UNIVERSITY OF SOUTHAMPTON

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES

School of Social Sciences

Applications of Order Statistics Based on Concomitant Variables in Survey Sampling

by

Ebrahim Khodaie-Biramy

Thesis for the degree of Doctor of Philosophy

Division of Social Statistics

July 2005

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

SCHOOL OF SOCIAL SCENCES
SOCIAL STATISTICS

Doctor of Philosophy

Applications of order statistics based on concomitant variables in survey sampling

By Ebrahim Khodaie-Biramy

Using auxiliary variables or concomitants to design a survey, to construct an estimator for unknown population parameters for given sample, to select an efficient sample or to make a complete data file for data analysis purposes is common in survey sampling. This thesis uses the theory of concomitants of order statistics to propose an imputation method which is called sequential taxonomy imputation (STI) and a variance estimator for a sample mean under ordered systematic sampling (OSY).

Let (X_i, Y_i) $i = 1, 2, ..., n$ be n independent and identically distributed random variables from a bivariate normal distribution. If $X_{(r:n)}$ denotes the r^{th} ordered X-variate then the Y-variate $Y_{[r:n]}$, paired with $X_{(r:n)}$ is called the concomitant of the r^{th} order statistic. In this thesis, we develop and evaluate a new imputation method for missing values. The method uses concomitants of order statistics. In particular, the method orders the data according to an auxiliary vector (X) and then selects k-nearest neighbours in order to impute a missing value in the variable Y. Under missing at random (MAR) and missing completely at random (MCAR) assumptions, this so-called single ordered sequential taxonomy imputation (SSTI) method is evaluated theoretically and empirically under a linear relationship between the auxiliary vector X and the variable Y. In particular we describe a generalised form of SSTI which is called doubly ordered sequential taxonomy imputation (DSTI). It is shown that, the bias of estimators for population parameters based on these imputed values is smaller than under other imputation methods. In addition, SSTI and DSTI preserve marginal distributions and individual values better than some commonly used imputation methods such as Hot deck imputation.

Applications of order statistics have introduced new sampling methods such as ranked-set and double sampling in recent years. In this research the statistical properties of ranked-set sampling are examined, the usual systematic sampling (SY) scheme is modified and a variance estimator for ordered systematic sampling (OSY) is suggested. Systematic sampling is a practical and efficient method for selecting samples from administrative registers or other logically arranged files. A proper sorting order of the population ensures that the sample obtained reflects the true population distributions. In this study, we use an auxiliary variable to order data and refer to this variable as a concomitant variable because of its ordering properties. By assuming a linear relationship between the variable of interest and its concomitant, we propose a variance estimator for the sample mean that is less biased compared to other variance estimators. In addition, we compare the statistical properties of ranked-set and ordered systematic sampling with simple random sampling. We justify the proposed variance estimator theoretically and demonstrate its properties using a simulation study.

Contents

List of Tables

List of Figures

Acknowledgment

I am deeply grateful to my supervisor, Prof. Ray Chambers, for his continued attention and support, for quality of his advice, criticism and encouragement and insight throughout research. I would also like to thank Prof. Chris Skinner (Social Statistics Division) for his valuable advices.

Thanks are also due to all members of the 'Social Statistics Division and S3RI team' for their support, as well as Dr. Ayoub Saei, Dr. Sandy Mackinnon, Darren Hampton, Tom Bayley, and Christina Thompson. My special gratitude is also due to my colleagues and friends from department and Iranian friends that they made life easier and happier for me and my family here, among them, Leslie, Marcel, Nikos, Reza, Mohammad, Gabi, Zoë, Leonardo, and Solange. Many great friends at EMEO supported me with enthusiasm and devoted attention, in particular, Behrooz, Ehsan, Kheirolah, Seif, and Hamid.

I am forever indebted to my parents, father, and mother in law, brothers, and sisters in Iran for their support and encouragement during all these years, specially my grandmother, and my brother Hamid. No words can express my gratitude to my wife Leila for her unconditional support and understanding encouragement and love throughout all this process. Finally, I dedicate this research to my daughter, Elly, as tribute to her love and patience.

Research for this thesis was generously supported by a grant from Ministry of Science and Technology of Iran, and by my employer the Educational Measurement and Evaluation Organisation of Iran (EMEO).

Chapter 1

Introduction

It is common in survey sampling to use information about the population in constructing a design, estimating procedures, and in selecting the sample. This information can be provided by auxiliary variables or concomitant variables, often from official sources, such as a national census. We can use this information in designing a survey, constructing an estimator for unknown population parameters for a given sample, selecting a sample, or making a complete data file for data analysis purposes. This thesis is divided into two parts.

Making a complete data file by imputation methods is one of the important areas in survey sampling. Concomitant variables can be used to construct a new imputation method or improve the quality of current imputation methods. In constructing a new imputation method, the relationship between the study variable and a set of concomitant variables plays an important role. Throughout the thesis, a linear relationship is assumed between the study variable and the concomitant variables. The proposed method uses concomitants of order statistics theory based on the normal distribution. In this method, the data is ordered, according to a constructed variable from concomitant variables, and then k-nearest neighbours are selected in order to impute a missing value in the study variable. In addition, we use multivariate data ordering methods and the sequential taxonomy technique to order data according to the concomitant variables. The proposed imputation methods are called single ordered sequential taxonomy imputation (SSTI) and double ordered sequential taxonomy imputation (DSTI).

Applications of order statistics have introduced new sampling methods, such as ranked-set and double sampling in recent years. In this thesis, the statistical properties

of ranked-set sampling are examined, the usual systematic sampling (SY) scheme is modified, and a variance estimator of the sample mean for ordered systematic sampling (OSY) is suggested.

The first part proposes a new k-nearest neighbour imputation procedure according to the concomitant variables, and the next part proposes an improved variance estimator for a sample mean in systematic and ranked set sampling based on concomitants of order statistics theory.

The thesis is organised as follows. In chapter 2, the theory of concomitants of order statistics based on normal distribution and the linear relationship between a concomitant variable and the study variable are reviewed. Chapter 3 briefly reviews the causes of missingness and methods for correcting missingness in surveys and censuses, and well-known imputation methods from different statistical aspects, such as classical and Bayesian analysis. Chapter 4 covers background information on statistical theory for sequential taxonomy imputation. Key definitions and the theory of sequential taxonomy imputation are provided in Chapter 5. In this chapter, a new multivariate statistical imputation method is developed and its purpose is discussed. Mathematical properties such as probability distributions, asymptotic behaviour, and the variance of estimates are derived. In addition, Chapter 5 provides some evaluation methods for different imputation methods based on the structure of the data. The results of different imputation methods are compared using simulated multivariate normal data. Chapter 6 briefly reviews the standard systematic, ranked set sampling techniques, and proposes a new variance estimator for the sample mean under ordered systematic sampling (OSY) and median ranked set sampling (MRSS). Moreover, Chapter 6 compares the statistical properties of the variance estimators with other sampling methods, such as simple random sampling and regression sampling, using a simulation study. Finally, Chapter 7 gives a summary and conclusions to the first and second parts of the thesis.

Chapter 2

Theory of concomitants of order statistics

2.1 Introduction

The sequential taxonomy imputation (STI) procedure and the variance estimation for ordered systematic sampling (OSY) to be investigated in this research involve the use of the theory of concomitants of order statistics. Therefore, this chapter briefly reviews the theory of concomitants according to our research purpose.

Suppose (X_r, Y_r) $(r = 1,...,n)$ is a random sample of n observations of a bivariate random variable *(X,Y)* with cumulative distribution function (cdf) *F(x,y)*. If we order the X variates in ascending order as

$$X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n} \quad ,$$

then, the Y-variates paired with these order statistics are denoted by

$$Y_{[1:n]}, Y_{[2:n]}, \cdots, Y_{[n:n]}$$

and termed the concomitants of the order statistics of X (David, 1973, 1981), while Bhattacharya (1974) termed $Y_{[r:n]}$ as the induced order statistics. The $Y_{[r:n]}$ are not necessarily ordered, but can be expected to reflect the association between X and Y, a strong positive association tending to lead to values of $Y_{[r:n]}$ in roughly ascending order, and similarly negative association to $Y_{[r:n]}$ in descending order. Concomitants have found a wide variety of applications in fields such as selection procedure (Yeo and David, 1984), ocean engineering (Castillo, 1988), inference problems (Do and Hall, 1991, Yang 1981), prediction analysis (Gross, 1973) and double sample plans (O'Connel and David, 1976). Under the assumption that X and Y are linearly related, apart from an independent error term, the small sample theory of concomitants has

been studied extensively by O'Connell (1974). The asymptotic distribution theory of the concomitants has been investigated by David and Galambos (1974), when the paired (X_r, Y_r) has a bivariate normal distribution. For a comprehensive review of this topic, see Bhattacharya (1984) and David and Nagaraja (1998).

2.2 Distribution theory of the concomitants of order statistics

This section reviews concomitants when the distribution of (X_r, Y_r) is a bivariate normal distribution: $N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, where μ_x and σ_x^2 are the mean and variance of X, μ_y and σ_y^2 are the mean and variance of Y and ρ is the correlation between X and Y. Before focusing on normal distribution, a general theory of concomitants is reviewed based on Yang (1977) as follows:

Let (X_r, Y_r) $(r = 1, 2, ..., n)$ be i.i.d. variates from a continuous bivariate distribution with cdf $F(x, y)$ and pdf $f(x, y)$. If $f(y \mid x)$ is the conditional distribution of y given x and $f_{r_1,...,r_n}(x_1, ..., x_k)$ is the joint pdf of $(X_{r_1:n}, ..., X_{r_k:n})$, where $k \geq 1$ and $1 \leq r_1 \leq .... \leq r_k \leq n$, then from the i.i.d. property of (X_r, Y_r), the conditional pdf of $Y_{[r:n]}$ given $X_{r:n} = x$ is $f_{Y_{[r:n]}}(y \mid X_{r:n} = x) = f(y \mid x)$. Therefore, the joint distribution of $X_{r:n}$ and $Y_{[r:n]}$ can be written as follows:

$$f_{X_{r:n}, Y_{[r:n]}}(x, y) = f(y \mid x) f_{r:n}(x) \quad . \tag{2.1}$$

Moreover, from (2.1) we have the marginal distribution of $Y_{[r:n]}$

$$f_{Y_{[r:n]}}(y) = \int_{-\infty}^{+\infty} f(y \mid x) f_{r:n}(x) dx \quad . \tag{2.2}$$

Formulae (2.1) and (2.2) can be generalised to a multivariate distribution as follows:

$$f_{Y_{[r_1:n]},...,Y_{[r_k:n]}}(y_1, ..., y_k) = \int_{-\infty}^{\infty} \int_{-\infty}^{x_k} .... \int_{-\infty}^{x_2} \prod_{i=1}^{k} f(y_i \mid x_i) f_{r_1,...,r_k:n}(x_1, ..., x_k) dx_1 ... dx_k \, , \tag{2.3}$$

$$f_{X_{s:n}, Y_{[r:n]}}(x, y) = \int_{-\infty}^{x} f(y \mid t) f_{r,s:n}(t, x) dt \quad , \tag{2.4}$$

where $1 \le r_1 \le .... \le r_k \le n$ and $r < s$.

From (2.3) and (2.4) we can easily show the following results:

$$
\begin{aligned}
&E(Y_{[r:n]}) = E[E(Y_1 \mid X_1 = X_{r:n})];\\
&Var(Y_{[r:n]}) = E[Var(Y_1 \mid X_1 = X_{r:n})] + Var[E(Y_1 \mid X_1 = X_{r:n})];\\
&Cov(Y_{[r:n]}, Y_{[s:n]}) = Cov[E(Y_1 \mid X_1 = X_{r:n}), E(Y_1 \mid X_1 = X_{s:n})](r \neq s);\\
&Cov(X_{s:n}, Y_{[r:n]}) = Cov[X_{s:n}, E(Y_1 \mid X_1 = X_{r:n})].
\end{aligned}
\tag{2.5}
$$

Now we focus on the special case of concomitants, where the distribution is bivariate normal, and the relationship between X and Y is linear or more generally nonlinear.

2.3 The simple linear model

We start with the simplest case of concomitants; suppose (X_r, Y_r) (r=*1,2,..,n*) is a random sample of n pairs drawn from a bivariate normal distribution $N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, where μ_x, μ_y, σ_x^2 and σ_y^2 are means and variances of X and Y and ρ is the correlation between X and Y. According to the normal distribution, X and Y have a linear relationship as follows:

$$Y_r = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(X_r - \mu_X) + \varepsilon_r \quad , \tag{2.6}$$

where X_r and ε_r are independent. Then, from (2.6), we have $E(\varepsilon_r) = 0$ and $Var(\varepsilon_r) = \sigma_Y^2(1-\rho^2)$. By ordering data based on X and using formula (2.6), we have:

$$Y_{[r:n]} = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(X_{r:n} - \mu_X) + \varepsilon_{[r]} \quad , \quad r = 1, 2, ..., n \tag{2.7}$$

where $\varepsilon_{[r]}$ denotes the particular ε_r associated with $X_{r:n}$. Independence of ε_r and X_r, implies that $X_{r:n}$ and $\varepsilon_{[r]}$ are independent. From (2.7) and by using

regression and normal distribution properties, the expectation and variance of $Y_{[r:n]}$ can be obtained as follows:

Let

$$\alpha_{r:n} = E(\frac{X_{r:n} - \mu_X}{\sigma_x}) \text{ and } \beta_{rs:n} = Cov(\frac{X_{r:n} - \mu_X}{\sigma_X}, \frac{X_{s:n} - \mu_X}{\sigma_X}),\ r,s = 1,2,...,n. \qquad (2.8)$$

Then, from (2.7) we have:

$$E(Y_{[r:n]}) = \mu_Y + \rho\sigma_Y\alpha_{r:n} \qquad (2.9)$$

$$Var(Y_{[r:n]}) = \sigma_Y^2(\rho^2\beta_{rr:n} + 1 - \rho^2), \qquad (2.10)$$

$$Cov(X_{r:n}, Y_{[s:n]}) = \rho\sigma_X\sigma_Y\beta_{rs:n}, \qquad (2.11)$$

$$Cov(Y_{[r:n]}, Y_{[s:n]}) = \rho^2\sigma_Y^2\beta_{rs:n} \qquad r \neq s, \qquad (2.12)$$

where $\alpha_{r:n}$ is the expectation of the rth of X order statistics and $\beta_{rs:n}$ is the covariance of rth and sth order statistics of the variable X from the standard normal distribution.

2.4 General models

In this section, following to David and Ngaraja (2003), we generalize (2.6). Suppose $Y_i = g(X_i) + \varepsilon_i$ is the general model of the regression of Y on X, where X_i and ε_i are independent. The general model for concomitants can be written as follows

$$Y_{[r:n]} = g(X_{(r:n)}) + \varepsilon_{[r:n]}, \quad r = 1,2,...,n.$$

Then from (2.5) and (2.3), formulae (2.9), (2.10), (2.11) and (2.12) can be written as follows:

$$E(Y_{[r:n]}) = E[g(X_{r:n})], \qquad (2.13)$$

$$Var(Y_{[r:n]}) = Var[g(X_{r:n})] + E[\sigma^2(X_{r:n})], \qquad (2.14)$$

$$Cov(X_{r:n}, Y_{[s:n]}) = Cov[X_{r:n}, g(X_{s:n})], \qquad (2.15)$$

$$Cov(Y_{[r:n]}, Y_{[s:n]}) = Cov[g(X_{r:n}), g(X_{s:n})]\ ,\ r \neq s. \qquad (2.16)$$

2.5 The asymptotic distribution of concomitants

This section shows the asymptotic distribution of concomitants when (X_i, Y_i) $(i=1,2,..,n)$ is a random sample of n pairs drawn from a bivariate normal distribution $N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. From (2.9) we have:

$$\mu_Y = E(Y_{[r:n]}) - \rho\sigma_Y\alpha_{r:n}, \tag{2.17}$$

where $\alpha_{r:n}$ is the expectation of the rth of X order statistics from the standard normal distribution and ρ is the correlation between X and Y. By replacing (2.17) in formula (2.7) we have:

$$Y_{[r:n]} = \left(E(Y_{[r:n]}) - \rho\sigma_Y\alpha_{r:n}\right) + \rho\frac{\sigma_Y}{\sigma_X}(X_{r:n} - \mu_X) + \varepsilon_{[r]},\ r=1,2,\ldots,n.$$

By simplifying the above formula we have:

$$Y_{[r:n]} - E(Y_{[r:n]}) = \rho\sigma_Y\left(\frac{(X_{r:n} - \mu_X)}{\sigma_X} - \alpha_{r:n}\right) + \varepsilon_{[r]},\ r=1,2,\ldots,n, \tag{2.18}$$

where $\alpha_{r:n} = E(\frac{X_{r:n} - \mu_X}{\sigma_x})$. If the first part on the right hand side of (2.18), which is the function of $X_{r:n}$, $\rho h(X_{r:n}) = \sigma_Y\left(\frac{(X_{r:n} - \mu_X)}{\sigma_X} - \alpha_{r:n}\right)$, converges to zero in probability as $n \to +\infty$ for all r. In other words if $\left(h(X_{r:n}) - c_n\right) \to 0$ as $n \to \infty$, $\left(Y_{[r:n]} - \rho c_n\right) \xrightarrow{d} \varepsilon$ and hence the limit distribution of $Y_{[r:n]}$ can be arbitrary (David and Ngaraja, 2003). Therefore for a bivariate normal distribution, assuming $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, the above condition holds with $c_n = (2\log n)^{\frac{1}{2}}$. Therefore, from (David and Nagaraja, 2003) the distribution of $Y_{[r:n]} - E(Y_{[r:n]})$ is asymptotically $N(0, \sigma_Y^2(1-\rho^2))$.

$$Y_{[r:n]} - E(Y_{[r:n]}) \sim N(0, \sigma_Y^2(1-\rho^2)). \tag{2.19}$$

The expectation of $Y_{[r:n]}$ in (2.19) is affected by how r is related to n. David and Nagaraja (2003) derived an asymptotic expression for $E\left(Y_{(r:n)}\right)$ in a simple case ($\mu_X = \mu_Y = 0,\ \sigma_X = \sigma_Y = 1$) under three different situations. These are the quantile case, where $\frac{r}{n} \to p,\ 0 < p < 1,$ the extreme case, where either r or n-r is fixed, and the intermediate case, where

$$\begin{aligned} E\left(Y_{[r:n]}\right) &\cong \rho\Phi^{-1}(p) && r = [np],\ 0 < p < 1 \\ &\cong \rho\left(2\log\ n\right)^{1/2} && r = n - k + 1,\ k \text{ fixed} \\ &\cong -\rho\left(2\log\ n\right)^{1/2} && r = k,\ k \text{ fixed} \end{aligned} \tag{2.20}$$

It should be emphasized that (2.20) is valid only under simple random sampling scheme.

Part I

Nearest Neighbour Imputation based on Concomitant Variables

Chapter 3

A survey of current imputation methods

3.1 Introduction

This chapter reviews the causes of missingness, methods for correcting missingness in surveys and censuses, and well-known imputation methods from different statistical aspects, such as classical and Bayesian analysis.

3.2 Imputation and missing values

Standard statistical methods and formula have been developed for analysing complete data. Incompleteness in a data file can lead to biased estimates, and standard methods of statistical data analysis cannot be used. In addition, possible biases exist when respondents are systematically different from non-respondents, and these biases are difficult to eliminate when reasons for non-response are not known.

Non-response is just one source of incompleteness. It is typically characterised as being of two types, unit and item non-response. Unit non-response occurs when no information is collected from a sample case, and item non-response occurs when some of the questions for a case are not answered. However, unit non-response can also occur because of data editing, where answers that fail the edits are suppressed and replaced by imputed values. Unit non-response is usually accounted for by weighting adjustment methods, while item non-response is commonly dealt with using imputation methods. This thesis will be concerned with an investigation of imputation methods based on ordering.

3.3. Missingness

Missing observations in sample surveys lead to non-sampling errors. Generally, total survey errors split into two components: sampling errors and non-sampling errors. Sampling error arises from the sampling process itself, when inferences are made from observation on a randomly chosen subset of units, rather than observing the whole population. In other words, sampling error is a penalty for incomplete enumeration. Non-sampling error includes all the errors not attributable to this incomplete enumeration. Every step in the survey process is a potential source of non-sampling error, such as imperfections in the initial specification, incomplete listing of the target population, failure to obtain complete information from all units drawn in the sample, failure to obtain correct information from the contacted units and errors in recording and managing the data after the survey has been completed. Sampling error is relatively easy to deal with, at least in theory, and increasing the sample size and modelling the proper choice of design and estimator can reduce it. The size of the sampling error may be estimated from the sample in this case. In contrast, non-sampling errors often increase when sample size increases or there is increased complexity in the sampling procedure. In addition, it is difficult to measure the size of most components of non-sampling errors without external information. The following section examines some specific sources of non-sampling errors from the missing values point of view. Missing values in sample surveys occur in three ways: non-coverage, unit non-response, and item non-response. Non-coverage arises when the population from which the sample is actually drawn differs from the target population. Unit non-response occurs when no information is collected from a sampled unit, and item non-response occurs when some but not all the information is collected. This thesis focuses on item non-response.

Item non-response may occur for a variety of reasons. The respondents may not know the answers to certain questions or they may refuse to answer some questions. They may consider them sensitive or irrelevant. The interviewer may skip over a question under pressure in an interview or for other reasons. A recorded answer on the survey may be rejected during editing because of inconsistency with other questions. Analysing missing values requires clear assumptions about the structure, pattern, and

mechanism for the occurrence of missing values. These are reviewed in the next two sections.

3.3.1 Missing data mechanisms

The performance of missing-data methods depends strongly on the missing data mechanism, which explains why values are missing, and in particular, whether missingness depends on the values of variables in the dataset. For example, subjects in longitudinal studies may drop out of a study because they feel that their involvement is ineffective, which might be related to a poor value of an outcome measure. Little and Rubin (1987) define three types of missing data mechanisms. For simplicity, we use Rubin's notation. Let $Y = (Y^{obs}, Y^{mis})$ be a data matrix that includes two parts, missing and observed; and let $R = (r_{ij})$ be the associated missing-data indicator matrix; that is $r_{ij} = 1$ if y_{ij} is observed, and $r_{ij} = 0$ if y_{ij} is missing. Little and Rubin (2003) assume that the missing-data indicator matrix is a random matrix, and characterises the missing-data mechanism by the conditional distribution of R given Y, $f(R \mid Y, \phi)$, where $\phi = (\theta, \psi)$ denotes unknown parameters for the distribution of Y and R respectively. Missing-data mechanisms then fall into three classes, the missing completely at random mechanism (MCAR), the missing at random mechanism (MAR) and non-ignorable missingness (NI). Here, we define missing data mechanisms according to Little and Rubin's (1987) definitions.

a) Missing completely at random (MCAR)

Under MCAR, the missing-data mechanism is unrelated to the values of any of the survey variables, whether missing or observed. In other words, the observations are missing completely at random, with $P(R = r \mid Y = y) = P(R = r)$ or equivalently $f(R \mid Y, \phi) = f(R \mid \phi)$ for all Y and ϕ; that is R and Y are independent. For example, if survey participants randomly skipped answering some questions or a researcher accidentally forgot to ask some questions or drops a case, in this case, the available data is a simple random sub-sample from the original dataset.

b) Missing at random (MAR)

A slightly weaker assumption than MCAR is missing at random (MAR). Under MAR the missing-data mechanism depends only on the observed values Y^{obs} and not on the missing values Y^{mis}. In other words, the observations are missing at random if $P(R=r \mid Y=y)=P(R=r \mid Y^{obs})$ or equivalently $f(R \mid Y,\psi)=f(R \mid Y^{obs},\psi)$ for all Y^{mis} and ψ. That is, knowledge about Y^{mis} does not provide any additional information about R if Y^{obs} is already known. As a consequence, R and Y^{mis} are conditionally independent, given Y^{obs}. For example, since people with low education may skip questions about household income, low education is one explanation of why income is missing. If we include education in any question as a mechanism variable, we will have the distribution of R in our survey. More generally, if we include all the mechanism variables in our model as controls, we have made the missing data mechanism MAR. However, Little (1995) introduces covariate dependent missingness (CD) as a extension to MAR. Both MAR and CD require that the cause of the missing data is unrelated to the missing values, but may be related to the observed values of other variables. MAR means that the missing values are related to either observed covariates or response variables; whereas CD means that, the missing values are related only to covariates. As an example of CD missing data, missing income data are only related to education.

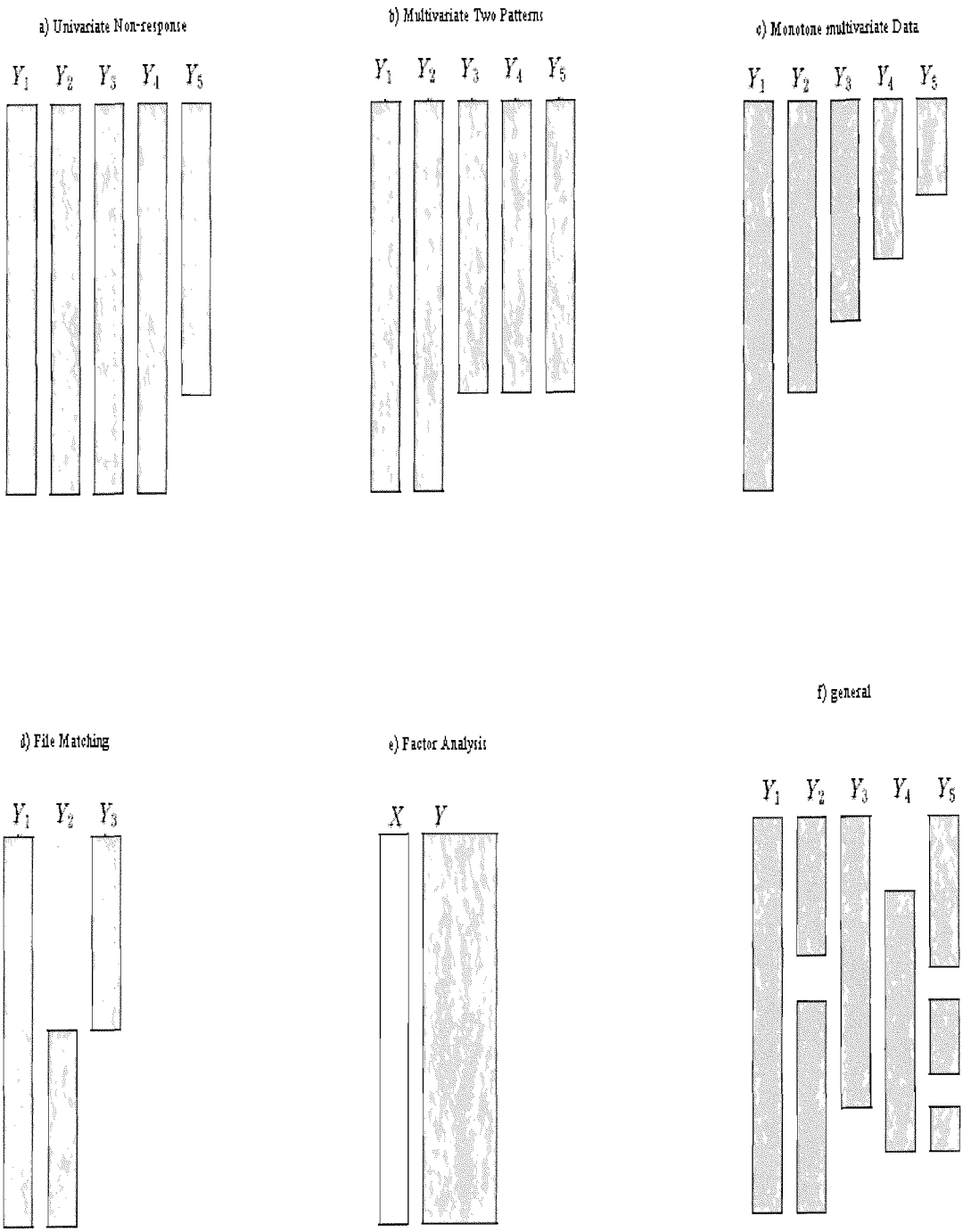
3) Non-ignorable (NI) missingness

If the MAR assumption does not hold, then we say that the missing data mechanism is non-ignorable (NI). NI means that the missing data mechanism is related to the missing values. For example, if individuals with higher incomes are less likely to reveal them on a survey than are individuals with lower incomes, the missing data mechanism for income is non-ignorable. Whether income is missing or observed is related to its value.

3.3.2 The missing data pattern

The missing data pattern simply indicates which values in the dataset are observed and which are missing. Specifically, let $Y=(y_{ij})$ denote an $(n\times p)$ data matrix without missing values. With missing values, the pattern of missing data is defined by the missing-data indicator matrix $R=(r_{ij})$, such that $r_{ij}=1$ if y_{ij} is observed and $r_{ij}=0$ if y_{ij} is missing. We assume throughout this thesis that individual values $y_i=(y_{i1},y_{i2},...,y_{ip})$ and $r_i=(r_{i1},r_{i2},...,r_{ip})$ are independent over i. Some missing data analysis methods are designed for any pattern of missing data, whereas other methods assume a special pattern. Certain patterns may allow simpler or more direct techniques to be applied: for example, "monotone" missingness patterns may allow ML estimates (under a "missing at random" mechanism assumption) to be obtained without resorting to data augmentation or imputation. Figure 3.1 shows the shape of different types of missing values in multivariate data. Obviously, for univariate data, the missing data pattern is the special case (3.1a).

Figure 3.1 Missing data patterns



Figures 3.1a-f display different types of missing data patterns. Figure 3.1a shows the simplest case of a missing data pattern; in this case, missing values occur only in one variable and most imputation methods can be used to handle the missing data. Figure

3.1b shows unit non-response for multivariate data, with the fully observed variables consisting of survey design variables. In other words Y_1 and Y_2 are survey design variables. The third type of missing data pattern (Figure 3.1c) displays a monotone pattern of missingness in Y. Y_1 is at least as observed as Y_2, which is at least as observed as Y_3, and so on. Such a pattern of missingness, or close approximation to it, is not uncommon in practice. For example, longitudinal studies collect information on a set of cases repeatedly over time. Different subjects drop out in each "wave" and do not return. Figures 3.1d and 3.1f are also common missing data patterns. Figure 3.1d shows two sets of variables that are never jointly observed. An important point is that attempts to estimate the association between these variables may yield misleading results. Figure 3.1e shows latent-variable patterns with some variables never observed. It will be useful to consider these kinds of unobserved latent variables as a missing-data problem, where the latent variables are completely missing, and then use missing data theory to estimate data parameters. In the Figure 3.1e, X is a latent variable that is completely missing and Y is a set of variables that are fully observed. Factor analysis is one of the suggested multivariate methods to deal with this kind of pattern. By using factor analysis, under specific assumptions, the data parameters can be estimated without using imputation methods.

Almost all imputation methods can be used to handle missing values, but they need more work in non-standard situations. For example, when handling missing values by regression imputation under a monotone pattern of missingness, there are two options: the first is to fit a multivariate regression model to all the fully observed variables. In this case, we have to discard a considerable amount of information, because the fully observed values are only those with observed values for the last variable. Therefore, the number of valid cases for a multiple regression imputation is the number of respondents for the last variable (Y_5). The second option is to carry out the regression imputation on a variable-by-variable basis, using only the most fully observed variables. For example, suppose Y_1 is a complete variable and missing values occur in Y_2, then missing values in Y_2 are imputed based on the regression of Y_2 on Y_1. Now variables Y_1 and Y_2 are complete and we can use the regression of Y_3 on Y_1 and Y_2 to

impute missing values for Y_3. This procedure continues until all missing values are imputed.

3.3.3 Strategies for analyzing data with item non-response

In the missing data literature, there are three general approaches to handling data containing missing values. These approaches depend on several factors, such as knowledge about the data distribution, the missing data pattern, the missing data mechanisms, and the number of missing values. To illustrate, consider a dataset made up of two variables $Y_1 = (Y_1^{obs1}, Y_1^{obs2})$ and $Y_2 = (Y_2^{obs}, Y_2^{mis})$, where Y_1 is a complete variable, while missing values occur in Y_2. The first strategy is based on the analysis of the complete, which focuses on Y_1^{obs1} and Y_2^{obs} only. This approach is designed to make inference simple. However, it can be inefficient and can produce biased estimates. The second approach is based on using all of the available data and is based on the application of either weighting or model-based procedures. Under both we use Y_1^{obs1}, Y_2^{obs} and Y_1^{obs2} to estimate unknown population parameters. The weighting method typically leads to simple estimates but complex inferences about parameters, and can be inefficient, whereas model-based methods typically lead to complex estimates and inferences about parameters that are statistically efficient. The third approach is imputation, which uses Y_1^{obs1}, Y_2^{obs}, Y_1^{obs2} and $\hat{Y}_2^{mis}$ to make an estimate and inferences about unknown parameters, where $\hat{Y}_2^{mis}$ denotes imputed values for the missing observations in Y_2^{mis}. Under this approach, we usually have simple estimates but complex inferences for unknown parameters, which can be less efficient than model-based methods. However, since fully specified models for data can be unavailable in many surveys, this loss of efficiency from using imputation methods to handle missing values is usually acceptable.

1) Complete case analysis

If the fraction of missing data is small, one way to handle it is to discard records or cases that have missing values. This procedure is known as list-wise deletion or case-wise deletion. A serious limitation of this approach is that relevant data are frequently

discarded. List-wise deletion is simple, but with a greater number of variables, increasing amounts of data are ignored, even though the total number of missing values remains constant. For example, in an extreme case, where all respondents in the sample have only one (not necessarily the same) variable missing, list-wise deletion would discard all cases. Hence, this approach leads to excessive loss of statistical power. As the number of complete cases decreases, there is a decrease in error degrees of freedom, leading to a loss of statistical power and a larger standard error.

An alternative is pair-wise deletion, which is attractive when there is a small number of missing cases for each variable relative to total sample size, and a large number of variables are involved. With this method, all available observations for each particular variable are used to compute means and variances, while all available pairs of values are used to compute covariances. Thus, correlations are computed using only those observations that have non-missing values for both variables. The problem with the use of pair-wise deletion is the potential inconsistency of the covariance matrix in a multivariate context. When correlations and other statistics are based on different but overlapping sub-samples of a larger sample, the population to which generalization is sought is no longer clear.

There are two obvious advantages to complete case analysis: (a) it can be used for any kind of statistical analysis without an increase in complexity; (b) no special computational methods are required. Nevertheless, discarding cases will decrease the sample size and this reduced sample size increases the variance of estimates. Note also that, under MCAR, complete case analysis does not lead to an increase in bias, since the distribution of the missing data is the same as the distribution of the complete case data. The same is not true under MAR, where these distributions can be quite different. Since MCAR has a very strong modelling assumption, the use of complete case analysis is not usually recommended unless the extent of missingness in the data is very small.

2) Available case analysis

This approach uses all of the available information for inference, and can be divided into two kinds: weighting methods and model-based methods. Weighting methods for non-response modify estimation weights in an attempt to adjust for non-response, treating this as part of the sample design. Typically, design weights are inversely proportional to the probability of selection. For example, if Y is a random variable defined on a specific population, and y_i is the value of this variable for unit i in the population, then the mean of Y in the population can be estimated via:

$$\hat{\bar{Y}} = \sum \pi_i^{-1} y_i \Big/ \sum \pi_i^{-1}$$

where the summation is over sample units, π_i is the probability of inclusion in the sample for unit i, and π_i^{-1} is the design weight for unit i. With non-response on Y, this estimate cannot be calculated. Weighting procedures modify the estimator, so that it is of the form:

$$\hat{\bar{Y}} = \sum (\pi_i \hat{p}_i)^{-1} y_i \Big/ \sum (\pi_i \hat{p}_i)^{-1},$$

where summation is over respondent units, and $\hat{p}_i$ is the estimated probability of response for unit i. Weighting procedures and some imputation methods, such as mean imputation methods are related. For example, if we have constant design weights in categories of the sample, both provide the same estimates for the mean of the population.

The second method of handling missing data using the entire available information is via model-based procedures. Here, a model for the data (including the missing data) and a model for the missing data mechanism are specified. Then, unknown parameters in the models are estimated using efficient estimation procedures, such as maximum likelihood or Bayesian methods. The main advantages of this model-based approach are flexibility, avoidance of ad hoc methods, the availability of assessment methods

for model assumptions and estimators, and the capacity to validate these procedures using the central limit theorem for large samples.

3) Imputation

A popular class of procedures for handling missing item response is imputation. Imputation methods are the focus of this thesis. Imputation is the assignment of likely or feasible values to remaining missing items. Imputation procedures use information that is available for a case to impute missing item values for that case. This implies that imputation is most successful when the amount of information to be imputed is small in relation to what is known about the case.

Before proceeding further, we shall define some common concepts used in the thesis. A donor is the record from which the value to be assigned to the missing item is normally taken. The records with missing items (for which imputation is carried out) are called recipients. Not all imputation methods assign imputed values to recipients from a donor (e.g. mean imputation). In other words, a donor may come from a model rather than the data file. Auxiliary variables (also called control variables, matching variables or assignment variables) are those related to a variable with missing values. These variables are used to find the distance between donors and recipients or can be used for defining models, for example regression models.

In general, let Y be the variable with missing values, with $X_1, X_2, ..., X_p$ the p auxiliary variables that are complete. Suppose we want to impute a missing value y_i. A broad class of imputation methods that lead to imputed values can be written as follows:

$$\hat{y}_i = g_i\left(X, Y^{obs}, e^{mis}\right),$$

where g (.) is a known function and e^{mis} is a part of some introduced perturbations. In the above formula, the specification of g (.) can distinguish the imputation methods (Lessler and Kalsbeek, 1992). For example, in deterministic imputation e^{mis} set to zero. We therefore have

$$\hat{y}_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} ,$$

where b_0 and b_j are estimated by standard regression methods from complete case data. In contrast, for stochastic regression imputation, a randomly generated residual e_i is added to this imputed value. For mean imputation, a missing value is imputed as:

$$\hat{y}_i = \bar{y} ,$$

where $\bar{y}$ is the mean of respondents for the variable Y. Again, this can be extended to stochastic imputation by addition of a randomly generated residual. This method can be generalized to categorical variables via either mean, median or mode imputation.

Imputation creates a complete dataset and has the following advantages.
1) Data collectors usually know the reasons for the missing values. This information can be used in imputation.
2) Missing values complicate the data structure, so sophisticated statistical tools are required to conduct analysis. Imputation may ease this difficulty.
3) Imputation can prevent the loss of information due to deletion of incomplete records if the statistical methods used (e.g. regression) require complete records.
4) Imputation can reduce non-response bias in some situations. Imputation attempts to reduce bias based on distinct assumptions, which specify the missing mechanisms and the relationships between response and non-response.

On the other hand, a major concern is that imputation can distort the distribution of variables, leading to underestimation of variance, or changes in the relationship between variables. A proper imputation method should therefore lead to plausibility and consistency with edits, should reduce bias and preserve the relationship between items properly, should work with every missing data pattern and mechanism, should be capable of being set up ahead of time, and should be capable of being evaluated in terms of impact on the biases and the precision of the estimates. However, this does not necessarily lead to estimates that are less biased than those obtained from the incomplete dataset; indeed the biases could be much larger, depending on the imputation procedure and the form of the estimate.

3.4 General categories of imputation methods

Several methods of data imputation are available, but some of the more popular methods for handling missing data appear below. The list is not exhaustive, but it covers some of the more widely recognized approaches for handling incomplete cases in databases. It should be noted that different imputation methods are acceptable and reliable in different circumstances, such as with discrete or continuous data structures and the various missing data mechanism.

There are two main types of imputation methods: the first is single imputation, and the second is multiple imputation. Single imputation assigns a single value to each of the missing values and multiple imputations impute each of the missing values by two or more values to reflect the distribution of the missing values (Rubin, 1987). Multiple imputation produces more than one complete dataset. The data analysis should combine the results from each component dataset. All methods are based on some assumptions about the missing data mechanism and pattern, even if they are derived from intuitive models. Without these assumptions, any imputation method cannot be justified. Multiple imputation has the advantage of compensating for the uncertainty of the assumptions, but it increases the complexity of the analysis. Therefore, single imputation probably remains the most widely used approach. In the literature, there is another categorizing procedure for imputation methods. In this procedure, imputation methods are divided into three groups: real-donor, model-donor, and a mixture of real-donor and model-donor. This section discusses imputation techniques under these three categories. The three categories are not exhaustive, and are used mainly for convenience of discussion.

3.5 Real-donor imputation methods

The real-donor imputation method imputes missing values by observed values without any changes. The popular real-donor imputation methods are as follows.

3.5.1 Deductive imputation

This method deduces missing values from available information such as similar items in previous surveys, related items in current surveys, etc. To apply this method, the user needs to find some deterministic relationship between the missing item and items from other sources. Generally, it is impossible to find enough information to impute all missing items in a survey using deductive imputation, but this method can be used to impute some of the missing items. Whenever possible, deductive imputation should be used before any other imputation method, because it provides accurate or approximately accurate imputations for missing cases. However, the performance of a deductive imputation method completely depends on the available sources.

3.5.2 Cold deck imputation

Cold deck imputation provides a set of values for each missing item, which are usually taken from previous surveys. The problem with this method is the compatibility between the former and the more up-to date data.

3.5.3 Hot deck imputation

The hot deck imputation method is a general class of procedures for the handling missing data, and there is no general agreement on the definition of this method. One of the more general definitions of hot deck method is as follows. A hot deck method is a duplication process, when a missing value occurs in a sample; a reported value is chosen and replaces the missing value. One of the advantages of hot deck methods is their simplicity and non-reliance on a strong statistical background. In addition, in the hot deck imputation method generally, there is no explicit statistical model. The general methodology of hot deck imputation employs many methods, and some popular methods are based on classification, finding the closeness of a missing value to an available case based on different closeness procedures, such as distance or similarity measures. In the classification process, data is classified into homogenous categories, based on prior knowledge about the structure of the data, or based on available information. In this case, for an efficient classification, we must have a reasonable correlation between auxiliary variables and the variable with missing values. After the classification process, hot deck imputation methods employ two

strategies for handling missing values: the first is deterministic and the second the random.

a) Deterministic hot deck imputation

The following are the most popular deterministic hot deck imputation methods.

a1) Sequential hot deck imputation

In the deterministic sequential hot deck method, data are ordered, then for each missing value the nearest value is chosen, and replaced. If the data ordering procedure is not a random method, we call this a deterministic sequential hot deck method. First, a single value, such as the mean or some pre-specified value, is assigned as an initial value. Then the records or cases in the data file are treated sequentially. If a record has a response for the target variable, this response replaces the previously stored value. If a record has a missing value for the target variable, it keeps the previously assigned value. This method can be used in categorized data. In other words, the data is categorised into homogeneous categories then the sequential hot deck is carried out inside each category.

A major attraction of this method is its computing economy, since all imputations are made in a single pass through the data file. A disadvantage is that it may easily give rise to multiple uses of donors, a feature that leads to a loss of precision for survey estimators (Kalton and Kasprzyk, 1982). In addition, there is a serial correlation between cases (see Bailar, Bailey and Corby, 1978).

a2) Multivariate matching

In this method, the respondents are divided into groups, and then each nonrespondent is matched to a group of respondents. Classification of respondents into groups is based on the values of a set of fully observed auxiliary variables. Each nonrespondent is matched to the group of respondents for whom the auxiliary variables take the same values, and the missing value is imputed using the value observed from the nearest respondent in the matched group. If no donor is found in a matched group, the group is combined with other groups to obtain donors.

While this method is not convenient to implement using computer programs, an approximately equivalent imputation algorithm may be used to replace it. The algorithm first sorts the data file by the same auxiliary variables, and then imputes the nearest response value for each missing case. This alternative method is very easy to implement. The donor and recipient will match on all auxiliary variables if such donors are available. Otherwise, it will automatically find a donor matched on some of the auxiliary variable, which is equivalent to collapsing the matched groups.

a3) Nearest neighbour imputation

In this method, the nearest value to a missing value is found, based on a distance or similarity function of covariates. In the literature, this method has been called as a distance function matching imputation method (Little and Rubin 1987). Therefore, it is assumed that auxiliary variables or covariates are fully or partially observed. Finding the distance or similarity between two cases depends on the type of variables. A variety of distance or similarity functions that can be used to find nearness based on the type of variables. (For more details, see section 4.2.3.)

b) Random hot deck method

The process of a random hot deck method is divided into three steps: first, defining the match method (in other words, in this step, we have to define the auxiliary variables that are going to be used in matching); secondly, the missing value is imputed using the value observed for a randomly drawn respondent from the class; thirdly, if there was no valid value for a missing value in this class, then combining some classes is an alternative strategy to reach a valid imputed value for the missing value. In this method, we categorise the data into homogenous categories or classes, then the random hot deck method is carried out inside each class. The following are the most popular random hot deck imputation methods.

b1) Random sequential hot deck imputation

In this method, the data is ordered randomly, while other processes are the same as in the deterministic hot deck imputation method. Bailar, Baily and Corby (1978) showed

that the sample mean with imputed values is an unbiased estimator of the population mean. In addition, they showed that the variance of sample mean with replacement sampling is larger than the variance in a hot deck procedure without replacement sampling.

b2) Hierarchical hot deck imputation

Similar to the sequential hot deck method, the data file is sorted into a much larger number of imputation classes or categories in a hierarchical structure. It is possible to add more auxiliary variables and have a greater number of imputation classes. If no suitable donor is found at the finest level of the classification, classes can be collapsed into broader groups until a donor is found. A pattern of 'hard' and 'soft' class boundaries can be programmed into a hierarchical structure, e.g. to ensure that an item is always imputed from a donor of the same group. It is assumed that the missing mechanism in each class is completely at random. Therefore, the imputation is just a random selection of the available values. This method is similar to the nearest neighbour imputation that is explained in section 3.6.2.

c) The mathematical properties of hot deck

In this section, the mathematical properties of the hot deck imputation method will be reviewed, because of its similarity with the sequential taxonomy imputation procedure. According to the traditional hot deck procedure, missing values are replaced randomly by values from the most similar respondents in the sample. Consider a finite population of size N of Y and a sample with a size of n from this population. Let y to be a sample from Y with size n, y_r are respondents with the size of r and y_{n-r} are non-respondents with the size of $n-r$. Given an equal probability to sample units (Rubin, 1987), the mean of sample can be written:

$$\bar{y}_{HD1} = \left\{ r\bar{y}_r + (n-r)\bar{y}^*_{n-r} \right\} / n \quad , \tag{3.1}$$

where $\bar{y}_r$ is the mean of respondents in the sample and

$$\bar{y}^*_{n-r} = \sum_{i=1}^{r} \frac{H_i y_i}{n-r} ,$$

where H_i is the number of times y_i is used for imputing a missing value. The statistical properties of $\bar{y}_{HD1}$ depends on the generation method of $(H_1, H_2, ..., H_r)$. In the traditional hot deck method, one of the suggested methods is selection by a probability sampling design; in this case, the distribution of $(H_1, H_2, ..., H_r)$ is known in the statistical literature. For example, suppose the probability sampling design for $(H_1, H_2, ..., H_r)$ is simple random sampling with replacement from the respondents. By conditioning over given respondents, the distribution of $(H_1, H_2, ..., H_r)$ is multinomial, with sample size $n-r$ and probabilities of $(1/r, 1/r, ..., 1/r)$. Therefore, according to the theory of multinomial distribution we have:

$$E(H_i \mid y_r) = \frac{n-r}{r},$$

$$Var(H_i \mid y_r) = (n-r)(1-1/r)/r,$$

$$Cov(H_i, H_{i'} \mid y_r) = -(n-r)/r^2, \quad \text{for } i \neq i'.$$

Now, suppose that $\bar{y}_{HD1}$ is the estimate of the population mean with respect to the above multinomial distribution, then:

$$E(\bar{y}_{HD1} \mid y_r) = \bar{y}_r$$

and

$$Var(\bar{y}_{HD1} \mid y_r) = (1-r^{-1})(1-r/n)s_r^2/n, \tag{3.2}$$

where s_r^2 is the variance of respondents in the sample. It can be shown that $\bar{y}_{HD1}$ is an unbiased estimator for $\bar{Y}$ (population mean) as follows:

$$E(\bar{y}_{HD1}) = E\left(E\left(\bar{y}_{HD1} \mid y_r\right)\right), \tag{3.3}$$

then, by assuming MCAR for the missing data mechanism:

$$E(\bar{y}_{HD1}) = \bar{Y}.$$

In addition, the variance of $\bar{y}_{HD1}$ given Y (Rubin, 1987) can be obtained as follows:

$$Var(\bar{y}_{HD1} \mid Y) = Var\left(E\left(\bar{y}_{HD1} \mid y_r\right)\right) + E\left(Var\left(\bar{y}_{HD1} \mid y_r\right)\right); \tag{3.4}$$

then, by assuming MCAR for the missing data mechanism

$$Var(\bar{y}_{HD1} \mid Y) = (1/r - 1/N)S_Y^2 + (1-1/r)(1-r/n)S_Y^2/n,$$

where S_Y^2 is the population variance.

The variance of sample mean by the traditional hot deck method is larger than the mean imputation method. The equation (3.2) is the added part to the variance estimation of $\bar{y}_{HD1}$. In other words, this part is added by hot deck imputation to the total variance. This added part of the variance of sample mean can be decreased again by selecting an appropriate sampling design, such as sampling without replacement. To conduct hot deck imputation by simple random sampling without replacement according to Little and Rubin (1987), the number of respondents should be greater than the half of sample size or number of non-respondents should be less than the number of respondents. In other words, we have $r \geq n/2$ or $n-r < r$ in r respondents cases. Missing values are replaced by sampling without replacement from respondent cases, and in this case H_i is zero or one because selected cases cannot be duplicated. To produce a general method, suppose that $n-r = kr+t$, where k is a nonnegative integer and $0 \leq t \leq r$. In other words, this method selects each respondent k times and then selects t additional cases for imputing the $n-r$ missing values. Hence,

$$\bar{y}_{n-r}^{*} = (kr\bar{y}_r + t\bar{y}_t)/(n-r),$$

where $\bar{y}_t$ is the mean of t additional cases of Y. The expectation and variance of $\bar{y}_t$ given y_r can be obtained as follows:

$$E(\bar{y}_t \mid y_r) = \bar{y}_r$$

and:

$$\mathrm{Var}(\bar{y}_t \mid y_r) = (1 - t/r)s_r^2 / t$$

Now, suppose that $\bar{y}_{HD2}$ is the estimator of the population mean $\bar{Y}$, as follows:

$$\bar{y}_{HD2} = (k+1)rn^{-1}\bar{y}_r + tn^{-1}\bar{y}_t ;$$

then, the expectation and variance of $\bar{y}_{HD2}$ given y_r are:

$$E(\bar{y}_{HD2} \mid y_r) = \bar{y}_r$$

and:

$$Var(\bar{y}_{HD2} \mid y_r) = t/n(1 - t/r)s_r^2 / n . \qquad (3.5)$$

Finally from (3.4), (3.5) and assuming missing completely at random (MCAR) for missing data mechanism, from Rubin, (1987) we have

$$E(\bar{y}_{HD2}) = \bar{Y}$$

and

$$Var(\bar{y}_{HD2}) = (1/r - 1/N)S_Y^2 + (t/n)(1 - t/r)S_Y^2 / n , \qquad (3.6)$$

where S_Y^2 is the population variance. Therefore, from (Little and Rubin, 1987) the variance of sample mean using without replacement sampling and hot deck imputation is smaller than the variance of sample mean with replacement.

Sequential hot deck is one of the popular hot deck imputation families. In this method, data is ordered according to similarities between cases, and then the nearest case in sequence is selected to impute a missing value. Estimation of data parameters depends on the type of data ordering. If the data is ordered randomly, according to Bailer, Bailey and Corby (1978), the estimator of $\bar{Y}$, say $\bar{y}_{HD3}$, is unbiased for $\bar{Y}$ and the variance of $\bar{y}_{HD3}$ is:

$$Var(\bar{y}_{HD3}) = \left[1 + \left(\frac{2(n-r)}{n}\right)\left(\frac{rn+n-1}{(r+1)(r+2)}\right)\right] S_Y^2 / n \,, \tag{3.7}$$

It can be seen from (3.7) that the variance of sequential hot deck with random ordering is larger than the variance of traditional hot deck methods if $r > 0$. Sometimes the missing value occurs at the beginning of a data file; hence, in this case an initial value is needed for imputation. This initial value is usually selected from other sources, such as previous similar studies, or can be selected from the population, independently of the other responses in the sample. Formula (3.7) can be approximated (Bailar, Baily, and Corby, 1978) for large n by:

$$Var(\bar{y}_{HD3}) = \left[1 + \frac{2(n-r)}{n}\right] S_Y^2 / n \quad . \tag{3.8}$$

Another type of sequential hot deck is when the data file is not randomly ordered. In this case there is a serial correlation between cases i and j. Suppose that the serial correlation structure can be modelled by:

$$Cov(y_i, y_j) = \sigma^2 \delta^{|i-j|}, \qquad i, j = 1,2,...,n \quad . \tag{3.9}$$

Bailer and Bailar (1978) showed that the variance of sample mean ($\bar{y}_{HD4}$) is:

$$\begin{aligned} Var(\bar{y}_{HD4}) = S_Y^2 / n^2 \Bigg[& n + 2\frac{(n-r)(nr+n+r)}{(r+1)(r+2)} \\ & + 2\Bigg\{ \frac{(n+1)(r)}{r+1} \frac{\delta - \delta^n}{1-\delta} \frac{\delta}{(1-\delta)^2} \left(1 - n\delta^{n-1} + (n-1)\delta^n\right) \\ & + \frac{\delta^{r+1}}{\binom{n}{n-r}(1-\delta)^{r+1}} \sum_{i=r+2}^{n} \binom{n}{i} (1-\delta)^i \delta^{n-1} \\ & - \frac{r}{\binom{n}{n-r}} \sum_{i=1}^{n} \binom{n+1-i}{r+1} \delta^i \Bigg\} \Bigg] \quad . \end{aligned} \tag{3.10}$$

Formula (3.10) can be approximated for a large n by

$$Var(\bar{y}_{HD4}) \cong S_Y^2 / n\left[1+\frac{2(n-r)}{r}+2\left(\frac{\delta}{1-\delta}-\frac{n-r}{n}\frac{2\delta}{2-\delta}\right)\right]. \qquad (3.11)$$

The formula (3.11) is not monotonic in δ, and as a result, hot deck imputation methods have a greater variance for the sample mean than the mean imputation method; however, this method preserves the distribution of the data better than mean imputation (Little and Rubin, 1987).

Reducing non-response bias is one of the important aims of hot deck imputation methods. In the random hot deck methods, the data are classified into homogenous categories; then for every missing value, a donor is chosen randomly among its category. Hence, it is expected that the non-response bias will reduce by increasing the homogeneity of categories and decreasing the number of cases in each category. We justify the reduction of non-response bias by the following example:

Consider a finite population of size N and a sample with size of n; let y be a variable of interest, and suppose s_r is the set of respondents to item y, of size r, and s_{n-r} the set of non-respondents with size n-r. In addition, suppose B is the non-response bias if we use only respondents to estimate the mean. Hence, we have:

$$B = E(\bar{y}_r) - \bar{Y},$$

where $\bar{y}_r = \frac{1}{r}\sum_{i \in s_r} y_i$. We use hot deck imputation to impute missing values in each category. Suppose the sample units are categorized into K homogenous groups; also suppose that $p_k = n_k / n$ and $\bar{y}_k = \frac{\sum_{i=1}^{n_k} y_i}{n_k}$ $(k=1,2,..,K)$ are the proportion and mean of sample units in group k. Let $\bar{y}_{HD}$ be the estimator of the sample mean by any hot deck method; then the non-response bias is

$$B_{HD} = E(\bar{y}_{HD}) - \bar{Y}.$$

The expectation of $\bar{y}_{HD}$, given constant $\bar{y}_k$ and p_k, can be written as follows:

$$E(\bar{y}_{HD}) = E\left[E\{\bar{y}_{HD} \mid sample\}\right] = E\left[\sum_{k=1}^{K} p_k \bar{y}_k\right] = \sum_{k=1}^{K} E\left[p_k \bar{y}_k\right]; \tag{3.12}$$

then, from the independence of p_k and $\bar{y}_k$, we have:

$$E(\bar{y}_{HD}) = \sum_{k=1}^{K} E(p_k)E(\bar{y}_k) = \sum_{k=1}^{K} p_k^* E(\bar{y}_k), \tag{3.13}$$

where $p_k^* = E(p_k)$. Therefore, the non-response bias for the hot deck estimator is:

$$B_{HD} = \sum_{k=1}^{K} p_k^* E(\bar{y}_k) - \bar{Y} = \sum_{k=1}^{K} p_k^* E(\bar{y}_k - \bar{Y}_k) = \sum_{k=1}^{K} p_k^* B_k, \tag{3.14}$$

where $\bar{Y}_k$ and B_k are the population mean and its bias in category k. From (3.14), it can be seen that, by increasing the homogeneity in each category, B_k decreases and the non-response bias (B_{HD}) decreases.

3.6 Model-donor imputation methods

In the model-donor based imputation methods, imputed values are directly derived from a model. This procedure builds a predictive model for a variable with missing values based on all of the available information. Some of the popular model-donor based procedures appear below.

3.6.1 Mean and mode imputation:

Substitute variables (mean or mode values) are computed from available cases to fill in for missing data values in the remaining cases, based on the structure of data. The rationale for assigning the mean and mode is that, without any other knowledge about a case, the best guess of their score on any variable is the mean (for skewed variables, it may be better to substitute the median). This method can provide unbiased estimates for the population means or totals in two situations. The first is when missing values are missing completely at random (MCAR), and the second is when data are

approximately normally distributed and missing values are missing at random (MAR) (see Little & Rubin 1987).

An improvement over mean or mode imputation is to impute the mean or the mode for groups that are known to be relatively homogenous. This approach is reasonable when the grouping variable is significantly correlated with the variable with missing data. This method can give unbiased estimates for the population mean, mode or total if the missing values only depend on the auxiliary variables that are used to construct the homogenous groups. From (3.14), it can be seen that by having homogenous categories, reduction in non-response bias is expected. The approach is much better than simple mean or mode imputation, because the imputed values and estimated parameters are more efficient; but the variance is underestimated. However, the distribution of data will be distorted substantially, and the concentration of all imputed values in group means, creates spikes in the distribution. Some mathematical properties of the mean imputation method are as follows:

Consider a finite population of size N and let y be a variable of interest. Suppose the aim is to estimate the mean and variance of the population. The formulae for the mean and variance of the population are $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$ respectively. Then a simple random sample without replacement of size n is derived. In the complete case, the unbiased estimators of the mean and variance of population are $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ respectively.

Suppose s_r is the set of respondents to item y, of size r, s_{n-r} is the set of non-respondents with size n-r and y_i^* denotes the imputed value for missing value y_i. Then the estimators for the sample mean and variance with imputed values are:

$$\bar{y}_{mi} = 1/n\left[\sum_{i\in s_r} y_i + \sum_{i\in s_{n-r}} y_i^*\right], \tag{3.15}$$

and

$$s_{mi}^2 = 1/n-1\left[\sum_{i\in s_r}(y_i - \bar{y}_{mi})^2 + \sum_{i\in s_{n-r}}(y_i^* - \bar{y}_{mi})^2\right], \tag{3.16}$$

where $\bar{y}_{mi}$ and s_{mi}^2 are the sample mean and variance with imputed values. By imputing the missing values with the mean of respondents that is $y_i^* = \bar{y}_r$, $i \in s_{n-r}$, and replacing $\bar{y}_r$ in (3.15) and (3.16) we have:

$$\bar{y}_{mi} = \bar{y}_r$$

and

$$s_{mi}^2 = 1/n-1\left[\sum_{i\in s_r}(y_i - \bar{y}_r)^2\right] = \frac{r-1}{n-1}s_{yr}^2,$$

where $s_{yr}^2 = 1/r-1\left[\sum_{i\in s_r}(y_i - \bar{y}_r)^2\right]$ is the variance of respondents. If we suppose that the respondents are simple random samples from the population, in other words, under the MCAR assumption for missing values, we have:

$$E(\bar{y}_{mi}) = \bar{Y} \tag{3.17}$$

and

$$E(s_{mi}^2) = \frac{r-1}{n-1}S_y^2 \approx \frac{r}{n}S_y^2 \leq S_y^2 \tag{3.18}$$

where $\bar{y}_{mi}$ is an unbiased estimator for $\bar{Y}$ but s_{mi}^2 is a biased estimator for S_y^2. In other words the mean imputation procedure, apart from creating a spike at $\bar{y}_r$, decreases the variance of y.

3.6.2. Nearest neighbour imputation (NNI)

Generally, NNI methods are divided into two categories, the first is the traditional NNI and the second is k-Nearest neighbour imputation. The following section briefly reviews both methods.

a) Nearest neighbour imputation

Nearest neighbour imputation is a type of hot deck imputation. The NNI procedure uses a metric to define the nearness between a recipient and a donor or donors; based on auxiliary variables (it is supposed that the auxiliary variable or variables are complete). Then a donor or donors are chosen according to the metric. There is a comprehensive discussion by Chen and Shao (2000) about NNI. They test the theory of NNI in a bivariate case by using the Euclidian distance between the donor and recipients. In the simplest case, suppose there is a bivariate sample $(x_1, y_1),...,(x_n, y_n)$ where X is complete and missing values occur only in Y. For simplicity, suppose the first r values of Y to be observed and the rest of them $(m = n - r)$ are missing. According to Chen and Shao (2000), in the NNI method, a missing y_j, $r+1 \le j \le n$ is imputed by y_i, $1 \le i \le r$ if:

$$\left|x_i - x_j\right| = \min\left|x_l - x_j\right| \ , \ 1 \le l \le r . \tag{3.19}$$

When more than one donor exists by criterion (3.19), a donor can be randomly selected. If $\tilde{y}_j$, $r+1 \le j \le n$ are imputed values, then the NNI sample mean is:

$$\bar{y}_{NNI} = \frac{1}{n}\left(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \tilde{y}_i\right) = \frac{1}{n}\sum_{i=1}^{r}(1+d_i)y_i , \tag{3.20}$$

where d_i is the number of times that case i is used as a donor. Chen and Shao (2000) showed that the NNI sample mean is asymptotically an unbiased estimator for the population mean under simple random sampling and stratified sampling. NNI was generalized to multivariate data by Murthy and Hossain (2003), with possibly mixed types of variable, such as nominal, ordinal, binary, categorical, and interval. They used the following formula to define the closeness between a complete case c and a missing case m.

$$D(c,m) = \frac{\sum_{j=1}^{p-1} \delta_{cm}^{j} d_{cm}^{j}}{\sum_{j=1}^{p-1} \delta_{cm}^{j}} \ , \tag{3.21}$$

where

$$d_{cm}^{j} = d_{cm}^{j}(x_{cj}, x_{mj}) = \begin{cases} 1 & \text{if } x_{cj} \neq x_{mj}, \text{ when } j^{th} \text{ covariate is binary or nominal} \\ 0 & \text{otherwise , when } j^{th} \text{ covariate is binary or nominal} \\ \dfrac{|x_{cj} - x_{mj}|}{R_j} & \text{when the } j^{th} \text{ covariate is continuous} \end{cases}$$

δ_{cm}^{j} is an indicator variable that is zero when the j^{th} covariate is symmetric and $x_{cj} = x_{mj} = 0$, otherwise it is one. R_j is the range of the j^{th} covariate and p is the number of variables. In (3.21), we assume p and p-1 variables are observed for case c and case m respectively. Therefore, there are p-1 terms in (3.21), because we have only p-1 available variables.

b) k-Nearest neighbour imputation

This method is similar to the nearest neighbour imputation in terms of finding the closest value to a missing value, based on the distance function. The only difference is that in this method we find k neighbours rather than one neighbour. Again, to find nearness, we need similarity and dissimilarity measures. After finding the k nearest neighbours for a missing value, a function of k neighbours is used to estimate the missing value. Where more than one case is being used, there are various possibilities, the simplest being a straightforward arithmetic mean and alternatively an inverse distance weighted algorithm. However, the common functions are the mean and median. For categorical variables, a similarity measure is used. Typically, k is one or two but as it tends to r, where r is the number of available cases, NNI coincides with mean imputation if all of available cases are used as nearest neighbours and the mean function is used to impute. The obvious advantage of this technique is that it enables the most similar cases to be the most influential. In addition, many distance metrics will allow the inclusion of categorical and continuous variables. A number of studies have reported good results using *k-NN,* including the investigation by Chen (2000).

As a result, we can categorize this method as a model donor based imputation method because of the use of the function of available cases for imputing missing values.

3.6.3 Regression imputation methods

One of the traditional imputation methods is the regression imputation procedure. In this method, a missing value is replaced by a value predicted from a regression model based on observed values. The main assumptions in regression imputation are the MAR assumption, and the specification of the regression function. In regression imputation, there is no distinction between response and independent variables, in other words every variable with missing values is considered as a dependent variable. Generally, RI (regression imputation) is divided into two main categories: deterministic regression imputation method and random or stochastic regression imputation method. In this section, two popular regression imputation methods are reviewed.

a) Deterministic and random linear regression imputation

We start with a simple regression with one response variable Y and one independent variable X and missing values occur in the variable Y. It is assumed that r cases are available and n-r cases are missing in the response variable. This method imputes the missing value by a prediction from the linear model built using the available cases. Here MAR is implicitly assumed.

Suppose the regression model is as follows:

$y_i \sim N(\mu_i, \sigma^2)$, $\mu_i = E(y_i \mid x_i) = \alpha + \beta x_i$ and the unknown parameters are $\theta=(\alpha, \beta, \sigma^2)$.

Then the regression model is:

$$y_i = \alpha + \beta x_i + \varepsilon_i \ , \ i=1,\ldots,n$$

where $\varepsilon_i \sim N(0, \sigma^2)$. The deterministic linear regression imputation of y_i $(i=r+1,...,n)$ is

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i, \; i=r+1,...,n. \tag{3.22}$$

This method underestimates the variance of the sample mean. Therefore, to avoid this underestimating problem, an error term is added to the imputed value. In this case, the method is known as a random linear regression imputation. Thus, we have:

$$\hat{y}_i = \hat{\alpha} + x_i\hat{\beta} + \hat{\varepsilon}_i, \; i=r+1,...,n, \tag{3.23}$$

where $\hat{\varepsilon}_i$ is added to the prediction by the regression model. This is done to conserve the distributional properties of the target variable Y. The residual term can be obtained from a random draw from the residuals of available cases. Finally, different regression imputation methods can be used according to the type of response variable, such as logistics regression imputation, multinomial regression imputation, and nonparametric imputation.

3.6.4 Multiple imputation

The general idea of multiple imputation (MI) is to impute missing data via a random process that reflects uncertainty about the actual value being imputed. Simply, we randomly impute M values for a missing value, and thus create M complete but different datasets. According to Rubin (1987) in the simplest cases, a Monte Carlo simulation approach is used to complete the dataset with M>1 simulated values (where typically 3<M<10), leading to M complete datasets. Each of the M complete datasets is then analysed, and the results are combined using simple rules into a result comprising estimates and standard errors. However, the choice of imputation model should be compatible with the analyses to be performed and should preserve relationships between the variables that are to be investigated. Thus, the imputation model should at least use all of the variables available to the analysis model. MI techniques generally assume data missing at random MAR.

In comparison with MI, single imputation has two advantages: first, standard complete data analysis methods can be used easily; secondly, in the context of public-use databases, the possibly substantial effort required to create sensible imputations needs to be made only once, by the data producer, and these imputations can

incorporate the data collector's knowledge. There is also an important disadvantage of single imputation: the single value being imputed reflects neither sampling variability about the actual value when one model for non-response is being considered, nor additional uncertainty when more than one model is being entertained.

An efficient MI is based on three assumptions. These assumptions are related to the population of data values, the prior distribution for the model parameters, and the missing data mechanism. The first assumption means that, to conduct MI, a probability model for data (observed and missing values) should be assumed. The popular models are the multivariate normal model, loglinear models, the general location model, and the two-level linear regression model. The second assumption means that a prior distribution for the parameters of the imputation model should be assumed. This assumption is directly related to Bayes theory. However, Schafer (1997) claims that an efficient MI is more sensitive to the choice of the data model than the choice of the prior. In addition, he says that, for large samples, any reasonable prior distribution gives the same results. Finally, the third assumption pertains to the missing data mechanism, and MI assumes MAR for missing values.

In conclusion, any "proper" imputation method can be used in MI, and it would be easier if the missing pattern were monotone. The suggested proper imputation methods for MI (Schafer, 1997) are the Markov chain Monte Carlo, data augmentation methods and the EM algorithm.

Chapter 4

Sequential taxonomy and its applications

4.1. Introduction

Nearest neighbour imputation methods use variety of techniques to find a donor or donors for a recipient. Examples are ordering methods, multivariate matching, distance function matching, and minimum distance function based on fully observed covariates techniques (see 3.5.2 and 3.6.2). In this thesis ordering methods are used to conduct a new imputation method. There is a general discussion about the ordering methods in appendix 1. Therefore, choosing an efficient ordering method plays an important role in increasing the efficiency of our imputation method. We use sequential taxonomy (ST) to find nearest neighbours for a missing value. ST is a multivariate ordering method which is used for ordering numerical data, based on similarities and dissimilarities between cases. ST uses a variety of measures to order data such as distance or similarity measures. We focus on Euclidian distance because of its simplicity and mathematical properties. This chapter reviews the theory of ST and statistical properties of ST and its applications. In addition, the relationship of ST with the theory of concomitants of order statistics, and the distribution of the distance measure when data are normally distributed will be discussed in this chapter.

4.2 Sequential taxonomy (ST)

Assume that a data matrix X exists with p variables ($p > 1$) and N objects or cases ($N > 1$) drawn from a specific multivariate distribution. The idea behind sequential

taxonomy comes from the ordering of cities based on their levels of development. The definition of a developed city or country is not an absolute concept and is related to the corresponding development of other cities. In order to make a comparison between these different cities, we therefore need to identify a "best city". Differences between any particular city and the best city can be characterised using an appropriate distance measure. In practice, the city with the best in each development indexes for all variables is defined as the ideal city. Note that, the definition of the best value for an index depends on the underlying variable, and can be a maximum value or a minimum value of this variable. The vector defined by the best values is called the ideal vector, and the distance from this vector to the corresponding vectors of the cities of interest can be calculated, and the cities then ordered based on this distance vector. This idea can be easily extended to any multivariate data matrix X with N cases and p variables.

An efficient ST contains four stages: the first is weighting and standardization, the second is finding an ideal vector, the third is calculating distances between cases and the ideal vector, and the final stage is ordering the data file based on these distances. We now describe each of these steps in the next sections.

4.2.1 Standardization

To calculate a distance in sequential taxonomy, we need to calculate the distance from the ideal vector. Most distance measures are sensitive to the scale of measurement variables, and, typically require same scales for all variables. When all the variables are measured on a continuous scale, the solution most often suggested is simply to standardize each variable to a unit variable before analysis. This can be done, for example by standard scoring:

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j}, \quad i=1,2,\ldots,N,\ j=1,2,\ldots,p \tag{4.1}$$

where $\bar{X}_j$ and S_j are the mean and standard deviation of the variable j, respectively. In general, standardization of variables to unit variability can be viewed as a special

case of weighting, and corresponds to giving it greater or lesser importance than other variables when using it to determine the proximity between two cases. The weights chosen for the variables reflect the importance that the investigator assigns to the variables for the classification task. This assignment or weighting is either the result of a judgement by the researcher or based on some aspect of the data matrix, X. In what follows we assume standardization, so the data matrix of X is replace by the data matrix Z, of z scores via (4.1).

4.2.2 Ideal vector

The second stage in sequential taxonomy is defining the ideal vector. This vector consists of specified function of the values of each variable, such as the maximum, minimum, mean or median. That is, the ideal vector is a $1 \times p$ vector.

$$\alpha = (\alpha_1, \alpha_2, ..., \alpha_p),$$

where $\alpha_j = h_j(z_j), j = 1, 2, ..., p$ and $h_j(\cdot)$ is specified function of the N values of variable Z_j. For example, if all the variables had the same direction then the ideal vector could be defined as

$$\alpha_j = \max_i(z_j), j = 1, 2, ..., p,$$

where i is the case index. It should be emphasized that the maximum is taken over cases not variables. Under this assumption, all variables represent a direction of increasing development. However, suppose a data file contains countries with three variables, adult literacy rate (ALR), infant mortality rate (IMR), and real GDP per capital (GDP). It seems reasonable to assume that that ALR and GDP have positive direction, but IMR has negative direction. In other words by increasing the values of ALR and GDP the development of a country increases, but by increasing the values of IMR the development of the country decreases. Hence, if the above ideal vector definition is to be used, then the direction of IMR should be changed (e.g. by multiplying by minus one).

4.2.3 Similarity, dissimilarity, and distance measures

The third stage of ST is to calculate the distance or similarity between the ideal vector and the cases of interest. In this section, we review similarity, dissimilarity, and distance measures. Generally, we use similarity measures for categorical variables and mixtures of categorical and continuous variables. However, dissimilarity or distance measures are usually applied with continuous data. In what follows we examine each of these situations separately.

a) Similarity measures for categorical data

When data is categorical, similarity measures are the most common measures used to measure the similarity between individuals. Generally, values of similarity measures lie between zero and one. We represent the similarity between case i and j by s_{ij}, where a zero value for s_{ij} corresponds to the maximum difference between case i and j for all variables. The dissimilarity of case i and case j is then given by $\delta_{ij} = 1 - s_{ij}$. To illustrate, Table 4.1 shows the cases classification of binary outcome variables for two cases as follows:

Table 4.1 frequency Table for case i and case j

		Case i		
	Outcome	*1*	*0*	Total
Case j	*1*	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$p=a+b+c+d$

Where p is the number of binary variables. Some similarity measures that have been proposed for the data in Table 4.1 are set out in Table 4.2.

Table 4.2 Similarity measures for binary data

Measure	Formula
S1 Matching coefficient	$s_{ij} = (a+d)/(a+b+c+d)$
S2 Jaccard coefficient (1901)	$s_{ij} = a/(a+b+c)$
S3 Rogers and Tanimoto (1960)	$s_{ij} = (a+d)/[(a+2(b+c)+d)]$
S4 Sokal and Sneath (1963)	$s_{ij} = a/[a+2(b+c)]$
S5 Gower and Legendre (1986)	$s_{ij} = (a+d)/[a+1/2(b+c)+d]$
S6 Gower and Legendre (1986)	$s_{ij} = a/[a+1/2(b+c)]$

For an extensive discussion of these and other measures of association for binary data, see Kaufman and Rousseeuw (1990).

A categorical variable is a generalisation of the binary variable, which has more than two categories. Here, we only generalise the matching approach (S1) to categorical variables. To define a similarity measure, suppose s_{ijk}, $k=1,2,..,p$ is zero or one depending on whether the two cases i and j disagree or agree on variable k. In other words, we have

$$s_{ijk} = \begin{cases} 1 & \text{if case } i \text{ and } j \text{ match in variable } k \\ 0 & \text{otherwise} \end{cases},$$

where $k=1,2,....,p$. Therefore from Gower and Legendre (1986) a generalized similarity measure for categorical variables based on the matching approach is

$$s_{ij} = \left(\sum_{k=1}^{p} s_{ijk} \right) \Big/ p . \qquad (4.2)$$

b) Dissimilarity and distance measures for continuous data

Generally, dissimilarity and distance measures are used when variables are continuous. Let δ_{ij} represent dissimilarity and d_{ij} a distance measure between case i and case j. We say that the dissimilarity is a distance measure if it fulfils the triangle inequality for pairs of cases (i,j), (i,m) and (j,m):

$$\delta_{ij} + \delta_{im} \geq \delta_{jm}$$
$$\delta_{ij} = \delta_{ji}, \forall i, j \qquad (4.3)$$
$$\delta_{ij} \geq 0$$

There is a variety of distance measures in the literature, see Gower and Legendre (1986). Table 4.3 shows the commonly used distance measures when $X_1, X_2, ..., X_P$ are continuous random variables with n observations.

Table 4.3 Distance measures for continuous data

Measure	Formula
D1 Euclidian	$d_{ij} = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{1/2}$
D2 City block	$d_{ij} = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert$
D3 Minkowski	$d_{ij} = \left(\sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert^r \right)^{1/r}$
D4 Canberra	$d_{ij} = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert / \left(\lvert x_{ik} \rvert + \lvert x_{jk} \rvert \right)$ for $x_{ik} \neq 0$ or $x_{jk} \neq 0$
D5 Pearson correlation	$\delta_{ij} = (1 - \phi_{ij})/2$ with $\phi_{ij} = \sum_{k=1}^{p} (x_{ik} - \bar{x}_{i.})(x_{jk} - \bar{x}_{j.}) \Big/ \left(\sum_{k=1}^{p} (x_{ik} - \bar{x}_{i.})^2 \sum_{k=1}^{p} (x_{jk} - \bar{x}_{j.})^2 \right)^{1/2}$ where $\bar{x}_{i.} = 1/p \sum_{k=1}^{p} x_{ik}$
D6 Angular separation	$\delta_{ij} = (1 - \phi_{ij})/2$ with $\phi_{ij} = \sum_{k=1}^{p} x_{ik} x_{jk} \Big/ \left(\sum_{k=1}^{p} x_{ik}^2 \sum_{k=1}^{p} x_{jk}^2 \right)^{1/2}$

Generally, dissimilarity measures can be divided into two main categories: distance measures and correlation-type measures. Distance measures are the physical distance between two p dimensional points. For example, the Euclidian distance (D1) is the distance between $X_i' = (x_{i1}, x_{i2}, ..., x_{ip})$ and $X_j' = (x_{j1}, x_{j2}, ..., x_{jp})$ in Euclidian

space. This measure is also known as the l_2 norm. Measures D1 to D4 are examples of distance measures, and D5 and D6 are examples of dissimilarity measures based on correlation coefficients. More details of the advantages and disadvantages of different dissimilarity measures can be found in Gower and Legendre (1986) and Anderberg (1973). This thesis will only use Euclidian distance, because of its simple mathematical properties when data are distributed as multivariate normal.

b) Similarity measures for data containing both continuous and categorical variables

Different methods have been proposed in the statistics literature for calculating similarity measures for mixed-mode data (continuous and categorical data). One of the simplest methods is to categorise the continuous variables and then similarity measures appropriate for categorical data. In this thesis, however, we consider Gower's general similarity measure (Gower and Legendre, 1986), which is defined as follows:

$$s_{ij} = \sum_{k=1}^{p} w_{ijk} s_{ijk} \Big/ \sum_{k=1}^{p} w_{ijk} \tag{4.4}$$

where s_{ijk} is the similarity between case i and case j as measured by the kth variable, and w_{ijk} is a weight that takes the value zero or one, depending on the validity of the comparison. For binary and categorical variables, s_{ijk} is one and zero. For continuous variables, Gower suggests using City block distance (D2), after scaling the kth variable to unit range.

$$s_{ij} = 1 - \left| x_{ik} - x_{jk} \right| / R_k , \tag{4.5}$$

where R_k is the range for the variable k.

4.2.4 Ordering the data

In what follows we assume that the data have a multivariate normal distribution and so use Euclidian distance to find the distance between cases and the ideal vector. In particular, this Euclidean distance is calculated as:

$$d_i(\alpha)=\left(\sum_{j=1}^{p}\left(z_{ij}-\alpha_j\right)^2\right)^{1/2} \tag{4.6}$$

where z_{ij} is the standardized matrix of x_{ij}, $(i=1,2,...,n),(j=1,2,...,p)$, and α_j denote the jth component of the ideal vector.

After obtaining the $d_i(\alpha)$, the data file is ordered according to these values. This type of ordering is actually a special case of reduced (aggregate) ordering, and is sometimes referred to as R ordering, where each multivariate observation is reduced to a single value by means of some combination of the component population or sample values (see Appendix 1, ordering methods). In general the $d_i(\alpha)$ can be obtained by any ordering method or dimension reduction technique, e.g. based on the first component of principal component analysis. In addition, the $d_i(\alpha)$ can be generalized to different types of data; if data is categorical, continuous or a mixture of continuous and categorical, we simply use a different distance measure appropriate to these data types to define the distance vector.

4.2.5 Example

The underlying data here are taken from a 1997 data file issued by the United Nations Statistics Division (UNSD), which covers 85 countries, and gives information on thirteen variables, namely region, HDI, adult literacy (ad_li) rate (%), population growth (popul), contraceptive prevalence (%) (contr), dependency ratio (depen), infant mortality rate (per 1000 live births) (inf_m), life expectancy (l_exp) at birth (years), maternal mortality rate (per 10000 live) (mater), real GDP per capital (PPP$) (gdp), education index, GDP index and life expectancy index. Our aim is to order these 85 countries using sequential taxonomy based on their development and to then

compare this ordering with the ordering of the same 85 countries based on these Human Development Index (HDI). To start, we note that adult literacy rate, contraceptive prevalence, life expectancy at birth and real GDP per capita all have a positive direction, while population growth, dependency ratio, infant mortality rate and maternal mortality rate all have a negative direction. These negative direction variables are therefore converted to positive direction by multiplying by -1. All variables are then standardized:

$$Z_{ij} = \frac{X_{ij} - \overline{X}_{\cdot j}}{s_{\cdot j}},\ i = 1,2,...,85 \text{ and } j = 1,2,..,8,$$

where $\overline{X}_{\cdot j}$ and $s_{\cdot j}$ are the mean and standard deviation of the j^{th} variable. The ideal vector is then defined as the vector of maximum values of each variable and the Euclidian distance between each of the 85 countries and the ideal vector is computed. These distances are referred to as ST in what follows. Tables 4.4 and 4.5 show the correlations between ST, the component variables, HDI, and a number of other UN indexes that can be used to order the 85 countries, such as education index (EDI), GDP index (GDPI) and life expectancy index (LEI). It can be seen from Table 4.4 that there are high negative correlations between ST and the main variables. Therefore, by increasing the development variables, ST decreases, which means small values of ST show more developed countries.

Table 4.4: Correlation matrix between main variables and ST

	ad_li	popul	contr	depen	inf_m	l_exp	mater	gdp	ST
Ad_li	1.000	0.644	0.733	0.643	0.786	0.697	0.780	0.667	-0.835
Popul	0.644	1.000	0.666	0.741	0.688	0.640	0.620	0.628	-0.826
Contr	0.733	0.666	1.000	0.775	0.798	0.753	0.755	0.658	-0.870
Depen	0.644	0.741	0.775	1.000	0.734	0.754	0.643	0.683	-0.869
Inf_m	0.786	0.687	0.798	0.734	1.000	0.846	0.811	0.693	-0.897
L_exp	0.697	0.640	0.753	0.754	0.846	1.000	0.772	0.670	-0.858
mater	0.780	0.620	0.755	0.643	0.811	0.772	1.000	0.671	-0.848
gdp	0.667	0.628	0.658	0.683	0.693	0.670	0.671	1.000	-0.864
ST	-0.835	-0.826	-0.870	-0.869	-0.897	-0.858	-0.848	-0.864	1.000

It can be seen from Table 4.5 that there is a very strong correlation between ST, HDI and other indexes. The correlation between ST and HDI is -0.955, which means that 92 % of the variation of HDI can be explained by ST.

Table 4.5: Correlation matrix between ST, HDI, and other indexes

	ST	HDI	EDI	GDPI	LEI
ST	1.0000000	-0.9545449	-0.8568659	-0.9244939	-0.8589800
HDI	-0.9545449	1.0000000	0.9196896	0.9265910	0.9077694
EDI	-0.8568659	0.9196896	1.0000000	0.7984540	0.7089641
GDPI	-0.9244939	0.9265910	0.7984540	1.0000000	0.7906557
LEI	-0.85898	0.9077694	0.7089641	0.7906557	1.0000000

Correlation may not be meaningful without linear association. Figure 4.1 and 4.2 show pairwise scatterplots for ST, the component variables, HDI and the other indices.

Figure 4.1: Pairwise plots for the main variables and ST

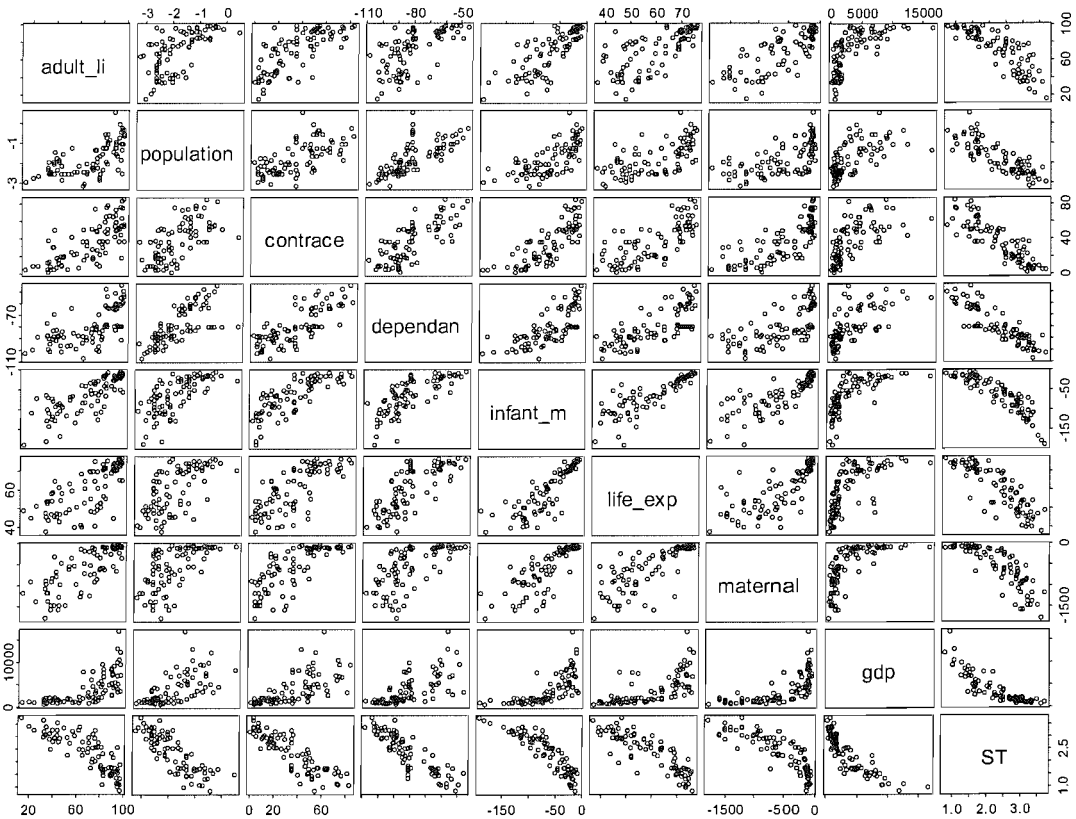


Figure 4.2: Pairwise plots for ST, HDI and other indexes

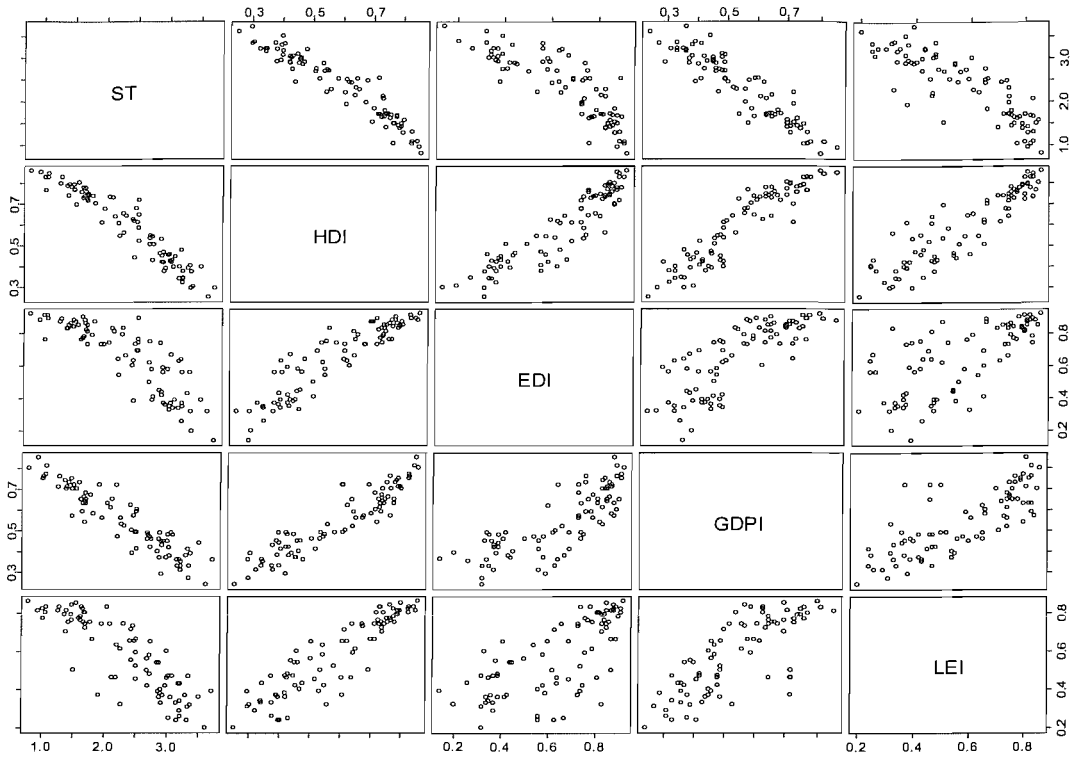


Figure 4.2 shows a strong linear relationship between ST, HDI and other indices. That is, for these data, sequential taxonomy preserves the ordering relationships defined by HDI and the other development indices.

4.3 Distribution theory for distance measures

Before we can develop a theory for sequential taxonomy imputation, we need to know the distribution of the distance vector. In this section we therefore obtain the distribution of d_i given X is normally distributed. For simplicity, we assume a bivariate situation where the data consist of independent measurements on two random variables X and Y. To find the distribution of (4.6), we start from the simple case where $X \sim N(0, \sigma_X^2)$, $Y \sim N(0, \sigma_Y^2)$ and $R = \sqrt{(X-h_X)^2+(Y-h_Y)^2}$ is the distance from an arbitrary and fixed point (h_X, h_Y) to (X,Y). Here (h_X, h_Y) corresponds to fixed maximum points of X and Y or to an ideal vector. Our aim is to find the cumulative distribution $F_R(r|h_X, h_Y)$ and density function $f_R(r|h_X, h_Y)$ for random variable R.

Case one:

Suppose $\sigma_X=\sigma_Y=\sigma$ and $h=\sqrt{h_X^2+h_Y^2}=0$ then

$$F_R(z)=P(R\le z)$$

$$=\left(\frac{1}{2\pi\sigma^2}\right)\iint_{x^2+y^2\le z^2}\exp\left(-\left(x^2+y^2\right)/2\sigma^2\right)d_x d_y$$

$$=\left(\frac{1}{2\pi\sigma^2}\right)\int_0^{2\pi}\int_0^z \exp\left(-r^2/2\sigma^2\right)\,r d_r d_\theta\,.$$

Integrating over θ and making the change of variable $u=r^2/2\sigma^2$, so $du=rdr/\sigma^2$ and $dr=\sigma^2 du/r$, we have:

$$F_R(z)=1/\sigma^2\int_{u=0}^{z^2/2\sigma^2}\exp(-u)\sigma^2 d_u$$

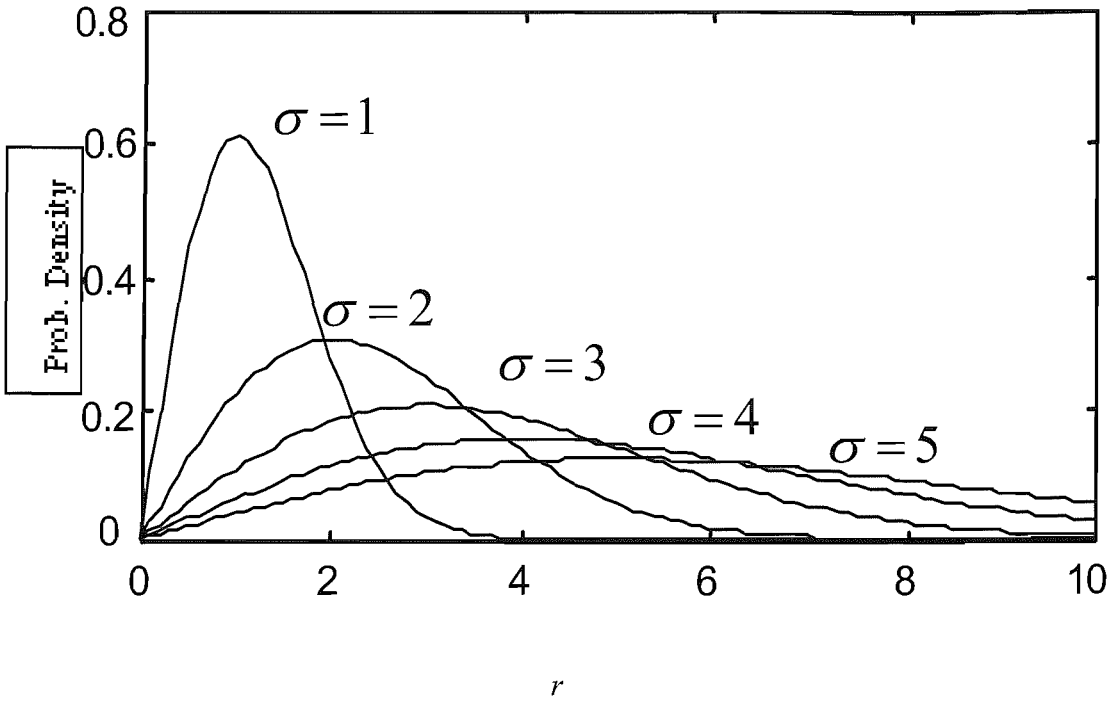
$$=1-\exp\left(-z^2/2\sigma^2\right).$$

Hence,

$$f_R(r)=\frac{d}{dr}F_R(r)=\frac{r}{\sigma^2}\exp\left(-r^2/2\sigma^2\right) \qquad (4.7)$$

This is the Rayleigh density function. $f_R(r)$ has a maximum value at $r=\sigma$ and $E[R]=\sigma\sqrt{\pi/2}\approx 1.2533\,\sigma$ and $Var(R)=\sigma^2(4-\pi)/2=0.4290\sigma^2$ (Johnson, Kotz and Balakrishnan, 1994).

Figure 4.3: Rayleigh distribution for some values of σ



It can be seen from Figure 4.3 that when σ has small values, the Rayleigh distribution is very close to the normal distribution.

Case two:

Ordering data via sequential taxonomy requires an ideal vector that is usually nonzero. Given $\sigma_x=\sigma_y=\sigma$ and $h=\sqrt{h_{x^2}+h_{y^2}}>0$, the distribution of R is as follows (Jagdish, 1996):

$$f_R(r)=\left(\frac{r}{\sigma^2}\right)\exp\left(\frac{-\left(r^2+h^2\right)}{2\sigma^2}\right)I_0\left(\frac{rh}{\sigma^2}\right), \tag{4.8}$$

where $I_0(x)$ is the modified Bessel function of order *0*. This distribution is also called Rice distribution (Johnson, Kotz and Balakrishnan, 1994). Numerical integration is required to evaluate $F_R(r)$. For large x, $I_0(x) \approx e^x/\sqrt{2\pi x}$.

Figure 4.4: Distribution of $f_R(r)$ for some values h and $\sigma = 5$

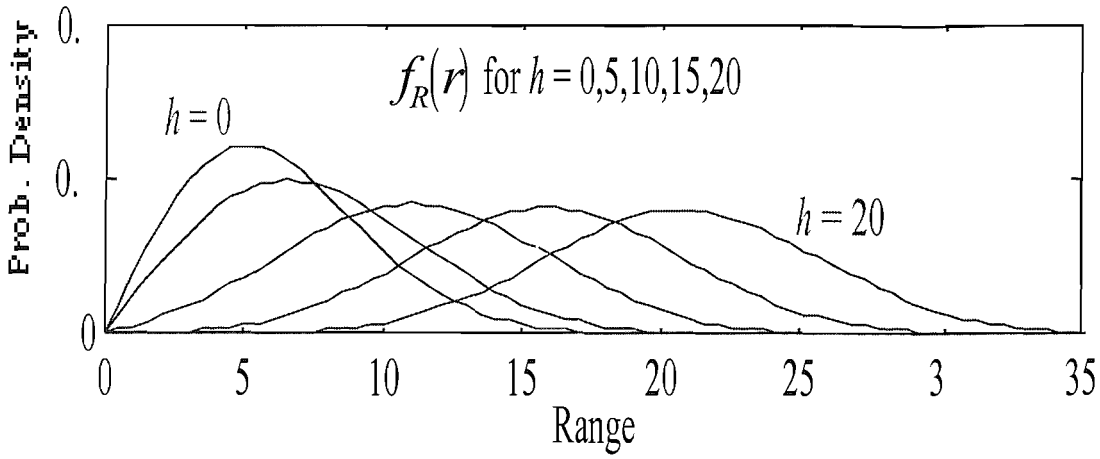


Figure 4.4 shows the distribution of $f_R(r)$ for some values h. when r is farout the tail of the curve or when h is large, the Rice density function in (4.8) behaves like a normal distribution with mean μ and variance σ^2 (Johnson, Kotz and Balakrishnan, 1994).

So far, we have shown that the distance measure or sequential taxonomy coefficients have approximately normally distributed under two assumptions. The first is $\sigma_X=\sigma_Y=\sigma$ and second is independence of X and Y. To calculate sequential taxonomy coefficients as seen in section 4.2.1, the data are standardised. Therefore, the variance of X and Y are equal ($\sigma_X=\sigma_Y=1$). In practice, X and Y are dependent; therefore, to meet the second condition, correlated variables are transformed to uncorrelated variables by standard statistical techniques such as principal component analysis, then sequential taxonomy coefficients are calculated. Finally, to justify using the normal distribution for taxonomy coefficients, we show the normality of these coefficients by a simulation study in section 4.5.

4.4 Sequential taxonomy (ST) and concomitants of order statistics (COS)

Now we focus onto the relationship between ST and COS. In chapter 2, we studied two variables with a joint distribution of $f(x,y)$. According to COS theory, the data are ordered according to the values of the variable X and the corresponding ordered values of Y are termed the concomitants of order statistics. In next chapter, it is assumed that the distance vector in the sequential taxonomy procedure is equivalent to the X variable and Y is a variable with missing values; therefore, the distance variable X is the complete variable and missing values occur only in the Y variable. We know from section 4.3 that the distribution of the distance vector is approximately normal, hence:

$$(X,Y) \sim N_2(\mu_1, \mu_2, \sigma_X^2, \sigma_Y^2, \rho),$$

where μ_1, μ_2 are the means of X and $Y, \sigma_X^2, \sigma_Y^2$ are variances of X and Y, respectively, and ρ is the correlation between X and Y. Based on the theory of concomitants of order statistics the X values are ordered as follows:

$$X_{1:n} \le X_{2:n} \le \cdots \le X_{r:n} \le \cdots \le X_{n:n} \quad .$$

Thus, the corresponding values of each Y based on the ordered X are:

$$Y_{[1:n]} \le Y_{[2:n]} \le \cdots \le Y_{[r:n]} \le \cdots \le Y_{[n:n]}.$$

Suppose there is a linear relationship between X and Y and, in addition, X is a complete and Y is an incomplete variable. More specifically, suppose $Y_{[r:n]}$ is a missing value. Thus, we can find the distribution of $Y_{[r:n]}$ based on the theory of concomitants of order statistics. In the next chapter, more details of imputing missing values and other characteristics of the estimation of missing values are given.

4.5 The simulation study

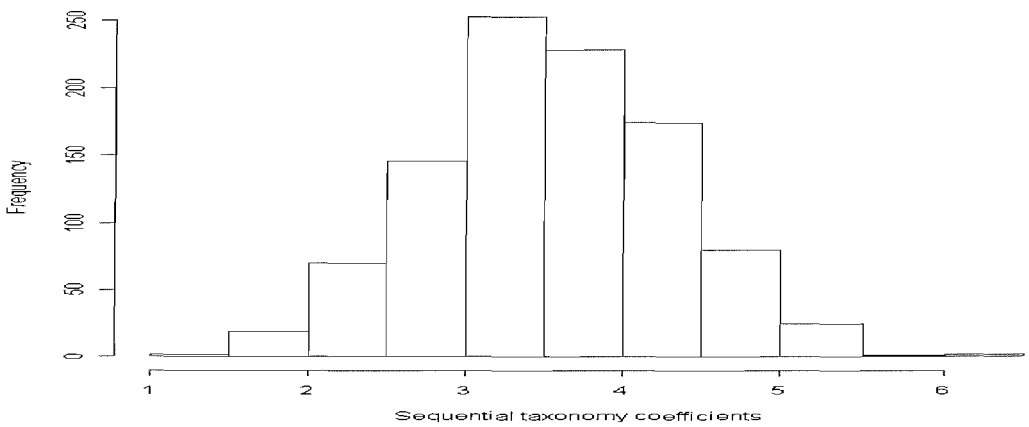
In this section, a simulation study is carried out to show the approximate normality of sequential taxonomy coefficients. A multivariate normal database with 1000 cases was generated according to the mean vector and the correlation matrix in section 5.8. The generated multivariate data contains five variables with no missing values. All variables are then standardized:

$$Z_{ij} = \frac{X_{ij} - \overline{X}_{.j}}{s_{.j}},\ i = 1,2,...,1000 \text{ and } j = 1,2,..,5,$$

where $\overline{X}_{.j}$ and $s_{.j}$ are the mean and standard deviation of the j^{th} variable. The ideal vector is then defined as the vector of maximum values of each variable and the Euclidian distance between each of the 1000 cases and the ideal vector is computed. These distances are referred to as ST in what follows. The number of iterations was 2000.

To test the normality of ST there are different normality tests such as the Chi-square good-ness of fit, and the Kolomogrov-Simirnov test, but here we use the Shapiro-Wilk test because of the simplicity and availability in the R package. The Shapiro-Wilk test, proposed by Shapiro and Wilk (1965), calculates a W statistic that tests whether a random sample, comes from (specifically) a normal distribution. Shapiro-Wilk test is one of the most powerful normality tests, especially for small samples. In each iteration, the p-value was calculated and if it was higher than a significance levels (0.05), then we accepted null hypothesis that is ST has a normal distribution. In 2000 iterations, the percentage of insignificant Shapiro-Wilk test was 94.94%. Therefore, our simulation result shows approximate normality of ST. Figure 4.5 shows the histogram of ST in the last iteration.

Figure 4.5: Histogram of sequential taxonomy coefficients in the last iteration.



Chapter 5

Sequential taxonomy imputation for normal data

5.1 Introduction

In this chapter, a new imputation method is developed based on an integration of sequential taxonomy (ST), concomitants of order statistics (COS) and nearest neighbour imputation (NNI) theory. The method is referred to as sequential taxonomy imputation (STI) in what follows. STI may be categorised as a nonparametric imputation method because of its similarities to NNI, but here we want to apply this method to parametric models. It is similar to hot deck imputation methods, especially sequential hot deck, when data are ordered non-randomly. STI can be implemented in five steps: (1) calculating a distance variable (vector) using all available data, (2) identifying the variable with missing values as a concomitant variable for the distance variable, (3) ordering both variables according to the distance variable, (4) finding the locations and nearest neighbours of the missing values in the ordered data, and (5) imputing missing values using nearest neighbour methods. It should be emphasised that the distance variable X is assumed to be complete with missing values only in its concomitant variable Y. For general definitions of nearness, see sections 3.5.2 and 3.6.2. It is also assumed that the k nearest neighbours for a case with missing value $Y_{[j:n]}$ are all available, consisting of $k/2$ cases "below" this case in the concomitant ordering and $k/2$ cases "above" this case in the ordering.

Like any imputation method, STI is based on assumptions. In this chapter, we assume a linear relationship between the distance variable X and its concomitant Y. Two versions of STI are developed in what follows, the first is single order sequential taxonomy imputation (SSTI), and the second is double order sequential taxonomy imputation (DSTI).

5.2 Single order sequential taxonomy imputation (SSTI)

An imputation method called single order sequential taxonomy imputation (SSTI) is proposed here. SSTI imputation uses one complete auxiliary variable to order data according to data ordering methods. We assume X is a complete variable and missing values only occur in Y. The method developed here assuming a bivariate normal distribution for X and Y. To formulize this method we us assume (X_i, Y_i) $(i = 1,2,...,n)$ is an i.i.d. random sample from a bivariate normal distribution

$$(X_i, Y_i) \sim N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho),$$

where $E(X) = \mu_X, E(Y) = \mu_Y, Var(X) = \sigma_X^2$, $Var(Y) = \sigma_Y^2$ and $Cor(X,Y) = \rho$. It then follows that we can write

$$Y_i = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X_i - \mu_X) + Z_i, \tag{5.1}$$

where X_i and Z_i are mutually independent, with $E(Z) = 0$ and $Var(Z) = \sigma_Y^2(1-\rho^2)$. The data is now ordered according to X, and, based on the concomitant statistics theory set out in Chapter 2, we have

$$Y_{[r:n]} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X_{r:n} - \mu_X) + Z_{[r:n]}, \tag{5.2}$$

where $Z_{[r:n]}$ denotes the concomitant of Z_i with respect to $X_{r:n}$. From the independence of X_i and Z_i, we can conclude that $X_{r:n}$ and $Z_{[r:n]}$ are independent. Here we use the same notation as in Chapter 2, so by setting

$$\alpha_{r:n} = E(\frac{X_{r:n} - \mu_X}{\sigma_x}) \text{ and } \beta_{rs:n} = Cov(\frac{X_{r:n} - \mu_X}{\sigma_X}, \frac{X_{s:n} - \mu_X}{\sigma_X}),\ r,s = 1,2,...,n,$$

then, from (5.2) we have:

$$E(Y_{[r:n]}) = \mu_Y + \rho\sigma_Y\alpha_{r:n}, \tag{5.3}$$

$$Var(Y_{[r:n]}) = \sigma_Y^2(\rho^2\beta_{rr:n} + 1 - \rho^2), \tag{5.4}$$

$$Cov(X_{r:n}, Y_{[s:n]}) = \rho\sigma_X\sigma_Y\beta_{rs:n}, \tag{5.5}$$

$$Cov(Y_{[r:n]}, Y_{[s:n]}) = \rho^2\sigma_Y^2\beta_{rs:n} \qquad r \neq s, \tag{5.6}$$

where $\alpha_{r:n}$ is the expectation of the r^{th} of order statistics and $\beta_{rs:n}$ is the covariance of r^{th} and s^{th} order statistics from standard normal distribution. In order to calculate the expectation and variance of $Y_{[r:n]}$, we therefore need to calculate the expectation and variance of the r^{th} order statistic of X.

To construct a mathematical theory under SSTI, suppose $Y_{[r:n]}$ is a missing value and we wish to impute $Y_{[r:n]}$ by the k nearest neighbours of $Y_{[r:n]}$, as defined by the concomitant ordering of Y in terms of X. A simple predictor of this missing value is the average of its two nearest Y values, or more generally, the average of the k nearest neighbours of $Y_{[r:n]}$, defined as the $k/2$ values above and the $k/2$ values below $Y_{[r:n]}$. If there are missing values in these k nearest neighbours of $Y_{[r:n]}$, then the next complete cases in the ordering are used until a total of k are identified. It should be noted that may be missing values in the k nearest neighbours of $Y_{[r:n]}$. In this case, we have to re-index the k neighbours and include the k nearest values with non-missing values. However, in this case, the k neighbours may no longer be balanced ($k/2$ values above and below of a missing value). Therefore, existence of these neighbours is important, and in the case of unbalanced situations it may decreases the efficiency of our proposed imputation method. The following developed theory for SSTI is based on the availability of all k neighbours, and the effect of missing values needs further research. We therefore assume that the k nearest neighbours of $Y_{[r:n]}$ are complete and given by

$$Y_{[r-i:n]} \text{ and } Y_{[r+i:n]}, \quad i = 1,2,\ldots,k/2. \tag{5.7}$$

An imputation for $Y_{[r:n]}$ based on these k neighbours is then

$$\hat{Y}_{[r:n]} = \frac{\sum_{i=1}^{k/2} (Y_{[r-i:n]} + Y_{[r+i:n]})}{k} \quad . \tag{5.8}$$

To illustrate (5.8), we investigate the simple case $k=2$, where

$$\hat{Y}_{[r:n]} = \frac{Y_{[r-1:n]} + Y_{[r+1:n]}}{2} . \tag{5.9}$$

That is, the imputation for $Y_{[r:n]}$ is the average of the Y values for the nearest complete cases above and below this case in the concomitant ordering. In the next two sections, some statistical properties of (5.8) and (5.9) are established.

5.2.1 Expectations of $\hat{Y}_{[r:n]}$ for two and k nearest neighbours

From (5.9) the expectation of $\hat{Y}_{[r:n]}$ when $k=2$ is

$$E(\hat{Y}_{[r:n]}) = \left[E(Y_{[r-1:n]}) + E(Y_{[r+1:n]}) \right] / 2 . \tag{5.10}$$

By substituting the expectation of $Y_{[r:n]}$ from (5.3) in (5.10), we see that

$$\begin{aligned} E(\hat{Y}_{[r:n]}) &= \left[E(\mu_Y + \rho\sigma_Y \alpha_{r-1:n}) + E(\mu_Y + \rho\sigma_Y \alpha_{r+1:n}) \right] / 2 \\ &= \mu_Y + \rho\sigma_Y \left(\frac{\alpha_{r-1:n} + \alpha_{r+1:n}}{2} \right) \quad , \end{aligned} \tag{5.11}$$

where $\alpha_{r-1:n} = E\left(\frac{X_{r-1:n} - \mu_X}{\sigma_X}\right)$ and $\alpha_{r+1:n} = E\left(\frac{X_{r+1:n} - \mu_X}{\sigma_X}\right)$.

The parameter we are interested in calculating is the mean of prediction error, which is $E\left(\hat{Y}_{[r:n]} - Y_{[r:n]}\right)$. By substituting (5.11) and (5.3) in the prediction error formula we have

$$E\left(\hat{Y}_{[r:n]} - Y_{[r:n]}\right) = E\left(\hat{Y}_{[r:n]}\right) - E\left(Y_{[r:n]}\right) \tag{5.11a}$$

$$= \mu_Y + \rho\sigma_Y\left(\frac{\alpha_{r-1:n} + \alpha_{r+1:n}}{2}\right) - \mu_Y - \rho\sigma_Y\alpha_{r:n}$$

$$= \rho\sigma_Y\left(\frac{\alpha_{r-1:n} + \alpha_{r+1:n}}{2} - \alpha_{r:n}\right)$$

As can be seen the prediction error is not necessarily zero. However, it easily can be shown that the prediction error is asymptotically equal to zero as follows:

For ease of writing take $\mu_Y = \mu_X = 0$, $\sigma_X = \sigma_Y = 1$, and by substituting (2.20) in (5.10) and (5.11a), and assuming a bivariate normal distribution for X and Y it immediately can be seen that

$$E\left(\hat{Y}_{[r:n]} - Y_{[r:n]}\right) \to 0 .$$

We can easily generalize (5.11) to the k nearest neighbours. Suppose k nearest neighbours are available for $Y_{[r:n]}$, by using (5.3) and (5.8), we have:

$$E(\hat{Y}_{[r:n]}) = \mu_Y + \rho\sigma_Y\left(\frac{\sum_{i=1}^{k/2}(\alpha_{r-i:n} + \alpha_{r+i:n})}{2}\right) , \tag{5.12}$$

where $\alpha_{r:n} = E(\frac{X_{r:n} - \mu_X}{\sigma_X})$, μ_Y is the mean, σ_Y is the variance of Y. Again as can be seen from (5.12), the prediction error of $\hat{Y}_{[r:n]}$ according to (5.11a) is necessarily not zero. However, by substituting (2.20) in (5.12), it immediately can be seen that the mean prediction error for $\hat{Y}_{[r:n]}$ is asymptotically zero.

5.2.2 Variance of $\hat{Y}_{[r:n]}$ for two and k nearest neighbours

In this section, we find the variance of $\hat{Y}_{[r:n]}$ under SSTI according to two and k nearest neighbours.

a) Variance of $\hat{Y}_{[r:n]}$ for two nearest neighbours

In this section we calculate the variance of $\hat{Y}_{[r:n]}$ when $k=2$. We use the following formula to calculate the variance of $\hat{Y}_{[r:n]}$.

$$Var(\sum W_i Y_i) = \sum W_i^2 Var(Y_i) + 2\sum_{i<j} W_i W_j Cov(Y_i, Y_j) \qquad (5.13)$$

By substituting (5.9) in (5.13), we have:

$$Var(\hat{Y}_{[r:n]}) = \frac{1}{4}Var(Y_{[r-1:n]}) + \frac{1}{4}Var(Y_{[r+1:n]}) + \frac{1}{2}Cov(Y_{[r-1:n]}, Y_{[r+1:n]}),$$

and by substituting (5.4) and (5.6) in the above formula we have:

$$Var(\hat{Y}_{[r:n]}) = \frac{1}{4}\left(\sigma_Y^2(\rho^2\beta_{(r-1)(r-1):n} + 1 - \rho^2\right) + \frac{1}{4}\left(\sigma_Y^2(\rho^2\beta_{(r+1)(r+1):n} + 1 - \rho^2\right) + \frac{1}{2}\sigma_Y^2\rho^2\beta_{(r-1)(r+1):n}.$$

Then, by simplifying, we have:

$$Var\left(\hat{Y}_{[r:n]}\right) = \sigma_Y^2\left\{\frac{1}{4}\left(\rho^2\beta_{(r-1)(r-1):n} + 1 - \rho^2\right) + \frac{1}{4}\left(\rho^2\beta_{(r+1)(r+1):n} + 1 - \rho^2\right) + \frac{1}{2}\rho^2\beta_{(r-1)(r+1):n}\right\}.$$

Hence, the variance of for $\hat{Y}_{[r:n]}$ can be written as follows:

$$Var(\hat{Y}_{[r:n]}) = \sigma_Y^2\left\{\rho^2\left(\frac{1}{4}\beta_{(r-1)(r-1):n} + \frac{1}{4}\beta_{(r+1)(r+1):n} + \frac{1}{2}\beta_{(r-1)(r+1):n}\right) + \frac{1-\rho^2}{2}\right\}. \qquad (5.14)$$

Finally, the variance for $\hat{Y}_{[r:n]}$ can be estimated as follows:

$$Var(\hat{Y}_{[r:n]}) = s_y^2 \left\{ r^2 \left(\frac{1}{4}\beta_{(r-1)(r-1):n} + \frac{1}{4}\beta_{(r+1)(r+1):n} + \frac{1}{2}\beta_{(r-1)(r+1):n} \right) + \frac{1-r^2}{2} \right\},$$

where

$$\hat{\beta}_{rs:n} = Cov\left(\frac{X_{r:n} - \bar{X}}{s_X}, \frac{X_{s:n} - \bar{X}}{s_X} \right),$$

$$\hat{\beta}_{(r-1)(r-1):n} = Cov\left(\frac{X_{r-1:n} - \bar{X}}{s_X}, \frac{X_{r-1:n} - \bar{X}}{s_X} \right) = s^2_{X_{r-1:n}} \text{ and}$$

$$\hat{\beta}_{(r+1)(r+1):n} = Cov\left(\frac{X_{r+1:n} - \bar{X}}{s_X}, \frac{X_{r+1:n} - \bar{X}}{s_X} \right) = s^2_{X_{r+1:n}}$$

are the covariance of $X_{r:n}$ and $X_{s:n}$, the variance of $X_{r-1:n}$ and the variance of $X_{r+1:n}$ respectively. In addition $\bar{X}$ and s_X are the sample mean and standard deviation of X and r is the correlation between X and Y. Therefore, to find the variance of $\hat{Y}_{[r:n]}$, we need to find the expectation and the variance and covariance of the r^{th} and s^{th} of order statistics from a standard normal distribution.

The prediction variance for $\hat{Y}_{[r:n]}$ when $k=2$ can be found as follows

$$Var\left(\hat{Y}_{[r:n]} - Y_{[r:n]}\right) = Var\left(\hat{Y}_{[r:n]}\right) + Var\left(Y_{[r:n]}\right) - 2Cov\left(\hat{Y}_{[r:n]}, Y_{[r:n]}\right).$$

By substituting (5.9) in above formula we have

$$Var\left(\hat{Y}_{[r:n]} - Y_{[r:n]}\right) = Var\left(\hat{Y}_{[r:n]}\right) + Var\left(Y_{[r:n]}\right) - Cov\left(Y_{[r-1:n]}, Y_{[r:n]}\right) - Cov\left(Y_{[r+1:n]}, Y_{[r:n]}\right),$$

where $Var\left(\hat{Y}_{[r:n]}\right)$ has been calculated in (5.14). Therefore by substituting (5.4), (5.5) and (5.14) we have

$$Var\left(\hat{Y}_{[r:n]} - Y_{[r:n]}\right) = \sigma_Y^2 \left\{ \rho^2 \left(\frac{1}{4}\beta_{(r-1)(r-1):n} + \frac{1}{4}\beta_{(r+1)(r+1):n} + \frac{1}{2}\beta_{(r-1)(r+1):n} + \beta_{rr:n} \right.\right.$$
$$\left.\left. + \beta_{(r-1)(r):n} + \beta_{(r+1)(r):n} \right) + \frac{3}{2}\left(1 - \rho^2\right) \right\}$$

b) Variance estimation for *k* nearest neighbours

An estimator that use k nearest neighbours for the missing value $Y_{[r:n]}$ is as follows:

$$\hat{Y}_{[r:n]} = \frac{\sum_{i=1}^{k/2}(Y_{[r-i:n]} + Y_{[r+i:n]})}{k} \quad . \tag{5.15}$$

By substituting (5.15) in (5.13) we have :

$$Var(\hat{Y}_{[r:n]}) = \frac{1}{k^2} Var\left[\sum_{i=1}^{k/2}\left(Y_{[r-i:n]} + Y_{[r+i:n]}\right)\right]$$

$$= \frac{1}{k^2}\left[\sum_{i=1}^{k/2} Var\left(Y_{[r-i:n]} + Y_{[r+i:n]}\right) + \right.$$

$$\left. 2\sum_{i<j}\left(Cov\left(Y_{[r-i:n]}, Y_{[r-j:n]}\right) + Cov\left(Y_{[r+i:n]}, Y_{[r+j:n]}\right) + Cov\left(Y_{[r-i:n]}, Y_{[r+j:n]}\right) + Cov\left(Y_{[r+i:n]}, Y_{[r-j:n]}\right)\right)\right]$$

(5.16)

Formula (5.16) is divided into two terms, the variance, and the covariance term. We expand the first term as follows:

$$I = \frac{1}{k^2}\left(\sum_{i=1}^{k/2}\left[\sigma_Y^2(\rho^2\beta_{(r-i)(r-i):n} + 1 - \rho^2) + \sigma_Y^2(\rho^2\beta_{(r+i)(r+i):n} + 1 - \rho^2) + 2\rho^2\sigma_Y^2\beta_{(r-i)(r+i):n}\right]\right)$$

$$= \frac{\sigma_Y^2}{k^2}\sum_{i=1}^{k/2}\left[\rho^2\beta_{(r-i)(r-i):n} + 1 - \rho^2 + \rho^2\beta_{(r+i)(r+i):n} + 1 - \rho^2 + 2\rho^2\beta_{(r-i)(r+i):n}\right]$$

$$= \frac{\rho^2\sigma_Y^2}{k^2}\sum_{i=1}^{k/2}\left[\beta_{(r-i)(r-i):n} + \beta_{(r+i)(r+i):n} + 2\beta_{(r-i)(r+i):n}\right] + \frac{\sigma_Y^2(1-\rho^2)}{k} \tag{5.17}$$

Now we turn to the second term of (5.16). Here, we have:

$$k^2 II = 2\sum_{i<j}\left(Cov\left(Y_{[r-i:n]}, Y_{[r-j:n]}\right) + Cov\left(Y_{[r+i:n]}, Y_{[r+j:n]}\right) + Cov\left(Y_{[r-i:n]}, Y_{[r+j:n]}\right) + Cov\left(Y_{[r+i:n]}, Y_{[r-j:n]}\right)\right)$$

$$= 2\sum_{i<j}\left(\rho^2\sigma_Y^2\beta_{(r-i)(r-j):n} + \rho^2\sigma_Y^2\beta_{(r+i)(r+j):n} + \rho^2\sigma_Y^2\beta_{(r-i)(r+j):n} + \rho^2\sigma_Y^2\beta_{(r+i)(r-j):n}\right)$$

$$= 2\rho^2\sigma_Y^2\sum_{i<j}\left(\beta_{(r-i)(r-j):n} + \beta_{(r+i)(r+j):n} + \beta_{(r-i)(r+j):n} + \beta_{(r+i)(r-j):n}\right).$$

By multiplying $\frac{1}{k^2}$ from (5.16), we have:

$$II = \frac{2\rho^2\sigma_Y^2}{k^2}\sum_{i<j}\left(\beta_{(r-i)(r-j):n} + \beta_{(r+i)(r+j):n} + \beta_{(r-i)(r+j):n} + \beta_{(r+i)(r-j):n}\right) . \tag{5.18}$$

Finally, from (5.17) and (5.18) we have

$$\begin{aligned} Var(\hat{Y}_{[r:n]}) = \frac{\rho^2\sigma_Y^2}{k^2}\Bigg[\sum_{i=1}^{k/2}\left(\beta_{(r-i)(r-i):n} + \beta_{(r+i)(r+i):n} + 2\beta_{(r-i)(r+i):n}\right) + \\ 2\sum_{i<j}\left(\beta_{(r-i)(r-j):n} + \beta_{(r+i)(r+j):n} + \beta_{(r-i)(r+j):n} + \beta_{(r+i)(r-j):n}\right)\Bigg] + \frac{\sigma_Y^2(1-\rho^2)}{k}, \end{aligned} \tag{5.19}$$

where ρ is the correlation coefficient between X and Y, σ_Y^2 is the variance of Y. In addition,

$$\beta_{rs:n} = Cov\left(\frac{X_{r:n}-\mu_X}{\sigma_X}, \frac{X_{s:n}-\mu_X}{\sigma_X}\right),$$

$$\beta_{(r-i)(r-i):n} = Cov\left(\frac{X_{r-i:n}-\mu_X}{\sigma_X}, \frac{X_{r-i:n}-\mu_X}{\sigma_X}\right) = \sigma^2_{X_{r-i:n}} \text{ and}$$

$$\beta_{(r+1)(r+1):n} = Cov\left(\frac{X_{r+i:n}-\mu_X}{\sigma_X}, \frac{X_{r+i:n}-\mu_X}{\sigma_X}\right) = \sigma^2_{X_{r+i:n}}$$

are the covariance of $X_{r:n}$ and $X_{s:n}$, the variance of $X_{r-i:n}$ and the variance of $X_{r+i:n}$, respectively. As we know $X_{r:n}$ is the r^{th} order statistic of the variable X. Therefore, to find the variance of $\hat{Y}_{[r:n]}$, we need to find the expectation and the variance and covariance of the r^{th} and s^{th} of order statistics from the standard normal distribution.

The prediction variance for $\hat{Y}_{[r:n]}$ under k nearest neighbours can be found as follows

$$Var\left(\hat{Y}_{[r:n]} - Y_{[r:n]}\right) = Var\left(\hat{Y}_{[r:n]}\right) + Var\left(Y_{[r:n]}\right) - 2Cov\left(\hat{Y}_{[r:n]}, Y_{[r:n]}\right).$$

By substituting (5.8) in above formula we have

$$Var\left(\hat{Y}_{[r:n]}-Y_{[r:n]}\right)=Var\left(\hat{Y}_{[r:n]}\right)+Var\left(Y_{[r:n]}\right)-\frac{2}{k}\sum_{i=1}^{k/2}\left(Cov\left(Y_{[r-i:n]},Y_{[r:n]}\right)-Cov\left(Y_{[r+i:n]},Y_{[r:n]}\right)\right),$$

where $Var\left(\hat{Y}_{[r:n]}\right)$ has calculated in (5.19). Therefore by substituting (5.4), (5.6) and (5.19) we have

$$Var\left(\hat{Y}_{[r:n]}-Y_{[r:n]}\right)=$$

$$\frac{\rho^2\sigma_Y^2}{k^2}\left(\frac{\beta_{rr:n}}{k^2}+\sum_{i=1}^{k/2}\left(\beta_{(r-i)(r-i):n}+\beta_{(r+i)(r+i):n}+2\beta_{(r-i)(r+i):n}\right)\right.$$

$$\left.-\frac{2}{k}\sum_{i=1}^{k/2}\left(\beta_{(r-i)(r):n}+\beta_{(r+i)(r):n}\right)+2\sum_{i<j}\left(\beta_{(r-i)(r-j):n}+\beta_{(r+i)(r+j):n}+\beta_{(r-i)(r+j):n}+\beta_{(r+i)(r-j):n}\right)\right)\cdot$$

$$+\frac{(k+1)\sigma_Y^2(1-\rho^2)}{k}.$$

5.3 Double order sequential taxonomy imputation (DSTI)

We generalize SSTI to double order sequential taxonomy imputation (DSTI). In this method, we use two components to order the data, and as a result, there are two sets of nearest neighbours for a missing value. In the previous section (5.2) for data imputation, we had two variables, X and Y, and the data were ordered based on X. Then the expectation and variance of $\hat{Y}_{[r:n]}$ were found with two and k nearest neighbours. In this section, there are three variables X_1, X_2 and Y, and the data is ordered according to X_1 and X_2, which are assumed to be complete variables. To conduct DSTI, two scenarios exist. In the first scenario, data is ordered based on a linear combination of X_1 and X_2, and in the second scenario data is separately ordered on these variables. This section is divided into two sub-sections, the first sub-section reviews imputed value properties of the linear combination of two orders, and, the second sub-section reviews the imputed value properties of two separate orders.

5.3.1 Imputation by a linear combination of two orders

Suppose three random variables X_1, X_2 and Y exist, where $X_1 \sim N(\mu_{X_1}, \sigma^2_{X_1})$, $X_2 \sim N(\mu_{X_2}, \sigma^2_{X_2})$ and $Y \sim N(\mu_Y, \sigma^2_Y)$. We want to order the data according to the linear combination of X_1 and X_2. That is

$$X = a_1 X_1 + a_2 X_2, \tag{5.20}$$

where a_1 and a_2 are fixed. This case is similar to the single order sequential taxonomy imputation. Based on statistical properties of normal distribution (Johnson and Kotz, 1994) the linear combination of two normal random variables (5.20) has normal distribution as follows:

$$X \sim N(a_1\mu_{X_1} + a_2\mu_{X_2}, a_1^2\sigma^2_{X_1} + a_2^2\sigma^2_{X_2} + 2a_1a_2\sigma_{X_1X_2}), \tag{5.21}$$

where $\sigma_{X_1X_2}$ is the covariance between X_1 and X_2. In addition, it is assumed that the X_1 and X_2 are complete, and missing values occur only in the variable Y. Thus, this case changes to SSTI, and all of the assumptions and formulae in section 5.2 are valid for imputation by the linear combination of two orders. Expectations, variances, and covariances of the order statistics X are obtained according to the distribution of X (5.21). This method is more efficient than SSTI if $\rho_{X,Y} \geq \max(\rho_{X_1,Y}, \rho_{X_2,Y})$ according to the regression properties.

5.3.2 Imputation with two separate orders

In this method, the data is ordered based on two complete variables and nearest neighbours under each ordering are obtained separately. Therefore, there are two sets of nearest neighbours for a missing value, where the first set comes from ordering by the first component, and the second set comes from ordering by the second component.

Let $X_1 \sim N(\mu_{X_1}, \sigma^2_{X_1})$, $X_2 \sim N(\mu_{X_2}, \sigma^2_{X_2})$ and $Y \sim N(\mu_Y, \sigma^2_Y)$. The data is divided into two sets: (X_1, Y) and (X_2, Y). Thus, two separate datasets exist, and the link between

them is the correlation between X_1 and X_2. Therefore, we have two sets of concomitants of order statistics:

$(X_{1(r1:n)}, Y_{1[r1:n]}), r1 = 1,2,...,n$,

$(X_{2(r2:n)}, Y_{2[r2:n]}), r2 = 1,2,...,n$,

where $Y_{1[r1:n]}$ is the concomitant of $X_{1(r1:n)}$ and $Y_{2[r2:n]}$ is the concomitant of $X_{2(r2:n)}$. In addition, $X_{1(r1:n)}$ and $X_{2(r2:n)}$ are complete and missing values occur in $Y_{1[r1:n]}$ and $Y_{2[r2:n]}$.

In order to develop a mathematical theory for DSTI suppose (X_{1i}, Y_i) and (X_{2i}, Y_i) $(i = 1,2,...,n)$ are i.i.d. random samples from a bivariate normal distribution, such that $(X_{1i}, Y_i) \sim N_2(\mu_{X_1}, \mu_Y, \sigma^2_{X_1}, \sigma^2_Y, \rho_1)$ and $(X_{2i}, Y_i) \sim N_2(\mu_{X_2}, \mu_Y, \sigma^2_{X_2}, \sigma^2_Y, \rho_2)$, where $E(X_1) = \mu_{X_1}, E(X_2) = \mu_{X_2}$, $E(Y) = \mu_Y$, $Var(X_1) = \sigma^2_{X_1}$, $Var(X_2) = \sigma^2_{X_2}$, $Var(Y) = \sigma^2_Y$, $Cor(X_1, Y) = \rho_1$ and $Cor(X_2, Y) = \rho_2$. From statistical properties of bivariate normal distribution we then have

$$Y_i = \mu_Y + \rho_1 \frac{\sigma_Y}{\sigma_{X_1}}(X_{1i} - \mu_{X_1}) + Z_{1i}, \tag{5.22}$$

$$Y_i = \mu_Y + \rho_2 \frac{\sigma_Y}{\sigma_{X_2}}(X_{2i} - \mu_{X_2}) + Z_{2i}, \tag{5.23}$$

where X_{1i} and Z_{1i}, and X_{2i} and Z_{2i} are mutually independent. Then from (5.22) and (5.23), it follows that $E(Z_1) = 0$, $Var(Z_1) = \sigma^2_Y(1 - \rho_1^2)$, $E(Z_2) = 0$ and $Var(Z_2) = \sigma^2_Y(1 - \rho_2^2)$.

Assume X_1 and X_2 are ordered, and Y_1 and Y_2 are concomitants of X_1 and X_2 respectively. Hence, from (5.22) and (5.23), we have:

$$Y_{1[r:n]} = \mu_Y + \rho_1 \frac{\sigma_Y}{\sigma_{X_1}}(X_{1r:n} - \mu_{X_1}) + Z_{1[r:n]}, \tag{5.24}$$

$$Y_{2[r:n]} = \mu_Y + \rho_2 \frac{\sigma_Y}{\sigma_{X_2}}(X_{2r:n} - \mu_{X_2}) + Z_{2[r:n]}, \tag{5.25}$$

where $Z_{1[r]}$ and $Z_{2[r]}$ are concomitants of Z_{1i} and Z_{2i} with respect to $X_{1r:n}$ and $X_{2r:n}$. From the independence of X_{1i} and Z_{1i} we can conclude that $X_{1r:n}$ and $Z_{1[r:n]}$ are independent. Similarly from the independence of X_{2i} and Z_{2i} it is concluded that $X_{2r:n}$ and $Z_{2[r:n]}$ are independent.

In order to establish an imputation method for the missing value under DSTI, we suppose that there are two sets of imputed values for a missing value, given by $\hat{Y}_{1[r:n]}$ and $\hat{Y}_{2[r:n]}$, which we then combine to form a unique imputed value as follows:

$$\hat{Y}_{[r:n]} = a\hat{Y}_{1[r:n]} + (1-a)\hat{Y}_{2[r:n]}\ , \tag{5.26}$$

where a is a fixed. From (5.26), the statistical properties of the missing value $\hat{Y}_{[r:n]}$ can be found. The expectation of $\hat{Y}_{[r:n]}$ is:

$$E(\hat{Y}_{[r:n]} \mid a) = aE(\hat{Y}_{1[r:n]}) + (1-a)E(\hat{Y}_{2[r:n]})\,, \tag{5.27}$$

where $E(\hat{Y}_{1[r:n]})$ and $E(\hat{Y}_{2[r:n]})$ are the expectations of imputed values based on a single order as discussed in section 5.2. Consequently $E(\hat{Y}_{[r:n]} \mid a)$ is asymptotically equal to the population mean of Y. As seen from section 5.2.1 $E(\hat{Y}_{1[r:n]}) \cong \mu_Y$ and $E(\hat{Y}_{2[r:n]}) \cong \mu_Y$, hence $E(\hat{Y}_{[r:n]} \mid a) \cong a\mu_Y + (1-a)\mu_Y \cong \mu_Y$.

The variance of $\hat{Y}_{[r:n]}$ can be found as follows:

$$\begin{aligned} Var(\hat{Y}_{[r:n]} \mid a) &= Var(a\hat{Y}_{1[r:n]} + (1-a)\hat{Y}_{2[r:n]})\,, \\ &= a^2 Var(\hat{Y}_{1[r:n]}) + (1-a)^2 Var(\hat{Y}_{2[r:n]}) + 2a(1-a)Cov(\hat{Y}_{1[r:n]}, \hat{Y}_{2[r:n]}) \end{aligned} \tag{5.28}$$

where $Var(\hat{Y}_{1[r:n]})$ and $Var(\hat{Y}_{2[r:n]})$ are the variances of the imputed values under the separate ordering. Then from (5.24), (5.25), and (5.28), we have:

$$Cov(\hat{Y}_{1[r:n]}, \hat{Y}_{2[r:n]}) = Cov(\mu_Y + \rho_1 \frac{\sigma_Y}{\sigma_{X_1}}(X_{1r:n} - \mu_{X_1}) + Z_{1[r]}, \mu_Y + \rho_2 \frac{\sigma_Y}{\sigma_{X_2}}(X_{2r:n} - \mu_{X_2}) + Z_{2[r]})$$

$$= \frac{\rho_1\rho_2\sigma_Y^2}{\sigma_{X_1}\sigma_{X_2}} Cov(X_{1r:n}, X_{2r:n}) + \frac{\rho_1\sigma_Y}{\sigma_{X_1}} Cov(X_{1r:n}, Z_{2[r]}) +$$

$$\frac{\rho_2 \sigma_Y}{\sigma_{X_2}} Cov(X_{2r:n}, Z_{1[r]}) + Cov(Z_{1[r]}, Z_{2[r]}). \qquad (5.29)$$

Now the important question is how to define "a" for equation (5.26). Equation (5.26) can be seen as a weighted imputation method for two different imputation procedures. The constant a in (5.26) is chosen to give the imputed values minimum variance. We calculate a by differentiating the prediction variance $Var(\hat{Y}_{[r:n]} - Y_{[r:n]} \mid a)$, with respect to a, and setting the result to zero. The prediction variance of $\hat{Y}_{[r:n]}$ under DSTI is

$$Var(\hat{Y}_{[r:n]} - Y_{[r:n]} \mid a) = Var\left(\hat{Y}_{[r:n]}\right) + Var\left(Y_{[r:n]}\right) - 2Cov\left(\hat{Y}_{[r:n]}, Y_{[r:n]}\right).$$

By substituting (5.26) in above formula we have

$$Var(\hat{Y}_{[r:n]} - Y_{[r:n]} \mid a) = a^2 Var(\hat{Y}_{1[r:n]}) + (1-a)^2 Var(\hat{Y}_{2[r:n]}) + 2a(1-a)Cov(\hat{Y}_{1[r:n]}, \hat{Y}_{2[r:n]})$$
$$+Var\left(Y_{[r:n]}\right) - 2aCov\left(\hat{Y}_{1[r:n]}, Y_{[r:n]}\right) - 2(1-a)Cov\left(\hat{Y}_{2[r:n]}, Y_{[r:n]}\right).$$

Then by differentiating, this formula with respect a and setting equals to zero we have

$$\partial\left(Var(\hat{Y}_{[r:n]})\right)/\partial a = 2aVar(\hat{Y}_{1[r:n]}) - 2(1-a)Var(\hat{Y}_{2[r:n]}) + (2-4a)Cov(\hat{Y}_{1[r:n]}, \hat{Y}_{2[r:n]})$$
$$-2Cov\left(\hat{Y}_{1[r:n]}, Y_{[r:n]}\right) + 2Cov\left(\hat{Y}_{2[r:n]}, Y_{[r:n]}\right) = 0.$$

Solving for a, we have

$$a = \frac{Var\left(\hat{Y}_{2[r:n]}\right) + Cov\left(\hat{Y}_{1[r:n]}, Y_{[r:n]}\right) - Cov\left(\hat{Y}_{2[r:n]}, Y_{[r:n]}\right) - Cov\left(\hat{Y}_{1[r:n]}, \hat{Y}_{2[r:n]}\right)}{Var\left(\hat{Y}_{1[r:n]}\right) + Var\left(\hat{Y}_{2[r:n]}\right) - 2Cov\left(\hat{Y}_{1[r:n]}, \hat{Y}_{2[r:n]}\right)}.$$

The above formula depends on the covariances of true and imputed values, which are unavailable. Therefore, in practice we use the following procedure to find a.

1) For each ordering, a mean vector based the values of nearest neighbours under that ordering is calculated. Then, two Euclidean distances of a missing value from the two mean vectors are calculated according to the available information of the missing value and these two vectors. These are d_{1i} and d_{2i}, $i=1,2,..,m$, where m is the number of missing values. In other words for each missing value we have two distance values.
2) a can be defined for each missing value as follows:

$$\begin{cases} d_{1i} < d_{2i} & a = 1 \\ d_{1i} > d_{2i} & a = 0 \\ d_{1i} = d_{2i} & a = 0 \text{ or } 1 \end{cases}$$

where in the case of two equal distances, we can choose $a=1$ or $a=0$ randomly.

5.4 Evaluation and simulation study

Different imputation methods have been designed for different types of data. In addition, every imputation method has its own assumptions, such as the missing data pattern, the missing data mechanism, and possibly the distribution of the data. Therefore, when comparing and evaluating imputation methods, it is necessary to determine and consider all the assumptions and the type of data required; otherwise, the comparisons and evaluations are not valid. There are two main purposes in the evaluation of imputation methods. The first aim is to identify the capability of any imputation method to predict true data in the sample and the second is to assess the statistical imputation effects on the data. In other words, an imputation method affects the quality of the estimates and data. In evaluating an imputation method, the most relevant concerns are the bias of point estimators, the variance of estimates, and the ability to predict missing values correctly. In the statistical literature on evaluation, there are two points of view. The first focuses on evaluation when true values are known, and is essentially a simulation-based approach (Chambers 2001), and the second focuses on evaluation when the true values are unknown. The first type of evaluation method is usually used for the establishment and comparison of different types of imputation methods, and for assessing the suitability of an imputation method for specific application. Which evaluation criteria to use depends on the purpose of on study. For example, if our aim is to estimate the sample mean, we then shall focus in the estimation accuracy and so on. This section is divided into three sections: the first is concerned with evaluation, when true values are available; the second is concerned with the establishment of assessment measures for imputation methods when true values are not available; and the third describes a simulation study that was used to

compare the imputation methods described in this chapter with a number of other commonly used methods.

5.5 Evaluation methods when true values are available

According to Chambers (2001), the evaluation criteria for comparing different imputation procedures include predictive, ranking, distribution and estimation accuracy as well as imputation plausibility. An efficient imputation method should have some of the following properties:

a) Predictive accuracy, which means that that the imputed values should be close to the true values
b) Ranking accuracy, which means the imputed values should preserve the ranks of true values
c) Distribution accuracy, which means that the imputation method should preserve the distribution of true values.
d) Estimation accuracy, which means that the imputation procedure should produce unbiased and efficient estimators for the parameters of the population distribution of true values.
e) Imputation plausibility, which means that the imputed values should be plausible with respect to other values and other variables; for example, the imputed values should be verified by all edit tests.

Note that some of the above properties are used for continuous data and some are used for ordinal or categorical variables. This section reviews only the predictive and distribution accuracy, and the estimation accuracy is reviewed in section 5.3.

5.5.1 Distribution accuracy

The consistency of marginal distributions between true and imputed values is a measure of the validity of the imputation method. Stuart (1955) offered a test statistic, which has a Chi-square distribution. Chambers (2001) adapts this statistic to imputation leading to a Wald statistic expression. This method is based on the following assumptions:

Suppose Y is a multinomial variable with $c+1$ categories. We assume the last category of Y as a reference category, which contains missing values. We map the first c categories of Y to a c-vector y. In addition, we use index i for a specific case. Therefore, $y_i = (y_{ij})$ for case i, where $y_{ik} = 1$ for $Y = k < c+1$ and $y_{ij} = 0$ for $j \neq k$. Let $\hat{y}_i$ be an imputer for y_i, p_i is the probability of observing category i, and $\hat{p}_i$ is the estimator of p_i. Then the first assumption is: $\{\hat{y}_i\}$ represents independent draws from $multinomial(\hat{p}_i)$ (imputed values for different individuals are independent of one another and drawn from some distributions). The second assumption is: imputed and actual values for an individual are independent of one another, as well as across different individuals. This assumption reflects the fact that

- Variation in the p_i's accounts for all dependence between different individuals (no residual dependence),
- The $\hat{p}_i$'s are based on responding individuals,
- The respondent sample size is large enough to ignore correlation between $\hat{p}_i$ and $\hat{p}_j$ when $i \neq j$.

We then have:

$$W = \left[\frac{1}{n-m}\sum_{i=1}^{n-m}(\hat{\mathbf{y}}_i - \mathbf{y}_i)^T\right]\left[\frac{1}{(n-m)^2}\sum_{i=1}^{n-m}(\hat{\mathbf{y}}_i - \mathbf{y}_i)(\hat{\mathbf{y}}_i - \mathbf{y}_i)^T\right]^{-1}\left[\frac{1}{n-m}\sum_{i=1}^{n-m}(\hat{\mathbf{y}}_i - \mathbf{y}_i)\right]$$

$$= [R-S]^T\left[diag(R+S) - T - T^T\right]^{-1}[R-S], \qquad (5.32)$$

where R and S are the marginal counts of $c \times c$ upper left matrix of contingency table of Y_i^I and Y_{mis} and $T = (n_{ij})$, $i, j = 1, 2, .., c$. Wald has asymptotically χ_c^2 distribution. The following is the contingency table.

	Y_i^I			
	1 ...	c	$c+1$	R
Y_{mis}				
1 $\vdots$ c $c+1$	n_{ij}			$n_{1.}$ $\vdots$ $n_{p.}$ $n_{p+1.}$
S	$n_{.1}$ ...	$n_{.p}$	$n_{.p+1}$	

For continuous variables, the Kolomogrov-Smirnov test (KS) can be used to test the equality of marginal distribution between true and imputed values. Let Y_i^I and Y_i^T be imputed and true values, respectively. The KS distance can be written as follows:

$$KS\left(F_{Y_n^T}, F_{Y_n^I}\right) = \max_t \left| F_{Y_n^T}(t) - F_{Y_n^I}(t) \right|, \tag{5.33}$$

where

$$F_{Y_n^T}(t) = \frac{\sum_{i=1}^{n} w_i I(Y_i^T \le t)}{\sum_{i=1}^{n} w_i},$$

$$F_{Y_n^I}(t) = \frac{\sum_{i=1}^{n} w_i I(Y_i^I \le t)}{\sum_{i=1}^{n} w_i} \quad \text{and}$$

w_i are the sampling weights. Marginal distributions of true values and imputed values are equal ($F_{Y_n^I}(t) = F_{Y_n^T}(t), \forall(t)$), when $KS = 0$

5.5.2 Predictive accuracy or preservation of individual data

The extent of preservation of individual values can be measured if true or validated data are available. Suppose Y is the variable with missing values, and let Y_i^T and Y_i^I

be true and imputed values of Y for $i = 1, 2, ..., n$, respectively. The definition of the predictive accuracy measure depends on the scale of measurement of Y. For a nominal variable Y, a measure of preservation of individual values can be written as follows:

$$d_{Y^T,Y^I} = \frac{1}{n}\sum_{i=1}^{n} I(Y_i^T, Y_i^I), \tag{5.34}$$

where

$$I(Y_i^T, Y_i^I) = \begin{cases} 1 & Y_i^T = Y_i^I \\ 0 & otherwise \end{cases}.$$

When formula (5.34) is equal to one, the imputation method preserves individual data perfectly. Chambers (2001) obtains the variance of $D = 1 - d_{Y^I,Y^T}$ as follows:

$$\hat{S}_D^2 = n^{-1} - n^{-2} 1^t \{diag(R+S) - T - diag(T)\} 1 = n^{-1}(1-D), \tag{5.35}$$

where 1 denotes a c-vector of ones, and the definitions of R, S and T are as given in section 5.5.1. In an ideal condition, when the imputation method preserves individual values, D is zero. Knowing this fact, we can build a statistical test for D. Hence, if $D > \varepsilon + 2\sqrt{\hat{S}_D^2}$, then it can be said that the imputation method does not preserve the individual values. There is a variety of ways to calculate ε. Chambers (2001) sets ε to zero and suggests $\varepsilon^* = \max\left(0, D - 2\sqrt{\hat{S}_D^2}\right)$. Therefore, the smaller values of ε^* show a better imputation method, and when ε^* is equal to zero it can be said that the imputation method preserves the individual values.

From Chambers (2001), for an ordinal variable Y, predictive accuracy can be measured by

$$d_{Y^T,Y^I} = \frac{1}{n \times m}\sum_{i=1}^{n} W(Y_i^T, Y_i^I), \tag{5.36}$$

where:

$$W(Y_i^T, Y_i^I) = \begin{cases} 0 & \text{if } Y_i^T = Y_i^I \\ \left|Y_i^T - Y_i^I\right| & \text{if } Y_i^T \neq Y_i^I \text{ and } Y_i^T, Y_i^I \neq blank \\ m & \text{if } Y_i^T \neq Y_i^I \text{ and } Y_i^T \text{ or } Y_i^I = blank \end{cases}$$

and $m = \left(\max(Y) - \min(Y)\right) + 1$ if the category blank is in the domain of Y, while $m = \left(\max(Y) - \min(Y)\right)$ if the category blank is not in the domain of Y.

For continuous variables there are three measures to obtain predictive accuracy: the first, transforming continuous data to categorical data, then using the Wald statistic and D to test preservation of marginal distribution and individual values; the second calculating the distance between true values and imputed values, the third, fitting regression line between the true values and the imputed values or calculating correlations between true and imputed values ($RHO = Cor(Y_i^T, Y_i^I)$). Hence, for the second case we have

$$d_{Y^T, Y^I} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left|Y_i^T - Y_i^I\right|^{\alpha} \right\}^{1/\alpha} \tag{5.37}$$

where $\alpha > 0$ while, $\alpha = 2$ corresponds to the Euclidian distance between the true and imputed values. In the third case, Y_i^I is considered a dependent variable and Y_i^T is an independent variable. Then the regression properties of the fitted model, such as R^2 and root mean square error (RMSE), are calculated. Large values of R^2 or RHO and small values of RMSE represent a good imputation method.

5.6 Evaluation methods when true values are not available

In reality, true values might not be available. Hence, performing the evaluation method for imputation is more difficult than in the case of knowing the true values. When conducting the evaluation method, knowing the purpose of the imputation method is very important. For example if the aim is to estimate the population parameter θ, then some aspects of classical statistical evaluation methods are involved, such as unbiasedness of the estimators, having the minimum variance and

standard errors for the estimator of θ, preservation of the distribution of data, and preservation of the relationship between variables.

The first measure of the performance of an imputation method is bias. In an ideal situation, the estimator of the unknown population parameter should be unbiased, but an estimator with low bias is also desirable. Let θ be the population parameter and $\hat{\theta}$ its estimator based on the imputed dataset. In addition, let $SE(\hat{\theta})$ be the standard error for $\hat{\theta}$. Unbiasedness of the unknown population parameter estimator means

$E(\hat{\theta}) = \theta$.

Interpreting the amount of bias is difficult; there are two ways of looking at it in order to get alternative information for the evaluation of imputation performance.
The first measure is relative bias, which can be defined as follows:

$$\frac{E(\hat{\theta}) - \theta}{\theta},$$

(expressed as a percentage of θ). The second measure is standardized bias, which is a percentage of standard error: $\frac{E(\hat{\theta}) - \theta}{SE(\hat{\theta})} * 100$. According to Schafer (1997), if the amount of the standardized bias exceeds 30%, it starts to adversely affect the coverage of confidence intervals.

The combination of unbiasedness and low variance creates a measure of accuracy, which is the mean square error for $\hat{\theta}$. The formula for mean square error is

$$MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right] = Var(\hat{\theta}) + \left(E(\hat{\theta}) - \theta\right)^2 \text{ or}$$

Mean Square Error = Variance + $(\text{Bias})^2$,

where high values of mean square error are not desirable, because they imply either big variance or big bias (or both).

5.7 The simulation study

The aim of this section is to assess the performance of single order and double order sequential taxonomy imputation methods (SSTI and DSTI), using simulated

multivariate normal data. In order to assess this performance, three simulations were carried out using a database that contains synthetic missing values. The simulated data involves cases with no missing information. Selected data items, for which full information was available, were artificially considered as missing and imputation procedures were applied to impute their values. In these simulations, different percentages of missing values (20% missing values under MCAR and approximately 20% and 30% missing values under MAR) for $k=10$ nearest neighbours and different types of missing data mechanism were used in order to compare their effect on results. Moreover, means and variances of population and imputed values, correlations between true and imputed values and Wald statistic were calculated in order to evaluate the properties of the estimators.

A brief description of all the steps taken in this simulation will be given in this section. This includes the use of the ranks, imputation and more general material as, for example, the generation of the database.

5.7.1 Generation of the multivariate normal database

A multivariate normal database was generated as shown in section 5.8, with different means and covariance matrices. The generated multivariate data contains five variables with no missing values. The missing values are generated by an ignorable missing data mechanism. In this case, the generation of missing values is based on two popular missing data mechanisms. Those are missing completely at random (MCAR) and missing at random (MAR). In addition, the rate of generated missing values was set to 20% of the data file with missing values generated to four of the five variables. In other words, one variable was assumed complete.

5.7.2 Ordering the data file

SSTI and DSTI use data ordering to impute missing values. Therefore, the data were ordered by three different techniques. These are the sequential taxonomy coefficients (see Chapter 4), the first component of principal component analysis and the Euclidian distance of each case from the centre (EDC).

5.7.3 Imputing

Given the different ordering, SSTI and DSTI imputation was carried out independently for each order. In addition, some other imputation methods were carried out to compare performance. The imputation methods that were investigated were single order sequential taxonomy imputation (SSTI), double order sequential taxonomy imputation (DSTI), regression imputation (RI), stratified hot deck imputation (SHDI), Euclidian distance from centre imputation (EDCI), and principal component analysis imputation (PCAI). Actually, SSTI, DSTI, EDCI, and PCAI are in the same family of imputation methods, which in these methods the data are ordered according to the coefficients calculated by these methods. Under these methods, the location of the missing values is specified and k nearest neighbours for missing values are selected. Regression imputation (RI) actually is a form of conditional mean imputation. Under multivariate normal assumptions, the imputed value will be the mean of the variable, multiplied by its associated coefficient. In this study, we assume the variable with missing values as a dependent variable. The dependent variable can be Y_1, Y_2, Y_3, or Y_4. A multivariate regression was carried out based on all available observations and then the regression models were used to predict missing values. To conduct SHDI the data categorized into 10 homogeneous categories based on a k-means technique, then missing value is replaced by a responding value from a donor randomly selected from a set of potential donors in each category. It is assumed that the number of nearest neighbours is $k=10$ in this simulation.

5.7.4 Evaluation of imputation methods

Different graphs, tests, biases, and variances are used to evaluate the imputation methods. These are mean and variance of the population and imputed values, correlation between true values and imputed values in order to compare preservation of individual values, and the Wald statistic test for comparison of the marginal distribution of true and imputed values. In addition, different graphs are produced to compare means, and the variances of imputed values with true values. Finally, real

orders of data (calculating data orders in the complete data) were calculated to assess the performance of the ordering based imputation methods.

5.8 Data

The database used for the analysis consists of five normally distributed variables Y_1, Y_2, Y_3, Y_4, Y_5, which are generated by simulation. The mean vector of these variables is $\mu = (50,40,60,35,20)$ and correlation matrix for this database is set out in Table 5.1.

Table 5.1: Correlation matrix for multivariate simulated data

	Y_1	Y_2	Y_3	Y_4	Y_5
Y_1	1.0000	0.5668	0.3960	0.4465	0.5090
Y_2	0.5668	1.0000	0.3863	0.3194	0.4343
Y_3	0.3960	0.3863	1.0000	0.4620	0.5273
Y_4	0.4465	0.3194	0.4620	1.0000	0.4159
Y_5	0.5091	0.4344	0.5273	0.4159	1.0000

In order to obtain the bias and variances of estimators as well as estimators for the variance in some cases, a simulation was carried out. The simulation involved several steps, which are explained below.

1. **Generation of the databases**. First 2000 databases each with 1000 cases were created according to the above mean vector and the correlation matrix. These 2000 databases are fully observed with no missing values.
2. **Generation of missing values**. In these files, the missing values were generated under two missing data mechanisms: MCAR and MAR. Missing values under the MCAR assumption are generated by uniform distribution and the occurrence of missing values does not depend on the other variables. In other words, each variable (Y_1, Y_2, Y_3, Y_4) has been indexed from 1 to 1000 and then 200 of observations in each variable were defined as missing values according a random sample of size 200 from 1000 cases. The missing values under the MAR assumption are generated under a logistic model conditioned on the complete variable Y_5. For the MAR mechanism, we used the follwing

two models, where the average response probability of observations are 0.80 and 0.70, respectively.

$$p_{0.80}(i)=\exp(0.1+0.07*Y_{5i})/(1+\exp(0.1+0.07*Y_{5i})),\ i=1,2,...,1000$$

$$p_{0.70}(i)=\exp(0.1+0.0385*Y_{5i})/(1+\exp(0.1+0.0385*Y_{5i})),\ i=1,2,...,1000.$$

Where $p_{0.80}(i)$ and $p_{0.70}(i)$ are the response probability of Y_1, Y_2, Y_3 and Y_4, when the probability of respondents are 0.80 and 0.70, respectively. We then generated a uniform random sample $p_a(i,j) \sim U(0,1)$, $i=1,2,...,1000$ and $j=1,2,3,4$ for each variable and each case separately. In the first simulation, if $p_a(i,j)$ was less than $p_{0.80}(i)$ case i was considered as a respondent for variable j. In this simulation, the observed number of missing values is 197 cases for all variables. In the second simulation, if $p_a(i,j)$ was less than $p_{0.70}(i)$ case i was considered as a respondent for variable j. In this simulation, the observed number of missing values is 304 cases for all variables.

3. **Ordination of data**. In order to carry out the imputation, the data file in each iteration is ordered according to the different ordering methods. The data are ordered according to sequential taxonomy coefficients, the first component of principal component analysis, and the Euclidian distance from the centre of the data. The centre in the Euclidian distance based ordering is defined to be the zero vector.
4. **Imputation**. After ordering the data, six different imputation methods are carried out in each iteration. These methods are set out in Table 5.2 below.

Table 5.2: Applied imputation methods

Single order sequential taxonomy imputation (from max points)	: SSTI
Euclidian distance from centre imputation	: EDCI
Regression imputation	: RI
Principal component analysis imputation	: PCAI
Stratified Hot Deck Imputation	: SHDI
Double ordered sequential taxonomy imputation	: DSTI

It should be noted that, SHDI was implemented using a classifier based on a k-means clustering technique.

5. Evaluation. There is no doubt that the evaluation of imputation methods depends on the aim of the study and available information for testing the results. This thesis evaluates the imputation methods according to the three following principles: the first is a comparison of marginal distributions of real and imputed values, the second is a comparison of the individual values and the third is an assessment of the properties of the estimators used in the study.

Preservation of marginal distributions is essential when the imputed data are going to be used for estimating aggregates or totals. In this case, preserving marginal distributions guarantees an accurate estimation of these aggregates, since individual values are not needed separately, such as in descriptive studies. Nevertheless, in some cases where micro data are required, it is important to maintain a relationship between variables for the subjects.

Therefore, in order to evaluate the three aspects mentioned at the beginning of this section, three different criteria were used. These are the correlation of real and imputed values for testing the preservation of individual values, the Wald statistic for testing the preservation of marginal distribution and finally comparison of the mean and the variance of real and imputed values.

5.9. Results

Before presenting the results, it is necessary to define the characteristics of the simulation. Firstly, the distribution of the generated data is multivariate normal; secondly, other characteristics of this simulation are as follows:

Size of generated databases	:	1000
Number of missing values in each variable	:	200
Number of variables with missing values	:	4
Number nearest neighbours (k)	:	10
Number of simulations	:	2000

Results are presented under two assumptions about the missing data mechanisms, MCAR and MAR. The expected values of number of missing values under MAR is 200.

5.9.1 Simulations under MCAR assumption

In this sub-section the simulation results are under MCAR.

5.9.1.1 Test of agreement or Wald statistic

Table 5.3 shows the percentages of insignificant Wald statistics as described in section 5.5.1, for variable Y_4. Only the Wald statistic results presented for variable Y_4, because of the similarity between the results of this variable with other variables.

Table 5.3: The percentages of insignificant Wald statistic for different imputation methods by real order of data and estimated order of data

estimated order		real order of data
SSTI	79.40	86.35
EDCI	81.90	80.45
RI	20.05	20.05
PCAI	67.70	78.70
SHDI	97.30	97.70
DSTI	85.25	90.55

The first column of this Table shows the percentages of insignificant Wald statistics when estimated data orders are used for imputation. For example, in 79.4% of the 2000 iterations, SSTI preserves the marginal distribution, but in 20.05% of iterations the marginal distributions are preserved by RI. However, SHDI preserves the marginal distributions in 97.3% of iterations and the percentage for DSTI is 85.25%.

The second column of Table 5.3, shows the percentages of insignificant Wald statistics when the real order of data is used. In other words, in this case, the data are ordered when there are no missing values in the data file. It can be seen from Table 5.3 that for the order based imputation methods, the percentage of insignificant Wald statistics increase, but for RI it remains constant. For example, for DSTI, if real orders of data rather than estimated orders are used, the percentage of the preservation of marginal distribution increases from 85.25% to 90.55%.

5.9.1.2 Comparison of population means

Table 5.4 and Figures 5.1 to 5.4 show the means of the database without considering missing values and the database with imputed values. These means are population means in 2000 iterations.

Table 5.4: Population means for simulated data and different imputation methods

	Y_1	Y_2	Y_3	Y_4
True	50.0000	40.0000	60.0000	35.0000
SSTI	50.0013	40.0018	60.0018	34.9981
EDCI	50.0048	40.0066	60.0090	35.0066
RI	49.9985	40.0013	60.0008	35.0005
PCAI	49.9925	40.0037	60.0039	35.0107
SHDI	49.9985	40.0013	60.0009	35.0003
DSTI	49.9964	40.0023	60.0026	35.0038

It can be seen from Table 5.4 and Figures 5.1 to 5.4 that all of the imputation methods create biased estimations for population means. However, generally, SSTI, RI, and SHDI have small biases with compared to the other imputation methods.

Figure5.1: Comparison of population means of Y_1 by different imputation methods

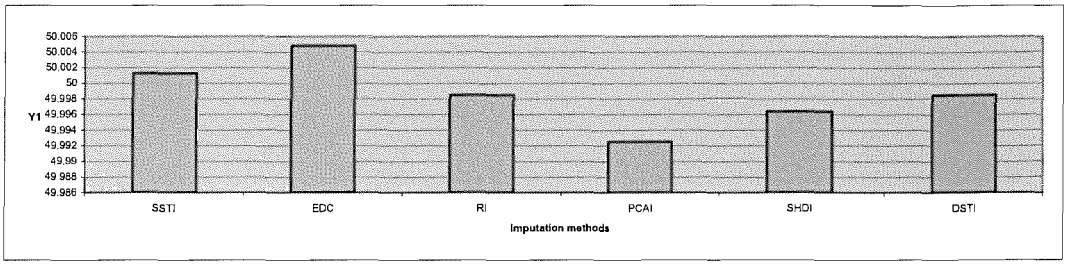


Figure5.2: Comparison of population means of Y_2 by different imputation methods

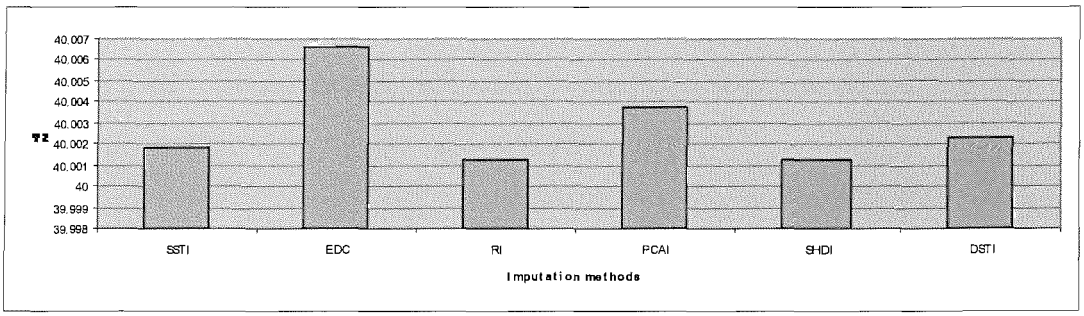


Figure5.3: Comparison of population means of Y_3 by different imputation methods

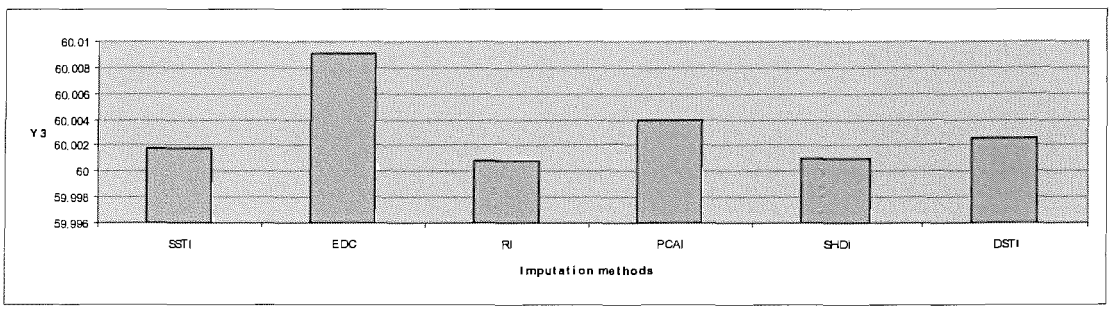
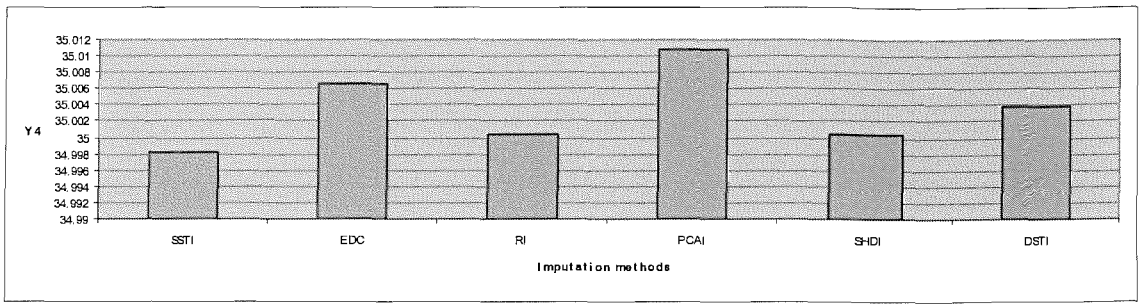


Figure5.4: Comparison of population means of Y_4 by different imputation methods



5.9.1.3 Comparison of population variances

Table 5.5 and Figures 5.5 to 5.8 show the variances of the database without considering missing values and the database with imputed values. These variances are population variances in 2000 iterations.

Table 5.5: Population variance for simulated data and different imputation methods

	Y_1	Y_2	Y_3	Y_4
True	4.0000	3.0000	7.0000	4.0000
SSTI	3.8918	2.9185	6.8095	3.8810
EDCI	3.8643	2.9062	6.7769	3.8753
RI	3.8137	2.8557	6.6548	3.7845
PCAI	3.8887	2.9174	6.8064	3.8822
SHDI	3.9948	3.0024	6.9982	3.9975
DSTI	3.8779	2.9278	6.8095	3.9105

Figure 5.5: Comparison of population variance of Y_1 by different imputation methods

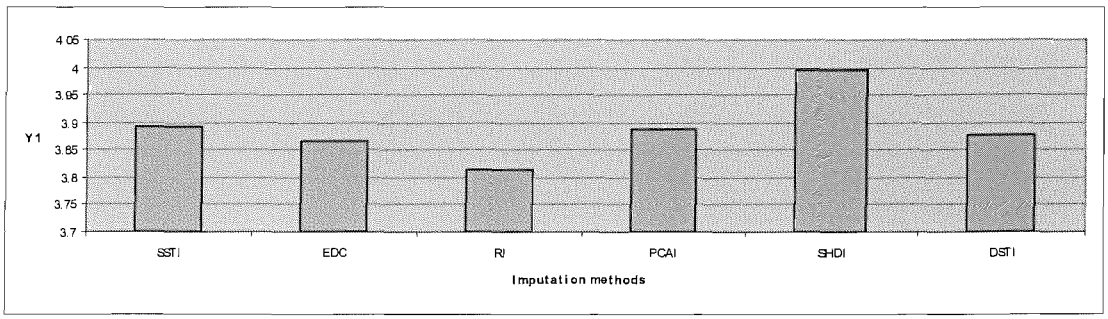


Figure 5.6: Comparison of population variance of Y_2 by different imputation methods

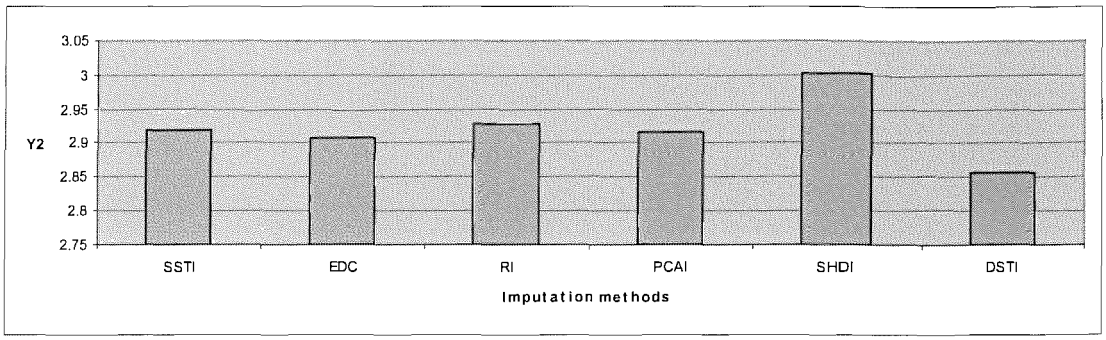


Figure 5.7: Comparison of population variance of Y_3 by different imputation methods

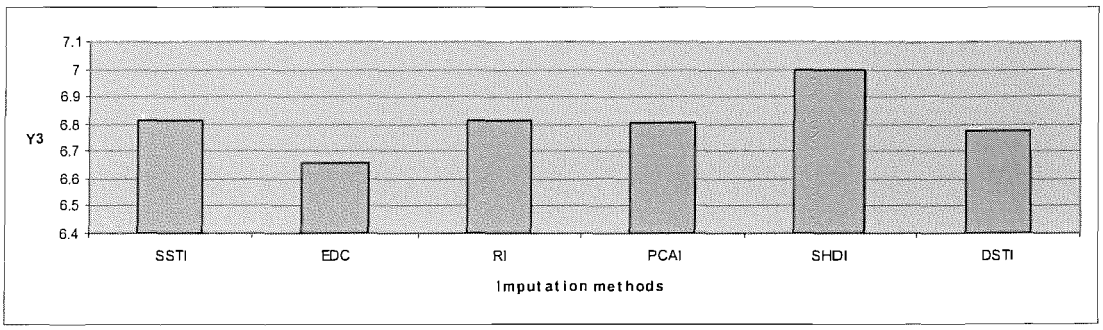
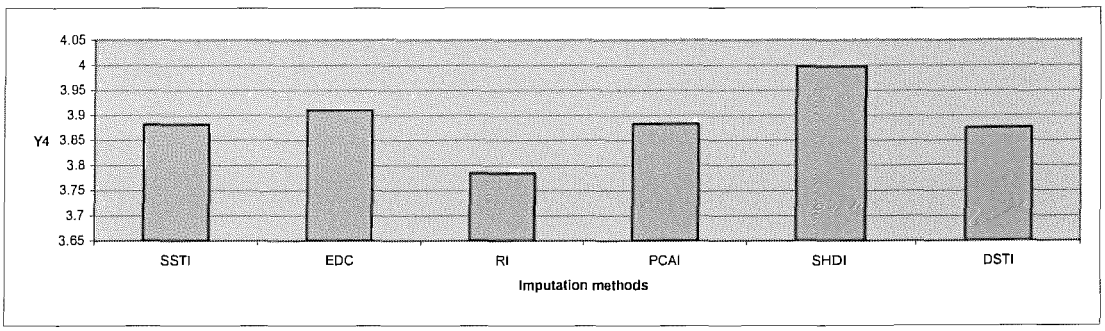


Figure 5.8: Comparison of population variance of Y_4 by different imputation methods



It can be seen from Table 5.5 and Figures 5.5 to 5.8, that apart from SHDI, all of the used imputation methods underestimate the variance of the population. Based on the theory of SHDI (see Chapter 3) this method creates asymptotically unbiased estimations for population variance under MCAR. According to the simulation results, DSTI, SSTI, PCAI create less bias for the estimation population variances. However, RI underestimates the population variance more, compared to the other imputation methods.

5.9.1.4 Comparing means of imputed and real values

Table 5.6 and Figures 5.9 to 5.12, show the true means and estimated means of imputed values in four variables by different imputation methods. It should be added that the following table and figures show only the means of imputed values and their correspondent true values. In other words, the means have been calculated from 200 imputed values rather than the whole data file. In addition, the means in this section are based on standardized data.

Table 5.6: Means of true and imputed values by different imputation methods

	Y_1	Y_2	Y_3	Y_4
True	5.0024	4.9963	4.9970	5.0002
SSTI	5.0169	5.0024	5.0035	4.9871
EDCI	5.0384	5.0379	5.0392	5.0398
RI	4.9988	4.9987	4.9998	5.0010
PCAI	4.9608	5.0163	5.0143	5.0654
SHDI	4.9977	4.9983	5.0003	4.9993
DSTI	4.9856	5.0060	5.0078	5.0222

Figure 5.9: Comparison of the means of imputed values and true values by different imputation methods

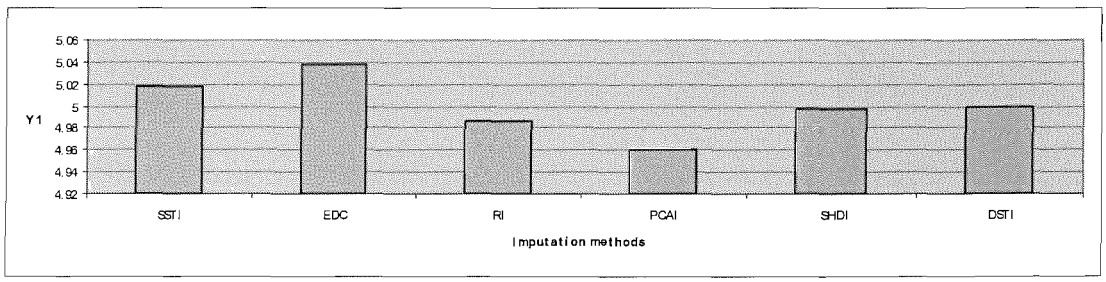


Figure 5.10: Comparison of the means of imputed values and true values by different imputation methods

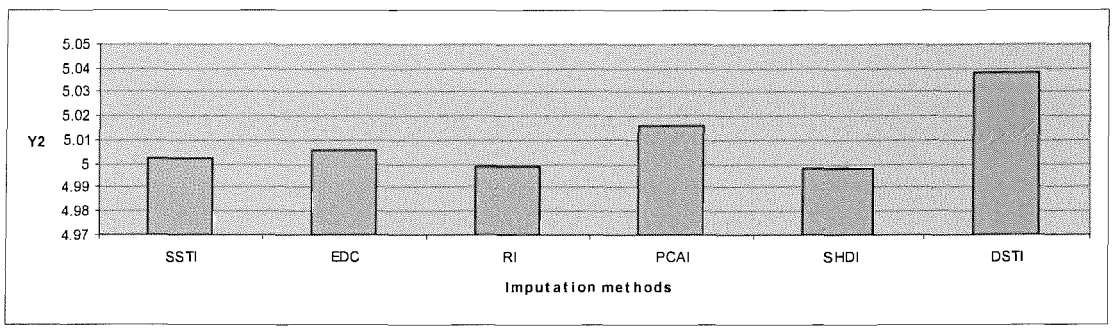


Figure 5.11: Comparison of the means of imputed values and true values by different imputation methods

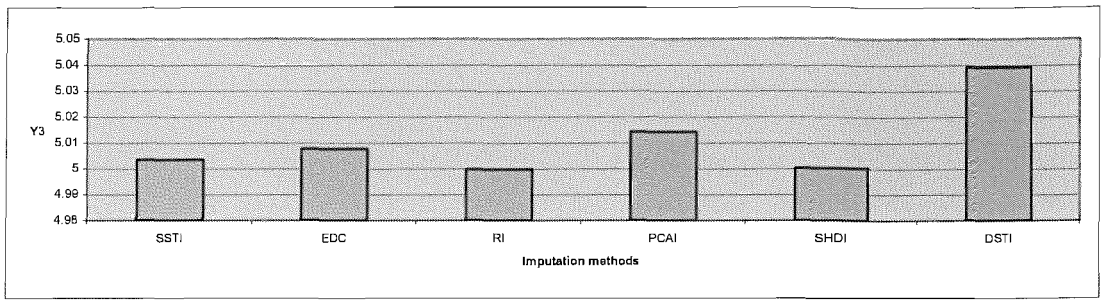
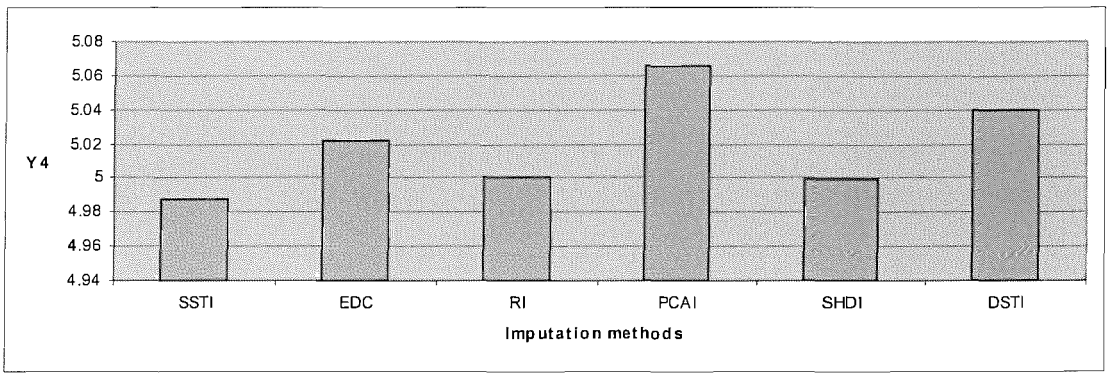


Figure 5.12: Comparison of the means of imputed values and true values by different imputation methods



It can be seen from the above table and figures that the estimation of means by imputed values is slightly better than the estimation of population means, but the structure is the same. However, SHDI, RI, SSTI, and DSTI create less biased estimations for means of imputed values.

5.9.1.5 Comparing variances of imputed and real values

Table 5.7 and Figures 5.13 to 5.16 show the variances of true and imputed values by different imputation methods.

Table 5.7: Variance of true and imputed values by different imputation methods

	Y_1	Y_2	Y_3	Y_4
True	1.00602	1.0002	1.0001	1.0050
SSTI	0.66543	0.6432	0.6447	0.6216
EDCI	0.58513	0.5967	0.5937	0.6091
RI	0.42655	0.3862	0.3747	0.3236
PCAI	0.66133	0.6438	0.6464	0.6270
SHDI	0.9999	1.0033	0.9956	0.9972
DSTI	0.6270	0.6886	0.6450	0.7215

Figure 5.13: Comparison of the variance of imputed values and true values by different imputation methods

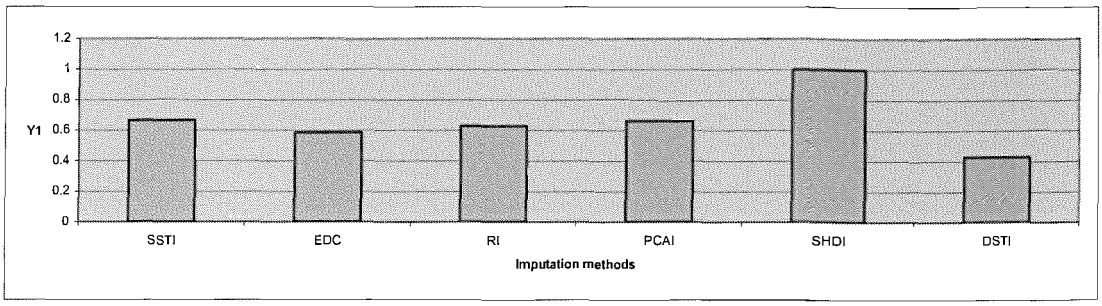


Figure 5.14: Comparison of the variance of imputed values and true values by different imputation methods

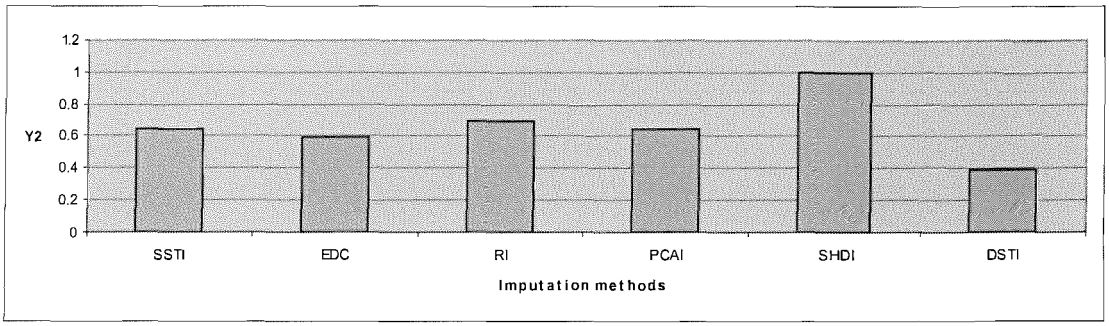


Figure 5.15: Comparison of the variance of imputed values and true values by different imputation methods

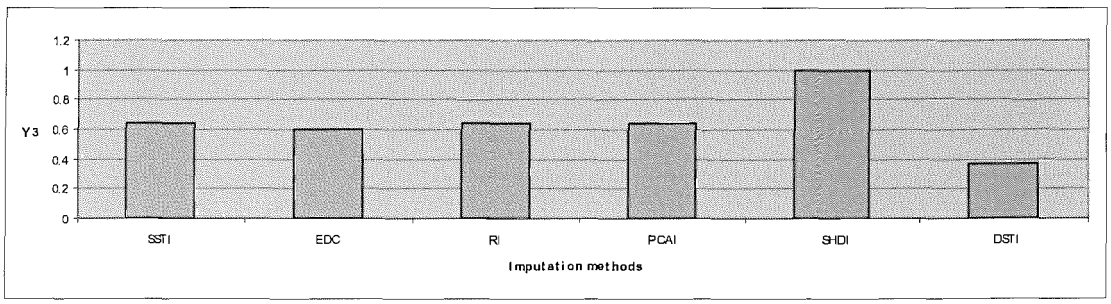
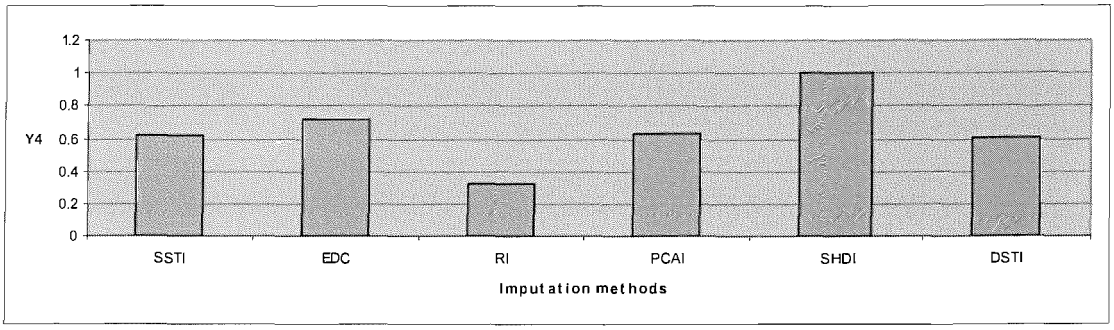


Figure 5.16: Comparison of the variance of imputed values and true values by different imputation methods



The estimation of variances by imputed values and population are the same. As can be seen from the Figures, SHDI has approximately unbiased estimation for variances and other imputation methods underestimate variances. However, DSTI underestimates variances slightly less than other imputation methods.

5.9.1.6 Comparing the correlations of imputed and real values

Table 5.8 and Figures 5.17 to 5.20; show the correlations between true and imputed values by different imputation methods in four variables.

Table 5.8: Correlation between true and imputed values

	Y_1	Y_2	Y_3	Y_4
SSTI	0.58707	0.5519	0.5541	0.5198
EDCI	0.60803	0.5702	0.5753	0.5424
RI	0.64642	0.6164	0.6028	0.5598
PCAI	0.59111	0.5533	0.5558	0.5234
SHDI	0.0024	-0.0048	-0.0044	0.0007
DSTI	0.5986	0.5492	0.5533	0.5332

Figure 5.17: Comparison of the correlation of imputed values and true values by different imputation methods

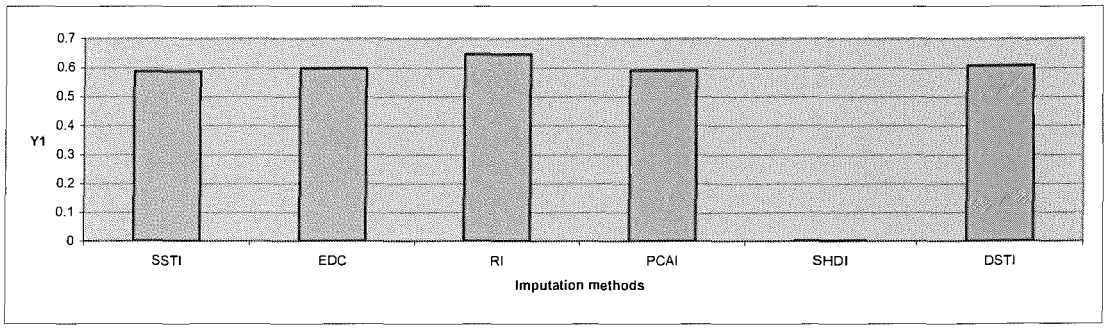


Figure 5.18: Comparison of the correlation of imputed values and true values by different imputation methods

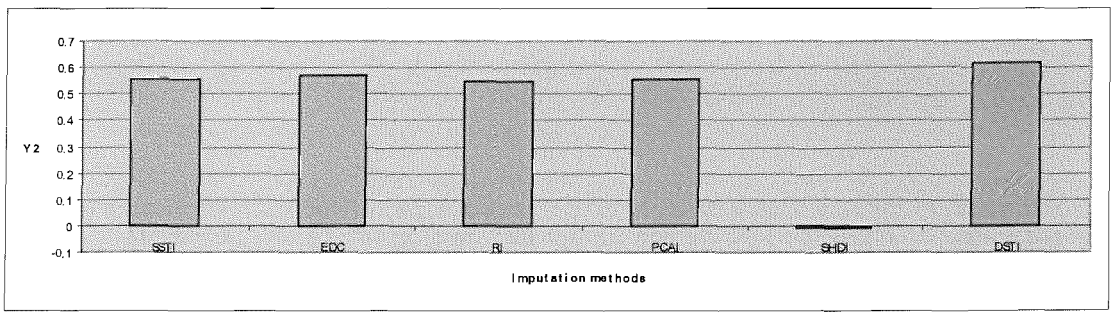


Figure 5.19: Comparison of the correlation of imputed values and true values by different imputation methods

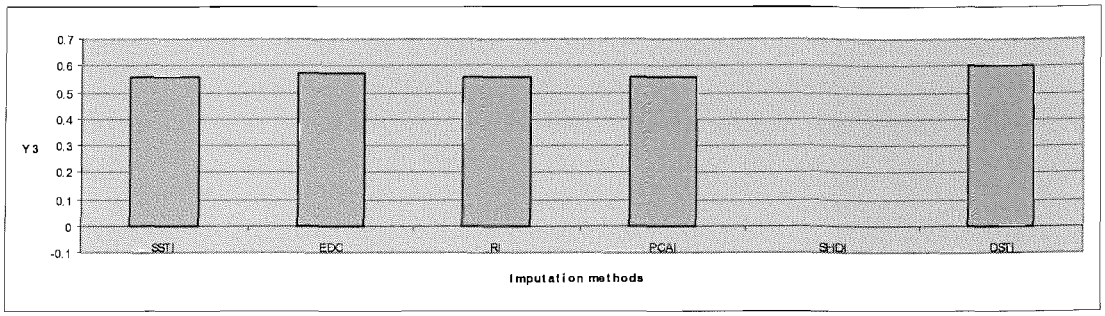
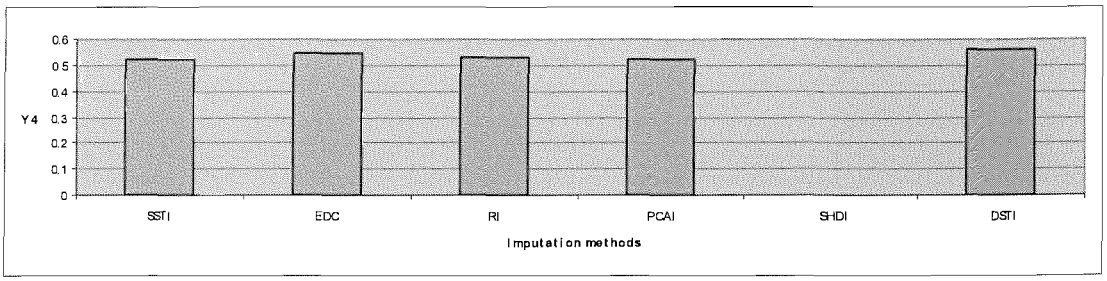


Figure 5.20: Comparison of the correlation of imputed values and true values by different imputation methods



As can be seen from Table 5.8 and Figures 5.17 to 5.20, all of the imputation methods apart from SHDI preserve individual values. For all of the four variables, RI has the best preservation of individual values, followed by DSTI.

5.9.2 Simulation under MAR assumption

This section is similar to section 5.6.1; the only difference being the missing data mechanism. The occurrence of missing values is under the MAR assumption and other simulation parameters such as imputation methods, the definition of the simulated data and ordering methods are the same as in the previous section.

5.9.2.1 Test of agreement or Wald statistic

Table 5.9 shows percentages of insignificant Wald statistics for two scenarios as described in section 5.5.1, for variable Y_4 under the MAR assumption. First and second columns show Wald statistics when the expected number of missing values is 197 and 304. Only the Wald statistic results for variable Y_4 are presented, because of the similarity between the results for this variable and other variables.

Table 5.9: The percentages of insignificant Wald statistics for different imputation methods under MAR

Number of missing values	197	304
SSTI	39.4	27.85
EDCI	48.6	27.9
RI	0.2	0
PCAI	47	31.3
SHDI	96	95.5
DSTI	75.6	61.15

It can be seen from Table 5.9 that when the percentage of missing values is approximately 20% then SHDI preserves the marginal distribution of Y_4 in 96% of iterations, followed by DSTI is in second place with 75.6% of iterations. In addition, EDCI, PCAI, SSTI, and RI preserve marginal distribution of Y_4 in 48.6%, 47%, 39.4%, and 0.2% of iterations, respectively. When the percentage of missing values is 30%, SHDI preserves the marginal distribution of Y_4 in 95.5% of iterations, followed by DSTI is in second place with 61.15% of iterations. In addition, PCAI, EDCI, SSTI, and RI preserve marginal distribution of Y_4 in 31.3%, 27.9%, 27.85%, and 0% of iterations, respectively.

5.9.2.2 Comparison of population means

Table 5.10 and Figures 5.21 to 5.24 show the means of the database without considering missing values and the database with imputed values. These means are population means in 2000 iterations under the MAR assumption.

Table 5.10: Population means by different imputation methods under MAR

	Y_1	Y_2	Y_3	Y_4
True	50.000	40.000	60.000	35.000
STI	50.003	39.999	60.001	34.983
EDCI	50.010	40.013	60.021	35.006
RI	49.997	40.002	60.003	34.995
PCAI	49.983	40.009	60.013	35.017
SHDI	49.996	40.003	60.005	35.023
DSTI	49.996	40.009	60.007	35.002

Figure 5.21: Comparison of the means of population

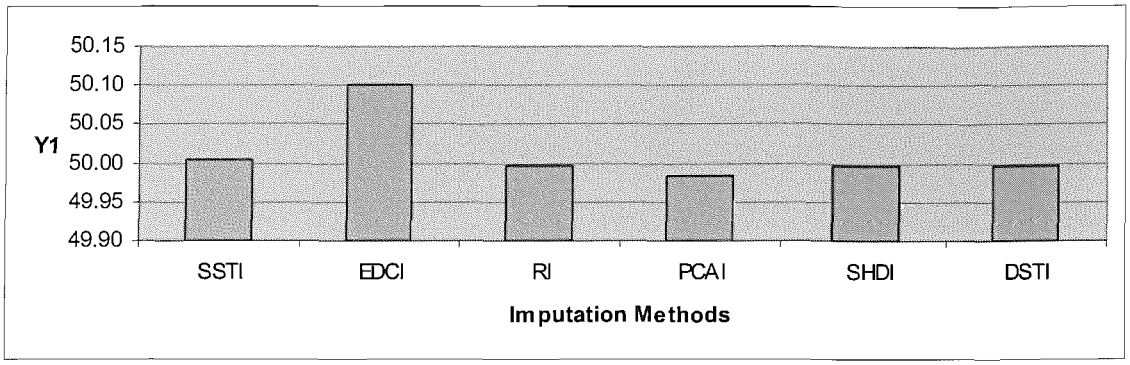


Figure 5.22: Comparison of the means of population

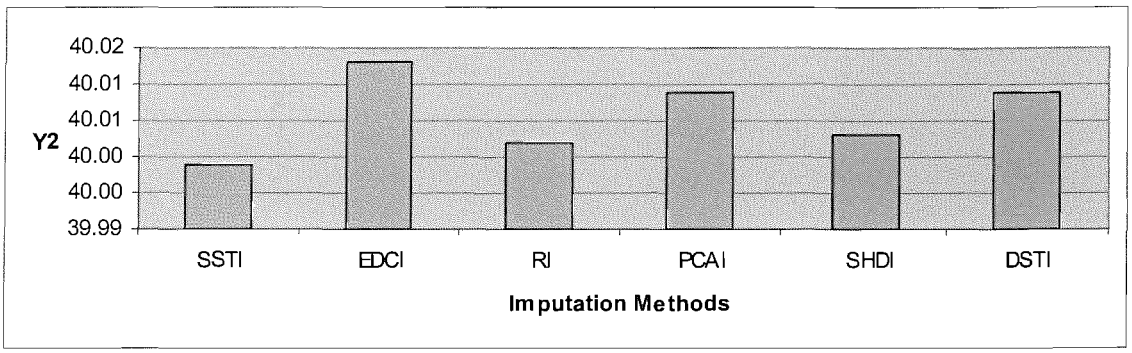


Figure 5.23: Comparison of the means of population

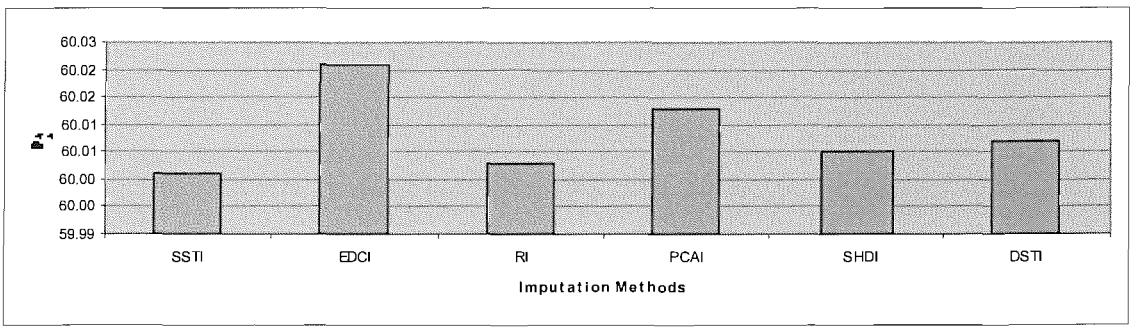
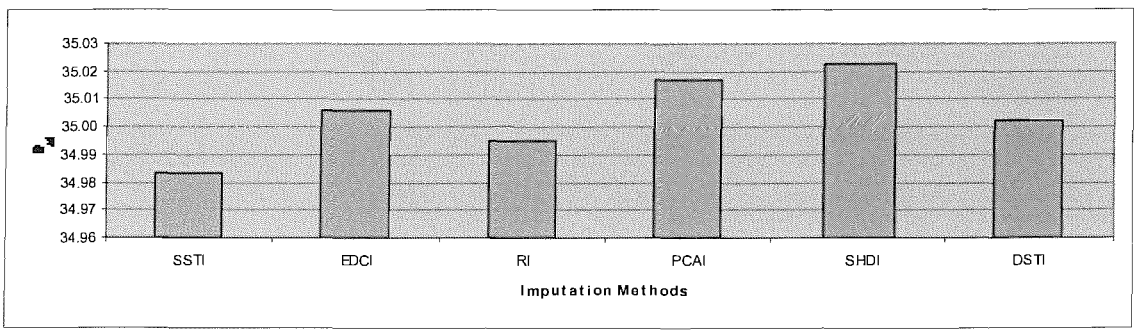


Figure 5.24: Comparison of the means of population



It can be seen from Table 5.9 and Figures 5.21 to 5.24 that the estimated means by almost all imputation methods are close to the population means.

5.9.2.3 Comparison of population variances

Table 5.11 and Figures 5.25 to 5.28 show the variances of the database without considering missing values and the database with imputed values. These variances are the population variances in 2000 iterations under the MAR assumption.

Table 5.11: Population variances by different imputation methods under MAR

	Y_1	Y_2	Y_3	Y_4
True	4.000	3.000	7.000	4.000
STI	3.772	2.813	6.530	3.725
EDCI	3.705	2.786	6.451	3.699
RI	3.547	2.637	6.119	3.458
PCAI	3.763	2.813	6.527	3.718
SHDI	4.002	3.005	6.969	3.984
DSTI	3.977	3.409	6.987	3.785

As can be seen from Table 5.11 and Figures 5.25 to 5.28, the variances of the sample mean by SHDI are close to the true means, but other imputation methods underestimate it for all variables. However, DSTI underestimates variances less than the other imputation methods and RI underestimates variances the most.

Figure 5.25: Comparison of the variance of population

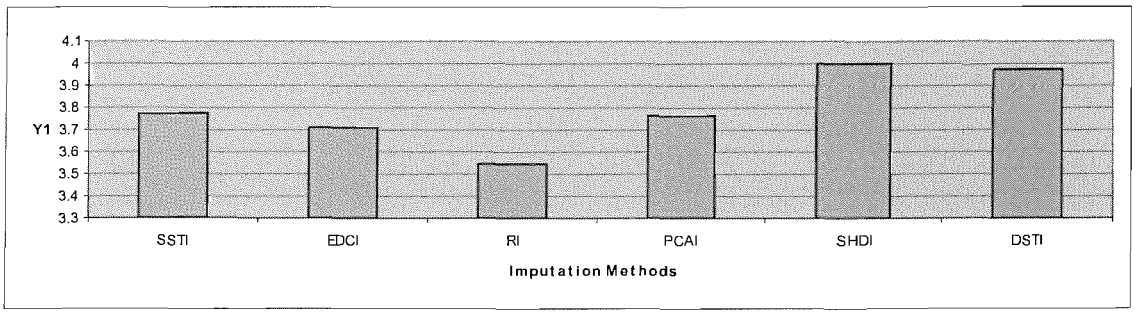


Figure 5.26: Comparison of the variance of population

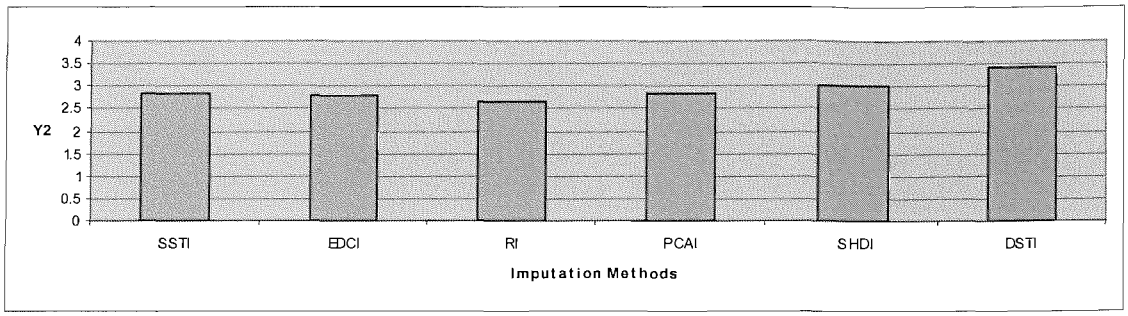


Figure 5.27: Comparison of the variance of population

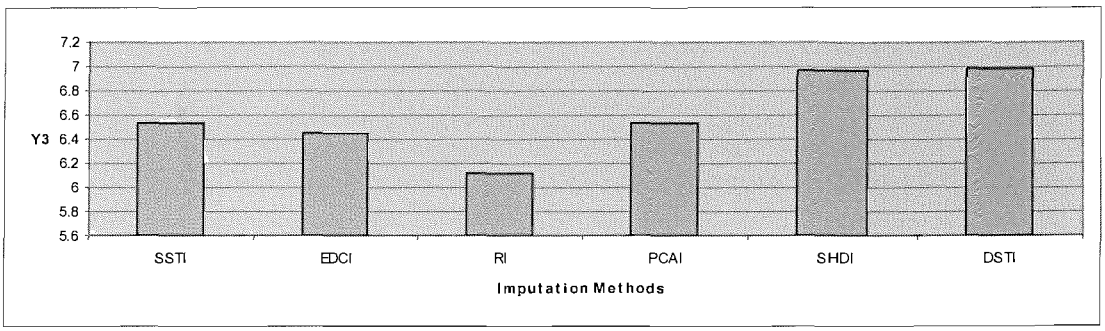
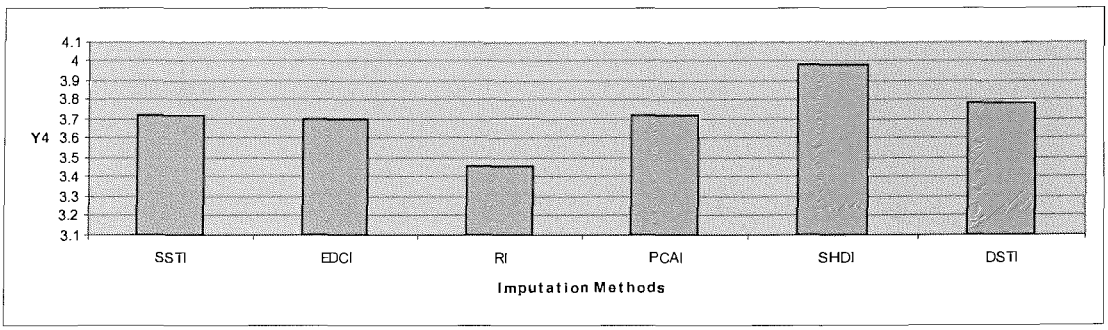


Figure 5.28: Comparison of the variance of population



5.9.2.4 Comparing means of imputed and real values

Table 5.12 and Figures 5.29 to 5.32 show means and the graphs of imputed values under MAR and their correspondent real values.

Table 5.12: Means of true and imputed values by different imputation methods under MAR

	Y_1	Y_2	Y_3	Y_4
True	5.007	4.995	4.997	4.933
STI	5.019	4.988	4.992	4.896
EDCI	5.037	5.030	5.030	4.958
RI	5.005	4.996	4.996	4.928
PCAI	4.968	5.018	5.015	4.986
SHDI	5.000	5.000	4.999	5.000
DSTI	5.002	5.019	5.003	4.946

Figure 5.29: Comparison of the means of imputed values and true values

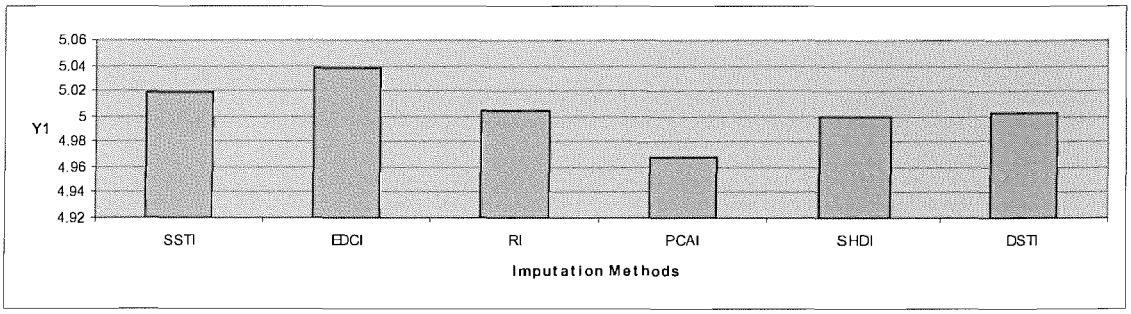


Figure 5.30: Comparison of the means of imputed values and true values

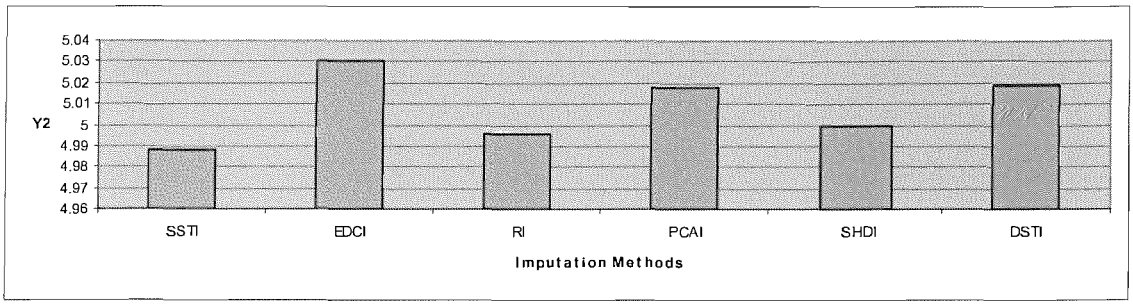


Figure 5.31: Comparison of the means of imputed values and true values

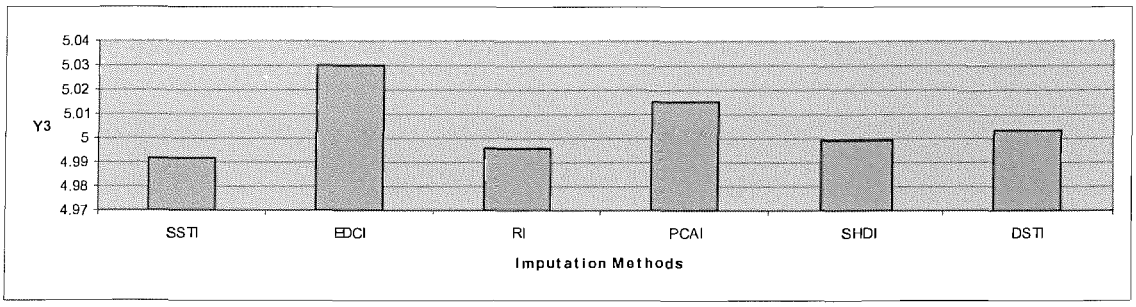
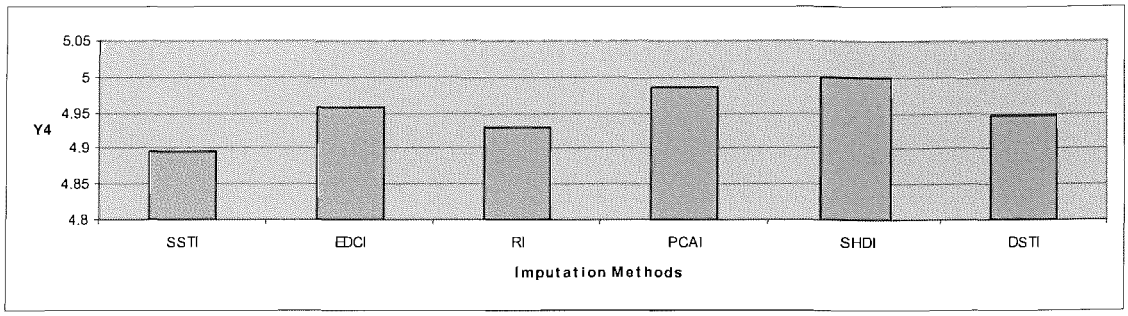


Figure 5.32: Comparison of the means of imputed values and true values



It can be seen from the above table and figures that the mean of imputed values by different imputation methods are close to the true means. DSTI gives better estimated means with compare to SHDI in all variables except in Y_2. In addition, estimated means by other imputation methods are very close to true means too.

5.9.2.5 Comparing variances of imputed and real values

Table 5.13 and Figures 5.33 to 5.36 show the variances of true and imputed values and their graphs by different imputation methods under MAR. It can be seen from the following tables and figures that SHDI slightly underestimates the variance and other imputation methods underestimate the variance in each variable. However, DSTI underestimates the variance less than other imputation methods and is very close to true variance.

Table 5.13: Variances of true and imputed values by different imputation methods under MAR

	Y_1	Y_2	Y_3	Y_4
True	0.9966	0.9993	1.0060	1.0050
STI	0.6929	0.6589	0.6673	0.6460
EDCI	0.6134	0.6181	0.6161	0.6253
RI	0.4084	0.3605	0.3679	0.3082
PCAI	0.6872	0.6633	0.6712	0.6502
SHDI	0.9980	0.9961	1.0017	0.9970
DSTI	0.9686	0.9560	1.0013	0.7356

Figure 5.33: Comparison of the variance of imputed values and true values

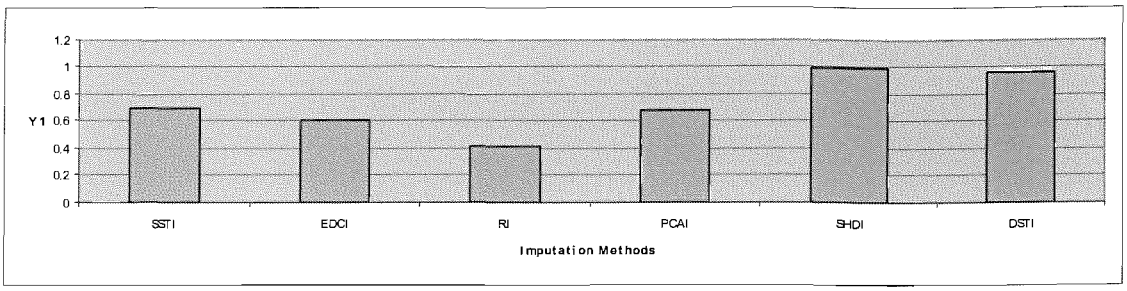


Figure 5.34: Comparison of the variance of imputed values and true values

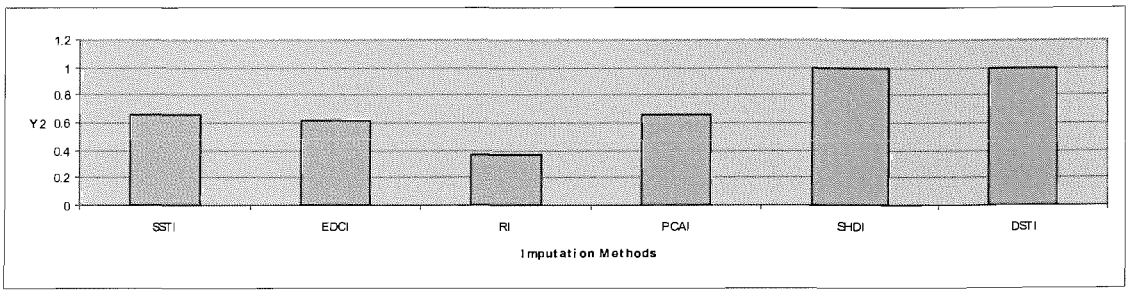


Figure 5.35: Comparison of the variance of imputed values and true values

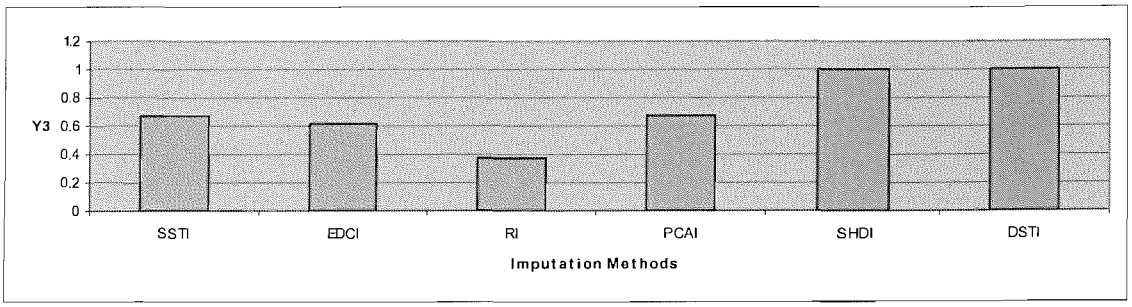
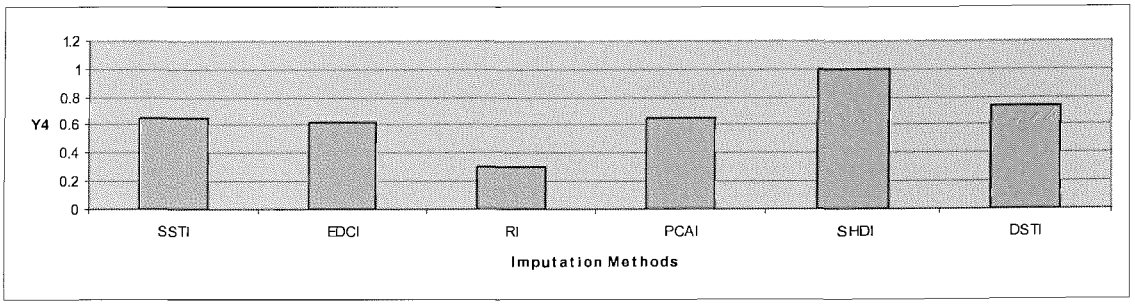


Figure 5.36: Comparison of the variance of imputed values and true values



5.9.2.6 Comparing correlations of imputed and real values

Table 5.14 and Figures 5.37 to 5.40 show the correlations between true and imputed values under MAR and their graphs by different imputation methods.

Table 5.14: Correlations between true and imputed values by different imputation methods under MAR

	Y_1	Y_2	Y_3	Y_4
STI	0.571	0.533	0.548	0.511
EDCI	0.594	0.557	0.572	0.538
RI	0.625	0.594	0.597	0.551
PCAI	0.574	0.537	0.552	0.518
SHDI	0.000	0.003	-0.002	-0.001
DSTI	0.561	0.506	0.519	0.527

Figure 5.37: Comparison of the correlation of imputed values and true values

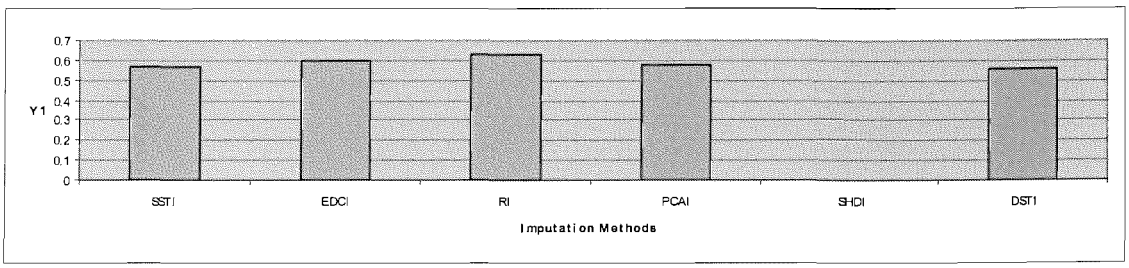


Figure 5.38: Comparison of the correlation of imputed values and true values

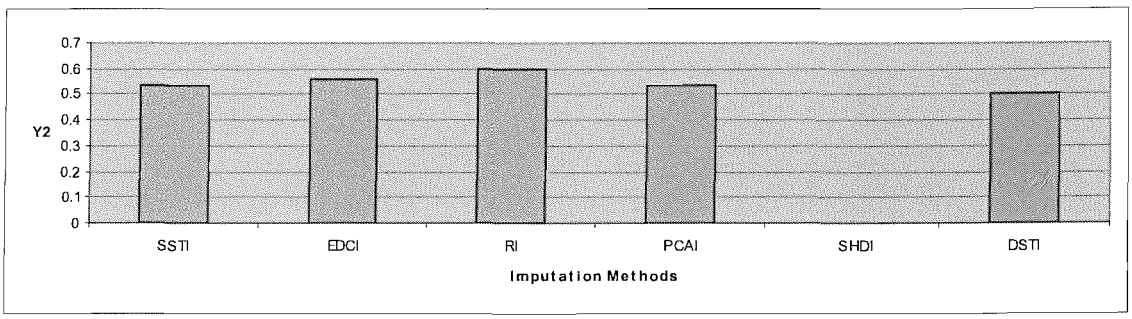


Figure 5.39: Comparison of the correlation of imputed values and true values

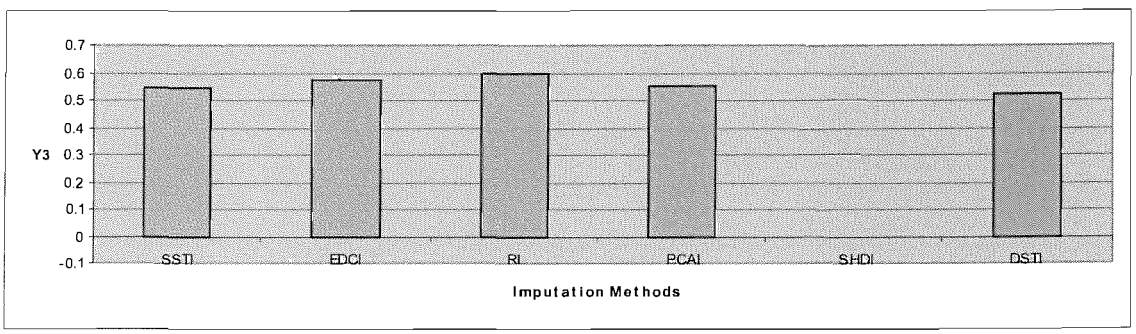
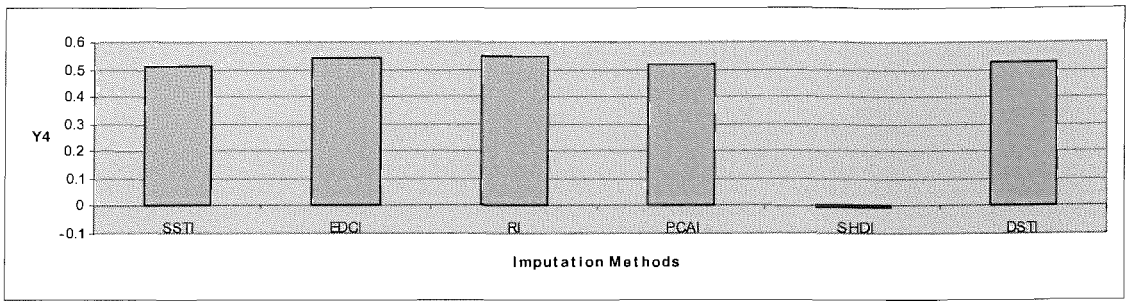


Figure 5.40: Comparison of the correlation of imputed values and true values



It can be seen from Table 5.14 and Figures 5.37 to 5.40 that correlations between true and imputed values under MAR have approximately the same pattern with correlations under the MCAR assumption. RI and SHDI have the highest and lowest correlations between true and imputed values for all variables respectively. In other words, RI preserves the relationship of imputed and true values the most, but on the other hand SHDI does not preserves the relationship between imputed and true values at all. However, other imputation methods have slightly less correlations than RI and they preserve the relationship between true and imputed values less than RI.

Part II

Systematic and Ranked Set Sampling based on Concomitants of Order Statistics

Chapter 6

Ordered systematic and ranked set sampling

6.1 Introduction

New sampling methods, such as ranked-set and double sampling, based on applications of order statistics have been introduced in recent years. In this chapter, the statistical properties of ranked-set sampling are examined, the usual systematic sampling (SY) scheme is modified and a variance estimator for ordered systematic sampling (OSY) is suggested. As in previous chapters, we assume that an auxiliary variable is used to order data. This variable is referred to as a concomitant variable, in what follows. By assuming a linear relationship between the variable of interest and its concomitant, a variance estimator can then be developed for the sample mean under OSY. This estimate turns out to be less biased compared to other variance estimators for the sample mean under OSY. In addition, we compare the statistical properties of ranked-set and OSY with those of simple random sampling. We justify the proposed variance estimator theoretically, and demonstrate its properties using a simulation study.

6.2 Ordered systematic sampling (OSY)

Systematic sampling is a practical and efficient method for selecting samples from administrative registers or other logically arranged files. A proper sorting order of the population ensures that the sample obtained reflects true population distributions. When a population is ordered, it is clear that the selection of a systematic sample will provide a heterogeneous sample and that the variance of the sample mean will generally be smaller than the variance of the sample mean under simple random sampling.

The reason for this is intuitively clear. A systematic sample will cover the whole population and will avoid the chances of selecting samples containing too many large or small values. That is to say, a systematic sample will tend to be more representative of the population than a random sample. In general, we may say that the sampled units in a systematic sample from an ordered population will generally be more heterogeneous than those in a simple random sample. Hence, the intraclass correlation will be small and, as will be explained in the following sections, the variance of the sample mean under systematic sampling can be expected to be smaller than the variance of a sample mean under simple random sampling.

This chapter is organised as follows. In the following section, the procedure of sample selection under OSY is described. In Section 4, the sample mean is shown to be an unbiased estimator for the population mean under OSY given appropriate assumptions. In section 5, the variance of the sample mean is obtained using concomitant order statistics theory. Finally, in Section 6, the relative precision of the variance of the sample mean under OSY is derived.

6.3 Notation and sample selection procedure

This section introduces appropriate assumptions and notation, and describes the OSY sample selection procedure.

As in the classic systematic sampling scheme, we assume the study variable is Y and the concomitant variable is X. The population size is N, the sample size is n and $k =$

N/n, the sampling step that, for simplicity, is assumed to be an integer. The values obtained for any specific item in the N units that comprise the population are denoted by $Y_1, Y_2, ..., Y_N$ and $X_1, X_2, ..., X_N$. In the following analysis, we assume x_{ij} is the jth member of the ith systematic sample, with $i = 1, 2, ..., k$, $j = 1, 2, ..., n$. Here, each systematic sample is considered as a class; therefore, we have k classes. In addition, we use the following notation for the super population moments of X and Y.

μ_Y = Population mean of Y , μ_X = Population mean of X

σ_Y = Population variance of Y , σ_X = Population variance of X

ρ = Correlation between X and Y , δ = Intraclass correlation of X

where

$$\delta = \frac{E\left(x_{ij} - \mu_X\right)\left(x_{ij'} - \mu_X\right)}{E\left(x_{ij} - \mu_X\right)^2}, \quad i = 1, 2, .., k \text{ and } j \neq j' = 1, 2, ..., n.$$

In other words, δ expresses the degree of homogeneity in a systematic sample which is the correlation coefficient between pairs of units in the same systematic sample under the following model:

$$Var(x_{ij}) = \sigma_X^2, \; Cov(x_{ij}, x_{ij'}) = \delta\sigma_X^2, \; i \neq j \text{ and } Cov(x_{ij}, x_{i'j'}) = 0, \; i \neq i'.$$

From Sarndal, Swensson and Wretman (1997), δ can be estimated as follows:

$$\hat{\delta} = 1 - \frac{n}{n-1} \cdot \frac{SSW}{SST},$$

where n is the sample size in each systematic sample, SSW is the variation within systematic samples, and SST is the total variation. Positive values of $\hat{\delta}$ show elements in the same sample tend to have similar values. $\hat{\delta}$ =1 if $SSW=0$, which means there is no variation within systematic samples or in other words there is complete homogeneity. On the other hand, if $SSB=0$, which means there is complete heterogeneity within samples, and our sample is a proper representative of population. A systematic sample from a perfectly ordered population has almost complete heterogeneity.

Finally, we assume a linear relationship between Y and X as follows:

$$Y = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X) + \varepsilon, \tag{6.1}$$

where ε and X are independently distributed.

6.3.2 The sample selection procedure

To select a sample via OSY the population is first ordered according to the concomitant X in ascending order as follows:

$$X_{(1:N)} \le X_{(2:N)} \le \cdots \le X_{(N:N)}.$$

The Y-variates paired with these order statistics are denoted by

$$Y_{[1:N]}, Y_{[2:N]}, \cdots, Y_{[N:N]}.$$

The $Y_{[j:N]}$ are not necessarily ordered, but can be expected to reflect the association between X and Y, a strong positive association tending to lead to values of $Y_{[j:N]}$ in roughly ascending order, and similarly a negative association tending to lead to $Y_{[j:n]}$ in descending order. A sample is then taken from the ordered population using a systematic sampling procedure. That is a random integer θ between 1 and N/n is selected and then every $\theta + j\,{}^{N}\!/_{n}$, $j=0,1,...,n-1$ unit in the ordered population is selected into the sample. The Y values associated with this sample are:

$$y_{[1:n]}, y_{[2:n]}, \cdots, y_{[n:n]},$$

and the corresponding values of the concomitant X are

$$x_{(1:n)} \le x_{(2:n)} \le \cdots \le x_{(n:n)}.$$

Note that (6.1) is still valid for this sample. Consequently

$$y_{[j:n]} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}\left(x_{(j:n)} - \mu_X\right) + \varepsilon_{[j:n]}, \tag{6.2}$$

where $\varepsilon_{[j:n]}$ and $x_{(j:n)}$ are independent.

6.4 Estimation of the population mean

By assuming $N=nk$, it can be proved that the sample mean is an unbiased estimator of the population mean under OSY. Let

$$\bar{y}_{OSY} = \frac{1}{n}\sum_{j=1}^{n} y_{[j:n]}, \tag{6.3}$$

be the sample mean under OSY.

Theorem 6.1. The sample mean $\bar{y}_{OSY}$ is an unbiased estimator of μ_Y.

Proof: By its definition,

$$E(\bar{y}_{OSY}) = E\left(\frac{1}{n}\sum_{j=1}^{n} y_{[j:n]}\right) = \frac{1}{n}E\left(\sum_{j=1}^{n} y_{[j:n]}\right) = \frac{1}{n}\sum_{j=1}^{n} E\left(y_{[j:n]}\right).$$

By substituting (2.9) we then have:

$$E(\bar{y}_{OSY}) = \frac{1}{n}\sum_{j=1}^{n}\left(\mu_Y + \rho\sigma_Y\alpha_{j:n}\right), \text{ where } \alpha_{j:n} = E\left(\frac{x_{(j:n)} - \mu_X}{\sigma_X}\right).$$

Hence,

$$E(\bar{y}_{OSY}) = \mu_Y + \rho\frac{\sigma_Y}{n}\sum_{j=1}^{n} E\left(\frac{x_{(j:n)} - \mu_X}{\sigma_X}\right),$$

or equivalently, since $\sum_{j=1}^{n} x_{(j:n)} = \sum_{j=1}^{n} x_j$, where $x_1, x_2, ..., x_n$ is a random sample from X, (David and Nagaraja, 2003) we have

$$E(\bar{y}_{OSY}) = \mu_Y + \rho\frac{\sigma_Y}{n\sigma_X}E\sum_{j=1}^{n}\left(x_j - \mu_X\right).$$

It immediately follows that

$$E(\bar{y}_{OSY}) = \mu_Y + \rho\frac{\sigma_Y}{n\sigma_X}\left(n\mu_X - n\mu_X\right) = \mu_Y. \tag{6.4}$$

Therefore $\bar{y}_{OSY}$ is an unbiased estimator of the population mean μ_Y.

6.5 The variance and other statistical properties of the sample mean

We now derive the variance of $\bar{y}_{osy}$ under OSY. In a classic systematic sampling scheme, the variance of the sample mean depends on the sample means generated by all k systematic samples. Hence, it cannot be used for practical applications. Instead the variance of the sample mean under simple random sampling (SRS) is used as the

classic systematic variance estimator, even though the variance estimator under SRS overestimates the variance of the sample mean under OSY. In this section we therefore develop a variance estimator for the sample mean under OSY that needs only the intraclass correlation of X and is more efficient than the variance estimator suggested by SRS.

Put

$$\delta = \frac{E\left(x_{(j:n)} - \mu_X\right)\left(x_{(j':n)} - \mu_X\right)}{E\left(x_{(j:n)} - \mu_X\right)^2}, \; j \neq j' = 1,2,..,n, \tag{6.5}$$

That is, δ is the intraclass correlation coefficient between the X values of all pairs of units in the same systematic sample. From (6.5) and using the definition of covariance we have:

$$\delta E\left(x_{(j:n)} - \mu_X\right)^2 = E\left(x_{(j:n)} - \mu_X\right)\left(x_{(j':n)} - \mu_X\right)$$

$$\delta\sigma^2_{X_{(j:n)}} = E\left(x_{(j:n)} - \mu_X\right)\left(x_{(j':n)} - \mu_X\right) = Cov\left(x_{(j:n)}, x_{(j':n)}\right) \; . \tag{6.6}$$

Theorem 6.2. The variance of sample mean under OSY is:

$$Var\left(\bar{y}_{OSY}\right) = \frac{\sigma_Y^2}{n}\left(1 + (n-1)\rho^2\delta\right) \; , \tag{6.7}$$

where ρ is the correlation between Y and X and δ is the intraclass correlation of X over all k possible systematic samples. (The following proof is based on a private communication from R. Sugden.)

Proof: By its definition

$$Var(\bar{y}_{OSY}) = Var\left(\frac{1}{n}\sum_{j=1}^{n} y_{[j:n]}\right) = \frac{1}{n^2} Var\left(\sum_{j=1}^{n} y_{[j:n]}\right).$$

By substituting (2.7) above we have:

$$n^2 Var(\bar{y}_{OSY}) = Var\left(\sum_{j=1}^{n}\left(\mu_Y + \frac{\rho\sigma_Y}{\sigma_X}\left(X_{(j:n)} - \mu_X\right) + \varepsilon_{[j:n]}\right)\right). \tag{6.8}$$

Simplifying,

$$n^2 Var(\bar{y}_{OSY}) = Var\left(n\mu_Y + \frac{\rho\sigma_Y}{\sigma_X}\sum_{j=1}^{n}\left(X_{(j:n)} - \mu_X\right) + \sum_{j=1}^{n}\varepsilon_{[j:n]} \right).$$

From the independence of X and ε we have:

$$n^2 Var(\bar{y}_{OSY}) = \frac{\rho^2\sigma_Y^2}{\sigma_X^2} Var\left(\sum_{j=1}^{n} X_{(j:n)} \right) + Var\left(\sum_{j=1}^{n}\varepsilon_{[j:n]} \right).$$

Since $\sum_{j=1}^{n} x_{(j:n)} = \sum_{j=1}^{n} x_j$, and $\sum_{j=1}^{n}\varepsilon_{[j:n]} = \sum_{j=1}^{n} e_j$ where $x_1, x_2, ..., x_n$ is a random sample from X (David and Nagaraja, 2003), we have:

$$n^2 Var(\bar{y}_{OSY}) = \frac{\rho^2\sigma_Y^2}{\sigma_X^2} Var\left(\sum_{j=1}^{n} X_j \right) + Var\left(\sum_{j=1}^{n}\varepsilon_j \right). \tag{6.9}$$

In (6.9) we need to calculate $Var\left(\sum_{j=1}^{n} X_j \right)$ and $Var\left(\sum_{j=1}^{n}\varepsilon_j \right)$. Now,

$$\begin{aligned} Var\left(\sum_{j=1}^{n} X_j \right) &= \sum_{j=1}^{n} Var\left(X_j\right) + 2\sum_{j<j'} Cov\left(X_j, X_{j'}\right) \\ &= n\sigma_X^2 + 2\frac{n(n-1)}{2}\delta\sigma_X^2 \\ &= n\sigma_X^2 + n(n-1)\delta\sigma_X^2 \end{aligned} \tag{6.10}$$

and, from the independence between ε_i and $\varepsilon_{i'}$ for all i, we have:

$$\begin{aligned} Var\left(\sum_{j=1}^{n}\varepsilon_j \right) &= \sum_{j=1}^{n} Var\left(\varepsilon_j\right) \\ &= n\sigma_Y^2\left(1-\rho^2\right). \end{aligned} \tag{6.11}$$

By substituting (6.10) and (6.11) in (6.9), we have:

$$Var(\bar{y}_{OSY}) = \frac{\sigma_Y^2}{n}\left[1 + (n-1)\rho^2\delta\right], \tag{6.12}$$

where σ_Y^2 is the population variance of Y, ρ is the correlation between X and Y variates, and δ is the intraclass correlation of all k possible samples in X. This completes the proof.

Now from Theorem 6.2, we can obtain a variance estimator of the sample mean based on the available sample data as follows:

$$Var\left(\bar{Y}_{OSY}\right)=\frac{s_y^2}{n}\cdot\frac{N-1}{N}\left[1+(n-1)r^2\delta\right], \tag{6.13}$$

where s_y^2 is the sample variance, r is the sample correlation and δ is the intraclass correlation of X over all k possible systematic samples.

6.6. The relative precision of the sample mean under OSY

In this section, we compare the variance of the sample mean under OSY with the variance of the same mean under classic systematic and simple random sampling. The variance of the sample mean under systematic sampling, (Cochran, 1978) is:

$$Var(\bar{y}_{SY})=\frac{\sigma_Y^2}{n}\left[1+(n-1)\delta_1\right], \tag{6.14}$$

where σ_Y^2 is the population variance of Y and δ_1 is the intraclass correlation of Y over all k possible samples. Calculating $Var(\bar{y}_{SY})$ requires all k possible systematic samples, and so (6.14) cannot be used for practical applications. Under certain conditions, however, we may consider a systematic sample to approximate a simple random sample and so we can use the variance under SRS as an approximation. On the other hand, $Var(\bar{y}_{OSY})$ is computable and, as shown in the next section, it is more efficient than $Var(\bar{y}_{SRS})$.

The relative precision of ordered systematic sampling in relation to simple random sampling is given by:

$$RP_{OSY}=\frac{Var\left(\bar{Y}_{SRS}\right)}{Var\left(\bar{Y}_{OSY}\right)}=\frac{\sigma_y^2/n}{\frac{\sigma_y^2}{n}\left[1+(n-1)\rho^2\delta\right]}=\frac{1}{1+(n-1)\rho^2\delta}, \tag{6.15}$$

As can be seen from (6.15), the performance of OSY in relation to SRS is therefore dependent on the correlation of X and Y and the intraclass correlation of X. The performance of OSY is better than SRS if, and only if, $\delta<0$. Sarndal, Swensson and

Wretman (1997) show that the efficiency of systematic sampling depends on the orders of population. In addition, they show that:

$$\delta = 1 - \frac{n}{n-1} \cdot \frac{SSW}{SST}, \tag{6.16}$$

In addition, Sarndal, Swensson and Wretman (1997) show that for populations ordered by X,

$$\delta \cong \frac{-1}{n-1} < 0.$$

We can therefore conclude that the precision of OSY is always better than that of SRS, except when the correlation of X and Y equals zero, in which case OSY and SRS have the same precision.

6.7 Ranked set and median ranked set sampling (RSS & MRSS)

Ranked set sampling was first suggested by McIntyre (1952), with Takashi and Wakimito (1968) giving the necessary mathematical theory. Ranked set sampling is designed for situations where the study variable Y is difficult or expensive to measure, but where ranking in small subsets is easy. Concomitants of order statistics enter when the ranking is subject to error or is not perfect. Applications of concomitants of order statistics in ranked set sampling were studied by Dell and Clutter (1972). They showed that when there is no ranking error the mean of RSS is an unbiased estimator for population mean. Moreover, they showed that the variance of the sample mean under RSS is smaller than the variance of the sample mean under SRS. Patil, Shina and Tailllie (1993) showed that the sample mean under RSS is more efficient than the regression estimator under SRS, when the correlation between the study variable and its concomitant is less than 0.85. Finally, Muttlak (1997) has suggested using MRSS to reduce errors in ranking, thereby increasing the efficiency of RSS.

The next section reviews ranked set and median ranked sampling schemes, and presents a variance estimator for the sample mean under these schemes when there is a concomitant variable. Note that there are many aspects of ranked set sampling that do not involve concomitants. See the comprehensive review paper by Patil , Shina and Tailllie (1992).

6.8 The RSS and MRSS algorithms

Suppose Y is the study variable, X is the concomitant variable, N is the population size, m is the set size, r is the cycle size, $n = mr$ is the sample size, and ρ is the correlation between X and Y. In addition, we assume a linear relationship between X and Y as specified by (6.1).

The RSS procedure can be summarized as follows:

a) Select m^2 sample units from the concomitant variable X.

b) Randomly allocate the m^2 selected units into m sets each of size of m.

c) Rank the units within each set.

d) Choose a sample of size m to include the smallest ranked unit from the first set, the second smallest unit from the second set and so on.

e) Measure the m associated values of the study variable Y.

f) Repeat the above steps r times until $n = mr$ sample values of Y are obtained.

For a graphic view of the above steps, see Tables 6.1 and 6.2.

Table 6.1 Summary of steps a to c of RSS

Sample→ Set ↓	1	2	⋯	i	⋯	m
1	$x_{(1:m)1}$	$x_{(2:m)1}$	⋯	$x_{(i:m)1}$	⋯	$x_{(m:m)1}$
2	$x_{(1:m)2}$	$x_{(2:m)2}$	⋯	$x_{(i:m)2}$	⋯	$x_{(m:m)2}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	$x_{(1:m)k}$	$x_{(2:m)k}$	⋯	$x_{(i:m)k}$	⋯	$x_{(m:m)k}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
m	$x_{(1:m)m}$	$x_{(2:m)m}$	⋯	$x_{(i:m)m}$	⋯	$x_{(m:m)m}$

Here $x_{(i:m)k}$ $i = 1,2,...,m$, $k = 1,2,...,m$ is the sample from the concomitant variable X.

Table 6.2 Summary of steps d to e of RSS

Sample→ Set ↓	1	2	⋯	i	⋯	m
1	$y_{[1:m]1}$	$y_{[2:m]1}$	⋯	$y_{[i:m]1}$	⋯	$y_{[m:m]1}$
2	$y_{[1:m]2}$	$y_{[2:m]2}$	⋯	$y_{[i:m]2}$	⋯	$y_{[m:m]2}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	$y_{[1:m]k}$	$y_{[2:m]k}$	⋯	$y_{[i:m]k}$	⋯	$y_{[m:m]k}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
m	$y_{[1:m]m}$	$y_{[2:m]m}$	⋯	$y_{[i:m]m}$	⋯	$y_{[m:m]m}$

Therefore, the first set of the RSS sample is the diagonal units from Table 6.2 That is $y_{[i:m]k} \quad i=k=1,2,...,m$, where m is the set and sample size in each set. In other words, the first set of all r (cycle size) possible sample sets are:

$y_{[1:m]1}, y_{[2:m]2}, \cdots, y_{[i:m]k}, \cdots, y_{[m:m]m}$.

For simplicity, we can drop the second index k, and show the first sample from the first cycle as $y_{[1:m]}, y_{[2:m]}, ..., y_{[m:m]}$. Note that we have to repeat sample selection process r times to select the sample size of $n=rm$ under RSS. Therefore, the RSS sample is $y_{[i:m]j \ \ i=1,2,...,m, \ \ j=1,2,...,r}$, where m is the set size and r is the cycle size.

The sample mean under RSS can therefore be written as follows:

$$\bar{y}_{RSS} = \frac{1}{rm}\sum_{i=1}^{m}\sum_{j=1}^{r} y_{[i:m]j}, \tag{6.17}$$

It can be proved that $E(\bar{y}_{RSS}) = \mu_Y$ (see David and Nagaraja, 2003). Moreover, the variance of $\bar{Y}_{RSS}$ can be obtained (David and Nagaraja, 2003) as follows

$$Var(\bar{y}_{RSS}) = \frac{1}{m^2 r}\left[m\sigma_Y^2 + \frac{\rho^2\sigma_Y^2}{\sigma_X^2}\sum_{i=1}^{m}\sigma_{X(i:m)}^2 \right], \tag{6.18}$$

where σ_Y^2 is the population variance of Y, ρ is the correlation between X and Y, σ_X^2 is the population variance of X and $\sigma_{X_{(i:m)}}^2$ is the variance of the i^{th} order statistics of X.

6.9 MRSS sample selection procedure

The procedure for selecting a median ranked set sample (MRSS) is similar to that used to select a RSS, the only difference being in step (d). Here instead of selecting the smallest ranked unit from the first set, we select the median observation for that set. This procedure is then repeated for all sets. That is, we select a median in each set instead of diagonal units (Table 6.2). This algorithm is displayed in Table 6.3.

Table 6.3. Summary of steps d to f of MRSS.

Sample→ Set ↓	1	2	⋯	median	⋯	m
1	$y_{[1:m]1}$	$y_{[2:m]1}$	⋯	$y_{\left[\frac{m}{2}:m\right]1}$	⋯	$y_{[m:m]1}$
2	$y_{[1:m]2}$	$y_{[2:m]2}$	⋯	$y_{\left[\frac{m}{2}:m\right]2}$	⋯	$y_{[m:m]2}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	$y_{[1:m]k}$	$y_{[2:m]k}$	⋯	$y_{\left[\frac{m}{2}:m\right]k}$	⋯	$y_{[m:m]k}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
m	$y_{[1:m]m}$	$y_{[2:m]m}$	⋯	$y_{\left[\frac{m}{2}:m\right]m}$	⋯	$y_{[m:m]m}$

Here $x_{\left(\frac{m}{2}:i\right)k}$ $i=1,2,...,m$, $k=1,2,...,m$ is the sample from the variable X under MRSS. Therefore, the first set of the MRSS sample is the median units from Table 6.3. That is

$y_{\left[\frac{m}{2}:m\right]k}$, $k=1,2,...,m$, where m is the set and sample size in each set.

We repeat the sample selection process r times to select the sample size of $n=rm$ under MRSS. The m index repeats for all cycles, so for simplicity, we can drop the m index and reindex the selected sample under MRSS as follows:

$Y_{[i:m/2]j}$ $i=1,2,...,m$, $j=1,2,...,r$,

where m is the set size and r is the cycle size. Note that, for simplicity we assume that set size m is an odd number. The median ranked set sample estimator of μ_Y is then the corresponding sample mean

$$\bar{y}_{MRSS} = \frac{1}{rm}\sum_{i=1}^{m}\sum_{j=1}^{r} y_{[i:m/2]j}, \tag{6.19}$$

The sample mean $\bar{y}_{MRSS}$ is an unbiased estimator of μ_Y if the underlying distribution is symmetric otherwise the sample mean is a biased estimator (Muttlak, 1998). The variance of sample mean under MRSS (Muttlak, 1998) is

$$Var(\bar{y}_{MRSS}) = \frac{1}{m^2 r}\left[m\sigma_Y^2 + \frac{\rho^2\sigma_Y^2}{\sigma_Y^2}\sum_{i=1}^{m}\sigma_{X(i:m/2)}^2 \right], \tag{6.20}$$

where σ_Y^2 is the population variance of Y, ρ is the correlation between X and Y, σ_X^2 is the population variance of X variates, and $\sigma_{X_{(i:m/2)}}^2$ is the variance of $\left(\frac{m}{2}\right)^{th}$ order statistic of X.

6.10 Relative precision of RSS and MRSS with respect to SRS

Clearly, the benefit of using concomitant variables will depend on the correlation between the variable of interest Y and the concomitant variable X. If Y and X are independent, the estimators $\bar{y}_{RSS}$ and $\bar{y}_{MRSS}$ will be equivalent in efficiency to the sample mean under SRS (assuming the same sample size). To compare the two estimators $\bar{y}_{RSS}$ and $\bar{y}_{MRSS}$ with respect to $\bar{y}_{SRS}$, a linear relationship between X and Y is assumed. The relative precision of $\bar{y}_{RSS}$ with respect to $\bar{y}_{SRS}$ can then be defined as

$$RP_{RSS} = \frac{Var(\bar{y}_{SRS})}{Var(\bar{y}_{RSS})} = \frac{\sigma_Y^2 / rm}{\frac{1}{m^2 r}\left[m\sigma_Y^2 + \frac{\rho^2\sigma_Y^2}{\sigma_X^2}\sum_{i=1}^{m}\sigma_{X(i:m)}^2 \right]}.$$

This can be simplified to

$$RP_{RSS} = \frac{1}{\left(1-\rho^2\right)+\frac{\rho^2}{m}\sum_{i=1}^{m}\sigma^2_{z(i:m)}} \tag{6.21}$$

where $\sigma^2_{z(i:m)}$ is the variance of the i^{th} order statistic from a standard normal distribution. Similarly, the relative precision of $\bar{y}_{MRSS}$ with respect to $\bar{y}_{SRS}$ is

$$RP_{MRSS} = \frac{Var\left(\bar{y}_{SRS}\right)}{Var\left(\bar{y}_{MRSS}\right)} = \frac{\sigma^2_Y / rm}{\frac{1}{m^2 r}\left[m\sigma^2_Y + \frac{\rho^2\sigma^2_Y}{\sigma^2_X}\sum_{i=1}^{m}\sigma^2_{X(i:m/2)}\right]}.$$

This can be simplified to

$$RP_{MRSS} = \frac{1}{\left(1-\rho^2\right)+\frac{\rho^2}{m}\sum_{i=1}^{m}\sigma^2_{z(i:m/2)}} \tag{6.22}$$

where $\sigma^2_{z(i:m/2)}$ is the variance of the i^{th} median statistic from a standard normal distribution. The relative precision of RSS and MRSS is compared in a simulation study in the next section.

6.11 A Simulation comparison of OSY, RSS and MRSS

6.11.1 Simulation study

A simulation study was carried out in order to assess the practical performances of the estimators of a population mean under OSY, RSS, and MRSS, relative to the performance of the usual SRS estimator. The simulation was a model-based simulation. In every iteration, a population was generated according to the model and a sample was selected.

The parameter of interest was the population mean of Y in all cases. Two kinds of simulation scenarios were considered. In the first the population was bivariate normal data with $N=6000$ and with five different correlation coefficients and a sample size of 60. The second was the same as the first, the only difference being the sample size, which in this case was 600. The five different correlation coefficients between the

target variable Y and its concomitant X were 0.07, 0.37, 0.50, 0.75 and 0.92. The number of sets and cycles for RSS and MRSS in the first scenario were 6 and 10 and for the second scenario were 60 and 100, respectively. The population means of Y and X were 9 and 3. Finally, the number of Monte Carlo simulation was 6000. Table 6.4 shows the key characteristics of the simulation study

Table 6.4 Key characteristics of the simulation study.

1- (X,Y) IS A BIVARIATE NORMAL DISTRIBUTION
2- N = 6000, n = 60 for first scenario and 600 for second scenario
3- $\mu_X = 3$, $\mu_Y = 9$
4- No. of sets = (6,60) , no. of cycles = (10,100) (for RSS and MRSS)
5- No. of Monte Carlo simulation = 6000

Tables 6.5 and 6.6 show the empirical means and variances of the sample means, the variance of sample means by formula (6.13), relative bias for the variance of sample means, and coverage percentages generated by ordered systematic sampling (OSY) under different correlation coefficients. Tables 6.7 and 6.8 then show the relative precisions under both scenarios for OSY, RSS, and MRSS for the different correlation coefficients.

Table 6.5: Scenario 1, statistical properties of the sample mean under OSY

n=60	Pop. Mean	Empirical Means	Empirical Variance	Formula Variance	Relative Bias % of variance estimator	Coverage %
ρ=0.07	9.00	9.0063	0.1187580	0.1143326	-3.73	94.24
ρ=0.37	9.00	9.0038	0.2060074	0.2000961	-2.86	94.18
ρ=0.50	9.00	8.99997	0.1195291	0.1206034	-0.89	94.68
ρ=0.75	9.00	8.998	0.05221673	0.05039186	-3.49	94.18
ρ=0.92	9.00	8.999	0.02341593	0.02195380	-6.24	94.14

It can be seen from Tables 6.5 and 6.6 that the OSY sample means are unbiased estimators for population means, and this unbiasedness is unchanged when we change the correlation between Y and X. We also show two different variances for the proposed sample means. The first is the Monte Carlo or empirical variance, and the second is the average of the sample values of (6.13), which is denoted by formula variance. To compare these two variances, we also show the relative bias of (6.13). As seen from Table 6.5, the calculated variance using (6.13) slightly underestimates the empirical variance, but the amount of underestimation is small enough to be ignored in practical applications. Finally, the last column shows the coverage of the 95% confidence intervals generated using (6.13), and, as can be seen this coverage is almost 95 percent in all cases.

Table 6.6: Scenario 2, statistical properties of the sample mean under OSY

n= 600	Pop. Mean	Empirical Means	Empirical Variance	Formula Variance	Relative Bias % of variance estimation	Coverage %
ρ=0.07	9.00	8.999	0.01151	0.01160	0.79	94.86
ρ=0.37	9.00	8.994	0.02062	0.02010	-2.52	94.66
ρ=0.50	9.00	9.002	0.01269	0.01201	-5.35	94.20
ρ=0.75	9.00	8.999	0.00528	0.00476	-9.77	94.00
ρ=0.92	9.00	8.999	0.00282	0.00251	-10.71	90.10

By increasing the sample size from 60 to 600, the same pattern is seen in Table 6.6 but with some decrease in coverage and increase in relative bias.

Table 6.7: Relative precisions of OSY, RSS, and MRSS with respect to SRS in the first scenario (n=60).

n=60	RP(RSS)	RP(MRSS)	RP(OSY)
ρ=0.07	0.98	1.01	1.01
ρ=0.37	1.09	1.13	1.15
ρ=0.50	1.22	1.25	1.40
ρ=0.75	1.63	1.79	2.27
ρ=0.92	2.39	2.70	4.99

Finally, from Tables 6.7 and 6.8, it is clear that substantial gains may be achieved by using the proposed OSY procedure. As seen from these results, the relative precision of OSY is more than that of RSS and MRSS. Furthermore, this precision increases as the correlation between Y and X increases, and is evident for both sample sizes considered in the simulation study.

Table 6.8: Relative precisions of OSY, RSS, and MRSS with respect to SRS in the second scenario (n=600).

n=600	RP(RSS)	RP(MRSS)	RP(OSY)
ρ=0.07	0.95	0.88	1.03
ρ=0.37	1.07	1.05	1.17
ρ=0.50	1.16	1.18	1.34
ρ=0.75	1.53	1.51	2.18
ρ=0.92	1.98	2.29	4.36

Chapter 7

Summary and Conclusions

7.1 Introduction

In this chapter we summarise the main conclusions of the work undertaken in this thesis and highlight the main contributions of this research. Further research areas and suggestions are given. This research is done within the framework defined by application of concomitants of order statistics in survey sampling. The core of the research consists in investigating two main topics:

1) Nearest Neighbour Imputation based on Concomitants.
2) Variance Estimation of Sample Mean based on Ordered Sampling using Concomitants.

Main conclusions and contributions associated with each of this two topics are now presented in detail.

7.2 Summary and conclusions for part one

The problem of missing data and using available information to impute missing values has been addressed in this thesis. In particular, the focus of the thesis was to develop an imputation method by using multivariate data ordering methods and concomitants of order statistics. The imputation method also used nearest neighbour imputation theory. Simply, the idea was that when multivariate data (with at least two variables) contains missing values, it is ordered according to an artificially calculated

variable. This variable is a function of auxiliary variables and can for example be a first component of principal component analysis, or sequential taxonomy coefficients. After ordering the data using the artificial variable, the concomitant orders of the missing values are specified and then the k nearest neighbours for the missing value in the ordered data defined by $k/2$ of available values above and below the missing values are used to define the imputed value.

In this thesis, three ordering methods were examined, based on (1) the first component of principal component analysis, (2) sequential taxonomy coefficients, (3), the Euclidian distance of each case from the centre of data.

Given this ordering, missing values can then be imputed under SSTI or DSTI. The statistical properties of these imputation methods are then under the assumption that the ordering variable and the variable with missing values have a linear relationship. Under this condition, mathematical properties of sequential taxonomy (ST) were established in Chapter 4 and statistical properties of SSTI and DSTI were investigated mathematically and by a simulation study in Chapter 5.

In this study, six different imputation methods were discussed and their performance evaluated with respect to estimation accuracy, preservation of marginal distributions and prediction of individual values. The results of the study showed that under MCAR and MAR, the methods that performed best with respect to preservation of marginal distributions were stratified hot deck imputation (SHDI), double order sequential taxonomy imputation (DSTI), and single order sequential taxonomy imputation (SSTI) in that order. From the mean estimation point of view, regression based imputation methods (RI) and SHDI resulted in better estimation than any of the ordering based methods. However, from the variance estimation point of view, SHDI, DSTI, and SSTI estimated the population variance better than the other imputation methods. Furthermore, when correlations between real and imputed values under MCAR and MAR are computed, it can be seen that RI gives the highest correlations with DSTI in second place. SHDI has the lowest correlations under both assumptions. In other words, RI and DSTI predict individual values better than the other imputation methods. Although the evaluation of imputation methods depends on the aims of the

study, our results provide evidence that with respect to all aspects of evaluation DSTI seems more stable than other imputation methods we investigated.

7.3 Summary and conclusions for part two

Ranked-set sampling (RSS), median ranked-set sampling (MRSS), and ordered systematic sampling (OSY) are three sampling methods based on the properties of direct and indirect order statistics. This research develops a new variance estimator for the sample mean under OSY based on these properties. In general, systematic sampling can be regarded as form of cluster sampling where only one cluster is selected, thus making it impossible to unbiasedly estimate the sampling variance without any assumption about the population. In practice, the variance of simple random sampling is therefore used instead, leading to overestimates of the variance of sample mean. Here we propose a variance estimator for sample mean under OSY based on a linear relationship between the concomitant variable X (always available for systematic sampling) and the variable of interest Y assuming all values of X are available in the population. This estimator is then compared theoretically and by simulation study with simple random sampling, ranked-set sampling, and median ranked-set sampling. Other statistical properties such as the relative precision of OSY compared with SRS, the coverage of sample mean and the relative bias of variance estimation are derived in this thesis.

In our study, RSS, MRSS, and OSY provided unbiased estimations for the population means, with the variance of the sample mean under OSY smaller than its variance under SRS, RSS, and MRSS. A major advantage of the new variance estimator is its applicability. In practice, this variance estimator can be used under standard systematic sampling and it needs assume only a linear relationship between the concomitant variable and the variable of interest. Under this assumption, the new variance estimator appears to be more efficient than previously used variance estimators.

7.4 Further research for part one

The research described in this thesis has shown that nearest neighbour imputation based on concomitants of order statistics such as SSTI, DSTI, and PCAI gives reasonable results when compared to other widely used imputation methods. These methods were developed under the assumption of a linear relationship between the variable with missing values and the concomitant variable. Furthermore, it was assumed that the distribution of the ordering was normal. This study has been developed under a bivariate normal distribution assumption of the variable with missing values and its concomitant because of the existence of linear relationship between both variables. However, the main assumption was the linearity not the normality. Further research is needed to resolve a number of issues connected to the variance estimation of a missing value. The variance estimator is a complicated formula and needs to made simpler and applicable in practice. Multiple concomitants were used to develop DSTI in the simplest form and this needs to be extended. In addition, to use multiple concomitants we defined weights according to minimize the variances of each imputation set, therefore further research is needed to investigate other weighting methods and to compare the efficiency of these weighting procedures. The proposed imputation methods are in the first stage of using order statistics theory to develop an imputation method; hence, the following directions are suggested to generalize SSTI and DSTI.

1) The relationship of the concomitant and the variable of interest with missing values can be assumed nonlinear.
2) Multiple concomitant order statistics can be used to improve the quality of SSTI and DSTI.
3) The distribution of the concomitant variable and the variable with missing values is assumed to be bivariate normal. Therefore, in future work, this imputation method can be extended to non-normal distributions.
4) The relationship between the concomitant variable and the variable with missing value can be extended to non-parametric relationships.
5) Comparison between the proposed methods and other imputation methods such as multiple imputation and neural network based imputation methods.

7.5 Further research for part two

This study represents just the first stage on the research of the use of concomitants order statistics for variance estimation under OSY. Further research should be done in order to assess more aspects. The further work can be divided into two different aspects: existence of nonlinear or nonparametric relationship between the concomitant variable and the variable of interest and having a balanced or an unbalanced sampling scheme. The proposed variance estimator for a sample mean under OSY was for a balanced sampling scheme. Further work is needed to investigate the properties this variance estimator under unbalanced sampling procedures. In addition, our new variance estimator was developed assuming of a linear relationship between the concomitant variable and the variable of interest. Therefore, further research should be done in order to assess more aspects of the variance estimator under nonlinear relationship. In summary, the following directions can be suggested as extension of this method.

1- Generalisation of this method to systematic sampling in two dimensions.
2- Comparison of the new variance estimator with other sampling techniques.
3- Generalizing to a broad class of super population model.
4- Generalisation of this method to using multi concomitants
5- Generalisation of this method to other aspects of systematic sampling such as having periodic or circle trend in the population.

Appendix 1: Ordering Multivariate Data

1.1 Introduction

In spite of the lack of a strong theory for data ordering, especially for multivariate data, we can find many examples of ordering in every day life, such as ordering cities or countries according to development, ranking of universities based on research activities, and so on. Ordering is a special case of dimension reduction of data into one dimension, and it has a large literature in statistics. Therefore, ordering and dimension reduction methods are related. Generally, high-dimensional datasets present many mathematical challenges as well as some opportunities, and are bound to give rise to new theoretical developments. This section briefly reviews ordering methods in five categories: marginal ordering, reduced (aggregate) ordering, partial ordering, conditional (sequential) ordering and sequential taxonomy. In addition, single components of dimension reduction techniques can be used as an ordered vector, for example in principal component analysis, the first component, or, in factor analysis, the first factor, and so on. Therefore, the dimension reduction technique can be seen and used as an ordering technique. Dimension reduction techniques can be found in multivariate analysis books, some popular methods are principal component analysis, projection pursuit, and projection pursuit regression, principal curves, and methods based on topologically continuous maps, such as Kohonen's maps, or generalized topographic mapping. Neural network implementations for several of these techniques are also reviewed, such as the projection pursuit-learning network and the theory of Bienenstock, Cooper and Munro (BCM) with an objective function.

1.2 Ordering techniques

Ordering techniques are divided into four groups: marginal ordering, reduced (aggregate) ordering, partial ordering and conditional (sequential) ordering. The first two ordering methods are reviewed in this section.

1.3 Marginal ordering (M ordering)

In M-ordering, the multivariate data is ordered according to a marginal distribution. Sometimes, for inference about M-ordering, certain order features in the marginal samples may be considered in combination (as in global, or component-based, concepts of median, range, extremes, etc). Another application of M-ordering is data transformation. In other words, M-ordering is sometimes used in the ordering of particular combinations (projections) of component values or of radial distances or angular deviations from some fixed point or direction. Then again, the sample points may each be initially reduced to a single value by some metric. However, most examples of ordering after initial transformation of data are best considered under the next heading of reduced ordering since their intention is not to represent marginal behaviour but to summarily express overall characteristics for the multivariate dataset

1.4 Reduced (aggregate) ordering (R-ordering)

In this method, each multivariate observation is reduced from p dimension to one dimension by an appropriate metric that is the component of sample values. This method is popular in distance based ordering, and usually data is reduced to a single value by a quadratic function:

$$(X-\alpha)'\Gamma^{-1}(X-\alpha)$$

for some convenient choice of α and Γ, where α is the origin, the mean, the extreme values or the set of marginal medians (sample or population) and Γ is

the identity matrix, the variance and covariance matrix of the sample or population, or perhaps the diagonal matrix component variances (see Wilk and Gnanadesikan 1964). Setting $\Gamma = I$ corresponds to the Euclidian distance of the sample points from some "centre" α. The raw Euclidian distance disregards second order moment structure (and location) and this is a big disadvantage of using $\Gamma = I$. If we knew the distribution of x, there would be some appeal in ordering in relation to probability concentration contours. For multivariate normal distribution, the above formula changes to:

$$(x-\mu)'\Sigma^{-1}(x-\mu),$$

where μ is the mean vector and Σ is the variance covariance or correlation matrix. This metric is very sensitive to the correlation of variables. For example, suppose we have two variables and the correlation between them is zero. In this case, the diagram of two variables is circular, but for a correlation of 0.5, the diagram is elliptical Barnett (1977). By increasing the correlation between the variables, the diagram between two variables becomes linear. In multivariate normal data, standardizing data and then using the above formula corresponds to use of Mahalanobis distance. Ordering data based on R-ordering is one of the main principles of sequential taxonomy method (see Chapter 4). For comprehensive discussion about ordering methods, please refer to Barnett (1977).

REFERENCES

Anderberg, M. R. (1973), "Cluster analysis for applications", *Academic Press*, New York.

Bailar, B. A. and Bailar, J. C. III (1978) "Comparison of two procedures for imputing missing survey values", *American Statistical Association Proceedings of the selection on Survey research Methods*, pp. 462-467.

Bailar, B. A., Bailar, J. C. III (1979), "Comparison of the biases of the Hot Deck imputation procedure with an equal weights imputation procedure", *National Academy of Science, Proceedings of the Symposium on Incomplete Data.*

Bailar, B. A., Baily, L. and Corby, C. (1978) " A comparison of some adjustment and weighting procedures for survey data. In N. Krishnam Namboodiri(ed), *Survey Sampling and Measurement*, pp. 175-198. New York: Academic Press.

Barnett, V. (1977), "The ordering of Multivariate Data", *Journal of the Royal statistics Society. Series A*, Volume 139, Issue 3, 318-355

Bassi, F. and Luigi, F. (1997), "Estimators of Nonsampling Errors in Interview-Reinterview Supervised Surveys with Interpenetrated Assignments", *Survey Methods and Process Quality*. Eds. Lars Lyberg, et al. New York, NY: Wiley and Sons, Inc.

Berger Yves G. (2002), " A variance estimator for systematic sampling from deliberately ordered population", *Communications in Statistics-Theory and Methods*, Vol. 34, No. 7.

Bhattacharya, P. K. (1974), "Convergence of sample paths of normalized sums of induced order statistics", *Ann. Statist.* 2, 1034-9.

Bhattacharya P. K. (1984), "Induced order statistics: Theory and applications", in Handbook of Statistics, Vol. 4, Nonparametric Methods (P. R. Krishnaiah and P. K. Sen, Eds.),pp. 383-403, North Holland, Amsterdam, 1984.

Bishop, Christopher M. (1996), "Neural Networks for Pattern Recognition".

Boland P. J. M., Hollander, K. Joag-Dev, and S. Kochar, "Bivariate dependence properties of order statistics", *J. Multivariate Anal.* 56 (1996), 75-89.

Castillo, E. (1988), "Extreme Value theory in engineering" , *Academic Press Inc.*

Chambers, R. (2001), "Evaluation Criteria for statistical editing and imputation", internal working paper , *EUREDIT project. www.cs.york..ac.uk/euredit/.*

Chaudhuri Arijit and Stenger Horst (1992), "Survey Sampling", *New York: marcel dekker, Inc.*

Chen, J. and Shao, J. (2000), "Nearest neighbourhood imputation for survey data", *Journal of Official statistics*, vol.16, 113-131.

Cochran, W. G. (1978), "Sampling techniques". *New York: Wiley*

David H. A. (1981), "Order Statistics". *New York: Wiley*

David H. A. and Nagaraja H. N. (2003), "Order Statistics". *New York: Wiley*

David, H. A. (1973), "Concomitants of order statistics", *Bull. Internat. Statist. Inst.* 45, 295-300.

David, H. A. and Galambos J. (1974), "The asymptotic theory of concomitants of order statistics", *J. Appl. Probab.* 11, 762-770.

David, H. A. and Nagaraja, H. N. (1998), "Concomitants of order statistics" in Handbook of Statistics, Vol. 16, Order Statistics: Theory and Methods (N. Balakrishnan and C. R. Rao, Eds.), pp. 487-513, Elsevier, New York, 1998.

David, H. A., O'Connel M. J., Yang S. S. (1977), "Distribution and Expected value of the Rank of a Concomitants of an Order Statistic", *The Annals of Statistics*, Volume 5, Issue 1, 216-223

Dell, T. R. and Clutter, J. L. (1972). "Ranked set sampling theory with order statistics background", *Biometrics*, 28, 545-555.

Dempster, A.P., Nan M. Laird, and Donald B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society* 39B, 1-22.

Do, K. A. and Hall, P. (1991), "Distribution theory using concomitants of order statistics with application to Monte Carlo simulation for the bootstrap., *Journal of Royal Statistical Society B* 54, 595-607

Fellegi, I. and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation." *Journal of the American Statistical Association*, 71, 353: 17–35.

Ford, B. (1983), "An Overview of Hot-Deck Procedures", *Incomplete Data in Sample Surveys, Vol. 2*. Burlington, MA: Academic Press, Inc.

Government Statistical Service Methods Committee (1996). Report of the Task Force on Imputation.

Gower, J. C. and Legendre, P. (1986), "Metric and Euclidean properties of dissimilarity coefficients", *Journal of Classification* 3: 5-48.

Greenlees, J., Reece, W., and Zeischang, K. (1982), "Imputation of Missing Values When the Probability of Response Depends Upon the Variable Being Imputed." *Journal of the American Statistical Association* 77: 251–261.

Gross, A. L. (1973), " Prediction in future samples studied in terms of the gain from selection". *Psychometrica* 138, 151-171.

Jaccard, P. (1901), "Étude comparative de la distribuition florale dans une portion des Alpes et de Jura." *Bulletin de la Societé Voudoise des Sciences Naturelles*(37): 547-579.

Jagdish, K. Patel and Campell, B. Read (1996), " Handbook of the Normal Distribution", *New York: MARCEL DEKKER INC.*

Johnson N. L. and Kotz S. (1972), "Distributions in Statistics: Continuous Multivariate Distributions," *Wiley, New York.*

Johnson N. L., Kotz S. and Balakrishnan, N. (1994), "Continuous univariate distribution", Volume 1, Second Edition, *Wiley, New York.*

Kalton, G. and Kasprzyk, D. (1982), "Imputing for Missing Survey Responses." *Proceedings of the Section on Survey Research Methods.* American Statistical Association.

Kaufman, L. and Rousseeuw, P. J. (1990), "Finding groups in data", *Wiley*, New York.

Kim, S. H. and David, H. A. (1990) "On the dependence structure of order statistics and concomitants of order statistics", *J. Statist. Plan. Inf.* 24 (1990), 363-368.

Lessler, J. and Kalsbeek, W. (1992) "Nonsampling error in surveys", Wiley Series in Probability and Mathematical Statistics. John Wiley.

Little, R. J. A. (1995), "Modelling the drop-out mechanism in repeated-measures studies", *JASA*, 90, 1112-1121.

Little, Roderick J. A. and Rubin, Donald B. (1987), "Statistical analysis with missing data". *New York: Wiley*

Little, R.J. and Rubin, D.B. (2003). "Statistical analysis with missing data", 2nd Edition. Wiley Series in Probability and Statistics. New Jersey: Wiley.

McIntyre, G. A. (1952). "A method of unbiased selective sampling, using ranked sets", *Australian J. Agricultural Research*, 3, 385-390.

Muttlak, H. A. (1997). "Median ranked set sampling", *Journal of Applied Statistics Sciences*, 6.

Muttlak, H. A. (1998). "Median ranked set sampling with concomitants variables and a comparison with ranked set sampling and regression estimators", *Environmetrics,* 9 255-267.

Murtly and Hossain (2003), "Multivariate nearest neighbourhood method of imputation", *Statistics in Transition*, Vol. 6, No. 1, pp. 55–66.

Nagaraja H. N. and David H. A., (1994), "Distribution of the maximum of concomitants of selected order statistics", *The Annals of Statistics*, 1994, Vol. 22, Issue 1, 478-494.

O'Connell M. J. (1974), "Theory and Applications of Concomitants of Order Statistics", Ph.D. dissertation, Iowa State University, Microfilm 75-10496.

O'Connel, M. J. and David, H. A. (1976), "Order statisticsand their concomitants in some double sampling situations", *Essays in Probability and Statistics.* Shinko Tsusho, Tokyo, 451-466.

Patil, G. P., Sinha, A. K. and Taillie, C. (1992), "Ranked Set Sampling: An Annotated Bibliography", Technical Report 91-1201, *Centre for Statistical Ecology and Environmental Statistics*, Pennsylvania State University.

Patil, G. P., Sinha, A. K. and Taillie, C. (1993), "Relative precision of ranked set sampling: a comparison with the regression estimator", *Environmentrics*, 4, 399-412.

Ripley, B. D. (1999), "Modern Applied Statistics with S-PLUS", Third Edition
Rogers, J. S. and Tanimoto, T. T. (1960), "A computer program for classifying plants." *Science* (132): 1115-1118.

Rubin, Donald B. (1987), "Multiple imputation for nonresponse in surveys", pp 1.

Rubin, D. B. (1977), "Formalising subjective notions about the effect of nonrespondents in sample surveys", *Journal of the American Statistical Association* 72,538-543.

Rubin, D. B. (1978), "Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse". *Proceedings of the survey research methods section of the American Statistical Association*, 20-34.

Rubin, Donald B. (1976), "Inference and Missing Data," *Biometrika* 63, 581-592.

Rubin, D. B. (1987), "Multiple Imputation for Nonresponse in Surveys". *New York: Wiley*

Sarndal Carl-Erik, Swensson B. and Wretman J. (1997), " Model assisted Survey Sampling", Fourth Edition *Springer Series in Statistics*

Sandstrom, A. (1987), "Asymptotic Normality of Linear Functions of Concomitants of Order Statistics" , *Metrika*, volume 34; pp. 129 - 142

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

Sen, P. K. (1976), "A note on invariance principles for induced order statistics", *Ann. Probab.* Vol. 4 (1976), 474_479.

Shaked M. (1977), "A family of concepts of dependence for bivariate distribution", *J. Amer. Statist. Assoc.* Vol. 72, 642-650.

Shapiro, S. S. and Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)", *Biometrika*, 52, 3 and 4, pages 591-611.

Sokal, R. R. and P. H. Sneath P. H. (1963), "Principles of numeric taxonomy". San Francisco, W.H. Freeman.

Stokes, S. L. (1977), "Ranked set sampling with concomitant variables", *Communications in Statistics*, A6, 1207-1211.

Stuart, A. (1955). "A test for homogeneity of the marginal distributions in a two-way classification". *Biometrika* 42, pg. 412.

Svein Nordbotten (1996), "Neural Network Imputation Applied to the Norwegian 1990 Population Census Data". *Journal of Official Statistics*, Vol. 12, No. 4, 1996, pp. 385-401.

Sukhatme, P. V. and Sukhatme, B. V. (1970), "Sampling Theory of Surveys with Applications". *Iowa State University Press*, Ames, Iowa.

Takahasi, K. and Wakimoto, K. (1968), "On unbiased estimates of the population mean based on the sample stratified by means of ordering", *Annals of the Institute of Statistical Mathematics*, 20, 1-31.

Wilk, M.B., & Gnanadesikan, R.(1964). "Graphical methods for internal comparisons in multiresponse experiments", *Ann. Math. Stat.*, 35, 613-631.

Yang, S. S. (1977), "General distribution theory of the concomitants of order statistics", *Ann. Statist.* 5 , 996-1002.

Yang, S. S. (1981), "Linear combination of concomitants of order statistics with application to testing and estimaton", *Ann. Inst. Statist Math.* 33 , 463-70.

Yeo, W. B. and David, H. A. (1984), "Selection through an associated characteristic with applications to random effects model". *Journal of American Association.* 79, 399-405.