WP.43
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (vii): New and emerging methods

ON IMPUTATION OF BINARY VARIABLES IN REGISTERS

Invited paper

Prepared by Thomas Laitila, Statistics Sweden

I. Introduction

1. One strategy for dealing with missing data in surveys is to adjust the data set to fit estimators designed for the complete data situation, i.e. imputation is made for missing data. There is numerous imputation techniques suggested in the literature (see e.g. Sande, 1982; Rubin, 1996; Särndal and Lundström, 2005), classifiable according to which data set is used, if a parametric model is used or not, and whether randomization is used for selection of imputed value. These imputation techniques have been developed for application to sample surveys, where the object is to generalize sample survey estimates to a larger population.

2. For surveys based on registers, random imputation for qualitative variables is suggested by Wallgren and Wallgren (2007). Also Fiedler and Schodl (2008) apply random imputation for persons occupation and education in a test of a register based census. Multiple imputation in registers are used by Abowd et al. (2006) in construction of an infrastructure for statistics production.

3. This paper contributes with theoretical results on the properties of register survey estimators when registers contain random imputations for missing values. Such results are missing in the literature despite its essentiality for appropriate use and interpretation of derived statistics. An illustrating problem is the assumption of a correct imputation “model” by which random imputations are generated. However, if the model generates random imputation estimates which are unbiased, it is possible to calculate the true population value. A major concern is therefore the loss of precision in estimates by introducing randomness via imputation. Estimators below are named as random or deterministic imputation estimators depending on whether random or deterministic imputations are used for missing values.

4. The relative precisions of random imputation estimators are considered in the next section, where estimation of population totals are considered. Efficiency of random imputation estimators visa vie deterministic imputation estimators is dealt with in Section 3. Section 4 contains a treatment of estimation problems under random imputation of binary variables. Imputation of link variables for aggregation of register units into objects are considered in Section 5. A discussion of results is saved for the final section.

II. Precision

5. Consider a population U of N units, each with associated scalar, real valued variables y_k and x_k ($k \in U$). Interest is in estimation of the unknown population total $t_{yx} = \sum_U y_k x_k$ using register information, assuming available register information on y_k and x_k to be without errors. Let $U_R \subseteq U$ be the set of units with register information on both variables. For $U_y \subseteq U$ and $U_x \subseteq U$ register information is not available for y_k and x_k, respectively. Units without register information on both variables are collected in $U_{yx} \subseteq U$. These subsets are all disjoint and cover the whole population U.

6. For estimation missing values are replaced by imputed values derived from auxiliary non-random information. Imputations for y_k in U_y are denoted $\hat{y}_k(x_k)$ allowing for dependence on the known values of x_k. Similarly, imputations for x_k in U_x are denoted $\hat{x}_k(y_k)$. For U_{yx}, imputations are denoted as $\hat{y}_k$ and $\hat{x}_k(\hat{y}_k)$, where $\hat{x}_k(\hat{y}_k)$ is generated conditionally on $\hat{y}_k$. The imputation based estimator of t_{yx} is then written as

$$\hat{t}_{yx} = \sum\nolimits_{U_R} y_k x_k + \sum\nolimits_{U_y} \hat{y}_k(x_k) x_k + \sum\nolimits_{U_x} y_k \hat{x}_k(y_k) + \sum\nolimits_{U_{yx}} \hat{y}_k \hat{x}_k(\hat{y}_k) \qquad (1)$$

7. Let $z \sim (\mu_z, \sigma_z^2)$ denote that z is a random variable with mean μ_z and σ_z^2. The following assumptions are made regarding the imputations made:

A1: Imputations are independent among units and generated from known models.

A2: For units in U_y, $\hat{y}_k(x_k) \sim (\mu_{yk}(x_k), \sigma_{yk}(x_k)^2)$

A3: For units in U_x, $\hat{x}_k(y_k) \sim (\mu_{xk}(y_k), \sigma_{xk}(y_k)^2)$

A4: For units in U_{yx}, $\hat{y}_k \sim (\mu_{yk}, \sigma_{yk}^2)$ and $\hat{x}_k(\hat{y}_k) | \hat{y}_k \sim (\mu_{xk}(\hat{y}_k), \sigma_{xk}(\hat{y}_k)^2)$

Assumption A1 means that the imputations are generated from the register itself, from auxiliary register data, or some combination thereof. No information from probability samples are utilized for making imputations. Although non-parametric imputation methods, like nearest neighbor imputation, does not involve a formulation of a parametric model, it is still considered as drawing a number from a model. Drawing a value at random from a set of values, a set generated by forming a group of similar units as in Fiedler and Schodl (2008), implies drawing a number from a probability distribution.

8. Under assumptions A1-A4, the estimator $\hat{t}_{yx}$ has expected value and variance

$$E(\hat{t}_{yx}) = \sum\nolimits_{U_R} y_k x_k + \sum\nolimits_{U_y} \mu_{yk}(x_k) x_k + \sum\nolimits_{U_x} y_k \mu_{xk}(y_k) + \sum\nolimits_{U_{yx}} \lambda_k$$

$$V(\hat{t}_{yx}) = \sum\nolimits_{U_y} x_k^2 \sigma_{yk}(x_k)^2 + \sum\nolimits_{U_x} y_k^2 \sigma_{xk}(y_k)^2 + \sum\nolimits_{U_{yx}} \left(V_{\hat{y}_k}(\hat{y}_k \mu_{xk}(\hat{y}_k)) + E_{\hat{y}_k}(\hat{y}_k^2 \sigma_{xk}(\hat{y}_k)^2) \right)$$

where $\lambda_k = E_{\hat{y}_k}(\hat{y}_k \mu_{xk}(\hat{y}_k))$. Rewrite $\hat{t}_{yx}$ as $\hat{t}_{yx} = \sum_{U_R} y_k x_k + \sum_{U_{\bar{R}}} \tilde{y}_k \tilde{x}_k$, where $\tilde{y}_k = \hat{y}_k(x_k)$ and $\tilde{x}_k = x_k$ if $k \in U_y$, $\tilde{y}_k = y_k$ and $\tilde{x}_k = \hat{x}_k(y_k)$ if $k \in U_x$, $\tilde{y}_k = \hat{y}_k$ and $\tilde{x}_k = \hat{x}_k(\hat{y}_k)$ if $k \in U_{yx}$, and $U_{\bar{R}} = U_y \cup U_x \cup U_{yx}$. The variance of the estimator can then be expressed as

$$V(\hat{t}_{yx}) = \sum\nolimits_{U_{\bar{R}}} V(\tilde{y}_k \tilde{x}_k) = (N - N_R) \bar{\sigma}_{\bar{R}}^2$$

where $\bar{\sigma}_{\bar{R}}^2 = (N - N_R)^{-1} \sum_{U_{\bar{R}}} V(\tilde{y}_k \tilde{x}_k)$ and N_R is the number of units in U_R. The variance of the random imputation estimator is of order $N - N_R$, the number of units with missing values. Since the expected value of the estimator is of order N, the coefficient of variation

$$cv(\hat{t}_{yx}) = \frac{\sqrt{V(\hat{t}_{yx})}}{E(\hat{t}_{yx})} = \sqrt{\frac{1-f_R}{N}} \cdot \frac{\bar{\sigma}_{\bar{R}}}{E(\hat{\mu}_{yx})} \qquad (2)$$

is of order $N^{-1/2}(1-f_R)^{1/2}$, where $\hat{\mu}_{yx} = \hat{t}_{yx}/N$ and $f_R = N_R/N$.

9. Equation (2) shows that the coefficient of variation for the random imputation estimator will be small already at moderately large population sizes, unless the mean variance of random imputations is extremely large compared with the expected value of the population mean estimator $\hat{\mu}_{yx} = \hat{t}_{yx}/N$.

III. Efficiency

10. As an alternative to random imputation, deterministic imputation can be used. Suppose the model mean $\lambda_k = E_{\hat{y}_k}(\hat{y}_k \mu_{xk}(\hat{y}_k))$ and the model means in A2 and A3 are used for imputation instead of random draws from the distributions. This deterministic imputation estimator equals

$$\hat{t}_{yx}^{D} = \sum\nolimits_{U_R} y_k x_k + \sum\nolimits_{U_y} \mu_{yk}(x_k) x_k + \sum\nolimits_{U_x} y_k \mu_{xk}(y_k) + \sum\nolimits_{U_{yx}} \lambda_k \qquad (3)$$

Table 1: Relative efficiency of $\hat{t}_{yx}$ compared with $\hat{t}_{yx}^{D}$ for different levels of relative bias and coefficient of variation.

	Coefficient of Variation			
Relative bias	0.01	0.05	0.1	0.2
-0.2	0.998	0.962	0.862	0.610
-0.1	0.992	0.832	0.552	0.236
-0.05	0.965	0.526	0.217	0.065
-0.01	0.505	0.039	0.010	0.003
0.01	0.495	0.038	0.010	0.002
0.05	0.958	0.476	0.185	0.054
0.1	0.988	0.768	0.452	0.171
0.2	0.996	0.917	0.735	0.410

Comparing this estimator with the expectation of $\hat{t}_{yx}$ reveals that

$$\hat{t}_{yx}^{D} = E(\hat{t}_{yx})$$

and the biases of the two estimators are the same, i.e.

$$Bias(\hat{t}_{yx}) = Bias(\hat{t}_{yx}^{D}) = \sum\nolimits_{U_y} (\mu_{yk}(x_k) - y_k) x_k + \sum\nolimits_{U_x} y_k (\mu_{xk}(y_k) - x_k) + \sum\nolimits_{U_{yx}} (\lambda_k - y_k x_k)$$

Bias of the deterministic estimator can be defined by assigning a degenerate distribution with mass one at the point $\hat{t}_{yx}^{D}$.

11. Since the biases are the same, the relative efficiency, measured in MSE terms, of the randomized imputation estimator compared with the deterministic imputation estimator is

$$\text{Reff}\left(\hat{t}_{yx};\hat{t}_{yx}^{D}\right)=\frac{Bias(\hat{t}_{yx})^2}{V(\hat{t}_{yx})+Bias(\hat{t}_{yx})^2} \tag{4}$$

Letting the relative bias be denoted as $\beta=\frac{Bias(\hat{t}_{yx})}{t_{yx}}$ the relative efficiency can be rewritten as

$$\text{Reff}\left(\hat{t}_{yx};\hat{t}_{yx}^{D}\right)=\frac{\left(\beta/(1+\beta)\right)^2}{cv(\hat{t}_{yx})^2+\left(\beta/(1+\beta)\right)^2} \tag{5}$$

The relative efficiency decreases when the relative bias decreases, in absolute terms, and when the coefficient of variation increases. Expression (5) is generic in the sense it applies to all comparisons of a random estimator to its mean.

12. Table 1 presents calculated values of (5) for a few combinations of relative bias and coefficient of variation values. From the table it is seen that the loss in efficiency is substantial if the relative bias is small, e.g. around ±1%. A pattern in the table indicates that the loss in efficiency is large if the relative bias, in absolute terms, is smaller than the coefficient of variation.

13. This section is closed by a short note on the construction of confidence intervals (CI) based on the random imputation estimator $\hat{t}_{yx}$. The estimator has expected value $\hat{t}_{yx}^{D}$, which is a value which can be calculated. Constructing a CI of the standard form $\hat{t}_{yx}\pm 1.96\sqrt{\hat{V}(\hat{t}_{yx})}$ will produce a CI for $\hat{t}_{yx}^{D}$, not for the unknown population total t_{yx}. Thus, the source of uncertainty in the register survey estimator is the bias introduced by imputation, not variance, and a standard CI gives no information on the uncertainty of a register survey estimate.

IV. Imputation for a binary variable

14. Consider the case where y_k is a binary variable, $y_k\in\{0,1\}$, e.g. a domain indicator, and where x_k is a bounded positive variable, $x_k\in\{0,M\}$, $M<\infty$. Random imputations for y_k can be interpreted as being made from Bernoulli distributions and assumptions A2 and A4 are replaced by A2*: For units in U_y, $\hat{y}_k(x_k)\sim bern\left(\mu_{yk}(x_k)\right)$, 0<$\mu_{yk}(x_k)$<1 and A4*: For units in U_{yx}, $\hat{y}_k\sim bern\left(\mu_{yk}\right)$, $\hat{x}_k(\hat{y}_k)|\hat{y}_k\sim\left(\mu_{xk}(\hat{y}_k),\sigma_{xk}(\hat{y}_k)^2\right)$, 0<$\mu_{yk}$<1. The expected values for the units in U_{yx} are $\lambda_k=\mu_{yk}\mu_{xk}(1)$, whereby the expected value of the random imputation estimator is

$$E\left(\hat{t}_{yx}\right)=M\left(\sum_{U_R}y_kz_k+\sum_{U_y}\mu_{yk}(x_k)z_k+\sum_{U_x}y_kE(\hat{z}_k(1))+\sum_{U_{yx}}\mu_{yk}E(\hat{z}_k(1))\right)$$

where $z_k=x_k/M$ and $\hat{z}_k(1)=\hat{x}_k(1)/M$ are both bounded to the unit interval. The variance is

$$V\left(\hat{t}_{yx}\right)=\sum_{U_y}\mu_{yk}(x_k)(1-\mu_{yk}(x_k))x_k^2+\sum_{U_x}y_k\sigma_{xk}(1)^2+\sum_{U_{yx}}\left(\mu_{yk}(1-\mu_{yk})\mu_{xk}(1)^2+\mu_{yk}\sigma_{xk}(1)^2\right)$$

Since 0<$(1-\mu_{yk}(x_k))$<1 and 0<$(1-\mu_{yk})$<1 the variance is bounded by

$$V\left(\hat{t}_{yx}\right) < M^2\left(\sum_{U_y}\mu_{yk}(x_k)z_k^2 + \sum_{U_x} y_k V(\hat{z}_k(1)) + \sum_{U_{yx}}\mu_{yk}E(\hat{z}_k(1)^2)\right)$$

Now, $z_k^2 \le z_k$, $V(\hat{z}_k(1)) \le E(\hat{z}_k(1))$ and $E(\hat{z}_k(1)^2) \le E(\hat{z}_k(1))$ whereby

$$V(\hat{t}_{yx}) < M^2\left(\sum_{U_y}\mu_{yk}(x_k)z_k + \sum_{U_x} y_k E(\hat{z}_k(1)) + \sum_{U_{yx}}\mu_{yk}E(\hat{z}_k(1))\right) = M \cdot t_{yx\bar{R}}^D$$

where $t_{yx\bar{R}}^D = \left(\sum_{U_y}\mu_{yk}(x_k)x_k + \sum_{U_x} y_k\mu_{xk}(1) + \sum_{U_{yx}}\mu_{yk}\mu_{xk}(1)\right)$ is the sum of imputed values $\tilde{y}_k\tilde{x}_k$.

15. Defining $t_{yxR} = \sum_{U_R} y_k x_k$, the total of the observed values $y_k x_k$, the coefficient of variation can be bounded as

$$\frac{V(\hat{t}_{yx})}{E(\hat{t}_{yx})^2} = \frac{V(\hat{t}_{yx})}{(t_{yxR} + t_{yx\bar{R}}^D)^2} = \frac{V(\hat{t}_{yx})}{M \cdot t_{yx\bar{R}}^D} \cdot \frac{M \cdot t_{yx\bar{R}}^D}{(t_{yxR} + t_{yx\bar{R}}^D)^2} < \frac{M \cdot t_{yx\bar{R}}^D}{(t_{yxR} + t_{yx\bar{R}}^D)^2}$$

Letting $\bar{t}_{yxR} = t_{yxR}/N_R$, $\bar{t}_{yx\bar{R}}^D = t_{yx\bar{R}}^D/(N - N_R)$, and $f_{\bar{R}} = 1 - f_R$, this inequality can be rewritten as

$$cv(\hat{t}_{yx}) < \sqrt{\frac{1-f_R}{N}}\frac{\sqrt{M \cdot \bar{t}_{yx\bar{R}}^D}}{(f_R\bar{t}_{yxR} + f_{\bar{R}}\bar{t}_{yx\bar{R}}^D)}$$

Unless $\sqrt{M}$ is extremely large compared with the expected value of the population mean estimator $\hat{t}_{yx}/N$, the coefficient of variation will be small for moderately large sample sizes.

16. If the mean of imputed values equals the mean of the observed values, i.e. $\bar{t}_{yx\bar{R}}^D = \bar{t}_{yxR}$ then

$$cv(\hat{t}_{yx}) < \sqrt{\frac{1-f_R}{N}}\sqrt{\frac{M}{\bar{t}_{yxR}}} \qquad (6)$$

Inequality (6) is illustrated with the following two examples:

Example 1: In a survey of establishments, let y_k be an indicator for a domain, e.g. an industry sector, and let x_k be an indicator of the size category of an establishments whereby *M*=1. Let the population be of size *N*=10000 and suppose a register contains observations on both y_k and x_k for 7500 of the units in the population, i.e. $f_R = 0.75$. Suppose 400 of the 7500 units with complete observations in the register are included in the domain and of those 400, 20 units belong to the size category of interest. Then $\bar{t}_{yxR}$=20/7500 and $cv(\hat{t}_{yx})$ < 0.0968.

Example 2: In a survey of household income, let y_k be a domain indicator and let x_k denote household income. Suppose the population contains 50 000 households and auxiliary information restricts household income to be less than 2 million SEK, i.e. *M* = 2000000. Suppose register information on both y_k and x_k can be obtained for

80% of the households in the population and that $\bar{t}_{yxR}$ = 150000. Then $cv(\hat{t}_{yx})$ < 0.0073. If M is increased 100 times to 200 million SEK, the bound on the coefficient of variation increases to 0.073.

V. Imputation for link variables

17. Sometimes the register contains units which can be grouped into larger units, here called objects, e.g. persons can be grouped into households or establishments can be grouped into enterprises. For such an aggregation into objects, the register must contain link variables indicating which units belong to which objects. Here the imputations of such link variables are considered.

18. The population U of N units is assumed to be grouped into a population V of B objects. Let v_h denote the set of units in U included in object $h \in V$. For each unit $k \in U$ there is an indicator variable y_{hk} which equals one if the unit belongs to object $h \in V$, y_{hk} is zero otherwise. For some units the link variable is missing and replaced by imputations $\hat{y}_{hk}$. Let a_h denote a vector of variables with characteristics of the object which is not a function of the characteristics of the units, e.g. location and living space of a household. Similarly, let x_k denote a vector of characteristics of unit *k*, e.g. income. Both a_h and x_k are assumed to be known via register information.

19. Let Y_h denote a vector containing the indicators y_{hk} for all $k \in U$. Also let X denote a matrix containing the vectors x_k, $k \in U$. A household characteristic g_h can then be expressed as a function $g_h = g(Y_h, X, a_h)$. For domain statistics, introduce a domain indicator $I_h = 1(q(v_h))$ equaling one if the restriction $q(v_h)$ is satisfied by household *h* and equaling zero otherwise. The estimation problem considered here is the estimation of the domain mean

$$\mu_{Ig} = \frac{t_I}{B_I} = \frac{\sum_V I_h g_h}{\sum_V I_h}$$

using random imputations of some of the link variables y_{hk} .

20. Let $\hat{Y}_h$ denote a vector with some of the elements y_{hk} imputed. The household characteristic is then defined by the imputations and denoted as $\hat{g}_h = g(\hat{Y}_h, X, a_h)$. Similarly the household domain indicator $I_h = 1(q(v_h))$ may be defined by imputations whereby it is denoted as $\hat{I}_h = 1(q(\hat{v}_h))$. The random imputation estimator considered is

$$\hat{\mu}_{Ig} = \frac{\hat{t}_{Ig}}{\hat{t}_I} = \frac{\sum_{V_R} I_h g_h + \sum_{V_{\bar{R}}} \hat{I}_h \hat{g}_h}{\sum_{V_R} I_h + \sum_{V_{\bar{R}}} \hat{I}_h} \qquad (7)$$

where V_R contain objects with complete information on links to the units in the population U while $V_{\bar{R}}$ contain objects with one or several missing link variables y_{hk}.

Using the Hartley-Ross identity, the expression

$$cv(\hat{\mu}_{Ig})^2 = \left(1 - \frac{1}{E(\hat{t}_{Ig})} Cov\left(\frac{\hat{t}_{Ig}}{\hat{t}_I}, \hat{t}_I\right)\right)^{-2} V\left(\frac{\hat{t}_{Ig} / E(\hat{t}_{Ig})}{\hat{t}_I / E(\hat{t}_I)}\right) = \kappa \cdot V\left(\frac{\hat{t}_{Ig} / E(\hat{t}_{Ig})}{\hat{t}_I / E(\hat{t}_I)}\right)$$

is obtained for the coefficient of variation of the estimator (7). If $Cov(\hat{t}_{Ig}, \hat{t}_I) \geq 0$, a result obtained if e.g. $\hat{t}_{Ig}$ is positive regression dependent on $\hat{t}_I$ (Esary et al, 1967), a Taylor expansion gives the bound

$$cv(\hat{\mu}_{Ig})^2 \leq \kappa\left(cv(\hat{t}_{Ig})^2 + cv(\hat{t}_I)^2\right) \tag{8}$$

21. Depending on the definitions $g_h = g(Y_h, X, a_h)$ and $I_h = 1(q(v_h))$, the estimators $\hat{t}_{Ig}$ and $\hat{t}_I$ can be made up of sums of dependent variables whereby the results of earlier sections are not directly applicable. To bound the rhs of (8), a result on negative association (Kumar and Proschan, 1983) is utilized. Let $\hat{Y}_h$ denote Y_h with randomly imputed values and let $\hat{\psi}_h = \psi_h(\hat{Y}_h)$ be a scalar function defined on the random vector $\hat{Y}_h$, e.g. $\hat{\psi}_h = \hat{I}_h \hat{g}_h$. The following result is used below.

Result 5.1: If $\hat{\psi}_h$ and $\hat{\psi}_{h'}$, $h \neq h' \in V_{\bar{R}}$, are both non-decreasing (non-increasing) functions of $\hat{Y}_h$, then $Cov(\hat{\psi}_h, \hat{\psi}_{h'}) \leq 0$.

Proof: For a given object *h*, imputations in $\hat{Y}_h$ are independent. For a given unit *k* the imputations $\hat{y}_{1k}, ..., \hat{y}_{Bk}$ follows a one trial multinomial distributions. The set of random variables $\hat{Y}_{\bar{R}} = \left\{\hat{y}_{hk} : h \in V_{\bar{R}}, k \in U_{\bar{R}}\right\}$ is then negatively associated. The functions $\hat{\psi}_h$ and $\hat{\psi}_{h'}$, $h \neq h' \in V_{\bar{R}}$ are defined on disjoint subsets of $\hat{Y}_{\bar{R}}$ whereby the result follows from Property P5 in Kumar and Proschan (1983) when the functions are non-decreasing. Since $Cov(-\hat{\psi}_h, -\hat{\psi}_{h'}) = Cov(\hat{\psi}_h, \hat{\psi}_{h'})$ the same result is obtained for non-increasing functions.□

Result 5.1. implies that $V(\hat{t}_{Ig}) \leq \sum_{V_{\bar{R}}} V(\hat{I}_h \hat{g}_h)$ and $V(\hat{t}_I) \leq \sum_{V_{\bar{R}}} V(\hat{I}_h)$ if $I_h g_h$ and I_h are both positive (negative) functions of $\hat{y}_{hk}$ for all $k \in U$. These inequalities in combination with the results in Section 4 yields the bound

$$cv(\hat{\mu}_{Ig}) \leq \kappa^{1/2}\sqrt{\frac{1-f_{VR}}{B}}\left(\frac{G\cdot\bar{t}^{D}_{Ig\bar{R}}}{\left(f_{VR}\bar{t}_{IgR}+(1-f_{VR})\bar{t}^{D}_{Ig\bar{R}}\right)^{2}}+\frac{\bar{t}^{D}_{I\bar{R}}}{\left(f_{VR}\bar{t}_{IR}+(1-f_{VR})\bar{t}^{D}_{I\bar{R}}\right)^{2}}\right)^{1/2}$$

where f_{VR} is the proportion of objects in the set V_R, G is an upper limit for g_h, $\bar{t}_{IgR}$ and $\bar{t}_{IR}$ are averages of observed values $I_h g_h$ and I_h, respectively, and $\bar{t}^{D}_{Ig\bar{R}}$ and $\bar{t}^{D}_{I\bar{R}}$ are means of expected values of $\hat{I}_h\hat{g}_h$ and $\hat{I}_h$, respectively. If $\bar{t}_{IgR}=\bar{t}^{D}_{Ig\bar{R}}$ and $\bar{t}_{IR}=\bar{t}^{D}_{I\bar{R}}$ the inequality simplifies to

$$cv(\hat{\mu}_{Ig}) \leq \kappa^{1/2}\sqrt{\frac{1-f_{VR}}{B}}\left(\frac{G}{\left(\bar{t}_{IgR}\right)^{2}}+\frac{1}{\left(\bar{t}_{IR}\right)^{2}}\right)^{1/2}$$

VI. Discussion

22. Random imputation is a practical way for producing a “full” register which later can be used for different purposes. The system of register and sample survey data files described in Abowd et al. (2006) provide with an illustrating example. When it comes to estimation of parameters defined as sums of population unit values using register data, the theoretical results partly give and partly do not give support to the use of random imputation. In terms of relative precision, already at moderate population sizes, the variance introduced by random imputation is small compared to the population total estimate. However, for a given random imputation technique, there is a corresponding deterministic imputation technique where the expected values of the specified distributions are imputed instead of random draws. The loss in efficiency compared with this deterministic imputation estimator can be substantial if random imputation is used. Results show that the smaller bias in the deterministic imputation estimator, the larger the loss in efficiency for the random imputation estimator.

23. The use of random imputation for constructing a full register data matrix can also be deceptive. The variance introduced is negligible in relative terms for moderately large population sizes, a consequence of the laws of large numbers. For small populations, the variance can be of a much larger magnitude. Such cases can occur in estimation of parameters in small domains, an estimation problem the register manager may not be in control over as the register is constructed for general application.

24. Another topic considered in the paper is imputation of link variables used for aggregation of population units to larger objects. Here the characteristic of the object considered is general and may constitute sums of object characteristics defined by non-linear functions of the link variables, e.g. per person household living space. For this estimation problem, the picture is a little bit different regarding the precision of the random imputation estimator. In terms of relative precision, the bound derived on the coefficient of variation shows on small coefficient of variations for moderately large populations. Again this result may not hold for small domains.

25. Regarding relative efficiency, the bias of the deterministic imputation estimator is generally not the same as the bias of the random imputation estimator. In the case when the model underlying the random imputations is correct, the random imputation estimator gives unbiased estimates. This is not generally obtained if imputations are deterministically made using the expected values instead of random draws. However, imputations of link variables imply imputations of object characteristics. A way of performing approximate deterministic imputation can be achieved by repeated random imputations of link variables, as in multiple

imputation, and imputing averages of the corresponding repeated household characteristics. The relative efficiency of the random imputation estimator in relation to this deterministic imputation estimator is expected to be similar to the results in Section 3.

26. In summary, results presented indicate that random imputation should be avoided for deriving full registers to be used for general applications. There is a loss in efficiency which can be substantial if the imputation method yields small bias, and if the population is small the random imputation estimator may be with both low efficiency and low relative precision. Imputations using the expected values in the models for random imputations are a better alternative. For continuous and binary variables such an imputation method causes no practical problems. For categorical variables in general, this imputation method causes a data storage problem. The expected value, i.e. the probability, for each category level must be stored in the register. On the other hand, the storage of expected values for category variables, e.g. link variables, increases the information available for sound statistical inference.

References

Abowd, J.M, Stephens, B.E, Vilhuber, L, Andersson, F, McKinney, K.L; Roemer, M. and S. Woodcock (2006). The LEHD infrastructure files and the creation of the quarterly workforce indicators, Technical Paper No. TP-2006-01, US Census Bureau.

Esary, J.-D, Proschan, F. and D.W. Walkup (1967). Association of Random Variables, with applications, *Annals of Mathematical Statistics*, **44**, 1466-1474.

Fiedler, R. and P. Schodl (2008). Data Imputation and Estimation for the Austrian Register-Based Census, Tech. rep. UN/ECE Work Session on Data Editing, Vienna, Austria, April 21-23.

Kumar, J.-D. and F. Proschan (1983). Negative Association of Random Variables with Applications, *The Annals of Statistics*, **11:1**, 286-295.

Rubin, D.B. (1996). *Multiple Imputation After 18+ Years, Journal of the American Statistical Association,* **91**, 473-489.

Sande, I.G. (1982). Inputation in Surveys: Coping with Reality, *The American Statistician*, **36:3**, 145-152.

Särndal, C.-E., Swensson, B. och J. Wretman (1992). Model Assisted Survey Sampling, Springer, New York.

Särndal, C.-E. and S. Lundström (2005). *Estimation in Surveys with Nonresponse*, Wiley, Chichester, England.

Wallgren, A. och B. Wallgren (2007). *Register-based Statistics*, Wiley, Chichester.