Statistica Sinica **13**(2003), 613-623

TWO-WAY CONTINGENCY TABLES UNDER CONDITIONAL HOT DECK IMPUTATION

Hansheng Wang and Jun Shao

Peking University and University of Wisconsin

Abstract: We consider the estimation of cell probabilities in a two-way contingency table where the two-dimensional categorical data have nonrespondents imputed by using a conditional hot deck imputation method. Under simple random sampling, we establish asymptotic normality of cell probability estimators based on imputed data and derive explicitly the form of their asymptotic covariance matrix, which can be used for large sample inference. We also show that estimators based on imputed data are more efficient than those obtained by ignoring nonrespondents and re-weighting when the proportion of nonrespondents is large. The results are extended to stratified sampling, under imputation, within each stratum or across strata. Two types of asymptotics are studied under stratified sampling. One deals with the case of a fixed number of strata with large stratum sizes and the other deals with the situation of a large number of strata with small stratum sizes. Some simulation results are presented to study finite sample properties of the proposed procedures.

Key words and phrases: Estimation of cell probability, imputation across strata, re-weighting, stratified sampling.

1. Introduction

Two-way contingency tables are widely used for the summarization of two-dimensional categorical data. Each cell (category) in a two-way contingency table is defined by a two-dimensional categorical variable (A, B), where A and B take values in $\{1, \ldots, a\}$ and $\{1, \ldots, b\}$, respectively. Cell probabilities $P(A = i, B = j)$, $i = 1, \ldots, a$, $j = 1, \ldots, b$, are estimated by the sample cell frequencies computed based on the observed responses of (A, B) from a sample of units (subjects).

In sample surveys or medical studies, it is not unusual that one or two of the categorical responses are missing (nonrespondents). Sampled units whose both components are missing (unit nonrespondents) can be handled by a suitable adjustment of sampling weights or sample sizes. If there are many sampled units having exactly one missing component in their responses (item nonrespondents), however, weight adjustment results in throwing away observed incomplete data and may lead to a decrease in efficiency of the statistical analysis. A popular

alternative approach to handle item nonresponse is imputation, which inserts values for nonrespondents. Statistical methods based on imputation are usually not the most efficient because of the artificial noise created during imputation. However, imputation is widely used for its simplicity and for many practical (non-statistical) reasons (Kalton and Kasprzyk (1986)).

Analysis of data with nonrespondents (with or without imputation) relies on an assumption on the response mechanism. Typically, the following assumption on response mechanism is valid or nearly valid. The population of interest can be divided into several sub-populations (referred to as imputation classes) according to the value of an auxiliary variable without nonresponse. Within an imputation class, sampled units independently have the same probability π_A to have observed A and missing B, π_B to have observed B and missing A, π_C to have observed A and B, and $1 - \pi_A - \pi_B - \pi_C$ to have missing A and B. (The probabilities π_A, π_B and π_C may be different in different imputation classes.) Once imputation classes are created, imputation or re-weighting is done within each imputation class. In many business surveys, imputation classes are strata or unions of strata; in medical studies, if data are obtained under several different treatments, then the treatment groups may be used as imputation classes. Another example is given in Section 5.2.

The main purpose of this paper is to study properties of a conditional hot deck imputation method for two-way contingency tables, which is described in Section 2. In hot deck imputation, nonrespondents are imputed by respondents of the same variable in the same dataset so that imputed values are actually occurring values, not constructed values. Hot deck imputation is popular in survey problems and is particularly useful when the variables of interest are categorical.

A basic requirement for any imputation method is that, after nonrespondents are imputed, approximately unbiased survey estimators (of cell probabilities) can be obtained by using formulas designed for the case of no nonresponse and by treating imputed values as observed data. In Section 2, we show that under simple random sampling, the conditional hot deck imputation method satisfies this requirement. Also, asymptotic normality of the estimated cell probabilities based on conditional hot deck imputation is established and explicit form of the asymptotic covariance matrix is derived so that consistent variance estimators can be obtained by substitution. In Section 3, we extend the results in Section 2 to the case where stratified sampling is used to obtain sampled units. Two types of situations are considered, the case of a small number of strata with large sizes and the case of a large number of strata with small sizes. Both imputation within each stratum and imputation across strata are considered. All the proofs are omitted and can be found in Wang (2001).

Some simulation results are presented in Section 4 to study finite sample properties of the conditional hot deck imputation method.

2. Results Under Simple Random Sampling

In this section we consider the case where a simple random sample is obtained. Simple random sampling is frequently used in medical studies. We consider the case of one imputation class so that the sampled units have probability π_A with observed A and missing B, π_B with observed B and missing A and π_C with observed A and B. The case of multiple imputation classes can be treated similarly, since imputation is carried out within each imputation class. Since units with missing A and B are ignored after a sample size adjustment, we assume for simplicity that $\pi_A + \pi_B + \pi_C = 1$.

Consider a sampled unit with observed $A = i$ and missing B. The conditional hot deck imputation method considered in this paper imputes B by the value of B of a unit randomly selected from all units with observed B and $A = i$. Let $\hat{p}_{ij}^C$ be the usual estimate of $p_{ij} = P(A = i, B = j)$ based on the two-way contingency table constructed using completers (data from units without nonresponse). Then, conditional hot deck imputation is equivalent to imputing B by j with probability $\hat{p}_{ij}^C/\hat{p}_{i\cdot}^C$, $j = 1, \ldots, b$, where $\hat{p}_{i\cdot}^C = \sum_{j=1}^c \hat{p}_{ij}^C$. Conditional hot deck imputation for a unit with observed B and missing A is similar. All nonrespondents are imputed independently.

After imputation, the cell probabilities p_{ij} are estimated using the standard formulas in the analysis of data from a two-way contingency table by treating imputed values as observed data. We denote these estimators by $\hat{p}_{ij}^I$, $i = 1, \ldots, a$, $j = 1, \ldots, b$. Let $\hat{p}^I = (\hat{p}_{11}^I, \ldots, \hat{p}_{1b}^I, \ldots, \hat{p}_{a1}^I, \ldots, \hat{p}_{ab}^I)'$ and $p = (p_{11}, \ldots, p_{1b}, \ldots, p_{a1}, \ldots, p_{ab})'$. The following result shows that $\hat{p}^I$ under conditional hot deck imputation is consistent, asymptotically unbiased, and asymptotically normal. Its proof can be found in Wang (2001). In sample surveys, sampling is usually without replacement from a finite population, but since the ratio of the sample size over the population size is nearly 0, we assume that sampling is with replacement in the asymptotic analysis.

Theorem 1. *Assume $\pi_C > 0$. Under conditional hot deck imputation,*

$$\sqrt{n}(\hat{p}^I - p) \to_d N(0, MPM' + (1 - \pi_C)P),$$

where $\to_d$ denotes convergence in distribution, $P = diag\{p\} - pp'$,

$$M = \frac{1}{\sqrt{\pi_C}} \left(I_{ab} - \pi_A diag\{p_{B|A}\} I_a \otimes U_b - \pi_B diag\{p_{A|B}\} U_a \otimes I_b \right), \qquad (1)$$

$$p_{A|B} = (p_{11}/p_{\cdot 1}, \ldots, p_{1b}/p_{\cdot b}, \ldots, p_{a1}/p_{\cdot 1}, \ldots, p_{ab}/p_{\cdot b})',$$
$$p_{B|A} = (p_{11}/p_{1\cdot}, \ldots, p_{1b}/p_{1\cdot}, \ldots, p_{a1}/p_{a\cdot}, \ldots, p_{ab}/p_{a\cdot})', \tag{2}$$

I_a is an $a \times a$ identity matrix, U_b is a $b \times b$ matrix with all entries 1, and $\otimes$ is the Kronecker product.

The asymptotic covariance matrix of $\hat{p}^I$, $\Sigma = MPM' + (1 - \pi_C)P$, can be estimated by replacing p_{ij}, π_A, π_B and π_C in Σ by $\hat{p}^I_{ij}$, $\hat{\pi}_A = n_A/n$, $\hat{\pi}_B = n_B/n$ and $\hat{\pi}_C = n_C/n$, respectively, where n_A is the number of sampled units with observed A and missing B, n_B is the number of sampled units with observed B and missing A, n_C is the number of sampled units with observed A and B, and $n = n_A + n_B + n_C$. The resulting estimator, denoted by $\hat{\Sigma}$, is a consistent estimator of Σ. This result together with the asymptotic normality of $\hat{p}^I$ can be used for large sample statistical inference.

Let $\hat{p}^C = (\hat{p}^C_{11}, \ldots, \hat{p}^C_{1b}, \ldots, \hat{p}^C_{a1}, \ldots, \hat{p}^C_{ab})'$ be the estimator of p obtained by using the two-way contingency table based on sampled units without nonresponse. Then $\sqrt{n}(\hat{p}^C - p) \to_d N(0, \pi_C^{-1}P)$. Intuitively, $\hat{p}^I$ is better than $\hat{p}^C$ when there are many nonrespondents. A general comparison of Σ and $\pi_C^{-1}P$ is not easy because of the complexity of Σ. As examples, we consider the following two extreme cases. The first example has $P(A = B) = 1$. In such a case conditional hot deck imputation actually recovers the original data (since $A = B$) and $\Sigma = P$, which is smaller than $\pi_C^{-1}P$ as long as $\pi_C < 1$. Hence, it is expected that $\hat{p}^I$ is better than $\hat{p}^C$ when A and B are highly correlated. The second example has A and B independent, which is the least favorable case for conditional hot deck imputation. Consider the 2×2 contingency table:

$$\begin{array}{ll} 0.28 & 0.12 \\ 0.42 & 0.18\,. \end{array} \tag{3}$$

The ratio of the variance of $\hat{p}^I_{ij}$ over the variance of $\hat{p}^C_{ij}$ for any (i, j), and some values of π_A, π_B and π_C, are listed in Table 1. It can be seen from Table 1 that with any fixed π_C, the ratio decreases as π_A increases; when π_C decreases, the ratio dose not always decrease because of different combinations for π_A and π_B, but there is a clear decreasing trend. In this example, the imputed estimator has a substantial advantage when $\pi_C < 0.5$. On the other hand, if $P(A = B) = 1$, the variance ratio is always equal to π_C.

3. Results Under Stratified Simple Random Sampling

In many business surveys conducted by agencies such as the U.S. Census Bureau, the U.S. Bureau of Labor Statistics, Westat, and Statistics Canada, stratified simple random sampling is adopted and imputation classes are either strata or unions of strata. In this section we extend the result in Section 2 to this situation.

Table 1. Ratio of the Var $(\hat{p}_{ij}^{I})$ over Var $(\hat{p}_{ij}^{C})$ for 2×2 contingency Table 3.

Response Probability			Variance Ratio (i,j)			
π_C	π_A	π_B	(1,1)	(1,2)	(2,1)	(2,2)
0.9	0.0	0.1	1.058	1.030	1.031	0.993
0.9	0.1	0.0	0.979	1.051	0.998	1.062
0.8	0.0	0.2	1.100	1.045	1.048	0.976
0.8	0.1	0.1	1.018	1.061	1.009	1.035
0.8	0.2	0.0	0.950	1.086	0.986	1.107
0.7	0.0	0.3	1.125	1.048	1.052	0.949
0.7	0.1	0.2	1.039	1.057	1.007	0.998
0.7	0.2	0.1	0.968	1.076	0.977	1.060
0.7	0.3	0.0	0.912	1.106	0.964	1.135
0.6	0.0	0.4	1.133	1.036	1.041	0.912
0.6	0.1	0.3	1.044	1.039	0.990	0.951
0.6	0.2	0.2	0.970	1.052	0.954	1.003
0.6	0.3	0.1	0.911	1.075	0.935	1.068
0.6	0.4	0.0	0.867	1.109	0.931	1.146
0.5	0.0	0.5	1.125	1.011	1.017	0.866
0.5	0.1	0.4	1.032	1.007	0.960	0.894
0.5	0.2	0.3	0.955	1.014	0.918	0.936
0.5	0.3	0.2	0.892	1.031	0.892	0.991
0.5	0.4	0.1	0.845	1.059	0.882	1.059
0.5	0.5	0.0	0.812	1.097	0.888	1.140
0.4	0.0	0.6	1.100	0.973	0.979	0.810
0.4	0.1	0.5	1.004	0.963	0.916	0.828
0.4	0.2	0.4	0.923	0.963	0.868	0.860
0.4	0.3	0.3	0.858	0.973	0.836	0.904
0.4	0.4	0.2	0.807	0.995	0.819	0.962
0.4	0.5	0.1	0.771	1.026	0.819	1.033
0.4	0.6	0.0	0.750	1.068	0.834	1.117
0.3	0.0	0.7	1.058	0.920	0.928	0.744
0.3	0.1	0.6	0.959	0.904	0.858	0.752
0.3	0.2	0.5	0.875	0.898	0.803	0.773
0.3	0.3	0.4	0.806	0.902	0.765	0.808
0.3	0.4	0.3	0.752	0.917	0.743	0.855
0.3	0.5	0.2	0.713	0.942	0.736	0.916
0.3	0.6	0.1	0.688	0.978	0.746	0.990
0.3	0.7	0.0	0.679	1.024	0.771	1.077

Under simple random sampling, n_C was the number of sampled units without nonresponse; under stratified sampling, $n_{h,C}$ denotes the same quantity but restricted to the hth stratum. Quantities $n_{h,A}, n_{h,B}, p_h, p_{h,ij}, p_{h,i\cdot}$ and $p_{h,\cdot j}$ are similarly defined. Within the hth stratum, we assume that a simple random

sample of size n_h is obtained (with or without replacement) and samples across strata are obtained independently. The total sample size is $n = \sum_{h=1}^{H} n_h$, where H is the number of strata. We assume that the total sampling fraction n/N is negligible, where N is the number of units in the population. Thus, we may assume that sampling is with replacement in the asymptotic analysis. The overall cell probability vector is $p = \sum_{h=1}^{H} w_h p_h$, where w_h is the hth stratum weight. Again, since unit nonresponse can be handled by re-weighting, we assume for simplicity that $n_{h,A} + n_{h,B} + n_{h,C} = n_h$.

In practice, we usually encounter one of two situations: H is fixed and all n_h's are large, or H is large and $\{n_h : h = 1, 2, \ldots\}$ is bounded.

When H is fixed and all n_h are large, if imputation classes are the same as strata, then imputation is carried out within each stratum (Section 3.1); if imputation classes are unions of several strata, then imputation is carried out across strata (Section 3.1). When H is large and all n_h's are small, each imputation class is typically a union of many strata (so that each imputation class contains enough respondents for imputation) and imputation is carried out across strata (Section 3.2).

3.1. The Case of Fixed H and Large n_i's

When H is fixed, all n_h's are large, and imputation classes are the same as strata, conditional hot deck imputation can be carried out as described in Section 2 within each stratum. Let $\hat{p}^I = \sum_{h=1}^{H} w_h \hat{p}_h^I$ be the estimator of p based on conditionally imputed values. Suppose that $n_h/n \to \rho_h > 0$ as $n \to \infty$, $h = 1, \ldots, H$. Then, a direct application of Theorem 1 leads to $\sqrt{n}(\hat{p}^I - p) \to_d N(0, \Sigma)$, where $\Sigma = \sum_{h=1}^{H} (w_h^2/\rho_h)\Sigma_h$ and Σ_h is the Σ in Theorem 1 but restricted to the hth stratum.

Consider now the case where H is fixed, all n_h's are large, and imputation classes are unions of strata. We propose the following conditional hot deck imputation procedure. For a sampled unit in the kth imputation class with observed $B = j$ and missing A, the missing value is imputed by i according to the conditional probability

$$\begin{aligned} p_{ij|B,k} &= P(A = i | B = j \text{ and } A \text{ is missing}) \\ &= \frac{P((A,B) = (i,j) \text{ and } A \text{ is missing})}{P(B = j \text{ and } A \text{ is missing})} \\ &= \frac{\sum_{h \in \mathcal{I}_k} P((A,B) \text{ from strata } h,\ A = i, B = j, \text{ and } A \text{ is missing})}{\sum_{h \in \mathcal{I}_k} P((A,B) \text{ from strata } h,\ B = j \text{ and } A \text{ is missing})} \\ &= \frac{\sum_{h \in \mathcal{I}_k} w_h \pi_{h,A} p_{h,ij}}{\sum_{h \in \mathcal{I}_k} w_h \pi_{h,A} p_{h,\cdot j}} \end{aligned} \tag{4}$$

with $p_{h,ij}$ replaced by $\hat{p}^C_{h,ij}$ and $\pi_{h,A}$ replaced by $\hat{\pi}_{h,A}$, where $\mathcal{I}_k$ contains all h's that are in the kth imputation class. Similarly, for a sampled unit in the kth imputation class with observed $A = i$ and missing B, the missing value is imputed by j according to the conditional probability

$$p_{ij|A,k} = \frac{\sum_{h\in\mathcal{I}_k} w_h \pi_{h,B} p_{h,ij}}{\sum_{h\in\mathcal{I}_k} w_h \pi_{h,B} p_{h,i\cdot}} \tag{5}$$

with parameters replaced by their estimates. Once nonrespondents are imputed, $\hat{p}^I$ can be computed by ignoring imputation classes and treating imputed values as observed data.

The asymptotic behavior of $\hat{p}^I$ is given in the following result.

Theorem 2. *Assume that H is fixed, there are K (a fixed number) imputation classes that are unions of strata, and $n_h/n \to \rho_h > 0$ as $n \to \infty$, $h = 1, \ldots, H$. For $\hat{p}^I$ based on conditional hot deck imputation across strata, $\sqrt{n}(\hat{p}^I - p) \to_d N(0, \Sigma)$, where*

$$\Sigma = \sum_{k=1}^{K} \sum_{h\in\mathcal{I}_k} \frac{w_h^2}{\rho_h}(M_h P_h M_h' + \pi_{h,A}\Sigma_h^A + \pi_{h,B}\Sigma_h^B),$$

$$M_h = \frac{1}{\sqrt{\pi_{h,C}}}\left[I_{ab} - \pi_{h,A}N_k^A(I_a \otimes U_b) - \pi_{h,B}N_k^B(U_a \otimes I_b)\right],$$

$$P_h = diag(p_h) - p_h p_h',$$

$$\Sigma_h^A = diag(a_h) - a_h a_h',$$

$$\Sigma_h^B = diag(b_h) - b_h b_h',$$

$$a_h = (p_{11|A,k}p_{h,1\cdot}, \ldots, p_{1b|A,k}p_{h,1\cdot} \ldots, p_{a1|A,k}p_{h,a\cdot}, \ldots, p_{ab|A,k}p_{h,a\cdot})',$$

$$b_h = (p_{11|B,k}p_{h,\cdot 1}, \ldots, p_{1b|B,k}p_{h,\cdot b}, \ldots, p_{a1|B,k}p_{h,\cdot 1}, \ldots, p_{ab|B,k}p_{h,\cdot b})',$$

$$N_k^A = diag\{(p_{11|A,k}, \ldots, p_{1b|A,k}, \ldots, p_{a1|A,k}, \ldots, p_{ab|A,k})\},$$

$$N_k^B = diag\{(p_{11|B,k}, \ldots, p_{1b|B,k}, \ldots, p_{1a|B,k}, \ldots, p_{ab|B,k})\}.$$

Note that the results in Theorem 2 holds even if $(\pi_{h,A}, \pi_{h,B}, \pi_{h,C})$ are different for different strata in the kth imputation class $\mathcal{I}_k$. Thus, the method of conditional hot deck imputation across strata described in this section is robust against the assumption that $(\pi_{h,A}, \pi_{h,B}, \pi_{h,C})$ is constant within imputation class k (Section 1). This is because $\pi_{h,A}$ and $\pi_{h,B}$ are incorporated in the imputation procedure through $p_{ij|B,k}$'s and $p_{ij|A,k}$'s in (4) and (5), and they can be consistently estimated when we have a large sample size for each stratum.

When imputation is carried out across strata, the conditional hot deck imputation method described in this section (and Section 3.2) imputes a nonrespondent by selecting a value from a set of respondents with probability depending on the sampling weights w_h, estimates of $p_{h,ij}$ and estimates of $\pi_{h,A}$ and $\pi_{h,B}$. Thus, it is different from the simple conditional hot deck imputation in Section 2 (which selects a value from a set of respondents with equal probability) and is similar to the weighted hot deck imputation considered in Rao and Shao (1992). When imputation classes are the same as strata, i.e., each $\mathcal{I}_k$ contains exactly one stratum, the two methods are the same, since $p_{ij|B,k} = p_{k,ij}/p_{k,\cdot j}$ and $p_{ij|A,k} = p_{k,ij}/p_{k,i\cdot}$, $k = 1, \ldots, H$.

3.2. The Case of Large H and Small n_h's

When all n_h's are small (bounded by a constant c_1) and H is large, imputation across strata is necessary. Since n_h's are small, $\hat{\pi}_{h,A}$'s and $\hat{\pi}_{h,B}$'s are not consistent estimators and the method described in Section 3.1 is not appropriate. Under the assumption that $(\pi_{h,A}, \pi_{h,B}, \pi_{h,C})$ is constant within each imputation class (Section 1), the missing value of a sampled unit in the kth imputation class with observed $B = j$ and missing A can be imputed by i according to the conditional probability $p_{ij|B,k} = \sum_{h\in\mathcal{I}_k} w_h p_{h,ij} / \sum_{h\in\mathcal{I}_k} w_h p_{h,\cdot j}$ with $p_{h,ij}$ replaced by $\hat{p}^C_{h,ij}$. Similarly, the missing value of a sampled unit in the kth imputation class with observed $A = i$ and missing B can be imputed by j according to the conditional probability $p_{ij|A,k} = \sum_{h\in\mathcal{I}_k} w_h p_{h,ij} / \sum_{h\in\mathcal{I}_k} w_h p_{h,i\cdot}$ with $p_{h,ij}$ replaced by $\hat{p}^C_{h,ij}$. Again, $\hat{p}^I$ can be computed by ignoring imputation classes and treating imputed values as observed data.

Let Σ_h have the $((i_1-1)b + j_1, (i_2-1)b + j_2)$th component

$$\frac{1}{\pi_{h,C}} p_{h,ij} + \frac{\pi^2_{h,A}}{\pi_{h,C}} p^2_{ij|A,k} p_{h,i\cdot} + \frac{\pi^2_{h,B}}{\pi_{h,C}} p^2_{ij|B,k} p_{h,\cdot j} - 2\frac{\pi_{h,A} p_{ij|A,k}}{\pi_{h,C}} p_{h,ij} - 2\frac{\pi_{h,B} p_{ij|B,k}}{\pi_{h,C}} p_{h,ij}$$
$$+2\frac{\pi_{h,A}\pi_{h,B}}{\pi_{h,C}} p_{ij|B,k} p_{ij|A,k} p_{h,ij} + p^2_{ij|B,k}\pi_{h,B} p_{\cdot j} + p^2_{ij|A,k}\pi_{h,A} p_{i\cdot} - p^2_{h,ij}$$

if $(i_1, j_1) = (i_2, j_2) = (i, j)$;

$$\frac{\pi^2_{h,A}}{\pi_{h,C}} p_{ij_1|A,k} p_{ij_2|A,k} p_{h,i\cdot} - \frac{\pi_{h,A}}{\pi_{h,C}} p_{ij_2|A,k} p_{h,ij_1} - \frac{\pi_{h,A}}{\pi_{h,C}} p_{ij_1|A,k} p_{h,ij_2}$$
$$+\frac{\pi_{h,A}\pi_{h,B}}{\pi_{h,C}} p_{ij_1|B,k} p_{ij_2|A,k} p_{h,ij_1} + \frac{\pi_{h,A}\pi_{h,B}}{\pi_{h,C}} p_{ij_2|B,k} p_{ij_1|A,k} p_{ij_2}$$
$$+\pi_{h,A} p_{ij_1|A,k} p_{ij_2|A,k} p_{h,i\cdot} - p_{h,ij_1} p_{h,ij_2}$$

if $i = i_1 = i_2$ and $j_1 \neq j_2$;

$$\frac{\pi^2_{h,B}}{\pi_{h,C}} p_{i_1 j|B,k} p_{i_2 j|B,k} p_{h,\cdot j} - \frac{\pi_{h,B}}{\pi_{h,C}} p_{i_2 j|B,k} p_{h,i_1 j} - \frac{\pi_{h,B}}{\pi_{h,C}} p_{i_1 j|B,k} p_{h,i_2 j}$$

$$+\frac{\pi_{h,A}\pi_{h,B}}{\pi_{h,C}}p_{i_1j|B,k}p_{i_2j|A,k}p_{h,i_2j}+\frac{\pi_{h,A}\pi_{h,B}}{\pi_{h,C}}p_{i_2j|B,k}p_{i_1j|A,k}p_{h,i_1j}$$
$$+\pi_{h,B}p_{i_1j|B,k}p_{i_2j|B,k}p_{h,\cdot j}-p_{h,i_1j}p_{h,i_2j}$$

if $i_1 \neq i_2$ and $j = j_1 = j_2$; and $-p_{h,i_1j_1}p_{h,i_2j_2}$ if $i_1 \neq i_2$ and $j_1 \neq j_2$, where $i_1, i_2 = 1, \ldots, a$, $j_1, j_2 = 1, \ldots, b$. Let $\Sigma_{k,A}$ have the $((i_1-1)b+j_1, (i_2-1)b+j_2)$th component $\pi_{k,A}(p_{ij|A,k} - p^2_{ij|A,k})\sum_{h\in\mathcal{I}_k} w_h^2 p_{h,i\cdot}/n_h$ if $(i_1, j_1) = (i_2, j_2) = (i, j)$; $-\pi_{k,A}p_{ij_1|A,k}p_{ij_2|A,k}\sum_{h\in\mathcal{I}_k} w_h^2 p_{h,i\cdot}/n_h$ if $i_1 = i_2 = i$ and $j_1 \neq j_2$; and 0 if $i_1 \neq i_2$ and $j_1 \neq j_2$. Let $\Sigma_{k,B}$ have entries $\pi_{k,B}(p_{ij|B,k} - p^2_{ij|B,k})\sum_{h\in\mathcal{I}_k} w_h^2 p_{h,\cdot j}/n_h$ if $(i_1, j_1) = (i_2, j_2) = (i, j)$; $-\pi_{k,B}p_{ij_1|B,k}p_{ij_2|B,k}\sum_{h\in\mathcal{I}_k} w_h^2 p_{h,\cdot j}/n_h$ if $j_1 \neq j_2$ and $i_1 = i_2 = i$; and 0 if $j_1 \neq j_2$ and $j_1 \neq j_2$.

The following result establishes the asymptotic normality of $\hat{p}^I$ based on the method of conditional hot deck imputation across strata.

Theorem 3. *Assume that there are K (a fixed number) imputation classes that are unions of strata and within the kth imputation class $\mathcal{I}_k$, $(\pi_{h,A}, \pi_{h,B}, \pi_{h,C}) = (\pi_{k,A}, \pi_{k,B}, \pi_{k,C})$ for all $h \in \mathcal{I}_k$ and $\pi_{k,C} > 0$. Assume further that $H \to \infty$, $n/N \to 0$, and that there are constants c_j, $j = 1, \ldots, 4$, such that $n_h \leq c_1$, $c_2 \leq Hw_h \leq c_3$, and $p_{h,ij} \geq c_4$ for all h. Then, $(l'Vl)^{-1/2}(\hat{p}^I - p) \to_d N(0,1)$ for any $l \in \mathcal{R}^{ab}$ with $l'Vl > 0$, where*

$$V = \sum_{k=1}^{K}\left(\sum_{h\in\mathcal{I}_k}\frac{w_h^2}{n_h}\Sigma_h + \Sigma_{k,A} + \Sigma_{k,B}\right).$$

A consistent estimator of the asymptotic covariance matrix V can be obtained by substituting $p_{h,ij}$, $\pi_{k,A}$, $\pi_{k,B}$, and $\pi_{k,C}$ by $\hat{p}^C_{h,ij}$, $\hat{\pi}_{k,A}$, $\hat{\pi}_{k,B}$, and $\hat{\pi}_{k,C}$, respectively, where $\hat{\pi}_{k,A} = \sum_{h\in\mathcal{I}_k} n_{h,A}/\sum_{h\in\mathcal{I}_k} n_h$ and $\hat{\pi}_{k,B}$ and $\hat{\pi}_{k,C}$ are similarly defined.

4. A Simulation Study

In this section we study by simulation the finite sample performances of the estimators discussed in Section 3.2, under stratified simple random sampling with a large number of strata. We created a population based on a dataset from the survey of victimization incidents conducted by the U.S. Department of Justice in 1989 (see Lohr (1999, p.443)). We considered three variables in the dataset, VIOLENT (= 1 if violent crime and = 2 if not violent crime), NUMOFF (number of offenders involved in crime; = 1 if only one offender, = 2 if more than one offenders) and SEX (= 1 if victim male and = 2 if victim female). The variable VIOLENT was used as variable A and NUMOFF was used as variable B in a 2×2 contingency table. That is, we were interested in estimating cell probabilities related to VIOLENT and NUMOFF. Since some victims reported "don't know" to the variables VIOLENT and/or NUMOFF, these variables were considered as variables with nonresponse. The variable SEX was used to create two imputation

Table 2. Simulation results for a 2×2 contingency table under stratified sampling.

Response Probability						Estimation of Cell Probability					
						Bias			Standard Deviation		
$\pi_{1,C}$	$\pi_{1,A}$	$\pi_{1,B}$	$\pi_{2,C}$	$\pi_{2,A}$	$\pi_{2,B}$	p_{11}	p_{12}	p_{21}	p_{11}	p_{12}	p_{21}
0.8	0.0	0.2	0.8	0.0	0.2	0.007	-0.002	0.016	0.013	0.013	0.017
0.8	0.0	0.2	0.8	0.1	0.1	-0.020	-0.005	-0.004	0.013	0.013	0.017
0.8	0.0	0.2	0.7	0.0	0.3	0.006	0.023	-0.034	0.013	0.013	0.017
0.8	0.0	0.2	0.7	0.1	0.2	0.001	-0.009	0.008	0.013	0.013	0.018
0.8	0.0	0.2	0.6	0.0	0.4	-0.010	-0.002	0.013	0.014	0.014	0.018
0.8	0.0	0.2	0.6	0.1	0.3	-0.013	-0.001	0.007	0.014	0.014	0.018
0.8	0.0	0.2	0.6	0.2	0.2	0.013	-0.003	0.005	0.014	0.014	0.018
0.8	0.0	0.2	0.5	0.0	0.5	0.006	0.003	-0.013	0.015	0.014	0.019
0.8	0.0	0.2	0.5	0.1	0.4	-0.005	0.021	-0.026	0.015	0.014	0.018
0.8	0.0	0.2	0.5	0.2	0.3	-0.007	0.006	0.007	0.015	0.014	0.018
0.8	0.1	0.1	0.8	0.1	0.1	-0.006	-0.006	0.019	0.013	0.013	0.017
0.8	0.1	0.1	0.7	0.0	0.3	-0.021	0.010	0.011	0.013	0.013	0.017
0.8	0.1	0.1	0.7	0.1	0.2	-0.006	-0.015	0.03	0.013	0.013	0.017
0.8	0.1	0.1	0.6	0.0	0.4	-0.011	0.010	0.023	0.014	0.014	0.018
0.8	0.1	0.1	0.6	0.1	0.3	0.000	-0.013	-0.025	0.014	0.014	0.018
0.8	0.1	0.1	0.6	0.2	0.2	-0.012	-0.020	0.021	0.014	0.014	0.018
0.8	0.1	0.1	0.5	0.0	0.5	-0.014	0.006	-0.014	0.015	0.015	0.018
0.8	0.1	0.1	0.5	0.1	0.4	0.002	-0.015	0.012	0.015	0.014	0.018
0.8	0.1	0.1	0.5	0.2	0.3	-0.001	-0.012	0.012	0.015	0.014	0.018
0.7	0.0	0.3	0.7	0.0	0.3	-0.016	0.010	0.001	0.014	0.014	0.018
0.7	0.0	0.3	0.7	0.1	0.2	0.013	-0.016	0.030	0.014	0.014	0.018
0.7	0.0	0.3	0.6	0.0	0.4	-0.025	0.029	0.006	0.014	0.014	0.019
0.7	0.0	0.3	0.6	0.1	0.3	-0.008	0.021	-0.006	0.014	0.014	0.019
0.7	0.0	0.3	0.6	0.2	0.2	0.010	-0.005	0.013	0.014	0.014	0.019
0.7	0.0	0.3	0.5	0.0	0.5	-0.007	0.005	0.005	0.015	0.015	0.019
0.7	0.0	0.3	0.5	0.1	0.4	0.032	0.008	-0.038	0.015	0.015	0.019
0.7	0.0	0.3	0.5	0.2	0.3	-0.007	0.015	-0.018	0.015	0.014	0.019
0.7	0.1	0.2	0.7	0.1	0.2	0.011	-0.002	-0.023	0.014	0.013	0.018
0.7	0.1	0.2	0.6	0.0	0.4	-0.013	0.007	-0.001	0.015	0.014	0.018
0.7	0.1	0.2	0.6	0.1	0.3	-0.004	-0.006	0.013	0.014	0.014	0.018
0.7	0.1	0.2	0.6	0.2	0.2	0.017	-0.027	-0.002	0.015	0.014	0.019
0.7	0.1	0.2	0.5	0.0	0.5	-0.009	0.010	-0.013	0.015	0.015	0.019
0.7	0.1	0.2	0.5	0.1	0.4	0.005	0.020	0.006	0.015	0.015	0.019
0.7	0.1	0.2	0.5	0.2	0.3	-0.011	0.014	0.013	0.015	0.015	0.019
0.6	0.0	0.4	0.6	0.0	0.4	0.007	-0.003	0.001	0.014	0.014	0.019
0.6	0.0	0.4	0.6	0.1	0.3	0.007	-0.010	0.029	0.015	0.014	0.019
0.6	0.0	0.4	0.6	0.2	0.2	-0.017	0.001	0.029	0.015	0.014	0.020
0.6	0.0	0.4	0.5	0.0	0.5	-0.017	0.036	-0.038	0.015	0.015	0.020
0.6	0.0	0.4	0.5	0.1	0.4	0.010	0.005	-0.013	0.015	0.015	0.020
0.6	0.0	0.4	0.5	0.2	0.3	-0.001	0.005	-0.008	0.015	0.015	0.020
0.6	0.1	0.3	0.6	0.1	0.3	-0.018	0.009	-0.010	0.015	0.014	0.019
0.6	0.1	0.3	0.6	0.2	0.2	0.002	-0.002	-0.009	0.015	0.014	0.019
0.6	0.1	0.3	0.5	0.0	0.5	-0.001	0.003	-0.014	0.016	0.015	0.019
0.6	0.1	0.3	0.5	0.1	0.4	0.004	0.002	-0.012	0.016	0.015	0.019
0.6	0.1	0.3	0.5	0.2	0.3	0.001	-0.007	-0.012	0.016	0.015	0.020
0.6	0.2	0.2	0.6	0.2	0.2	0.003	-0.019	0.019	0.015	0.014	0.019
0.6	0.2	0.2	0.5	0.0	0.5	-0.026	0.007	0.018	0.016	0.015	0.019
0.6	0.2	0.2	0.5	0.1	0.4	0.015	-0.016	-0.001	0.016	0.015	0.019
0.6	0.2	0.2	0.5	0.2	0.3	0.020	-0.016	-0.032	0.016	0.015	0.019
0.5	0.0	0.5	0.5	0.0	0.5	-0.009	0.002	-0.011	0.016	0.016	0.021
0.5	0.0	0.5	0.5	0.1	0.4	-0.034	0.000	0.001	0.016	0.016	0.021
0.5	0.0	0.5	0.5	0.2	0.3	0.013	-0.026	0.006	0.016	0.015	0.021
0.5	0.1	0.4	0.5	0.1	0.4	-0.012	-0.024	0.035	0.016	0.016	0.020
0.5	0.1	0.4	0.5	0.2	0.3	0.013	-0.018	-0.009	0.016	0.015	0.021
0.5	0.2	0.3	0.5	0.2	0.3	0.011	0.002	0.004	0.016	0.016	0.020

classes. The original dataset contains information about sampling weights, but not strata. For the purpose of running a simulation for stratified sampling with many strata, we created some artificial strata by using the variable SEX and combining the units with similar sampling weights. For SEX = 1, there are 54 strata and for SEX = 2, there are 48 strata (i.e., $K = 2$ and $H = 102$). Within each stratum, the true cell probabilities $p_{h,ij}$ were obtained from the original dataset.

Stratified simple random samples with $n_h = 10$ for all h were generated from the constructed population. For each sample, nonrespondents were created according to response probabilities in two imputation classes (SEX = 1 or 2) (see Table 2). For simplicity, we still consider only the case of no unit nonresponse. Nonrespondents were imputed according to the method in Section 3.2, i.e., conditional hot deck imputation across stratum (but within two imputation classes, SEX = 1 and 2).

Table 2 reports results (based on 10,000 simulation runs) on the bias and standard deviation in the estimation of the cell probabilities. The true overall cell probabilities (which are weighted averages of stratum cell probabilities) are $p_{11} = 0.1649$, $p_{12} = 0.1596$, and $p_{21} = 0.4897$. Overall, the performance of cell probability estimators are reasonably good.

Acknowledgement

The research was partially supported by National Science Foundation Grant DMS-01-02223. Hansheng Wang's research was also partially supported by Research Foundation, Guanghua School of Management, Peking University.

References

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing data. *Survey Methodology* **12**, 1-16.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA.

Moore, D. S. and McCabe, G. P. (1989). *Introduction to the Practice of Statistics*. 3rd edition. W. H. Freeman and Company, New York.

Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables (in applications). *J. Amer. Statist. Assoc.* **76**, 221-230.

Rao, J. N. K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79**, 811-822.

Wang, H. (2001). Two-way contingency tables with marginally and conditionally imputed nonrespondents. Ph.D. Thesis, Department of Statistics, University of Wisconsin-Madison.

Guanghua School of Management, Peking University, Beijing 100871, P. R. China.
E-mail: hansheng@gsm.pku.edu.cn

Department of Statistics, University of Wisconsin-Madison, 1210 West Dayton Street, Madsion, WI 53706-1685,, U.S.A.
E-mail: shao@stat.wisc.edu

(Received June 2001; accepted February 2003)