DACSEIS
research paper series
No. 5

The Impact of Multiple Imputation for DACSEIS

Susanne Rässler

2004

IST–2000–26057–DACSEIS

The DACSEIS research project is financially supported within the IST programme of the European Commission. Research activities take place in close collaboration with Eurostat.

`http://europa.eu.int/comm/eurostat/research/`

`http://www.cordis.lu/ist/`

`http://www.dacseis.de/`

The Impact of Multiple Imputation for DACSEIS

Susanne Rässler

University of Erlangen-Nürnberg
Department of Statistics and Econometrics
Lange Gasse 20, D-90403 Nürnberg, Germany
email: susanne.raessler@wiso.uni-erlangen.de

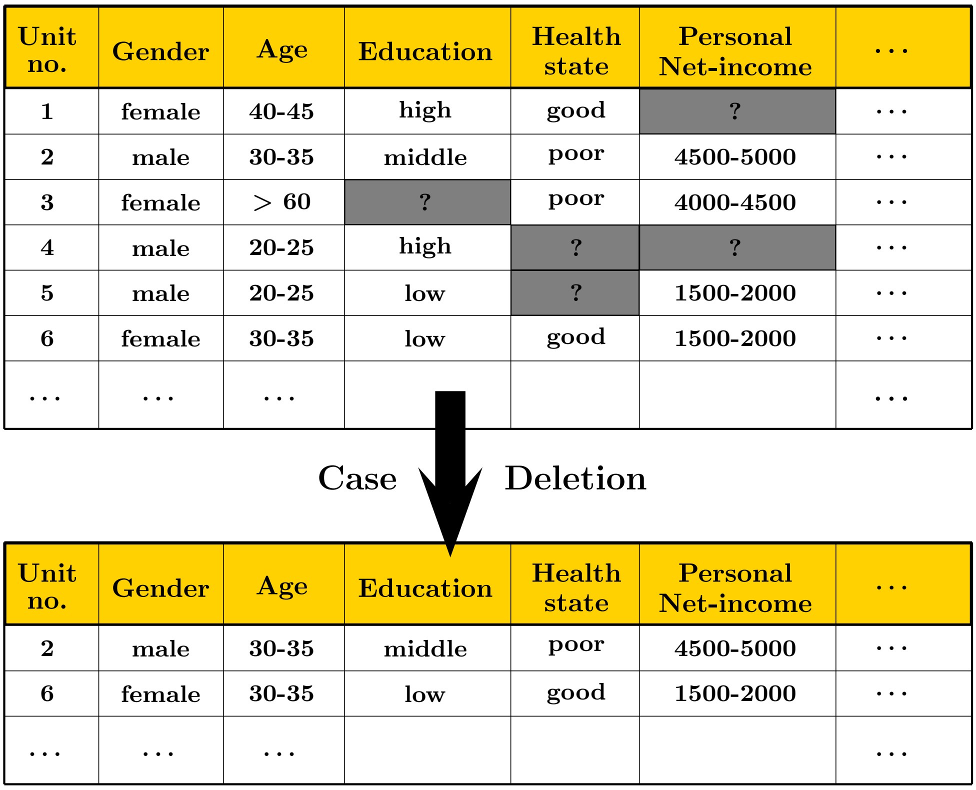
Abstract: This paper is designed to provide an extensive introduction to the principles of multiple imputation and to give some general recommendations of using multiple imputation techniques in the DACSEIS universes. The definition of an ignorable missingness mechanism is explained, and the concept of the observed-data likelihood is discussed. To introduce the multiple imputation principle a short introduction of Bayesian statistics is provided. A small simulation study is performed comparing different approaches to illuminate the advantages and disadvantages of different imputation techniques. Finally, an overview about recently available multiple imputation software is given and violations of the assumptions made are addressed.

Keywords: Complex survey, missing data, ignorable missingness, observed-data likelihood, observed-data posterior, Monte-Carlo techniques

1 Introduction

Often empirical researchers are confronted with missing values in their data sets. As the phenomenon is often not seen as a possible threat to the validity of the research, the most common approach to this problem is simply to deny it. However, a closer look to the data often reveals 5% to 20% of missing values in a few variables, reducing the complete data for any multivariate analysis considerably, see Figure 1.

Moreover, often these blind spots were not dropped randomly all over the responses. We find special socio-economic groups or minorities disproportionately struck by missing values. Even worse, if the missingness depends on the variable of interest itself, like it is common that the highest income appears to be unknown. The same happens when, e.g., populations with worst health conditions or high at risk refuse to be sampled. Finally, the quality of response deteriorates with long and boring questionnaires like they are common practice in media research. In all these cases, missing data can be a threat to the research and the remaining data are all but representative for the population of interest.



Unit no.	Gender	Age	Education	Health state	Personal Net-income	...
1	female	40-45	high	good	?	...
2	male	30-35	middle	poor	4500-5000	...
3	female	> 60	?	poor	4000-4500	...
4	male	20-25	high	?	?	...
5	male	20-25	low	?	1500-2000	...
6	female	30-35	low	good	1500-2000	...
...	...	...				...

Unit no.	Gender	Age	Education	Health state	Personal Net-income	...
2	male	30-35	middle	poor	4500-5000	...
6	female	30-35	low	good	1500-2000	...
...	...	...				...

Figure 1: Loss of information due to case deletion

Missing data are common in practice and usually complicate data analyses for scientific investigations. A rather general method for handling missing values in a data set is to impute, i.e., fill in one or more plausible values for each missing datum so that one or more completed data sets are created. Often it is easier to first impute for the missing values and then use a standard complete-data method of analysis than to develop special statistical techniques that allow the analysis of incomplete data directly.

Imputing a single value for each missing datum and then analyzing the completed data using standard techniques designed for complete data will usually result in standard error estimates that are too small, confidence intervals that undercover, and p-values that are too significant; this is true even if the modeling for imputation is carried out carefully. The usual single imputation strategies such as mean imputation, hot deck, or regression imputation typically result in confidence intervals and p-values that ignore the uncertainty due to the missing data, because the imputed data were treated as if they were fixed known values.[1]

[1] A discussion of advantages and disadvantages of single and multiple imputation procedures may be found by the interested reader in Marker et al. (2002) and Meng (2002). Approaches for obtaining frequency valid standard errors under single imputation procedures are discussed, e.g., by Lee et al. (2002) and Shao (2002).

Multiple imputation (MI), introduced by Rubin (1978) and discussed in detail in Rubin (1987), is an approach that retains the advantages of imputation while allowing the data analyst to make valid assessments of uncertainty. The concept of multiple imputation reflects uncertainty in the imputation of the missing values through resulting in theoretically wider confidence intervals and thus p-values suggesting less significance than single imputation would. MI is a Monte Carlo technique that replaces the missing values by $m > 1$ simulated versions, generated according to a probability distribution indicating how likely the true values are given the observed data, see Figure 2.

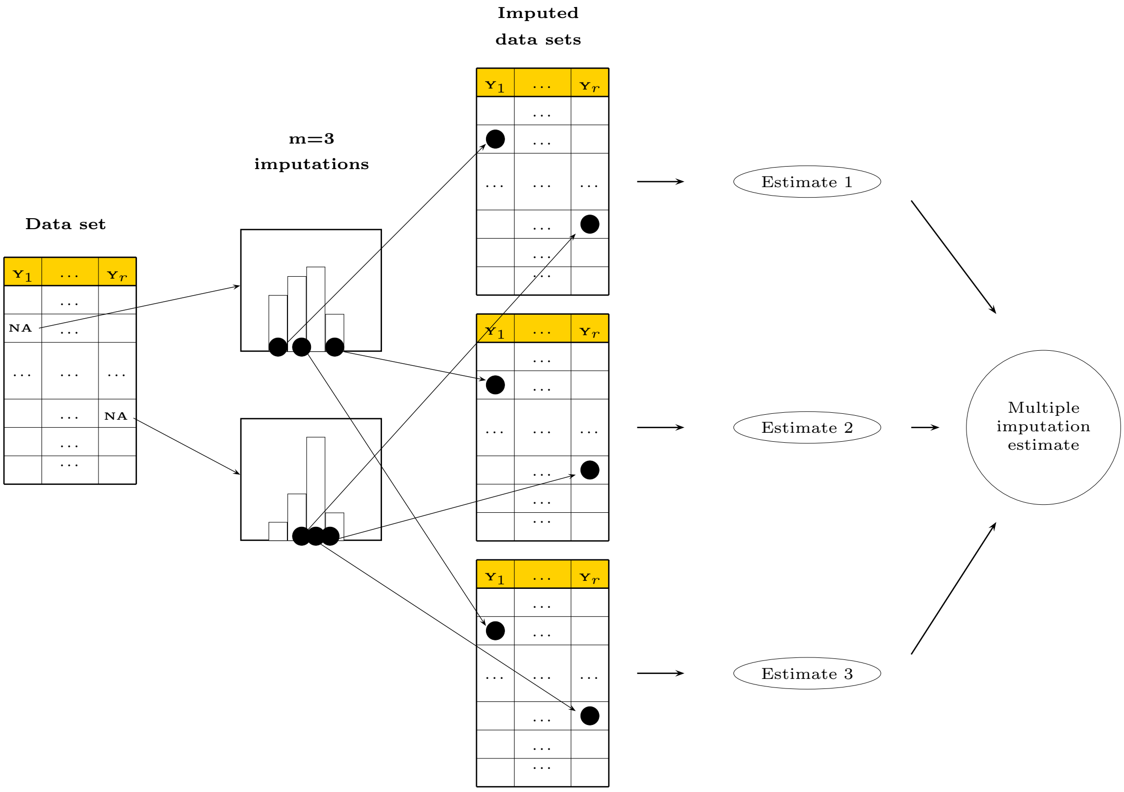


Figure 2: Multiple imputation

Typically m is small, e.g., $m = 5$, although with upcoming computational power m can be 10 or 20, in general, this depends on the amount of missingness and on the distribution of the parameter to be estimated, especially if analyst's model and imputer's model differ. Each of the imputed (and thus completed) data sets is first analyzed by standard methods; the results are then combined to produce estimates and confidence intervals that reflect the missing data uncertainty.

2 Missingness mechanisms

Following the terminology introduced by Rubin (1987) and Little and Rubin (1987, 2002) the missing-data mechanism can be classified according to the probability of response. The missing data are said to be as follows.

- MCAR – missing completely at random. In this case, the unobserved values form a random subsample of the sampled values. If, for instance, the probability that income is recorded is the same for all individuals, regardless of, e.g., their age or income itself, then the data are said to be MCAR.
- MAR – missing at random. In this case, the unobserved values form a random subsample of the sampled values within classes defined by the observed values. For example, if the probability that income is recorded varies according to the age of the respondent but does not vary according to the income of the respondent within an age group, then the data are MAR.
- MNAR – missing not at random. If the data are neither MCAR nor MAR, the mechanism is nonignorable. If the probability that income is recorded varies according to income itself, then the data are MNAR.

We will discuss these definitions more technically now. Let $\mathcal{U}$ be any population of interest, finite or not, and $u_i = (u_{i1}, u_{i2}, \dots, u_{ir})$ denote the value of a random vector $U = (U_1, U_2, \dots, U_r)$ for each unit $i \in \mathcal{U}$. Without loss of generality let $f_U(u_i; \theta)$ be the probability of drawing a certain unit i, $i \in \mathcal{U}$ with observations $u_i = (u_{i1}, u_{i2}, \dots, u_{ir})$ depending on the parameter $\theta \in \Omega_\theta$ which may be regarded as a scalar or vector. In the case of continuous random variables U, f_U may be taken as the density function instead of the probability function. To be more general, f_U may also describe a finite mixture of densities. Finally, let a random sample of n independently observed units from $\mathcal{U}$ be given with probability or, more generally, with density function $\prod_{i=1}^n f_U(u_i; \theta), \theta \in \Omega_\theta$.

Now denote the observed part of the random vector U by U_{obs}, and the missing part by U_{mis}, so that $U = (U_{obs}, U_{mis})$. The joint distribution of U_{obs} and U_{mis} is given by $f_U(u_i; \theta) = f_{U_{obs},U_{mis}}(u_{obs,i}, u_{mis,i}; \theta)$ for each unit $i \in \mathcal{U}$.

Furthermore, let R be an indicator variable being zero or one depending on whether the corresponding element of U is missing or observed; i.e.,

$$R_{ij} = \begin{cases} 1 & , \text{ if variable } U_j \text{ is observed on the unit } i, \\ 0 & , \text{ else,} \end{cases}$$

for all units $i \in \mathcal{U}$ and variables $U_j, j = 1, 2, \dots, r$. Generally a probability model for R with $f_R(r; \xi)$ is assumed, which depends on some unknown nuisance parameter $\xi \in \Omega_\xi$. Hence, the joint distribution of the response indicator R and the interesting variables U is given by

$$f_{U,R}(u, r; \theta, \xi) = f_U(u; \theta) f_{R|U}(r|u; \xi), \quad (\theta, \xi) \in \Omega_{\theta,\xi}.$$

The density or probability function describing the observed data of any unit $i \in \mathcal{U}$ and, thus, their likelihood may actually be written

$$\begin{aligned} L(\theta, \xi; u_{obs}, r) &= f_{U_{obs},R}(u_{obs}, r; \theta, \xi) \\ &= \int f_{U_{obs},U_{mis}}(u_{obs}, u_{mis}; \theta) f_{R|U_{obs},U_{mis}}(r|u_{obs}, u_{mis}; \xi) du_{mis}, \end{aligned} \qquad (1)$$

with $(\theta, \xi) \in \Omega_{\theta,\xi}$. For simplicity we want the integral to be understood as the sum for discrete distributions. To ease reading, we usually refer to f_U as the density function of U hereinafter.

Unit no.	$\boldsymbol{U_1}$	$\boldsymbol{U_2}$	$\boldsymbol{R_1}$	$\boldsymbol{R_2}$
1			**1**	**1**
			1	**1**
			1	**1**
$\boldsymbol{r}$			**1**	**1**
$\boldsymbol{r}$**+1**		**missing**	**1**	**0**
			1	**0**
$\boldsymbol{n}$			**1**	**0**

Figure 3: Missing data example

Consider, e.g., an iid sample with two random variables U_1 and U_2 observed from n (U_1) or $r < n$ (U_2) units, respectively. Then the observed-data likelihood according to the data presented in Figure 3, is

$$L(\theta, \xi; u_{obs}, r) = \prod_{i=1}^{r} f_{U_1,U_2}(u_{i1}, u_{i2}; \theta) f_{R|U_1,U_2}(r_i|u_{i1}, u_{i2}; \xi)$$
$$\times \prod_{i=r+1}^{n} \int f_{U_1,U_2}(u_{i1}, u_{i2}; \theta) f_{R|U_1,U_2}(r_i|u_{i1}, u_{i2}; \xi) du_{i2}$$

which is also called the full likelihood by Little and Rubin (2002), p. 119. Notice that the integration of the second term is not easily done without further assumptions.

Now the assumptions concerning the missing-data mechanisms can be explained in more detail.

- First of all it is assumed that θ and ξ are "distinct"; i.e., knowing θ will provide no information about ξ and vice versa (see Schafer (1997), p. 11). Then the joint parameter space $\Omega_{\theta,\xi}$ is the product of the parameter space of θ and the parameter space of ξ, i.e., $\Omega_{\theta,\xi} = \Omega_\theta \times \Omega_\xi$. Thus, the conditional distribution of R given a value $U = u$, i.e., $R|U = u$ or, for short, $R|u$, does not depend on θ and can therefore be written as $f_{R|U}(r|u;\xi)$.

- Under the MCAR mechanism the response indicator R and the interesting variables U are assumed to be independent with $f_{R|U}(r|u;\xi) = f_R(r;\xi)$ for all U.

- Under the MAR mechanism the conditional distribution of $R|U = u$ does not depend on the missing data U_{mis} and is given by $f_{R|U}(r|u;\xi) = f_{R|U_{obs}}(r|u_{obs};\xi)$ for all U_{mis}.

Thus we have seen, if the parameters ξ and θ are distinct and the missing-data mechanism is at least MAR, then the conditional distribution of $R|u$ is given by $f_{R|U}(r|u;\xi) = f_{R|U_{obs}}(r|u_{obs};\xi)$. The conditional distribution of $R|u$ is independent of U_{mis} and θ; the missingness mechanism is said to be ignorable. The likelihood (1) of the observed data of any unit $i \in \mathcal{U}$ under MAR can now be factorized into

$$\begin{aligned} L(\theta, \xi; u_{obs}, r) &= \int f_{U_{obs},U_{mis}}(u_{obs}, u_{mis}; \theta) f_{R|U_{obs},U_{mis}}(r|u_{obs}, u_{mis}; \xi) du_{mis} \\ &= \underbrace{\int f_U(u; \theta) du_{mis}}_{= L(\theta; u_{obs})} f_{R|U_{obs}}(r|u_{obs}; \xi) \, , \quad \theta \in \Omega_\theta, \, \xi \in \Omega_\xi . \end{aligned} \qquad (2)$$

According to Little and Rubin (2002) and illustrated by (2) under ignorable missingness, it is not necessary to consider a model for R if likelihood-based inference about θ is intended.

For the above example as shown in Figure 3, if θ and ξ are distinct and $f_{R|U_1,U_2}(r|u_1, u_2; \xi)$ does not dependent on the missing data, i.e., the MAR assumptions holds, then the observed-data likelihood reduces to

$$\begin{aligned} L(\theta, \xi; u_{obs}, r) &= \prod_{i=1}^{r} f_{U_1,U_2}(u_{i1}, u_{i2}; \theta) f_{R|U_1,U_2}(r_i|u_{i1}, u_{i2}; \xi) \times \prod_{i=r+1}^{n} f_{U_1}(u_{i1}; \theta) f_{R|U_1}(r_i|u_{i1}; \xi) \\ &= \underbrace{\prod_{i=1}^{r} f_{U_1,U_2}(u_{i1}, u_{i2}; \theta) \times \prod_{i=r+1}^{n} f_{U_1}(u_{i1}; \theta)}_{= L(\theta; u_{obs})} \\ &\quad \times \prod_{i=1}^{r} f_{R|U_1,U_2}(r_i|u_{i1}, u_{i2}; \xi) \times \prod_{i=r+1}^{n} f_{R|U_1}(r_i|u_{i1}; \xi) \end{aligned}$$

and maximizing $L(\theta; u_{obs})$ with respect to θ gives the correct ML estimate of θ. Thus, given n observations independently drawn from the underlying population, the likelihood ignoring the missing-data mechanism is

$$L(\theta; u_{obs}) = \prod_{i=1}^{n} L(\theta; u_{obs,i}) = \prod_{i=1}^{n} f_{U_{obs}}(u_{obs,i}; \theta) = \prod_{i=1}^{n} \int f_U(u_i; \theta) du_{mis,i} \, .$$

Notice that $u_{obs,i}$ describes the observed value of unit i for $i = 1, 2, \ldots, n$. Concerning the example above, $u_{obs,i} = (u_{i1}, u_{i2})$ for units $i = 1, 2, \ldots, r$ and $u_{obs,i} = (u_{i1})$ for units $i = r + 1, r + 2, \ldots, n$.

Hence, we have seen that all relevant statistical information about the parameters incorporated by θ should be contained in the observed-data likelihood $L(\theta; u_{obs})$, if the complete-data model, i.e., the data generating process assuming no missingness, and the ignorability assumption is correct.

3 Multiple imputation

Since the theoretical motivation for multiple imputation is Bayesian, a short introduction to the Bayesian way of argumentation is given here first.

3.1 Bayesian Inference

The Bayesian paradigm is based on specifying a probability model for the observed data U with joint density $f_{U|\Theta}(u|\theta)$ given a vector of unknown parameters $\Theta = \theta$ which is identical to the likelihood function $L(\theta; u)$ understood as a function of θ. Then we assume that Θ is random[2] and has a prior distribution with density or probability function f_Θ. Inference about Θ is then summarized in the function $f_{\Theta|U}$, which is called the posterior distribution of Θ given the data. The posterior distribution is derived from the joint distribution $f_{U,\Theta} = f_{U|\Theta} f_\Theta$ according to Bayes' formula

$$f_{\Theta|U}(\theta|u) \;=\; \frac{f_{\Theta,U}(\theta,u)}{f_U(u)} = \frac{f_{U|\Theta}(u|\theta)f_\Theta(\theta)}{\int_\Omega f_{\Theta,U}(\theta,u)d\theta} = \frac{L(\theta;u)f_\Theta(\theta)}{\int_\Omega L(\theta;u)f_\Theta(\theta)d\theta}, \tag{3}$$

where Ω denotes the parameter space of Θ. Notice that from a Bayesian perspective the joint distribution $f_{U|\Theta}(u|\theta)$ equates the likelihood $L(\theta; u)$ when the data are observed and only Θ is still variable.

For brevity again the integral is used, although Θ may also be discrete. In such cases the integral should be understood as the sum. From (3) it is easily seen that $f_{\Theta|U}(\theta|u)$ is proportional to the likelihood multiplied by the prior; i.e.,

$$f_{\Theta|U}(\theta|u) \;=\; c(u)^{-1}L(\theta;u)f_\Theta(\theta) \propto L(\theta;u)f_\Theta(\theta) = f_{U|\Theta}(u|\theta)f_\Theta(\theta),$$

and thus involves a contribution from the observed data through $L(\theta; u)$ and a contribution from prior information quantified through $f_\Theta(\theta)$. The quantity

$$c(u) = \begin{cases} \int_\Omega f_{U|\Theta}(u|\theta)f_\Theta(\theta)d\theta & \text{if } \Theta \text{ is continuous,} \\ \sum_\Omega f_{U|\Theta}(u|\theta)f_\Theta(\theta) & \text{if } \Theta \text{ is discrete,} \end{cases}$$

is usually treated as the normalizing constant of $f_{\Theta|U}(\theta|u)$ ensuring that it is a density or probability function, i.e., to integrate or sum to one. Notice that $c(u)$ is a constant when the data $U = u$ are observed. Before the data U are obtained, their distribution $f_U(u)$ is called the marginal density of U or the prior predictive distribution, which is not conditioning on previous observations. To predict a future observation value $\widehat{u}$ when the data $U = u$ have been observed, we condition on these previous observations u. The distribution $f_{\widehat{U}|U}(\widehat{u}|u)$ of $\widehat{U}|U = u$ is called the posterior predictive distribution with

$$f_{\widehat{U}|U}(\widehat{u}|u) \;=\; \int_\Omega f_{\widehat{U}|\Theta,U}(\widehat{u}|\theta,u)f_{\Theta|U}(\theta|u)d\theta = \int_\Omega f_{\widehat{U}|\Theta}(\widehat{u}|\theta)f_{\Theta|U}(\theta|u)d\theta \tag{4}$$

if Θ is continuous, otherwise the sum is taken instead of the integral. Notice that usually $\widehat{U}$ and U are assumed to be conditionally independent given Θ; thus $f_{\widehat{U}|\Theta,U}(\widehat{u}|\theta,u) = f_{\widehat{U}|\Theta}(\widehat{u}|\theta)$ holds. Hence the posterior predictive distribution is conditioned on the values $U = u$ already observed and predicts a value $\widehat{U} = \widehat{u}$ that is observable.

A classical and extensive introduction to Bayesian inference is given by Box and Tiao (1992); for deeper insights into Bayesian inference and computation the interested reader is referred thereto. For further reading concerning Bayesian inference we recommend

[2] To make clear that the parameter θ is treated as a random variable in the Bayesian context, we use capital letters for the random variable Θ as far as possible.

Berger (1985), Gelman et al. (2000), and Carlin and Louis (2000). Frequentist methods, however, do not tell us what our belief in a theory should be, given the data we have actually observed. This question is usually answered by the posterior distribution $f_{\Theta|U}$. To work out this value we must first establish f_{Θ}; i.e., we have to formulate some "prior probability" for the theory in mind. In contrast to classical Bayesian inference we do not focus further on what we can learn about our theory given the data. Our objective is the derivation of a suitable imputation procedure that has good properties under the frequentist's randomization perspective.

In the Bayesian framework all inference is based on a posterior density function for the unknown parameters conditioning on the quantities observed. Returning to our notation the unknown parameters are (Θ, Ξ) and the observed quantities are (U_{obs}, R). According to Bayes' theorem the posterior distribution of (Θ, Ξ) given $(U_{obs} = u_{obs}, R = r)$, i.e., the observed-data posterior distribution $f_{\Theta,\Xi|U_{obs},R} = f_{\Theta,\Xi,U_{obs},R}/f_{U_{obs},R}$, may be written as

$$\begin{aligned} f_{\Theta,\Xi|U_{obs},R}(\theta,\xi|u_{obs},r) &= c^{-1} f_{U_{obs},R|\Theta,\Xi}(u_{obs},r|\theta,\xi) f_{\Theta,\Xi}(\theta,\xi) \\ &= c^{-1} L(\theta,\xi;u_{obs},r) f_{\Theta,\Xi}(\theta,\xi) \qquad (5) \end{aligned}$$

with normalizing constant

$$c = \int\int f_{U_{obs},R,\Theta,\Xi}(u_{obs},r,\theta,\xi)d\theta\, d\xi = \int\int f_{U_{obs},R|\Theta,\Xi}(u_{obs},r|\theta,\xi) f_{\Theta,\Xi}(\theta,\xi) d\theta\, d\xi\,.$$

Note that $L(\theta,\xi;u_{obs},r)$ denotes the likelihood (2) of the observed data considering the missingness mechanism, and $f_{\Theta,\Xi}(\theta,\xi)$ is the joint prior distribution of the parameters.

Under the assumption of MAR and the distinctness of (Θ, Ξ), which means prior independence of Θ and Ξ in Bayesian inference, i.e., $f_{\Theta,\Xi}(\theta,\xi) = f_{\Theta}(\theta) f_{\Xi}(\xi)$, according to (2) the observed-data posterior (5) reduces to

$$\begin{aligned} f_{\Theta,\Xi|U_{obs},R}(\theta,\xi|u_{obs},r) &= c^{-1} f_{U_{obs}|\Theta,\Xi}(u_{obs}|\theta,\xi) f_{R|U_{obs},\Theta,\Xi}(r|u_{obs},\theta,\xi) f_{\Theta}(\theta) f_{\Xi}(\xi) \\ &= c^{-1} f_{U_{obs}|\Theta}(u_{obs}|\theta) f_{R|U_{obs},\Xi}(r|u_{obs},\xi) f_{\Theta}(\theta) f_{\Xi}(\xi)\,. \qquad (6) \end{aligned}$$

From the Bayesian point of view the MAR assumption requires the independence of R and Θ, i.e., $f_{R|U_{obs},\Theta,\Xi}(r|u_{obs},\theta,\xi) = f_{R|U_{obs},\Xi}(r|u_{obs},\xi)$, as well as the independence of U_{obs} and Ξ, i.e., $f_{U_{obs}|\Theta,\Xi}(u_{obs}|\theta,\xi) = f_{U_{obs}|\Theta}(u_{obs}|\theta)$ leading to (6) finally.

Hence the marginal posterior distribution of Θ is achieved by integrating (6) over the nuisance parameter Ξ with

$$\begin{aligned} f_{\Theta|U_{obs},R}(\theta|u_{obs},r) &= \int f_{\Theta,\Xi|U_{obs},R}(\theta,\xi|u_{obs},r)d\xi \\ &= \int c^{-1} f_{U_{obs}|\Theta}(u_{obs}|\theta) f_{R|U_{obs},\Xi}(r|u_{obs},\xi) f_{\Theta}(\theta) f_{\Xi}(\xi) d\xi \\ &= c^{-1} \underbrace{f_{U_{obs}|\Theta}(u_{obs}|\theta)}_{= L(\theta;u_{obs})} f_{\Theta}(\theta) \int f_{R|U_{obs},\Xi}(r|u_{obs},\xi) f_{\Xi}(\xi) d\xi \\ &\propto L(\theta;u_{obs}) f_{\Theta}(\theta) \qquad (7) \\ &\propto f_{\Theta}(\theta) \prod_{i=1}^{n} \int f_{U|\Theta}(u_i|\theta) du_{mis,i}. \end{aligned}$$

Now use $c = f_{U_{obs},R}(u_{obs}, r) = f_{R|U_{obs}}(r|u_{obs})f_{U_{obs}}(u_{obs}) = c_1 \cdot c_2$, then, from (7) we realize that $f_{\Theta|U_{obs},R}(\theta|u_{obs}, r)$ is also proportional to

$$c_2^{-1} L(\theta; u_{obs}) f_\Theta(\theta) = f_{\Theta|U_{obs}}(\theta|u_{obs}).$$

Thus, under ignorability all information about Θ is included in the posterior that ignores the missing-data mechanism,

$$f_{\Theta|U_{obs}}(\theta|u_{obs}) \;=\; c_2^{-1} L(\theta; u_{obs}) f_\Theta(\theta) = c_2^{-1} f_\Theta(\theta) \prod_{i=1}^{n} f_{U_{obs}}(u_{obs,i}; \theta). \tag{8}$$

For brevity we refer to (8) as the "observed-data posterior" according to Schafer (1997), p. 17. With incomplete data, however, the usual conjugate prior distributions no longer lead to posterior distributions that are recognizable or easy to summarize.

3.2 Multiple imputation paradigm

The theoretical motivation for multiple imputation is Bayesian, although the resulting multiple imputation inference is usually valid also from a frequentist viewpoint. Basically, MI requires independent random draws from the posterior predictive distribution $f_{U_{mis}|U_{obs}}$ of the missing data given the observed data analogous to (4). Since $f_{U_{mis}|U_{obs}}$ itself often is difficult to draw from directly, a two-step procedure for each of the m draws is useful:

(a) First, we make random draws of the parameters Θ according to their observed-data posterior distribution $f_{\Theta|U_{obs}}$ according to (8),

(b) then, we perform random draws of U_{mis} according to their conditional predictive distribution $f_{U_{mis}|U_{obs},\Theta}$.

Because

$$f_{U_{mis}|U_{obs}}(u_{mis}|u_{obs}) = \int f_{U_{mis}|U_{obs},\Theta}(u_{mis}|u_{obs}, \theta) f_{\Theta|U_{obs}}(\theta|u_{obs}) d\theta \tag{9}$$

holds, analogous to (4), with (a) and (b) we achieve imputations of U_{mis} from their posterior predictive distribution $f_{U_{mis}|U_{obs}}$. For many models the conditional predictive distribution $f_{U_{mis}|U_{obs},\Theta}$ is rather straightforward due to the data model used; see as an example Figure 4. It often may easily be formulated for each unit with missing data.

On the contrary, the corresponding observed-data posteriors $f_{\Theta|U_{obs}}$ usually are difficult to derive for those units with missing data, especially when the data have a multivariate structure and different missing data patterns, see for illustration Figure 4. The observed-data posteriors are often not standard distributions from which random numbers can easily be generated. However, simpler methods have been developed to enable multiple imputation on the grounds of Markov chain Monte Carlo (MCMC) techniques; they are extensively discussed by Schafer (1997). In MCMC the desired distributions $f_{U_{mis}|U_{obs}}$ and $f_{\Theta|U_{obs}}$ are achieved as stationary distributions of Markov chains which are based on the easier to compute complete-data distributions.

To proceed further, let θ denote a scalar quantity of interest that is to be estimated, such as a mean, variance, or correlation coefficient. Notice that now θ can be completely

Unit no.	**Variable**								
	$\boldsymbol{U_1}$	$\cdots$	$\boldsymbol{U_l}$	$\boldsymbol{U_{l+1}}$	$\cdots$	$\boldsymbol{U_k}$	$\boldsymbol{U_{k+1}}$	$\cdots$	$\boldsymbol{U_r}$
$\boldsymbol{i}$	$\boldsymbol{u}_{obs,i}$			$\boldsymbol{u}_{mis,i}$					
$\boldsymbol{j}$	$\boldsymbol{u}_{obs,j}$						$\boldsymbol{u}_{mis,j}$		

$$\begin{array}{llllll}
U|\theta & \sim f_{U|\Theta}(u|\theta) & \Rightarrow & U_{mis}|u_{obs},\theta & \sim & f_{U_{mis}|U_{obs},\Theta}(u_{mis}|u_{obs},\theta) \text{ , e.g, for units } i,j \\
U|\theta & \sim N_r(\mu_U,\Sigma_U) & \Rightarrow & U_{mis,i}|u_{obs,i},\theta & \sim & N_{r-l}(\mu_{U_{mis,i}|U_{obs,i}},\Sigma_{U_{mis,i}|U_{obs,i}}) \\
 & & \Rightarrow & U_{mis,j}|u_{obs,j},\theta & \sim & N_{r-k}(\mu_{U_{mis,j}|U_{obs,j}},\Sigma_{U_{mis,j}|U_{obs,j}})
\end{array}$$

Figure 4: Example of a conditional predictive distribution

different from the data model used before to create the imputations. In the remainder of this section, the quantity θ to be estimated from the multiply imputed data set, has to be distinguished from the parameter θ used in the model for imputation.

Consider, for example, the DACSEIS expenditure survey. Let us assume that only the income (inc) information is missing for some units. If the imputation of the missing income is based on the expenditure (exp) information, e.g., by applying a simple linear regression $inc_i = \alpha_0 + \alpha_0\alpha_1 exp_i + \epsilon_i$ for $i = 1, 2, \ldots n$, then θ (imputation) $= (\alpha_0, \alpha_1)$. If, on the other hand, the analyst's model explains expenditure by income, such that $exp_i = \beta_0 + \beta_1 inc_i + \nu_i$ for $i = 1, 2, \ldots n$ holds, then θ analysis) $= (\beta_0, \beta_1)$.

Although θ (analysis) could be an explicit function of θ (imputation), as it is the case in the example above, one of the strengths of the multiple imputation approach is that this need not be the case. In fact, θ (analysis) could even be the parameter of the imputation model, then imputation and analysis model are the same and are said to be congenial, a term coined by Meng (1995). However, multiple imputation is designed for situations when the analyst and the imputer are different, thus, the analyst's model could be quite different from the imputer's model. As long as the two models are not overly incompatible or the fraction of missing information is not high, inferences based on the multiply imputed data should still be approximately valid. Even more, if the analyst's model is a sub-model of the imputer's model, i.e., the imputer uses a larger set of covariates than the analyst and the covariates are good predictors of the missing values, then MI inference can beat the best inference possible using only the variables in the analyst's model. This property is called superefficiency by Rubin (1996). On the other hand, if the imputer ignores some important correlates of variables with missing data, but these variables are used in the analyst's model, then the result will be biased. Consider again the DACSEIS expenditure surveys and the situation of imputing income without using expenditure. This refers to an imputation being done under the hypotheses of zero correlation between income and expenditure which is surely not the case, thus, results will be biased.[3] Moreover,

[3] Rubin (1987) and Schafer (1997, Chapter 4) and their references therein discuss the distinction between θ (analysis) and θ (imputation) more fully.

the imputer's model also allows to use in-house variables such as additional information from the interviewers (area of living, neighborhood, house size, number of car garages etc.) which are typically not available to the analyst but may show some correlation with the missing variables. In general and specifically for the DACSEIS recommendations on MI[4], we suggest to use as much variables in the imputation model as are available or are reasonable. Thus, an inclusive strategy is nearly always better than a restrictive one.

As described before, $U = (U_{obs}, U_{mis})$ denotes the random variables concerning the data with observed and missing parts, and $\widehat{\theta} = \widehat{\theta}(U)$ denotes the statistic that would be used to estimate θ if the data were complete. Furthermore, let $\widehat{var}(\widehat{\theta}) = \widehat{var}(\widehat{\theta}(U))$ be the variance estimate of $\widehat{\theta}(U)$ based on the complete data set.

The MI principle assumes that $\widehat{\theta}$ and $\widehat{var}(\widehat{\theta})$ can be regarded as an approximate complete-data posterior mean and variance for θ, with

$$\widehat{\theta} \approx E(\Theta | u_{obs}, u_{mis})$$

and

$$\widehat{var}(\widehat{\theta}) \approx var(\Theta | u_{obs}, u_{mis})$$

based on a suitable complete-data model and prior; see also Schafer (1997), p. 108. Moreover, we should assume that with complete data, tests and interval estimates based on the normal approximation

$$(\widehat{\theta} - \theta)/\sqrt{\widehat{var}(\widehat{\theta})} \sim N(0,1) \tag{10}$$

should work well. Hence, we assume that the complete-data inference can be based on $\widehat{\theta} \sim N(\theta, var(\widehat{\theta}))$ and that $\widehat{var}(\widehat{\theta})$ is of lower-order variability than $var(\widehat{\theta})$; see Li et al. (1991). Notice that the usual maximum-likelihood estimates and their asymptotic variances derived from the inverted Fisher information matrix typically satisfy these assumptions. Sometimes it is necessary to transform the estimate $\widehat{\theta}$ to a scale for which the normal approximation can be applied. For example, we can use the so-called z-transformation for the correlation coefficient estimate, with $z(\widehat{\rho}) = (1/2)\ln((1+\widehat{\rho})/(1-\widehat{\rho}))$, which makes $z(\widehat{\rho})$ approximately normally distributed with mean $z(\rho)$ and constant variance $1/(n-3)$; see Schafer (1997), p. 216 and Brand (1999), p. 116.

Suppose now that the data are missing and we make $m > 1$ independent simulated imputations $(U_{obs}, U_{mis}^{(1)})$, $(U_{obs}, U_{mis}^{(2)})$, $\ldots$, $(U_{obs}, U_{mis}^{(m)})$ enabling us to calculate the imputed data estimate $\widehat{\theta}^{(t)} = \widehat{\theta}(U_{obs}, U_{mis}^{(t)})$ along with its estimated variance $\widehat{var}(\widehat{\theta}^{(t)}) = \widehat{var}(\widehat{\theta}(U_{obs}, U_{mis}^{(t)}))$, $t = 1, 2, \ldots, m$. Figure 5 illustrates the multiple imputation principle. From these m imputed data sets the multiple imputation estimates are computed.

The MI point estimate for θ is simply the average

$$\widehat{\theta}_{MI} = \frac{1}{m} \sum_{t=1}^{m} \widehat{\theta}^{(t)}. \tag{11}$$

[4] For details of the proposed MI algorithms in DACSEIS see deliverable D11.2 by Laaksonen et al. (2003).

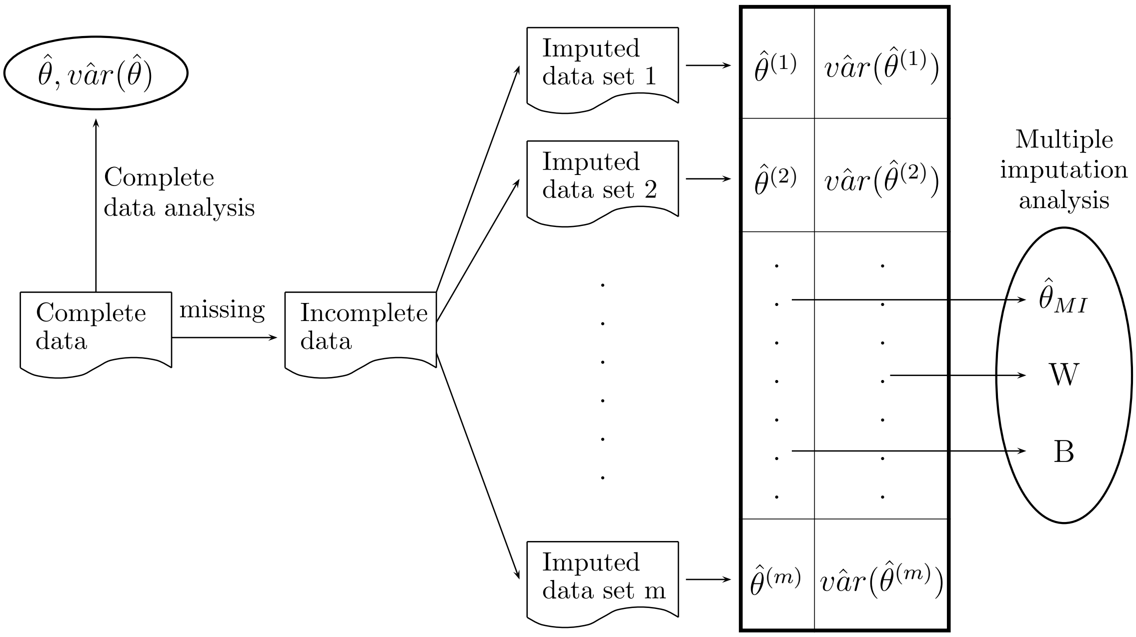


Figure 5: The multiple imputation principle

To obtain a standard error $\sqrt{\widehat{var}(\widehat{\theta}_{MI})}$ for the MI estimate $\widehat{\theta}_{MI}$ we first calculate the "between-imputation" variance

$$\widehat{var}(\widehat{\theta})_{between} = B = \frac{1}{m-1}\sum_{t=1}^{m}(\widehat{\theta}^{(t)} - \widehat{\theta}_{MI})^2,$$

and then the "within-imputation" variance

$$\widehat{var}(\widehat{\theta})_{within} = W = \frac{1}{m}\sum_{t=1}^{m}\widehat{var}(\widehat{\theta}^{(t)})\,.$$

Finally, the estimated total variance is defined by

$$\begin{aligned}\widehat{var}(\widehat{\theta}_{MI}) &= T = \widehat{var}(\widehat{\theta})_{within} + (1+\frac{1}{m})\widehat{var}(\widehat{\theta})_{between} \\ &= W + \frac{m+1}{m}B. \qquad (12)\end{aligned}$$

Notice that the term $((m+1)/m)B$ enlarges the total variance estimate T compared to the usual analysis of variance with $T = B + W$; $(m+1)/m$ is an adjustment for finite m. An estimate of the fraction of missing information γ about θ due to nonresponse is given by

$$\widehat{\gamma} = \frac{(1+1/m)B}{T}.$$

For large sample sizes, tests and two-sided $(1-\alpha)100\%$ interval estimates can be based on the Student's t-distribution

$$(\widehat{\theta}_{MI} - \theta)/\sqrt{T} \sim t_v \quad \text{and} \quad \widehat{\theta}_{MI} \pm t_{v,1-\alpha/2}\sqrt{T} \qquad (13)$$

with the degrees of freedom,

$$v = (m-1)\left(1 + \frac{W}{(1+m^{-1})B}\right)^2, \tag{14}$$

which are based on a Satterthwaite approximation; see Rubin and Schenker (1986) or Rubin (1987), pp. 76–77. For small data sets an improved expression for the degrees of freedom is given by Barnard and Rubin (1999). They relax the assumption of a normal reference distribution of (10) for the complete-data interval estimates and tests to allow a t distribution, and they derive the corresponding degrees of freedom for the MI inference to replace the formula (14) given here. Moreover, additional methods are available for combining vector estimates and covariance matrices, p-values, and Likelihood-ratio statistics (see Little and Rubin, 2002).

From (13) we realize that the multiple imputation interval estimate is expected to produce a larger interval than an estimate based only on the observed cases or based only on one single imputation. The multiple imputation interval estimates are widened to account for the missing data uncertainty and simulation error (see Schafer, 1999).

3.3 Efficiency of the multiple imputation estimates

Rubin (1987), p. 114, shows that the relative efficiency of an estimate based on m imputations to one based on $m = \infty$ number of imputations is approximately $1 + \gamma/m$ to 1, where γ is the rate of missing information. Taking $m = 3$ multiple imputations and assuming a fraction of 50% missing information an estimate based on this $m = 3$ imputations has a standard error that is about 8% higher than one based on $m = \infty$, because $\sqrt{1 + 0.5/3} = 1.0801$. Schafer (1999) states that unless the fraction of missing information is unusually high (i.e., far more than 50%), there is little benefit in using more than 5 to 10 imputations.

Notice that the multiple imputation theory is developed under the assumption that the imputer and the analyst use a common Bayesian model. To account for a wider variety of applications, Rubin (1987), pp. 113-147, addressed the frequency properties of multiple imputation methods and, therefore, defined the term “proper”. Roughly speaking, a proper multiple imputation method leads to inferences that are valid also from the (random-response) randomization-based perspective, if the complete-data inference is a valid randomization-based inference. In general, if (10) for the complete case estimator holds and the imputations are proper, then the multiple imputation estimate (11) is a consistent, asymptotically normal estimator in the frequentist sense, and an estimator of its asymptotic variance is given by (12), for a recent discussion see Nielson (2003), Meng and Romero (2003) and Rubin (2003a). Usually it is to be expected that imputations which are independently drawn from (9) should be proper or at least approximately proper. However, it can be quite difficult to prove that an imputation procedure is proper in Rubin’s sense. Therefore, Schafer (1997), p. 105, coined the term Bayesianly proper for an imputation procedure that generates independent realizations of the posterior predictive distribution $f_{U_{mis}|U_{obs}}$. Bayesianly proper imputations do not necessarily imply proper imputations. For example, the multiple imputation method based on a regression model proposed by Schenker and Welsh (1988) is Bayesianly proper, but not proper in

Rubin's frequentist sense. But, once a procedure is Bayesianly proper, at least some of the conditions of being proper are automatically satisfied.

It should be mentioned here that there is an ongoing discussion about how to account for design features, such as clustering and weighting. In complex survey designs it can be extremely difficult to find proper imputation methods allowing to construct confidence intervals that are frequency valid, see Binder and Sun (1996) or Marker et al. (2002). Moreover, a multiple imputation method that leads to valid frequentist inference for one complete data estimator may not be frequency valid for another one. However, as Rubin (2003a) states "it's not that MI is so great, it's that other generally available methods are worse, either computationally, analytically, or with respect to statistical validity." There is a growing body of evidence, see, e.g., Schafer (1997), pp. 372-377, p. 383, Schafer and Yucel (2002), Schafer (2003), Rubin (2003b), or Rässler and Schnell (2004), that also with complex survey designs MI works well and possibly better than traditional techniques, such as weighting. The general advice for creating multiple imputations accounting for complex survey designs is to include design variables or also weights in the imputer's model. Of course, more work needs to be done to develop flexible MI algorithms especially for complex panel surveys. However, it is its broad applicability which makes MI so appealing for the DACSEIS project. Finally, even if data are multiply imputed using a sensitive but imperfect model, then according to Rubin (2003a) MI "will typically lead to slightly conservative inference, that is, inferences that have coverage that is slightly larger than nominal." The proposed MI models for DACSEIS as specified in D11.2, Laaksonen et al. (2003), are sensitively chosen, such that any arbitrary inference should be, at least, approximately valid.

We do not want to replicate all the inferential questions and justifications for the MI principle in general here; they are extensively described by Rubin (1987, 1996), Schafer (1997), and Brand (1999). An illustration of the verification of a proper imputation method is shown in Figure 6 which is adopted from Brand (1999), p. 115, and extended. For the DACSEIS simulation studies similar procedures are used to evaluate different imputation techniques. Remember that the concept of proper imputation methods requires the complete-data inference to be randomization-valid; i.e., for repeated sampling from the underlying population the following should approximately hold,

$$\widehat{\theta} \sim N_1(\theta, var(\widehat{\theta})), \qquad E(\widehat{var}(\widehat{\theta})) = var(\widehat{\theta}),$$

and $\widehat{var}(\widehat{\theta})$ has less variability than $\widehat{\theta}$; see Rubin (1987), p. 118.

To evaluate whether a MI technique is proper for a set of complete-data statistics ($\widehat{\theta}(U)$, $\widehat{var}(\widehat{\theta}(U))$) by means of Monte Carlo simulation the following simplified validity conditions may be discussed.

$$\begin{aligned} E(\widehat{\theta}_{MI}) &= \widehat{\theta}(U), \\ E(W) &= \widehat{var}(\widehat{\theta}(U)), \\ E(B) &= var(\widehat{\theta}_{MI}). \end{aligned} \tag{15}$$

As the number of imputations becomes large (i.e., $m \to \infty$,) these equations (15) demand that $\widehat{\theta}_{MI}$, W, and B are unbiased estimates of the statistics $\widehat{\theta}(U)$, $\widehat{var}(\widehat{\theta}(U))$, and $var(\widehat{\theta}_{MI})$. Notice that $\widehat{\theta}(U)$ and $\widehat{var}(\widehat{\theta}(U))$ are based on the hypothetical complete data

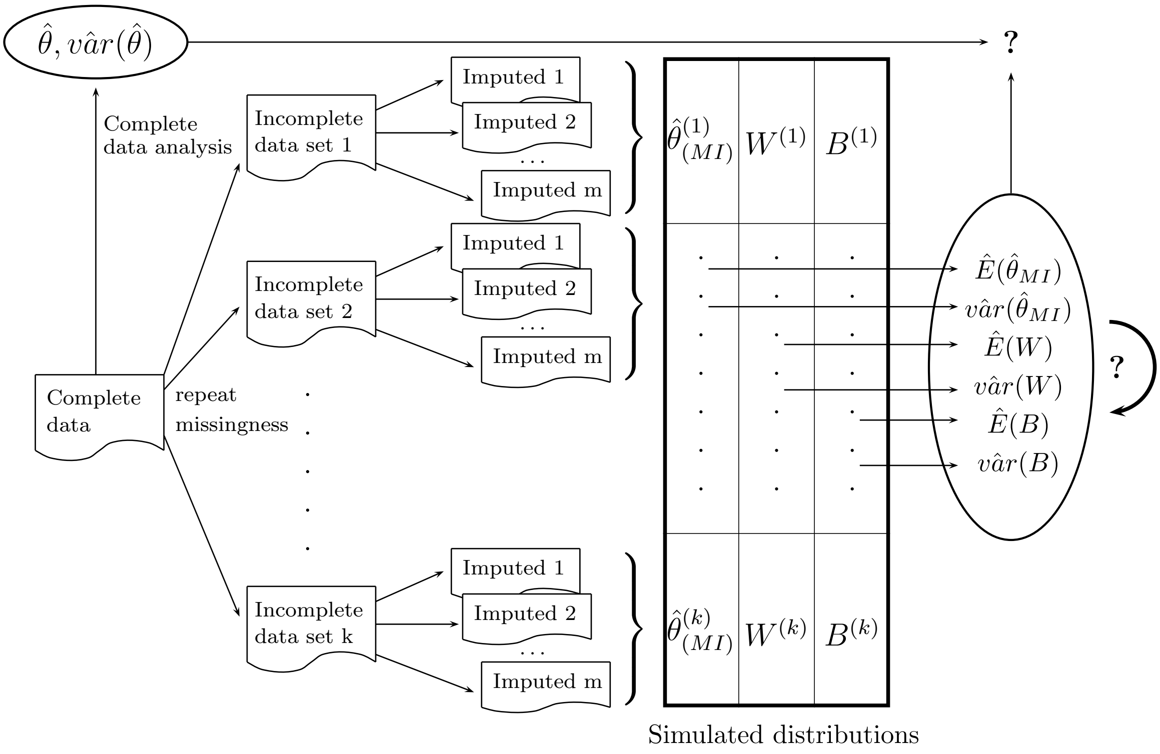


Figure 6: Simulation process to verify a proper imputation method

set when no data are missing and $var(\widehat{\theta}_{MI})$ describes the true variance of the MI estimate for a given observed data set and $m = \infty$ number of imputations.

More precisely, it is required that $(\widehat{\theta}_{MI} - \widehat{\theta}(U))/\sqrt{B}$ should become approximately $N(0,1)$, when the number of imputations gets large and the data U are held fixed. The within-imputation variance W should be an unbiased estimate of the complete-data variance estimate $\widehat{var}(\widehat{\theta}(U))$, and the variances of W and B should be smaller than the true variance $var(\widehat{\theta}_{MI})$ of the MI estimate, if again the data are regarded as fixed and $m = \infty$. Finally, over repeated samples the true variance $var(\widehat{\theta}_{MI})$ should be of lower variability than $\widehat{\theta}(U)$.[5] By means of Monte Carlo simulations we may recognize that an imputation procedure can be, at least approximately, proper for a set of complete-data statistics $(\widehat{\theta}(U), \widehat{var}(\widehat{\theta}(U)))$ by verifying whether

$$\begin{aligned}\widehat{E}(\widehat{\theta}_{MI}) &\approx \widehat{\theta}(U),\\ \widehat{E}(W) &\approx \widehat{var}(\widehat{\theta}(U)),\\ (1+m^{-1})\widehat{E}(B) &\approx \widehat{var}(\widehat{\theta}_{MI}) \quad \text{with} \quad m \ll \infty\end{aligned}$$

holds; see Brand (1999), pp. 114-117. A schematic overview of the simulation process to verify whether an imputation method may be proper is presented in Figure 6.

[5] For the original definition of a proper imputation method see Rubin (1987), pp. 118–119; a summary is given by Schafer (1997), pp. 144–145. A detailed discussion of the verification of proper multiple imputations is provided by Brand (1999), pp. 75–91 and pp. 114–117.

4 A simple simulation study

In this section we present a simple simulation study just for the purpose of illustration. The focus is to compare MI with results produced by a simple single mean imputation (SI) as well as a single mean imputation within classes (SI CM) or by using only the available cases (AC).

Assume that age (AGE) is normally distributed with mean 40 [years] and standard error of 10 [years], then take income (INC) as normally distributed with mean 1500 [EURO] and standard error of 300 [EURO]. Moreover, let the correlation between age and income be about 0.8. So we let

$$(AGE, INC) \sim N\left(\left(\begin{array}{c} 40 \\ 1500 \end{array}\right), \left(\begin{array}{cc} 10^2 & 0.8 \cdot 3000 \\ 0.8 \cdot 3000 & 300^2 \end{array}\right)\right)$$

A sample of n = 2000 is drawn from this universe. After being generated, the AGE variable is recoded into 6 categories, 1 <= 20 years, 2 = 20 - 30 years, ..., 6 > 60 years. First, the complete cases are analyzed, the mean income estimate, its standard error (s.e.), and the 95% confidence interval are calculated. Then different missingness mechanisms (MCAR, MAR, MNAR) are applied on income. Under MAR, income is missing with higher probability when age is higher, under MNAR, the probability that income is missing is higher the higher income is itself.

After discarding 30% of the income data, the available cases are analyzed, then a simple mean imputation is performed, and, finally, a proper multiple imputation procedure is applied according to Rubin (1987), p. 167. The whole simulation process of creating the data, applying the missingness, performing the imputations, and analyzing the sample is repeated 1000 times. The coverage (cvg.) is counted, i.e., the number of confidence intervals out of 1000 that cover the true mean value. The average width of the 95% confidence interval is reported and the usual correlation estimate between age (recoded) and income is given.

Very clearly the Table 1 shows how precision is reduced when only the available cases are used under MCAR, and how biased the available case estimate gets when the missingness is MAR or MNAR. The table also shows how biased a simple mean imputation is and how this bias is corrected when conditional means are imputed instead of the overall mean. However, this only works when the missingness depends on the variable conditioned on. The single mean imputation within classes also leads to an overestimation of the correlation between recoded AGE and INC though the simple single imputation underestimates it. Moreover, with single imputation the standard errors are always too small to get the nominal coverage.

Though the missingness is MCAR, a simple mean imputation is quite harmful to standard errors and correlation. Under MAR and even under MNAR, multiple imputation works by far better than the other alternatives, in the latter case borrowing strength from the correlation between age and income. Standard errors, correlation and the nominal coverage are well reproduced by MI. Notice that confidence intervals under MI can be shorter than confidence intervals based only on the complete or available cases (AC). This is especially true if the imputed sample is substantially larger than the complete case sample. Therefore, typically, the following comparisons hold for most surveys and most estimates of standard errors:

$$\text{s.e.(SI)} < \text{s.e.(truth)} < \text{s.e.(MI)} < \text{s.e.(AC)}.$$

Missing	Proc.	Cvg.	Mean(INC)	S.e. (INC)	CIwidth	Cor(AGE, INC)
None	CC	0.96	1500.21	6.71	26.3	0.77
MCAR	AC	0.95	1500.14	8.01	31.44	0.77
MCAR	SI	0.82	1500.14	5.61	22.01	0.64
MCAR	SI CM	0.91	1500.20	6.28	24.63	0.82
MCAR	MI	0.95	1500.24	7.34	29.1	0.77
MAR	AC	0.04	1470.35	7.98	31.31	0.77
MAR	SI	0.01	1470.35	5.58	21.90	0.63
MAR	SI CM	0.88	1499.90	6.28	24.65	0.82
MAR	MI	0.93	1499.82	7.43	29.50	0.77
MNAR	AC	0.11	1474.29	7.99	31.34	0.77
MNAR	SI	0.03	1474.29	5.59	21.91	0.64
MNAR	SI CM	0.59	1489.33	6.26	24.56	0.82
MNAR	MI	0.71	1489.30	7.36	29.20	0.77

Table 1: Results of the simulation study

5 Final comments

With the increasing computational power, more and more multiple imputation techniques are being implemented, making multiple imputation inference quite easy to perform. For an overview see Table 2.

There are programs and routines available for free, such as the stand-alone Windows program NORM or the S-PLUS libraries NORM, CAT, MIX, PAN, and MICE (also available for R now) wich are all basically data augmentation algorithms. NORM uses a normal model for continuous data, CAT a log-linear model for categorical data. MIX relies on a general location model for mixed categorical and continuous data. PAN is created for panel data applying a linear mixed-effects model. The new missing data library in S-PLUS 6 features these models and simplifies consolidating the results of multiple complete-data analyses after multiple imputation. Moreover, there is the free SAS-callable application IVEware, the SAS procedures PROC MI, PROC MIANALYZE, as well as the free Windows or Gauss version AMELIA. PROC MI provides a parametric and a nonparametric regression imputation approach, as well as the multivariate normal model. MICE as well as IVEware are very flexible tools for generating multivariate imputations for different kinds of variables by using chained equations. Finally, SOLAS for Missing Data Analysis 3.0 is a commercial Windows program provided by Statistical Solutions Limited. For links and further details see www.multiple-imputation.com or Horton and Lipsitz (2001).

As already mentioned, Meng (1995, 2002) coined the term “congeniality” which basically means that the imputer’s model should agree with the analyst’s model. In some sense this is assured if the imputer uses, at least, the same set of input data, i.e., variables and

Programme	Inf. Prior	Available at	Speciality
NORM, CAT, MIX, PAN	yes	`http://www.stats.psu.edu/~jls` `http://cran.r-project.org/`	S-Plus / R library
Missing Data Library	ridge prior	Insightful Corporation	S-Plus library
NORM 2.03	ridge prior	`http://www.stat.psu.edu/~jls`	standalone (WIN)
AMELIA	no	`http://gking.harvard.edu/stats.shtml`	standalone (DOS)
Hmisc	no	`http://hesweb1.med.virginia.edu/biostat/s` `http://cran.r-project.org/`	S-Plus / R library
IVEware	no	`http://www.isr.umich.edu/src/smp/ive/`	SAS procedure
MICE V1.0	no	`http://www.multiple-imputation.com`	S-Plus / R library
PROC MI	no	SAS Institue Inc.	SAS procedure
SOLAS 3.0	no	Statistical Solutions Ltd.	standalone (WIN)

Table 2: Currently available Software for MI

observations. Thus, problems of misspecification should be avoided. Schafer (2003) points out that MI performs well when many variables are incorporated in the imputer‘s model although assumptions may be violated. Finally, empirical evidence suggests that multiple imputation under MAR often is quite robust against violations of this assumption. Even an erroneous assumption of MAR may have only minor impact on estimates and standard errors computed using multiple imputation strategies. Only when MNAR is a serious concern and the fraction of missing information is substantial, does it seem necessary to model jointly the data and the missingness. Moreover, since the missing values cannot be observed, there is no direct evidence in the data to address the MNAR-assumption. It can be more helpful, therefore, to consider several alternative models and to explore the sensitivity of resulting inferences.

Thus, we like to conclude that a multiple imputation procedure seems to be the best alternative at hand to account for missingness and to exploit all valuable available information. In general, it is crucial to use multiple rather than single imputation so that subsequent analyses will be statistically valid. Otherwise, with single imputation, effort has to be placed on correcting variance estimates to assure valid inference. Notice that an extensive comparison of different variance estimation methods when values are (singly or multiply) imputed will be a result of the DACSEIS project.

Acknowledgement

The author wants to thank all of the members of the DACSEIS team for their valuable comments and their interest in multiple imputation. Special thanks go to Ralf Münnich, the coordinator of the DACSEIS project, for valuable discussions and his interest in the implementation of the imputation routines. Finally, the author is indebted to Don Rubin and his ”MI-team” for helpful insights and advice proffered during numerous discussions.

References

Barnard, J., Rubin, D.B. (1999), Small-Sample Degrees of Freedom with Multiple Imputation, Biometrika, 86, 948-955.

Berger, J.O. (1985), Statistical Decision Theory and Bayesian Analysis, Springer, New York.

Binder D.A., Sun, W. (1996), Frequency Valid Multiple Imputation for Surveys with a Complex Design, Proceedings of the Section on Survey Research Methods of the American Statistical Association, 281-286.

Box, G.E.P., Tiao, G.C. (1992), Bayesian Inference in Statistical Analysis, Wiley, New York.

Brand, J.P.L. (1999), Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets, Thesis Erasmus University Rotterdam, Print Partners Ispkamp, Enschede, The Netherlands.

Carlin, B.P., Louis, T.A. (2000), Bayes and Empirical Bayes Methods for Data Analysis, Chapman and Hall, London.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (1995), Bayesian Data Analysis, Chapman and Hall, London.

Horton, N.J., Lipsitz, S.R. (2001), Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables, The American Statistician, 55, 244-254.

Laaksonen, S., Rässler, S., Skinner, C. (2003), Documentation of Pseudo Code of Imputation Methods for the Simulation Study, DACSEIS Deliverable D11.2, Tübingen.

Lee, H., Rancourt, E., Särndal C.E. (2002), Variance Estimation from Survey Data under Single Imputation, Survey Nonresponse (eds. Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A.), 315-328, Wiley, New York.

Li, K.H., Raghunathan, T.E., Rubin, D.B. (1991), Large-Sample Significance Levels from Multiply Imputed Data Using Moment-Based Statistics and an F Reference Distribution, Journal of the American Statistical Association, 86, 1065-1073.

Little , R.J.A., Rubin, D.B. (2002), Statistical Analysis with Missing Data, Wiley, New York.

Marker, D.A., Judkins, D.R., Winglee, M. (2002), Large-Scale Imputation for Complex Surveys, Survey Nonresponse (eds. Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A.), 329-341, Wiley, New York.

Meng, X.L. (1995), Multiple-Imputation Inferences with Uncongenial Source of Input (with discussion), Statistical Science, 10, 538-573.

Meng, X.L. (2002), A Congenial Overview and Investigation of Multiple Imputation Inferences under Uncongeniality, Survey Nonresponse (eds. Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A.), 343–356, Wiley, New York.

Meng, X.L., Romero, M.(2003), Discussion: Efficiency and Self-efficiency With Multiple Imputation Inference, International Statistical Review, 71, 607-618.

Nielson, S.F. (2003), Proper and Improper Multiple Imputation, International Statistical Review, 71, 593-607.

Rässler, S., Schnell, R. (2004), Multiple Imputation for Unit-Nonresponse versus Weighting Including a Comparison with a Nonresponse Follow-Up Study, submitted to Allgemeines Statistisches Archiv.

Rubin, D.B. (1978), Multiple Imputation in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse, Proceedings of the Survey Research Methods Sections of the American Statistical Association, 20-40.

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

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

Rubin, D.B. (2003a), Discussion on Multiple Imputation Inference, International Statistical Review, 71, 619-627.

Rubin, D.B. (2003b), Nested Multiple Imputation of NMES via Partially Incompatible MCMC, Statistica Neerlandica, 57, 3-18.

Rubin, D.B., Schenker, N. (1986), Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse, Journal of the American Statistical Association, 81, 366-374.

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

Schafer, J.L. (1999), Multiple Imputation: a Primer, Statistical Methods in Medical Research, 8, 3-15.

Schafer, J.L. (2003), Multiple Imputation in Multivariate Problems When the Imputation and the Analysis Models Differ, Statistica Neerlandica, 57, 19-35.

Schafer, J.L., Yucel, R. (2002), Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values, Journal of Computational and Graphical Statistics, 11, 437-457.

Schenker, N., Welsh, A.H. (1988), Asymptotic Results for Multiple Imputation, Annals of Statistics, 16, 1550-1566.

Shao, J. (2002), Replication Methods for Variance Estimation in Complex Surveys with Imputed Data, Survey Nonresponse (eds. Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A.), 303–314, Wiley, New York.

The following papers are already published in the

DACSEIS research paper series

No.1 **Münnich, Ralf; Wiegert, Rolf (2001)**

The DACSEIS Project

http://w210.ub.uni-tuebingen.de/dbt/volltexte/2001/428

No.2 **Zhang, Li-Chun (2002)**

A method of weighting adjustment for survey data subject to nonignorable nonresponse

http://w210.ub.uni-tuebingen.de/dbt/volltexte/2002/451

No.3 **Quatember, Andreas (2002)**

A comparison of the five Labour Force Surveys of the DACSEIS project from a sampling theory point of view

http://w210.ub.uni-tuebingen.de/dbt/volltexte/2002/547

No.4 **Münnich, Ralf; Schürle, Josef (2003)**

On the Simulation of Complex Universes in the Case of Applying the German Microcensus

http://w210.ub.uni-tuebingen.de/dbt/volltexte/2003/979