

WORKING PAPER NO. SSMD-88-029 E

METHODOLOGY BRANCH

**11-613E**

**no.88-29**

**c. 3**



**LOG-LINEAR IMPUTATION**

SSMD-88-029 E

A.C. Singh

Social Survey Methods Division, Statistics Canada, Ottawa

## ABSTRACT

A method of imputation based on log-linear methodology is proposed. For this purpose, an initial categorical transformation of all variables is made. Like hot deck imputation (HDI) method, the proposed log-linear imputation (LLI) method is applicable to both discrete and continuous variables. The LLI method generalizes HDI in several ways: (i) chi-square type measure of association is used to choose suitable predictors  $X$  for forming "optimal" imputation classes, (ii) the categorical distribution of the variable of interest,  $Z$  within an imputation class is model-based; and (iii)  $Z$  values are imputed under the constraint of proportional allocation to categories according to imputed proportions. As compared to the linear regression imputation (LRI) method, LLI requires a less restrictive framework. Thus LLI can be placed somewhere between HDI and LRI. Furthermore, since LLI uses model-based procedures for imputing counts corresponding to missing data, imputation variance can be assessed in estimating parameters within a certain class. This class of parameters describes characteristics of the population frequency distribution under the categorical framework. A modification of LLI is also proposed for the problem of statistical matching. This approach offers some control when the commonly used assumption of conditional independence is not valid. Use of supplementary information about conditional dependence, if available, can also be incorporated. Finally, some possible generalizations to the cases of general missing patterns and nonignorable nonresponse are indicated.

**KEYWORDS:** Categorical transformation; Missing at random; Imputed proportions; Imputation variance; Statistical matching; Conditional independence

## RÉSUMÉ

Cet article présente une méthode d'imputation faisant recours à la théorie des modèles log linéaires. Cette méthode nécessite que toutes les variables soient préalablement catégorisées. Tout comme la méthode d'imputation "hot deck" (IHD), la méthode d'imputation log linéaire (ILL) proposée ici peut être appliquée aussi bien aux variables discrètes qu'aux variables continues. La méthode IHD généralise la méthode ILL de plusieurs façons: (i) elle utilise une mesure d'association du type chi-carré pour choisir les variables auxiliaires  $X$  qui détermineront les classes d'imputation "optimales"; (ii) l'estimation de la distribution de fréquences de la variables d'intérêt,  $Z$  à l'intérieur d'une classe d'imputation repose sur approche de modélisation. ; et (iii) les valeurs de la variable  $Z$  sont assignées à chacune des catégories selon les proportions imputées par la méthode ILL. Il est à noter que la méthode ILL est applicable à des conditions beaucoup plus générales que la méthode d'imputation par la régression linéaire (IRL). Ainsi, la méthode ILL se situe quelque part entre les méthodes IHD et IRL. De plus, comme la méthode ILL se sert d'une approche de modélisation afin d'imputer les fréquences des données manquantes dans chacune des catégories, la composante de variance due à l'imputation associée à l'estimation d'une certaine classe de paramètres peut être évaluée. Cette classe de paramètres décrit les caractéristiques de la distribution de fréquences dans la population selon le regroupement en catégories qu'on aura choisi. On montre également qu'une modification de la méthode ILL peut s'appliquer au problème de l'appariement statistique. De fait, cette approche nous offre une protection advenant le cas où l'hypothèse classique d'indépendance conditionnelle n'est pas vérifiée. Il est également possible d'incorporer l'information additionnelle de la structure de dépendance conditionnelle si celle-ci est disponible. On indique finalement comment la méthode peut s'étendre aux cas où la structure de non-réponse est générale ainsi qu'aux cas où le mécanisme de non-réponse n'est pas aléatoire à l'intérieur d'une catégorie.

**MOTS CLÉS:** Catégorisations de variables; non-réponse aléatoire; proportions imputées; variance due à l'imputation; appariement statistique; indépendance conditionnelle.

## 1. INTRODUCTION

Imputation is a popular class of methods for handling nonresponse especially with large data sets subject to multi-purpose use. In the proposed Log-Linear Imputation (LLI) method, we first perform a suitable categorical transformation of variables. This may entail regrouping for those variables which are already discrete or qualitative, whereas for continuous (or quantitative) variables, groups or class-intervals based on practical considerations are chosen. We then use a log-linear model-based procedure for imputing counts at the aggregate level and not at the unit level. That is, the missing data is distributed according to imputed proportions for completing the cross-classified categorical data. Imputation at the unit level, if desired, can also be done as a second step. The categorical transformation allows LLI to be applicable to both discrete and continuous variables. Although categorization of variables in general involves loss of information, variables are often interpreted in practice in terms of a few categories. Moreover, loss of information due to an initial categorization could be considered to be offset by an increase in the capability for adequate modelling and analysis. The LLI methodology is based on a combination of theoretical developments in three areas: (i) Log-linear model selection for simple random samples; (ii) Complex survey categorical data analysis; and (iii) Models for partially classified contingency tables.

The framework for LLI is similar to that of hot deck imputation (HDI); see Ford (1983). It is simple but commonly used in practice. We consider two types of records corresponding to two patterns of response by units in the survey. One type of record contains complete information about the variable of interest  $Z$  and auxiliary variable  $X$ . The other type of record contains responses for  $X$  but  $Z$  is missing. Both  $X$  and  $Z$  are in general vector-valued. We shall assume that missing values of  $Z$  are missing at random (MAR) when the values of  $X$  are taken into account; alternatively, the nonresponse mechanism is ignorable as defined in Rubin (1976, 1983) and Little (1982). This implies that item nonrespondents behave the same way as respondents within imputation classes formed by subsets of  $X$  values. This would be a reasonable assumption whenever the auxiliary variables  $X$  can explain nonresponse. We also assume for the asymptotic inference based on probability samples that the number of complete records is large in expectation. For nonprobability samples such as an administrative data source, an appropriate working model would be required for decision purposes as indicated in section 2.

The main purpose of LLI is to generalize HDI in order to overcome certain limitations. For instance, in the HDI method, imputation classes are not usually formed using any optimality criterion and the underlying implicit model for imputation has a rather general structure. The model simply assumes that after matching for imputation classes, the distributions of complete and missing values are the same. Clearly, it would be desirable to use a model that draws strength from other classes whenever an imputation class has a small number of donors. The imputation classes in LLI are formed by considerations of selection of suitable predictors and an optimality criterion as defined in section 4. Moreover, in the interests of reducing distortion in the distribution of the completed data,  $Z$  values are imputed under the condition of proportional allocation to categories according to imputed proportions. This task involves two parts. Firstly a model-based procedure is used to impute missing counts for  $Z$ -categories. Secondly,  $Z$ -categories are assigned to missing records within each imputation class and then  $Z$  values (if quantitative) within imputed categories. LLI assigns  $Z$ -categories by HDI under the constraint of proportional allocation; i.e. according to the conditional categorical distribution  $f(Z|X)$  obtained in the process of filling in the table. It is possible, however, to have fractional counts in some categories and in those cases, more than one  $Z$ -category can be assigned with partial weights summing to unity. Once  $Z$ -categories are imputed and if  $Z$  is quantitative, a value within a category can be obtained in a variety of ways; e.g., by cold deck such as the midpoint of the class-interval or by a hot deck method. Thus, LLI differs from HDI in many respects. However, HDI can be obtained as a special case if imputation classes are not required to be optimal, if a saturated log-linear model is employed (i.e. no smoothing to the empirical distribution function for each class is done) and if the condition of proportional allocation is not enforced.

The LLI method can be regarded as a compromise between HDI and linear regression imputation (LRI) because LLI uses stronger structural assumptions than HDI does but they are not as strong as those of LRI; see Kalton and Kasprzyk (1986) for a review of imputation-based methods. We make the following general observation. The general imputation problem involves characteristics of the conditional distribution  $f(Z|X)$ . This can be studied by using a multivariate histogram estimator for the density  $f(Z|X)$ . Such an approach naturally leads to multi-dimensional contingency tables whose dimensions correspond to categorized  $X$  and  $Z$  variables. Log-linear (including logit) models can then be used to parametrize  $f(Z|X)$  which would provide a unified framework for all kinds of underlying distributions. On the other hand, the LRI method assumes that  $Z$  is continuous and involves strong assumptions about the functional form of the regression model and

covariance structure. This model is especially difficult to specify when the missing variable is multivariate. Thus, an approach placed somewhere between HDI and LRI with regard to imputation model assumptions would be desirable, see also Rubin (1987, p. 157). The LLI approach seems to fall in this area. In this case, since a saturated model is easy to specify and a covariance matrix can be generally estimated under complex designs, the modelling task is simpler than in the LRI case. Moreover, the theory of log-linear modelling for multivariate categorical data is already well developed and used widely.

If it is sufficient to fill the missing counts in the categorical framework, then with LLI an assessment of variability due to imputation can be made. If the standard methods are used for analysing completed data, the resulting inferences are invalid; in fact, they are too sharp because they do not allow for differing status of the real and imputed values (Rubin, 1978). Modification to standard analysis via Multiple Imputation as proposed by Rubin (1978, 1987) can be applied to overcome this problem. It may be noted that the problem of assessing variability of estimates in the presence of missing data is generally easy to handle with model-based procedures where inference is based on the likelihood under a given model for partially missing data. Now, in a categorical framework, it may be fair to assume that most parameters of interest can be reasonably approximated by corresponding parameters of the joint categorical distribution when class intervals for quantitative variables are approximated by scalars such as midpoints. Thus, completed cross-classified data at an aggregate level would generally meet the demands for multipurpose analyses. It can then be observed that the original imputation problem at the unit level is transformed to the problem of imputing cell proportions at an aggregate level and hence imputation variance can be assessed via model-based procedures.

The LLI method with a suitable modification can be applied to the problem of statistical matching of two or more data files which can be viewed as an imputation problem (Rubin, 1986). In the case of two microdata files, file A contains information on vectors  $X$  and  $Y$ , and file B contains information on vectors  $X$  and  $Z$ . For the purposes of analysis at the microlevel, we are interested in constructing file C that contains for each micro-unit on the original file A, the completed information about  $Z$ . In statistical matching procedures for constructing file C, the assumption of conditional independence of  $Y$  and  $Z$  given  $X$  is often made although it is known to be unreasonable (Rubin 1986, 1987). It would be useful to parametrize departure from the assumption of conditional independence for arbitrary underlying unknown joint distribution of  $(X, Y, Z)$ . The log-linear parametrization does indicate in general which parameters are necessarily omitted under the categorical

conditional independence assumption. We present a modification of LLI in section 5 in which a grid of  $(X,Y,Z)$  space for categorical transformation is chosen such that some protection against violation of the conditional independence assumption is provided in the categorical sense.

Literature on the use of log-linear methodology for modelling partially missing categorical data is relatively recent, see e.g. Fuchs (1982), Nordheim (1984), Little (1985), Fay (1986), and Baker and Laird (1988) among others. It may be of interest to note that Rubin, Schafer, and Schenker (1988) also propose imputation strategies based on log-linear models in the context of census undercount estimation; see also Schenker (1988). The development of LLI, on the other hand, was motivated from considerations of evaluating a micro-economic database termed SPSP (Social Policy Simulation Database by Wolfson et al. 1987) which was constructed at Statistics Canada for use in economic policy analysis. SPSP was built in part by statistical matching of information from Revenue Canada with the Survey of Consumer Finance. Some preliminary results are reported in Singh, Armstrong, and Lemaitre (1988). While Rubin, Schafer, and Schenker (1988) propose an imputation strategy based on log-linear models with emphasis on the Bayesian method for the nonignorable nonresponse situation, the LLI method is mainly developed for the ignorable nonresponse case in which several related issues are also addressed.

This article is organized as follows: Section 2 contains the underlying theory of LLI and in section 3, assessment of imputation variance is presented. The steps of LLI are described in section 4. In section 5, modification of LLI for application to the problem of statistical matching is described. Use of auxiliary information about conditional dependence from a small supplementary survey is explained in section 6. Finally, in section 7 extensions of LLI to the cases of nonignorable missing data and general missing patterns are indicated as possible directions for further work.

## 2. UNDERLYING THEORY

For completing missing records by LLI, we first need an initial partition  $P_0$  that provides a fairly fine (from subject matter point of view) grid of  $(X,Z)$  space. Let  $X_0, Z_0$  denote the corresponding categorized  $X, Z$  variables and represent respectively the rows and columns of an  $r_0 \times c_0$  table for convenience. Let  $p^R(0)$  denote the vector of observed cell proportions arranged in the lexicographic order and based on the complete data. Let  $\pi^R(0)$  be the corresponding true or population vector for complete respondents. The

vector  $p^R(0)$  would generally be calculated from adjusted counts based on sampling design weights. We shall use  $n$  to denote the total sample size,  $n_R$  for the number of complete respondents and  $n_M$  for the number of nonrespondents. Similarly, let  $N$ ,  $N_R$  and  $N_M$  denote respectively the sum of design weights for total, respondent and nonrespondent samples. For the row marginal proportions, we will use  $p_+^R(0)$  and  $\pi_+^R(0)$  to denote respectively the observed and population proportion vectors for complete respondents while  $p_+^M(0)$  and  $\pi_+^M(0)$  will be used for nonrespondents. For the corresponding vectors for the total sample, the superscripts will be dropped. Finally, for complete data within each row  $i$ , the vectors  $q_i(0)$  and  $\psi_i(0)$  will denote respectively the observed and population proportion vectors conditional on the row marginal corresponding to the given partition  $P_0$ .

Consider the following asymptotic framework when  $n_R$  is large in expectation. In the following, the symbol " $\dot{\sim}$ " stands for "asymptotically distributed as" and the symbol " $\dot{=}$ " is used to indicate that the difference between the two sides is negligible in probability as  $n_R$  increases. We assume

$$(p^R(0) - \pi^R(0)) \dot{\sim} N_{r_0 c_0}(0, V_0^R), \quad (2.1)$$

where the right hand side denotes a  $r_0 c_0$ -dimensional multivariate normal distribution with mean 0 and covariance matrix  $V_0^R$ . For the row marginals,

$$(p_+(0) - \pi_+(0)) \dot{\sim} N_{r_0}(0, V_{0+}), \quad (2.2)$$

and for the conditional row proportions, we have for  $i=1, 2, \dots, r_0$ ,

$$(q_i(0) - \psi_i(0)) \dot{\sim} N_{c_0}(0, W_i(0)). \quad (2.3)$$

Also, let  $S_0$  denote the asymptotic covariance of  $(p_+(0) - \pi_+(0))$  and  $(q(0) - \psi(0))$ . This would be zero for multinomial case because the likelihood could be factored (Little and Rubin, 1987, p. 98). Further we will assume that under a suitable replication method, consistent estimates of the covariance matrices defined above are available for the sampling design under consideration, and for convenience, the same notation will be used for estimated covariances as well. Also, we will drop the partition indicator 0 when we use the optimal partition  $P_*$  defined later in section 4.

For testing hypotheses about  $\pi^R(0)$ ,  $\pi_+(0)$  and  $\psi(0)$ , we can use  $\chi^2$  tests adjusted by methods proposed by Rao and Scott (1984). If the sampling design were simple which implies multinomial or product multinomial in the categorical framework, then no adjustments to the usual  $\chi^2$  would be required. Thus, for testing a hypothesis  $H_1$  about  $\pi^R(0)$ , let  $p^R(1)$  denote the estimated expected proportions under  $H_1$ . The  $p^R(1)$  is generally obtained as a pseudo MLE (maximum likelihood estimate) using multinomial likelihood. We can use the following rule to decide about  $H_1$ :

$$\text{Reject } H_1 \text{ if } I(p^R(0), p^R(1)) \geq \delta_0 \quad (2.4)$$

where  $\delta_0$  is some small positive number and  $I$  denotes the I-divergence distance between  $p^R(0)$  and  $p^R(1)$  (Csiszár 1975), defined by  $\sum_t p_t^R(0) \log(p_t^R(0)/p_t^R(1))$  when summation is over all cells. In the case of simple random sampling,  $2n_R I$  would be asymptotically chi-square with degrees of freedom given by the number of parameters specified under  $H_1$  and so  $\delta_0$  can be easily calculated for a given level  $\alpha$ . For complex designs, one can use, for instance, Rao-Scott corrections to find an adjusted  $\delta_0$  from the generalized design effects using  $V_0^R$ .

The donor or complete data may not correspond to a probability sample e.g. in the context of statistical matching (section 5), administrative data are commonly used as donors for imputation purposes. We may still use a distance measure such as I-divergence and the rule (2.4) as a working guideline for modelling  $\pi^R(0)$ . The choice of I-divergence as a metric is convenient because it easily lends itself to partitioning for nested hypotheses. The specification of  $\delta_0$  would, however, require some considerations other than distributional. In practice, the following observation would be helpful in choosing  $\delta_0$ . Suppose, we decide to say that  $p^R(0)$  and  $p^R(1)$  are close if for all  $t$ , the distance between  $p_t^R(0)$  and  $p_t^R(1)$  relative to their average is small. That is,

$$(p_t^R(0) + p_t^R(1))^{-1} |p_t^R(0) - p_t^R(1)| < \epsilon_0 \text{ for all of } t, \quad (2.5)$$

where  $\epsilon_0$  is chosen arbitrarily e.g. .01 as a working value. This implies that

$$(\frac{1}{4}) I(p^R(0), p^R(1)) < \epsilon_0 \text{ (approximately for large } n_R). \quad (2.6)$$

Thus,  $\delta_0$  can be set equal to  $4\epsilon_0$ . To see (2.6), first note that  $I/4$  is asymptotically (for large  $n_R$ ) equivalent to the Hellinger distance (Bishop, Fienberg, and Holland, 1975, p. 513)

and consequently it is approximately bounded above by unity. Furthermore, defining Hellinger distance,  $H(p(0), p(1))$  by  $(\frac{1}{2})\sum_t (\sqrt{p_t(0)} - \sqrt{p_t(1)})^2$  and the sup-norm distance  $D(p(0), p(1))$  by  $(\frac{1}{2})\sum_t |p_t(0) - p_t(1)|$ , we have from Le Cam's lemma (1970, p. 803),

$$H(p(0), p(1)) \leq D(p(0), p(1)). \quad (2.7)$$

It then follows from (2.5) that  $H(p(0), p(1)) < \epsilon_0$ , which establishes (2.6). The result (2.6) can be used in providing some practical guideline in choosing  $\delta_0$  in the absence of any probability consideration.

We now consider the problem of selecting a subset of  $X$  as suitable predictors for  $Z$ . This is required in the first step of LLI to be described in section 4. This problem is similar to the predictor selection problem in multiple linear regression. In log-linear analysis, Goodman's partitioning of the likelihood ratio statistic  $G^2$  (see Fienberg, 1977, p. 51) is often used in model selection. This needs to be modified for our purpose because we distinguish between the target ( $Z$ ) and predictor or auxiliary ( $X$ ) variables and are interested in choosing a subset of  $X$ . First we define a chi-square type measure of association between  $Z_0$  and  $X_0$  by means of I-divergence. We then develop a partitioning of this measure for a sequence of nested hypotheses by adapting Goodman's partitioning of  $G^2$ . This partitioning can be used to rank  $X$  variables in an increasing order of importance and provides a step-wise method for eliminating  $X$  variables from a model.

Suppose that  $X_0$  includes three variables,  $X_{01}$ ,  $X_{02}$  and  $X_{03}$ . Let  $H_3$  be the hypothesis of independence of  $X_0$  and  $Z_0$  denoted by  $Z_0 \perp\!\!\!\perp X_0$ . Then I-divergence distance for testing  $H_3$  gives a measure of association between  $Z_0$  and  $X_0$ . We shall use  $I(H_3)$  to denote this measure. Similarly, for the collapsed table over  $X_{03}$ , let  $H_2$  denote the hypothesis of independence of  $Z_0$  and the reduced vector  $(X_{01}, X_{02})$ . Let  $I(H_2)$  be the corresponding measure of association. All I-measures are computed, of course, under the pseudo-multinomial assumption.

The above I-measures for testing associations are easy to compute for various subsets of  $X_0$  variables. It then follows from proposition 2.1 given below that they can be conveniently used in practice to compute conditional test statistics for selecting predictors. For example, suppose  $I(H_3) \geq \delta_0$ , so that  $Z_0$  and  $(X_{01}, X_{02}, X_{03})$  jointly have strong association. In order to decide whether an  $X_0$  variable, say  $X_{03}$ , can be dropped, I-divergence for conditional independence hypothesis,  $K_3: Z_0 \perp\!\!\!\perp X_{03}$  given  $(X_{01}, X_{02})$  is

required.  $I(K_3)$  can, of course, be computed directly. However, it would be easier to compute from the following proposition.

Proposition 2.1 Let hypotheses  $H_1, H_2$  and  $K_3$  be defined as above. We have

$$I(K_3) = I(H_1) - I(H_2). \quad (2.8)$$

The proof of this proposition follows easily from Goodman's partitioning calculus (see Fienberg 1977, p. 52). First note that  $H_1$  is nested in  $K_3$  and so  $I(K_3)$  can be obtained as  $I(H_1) - I(H_1|K_3)$ . Now use the fact that the conditional test statistic  $I(H_1|K_3)$  indeed coincides with  $I(H_2)$ . As a point of interest, it may be noted that here Goodman's partitioning is used in a reverse order because an unconditional test statistic is being computed as a difference of unconditional and conditional test statistics. Based on the above proposition, one can compute suitable difference of  $I(H_i)$ 's where  $i$  varies over the number of predictors, to decide whether or not to retain an  $X_0$  variable or a subset of  $X_0$ . Analogous to multiple regression, a step-wise procedure could be employed for choosing  $X$  variables.

Now suppose a partition  $P_\star$  is chosen and corresponding categorical variables are  $X_\star$  and  $Z_\star$  which form an  $r \times c$  table with rows and columns representing  $X_\star$  and  $Z_\star$  categories respectively. In LLI, the conditional categorical distributions  $f(Z_\star|X_\star)$  are smoothed jointly over all  $X_\star$  categories using log-linear modelling based on the complete data. To do this, we first model the joint distribution  $f(X_\star, Z_\star)$  by I-divergence or some other  $\chi^2$  type measure under the condition that all  $X_\star$ -effects are retained. This approach would lead to an appropriate model for  $r$  conditional probability distributions  $\psi_j$ 's defined earlier in (2.3) such that  $X_\star$  marginals from complete data remain conditionally fixed. Let  $H_\star$  denote the chosen model for  $\psi$ . After  $H_\star$  is determined and the corresponding estimates of  $\psi$  parameters, the supplementary column of missing data is distributed over  $Z_\star$  categories according to estimated  $\psi$  or  $f(Z_\star|X_\star)$  for each  $X_\star$  category. Thus, the new marginals of the smoothed counts in the completed  $r \times c$  table match with the observed  $X_\star$  configuration based on all  $n$  units. In other words, the  $X_\star$  data which is not subject to nonresponse is not smoothed. As well the counts in  $(X_\star, Z_\star)$  table for the complete data are not smoothed. These are not restrictions but may be desirable from practical considerations. We can therefore express the imputed proportion vector  $\hat{\pi}^M$  and the completed proportion vector  $\hat{\pi}$  for the  $r \times c$  table corresponding to partition  $P_\star$  as

$$\hat{\pi}_{ij}^M = p_{i+}^M \hat{\psi}_{ij}, \quad i=1, \dots, r; j=1, \dots, c,$$

$$\hat{\pi}_{ij} = (N_R/N) p_{ij}^R + (N_M/N) \hat{\pi}_{ij}^M \quad (2.9)$$

where  $p_{i+}^M$  is the observed proportion (generally based on design weighted counts) in the  $i$ -th  $X_*$  category for the missing data, and  $\hat{\psi}_{ij}$  is the smoothed proportion under  $H_*$ .

Alternatively, we can express (2.9) in vector notation in two ways, namely

$$\hat{\pi} = \frac{N_R}{N} (D_{p_+^R} \otimes I_C) q + \frac{N_M}{N} (D_{p_+^M} \otimes I_C) \hat{\psi} \quad (2.10a)$$

$$= \frac{N_R}{N} D_q (p_+^R \otimes 1_C) + \frac{N_M}{N} D_{\hat{\psi}} (p_+^M \otimes 1_C) \quad (2.10b)$$

where  $I_C$  is the  $c \times c$  identity matrix,  $1_C$  is a  $c$ -vector of ones,  $\otimes$  denotes the usual Kronecker product and  $D$  denotes a diagonal matrix. Clearly if  $n_M = 0$ , then  $\hat{\pi}$  coincides with the observed vector  $p$ . The estimate  $\hat{\pi}^M$  can be justified as a pseudo MLE (maximum likelihood estimate) of  $\pi$  assuming multinomial sampling (see Little and Rubin 1987, Ch. 9). The estimate  $\hat{\pi}^M$  is computed under the saturated model for  $X_*$  table obtained from partially classified data and  $H_*$  model for  $(X_*, Z_*)$  table corresponding to complete data with  $X_*$  counts fixed. The above observation follows from the fact that the missing pattern is monotone and that the method of factored likelihood can be applied for finding MLE. The following proposition 2.2 shows that the estimate  $\hat{\pi}$  has reasonable large sample properties.

Proposition 2.2 Under model  $H_*$  for  $\psi$  and saturated model for  $\pi_+^M$ , we have as  $n_R$  gets large in expectation,

$$(\hat{\pi} - \pi) \sim N_{rc}(0, V_*),$$

where  $V_* = V_{*1} + V_{*2} + V_{*3}$ ,  $V_{*1} = U W U'$ ,

$$V_{*2} = D_{\psi} (V_+ \otimes 1_C 1_C') D_{\psi}, \quad V_{*3} = U (S' \otimes 1_C') D_{\psi} + D_{\psi} (S \otimes 1_C) U,$$

$W = \text{block diag } (W_1, W_2, \dots, W_r)$  and  $U$  is defined below by (2.16).

Proof. Let  $H_*$  define the parameters  $\log \psi$  as a linear function of a reduced set of  $\theta$  parameters and let  $B$  denote the matrix of derivative  $(\partial \psi / \partial \theta)$ . Then the pseudo MLE  $\hat{\theta}$  is obtained as a solution of

$$B' D_{\psi}^{-1} (q - \psi) = 0. \quad (2.11)$$

It now follows from (2.3), that

$$\begin{aligned} \hat{\theta} - \theta &\approx (B' D_{\psi}^{-1} B)^{-1} B' D_{\psi}^{-1} (q - \psi) \\ &\approx N_s^{-1} (0, (B' D_{\psi}^{-1} B)^{-1} B' D_{\psi}^{-1} W D_{\psi}^{-1} B (B' D_{\psi}^{-1} B)^{-1}) \end{aligned} \quad (2.12)$$

where  $s$  is the number of  $\theta$  parameters.

Next note that expanding  $\hat{\pi}_+^M$  about  $(\pi_+^M, \psi)$ , we get from (2.10),

$$\begin{aligned} \hat{\pi}_+^M - \pi_+^M &\approx (D_{\pi_+^M} \otimes I_C) (\hat{\psi} - \psi) + D_{\psi} ((p_+^M - \pi_+^M) \otimes 1_C) \\ &\approx (D_{\pi_+^M} \otimes I_C) A(q - \psi) + D_{\psi} ((p_+^M - \pi_+^M) \otimes 1_C), \end{aligned} \quad (2.13)$$

in view of the fact that  $\hat{\psi} = \psi(\hat{\theta})$  and  $\hat{\psi} - \psi \approx A(q - \psi)$  where  $A$  is obtained from (2.12) as  $B (B' D_{\psi}^{-1} B)^{-1} B' D_{\psi}^{-1}$ . Similarly,

$$\begin{aligned} \hat{\pi}_+^R - \pi_+^R &= p_+^R - \pi_+^R \\ &\approx (D_{\pi_+^R} \otimes I_C) (q - \psi) + D_{\psi} ((p_+^R - \pi_+^R) \otimes 1_C). \end{aligned} \quad (2.14)$$

Now using the relation  $(N_R/N) \pi_+^R + (N_M/N) \pi_+^M \approx \pi$ , we have from (2.13) and (2.14),

$$\hat{\pi} - \pi \approx U(q - \psi) + D_{\psi} ((p_+ - \pi_+) \otimes 1_C), \quad (2.15)$$

where

$$U = \frac{N_R}{N} (D_{\pi_+^R} \otimes I_C) + \frac{N_M}{N} (D_{\pi_+^M} \otimes I_C) A. \quad (2.16)$$

The proposition 2.2 follows from (2.2), (2.3) and (2.15).

### 3. ASSESSMENT OF IMPUTATION VARIANCE

It can be seen from proposition 2.2 that the covariance expression  $V_{\star}$  of the estimated proportion vector  $\hat{\pi}$  for the completed table takes into account of the variability due to imputation via model-based computations. This implies that for any population parameter defined as a smooth function of  $\pi$ , asymptotic variance incorporating imputation effect and its estimate can be obtained. For instance, if  $Z$  is categorical whose categories coincide with those of  $Z_{\star}$ , then asymptotic covariance matrix of the estimate  $\hat{\phi}$  for the  $c$ -vector  $\phi$  of  $Z_{\star}$ -category proportions can be calculated from the linear transformation  $\hat{\phi} = (1_r' \otimes I_c) \hat{\pi}$ . Moreover, if  $Z$  is univariate continuous and the  $c$ -vector  $m$  represents midpoints of  $Z_{\star}$ -categories, then variance of the estimate  $m' \hat{\phi}$  of the mean for the grouped frequency distribution of  $Z$  can easily be calculated from the covariance matrix of  $\hat{\phi}$ . The asymptotic variances of  $\hat{\phi}$  and  $m' \hat{\phi}$  are summarized in the following proposition, and are denoted by subscript " $\infty$ ".

#### Proposition 3.1

Let  $V_{\star 1}$ ,  $V_{\star 2}$ , and  $V_{\star 3}$  be the same as defined before in proposition 2.2. Then,

$$\text{cov}_{\infty}(\hat{\phi}) = (1_r' \otimes I_c)(V_{\star 1} + V_{\star 2} + V_{\star 3})(1_r \otimes I_c),$$

and

$$\text{var}_{\infty}(m' \hat{\phi}) = m' \text{cov}_{\infty}(\hat{\phi}) m.$$

#### Remark 3.1

If one ignores the effect of imputation and treats the estimate  $\hat{\pi}$  as if it were based on  $n$  complete records, then it would imply using another estimate of  $\text{cov}(\hat{\pi})$ , namely, the one obtained by a suitable replication method when the completed data are used. In other words, no distinction is made between real and imputed values. The effect of this, in general, would be to decrease  $\text{var}_{\infty}(l' \hat{\phi})$  for arbitrary linear transformation vector  $l$ . In particular, the estimate  $m' \hat{\phi}$  would appear to be more efficient than it actually is. It follows that by ignoring the imputation effect,  $\chi^2$  type tests based on  $\hat{\pi}$  even after adjustments for complex designs would not have asymptotically correct chi-square distributions but that of linear combinations of chi-square variables. This is in the same spirit as the results of Gimotty (1987).

### Remark 3.2

If one were to ignore missing data completely, and consider a pseudo MLE  $\tilde{\pi}$  based on complete records only, then the estimate would not be consistent in general. It would be so under the assumption of missing completely at random (MCAR), see Little and Rubin (1987, p175). Even if data were MCAR, the effect of ignoring missing data would essentially amount to ignoring  $p_+^M$ , resulting in less efficient estimation of  $\pi$ .

### Remark 3.3

If estimates of interest are not functions of  $\hat{\pi}$ , but require imputation at the unit level, then the general procedure of Multiple Imputation (Rubin, 1978, 1987) can be used to assess imputation variance.

## **4. THE PROPOSED METHOD: LOG-LINEAR IMPUTATION**

Consider the initial grid  $P_0$  and the corresponding categorical variables  $(X_0, Z_0)$  as defined in section 2. The LLI method can be described in the following four steps.

### Step I Choice of X variables

Reduce the dimension of  $X_0$ , if possible, by choosing a parsimonious subset that can be effectively used to predict  $Z_0$ . It is assumed that  $X_0$  and  $Z_0$  would have strong association, i.e. the measure of association  $I(Z_0 \parallel X_0) \geq \delta_0$ . If this were not so, then choice of  $X$  as auxiliary information for imputing  $Z$  would be questionable. Next one could determine in cases where  $X$  is multivariate whether all components are needed. To do this, compute I-divergence measures for conditional independence hypotheses as shown in proposition 2.1. This way one can choose a set of  $X$  variables which are deemed important. Let  $X_1$  denote the chosen subset of  $X_0$ . If there was no reduction, then  $X_1$  would coincide with  $X_0$ . Let  $P_1$  denote the revised partition. Also, set  $Z_1$  equal to  $Z_0$  for notational convenience only because it is not affected in changing  $P_0$  to  $P_1$ .

### Step II Choice of Optimal Partition $P_*$

Let  $G$  denote a class of partitions  $P_1, P_2, \dots$  such that for each  $P_i$ , the association between the corresponding categorical variables  $X_i$  and  $Z_i$  is high, i.e. I-measure of

association is  $\geq \delta_0$ . The  $P_i$ 's represent modified versions of  $P_1$  which may be coarser or have different cell boundaries. We then define an instability measure  $R(\epsilon)$  related to the coarseness of a partition that will allow us to choose the optimal partition  $P_\star$  from the class  $G$ . We set  $r(\epsilon) = n(\epsilon)/T$ , where  $n(\epsilon)$  is the number of cells with proportions less than or equal to  $\epsilon$ , a small predetermined positive constant and  $T$  denotes the total number of cells. Note that cells with zero counts do contribute to  $r(\epsilon)$ . Now, the optimal partition  $P_\star$  is the partition in  $G$  for which  $r(\epsilon)$  is smallest. The use of  $r(\epsilon)$  has an heuristic justification. For a particular choice of  $X_1$  and  $Z_1$ , modification of a partition to make it finer will generally increase association between the corresponding categorical variables. Use of  $r(\epsilon)$  guards against selection of fine partitions containing many cells with small proportions. For the chosen optimal partition  $P_\star$ , let  $(X_\star, Z_\star)$  denote the corresponding categorical variables.

#### Step III Smoothing of $f(Z_\star|X_\star)$

It is easier to work with the joint distribution  $f(X_\star, Z_\star)$  where  $X_\star$  marginal counts are fixed conditionally. Use I-divergence or some other chi-square type measure to choose a log-linear model  $H_\star$  containing all  $X_\star$  effects using complete records. In this case, the  $RX^2$  method for model selection described in Singh (1988) may be conveniently used. While a parsimonious model is desirable, the saturated model can be retained if it is not feasible to reduce it. Finally, a smoothed version of conditional categorical distributions  $f(Z_\star|X_\star)$  can be obtained simply by rescaling expected counts under  $H_\star$  in order that row proportions sum to unity.

#### Step IV Imputation task

##### (a) Imputing counts

The  $X_\star$  marginal counts in the missing data are distributed over  $Z_\star$  categories according to the smoothed version of  $f(Z_\star|X_\star)$  obtained in the previous step. This process amounts to imputing at an aggregate level for the units in the  $X_\star$  category corresponding to row  $i$ .

The resulting completed proportion vector  $\hat{\pi}$  was given earlier by (2.10).

(b) Imputation of missing Z values under proportional allocation

This is the second part of the final step and was already explained in the introduction. Thus, under LLI each incomplete record would be assigned an imputed category as well as a value (if quantitative) within the category. More than one imputed categories (and values) with partial weights is also allowed whenever necessary.

## 5. MODIFICATION OF LLI FOR THE PROBLEM OF STATISTICAL MATCHING

The problem of statistical matching was briefly described in the introduction. Some useful references are Kadane (1978), Sims (1978), U.S. Department of Commerce (1980), Rodgers (1984), and Rubin (1986) among others. The process of statistical matching for merging of two files A and B can be viewed as a process of imputing Z values for the candidate records (X,Y) in file A using (X,Z) records from file B as donors in a single merged file obtained by combining files A and B. As before, we assume that the Z values are missing at random in the combined file within certain imputation classes formed by (X,Y) values. However, it differs from the usual imputation problem because there are no donor records containing the complete set of values (X,Y,Z).

If one can assume that Y and Z are conditionally independent given X i.e.  $f(Z|X,Y)$  equals  $f(Z|X)$ , then the information in Y can be ignored. The problem of completing records with missing Z values in file A reduces to the usual imputation problem in a single file. Thus, the LLI method described earlier can be applied. We shall denote this method of statistical matching by LLI-S where "S" stands for single file approach. However, as shown in Rubin (1986), the relationship between Y and imputed Z values in file A may differ substantially from the true relationship when there is departure from the assumption of conditional independence. This is a major problem since matching was conducted in first place to analyse Y,Z relationship. This leads to the following modification of LLI denoted here by LLI-M where 'M' stands for multiple file case.

For the log-linear imputation approach to statistical matching, Singh, Armstrong, and Lemaitre (1988) give an illustrative example along with some preliminary simulation results. In LLI-M, the Y information is not ignored in the process of statistical matching. It is known (see e.g. Mosteller 1968) that categorical association differs according to breaking points or boundaries chosen for various fixed values of the correlation coefficient in bivariate normals. This forms the basic idea of LLI-M. We transform the

statistical matching problem to one involving categorical variables  $X_*, Y_*, Z_*$  so that the unavoidable assumption of conditional independence holds approximately in the transformed framework. This of course cannot be checked directly because there is no information on the joint distribution of  $X_*$ ,  $Y_*$  and  $Z_*$ . However, an important advantage of the categorical approach is that a suitable criterion can be constructed to control possible violation of the conditional independence assumption. This criterion is used to choose categories and thus  $Y$  information ends up being used in the process. The LLI-M method can be described in four steps. As before, consider an initial partition  $P_0$  with corresponding categorical variables  $X_0$ ,  $Y_0$ , and  $Z_0$ .

### Step I Choice of $X$ variables

These variables should be chosen separately for File A and File B using the methodology described in section 2. Variables that are used to predict both  $Y$  and  $Z$  should be categorized in the same way but not all variables used to predict  $Y$  need also be used to predict  $Z$ , and vice versa. Let  $X_1$  denote the vector formed as the union of the variables used to predict  $Y$  and the variables used to predict  $Z$ . Let  $P_1$  denote the corresponding partition.

### Step II Choice of optimal partition $P_*$

We need to check departures from the conditional independence assumption of  $Y$  and  $Z$  given  $X$  in the categorical framework. There is of course no information on the joint categorical distribution. However, under pseudo-multinomial assumptions, we can estimate the expected proportion vector  $\tau$  corresponding to the conditional independence hypothesis. This allows us to construct an upper bound  $\eta$  on the  $\chi^2$  distance corresponding to the hypothesis of conditional independence. We have

$$\eta = \text{tr} (D_{\tau}^{-1}), \quad \tau = \xi \otimes \psi \quad (5.1)$$

where  $\psi$  is the vector of conditional proportions correspondings to  $f(Z|X)$ , and  $\xi$  is the vector of cell proportions for joint distribution  $f(X, Y)$ .

The measure  $\eta$  can be termed as the sensitivity measure. A good choice of grid or partition is one that tends to make  $\eta$  small. However, use of  $\eta$  exclusively leads to a very coarse grid and a trivialization of the problem. For this reason, we introduce a balancing

factor that requires high categorical association between  $X$  and  $Y$ , as well as between  $X$  and  $Z$ . Note that the requirement of high association favours use of fine grids. Thus there is a tradeoff between low sensitivity and high association. We first generate a class of grids by defining categories in various ways. Next we restrict attention to the class  $G$  of grids for which association measures are above a threshold value,  $\delta_0$ . Finally, the optimal partition  $P_*$  is that grid in the class  $G$  for which  $n$  is minimized. It should be noted that in practice, some elements of  $\tau$  may be zero, leading to computational difficulties. In such cases, zero elements of  $\tau$  can be replaced by some small positive constant, say  $\gamma$ , and all entries rescaled such that their total is unity.

### Step III Estimation of $f(Z_*|X_*,Y_*)$

Since we assume that categorically,  $Y_*$  and  $Z_*$  are approximately conditionally independent given  $X_*$ , the smoothed estimate of  $f(Z_*|X_*)$  can be used as an estimate of  $f(Z_*|X_*,Y_*)$  for all  $Y_*$  categories. Now, the smoothing of  $f(Z_*|X_*)$  using the donor data from file B is analogous to the step III of section 4.

### Step IV Statistical Matching Task

#### (a) Imputing Counts

For file A of candidate records, the counts within each  $(X_*,Y_*)$  category are distributed over  $Z_*$  categories according to the smoothed version of  $f(Z_*|X_*,Y_*)$  obtained in the previous step. Thus, we impute counts (or proportions) for  $Z_*$  categories at an aggregate level within each subset of file A records corresponding to  $(X_*,Y_*)$  category.

#### (b) Imputation of missing $Z$ values under proportional allocation

This part of step IV is similar to that in section 4. The only difference is that  $X_*$  category or the imputation class is replaced by  $(X_*,Y_*)$  category and the complete data is defined by file B. The donor records for an imputation class are obtained by matching the  $X_*$  characteristic only, i.e. by ignoring  $Y_*$ .

Besides the problem of having to assume conditional independence necessarily because observations containing the complete set of variables  $(X,Y,Z)$  are not available, there are several other issues that need to be addressed in practice. The main ones are:

- (i) Universe Differences — the two files may represent different populations with varying degrees of overlap, including no overlap.
- (ii) Unit differences — the records on the two files may correspond to different conceptual units.
- (iii) Differences in Linking Information — The distributions of  $X$  values for files A and B could differ due to differences in definitions and response error components as well as universe differences. Also records that refer to the same entity may have different  $X$  values in the two files.
- (iv) Constrained Matching — Statistical matching procedures are sometimes constrained so that certain characteristics of the distribution of imputed  $Z$  values coincide with those of the original distribution of  $Z$  in file B.

The above mentioned issues do not arise in the usual imputation problem. The point (iv) may not really be an issue because it would be better to preserve the conditional distribution  $f(Z|X)$  rather than the marginal distribution of  $Z$ . Although we do not address points (i), (ii) and (iii) in this article, it is assumed that suitable adjustments to the values of variables in one or both files have been made before LLI is applied. With regard to variance calculations in the case of LLI-M, results similar to those given in section 4 can be developed.

## 6. USE OF SUPPLEMENTARY SAMPLE FOR AVOIDING THE CONDITIONAL INDEPENDENCE ASSUMPTION

In LLI-M described in the previous section, we assumed that  $P_{\star}$  can be chosen such that at least in the categorical sense,  $Y_{\star}$  and  $Z_{\star}$  are approximately conditionally independent given  $X_{\star}$ . So, we do not need to look for an extra source of information for estimating  $f(Z_{\star}|X_{\star}, Y_{\star})$  in addition to  $(X_{\star}, Z_{\star})$  data from file B. However, an improvement over LLI-M could be made if a small supplementary survey or some other source was available with data on all three variables  $(X, Y, Z)$ . In fact, the variables in the supplementary sample could be different. All that is needed is that in the log-linear modelling, the parameters denoting  $(Y, Z)$  and  $(X, Y, Z)$  factor effects can be estimated from the supplementary sample. Thus, we would not necessarily drop these factor effects in the step III for estimating  $f(Z_{\star}|X_{\star}, Y_{\star})$ . It may be noted that Rubin (1986) also used extra

information in the regression imputation approach to statistical matching for illustrating sensitivity of matching results to departures from conditional independence assumption.

Let file D denote the supplementary information containing all three variables  $(x, y, z)$ . With this extra information, we need only to modify step III of LLI-M. This can be done as follows. In the absence of any extra information about  $(X, Y, Z)$ , we had to ignore  $Y_*$  in step III for smoothing  $f(Z_* | X_*, Y_*)$  based on only  $(X_*, Z_*)$  data. The extra information in file D can be used to put certain constraints on the estimate of  $f(Z_* | X_*, Y_*)$ . We have three configurations, namely,  $(X_*, Y_*)$  from file A,  $(X_*, Z_*)$  from file B and  $(X_*, Y_*, Z_*)$  from file D. It may be noted that these three tables would not in general be consistent with regard to marginal totals. We wish to construct a new table for  $(X_*, Y_*, Z_*)$  for finding  $f(Z_* | X_*, Y_*)$  such that (i) it matches with the observed counts in  $(X_*, Y_*)$  table for file A, (ii) it preserves the  $f(Z_* | X_*)$  distribution ie  $(X_*, Z_*)$  interaction from the table in file B, and (iii) it preserves  $(Y_*, Z_*)$  and  $(X_*, Y_*, Z_*)$  interaction effects from the table in file D. The distribution  $f(Z_* | X_*)$  may very well be a smoothed version via log linear modelling. The above table can be constructed by using the raking-ratio method of survey data analysis in two steps. First, rake the smoothed version of  $(X_*, Z_*)$  table so that it matches with the  $X_*$  marginal of table in file A. Next, rake the  $(X_*, Y_*, Z_*)$  table of file D so that it matches the  $(X_*, Y_*)$  table in file A and the already raked  $(X_*, Z_*)$  table of file B. The resulting  $(X_*, Y_*, Z_*)$  table will ensure that the  $(X_*, Y_*)$  counts are the same as the observed counts for candidate records in file A. We then compute  $f(Z_* | X_*, Y_*)$  for the imputation task of step IV.

For the assessment of imputation variance in the case of LLI method with supplementary information, it should be possible to use the existing results on variance calculations for raking-ratio estimators, see Binder and Th  berge (1988) and other references contained therein.

## 7. CONCLUDING REMARKS

Suppose one is willing to compromise in the sense that characteristics of continuous variables of interest will be approximated by suitable analogs for discrete (or grouped) distributions obtained by a categorical transformation. Then the method of log-linear imputation (LLI) may have potential benefits for filling in the partially classified data at an aggregate level. The effect of imputation on variance of estimators based on completed data via imputed proportions for cells under the categorical transformation can

also be determined. This would, of course, be valid for the chosen categorical transformation only. Furthermore, the LLI method can be used to impute a category or a value within a category at the unit level. With respect to underlying model assumptions, it provides a compromise between the hot deck methods and linear regression methods for imputation. For future investigation, it would be of interest to perform a simulation study for evaluating performance of LLI in comparison to other methods. Evaluation measures for both levels, namely, imputing proportions at the aggregate level and imputing values at the unit level should be calculated.

An important direction in which LLI can be generalized is when there is a more general pattern of missing data. If the missing data had a more general monotone pattern than the simple case considered in this article, it would be relatively easy to extend the results by following the treatment in Little and Rubin (1987, Ch. 9). However, for non-monotone missing patterns, one would require an iterative method such as EM algorithm of Dempster, Laird, and Rubin (1977) to compute the estimated proportions for the missing cells. Extension of LLI to this case and development of the corresponding variance formulas incorporating imputation effects need to be investigated. Another important direction for further development of LLI is to assume nonignorable nonresponse. Analyses for nonignorable nonresponse models for categorical data have been considered by Nordheim (1984), Little (1985), Fay (1986), Baker and Laird (1988) and Rubin, Schafer, and Schenker (1988) among others. In view of these recent results, generalizations of LLI could be considered.

Some other interesting directions for further work arise from the area of statistical matching. Work on developing a Bayesian LLI is currently being investigated by Stroud (1988).

#### ACKNOWLEDGEMENTS

This research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada. A version of this article will also appear in the proceedings of the Fifth Annual Research Conference, U.S. Bureau of the Census, Washington, March 19-22, 1989. I would like to thank Doug Drew, John Armstrong, and Georges Lemaitre for helpful discussions and Tom Stroud and Jon Rao for their useful comments. I would also like to thank Judy Clarke and Pat Pariseau for processing the manuscript in a very short notice.

## REFERENCES

- BAKER, S.G. and LAIRD, N.M. (1988). Regression Analysis for categorical variables with outcome subject to noignorable nonresponse. *J. Amer. Statist. Assoc.*, 83, 62-69.
- BINDER, D.A. and THÉBERGE, A. (1988). Estimating the variance of raking-ratio estimators. *Can. J. Statist.*, 16, Supplement, 47-55.
- BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Massachusetts: MIT Press.
- CSISZÁR, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3, 146-158.
- DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM Algorithm. *J. Roy. Statist. Soc., B*, 39, 1-38.
- FAY, R.E. (1986). Causal models for patterns of nonresponse. *J. Amer. Statist. Assoc.*, 81, 354-364.
- FIENBERG, S.E. (1977). *The Analysis of Cross-classified Categorical Data*. Cambridge, Massachusetts: MIT Press.
- FORD, B.L. (1983). An overview of hot-deck procedures. In *Incomplete Data in Sample Surveys* Vol. 2. (W.G. Madow, I. Olkin, and D.B. Rubin, eds.), Academic, New York, 185-207.
- FUCHS, C. (1982). Maximum likelihood estimation and model selection in contingency tasks with missing data. *J. Amer. Statist. Assoc.* 65, 225-256.
- GIMOTTY, P.A. (1987). The asymptotic distribution of the goodness-of-fit chi-square statistics computed from imputed data. *Comm. Statist. Theor. Meth.*, 16(1), 45-60.
- KADANE, J.B. (1978). Some statistical problems in merging data files. In *1978 compendium of Tax Research*, Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: Government Printing Office, 159-171.
- KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- LE CAM, L.M. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Annals of Mathematical Statistics*, 41, 802-828.
- LITTLE, R.J.A. (1982). Models for non-response in sample surveys. *J. Amer. Statist. Assoc.*, 77, 237-250.
- LITTLE, R.J.A. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bulletin of the International Statistical Institute*, 51, 15.1, 1-17.
- LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.
- MOSTELLER, F. (1968). Association and estimation in contingency tables. *J. Amer. Statist. Assoc.* 63, 1-28.

- NORDHEIM, E.V. (1984). Inference from nonrandomly missing data : An example from a genetic study on Turner's Syndrome. *J. Amer. Statist. Assoc.* 79, 772-780.
- RAO, J.N.K. and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *J. Amer. Statist. Assoc.* 12, 1, 46-60.
- RODGERS, W.L. (1984). An evaluation of statistical matching. *J. Bus. Econ. Statist.*, 2, 91-102.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- RUBIN, D.B. (1978). Multiple imputations in sample surveys — a phenomenological Bayesian approach to nonresponse. *Proc. Surv. Res. Meth. Sec., Amer. Statist. Assoc.* 20-34.
- RUBIN, D.B. (1983). Conceptual issues in the presence of nonresponse. In *Incomplete Data in Sample Surveys*, Vol. 2 (W.G. Madow, I. Olkin, and D.B. Rubin, eds.), Academic, New York, 123-142.
- RUBIN, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 4, 87-94.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.
- RUBIN, D.B., SCHAFER, J.L., and SCHENKER, N. (1988). Imputation Strategies for estimating the undercount. *Proceedings of the Fourth Annual Research Conference*, U.S. Bureau of the Census, 151-159.
- SCHENKER, N. (1988). Handling missing data in coverage estimation, with application to the 1986 test of adjustment related operations. *Survey Methodology*, 14.1, 87-98.
- SIMS, C.A. (1978). Comments on Kadane's work on matching to create synthetic data. In *1978 Compendium of Tax Research*, Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 172-177.
- SINGH, A.C., (1988). On components of chi-square for log-linear categorical analysis. *Proceedings of the Section on Social Statistics, Amer. Statist. Assoc.* (forthcoming).
- SINGH, A.C., ARMSTRONG, J.B., and LEMAÎTRE, G.E., (1988). Statistical matching using log-linear imputation *Proceedings of the Section on Survey Research Methods, Amer. Statist. Assoc.* (forthcoming).
- STROUD, T.W.F. (1988). Personal communication.
- U.S. DEPARTMENT OF COMMERCE (1980). Report on exact and statistical matching techniques. Statistical Policy Working Paper 5, *Federal Committee on Statistical Methodology*.
- WOLFSON, M., GRIBBLE, S., BORDT, M., MURPHY, B. and ROWE, G. (1989). The Social Policy Simulation Database: an example of survey and administrative data integration. *Proceeding of the Symposium on Statistical Uses of Administrative Data*, Ottawa November 23-25, 1987 (J.W. Coombs and M.P. Singh, eds.), 201-229.

Ca 008

STATISTICS CANADA LIBRARY  
BIBLIOTHEQUE STATISTIQUE CANADA



1010248647

03