WP. 23

ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and**
**ECONOMIC COMMISSION FOR EUROPE**
**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**
(Madrid, Spain, 20-22 October 2003)

Topic (ii): Developments related to new methods and techniques

# IMPUTATION OF MISSING DATA ITEMS UNDER LINEAR RESTRICTIONS

**Supporting Paper**

Submitted by Statistics Netherlands[1]

**ABSTRACT:** Item nonresponse is a problem that is often encountered when dealing with economic data. Imputation is a popular strategy to handle item nonresponse. However, imputations obtained by common imputation methods, such as hot deck and regression imputation, mostly do not satisfy the balance edit constraints imposed on economic data. In this paper we therefore propose to use the Dirichlet distribution in order to impute missing items while satisfying balance edits.

## I. INTRODUCTION

1. Economic data consist of many logical constraints on the data items, such as the fact that company profits must equal turnover minus expenses. Commonly used imputation methods such as hot deck and (random) regression imputation mostly do not provide imputations that satisfy these constraints. In this paper we therefore suggest the use of another imputation scheme to obtain imputations that will satisfy these linear restrictions. We will first discuss the distribution used to model these economic data, which is the Dirichlet distribution. Next random number generation and parameter estimation will be treated as well as the EM algorithm. Finally, some preliminary results and areas of future research will be discussed.

2. Missing data is a prevalent problem in survey analysis. At Statistics Netherlands imputation is often used to estimate and fill in missing data items, since missing values result in less efficient estimates due to the reduction of the sample size of the dataset. Besides, if the nonrespondents are judged to be significantly different from the respondents on the basis of auxiliary information, imputation reduces the nonresponse bias. Finally, imputation is applied because standard complete data methods such as regression analysis cannot immediately be used to analyse data when items are missing.

3. Several imputation methods have been developed, see for an overview of the methods that are frequently used, for example, Kalton and Kasprzyk (1986). Imputation methods can be either deterministic or stochastic. Deterministic methods determine imputed values uniquely, this means that when the imputation process is repeated the same value will be imputed. Stochastic methods depend on some sort of randomness, which means that when the process is repeated, other values may be imputed. Deterministic imputation methods avoid the loss in precision associated with the added randomness as opposed to stochastic methods. Therefore these methods are well suited to estimate means or totals.

[1] Prepared by Caren Tempelman, DTMN@cbs.nl.

However, the variance will be underestimated and the shape of the distribution will be distorted. So for the creation of general purpose datasets, stochastic imputation is preferred.

## II. THE STATISTICAL DISTRIBUTION OF ECONOMIC DATA

4. Examining realistic data from Statistics Netherlands leads to the conclusion that the data are rarely normally distributed and that they are mostly very skew. Distributions for models are often chosen on the basis of the range within which the random variable is constrained to vary. For a variable constrained between zero and one the beta distribution has proved useful.

5. The beta distribution is defined by the probability density function (pdf)

$$f(x\,|\,\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1},\ \ 0<x<1,\ \ \alpha,\beta>0$$

where $\Gamma(\cdot)$ is the gamma function defined by $\Gamma(x)=\int_0^\infty u^{x-1}e^{-u}du$ .

6. This functional form is extremely flexible in the shapes it will accommodate. It is symmetric if $\alpha=\beta$ and asymmetric otherwise. Besides it can be hump-shaped or U-shaped. Note that it reduces to the uniform distribution if $\alpha=\beta=1$. Also note that the beta distribution is symmetrical, that is if $X\sim\text{beta}(\alpha,\beta)$ then $1-X\sim\text{beta}(\beta,\alpha)$. An extension of the beta distribution is the so-called Dirichlet distribution, also referred to as the multivariate beta. Its pdf is

$$f(x_1,\ldots,x_k\,|\,\alpha_1,\ldots,\alpha_k)=\frac{\Gamma(\sum_{i=1}^k\alpha_i)}{\prod_{i=1}^k\Gamma(\alpha_i)}\prod_{i=1}^k x_i^{\alpha_i-1} \tag{1}$$

where

$$x_i\geq 0,\ \ \alpha_i>0,\ \ i=1,\ldots,k,\ \ x_k=1-\sum_{i=1}^{k-1}x_i\ .$$

We shall refer to the pdf of Dirichlet distribution given by (1) with $\text{Dir}_{k-1}(\alpha_1,\ldots,\alpha_k)$. Note that because $\sum_{i=1}^k x_i=1$, this is actually an *(k-1)*-dimensional distribution, since $x_k$ is redundant and can be replaced by $1-\sum_{i=1}^{k-1}x_i$ . Consequently, the pdf is sometimes written as

$$f(x_1,\ldots,x_{k-1}\,|\,\alpha_1,\ldots,\alpha_k)=\frac{\Gamma(\sum_{i=1}^k\alpha_i)}{\prod_{i=1}^k\Gamma(\alpha_i)}\prod_{i=1}^{k-1}x_i^{\alpha_i-1}(1-\sum_{i=1}^{k-1}x_i)^{\alpha_k-1}$$

7. The Dirichlet is a convenient distribution on the simplex: it is an exponential family and has finite sufficient statistics. The first and second order moments of the Dirichlet distribution are

$$\text{E}(X_i)\ =\frac{\alpha_i}{\alpha},\qquad\qquad i=1,\ldots,k$$

$$\text{Var}(X_i)\ =\frac{\alpha_i(\alpha-\alpha_i)}{\alpha^2(\alpha+1)},\ \ i=1,\ldots,k$$

where $\alpha=\sum_{j=1}^k\alpha_j$ .

The covariances of the *X*'s are

$$\text{Cov}(X_i,X_j)=\frac{-\alpha_i\alpha_j}{\alpha^2(\alpha+1)},\quad i,j=1,\ldots,k,\ i\neq j$$

8. The following theorems apply.

**Theorem 1 (Marginal Dirichlet)**
*If* $(X_1,\ldots,X_k)$ *is a random variable vector having the (k-1)-variate Dirichlet distribution* $\mathrm{Dir}_{k-1}(\alpha_1,\ldots,\alpha_k)$, *then the marginal distribution of* $(X_1,\ldots,X_{k_1})$, $k_1 < k$ *is the* $k_1$*-variate Dirichlet distribution* $\mathrm{Dir}_{k_1}(\alpha_1,\ldots,\alpha_{k_1},\alpha_{k_1+1}+\cdots+\alpha_k)$.

**Theorem 2 (Conditional Dirichlet)**
*If* $\mathbf{X}=(\mathbf{X}_1' \ ,\mathbf{X}_2' \ )' \sim \mathrm{Dir}_{k-1}(\boldsymbol{\alpha}_1' \ ,\boldsymbol{\alpha}_2' \ )$ *where* $\mathbf{X}_1$ *and* $\boldsymbol{\alpha}_1$ *consist of* $r$ *elements and* $\mathbf{X}_2$ *and* $\boldsymbol{\alpha}_2$ *consist of* $s$ *elements and* $k=r+s$, *then*

$$(1-\mathbf{X}_2' \ \mathbf{1})^{-1}\mathbf{X}_1 \mid \mathbf{X}_2,\boldsymbol{\alpha} \sim \mathrm{Dir}_{r-1}(\boldsymbol{\alpha}_1' \ )$$

See for a proof of these theorems Wilks (1962).

## III. IMPUTATION OF MULTIVARIATE MISSING ITEM VALUES

### A. The edit constraints

9. In general there are two types of linear edit constraints: balance and inequality edits.

$$c_1X_1+\cdots+c_kX_k \ = X_{k+1} \qquad (2)$$
$$c_1X_1+\cdots+c_kX_k \ \le X_{k+1}$$

We will first consider balance edit constraints. We believe that we can impute the missing items, satisfying the edit constraint directly and preserving the distribution of the data, by making use of the Dirichlet distribution.

10. Consider edit rule (2) and transform it by dividing the different parts by the total $X_{k+1}$, this is done in order to restrict the domain of the variables $\tilde{X}_1,\ldots,\tilde{X}_k$ to the simplex; we take $\tilde{X}_i = c_iX_i / X_{k+1}$.

$$\frac{c_1X_1}{X_{k+1}}+\cdots+\frac{c_kX_k}{X_{k+1}} \ = 1, \ \ X_{k+1} > 0$$
$$\tilde{X}_1+\cdots+\tilde{X}_k \qquad = 1$$

11. Note that we assume that the total, $X_{k+1}$, is known. This is done for two reasons. First of all since it is an aggregate the nonresponse rate will probably be low. And secondly, if it is indeed missing we expect to be able to estimate this value very well based on the other variables in the survey, whereas the different subtotals are far more difficult to estimate this way.

### B. Imputation

12. A special case arises when only one $\tilde{X}_i$, $i=1,\ldots,k$ is missing. In this instance deductive imputation can be used. Deductive imputation means that the value of the missing item can be established with certainty based on the other items in the survey.

13. If all items of $\tilde{\mathbf{X}}$ are missing, we can obtain imputations by drawing from $\mathrm{Dir}_{k-1}(\alpha_1,\ldots,\alpha_k)$. However, a common circumstance is that a few item values are missing and the others are observed. In this case one needs to draw values from the conditional distribution of the missing items given the observed ones, which is also a Dirichlet distribution as was established in Theorem 2.

14. Partition $\tilde{\mathbf{X}}$ in $\tilde{\mathbf{X}}^{mis}$ and $\tilde{\mathbf{X}}^{obs}$, where $\tilde{\mathbf{X}}^{mis}$ represents the missing items and $\tilde{\mathbf{X}}^{obs}$ represents the observed items. The vector with missings, $\tilde{\mathbf{X}}^{mis}$, consists of $m$ elements and $\tilde{\mathbf{X}}^{obs}$ consists of $o$ elements, which are the number of observed items and $k = m + o$. Partition $\boldsymbol{\alpha}$ accordingly. Then it holds that

$$(1 - \mathbf{1'}\tilde{\mathbf{X}}^{obs})^{-1}\tilde{\mathbf{X}}^{mis} \mid \tilde{\mathbf{X}}^{obs}, \boldsymbol{\alpha} \sim \mathrm{Dir}_{m-1}(\alpha_1,\ldots,\alpha_m) \qquad (3)$$

Thus imputations for missing items can be obtained by drawing from the conditional Dirichlet distribution mentioned in (3).

## IV. RANDOM NUMBER GENERATION

15. In order to impute missing values we need to generate random values from the Dirichlet distribution. This can be done as follows. Recall that when $U_1 \sim \mathrm{gamma}(\alpha,\lambda)$ and $U_2 \sim \mathrm{gamma}(\beta,\lambda)$, then $Z = \frac{U_1}{U_1+U_2} \sim \mathrm{beta}(\alpha,\beta)$. This can be generalised to the Dirichlet distribution, see for example Wilks (1962). Suppose $U_1,\ldots,U_k$ are independent random variables having gamma distributions: $\mathrm{gamma}(\alpha_1,\lambda),\ldots,\mathrm{gamma}(\alpha_k,\lambda)$. Let $Z_i = \frac{U_i}{U_1+\cdots+U_k}$, for $i = 1,\ldots,k$. Then $Z_1,\ldots,Z_k$ has the *(k-1)*-variate Dirichlet distribution $\mathrm{Dir}_{k-1}(\alpha_1,\ldots,\alpha_k)$. Thus random values from the Dirichlet distribution can be obtained by drawing independently from gamma distributions.

## V. PARAMETER ESTIMATION

### A. The method of moments estimator

16. The parameters $\alpha_1,\ldots,\alpha_k$ can be estimated by a method of moments estimator. The method of moments estimator is consistent. Recall that the first and second order moments of the Dirichlet distribution are

$$\mu_i = \frac{\alpha_i}{\alpha}, \qquad i = 1,\ldots,k \qquad (4)$$

$$\sigma_i^2 = \frac{\alpha_i(\alpha - \alpha_i)}{\alpha^2(\alpha+1)}, \quad i = 1,\ldots,k \qquad (5)$$

where $\alpha = \sum_{j=1}^{k}\alpha_j$. Rewrite (4) as $\alpha = \frac{\alpha_i}{\mu_i}$ and fill this in in equation (5). Then

$$\sigma_i^2 = \frac{\alpha_i\left(\frac{\alpha_i}{\mu_i} - \alpha_i\right)}{\left(\frac{\alpha_i}{\mu_i}\right)^2\left(\frac{\alpha_i}{\mu_i} + 1\right)}$$

$$\left(\frac{\alpha_i}{\mu_i} + 1\right)\sigma_i^2 = (1 - \mu_i)\mu_i$$

Solving for $\alpha_i$ gives

$$\alpha_i = \mu_i \left( \frac{\mu_i}{\sigma_i^2}(1-\mu_i) - 1 \right), \quad i = 1, \ldots, k$$

So

$$\hat{\alpha}_{MM,i} = \hat{\mu}_i \left( \frac{\hat{\mu}_i}{\hat{\sigma}_i^2}(1-\hat{\mu}_i) - 1 \right), \quad i = 1, \ldots, k$$

17. Although the method of moments is straightforward, estimation based on the method of moments generally is not statistically efficient. That is, the asymptotic variance-covariance matrix of estimates is usually larger that the inverse of the information matrix. However, estimation based on the method of moments can serve as an excellent initial guess to start iterations in the Newton-Raphson algorithm, which we use to maximise the likelihood.

## B. Maximum likelihood estimation

18. In order to find a consistent estimator that is statistically efficient, maximum likelihood estimation can be applied. When sampling from a distribution that is a member from an exponential family of distributions, the maximum likelihood estimators will be a function of the sufficient statistics. The likelihood is defined as follows

$$L(\boldsymbol{\alpha} \mid \mathbf{X}) = \prod_{j=1}^{n} f(\mathbf{X}, \boldsymbol{\alpha})$$

19. The joint density of $X_1, \ldots, X_k$ is given by (1). Then the likelihood will be

$$L(\boldsymbol{\alpha} \mid \mathbf{X}) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)^n}{\prod_{i=1}^{k} \Gamma(\alpha_i)^n} \prod_{j=1}^{n} \prod_{i=1}^{k} X_{ji}^{\alpha_i - 1}$$

20. Taking the natural logarithm leads to the following loglikelihood function

$$l(\boldsymbol{\alpha} \mid \mathbf{X}) = \ln L(\boldsymbol{\alpha} \mid \mathbf{X}) = n \ln \Gamma(\sum_{i=1}^{k} \alpha_i) - n \sum_{i=1}^{k} \ln \Gamma(\alpha_i) + \sum_{j=1}^{n} \sum_{i=1}^{k} (\alpha_i - 1) \ln X_{ji}$$

Taking the first derivative results in

$$\frac{\partial l(\boldsymbol{\alpha} \mid \mathbf{X})}{\partial \alpha_t} = n \frac{\partial \ln \Gamma(\sum_{i=1}^{k} \alpha_i)}{\partial \alpha_t} - n \frac{\partial \ln \Gamma(\alpha_t)}{\partial \alpha_t} + \sum_{j=1}^{n} \ln X_{jt}, \quad t = 1, \ldots, k$$

This leads to

$$g_t(\boldsymbol{\alpha} \mid \mathbf{X}) = \frac{\partial l(\boldsymbol{\alpha} \mid \mathbf{X})}{\partial \alpha_t} = n \Psi(\sum_{i=1}^{k} \alpha_i) - n \Psi(\alpha_t) + \sum_{j=1}^{n} \ln X_{jt}, \quad t = 1, \ldots, k \tag{6}$$

where $\Psi(\cdot)$ is the digamma function defined by $\Psi(x) = \frac{\partial \ln \Gamma(x)}{\partial x} = \frac{\Gamma'(x)}{\Gamma(x)}$.

21. We need some iterative scheme to solve the equation $g_t(\boldsymbol{\alpha} \mid \mathbf{X}) = 0, \ t = 1, \ldots, k$. A commonly used method is the Newton-Raphson method. Determine an initial value for $\boldsymbol{\alpha}$, for example by means of the method of moments estimator. To find the $\boldsymbol{\alpha}_{MLE}$ that solves $\mathbf{g}(\boldsymbol{\alpha} \mid \mathbf{X}) = \mathbf{0}$, calculate by iteration and until convergence

$$\boldsymbol{\alpha}^{new} = \boldsymbol{\alpha}^{old} - \mathbf{H}^{-1}\mathbf{g}$$

where $\mathbf{H}$ is the Hessian, the matrix of second-derivatives of $l$ given by

$$\frac{\partial^2 l(\boldsymbol{\alpha}\,|\,\mathbf{X})}{\partial \alpha_t^2} = n\Psi'(\sum_{i=1}^{k}\alpha_i) - n\Psi'(\alpha_t),\ t = 1,\ldots,k$$
$$\frac{\partial^2 l(\boldsymbol{\alpha}\,|\,\mathbf{X})}{\partial \alpha_t \partial \alpha_s} = n\Psi'(\sum_{i=1}^{k}\alpha_i), \qquad t,s = 1,\ldots,k,\ t \neq s$$

where $\Psi'$ is known as the trigamma function.

22. Under some regularity conditions the likelihood function is strictly concave for exponential families, and the MLE exists and is unique, since the Dirichlet distribution is an exponential family this holds true for the Dirichlet distribution. A direct proof has been given by Ronning (1989).

23. However, when we are faced with missing item values $\mathbf{X}$ is not completely observed and the loglikelihood cannot be calculated directly. In order to estimate the maximum likelihood estimates when some of the variables involved are not observed the EM algorithm has been developed.

## C. The Expectation-Maximization (EM) algorithm

24. The EM algorithm (Dempster, Laird and Rubin, 1977) is a popular tool in statistics. The EM algorithm is based on the assumption that the missing data $\mathbf{X}^{mis}$ and the parameters, in this case $\alpha_1,\ldots,\alpha_k$, are interdependent. The intuition behind the EM is that the $\mathbf{X}^{mis}$ will be filled in based on $\mathbf{X}^{obs}$ and an initial estimate of $\boldsymbol{\alpha}$, next $\boldsymbol{\alpha}$ will be re-estimated based on $\mathbf{X}^{obs}$ and the filled in $\mathbf{X}^{mis}$. This process will be iterated until the estimates converge. The idea is that we would like to maximize the complete data likelihood but since we do not know it, we maximize its expectation instead. The EM algorithm consists of two steps, an expectation step and a maximization step.

25. *Expectation step*

For each iteration $j$ compute $l_j(\boldsymbol{\alpha}) = E(l(\boldsymbol{\alpha})\,|\,\mathbf{X}^{obs}, \boldsymbol{\alpha}^{j-1})$, where $\boldsymbol{\alpha}$ is the complete data loglikelihood and the expectation is taken with respect to the conditional distribution of the missing data given the observed data and the parameter $\boldsymbol{\alpha}^{j-1}$. In the case of distributions from an exponential family we only need to calculate the expected sufficient statistics.

26. *Maximization step*

Now that we have $l_j(\boldsymbol{\alpha})$, we can calculate the maximum likelihood estimates based on the complete data loglikelihood, and thus re-estimate $\boldsymbol{\alpha}$.

27. In order to apply the EM algorithm when the data are Dirichlet distributed the expected sufficient statistics need to be calculated, since the Dirichlet is an exponential family. These values can be easily computed from the natural parameterisation of the exponential family representation of the Dirichlet distribution. Recall that a distribution is an exponential family if it can be written in the form:

$$p(\mathbf{X}\,|\,\boldsymbol{\theta}) = g(\mathbf{X})\exp\{\boldsymbol{\theta}' T(\mathbf{X}) - A(\boldsymbol{\theta})\}$$

where $\boldsymbol{\theta}$ is the natural parameter, $T(\mathbf{X})$ is the sufficient statistic, and $A(\boldsymbol{\theta})$ is the natural logarithm of the normalization factor. The density function of the Dirichlet (equation (1)) can be written in this form as follows

$$p(\mathbf{X}\,|\,\boldsymbol{\alpha}) = \exp\{\ln\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\ln\Gamma(\alpha_i) + \sum_{i=1}^{k}(\alpha_i - 1)\ln X_i\}$$

28. The natural parameter of the Dirichlet is therefore $\theta_i = \alpha_i - 1$ and the sufficient statistic is $T(X_i) = \ln X_i$, $i = 1,\dots,k$ . Using the general fact that the derivative of the ln normalization factor with respect to the natural parameter is equal to the expectation of the sufficient statistic, we obtain

$$E[\ln X_i \mid \boldsymbol{\alpha}] = \Psi(\alpha_i) - \Psi(\sum_{j=1}^{k} \alpha_j), \quad i = 1,\dots,k$$

29. The expectation of the sufficient statistic conditional on the observed values can also be easily calculated since the conditional distribution of Dirichlet distributed variables is also a Dirichlet distribution (see Theorem 2). Consider $\mathbf{X} = (X_1,\dots,X_k)'$ which is *(k-1)*-variate Dirichlet distributed with parameters $\alpha_1,\dots,\alpha_k$. The missing $X$ 's are represented by the vector $\mathbf{X}^{mis} = (X_1,\dots,X_m)'$ and the observed $X$ 's are represented by $\mathbf{X}^{obs} = (X_{m+1},\dots,X_k)'$. Then

$$\hat{\mathbf{X}}^{mis} \mid \mathbf{X}^{obs}, \boldsymbol{\alpha} \sim \mathrm{Dir}_{m-1}(\alpha_1,\dots,\alpha_m)$$

where

$$\hat{\mathbf{X}}^{mis} = (1 - \mathbf{1}'\mathbf{X}^{obs})^{-1} \mathbf{X}^{mis}$$

30. It follows that

$$\mathrm{E}[\ln \hat{X}_i^{mis} \mid \mathbf{X}^{obs}, \boldsymbol{\alpha}] = \Psi(\alpha_i) - \Psi(\sum_{j=1}^{m} \alpha_j), \qquad i = 1,\dots,m$$

$$\mathrm{E}[\ln \frac{X_i^{mis}}{1 - \mathbf{1}'\mathbf{X}^{obs}} \mid \mathbf{X}^{obs}, \boldsymbol{\alpha}] = \Psi(\alpha_i) - \Psi(\sum_{j=1}^{m} \alpha_j), \qquad i = 1,\dots,m$$

$$\mathrm{E}[\ln X_i^{mis} \mid \mathbf{X}^{obs}, \boldsymbol{\alpha}] = \ln(1 - \mathbf{1}'\mathbf{X}^{obs}) + \Psi(\alpha_i) - \Psi(\sum_{j=1}^{m} \alpha_j), \quad i = 1,\dots,m$$

Plug this value into equation (6) for the missing items, and calculate the parameter estimates.

31. For those instances that a solution to the M-step does not exist in closed form a generalized EM algorithm (GEM) has been developed. Generalizations appear by strictly increasing the complete data loglikelihood, $l_{j+1}(\boldsymbol{\alpha}^{j+1}) \geq l_j(\boldsymbol{\alpha}^j)$, rather than maximising it. This could, for example, be achieved by calculating only one Newton Raphson step.

## VI. SOME PRELIMINARY RESULTS

32. In this section we want to compare the method described above with several imputation methods that are frequently used at Statistics Netherlands. This comparison will be carried out using a realistically generated dataset based on actual data from Statistics Netherlands, in which missing values will be created and imputed.

### A. Missing data mechanism

33. The missing data mechanism concerns the reasons why values are missing, and in particular whether these reasons relate to values in the data set. Any analysis of data involving item nonresponse requires some assumption about the missing data mechanism. In our case we want to compare the different imputation methods given a certain missing data mechanism. Because of this and the fact that it concerns a preliminary investigation of the proposed method, we will assume that the data are missing completely at random (MCAR), that is the probability that an item is missing does not depend on the other items in the dataset and is equal for all items.

## B. Imputation methods

34. The imputation methods that we apply to the missing data are

- *Mean imputation (MI),*
  The respondent mean is imputed for each missing item. Next these imputed values will be proportionally adjusted in order to satisfy the edit constraints.
- *Nearest neighbour ratio hot deck (NNHD),*
  The missing items will be imputed using a complete donor record that is most similar to the record with missings. This donor will be found by means of a Euclidian distance measure, based on the responses that were observed. Next the ratios between items of this donor, rather than the reported items, are used for imputation in order to immediately satisfy the edit constraint.
- *Random ratio hot deck (RHD),*
  In this case a donor will be chosen randomly and again the ratios between items of this donor are used for imputation.
- *Dirichlet imputation (DIR),* the aforementioned method.

35. Because some of these imputation methods are stochastic (RHD and DIR) we apply those *50* times in order to rule out differences occurring due to the random nature of the method.

## C. Results

36. The dataset used consists of *12* subtotals regarding business expenses, which add up to the total business expenses reported. The number of records, *n*, is *200*. We assume that 30 percent of the items are missing. The missing items are generated by means of independent draws from the Bernouilli distribution with parameter $p = 0.3$.

37. In order to correct for differences occurring due to the fact that the nonresponse is random we repeat this process several *(25)* times. In this way we can judge the methods independently of the way the nonresponse occurred given a certain missing data mechanism.

38. The performance of the imputation methods is measured by means of three different measures describing location and shape of the distribution. First of all for each imputed dataset we compute the average relative absolute deviation from the actual mean. That is:

$$\text{average relative absolute deviation from the actual mean } = \frac{1}{12}\sum_{i=1}^{12}\left|(t_i^{imp} - t_i^{true})/t_i^{true}\right|,$$

where $t_i^{imp}$, $i = 1,\ldots,12$, is the aggregate over all records for the imputed dataset, and $t_i^{true}$ is the aggregate for the actual dataset. The deviation from the actual variance is calculated similarly:

$$\text{average relative absolute deviation from the actual variance} = \frac{1}{12}\sum_{i=1}^{12}\left|(v_i^{imp} - v_i^{true})/v_i^{true}\right|,$$

where $v_i^{imp}$, $i = 1,\ldots,12$, is the variance of variable $i$ after imputation, and $v_i^{true}$ is the true variance of variable $i$. Finally, to assess the differences in distribution between the imputed and actual dataset, the Kolmogorov Smirnov statistic is calculated, which is based on the greatest absolute vertical distance between the distribution functions.

39. The results of these statistics are summarized in Table 1. Obviously mean imputation will lead to the smallest deviation from the true mean. The other methods seem to perform equally well.

40. With regard to the relative deviation from the true variance we observe somewhat unusual behaviour. One would expect the MI method to have the largest deviations from the true variance. Probably this is not the case due to the fact that the imputations are adjusted after imputation, and

therefore attribute to the calculated variance. Besides, since the variance of this measure is relatively large it is hard to draw any conclusions from this measure. The variance is mostly large due to the fact that there are one or two large deviations from the actual variance, which also strongly influences the mean deviation from the variance. Therefore more research should be done and other measures should be applied to investigate the ability of these methods to preserve the true variance.

41. Finally, the average Kolmogorov Smirnov statistic behaves as expected. The lower the statistic, the closer the distributions are. The Dirichlet method results in an imputed dataset that describes the distribution of the actual dataset best.

| Relative deviation from the actual mean | | |
|---|---|---|
| | Average | Variance |
| MI | 0.11 | 0.03 |
| NNHD | 0.15 | 0.06 |
| RHD | 0.15 | 0.08 |
| DIR | 0.15 | 0.07 |
| Relative deviation from the actual variance | | |
| | Average | Variance |
| MI | 0.33 | 0.90 |
| NNHD | 0.47 | 3.38 |
| RHD | 0.58 | 5.60 |
| DIR | 0.63 | 6.80 |
| Kolmogorov Smirnov statistic | | |
| | Average | Variance |
| MI | $6.32*10^{-2}$ | $7.15*10^{-4}$ |
| NNHD | $4.44*10^{-2}$ | $3.56*10^{-4}$ |
| RHD | $4.31*10^{-2}$ | $3.31*10^{-4}$ |
| DIR | $3.70*10^{-2}$ | $1.81*10^{-4}$ |

**Table 1**. Comparison of the four imputation techniques based on three different performance measures.

## VII. CONCLUSIONS

42. The process of adjusting imputations in order to satisfy edit constraints may seriously distort the distribution of the imputed values. Using a method that imputes satisfying the edit constraints directly would therefore be desirable. The simulations in the previous section show that mean imputation performs worst in preserving the distribution. The method using the Dirichlet distribution succeeds to preserve the actual distribution best, and therefore seems to be a promising method. However, a more elaborate simulation study needs to be done. This study should make use of real economic data and create missing items according to several missing data mechanisms, while the percentage of missing items is varied.

43. Moreover, future research is needed to extend the described method, in order to handle complex edits, that is, the fact that variables are present in several edits and inequality edits.

## References

Dempster, A.P., Laird, N.M. and D.B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion)," *Journal of the Royal Statistical Society B*, 39, 1-38.

Kalton, G. and D. Kasprzyk (1986), "The Treatment of Missing Survey Data," *Survey Methodology*, 12, 1-16.

Ronning, G. (1989), "Maximum Likelihood Estimation of Dirichlet Distributions," *Journal of Statistical Computation and Simulation*, 32, 215-221.

Wilks, S. S. (1962), *Mathematical Statistics*, New York: Wiley series.