the constraints on probabilities directly into the likelihood function.

Maximum likelihood estimates of population proportions can be obtained from the partial log-likelihood function proposed by Chen and Fienberg (1974) for the cell probabilities of two way incomplete contingency tables. Kang (2006) proposed the partial log-likelihood function with a Lagrangian multiplier incorporating the constraint on cell probabilities.

Both the MLE method and the MI method should be appropriate when data are missing completely at random (MCAR) or missing at random (MAR) in the sense of Rubin (1976). The complete case method can cause bias for estimating cell probabilities when data are in fact MAR.

Less computation is required to get the matrix of second derivatives of the log-likelihood function with respect to cell probabilities when a Lagrangian multiplier is used than when it is not used. In this paper, variances of MLE estimators of population proportions are derived from the matrix of second derivatives. Wald test of independence are derived using the variances of MLEs. Likelihood ratio tests of independence are derived based on the likelihood function evaluated at the MLEs.

Section 2 presents notation and the likelihood function. Section 3 reviews MLEs and their variances proposed by Kang (2006). Section 4 derives tests of independence for incomplete contingency tables. The performance of tests of independence provided by the complete-case analysis, multiple imputation, and maximum likelihood approaches are examined through Monte Carlo simulation studies in Section 5 considering both type I error level and power. Section 6 contains a summary.

2. Notation and Likelihood Function

Consider an $I \times J$ contingency table where the row factor X_1 has I categories and the column factor X_2 has J categories. Assume simple random sampling with replacement. In a complete table, where the row and column categories are observed for every case in the sample, the counts have a multinomial distribution with sample size N and probability vector θ. Let θ_{ij}, an element of θ, denote the population proportion for the (i, j) cell.

When information on either the row or column classification is missing, we can construct a table of counts for the completely classified cases where x_{ij} denotes the number of cases observed in the (i, j) cell. We can also construct one-way tables of counts for partially classified cases. Let $x_i m$ denote the number of cases in the i^{th} row category, $i = 1, 2, \cdots, I$, where the column category is unknown, and let x_{mj} denote the number of cases in the j^{th} column category, $j = 1, 2, \cdots, J$, where the row category is unknown. Then, x_{im} and x_{mj} are marginally observed counts on a single variable. Let x_{mm} denote the number of cases where both the row and column categories are missing. The total sample size is

$$\begin{aligned} N &= \sum_{ij} x_{ij} + \sum_{i} x_{im} + \sum_{j} x_{mj} + x_{mm} \\ &= x_{cc} + x_{\bullet m} + x_{m\bullet} + x_{mm}. \end{aligned}$$

Discarding the x_{mm} cases for which both variables are missing does not affect any results in this paper except that it necessitates changing N to $n = N - x_{mm}$. Those cases do not contain any information about the joint distribution or marginal distributions of X_1 and X_2.

The log-likelihood function for the cell probabilities θ presented by Chen and Fienberg (1974) is the following;

$$l(\theta) = \sum_{i} \sum_{j} x_{ij} \log \theta_{ij} + \sum_{j} x_{mj} \log \theta_{\bullet j} + \sum_{i} x_i m \log \theta_{i\bullet}. \tag{2.1}$$

Note that the log-likelihood function in (2.1) does not include x_{mm}.

In the general case, there are r variables, $X_a, X_b, \cdots, X_r$ with levels $A, B, \cdots,$ R, respectively. A cell in the cross-classified table is identified by a r-tuple $(a, b, \cdots, r)$. Let $\theta_{a,b,\cdots,r}$ and $x_{a,b,\cdots,r}$ be the proportion and count, respectively, in cell $(a, b, \cdots, r)$. The complete-data log likelihood for a multinomial model is

$$l(\theta) = \sum_{a=1}^{A} \sum_{b=1}^{B} \cdots \sum_{r=1}^{R} x_{a,b,\cdots,r} \log \theta_{a,b,\cdots,r}.$$

When some variables are missing, partially classified counts appear in the observed-data log likelihood multiplying the log of aggregated probabilities for the corresponding cells, as in equation 2.1. In the case that $r = 3$, the log likelihood of this sort with the largest number of terms, ignoring the x_{mmm} cases with all three variables missing, is

$$\begin{aligned} l(\theta) = & \sum_a \sum_b \sum_c x_{abc} \log \theta_{abc} + \sum_a \sum_b x_{abm} \log \theta_{ab\bullet} \\ & + \sum_a \sum_c x_{amc} \log \theta_{a\bullet c} + \sum_b \sum_c x_{mbc} \log \theta_{\bullet bc} \\ & + \sum_a x_{amm} \log \theta_{a\bullet\bullet} + \sum_b x_{mbm} \log \theta_{\bullet b\bullet} + \sum_c x_{mmc} \log \theta_{\bullet\bullet c}. \qquad (2.2) \end{aligned}$$

3. Maximum Likelihood Estimation of Cell Probability and Variances

The EM algorithm (Dempster, Laird and Rubin, 1977) can be used to get maximum likelihood estimates (MLEs) of proportions in an incomplete contingency table. In the case of an $I \times J$ table, let $\theta_{ij}^{(0)}$ be an initial estimate of θ_{ij}, such as x_{ij}/x_{cc}. The estimate of θ_{ij} at the t^{th} iteration of the algorithm is

$$\theta_{ij}^{(t)} = 1/n \left(x_{ij} + x_{im} \times \frac{\theta_{ij}^{(t-1)}}{\theta_{i\bullet}^{(t-1)}} + x_{mj} \times \frac{\theta_{ij}^{(t-1)}}{\theta_{\bullet j}^{(t-1)}} \right). \qquad (3.1)$$

The algorithm converges to the MLEs of θ. Little (1982) presented examples of the algorithm for 2×2 tables. In the general case of multidimensional tables, the EM algorithm was described by Fuchs (1982). See also Little and Rubin (2002) and Schafer (1997).

The methodology to produce variances of MLEs (Kang, 2006) is reviewed in this section. Since the proportions are constrained to sum to one ($\sum_{ij} \theta_{ij} = 1$), the likelihood function (2.1) incorporating the constraint can be expressed with a Lagrangian multiplier as

$$l(\theta^*) = \sum_i \sum_j x_{ij} \log \theta_{ij} + \sum_j x_{mj} \log \theta_{\bullet j} + \sum_i x_{im} \log \theta_{i\bullet} + \gamma(1 - \sum_{ij} \theta_{ij}), \quad (3.2)$$

where $\theta^* = (\theta', \gamma)'$.

The first derivative of $l(\theta^*)$ in (3.2) with respective to θ_{ij} and to γ are

$$\begin{aligned}\frac{\partial l(\theta^*)}{\partial\theta_{ij}} &= \frac{x_{ij}}{\theta_{ij}} + \frac{x_{mj}}{\theta_{\bullet j}} + \frac{x_{im}}{\theta_{i\bullet}} - \gamma,\\ \frac{\partial l(\theta^*)}{\partial\gamma} &= 1 - \sum_{ij}\theta_{ij}.\end{aligned}$$

The second partial derivatives are

$$\begin{aligned}&\frac{\partial^2 l(\theta^*)}{\partial\theta_{ij}^2} = \frac{-x_{ij}}{\theta_{ij}^2} - \frac{x_{mj}}{\theta_{\bullet j}^2} - \frac{x_{im}}{\theta_{i\bullet}^2},\\ &\frac{\partial^2 l(\theta^*)}{\partial\theta_{ij}\partial\theta_{is}} = -\frac{x_{im}}{\theta_{i\bullet}^2}, \quad \frac{\partial^2 l(\theta^*)}{\partial\theta_{ij}\partial\theta_{kj}} = -\frac{x_{mj}}{\theta_{\bullet j}^2},\\ &\frac{\partial^2 l(\theta^*)}{\partial\theta_{ij}\partial\theta_{ks}} = 0, \quad \frac{\partial^2 l(\theta^*)}{\partial\theta_{ij}\partial\gamma} = -1, \quad \frac{\partial^2 l(\theta^*)}{\partial\gamma^2} = 0.\end{aligned}$$

Let $I(\theta^*)$ be the matrix of second derivatives of the log likelihood with respect to θ^*. An estimate of the covariance matrix of $\hat\theta^*$, the MLE of θ^*, is $-I^{-1}(\hat\theta^*)$. Let $\hat\Sigma_M$ be a matrix omitting the last row and column of $-I^{-1}(\hat\theta^*)$. The matrix $\hat\Sigma_M$ gives an estimate of the covariance matrix of $\hat\theta_M$, where $\hat\theta_M$ is the MLE of θ. For a 2×2 table, $\theta^* = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}, \gamma)'$ and $I(\theta^*)$ is

$$I(\theta^*) = -\begin{pmatrix} d_{11} & \frac{x_{1m}}{\theta_{1\bullet}^2} & \frac{x_{m1}}{\theta_{\bullet 1}^2} & 0 & 1\\ \frac{x_{1m}}{\theta_{1\bullet}^2} & d_{12} & 0 & \frac{x_{m2}}{\theta_{\bullet 2}^2} & 1\\ \frac{x_{m1}}{\theta_{\bullet 1}^2} & 0 & d_{21} & \frac{x_{2m}}{\theta_{2\bullet}^2} & 1\\ 0 & \frac{x_{m2}}{\theta_{\bullet 2}^2} & \frac{x_{2m}}{\theta_{2\bullet}^2} & d_{22} & 1\\ 1 & 1 & 1 & 1 & 0\end{pmatrix},$$

where $d_{ij} = \frac{x_{ij}}{\theta_{ij}^2} + \frac{x_{mj}}{\theta_{\bullet j}^2} + \frac{x_{im}}{\theta_{i\bullet}^2}$.

The likelihood function with a Lagrangian multiplier and corresponding derivatives for a three-way table are given in the Appendix.

4. Tests of Independence

A Wald test and a likelihood ratio test for independence in an incomplete two-way contingency table are described in this section. Extensions for three-way tables also are described.

4.1 A Wald Test Using MLE in an Incomplete Contingency Table

For a complete two-dimensional contingency table with sample size N, the null hypothesis of statistical independence is

$$H_0 : \theta_{ij} = \theta_{i\bullet}\theta_{\bullet j}, \quad \text{for all } i \text{ and } j. \tag{4.1}$$

Defining

$$g_{ab}(\theta) \equiv \left(\sum_{j=1}^{J} \theta_{aj}\right)\left(\sum_{i=1}^{I} \theta_{ib}\right) - \theta_{ab},$$

the null hypothesis of statistical independence can be expressed as

$$H_0 : g_{ab}(\theta) = 0, \tag{4.2}$$

for all $a = 1, 2, \cdots, I$ and $b = 1, 2, \cdots, J$. Let

$$g(\theta) = (g_{11}(\theta), \cdots, g_{1J}(\theta), g_{21}(\theta), \cdots, g_{2J}(\theta), \cdots, g_{IJ}(\theta))'.$$

Then (4.2) can be expressed as $H_0 : g(\theta) = \underset{\sim}{0}$.

For a complete three-way table with sample size N, the null hypothesis is

$$H_0 : \theta_{abc} = \theta_{a\bullet\bullet}\theta_{\bullet b\bullet}\theta_{\bullet\bullet c}, \quad \text{for all } a, b, \text{ and } c. \tag{4.3}$$

Defining

$$g_{abc}(\theta) \equiv \left(\sum_{j,k} \theta_{ajk}\right)\left(\sum_{i,k} \theta_{ibk}\right)\left(\sum_{i,j} \theta_{ijc}\right) - \theta_{abc},$$

the null hypothesis can be expressed as

$$H_0 : g_{abc} = 0 \quad \text{for all } a, b, \text{ and } c,$$

or $H_0 : g(\theta) = \underset{\sim}{0}$, where $g(\theta) = (g_{abc}(\theta), a = 1, \cdots, A, b = 1, \cdots, B, c = 1, \cdots, C)$.

The estimator, $\hat{\theta}$, has an approximate multivariate normal distribution with variance $\Sigma = V_\theta/N$ by the Central Limit Theorem, where $V_\theta = (\Delta_\theta - \theta\theta')$ and Δ_θ is a diagonal matrix with the elements of θ on the main diagonal. Under H_0, for a two-dimensional table, $g(\hat{\theta})$ has an approximate $p = I \times J$ dimensional normal distribution with variance $G\Sigma G'$, where $G_{p\times p}$ is the matrix of first partial derivatives of $g(\theta)$ with respect to θ_{ij}. The elements of $G_{p\times p}$ are

$$\frac{\partial g_{ab}(\theta)}{\partial \theta_{ij}} = \begin{cases} \sum_{i=1}^{I} \theta_{ib} + \sum_{j=1}^{J} \theta_{aj} - 1, & \text{for } a = i, \text{ and } b = j, \\ \sum_{i=1}^{I} \theta_{ib}, & \text{for } a = i, \text{ and } b \neq j, \\ \sum_{j=1}^{J} \theta_{aj}, & \text{for } a \neq i, \text{ and } b = j, \\ 0, & \text{for } a \neq i, \text{ and } b \neq j. \end{cases}$$

A Wald statistic for testing H_0 is

$$\hat{Q} = g(\hat{\theta})'\hat{T}^{-}g(\hat{\theta}), \tag{4.4}$$

where $\hat{T} = (\hat{G}\hat{\Sigma}\hat{G}')$ is obtained by substituting $\hat{\theta}$ for θ and $^{-}$ denotes generalized matrix inverse. For a complete table, $\hat{Q}$ has a distribution converging to a central chi-squared distribution with $df = k = (I-1)(J-1)$ when H_0 is true.

Results for a three-dimensional table are analogous. Under H_0, $g(\hat{\theta})$ has dimension $p' = A \times B \times C$. The elements of $G_{p' \times p'}$ depend on the overlap in dimensions between g_{abc} and θ_{ijk}. Examples are given below with overlap in three, two, one, or zero dimensions:

$$\begin{aligned}
\frac{\partial g_{abc}(\theta)}{\partial \theta_{abc}} &= \left(\sum_{j,k} \theta_{ajk}\right)\left(\sum_{i,k} \theta_{ibk}\right) + \left(\sum_{j,k} \theta_{ajk}\right)\left(\sum_{i,j} \theta_{ijc}\right) \\
&\quad + \left(\sum_{j,k} \theta_{ajk}\right)\left(\sum_{i,k} \theta_{ibk}\right) - 1, \\
\frac{\partial g_{abc}(\theta)}{\partial \theta_{abt}} &= \left(\sum_{j,k} \theta_{ajk}\right)\left(\sum_{i,j} \theta_{ijc}\right) + \left(\sum_{i,k} \theta_{ibk}\right)\left(\sum_{i,j} \theta_{ijc}\right), \\
\frac{\partial g_{abc}(\theta)}{\partial \theta_{ast}} &= \left(\sum_{i,k} \theta_{ibk}\right)\left(\sum_{i,j} \theta_{ijc}\right), \\
\frac{\partial g_{abc}(\theta)}{\partial \theta_{rst}} &= 0,
\end{aligned}$$

where $r \neq a$, $s \neq b$, and $t \neq c$. The degrees of freedom are $df = k' = (A-1)(B-1)(C-1)$.

For an incomplete contingency table, $g(\hat{\theta})$ and $\hat{G}$ are obtained by substituting $\hat{\theta}_M$ for $\hat{\theta}$. Then $\hat{T} = (\hat{G}\hat{\Sigma}\hat{G}')$ is obtained by substituting $\hat{\Sigma}_M$ for $\hat{\Sigma}$ in (4.4). Thus a Wald statistic for testing H_0 is

$$\hat{Q}_M = g(\hat{\theta}_M)'\hat{T}_M^{-}g(\hat{\theta}_M), \tag{4.5}$$

where $\hat{T}_M = (\hat{G}\hat{\Sigma}_M\hat{G}')$ and $\hat{G}$ is obtained by substituting $\hat{\theta}_M$ for θ. Then for a incomplete two-way table, $\hat{Q}_M$ has an approximate central chi-square distribution with $df = k$ when H_0 is true. For an incomplete three-way table the analogous statistic has $df = k'$ when H_0 is true.

4.2 Likelihood Ratio Test

Under the model of independence in (4.1), the MLE of θ_{ij} is

$$\hat{\theta}^0_{ij} = \left(\frac{x_{i\bullet} + x_{im}}{x_{cc} + x_{\bullet m}}\right)\left(\frac{x_{\bullet j} + x_{mj}}{x_{cc} + x_{m\bullet}}\right). \tag{4.6}$$

Let L_0 be the maximized value of the log-likelihood function under the null hypothesis of (4.1) and L_1 be the maximized value under $H_0 \bigcup H_a$. L_0 can be calculated directly by substituting $\hat{\theta}^0_{ij}$ for θ_{ij} in (3.2) and L_1 can be obtained by using $\hat{\theta}_M$. Then $2L_1 - 2L_0$ has a limiting null chi-squared distribution as n goes to infinity when H_0 is true. A size α test is implemented by rejecting H_0 if $2L_1 - 2L_0 > \chi^2_{k,\alpha}$.

For a three-way table, under independence, the MLE of θ_{abc} is

$$\hat{\theta}^0_{abc} = \left(\frac{x_{a\bullet\bullet} + x_{a\bullet m} + x_{am\bullet} + x_{amm}}{x_{\bullet\bullet\bullet} + x_{\bullet\bullet m} + x_{\bullet m\bullet} + x_{\bullet mm}}\right)\left(\frac{x_{\bullet b\bullet} + x_{\bullet bm} + x_{mb\bullet} + x_{mbm}}{x_{\bullet\bullet\bullet} + x_{\bullet\bullet m} + x_{m\bullet\bullet} + x_{m\bullet m}}\right)$$
$$\cdot \left(\frac{x_{\bullet\bullet c} + x_{\bullet mc} + x_{m\bullet c} + x_{mmc}}{x_{\bullet\bullet\bullet} + x_{m\bullet\bullet} + x_{\bullet m\bullet} + x_{mm\bullet}}\right).$$

A large-sample size α test is implemented by rejecting H_0 if $2L_1 - 2L_0 > \chi^2_{k',\alpha}$, where $k' = (A-1)(B-1)(C-1)$ and L_0 and L_1 are defined analogously to how they were defined for two-way tables.

5. Simulation Comparing Methods

The performance of tests of independence on incomplete two-way tables using maximum likelihood estimation (MLE), multiple imputation (MI), and complete case analysis (CC) are compared through Monte Carlo simulations. Two missing data mechanisms corresponding to the missing at random (MAR) assumption are used in simulations. Type I error levels are estimated from 1,000 tables simulated tables under the independence assumption. Power levels are examined by simulating 1,000 tables under an alternative to independence.

For multiple imputation, the Wald statistic proposed by Li, Raghunathan, and Rubin (1991) and the likelihood ratio test statistic proposed by Meng and Rubin (1992) based on five imputed data sets were applied to test independence. The algorithms for MI were programmed through S-PLUS 6.1 (2001) functions for missing values.

5.1 Type I Error Levels

The 2×2 incomplete contingency tables generated for this study to check type I error level were generated with equal cell probabilities and data missing at random (MAR). X_1 and X_2 were independently generated as Bernouli(0.5) random variables with sample size 500.

There are two cases with different missing at random mechanisms; 1,000 tables were generated for each case. The missing mechanism for each case is as follows:

$$\begin{aligned} Pr(X_1 \text{ is missing}|X_2 = 1) &= m_1, \\ Pr(X_1 \text{ is missing}|X_2 = 0) &= m_2, \\ Pr(X_2 \text{ is missing}) &= m_3. \end{aligned} \tag{5.1}$$

For the first case $m_1 = 0.1$, $m_2 = 0.3$, $m_3 = 0.2$ in (5.1). In the second case $m_1 = 0.2$, $m_2 = 0.4$, $m_3 = 0.3$. The percentages of cases with missing information on at least one variable are expected to be 36% and 51% for case 1 and 2 respectively.

Table 1 shows the numbers of tables for which the independence null hypothesis was falsely rejected out of 1000 tables for three nominal Type I error levels. The results using MLE and CC seem to have appropriate Type I error levels on both tests, but Type I error levels tend to be inflated for the MI method.

Table 1: Comparison of Type I error levels

				Wald test					
	MLE			**MI**			**CC**		
Case \ α	1%	5%	10%	1%	5%	10%	1%	5%	10%
1	10	53	91	10	45	96	9	51	95
2	9	58	109	16	68	110	8	56	109
				Likelihood ratio test					
1	9	52	91	10	44	96	9	51	96
2	7	57	109	16	66	108	8	57	109

5.2 Power Study

An alternative to independence for 2 × 2 tables with equal probability margins was used to compare the power of the various procedures. The generated multinomial variables have the cell probabilities

$$(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}) = (0.2, 0.3, 0.3, 0.2), \tag{5.2}$$

with sample size 500. The missing data mechanisms (5.1) are same as for the two cases used previously.

Before we study power of independence test, let's compare three methods with checking point estimations of θ_{11} and $\theta_{1+}\theta_{+1} - \theta_{11}$. Table 2 shows means and standard deviations of 1,000 values for the estimates of θ_{11} from the generated

1,000 tables in this subsection. The true value of θ_{11} is 0.2. MLE and MI methods provide essentially unbiased estimates for the cell probabilities but CC does not. The standard deviations of the estimates differ across methods. Complete-case analysis provides the estimate of θ_{11} with the largest variance. For all methods, variation increases as the proportion of missing values increases. MLE tends to provide smaller standard deviations of cell proportion than MI.

Table 2: Estimation of θ_{11}

	MLE		**MI**		**CC**	
Case	Mean	S.D.	Mean	S.D.	Mean	S.D.
1	0.1984	0.02063	0.1988	0.02118	0.2242	0.02377
2	0.1978	0.02193	0.1977	0.02281	0.2281	0.02697

Table 3 shows means and standard deviations of 1,000 simulated values for the estimates of $\theta_{1+}\theta_{+1}-\theta_{11}$, a measure of association between the two variables. The true value of $\theta_{1+}\theta_{+1}-\theta_{11}$ is 0.05. The averages of the estimates are similar for all methods. The complete-case and MLE have similar standard deviations and they exhibit smaller standard deviations than MI. Results on point estimation are helpful in interpreting power simulation results.

Table 3: Estimation of $\theta_{1+}\theta_{+1} - \theta_{11}$

	MLE		**MI**		**CC**	
Case	Mean	S.D.	Mean	S.D.	Mean	S.D.
1	0.0495	0.01376	0.0493	0.01422	0.0488	0.01358
2	0.0498	0.01608	0.0496	0.01691	0.0490	0.01585

Table 4: Power Comparison

				Wald test					
	MLE			**MI**			**CC**		
Case \ α	1%	5%	10%	1%	5%	10%	1%	5%	10%
1	848	955	976	718	896	953	840	956	976
2	702	874	928	525	765	865	698	876	931
				Likelihood ratio test					
1	846	955	976	709	894	953	840	956	976
2	690	870	928	503	752	862	701	877	932

The numbers in Table 4 indicate the number of tables out of 1,000 for which the independence null hypothesis was rejected under the given α levels among 1,000 tables in each type.

Table 4 shows MLE and CC have more power than MI. MLE does not show much improvement on the power levels of the tests of independence over complete-case analysis. Although MLE is often more conservative than MI with respect to Type I error levels, the test using MLE exhibited more power than MI.

6. Summary

It has been an issue to estimate the variance of MLE for the cell probabilities of an incomplete contingency table because it is very complicated to get the second derivatives of the likelihood. The likelihood including a Lagrangian multiplier related to a constraint can solve this problem.

Complete-case (CC) analysis produces biased estimates of joint probabilities under MAR and is less efficient than either MLE or MI. MLE and MI provides consistent results under either MAR situation used in simulations.

When data are missing at random, simulation results reveal that MLE provides more efficient estimates of population proportions than either multiple imputation (MI) based on data augmentation or complete case analysis, but neither MLE nor MI provides an improvement over complete-case (CC) analysis with respect to power of tests for independence.

If the missing mechanism does satisfy missing completely at random (MCAR) criterion, CC analysis can produce unbiased estimates of joint probabilities and moderate type I error level and power of tests for independence.

Appendix: Derivatives for 3-Way Tables

The log likelihood for a three-way table was given in (2.2). In the case of a three-way table, the log likelihood function incorporating a Lagrangian multiplier is

$$l(\theta^*) = l(\theta) + \gamma(1 - \sum_{a,b,c} \theta_{a,b,c}), \tag{A.1}$$

where $\theta' = (\theta_{abc}, a = 1, \cdots, A, b = 1, \cdots, B, c = 1, \cdots, C)$, $\theta^* = (\theta', \gamma)'$, and $l(\theta)$ is given in (2.2).

The first derivative of $l(\theta^*)$ in (A.1) with respective to θ_{abc} and to γ are

$$\frac{\partial l(\theta^*)}{\partial \theta_{abc}} = \frac{x_{abc}}{\theta_{abc}} + \frac{x_{mbc}}{\theta_{\bullet bc}} + \frac{x_{amc}}{\theta_{a\bullet c}} + \frac{x_{abm}}{\theta_{ab\bullet}} + \frac{x_{mmc}}{\theta_{\bullet\bullet c}} + \frac{x_{mbm}}{\theta_{\bullet b\bullet}} + \frac{x_{amm}}{\theta_{a\bullet\bullet}} - \gamma,$$

$$\frac{\partial l(\theta^*)}{\partial \gamma} = 1 - \sum_{abc} \theta_{abc}.$$

The second partial derivatives involving γ are

$$\begin{aligned}\frac{\partial^2 l(\theta^*)}{\partial\theta_{abc}\partial\gamma} &= -1,\\ \frac{\partial^2 l(\theta^*)}{\partial\gamma^2} &= 0.\end{aligned}$$

The second partial derivatives for the proportions are determined by the overlap in the three-tuples identifying the parameters. Below are illustrations for overlap of three, two, one, and zero dimensions.

$$\begin{aligned}\frac{\partial^2 l(\theta^*)}{\partial\theta_{abc}\theta_{abc}} &= \frac{-x_{abc}}{\theta^2_{abc}} - \frac{x_{mbc}}{\theta^2_{\bullet bc}} - \frac{x_{amc}}{\theta^2_{a\bullet c}} - \frac{x_{abm}}{\theta^2_{ab\bullet}} - \frac{x_{mmc}}{\theta^2_{\bullet\bullet c}} - \frac{x_{mbm}}{\theta^2_{\bullet b\bullet}} - \frac{x_{amm}}{\theta^2_{a\bullet\bullet}},\\ \frac{\partial^2 l(\theta^*)}{\partial\theta_{abc}\theta_{abt}} &= \frac{-x_{abm}}{\theta^2_{ab\bullet}} - \frac{x_{mbm}}{\theta^2_{\bullet b\bullet}} - \frac{x_{amm}}{\theta^2_{a\bullet\bullet}},\\ \frac{\partial^2 l(\theta^*)}{\partial\theta_{abc}\theta_{ast}} &= \frac{-x_{amm}}{\theta^2_{a\bullet\bullet}},\\ \frac{\partial^2 l(\theta^*)}{\partial\theta_{abc}\theta_{rst}} &= 0,\end{aligned}$$

where $r \neq a$, $s \neq b$, and $t \neq c$. The estimated covariance matrix is $-I^{-1}(\hat{\theta}^*)$, where $I(\theta^*)$ is the matrix of second derivatives of the log likelihood with respect to θ^*.

References

Chen, T. T., and Fienberg, S. E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics* **32**, 133-144.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1-38.

Finch, W. H. (2010). Imputation methods for missing categorical questionnaire data: A comparison of approaches. *Journal of Data Science* **8**, 361-378.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association* **77**, 270-278.

Kang (2006). MLE for incomplete contingency tables with lagrangian multiplier. *Journal of Korean Data & Information Science Society* **17**, 919-925.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991). Large-sample significance levels from multiple imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association* **86**, 1065-1073.

Lipsitz, S. R., Parzen, M., and Molenberghs, G. (1998). Obtaining the maximum likelihood estimates in incomplete $R \times C$ contingency tables using a Poisson generalized model. *Journal of Computational and Graphical Statistics* **7**, 356-376.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association* **77**, 237-250.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* John Wiley & Sons, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. Chapman and Hall, London.

Meng, X. L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79**, 103-111.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63**, 581-592.

Rubin, D. B. (1978). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20-34.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, New York.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London.

Insightful Corporation. (2001). S-Plus 6.1 Manual: Analyzing Data with Missing Values in S-Plus. Seattle, Washington.

Received September 11, 2010; accepted December 20, 2010.

Shin-Soo Kang
Department of Information and Statistics
KwanDong University
Kangwon-Do, 210-701, South Korea
sskang@iastate.edu

Michael D. Larsen
Department of Statistics
George Washington University
Washington, D.C. 20052, U.S.A.
mlarsen@bsc.gwu.edu