# Modeling nonresponse in a synthetic population: An introduction to missing data and the evaluation of correction methods

Nicolas Kaiser
St. Mergener Straße 10, 54292 Trier
University of Trier
s4nikais@uni-trier.de
Third semester
M.sc Survey Statistics
1448043

Course: Weighting and calibration (Second Semester)
Supervisor: Anne Konrad

October 29, 2019

# Contents

# 1 Introduction

For a long time now, a largely well-founded and proven methodology has been in place to draw conclusions about the underlying population from sampled data, which can be subsumed under the term inference statistics (cf. Bethlehem, Cobben, Schouten: 2011, 2). However, here we are initially assuming an ideal situation of sampling, which is hardly ever encountered in reality (Rubin: 1976; 2009, Leeuw: 2001). Missing data occur in almost every data set and can distort corresponding statements about the underlying population. This problem has been exacerbated by increasing non-response rates in recent years (Matai and Ranalli: 2018, Kreuter: 2013). Our data can be almost useless in worst cases. For the founders of the probability sampling approach this was perfectly clear (cf. Brick: 2013, 330). Since then, there have been many suggestions on how to deal with missing data and numerous methods have been developed to reduce possible distortions (see for example Särndal and Lundström: 2006). However, some of these methods may even increase the bias in cases of non-ignorable nonresponse (cf. Graham: 2009, 570). Until recently, a much bigger problem was that a common methodological approach to artificially generate missing data was not found. As a result, certain correction methods may be evaluated incorrectly. This article aims to achieve three objectives: 1. to provide a brief introduction to the problem of missing data. 2. to explain how missing data can be simulated and what criteria can should be used to evaluate the performance of different correction methods. 3. to demonstrate, by means of a simulation study, how to proceed with such a study and how the results can be interpreted. This will also allow us to verify if the simulation of missing data worked properly.

Monte Carlo simulations are suitable for investigating the efficacy of different methods that aim to reduce a possible bias in different situations (see for example Münnich and Schürle: 2003). First, a synthetic population is generated from which samples are taken by a specific selection scheme. In this way, the data-generating process as it takes place in reality is mapped as similarly as possible. We ensure that the sampling, as in reality, leads to various forms of missing data. Since we have generated the synthetic population ourselves, we know the true values and can therefore determine how much our estimates deviate from the true value in the long term. As there is now a proliferating number of methods for missing data, this article will concentrate on the best known and most commonly used approaches to give a basic overview. However, the focus will be on how missing data in a synthetic population has to be generated in order to be able to use such a Monte Carlo simulation. This seems appropriate, since writing about all forms of non-response and counter-strategies would clearly go beyond the scope of this article. Instead, the interested reader can use these explanations for his own study purposes. It should be noted that this article will simulate a cross-sectional survey on individuals, although there are of course many other forms like Household Surveys or Establishment Surveys (Earp et. al: 2014).

# 2 The Problem of Nonresponse

## 2.1 Sampling without missing data

Let us now briefly look at a frequently encountered selection mechanism to be able to make unbiased estimates about a target population, a sample obtained using a simple random sample (SRS). A finite target population $U$ with $N$ elements is considered

$$U = \{1, 2, \ldots, N\} \tag{1}$$

and a SRS $s$ is drawn out of a possible set of samples $S$. The population mean $\overline{y}$ can be obtained by the unbiased Horvitz-Thompson-Estimator (HT-Estimator). We assume complete response without any missing data (Horvitz and Thompson: 1952).

$$\overline{y}_{HT} = \frac{1}{N} \sum_{k=1}^{N} a_k \frac{Y_k}{\pi_k} \tag{2}$$

A certain value for an element $Y_k$ is divided by its inverse selection probability $\pi_k$. The indicator variable $a_k$ shows if a certain unit from the population became part of the sample (Betlehem, Cobben, Schouten: 2011, 31 f.). An estimator for an estimate $\hat{\theta}$ in general is referred as design unbiased if

$$E(\hat{\theta}) = \sum_{s \in \mathcal{S}} \hat{\theta}(s) p(s) = \theta \tag{3}$$

which is true for the HT-Estimator (Särndal et. al: 1978, 30). It is important to mention that only the distribution of $\hat{\theta}$ among all possible samples under a sampling scheme p($\cdot$) is relevant. If we let $n \to \infty$ in our sample the estimate will reach the true value $\theta$ of the population. However, unbiasedness alone is not enough. Our estimator must also be precise. Therefore the variance of our estimate is calculated. In the realistic case of a fixed sample size without replacement this can be obtained by

$$V\left(\overline{y}_{HT}\right) = \frac{1}{2N^2} \sum_{k=1}^{N} \sum_{l=1}^{N} \left(\pi_k \pi_l - \pi_{kl}\right) \left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2 \tag{4}$$

for the HT-Estimator. Additionally, we want to interpret the variance in terms of precision. Typically, to answer this question a 95-Percent confidence interval is calculated:

$$\left(\overline{y}_{HT} - 1.96 \times S\left(\overline{y}_{HT}\right); \overline{y}_{HT} + 1.96 \times S\left(\overline{\gamma}_{HT}\right)\right) \tag{5}$$

For what reason was it necessary to recall these simplest principles? In a nutshell: Although this well-proven methodology is permanently applied in surveys everyday, the equations merely reflect an ideal situation. In the real world, however, missing data appear for a variety of reasons which lead to bias. The above equations do not include this phenom because they "were not designed for them" (Schafer and Graham: 2002, 147). Consequently, we need to expand our theory.

## 2.2 Modeling nonresponse

To study the effect of non-response, we can use two different models which differ in their fundamental assumptions but lead to the same results if the sample size is relatively small compared to the population size ($n \ll N$) (cf. Münnich an Schürle: 2003, 42). These are the *fixed response model* and the *random response model*. They help us determine which conditions lead to biased estimators.

In the fixed response model (Bethlehem; Cobben; Schouten: 2011, 40 ff.) we assume the population to be divided into two mutually exclusive strata: A response stratum and a non-response stratum. At this point we introduce a set of response indicators

$$R_1, R_2, \ldots, R_N \tag{6}$$

which will also be relevant for further analysis. $R_k = 1$ for an element $k$ being part of the response stratum while $R_k = 0$ for an element $k$ in the non-response stratum. Therefore the mean for the target variable in the response stratum and a separate mean for the target variable in the non-response stratum are calculated. If both values were the same a bias would not appear. The bias itself can be expressed the following way

$$B\left(\overline{y}_R\right) = \overline{Y}_R - \vec{Y} = \frac{N_{NR}}{N}\left(\overline{Y}_R - \overline{Y}_{NR}\right) = QK \tag{7}$$

where $K = \overline{Y}_R - \overline{Y}_{NR}$ and $Q = N_{NR}/N$. As we can see, the larger the difference between reponse-mean and non-response-mean the larger the bias. Also a larger relative amount of non-respondents will increase the bias.

The random response model (e.g, 43 ff.) chooses a different approach. Here we assume an (unknown) response probability $\rho_k$ for every element $k$ in the population. Therefore, we stick to our concept of response indicators (6) meaning that $R_k = 1$ for the response of a corresponding element $k$ and $R_k = 0$ if the opposite is true. Again we can calculate the bias

$$B\left(\overline{y}_R\right) = \tilde{Y} - \overline{Y} = \frac{R_{\rho Y} S_\rho S_Y}{\overline{\rho}} \tag{8}$$

where $R_{\rho Y}$ shows the correlation between the target variable and the response probability. We have a standard deviation for the response probabilities $S_\rho$ and such for the standard deviation of the variable $Y$ which is $S_Y$. This equation leads to three conclusions: The stronger target variable and response probability are correlated, the larger the bias will be. Obviously, without a relationship between those variables the bias will disappear. Second, equal response probabilities will also lead to vanishing bias. Lower response rate will lead to higher bias such as in the fixed response model.

Non-response bias also influences our confidence interval (5). As a result the confidence level is much lower with increasing bias since we can only use the response mean $\overline{y}_R$. The resulting confidence interval under non-response bias changes to

$$P\left(\overline{Y} \in I_R\right) = \Phi\left(1.96 - \frac{B\left(\overline{y}_R\right)}{S\left(\overline{y}_R\right)}\right) - \Phi\left(-1.96 - \frac{B\left(\overline{y}_R\right)}{S\left(\overline{y}_R\right)}\right) \tag{9}$$

where $P(\overline{Y} \in I)$ gives the probability that our confidence interval contains the true value. $\Phi$ is the standard normal distribution while $B\left(\overline{y}_R\right)/S\left(\overline{y}_R\right)$ gives the *relative bias* (Bethlehem; Cobben; Schouten: 2011, 45 f.).

## 2.3 Nonreponse-mechanism

So far we have only talked about non-response in general without going into the different patterns of non-response and their effects. Nevertheless, this is a major point since different missing-data-mechanisms require different treatments. One should not think of those mechanisms as "mutually exclusive categories of missingness" (Graham: 2009, 567) but for the sake of theoretical clarity we will clearly distinguish between them. This concept was developed by Rubin (1976). It starts again with the missing data indicator R (6) having values 1 or 0 depending on units being observed or not. Often the distribution of R is called *response mechanism* although it is important to mention that this does not imply a causal relationship (cf. Schafer and Graham: 2002, 150).

To formalize the different non-response mechanisms we divide a dataset $Y$ into an observed part $Y_{obs}$ and a missing part $Y_{mis}$. The conditional distribution of $R$ given $Y$ is our *distribution of the missingsness.*

$$P(R|Y) = P\left(R|Y_{obs}, Y_{mis}\right) \tag{10}$$

Three different non-response-mechanism are the result. The first one is called *missing completely at random* (MCAR). In this case the distribution of $R$ given $Y$ is not influenced by the data at all. Consequently the equation is $P(R|Y) = P(R)$. As a result, a dataset which missing data were caused exclusively by MCAR would not be affected by bias. Instead, the more missing values according to this mechanism would lead to decreasing precision meaning higher variance of the estimator. In real world situations we can hardly ever assume that our missing data are MCAR.

If our distribution of the missingness only depends on the observed data $Y_{obs}$ but not on the missing data $Y_{mis}$ we call our non-response-mechanism *missing at random* (MAR). The equation changes to $P(R|Y) = P\left(R|Y_{obs}\right)$. Some correction methods for missing data are based on the assumption of MAR.

In the last case we are not able to simplify equation 10 any further. R is depending not only on the observed data but also on the missing data leading to a result called *missing not at random* (MNAR). This usually is referred to as *non ignorable nonresponse* because it can dramatically bias the results while at the same time correction methods only work under correct assumptions. Sullivan and Andridge (2015) give more details relating to recent discussions about MNAR.

At this point one must be expressly mentioned: The underlying assumptions about the existing non-response mechanism cannot be verified in reality (cf. Groves: 2006, 653). For reasons of scientific correctness, however, the mechanisms assumed should be made transparent. Alternative assumptions should lead to similar conclusions (cf. Schafer and Graham: 2002, 149).

## 2.4 Dealing with Unit-nonresponse

If an element in our sample did not participate in the survey (for whatever reasons) we call it $Unit\ -nonreponse$. The typical correction method in such a case is *weighting adjustment* (Kalton and Flores-Cervantes: 2003, Li et. al: 2013). If our Unit-non-response would completely be caused by MCAR this would, as mentioned before, not lead to bias but only to a less precise estimate with broader confidence intervals. Obviously, this is not a reasonable assumption. The scientific literature on weighting adjustment proposes numerous methods. Therefore, we describe the basic idea of weighing and refer to further literature for more detailed insights into specific techniques.

To perform weighting adjustment we need $auxiliary\ information$. Those information can be part of the sample but also be from other sources. For example, we could compare the distribution of a certain property in our sample with its distribution in the population (if we have such information) (cf. Groves: 2006, 655). If they differ too much one can assume a bias. How can these theoretical considerations be formalized?

What we need are so called *adjustment weights* which can be assigned to our elements in the sample to re-weight our estimate. We start again with the HT-Estimator (2) but this time we want to include our weights. We obtain a new weighted estimator

$$\overline{y}_W = \frac{1}{N}\sum_{i=1}^{n} w_i y_i \tag{11}$$

where $w_i = g_i \times d_i$ with $g_i$ being a correction weight for a certain weighting technique and an inclusion weight $d_i = 1/\pi_i$. This weighting correction should now lead to an unbiased estimate (cf. Bethlehem; Cobben; Schouten: 2011, 212 ff.).

Let us now take a brief look at the frequently used method of $post-stratification$ (e.g, 214 ff.). After the data are collected we assign stratum weights to over- and underrepresented groups. To obtain those non-overlapping strata we need $auxiliary\ information$. The idea is that this leads to *homogeneous* strata which is necessary to increase precision and reduce non-response bias. By using the inclusion probabilities and correction weights on equation 11 we obtain a new $post-stratification\ estimator$

$$\overline{y}_{PS} = \frac{1}{N}\sum_{h=1}^{L} N_h \overline{y}^{(h)} \tag{12}$$

where $\overline{y}^{(h)}$ gives the mean of the observed values in stratum $h$. Of course this estimator is again influenced by bias which is equal to

$$B\left(\overline{y}_{R,PS}\right)=\frac{1}{N}\sum_{h=1}^{L}N_hB\left(\overline{y}_R^{(h)}\right)\tag{13}$$

What do we need for small bias using our post-stratification estimator? The answer is small bias within the strata. This is accomplished by a small correlation between the target variable and the response behaviour within all strata. Additionally, low standard errors four our response probabilities as well as for the values of our target variables are needed.

The last method we look at for unit-non-response is *linear weighting*, a technique which is based on *generalized regression estimation* (e.g., 221 ff.). Here auxiliary variables are used to obtain a linear estimate for a target variable. Again, we start with the ideal case of full response. We assume a set of $p$ continuous auxiliary variables. For every element $k$ we obtain values which are stored in a $p-vector$

$$X_k=(X_{k1},X_{k2},\ldots,X_{kp})'\tag{14}$$

All values for a certain target variable are stored in an N-vector $Y$ while the values of the auxiliary variables can be found in a $N\times$ p-matrix $X$. The correlation of our auxiliary variable with the target variable will lead to a vector of regression coefficients $B$ for a best fit of $Y$ on $X$. We estimate this vector $B$ by

$$b=\left(\sum_{k=1}^{N}a_kX_kX_k'\right)^{-1}\left(\sum_{k=1}^{N}a_kX_kY_k\right)=\left(\sum_{i=1}^{n}x_ix_i'\right)^{-1}\left(\sum_{i=1}^{n}x_iy_i\right)\tag{15}$$

where $a_k$ serves as an dichotomous variable indicating if an element $k$ is selected in the sample. $x_i=(x_{i1},x_{i2},\ldots,x_{ip})'$ is the p-vector of values for the $p$ auxiliary variables. The sample elements are denoted as $i$ for $i=1,2,\ldots,n$. In a short way, we define our *generalized regression estimator* as

$$\overline{y}_{GR}=\overline{y}+(\overline{X}-\overline{x})'b\tag{16}$$

and extend our theory to the case of non-response. This leads to a modified version

$$\overline{y}_{GR,R}=\overline{y}_R+\left(\overline{X}-\overline{x}_R\right)'b_R=\overline{X}'b_R\tag{17}$$

where $b_R$ is the analogue to $b$ except it is only based on the response data. The bias of this new introduced estimator is approximately equal to

$$B\left(\overline{y}_{GR,R}\right)=\overline{X}B_R-\overline{Y}\tag{18}$$

where $B_R$ is given by

$$B_R=\left(\sum_{k=1}^{N}\rho_kX_kX_k'\right)^{-1}\left(\sum_{k=1}^{N}\rho_kX_kY_k\right)\tag{19}$$

Practically, our regression estimator is unbiased if the non-response is not influencing our regression coefficients. Formally expressed: If $B_R=B$ the bias vanishes.

## 2.5 Dealing with Item-nonresponse

If a person selected by our sampling scheme did participate but left some questions unanswered we call this $Item- \ non - response$. This leads to the fundamental problem of being unable to analyse the relationship between two variables $X$ and $Y$. It may come as a surprise, but in reality in such a case no attempts are usually made to estimate the missing values. Instead, the missing values are simply thrown out by $listwise \ -deletion$ also called *complete case analysis* (CCA) (cf. Gustavson et al.: 2019, 2). This approach has some drawbacks. Firstly, the number of elements in the sample is drastically reduced. On the other hand, it also deletes possible hints for the reasons why the data are missing. Since we also have to assume that the missing values are not MCAR, a bias has to be assumed by deleting the values.

The *pairwise-deletion* offers a slightly less serious intervention in our data. To estimate the relationship between our variables, only the cases are selected where $X$ and $Y$ both have values for. Nevertheless, neither CCA nor listwise-deletion are recommended in the scientific literature (see for example Little and Rubin: 2002, 39).

One can find a lot of debate in the scientific literature about the question if Unit- and Item-non-response have to be viewed as two different phenomena or if they are theoretically the same. For an overview about this question see Yan and Curtin (2010).

If we dont wish to delete our data we need an alternative approach. Therefore, so called $Imputation$-techniques are applied. This is also a broad field, which is why we will limit ourselves to the best known and most frequently applied methodologies (for an overview see Chen and Haziza 2019). If it is possible to apply imputation, why is it not always done that way? The answer is that imputation is no guarantee that the bias will disappear. Additionally, the underlying structure of the data can take damage. Non-response mechanisms as discussed in Section 2.2 are of crucial importance.

Lets start with a set of basic imputation-methods, subsumed under the term $single \ -imputation$ (Bethlehem; Cobben; Schouten: 2011, 421 ff.). One possible method is the $Mean - Imputation$. A value $Y_k$ for an element $k$ in the sample is missing. We impute this value by

$$\hat{Y}_k = \overline{y}_R = \frac{\sum_{k=1}^{N} a_k R_k Y_k}{\sum_{k=1}^{N} a_k R_k} \tag{20}$$

A well know disadvantage of this method is that it will reduce the variation of our distribution.

A less deterministic method is called $random \ -imputation$ (RI). We replace the missing value by randomly selecting values from the available values of the variable. The available values are nothing else but

$$\{Y_k | a_k = 1 \land R_k = 1\} \tag{21}$$

If our missing values are not missing randomly the distribution will differ from the real distribution.

Again we can use a model to estimate missing values. The idea of *regression imputation* (REG-I) is to assume a linear relationship between an auxiliary Variable $X$ and the target variable $Y$. We estimate the missing value for an element $k$ by

$$\hat{Y}_k = a + bX_k \tag{22}$$

where $b$ is estimated using ordinary least squares by

$$b = \frac{\sum_{k=1}^{N} a_k R_k \left(Y_k - \overline{y}_r\right)\left(X_k - \overline{x}_r\right)}{\sum_{k=1}^{N} a_k R_k \left(X_k - \overline{x}_r\right)^2} \tag{23}$$

and

$$a = \overline{y}_r - b\overline{x}_r \tag{24}$$

Such a regression model can of course be extended to a multivariate model with a couple of auxiliary variables.

Even if single-imputation has the advantage of having a complete data set at first, the application of this method can still cause more problems than it solves. At the end of this brief theory overview, we therefore take a look at *multiple imputation* (MI), a technique proposed by Rubin (1987). For each missing value we create $m > 1$ values with $m$ resulting complete datasets. To apply MI we start with the construction of a regression model which is used to create our synthetic values to replace the variable $Y$ with our missing values. It takes the form of

$$\hat{Y}_k = B_0 + \sum_{j=1}^{p} B_j X_{kj} + E_k \tag{25}$$

This model can be used very efficiently in the case of MAR. If our non-response is caused by MNAR we are very likely to draw wrong conclusions. Since we have multiple data-sets, how do we get to an overall estimator for the population mean? For every dataset $j$ we have an estimator $\overline{y}_j$ as well as an standard error $S\left(\overline{y}_j\right)$. Now we are able to estimate our population mean by

$$\overline{y}_{MI} = \frac{1}{m}\sum_{j=1}^{m} \overline{y}_j \tag{26}$$

The estimators variance is calculated by

$$V\left(\overline{y}_{MI}\right) = \frac{1}{m}\sum_{j=1}^{m} S^2\left(\overline{y}_j\right) + \left(1 + \frac{1}{m}\right)\frac{1}{m-1}\sum_{j=1}^{m}\left(\overline{y}_{MI} - \overline{y}_j\right)^2 \tag{27}$$

Rubin (1987) himself recommended not to exceed $m = 10$. However, Van Buuren (2018) gave an overview about the literature and finds, that other authors use a value between 20 and 100 for $m$. For our simulation we will stick to $m = 10$. Murray (2018) gives more insights on the complexity of MI.

# 3 Non-response in a synthetic population

## 3.1 The synthetic population

Before we can draw our samples, a synthetic population must first be created. Here we want to obtain data which can be generalized for the evaluation of statistical inference in cases of massing data. One possibility would be to use a $close-to-reality$ approach (Münnich and Schürle: 2003) for certain special situations like the german microcensus. A general approach to create a synthetic population based on sample data can be found in Alfons et. al (2011). The advantage would be that the approach presented there can be used for the preparation of a wide variety of household surveys with different levels of complexity. We call such a study $design-based\ simulation$ (Vink: 2016, 7).

In the past researchers often created a dataset based on the simple principle of drawing a random vector from a predefined distribution with certain patterns, structures and correlations, see for example Beckmann et. al (1996) or Barthelemy and Toint (2013). We stick to this approach which is known as $model-based\ simulation$ (Vink: 2016, 7). What kind of variables do we need for our simulation study?

- A target variable $Y_1$ we want to estimate
- A (observed) variable $Y_2$ more or less correlated with the target variable. We will use it for the MAR mechanism as $Y_{obs}$
- An (observed) auxiliary variable $X$ more ore less correlated with the target variable. It can be used for weighting or imputation techniques which rely on auxiliary information.

The data are generated by repeatedly drawing random numbers from a multivariate normal distribution using *mvrnorm* from the package $MASS$ (Venables and Ripley: 2002) in R. This is important since *parametric imputation*-techniques rely on an assumed distribution. We use a mean structure according to

$$\mu = \begin{array}{c} Y_1 \\ Y_2 \\ X_1 \end{array} \begin{pmatrix} 5 \\ 10 \\ 0 \end{pmatrix} \tag{28}$$

while the covariance pattern follows

$$\sum = \begin{array}{c} Y_1 \\ Y_2 \\ X_1 \end{array} \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} \tag{29}$$

This basic structure allows us to freely vary the number of elements in the population as well as certain correlation structures. Of course one could create far more complex synthetic populations with unique characteristics but like mentioned before we want to stick to a basic and easy concept.

For a more complex synthetic population one could create unobserved variables or factors which are influencing the response rates as well as other characteristics. We will solve the problem of unobserved influence by letting the response probability for a systematic error be based on the variable itself.

## 3.2 Adding nonresponse

If we want to imitate the mechanism of missing data, this should be as independent as possible from the data set we are using. Otherwise we run the risk of creating unique situations that cannot be generalized (Schouten et al. 2018). According to Schouten et. al (cf. 2018, 2190), a missing data simulation in general should include four steps:

- 1. A multivariate and complete data set has to be generated (3.1). It will be treated as the population of interest.
- 2. The synthetic population is made incomplete (3.2).
- 3. A certain method is used to estimate the incomplete data set (4.1-2).
- 4. Statistical inferences are compared for the original dataset and the missing dataset. With this information we evaluate the performance of a certain missing data method (4.1-2).

In our simulations, we found that it makes no significant difference whether the missing data is inserted into the population and then sampled, or whether the missing values are inserted into the sample after sampling from the complete synthetic population.

Overall, however, it has to be said that until recently there was no common methodology for generating missing data. Some authors did not even mention the mechanism they used (cf. Vink 2016, 10). If we look at the scientific literature to date, we see that researchers usually generate missing data using a logit model. Nishimura et. al (2016) use this approach to model the response probabilities of their sample units according to $logit\,(\rho_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i, i = 1, \ldots, n$. The coefficients vary to meet the needed response rate as well as the missing data-mechanism. Gustavson et. al (2019) model the "liability of non-response" according to $\mathrm{L} = \mathrm{b0} + \mathrm{b}_{\mathrm{non}}1^* \times 1 + \mathrm{b}_{\mathrm{non}}\mathrm{l}^* \times 2 + \mathrm{b}_{\mathrm{non}}1^* \times 3 + \mathrm{b}_{\mathrm{non}}2^*\mathrm{y} +$ random normal component.

The idea of creating non-response with a logit-model is a direct consequence of the existing missing-data mechanism introduced in section 2.2. A basic missing data model for MCAR, MAR and MNAR is given by Van Buuren (2018). From a standard bivariate normal distribution we generate our data $Y = (Y_1, Y_2)$ and let $Y_1$ be correlated with $Y_2$ (0.5). To create missing Values on $Y_2$ we calculate probability for non-response with

$$Pr\,(R_2 = 0) = \psi_0 + \frac{e^{Y_1}}{1 + e^{Y_1}}\psi_1 + \frac{e^{Y_2}}{1 + e^{Y_2}}\psi_2 \qquad (30)$$

where $\psi = (\psi_0, \psi_1, \psi_2)$ are different parameter settings to specify the missing data mechanism. Therefore we specify $\psi_{\text{MCAR}} = (0.5, 0, 0)$, $\psi_{\text{MAR}} = (0, 1, 0)$ and $\psi_{\text{MNAR}} = (0, 0, 1)$ resulting in three missing data models

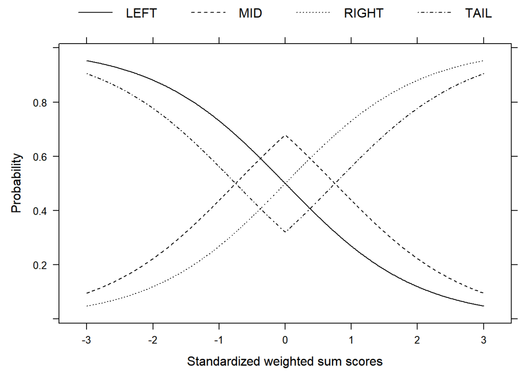
$$\begin{array}{rl} \text{MCAR}: & Pr\left(R_2 = 0\right) = 0.5 \\ \text{MAR}: & logit\left(Pr\left(R_2 = 0\right)\right) = Y_1 \\ \text{MNAR}: & logit\left(Pr\left(R_2 = 0\right)\right) = Y_2 \end{array} \tag{31}$$

Unfortunately, this is not enough to describe the generation process completely. The logit function used itself also influences the course of our missing data. But if we shift the logit-function we also influence other missing data parameters. Van Buuren (2018) gives an example for this describing a MAR-mechanism with three different logit-functions

$$\begin{array}{rl} MARRIGHT: & logit\left(Pr\left(R_2 = 0\right)\right) = -5 + Y_1 \\ MARMID: & logit\left(Pr\left(R_2 = 0\right)\right) = 0.75 - |Y_1 - 5| \\ MARTAIL: & logit\left(Pr\left(R_2 = 0\right)\right) = -0.75 + |Y_1 - 5| \end{array} \tag{32}$$

which can be visualized with a fourth possibility (Figure 1).

Figure 1: Visualisation of four possible logit-functions (Schouten et. al: 2018)



Schouten et. al (2018) have recently described such approaches as "*stepwise univariate amputation*" and criticized it for possibly misleading interpretation of the effectiveness of certain correction methods especially in cases where multiple values are amputed. This is because researchers so far had to apply univariate imputation on different variables over and over again to create multivariate missingness. Therefore, the missingness on two or more variables was independent from each other. To avoid this, Schouten et. al propose a solution to this problem called *multivariate amputation* which promises to be applicable on any dataset. This method builds upon an idea developed by Brand (1999, 110-113). Brand proposed a method for creating multivariate MAR-missingness. It can be described as follows: We want to generate non-monotone multivariate missing data in $p$ variables $Y_1, \ldots, Y_p$ and assume that

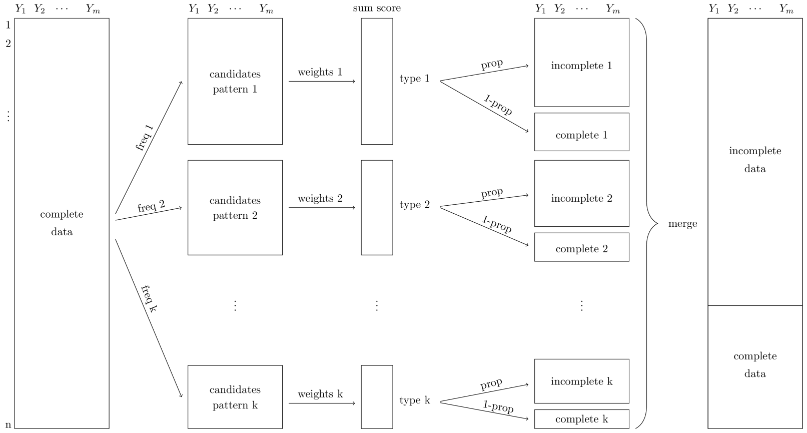
$Y = (Y_1, \ldots, Y_p)$ is completely known. Before we apply our method we specify a set of parameters. $\alpha$ is the desired proportion of incomplete cases. Possible response patterns are stored inside $R_{\text{pat}}$, a binary $n_{\text{pat}} \times p$ matrix defining $n_{\text{pat}}$ allowed patterns. We store the relative frequencies of those pattern in a vector $f = \left(f_{(1)}, \ldots, f_{(n_{\text{pat}})}\right)$. For each pattern, a set of $n_{\text{pat}}$ response probability models is defined as $P(R|Y) = \left(P\left(R_{(1)}|Y(1)\right), \ldots, P\left(R_{(n_{\text{pat}})}|Y_{(n_{\text{pat}})}\right)\right)$.

Before it is decided which units in the sample do not respond the dataset is ”randomly divided into $k$ subsets” (Schouten e. al: 2018, 2913) according to the pre-defined frequency value representing the nonresponse rate. The randomly allocated units in the subset are denoted as *candidates* because it is not yet decided if they will receive missing values. The decision about this is made by the new introduced concept of *weighted sum scores*. Their value can be obtained by a linear regression where the researcher determines the coefficients. For a case $i$ we calculate the weighted sum scores by

$$wss_i = w_1 \cdot y_{1i} + w_2 \cdot y_{2i} + \cdots + w_m \cdot y_{mi} \tag{33}$$

where a set of variables values in case $i$ is given with $\{y_{1i}, y_{2i}, \ldots, y_{mi}\}$ while $\{w_1, w_2, \ldots, w_m\}$ are the corresponding pre-specified weights. If we, for instance, choose a weight value of zero for all variables we create MCAR missingness. For creating MAR missingness weights of zero should be assigned to the variables we want to ampute. The simulation of an MNAR situation is fulfilled by assigning non-zero weights to the variables we like to ampute. The whole missing-data generation procedure as described is visualised in Figure 2.

Figure 2: Visualisation of the underlying procedure of ”ampute” (Schouten et. al: 2018)



Schouten et. al (2018) implemented this methodology with the function *ampute* inside the R-package *mice* (van Buuren and Groothuis-Oudshoorn: 2011). We will use this function for the generation of our own missing data pattern while keeping the complexity of the missing data at a simple level.

Although the R-package ”ampute” can generate complex multivariate missing data, we will only focus on missing values on one variable (Y1) caused by an exclusive missing data mechanism. The reason we want to give for this is that this article can only give an introduction to the various possibilities of analysis and implementation of missing data in Monte Carlo simulations. The fact that problems of common imputation methods can already be pointed out at such a low complexity level should make the inclined reader think.

# 4 Simulation Study

## 4.1 Settings and Evaluation Criteria

Before we can start with the simulation, we first have to define the criteria we want to use to evaluate our results. At first, we calculate the *relative Bias* (RB) as follows

$$RBias := \frac{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_r - \theta\right)}{\theta} \tag{34}$$

which gives us the deviation of the estimate $\hat{\theta}$ for the true value of our target variable $\theta$ in the $r$th simulation run. If we have no bias the result should converge against zero. In cases where we apply a certain correction method on an estimate $\widehat{\theta}_r$ this new corrected value is used. We use a threshold of 5-percent as criteria for an acceptable performance (Demirtas; Freels; Yucel: 2008). Additionally, we want to evaluate the $confidence\ interval\ coverage\ rate$ (CI-Rate), since we learned that the coverage rate changes under bias (9). We obtain it by

$$\text{CI}rate_{\alpha} := \frac{1}{R}\sum_{r=1}^{R} I\left(\theta \in CI\left(\widehat{\theta}_r, \alpha\right)\right) \tag{35}$$

and will stick to the often used 95 percent confidence level. A coverage rate way below 95-percent gives an too ”optimistic” estimate and will lead to bias in estimating p-values for significance tests. Now we also want a measure to evaluate the combination of bias and precision. Therefore we calculate the *relative root mean squared error* (RRMSE)

$$RRMSE := \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_r - \theta\right)^2}}{\theta} \tag{36}$$

Every combination of different circumstances needs its own simulation run. Therefore lets have a look at the different conditions we like to vary:

- We use four different nonresponse frequencies (0.10, 0.25, 0.5, 0.75)
- The three nonresponse-mechanisms (MCAR, MAR, MNAR) vary
- We vary the strength of the correlation between our variables according to $\rho \in \{0.2, 0.4, 0.6, 0.8\}$

- We use the before mentioned imputation-techniques (2.4) resulting in 5 different methods

This leads to 4 x 3 x 4 x 5 = 240 different simulations. Every Monte-Carlo-Simulation has $R = 1000$ runs. The synthetic Population consists of $N =$ 100.000 elements while in every simulation run a simple random sample of size $n = 1000$ is drawn from the population. According to our non-response-rate, this sample size will be reduced because of the missingness. In addition, we must note that the strength of the MAR mechanism depends on the strength of the correlations of the variables. A MAR mechanism with weak correlations is more similar to a MCAR mechanism (Vink: 2016, 5). Therefore, four different results are given for the MAR mechanism, depending on the strength of the correlations among the variables. In cases of MCAR or MNAR there is no reason to vary the correlation coefficient, expect for the use of auxiliary variables like in regression imputation. Therefore, such results are listed as "-" which means that there is only random variation compared to results with different correlation coefficients.

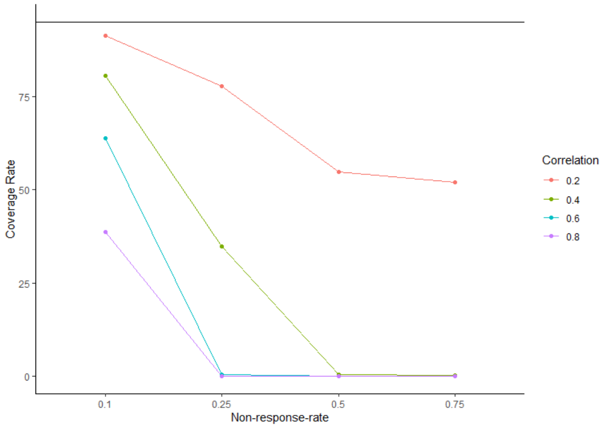
## 4.2 Results

Many authors do not give the output of the MCAR results in missing data simulations. However, we will do this because the results there reveal whether the Missing Data algorithm works at all. In addition, we can show that some imputation methods already have problems under MCAR at least considering coverage rates (see as well van Buuren: 2018). Lets have a look at the common method of *complete case analysis* (CCA). As expected, under MCAR there is almost no relative Bias (RB) and the coverage rates (CI-Rate) are very good. The loss in precision (RRMSE) is not notable at all. Even under a non-response-rate of 75-percent the CI-Rate stays close to 95-percent. In cases of MAR or MNAR the picture is quiet the opposite: Even a weak MAR-mechanism leads to notable RB. In the typical case of a non-response of 50-percent, the CI-Rates become almost useless under a medium MAR-mechanism. The stronger the MAR-mechanism, the faster RB increases and CI-Rates drop - very often to zero. We visualized this effect in Figure 3. The results for the different results under MAR in cases of different methods and correlations are visualized in figures 4 and 5.

The MNAR-mechanism gives us quiet similar results which are a bit worse than the MAR-mechanism. This is caused by the fact, that the response probability is a function of the variable itself, which makes it a stronger correlation than in the used 0.8 correlation MAR-mechanism.

The method of *mean imputation* shows a before mentioned typical problem. While it is true that mean imputation under MCAR will not lead to RB, the variance becomes highly skewed leading to under-coverage. As by definition mean imputation as a deterministic method will lead to no improvements in bias reduction compared to CCA when the missing data are MAR or MNAR. In those cases the CI-Rates converge fast to zero making them problematic for tests of significance.

Figure 3: Drop of CI-Rates under different correlations in case of MAR and CCA

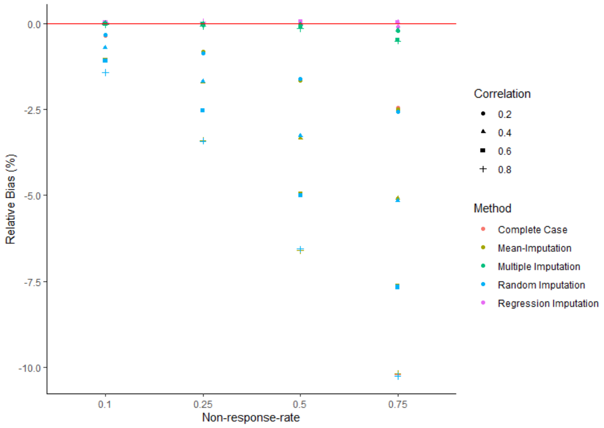


Using *random imputation* (RI) as an alternative to mean-imputation can be viewed as an improvement. RRMSE and RB will perform similarly to mean-imputation but CI-Rate is much better.

For our analysis of regression imputation a word of warning must be uttered. In our simulations this method works very well in MCAR and MAR situations with strongly correlated auxiliary variables. This is no surprise, as the auxiliary information can perfectly estimate the original variable in strong correlations. In the case of MAR, the simulation is based on the extremely unlikely assumption that the response mechanism can be explained by a single variable (Y2), that this variable is also part of our data set and can be estimated as a simple linear regression model. So the excellent results of the simulation are more evidence that the algorithm works than the idea that regression imputation is always a superior method. A more realistic insight can be obtained by looking at regression imputation under MNAR. The use of auxiliary variables do not lead to a perfect CI rate or even a disappearance of the RB, but strong correlations lead to a significant decrease of the before mentioned. This is visualized in Figure 6. Here, conclusions can be drawn about weighting methods for unit-nonreponse, which are also based on strong auxiliary variables.

If we take a look at the simulation results of *multiple imputation* (MI) using predictive mean matching (Little: 1988) it becomes quiet clear why this method is treated as a "principled and practical solution" to missing data problems (Murray: 2018, 156). Like the theoretical assumptions stated, in cases of MAR-missingness MI makes the bias vanish almost completely while simultaneously the CI-Rates drop but still remain quiet good for the amount of missingness in the data. Only if we choose a non-response frequency over 50 percent a massive drop of performance has to be faced. (Figure 7). It can also be confirmed that under MNAR-missingness MI will fail and produce similar results to complete case analysis. This shows us again that systematic missingness remains a big

Figure 4: Relative Bias under different methods and correlations (MAR)



problem in survey methodology. In contrast to the methods presented here, which do not improve RB but also do not make it worse, methods for MNAR missingness may well lead to a deterioration in the accuracy of our estimate.

All in all, the results show that ampute is excellent for generating missing values. The tested imputation methods led under appropriate conditions continuously to the expected results. Thus, this article could simultaneously evaluate common imputation methods and empirically test the method of generating missing data.

## 4.3 Discussion

Looking at the results of our simulation, one might ask why the bias for the different scenarios is not much larger. One reason for this is certainly that our population follows a perfect multivariate normal distribution, which is hardly the case in reality. A simulation based on non-normally distributed data would be interesting here. In addition, we assume that our target population is congruent with the basic population. In reality, however, it is difficult to get a perfect random sample from a population, as it is best to have a list of the elements in the population (cf. Lohr: 2009, 99 ff.) The various additional errors in sampling that arise here (Campanelli: 2009), as well as the problem of validly and reliably measuring an object by means of indicators, which is particularly common in the social sciences, should in reality significantly increase the bias (Hox: 2009). From this we can conclude that even under very good basic conditions, which are often hardly to be found in reality, missing data often lead to serious distortions. At this point, reference should be made to the further literature in which it is discussed how problematic this must be regarded as for research.

Furthermore, we should bear in mind that mixed forms of missing data occur in reality. MCAR, MAR and MNAR occur simultaneously in different propor-

Figure 5: Coverage Rates under different methods and correlations (MAR)

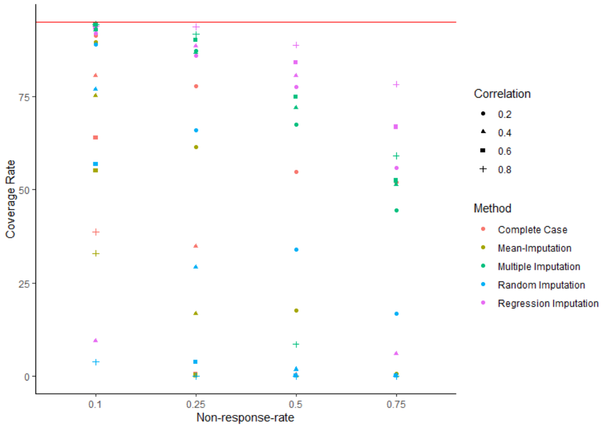
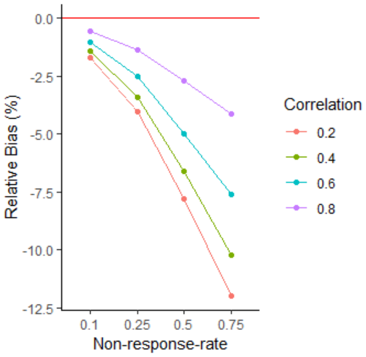


Figure 6: Relative bias under regression imputation with different correlations (MNAR)

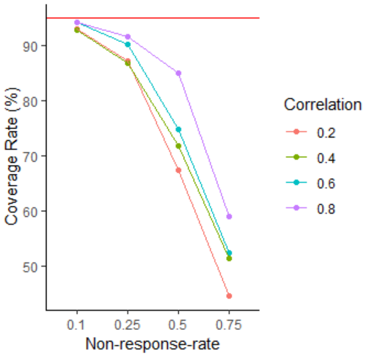


tions. This can be very well simulated with the ampute function and should be used for more complex simulations. For example, the fact that missing data in our simulation was only estimated on a single variable leads to an overestimation of the performance of complete case analysis. If missing values occur on multiple variables at the same time, the sample often will be dramatically reduced.

A core issue, which this article deliberately avoided, is the prevention of missing data. Prevention is always preferable to compensation. Nevertheless, it should at least be mentioned that increasing response rates through incentives can even lead to worse estimates if the inclusion-probabilities are changed by incentives (Groves: 2006, 666).

The right-shifted logit function was used for all simulations. For a comparison, the other possible shifts could also be used and their performance examined.

Figure 7: CI-Rates under MI in case of MAR with different correlations



As the reader will have noticed, the basics of weighting were discussed in section (2.3), but no separate simulation study was carried out. We argue that weighting methods are mainly based on auxiliary information, as is the case with regression imputation, which was discussed and simulated. The equations mentioned in section 2.3 show under which circumstances a high RB occurs and when not. The better the auxiliary information, the better the estimate. Our article should be sufficient to conduct such a study itself.

# 5 Conclusion

One of the aims of this article was to show the virulent problem of missing data and its consequences for the analysis of data based on inference statistical methods. The main question, however, was how to realistically simulate missing data in a synthetic population and evaluate its impact. Therefore, the theoretical assumptions about sampling were first explained and related to the occurrence of missing data. Three different mechanisms that lead to missing data were explained. We then looked at common methods for dealing with missing data. Subsequently, recent developments in the scientific literature that deal with the generation of missing data and why previous approaches may be inadequate were investigated. The main point of this article was to generate missing data in a realistic way using the function "ampute" inside the R-package *mice*. At the same time, it was possible to demonstrate which methods can be used to evaluate imputation methods and to test the functionality of "ampute". In summary, it can be said that for the generation of missing data, the R-package mentioned above can be used reliably. The evaluation of common imputation methods showed the picture known from the scientific literature: Single imputation can be a justified method in individual cases, but in most cases leads to serious disadvantages and is not able to reduce bias. Multiple imputation is a superior method under the assumption of MAR. The simulation results are further proof that there are good reasons why this method has become more and more accepted in recent years. Nevertheless, systematically missing data remains a serious problem that cannot be solved by the methods discussed in this article. After all, the methods used here do not lead to a worsening of the bias compared to complete case analysis. For the generation of the missing data a very simple pattern was chosen, which simulated missing values only on a variable. At the same time, the synthetic population was a perfect multivariate normal distribution - an ideal state that is unlikely to be found in reality. This article can therefore only be an introduction to the simulation and evaluation of missing data. The interested reader should thus be able to conduct his own research for more specific and complex simulations.

# List of Figures

# Nomenclature

| | |
|---|---|
| CCA | complete case analysis |
| CI-Rate | confidence interval coverage rate |
| HT | Horvitz-Thompson |
| MAR | missing at random |
| MCAR | missing completely at random |
| MI | multiple imputation |
| MNAR | missing not at random |
| RB | relative bias |
| REG-I | regression imputation |
| RI | random imputation |
| RRMSE | relative root mean squared error |
| SRS | simple random sample |

# List of Symbols

| | |
|---|---|
| $B$ | bias |
| $d_i$ | inclusion weight |
| $E$ | expected value |
| $GR$ | generalized regression |
| $k$ | an element |
| $m$ | number of imputations |
| $n$ | number of elements in sample |
| $N$ | number of elements in population |
| $R$ | response indicator |
| $P/Pr$ | probability |
| $PS$ | post-stratification |
| $s$ | sample |
| $S$ | set of possible samples |
| $S$ | standard deviation |
| $\overline{y}$ | mean |
| $w_i$ | correction weight |
| $wss_i$ | weighted sum score |
| $X_k$ | continuous auxiliary variables |
| $Y_k$ | element in the population with a certain value |
| $Y_{obs}$ | observed part of a dataset |
| $Y_{mis}$ | unobserved part of a dataset |
| $V$ | variance |
| $U$ | target population |
| $\alpha$ | proportion of missingness |
| $a_k$ | indicator variable showing if element in population was included in the sample |
| $\mathrm{p}(\cdot)$ | sampling scheme |
| $\rho$ | correlation coefficient |
| $\rho_k$ | response probability |
| $\pi_k$ | inclusion probability |
| $\hat{\theta}$ | estimation for a certain value |
| $\theta$ | true value in the population |
| $\Phi$ | standard normal distribution |
| $\psi = (\psi_0, \psi_1, \psi_2)$ | parameter settings for specification of missing data mechanism |

# A Appendix

MCAR

| Nonresponse | Cor | RRMSE (%) | CI-Rate | R-Bias (%) | Correction Method |
|---|---|---|---|---|---|
| 0.10 | 0.2 | 1,53 | 94,7 | -0,02 | Complete Case |

| | | | | | |
|---|---|---|---|---|---|
| 0.25 | | 1,68 | 94,5 | -0,03 | |
| 0.5 | | 2,09 | 93,3 | -0,05 | |
| 0.75 | | 2,85 | 94,6 | 0,00 | |
| 0.10 | 0.4 | 1,5 | 94,6 | -0,03 | |
| 0.25 | | 1,65 | 94,3 | -0,04 | |
| 0.5 | | 2,04 | 94,3 | -0,05 | |
| 0.75 | | 2,94 | 94,5 | 0,01 | |
| 0.10 | 0.6 | 1,51 | 94,9 | -0,04 | |
| 0.25 | | 1,67 | 94,4 | -0,04 | |
| 0.5 | | 2,09 | 94,4 | -0,06 | |
| 0.75 | | 2,95 | 94,7 | 0,01 | |
| 0.10 | 0.8 | 1,51 | 94,3 | -0,04 | |
| 0.25 | | 1,59 | 94,5 | -0,02 | |
| 0.5 | | 2,09 | 94 | -0,06 | |
| 0.75 | | 2,94 | 94,5 | 0,01 | |
| 0.10 | 0.2 | 1,46 | 93,1 | 0,00 | Mean-Imputation |
| 0.25 | | 1,63 | 85,4 | 0,01 | |
| 0.5 | | 2,01 | 66,9 | -0,02 | |
| 0.75 | | 2,81 | 38,6 | -0,03 | |
| 0.10 | 0.4 | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | 0.6 | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | 0.8 | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | 0.2 | 1,56 | 92,4 | 0,01 | Random Imputation |
| 0.25 | | 1,78 | 87,3 | 0,03 | |
| 0.5 | | 2,22 | 78,8 | 0,00 | |
| 0.75 | | 3,12 | 62,5 | 0,01 | |
| 0.10 | 0.4 | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | 0.6 | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | 0.8 | - | - | - | |

| | | | | | |
|---|---|---|---|---|---|
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | 0.2 | 1,56 | 92,7 | 0,01 | Regression Imputation |
| 0.25 | | 1,76 | 88 | -0,01 | |
| 0.5 | | 2,22 | 78,2 | -0,03 | |
| 0.75 | | 3,11 | 63,6 | 0,04 | |
| 0.10 | 0.4 | 1,49 | 93,7 | 0,03 | |
| 0.25 | | 1,73 | 88,9 | 0,03 | |
| 0.5 | | 2,14 | 80,9 | -0,02 | |
| 0.75 | | 2,82 | 69,2 | 0,01 | |
| 0.10 | 0.6 | 1,5 | 93,5 | 0,02 | |
| 0.25 | | 1,63 | 90,9 | 0,02 | |
| 0.5 | | 2,01 | 83,3 | -0,02 | |
| 0.75 | | 2,54 | 73,2 | 0,03 | |
| 0.10 | 0.8 | 1,45 | 93,9 | 0,02 | |
| 0.25 | | 1,56 | 92,2 | 0,02 | |
| 0.5 | | 1,77 | 87,7 | -0,02 | |
| 0.75 | | 2,08 | 81,5 | 0,02 | |
| 0.10 | 0.2 | 1,56 | 91,9 | -0,01 | Multiple Imputation |
| 0.25 | | 1,85 | 85,8 | -0,05 | |
| 0.5 | | 2,47 | 74,3 | -0,05 | |
| 0.75 | | 3,7 | 54,5 | 0,01 | |
| 0.10 | 0.4 | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | 0.6 | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | 0.8 | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |

MAR

| Nonresponse | Cor | RRMSE | CI-Rate | R-Bias | Correction Method |
|---|---|---|---|---|---|
| 0.10 | 0.2 | 1,7 | 91,4 | -0,35 | Complete Case |
| 0.25 | | 2,57 | 77,7 | -0,87 | |
| 0.5 | | 4,15 | 54,8 | -1,62 | |
| 0.75 | | 6,15 | 51,9 | -2,45 | |
| 0.10 | 0.4 | 2,21 | 80,6 | -0,7 | |

| | | | | | |
|---|---|---|---|---|---|
| 0.25 | | 4,41 | 34,8 | -1,7 | |
| 0.5 | | 7,71 | 0,03 | -3,34 | |
| 0.75 | | 11,73 | 0,02 | -5,09 | |
| 0.10 | 0.6 | 2,8 | 63,9 | -1,07 | |
| 0.25 | | 5,9 | 0,05 | -2,54 | |
| 0.5 | | 11,23 | 0 | -4,95 | |
| 0.75 | | 17,28 | 0 | -7,63 | |
| 0.10 | 0.8 | 3,51 | 38,7 | -1,43 | |
| 0.25 | | 7,78 | 0 | -3,4 | |
| 0.5 | | 14,89 | 0 | -6,6 | |
| 0.75 | | 22,98 | 0 | -10,2 | |
| 0.10 | 0.2 | 1,64 | 89,7 | -0,34 | Mean-Imputation |
| 0.25 | | 2,43 | 61,4 | -0,82 | |
| 0.5 | | 4,2 | 17,5 | -1,66 | |
| 0.75 | | 6,26 | 0,07 | -2,51 | |
| 0.10 | 0.4 | 2,17 | 75,3 | -0,71 | |
| 0.25 | | 4,14 | 16,7 | -1,71 | |
| 0.5 | | 7,71 | 0 | -3,34 | |
| 0.75 | | 11,73 | 0 | -5,09 | |
| 0.10 | 0.6 | 2,8 | 55,1 | -1,07 | |
| 0.25 | | 5,9 | 0,2 | -2,54 | |
| 0.5 | | 11,23 | 0 | -4,97 | |
| 0.75 | | 17,28 | 0 | -7,63 | |
| 0.10 | 0.8 | 3,52 | 32,9 | -1,43 | |
| 0.25 | | 7,78 | 0 | -3,4 | |
| 0.5 | | 14,89 | 0 | -6,6 | |
| 0.75 | | 22,93 | 0 | -10,18 | |
| 0.10 | 0.2 | 1,74 | 88,9 | -0,34 | Random Imputation |
| 0.25 | | 2,69 | 65,9 | -0,88 | |
| 0.5 | | 4,23 | 33,9 | -1,61 | |
| 0.75 | | 6,55 | 16,7 | -2,57 | |
| 0.10 | 0.4 | 2,24 | 76,9 | -0,7 | |
| 0.25 | | 4,18 | 29,3 | -1,69 | |
| 0.5 | | 7,64 | 1,8 | -3,27 | |
| 0.75 | | 11,93 | 0,2 | -5,16 | |
| 0.10 | 0.6 | 2,89 | 56,9 | -1,09 | |
| 0.25 | | 5,94 | 3,8 | -2,53 | |
| 0.5 | | 11,46 | 0 | -5,01 | |
| 0.75 | | 17,42 | 0 | -7,67 | |
| 0.10 | 0.8 | 3,56 | 38 | -1,43 | |
| 0.25 | | 7,86 | 0 | -3,42 | |
| 0.5 | | 14,84 | 0 | -6,56 | |
| 0.75 | | 23,1 | 0 | -10,25 | |
| 0.10 | 0.2 | 1,58 | 91,9 | 0 | Regression Imputation |
| 0.25 | | 1,79 | 85,9 | -0,01 | |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 0.5 |  | 2,33 | 77,6 | 0,02 |  |
| 0.75 |  | 3,57 | 55,8 | -0,11 |  |
| 0.10 | 0.4 | 1,47 | 94 | 0 |  |
| 0.25 |  | 1,78 | 88,5 | -0,01 |  |
| 0.5 |  | 2,22 | 80,5 | -0,01 |  |
| 0.75 |  | 3,32 | 59 | 0,03 |  |
| 0.10 | 0.6 | 1,48 | 93,3 | 0,02 |  |
| 0.25 |  | 1,68 | 90,1 | -0,02 |  |
| 0.5 |  | 1,98 | 84,2 | 0,04 |  |
| 0.75 |  | 2,89 | 66,8 | 0,02 |  |
| 0.10 | 0.8 | 1,51 | 93,9 | 0,01 |  |
| 0.25 |  | 1,51 | 93,7 | 0,04 |  |
| 0.5 |  | 1,72 | 88,8 | 0 |  |
| 0.75 |  | 2,25 | 78,3 | 0 |  |
| 0.10 | 0.2 | 1,57 | 93,1 | -0,01 | Multiple Imputation |
| 0.25 |  | 1,83 | 87,2 | -0,05 |  |
| 0.5 |  | 2,83 | 67,4 | -0,07 |  |
| 0.75 |  | 4,83 | 44,5 | -0,22 |  |
| 0.10 | 0.4 | 1,53 | 92,8 | 0,01 |  |
| 0.25 |  | 1,86 | 86,8 | 0 |  |
| 0.5 |  | 2,56 | 71,9 | -0,04 |  |
| 0.75 |  | 4,02 | 51,3 | -0,2 |  |
| 0.10 | 0.6 | 1,46 | 94,3 | -0,02 |  |
| 0.25 |  | 1,69 | 90,2 | -0,03 |  |
| 0.5 |  | 2,32 | 74,8 | -0,12 |  |
| 0.75 |  | 3,77 | 52,4 | -0,48 |  |
| 0.10 | 0.8 | 1,47 | 94,3 | -0,04 |  |
| 0.25 |  | 1,56 | 91,7 | -0,08 |  |
| 0.5 |  | 1,93 | 85 | -0,15 |  |
| 0.75 |  | 3,11 | 59,1 | -0,51 |  |

MNAR

| Nonresponse | Cor | RRMSE | CI-Rate | R-Bias | Correction Method |
|---|---|---|---|---|---|
| 0.10 | 0.2 | 4,25 | 20,1 | -1,78 | Complete Case |
| 0.25 |  | 9,64 | 0 | -4,25 |  |
| 0.5 |  | 18,49 | 0 | -8,22 |  |
| 0.75 |  | 28,64 | 0 | -12,77 |  |
| 0.10 | 0.4 | 4,32 | 19,2 | -1,81 |  |
| 0.25 |  | 9,67 | 0 | -4,26 |  |
| 0.5 |  | 18,62 | 0 | -8,28 |  |
| 0.75 |  | 28,68 | 0 | -12,77 |  |
| 0.10 | 0.6 | 4,27 | 21,2 | -1,79 |  |
| 0.25 |  | 9,66 | 0 | -4,26 |  |
| 0.5 |  | 18,55 | 0 | -8,25 |  |

| | | | | | |
|---|---|---|---|---|---|
| 0.75 | | 28,59 | 0 | -12,73 | |
| 0.10 | 0.8 | 4,23 | 20,2 | -1,79 | |
| 0.25 | | 9,78 | 0 | -4,31 | |
| 0.5 | | 18,54 | 0 | -8,25 | |
| 0.75 | | 28,6 | 0 | -12,73 | |
| 0.10 | | 4,25 | 15,5 | -1,78 | Mean Imputation |
| 0.25 | | 9,64 | 0 | -4,25 | |
| 0.5 | | 18,49 | 0 | -8,23 | |
| 0.75 | | 28,7 | 0 | -12,77 | |
| 0.10 | | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | 0.2 | 4,38 | 18,8 | -1,82 | Random Imputation |
| 0.25 | | 9,62 | 0 | -4,26 | |
| 0.5 | | 18,57 | 0 | -8,25 | |
| 0.75 | | 28,9 | 0 | -12,86 | |
| 0.10 | 0.4 | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | 0.6 | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | 0.8 | - | - | - | |
| 0.25 | | - | - | - | |
| 0.5 | | - | - | - | |
| 0.75 | | - | - | - | |
| 0.10 | 0.2 | 4,06 | 23,1 | -1,69 | Regression Imputation |
| 0.25 | | 9,17 | 0 | -4,02 | |
| 0.5 | | 17,55 | 0 | -7,79 | |
| 0.75 | | 27,01 | 0 | -12,01 | |
| 0.10 | 0.4 | 3,52 | 37,2 | -1,41 | |
| 0.25 | | 7,76 | 0,3 | -3,39 | |
| 0.5 | | 14,94 | 0 | -6,62 | |
| 0.75 | | 23,06 | 0 | -10,25 | |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 0.10 | 0.6 | 2,75 | 59,6 | -1,03 |  |
| 0.25 |  | 5,8 | 3,6 | -2,49 |  |
| 0.5 |  | 11,26 | 0 | -4,97 |  |
| 0.75 |  | 17,23 | 0 | -7,62 |  |
| 0.10 | 0.8 | 1,84 | 86,8 | -0,54 |  |
| 0.25 |  | 3,4 | 39,9 | -1,36 |  |
| 0.5 |  | 6,3 | 2 | -2,7 |  |
| 0.75 |  | 9,55 | 0,1 | -4,15 |  |
| 0.10 | 0.2 | 4,05 | 24,8 | -1,66 | Multiple Imputation |
| 0.25 |  | 9,32 | 0 | -4,09 |  |
| 0.5 |  | 17,67 | 0 | -7,85 |  |
| 0.75 |  | 27,11 | 0 | -12,01 |  |
| 0.10 | 0.4 | - | - | - |  |
| 0.25 |  | - | - | - |  |
| 0.5 |  | - | - | - |  |
| 0.75 |  | - | - | - |  |
| 0.10 | 0.6 | - | - | - |  |
| 0.25 |  | - | - | - |  |
| 0.5 |  | - | - | - |  |
| 0.75 |  | - | - | - |  |
| 0.10 | 0.8 | - | - | - |  |
| 0.25 |  | - | - | - |  |
| 0.5 |  | - | - | - |  |
| 0.75 |  | - | - | - |  |

# References

Alfons, Andreas; Kraft, Stefan; Templ, Matthias; Filzmoser, Peter (2011): Simulation of close-to-reality population data for household surveys with application to EU-SILC. In: Stat Methods Appl 20 (3), p. 383–407.

Barthelemy, Johan; Toint, Philippe L. (2013): Synthetic Population Generation Without a Sample. In: Transportation Science 47 (2), p. 266–279.

Beckman, Richard J.; Baggerly, Keith A.; McKay, Michael D. (1996): Creating synthetic baseline populations. In: Transportation Research Part A: Policy and Practice 30 (6), p. 415–429.

Bethlehem, Jelke G.; Cobben, Fannie; Schouten, Barry (Hg.) (2011): Handbook of nonresponse in household surveys. Hoboken, NJ: Wiley (Wiley series in survey methodology).

Brand, J.P.L. (1999). Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets (pp. 110-113). Dissertation. Rotterdam: Erasmus University.

Brick, J. Michael (2013): Unit Nonresponse and Weighting Adjustments: A Critical Review. In: Journal of Official Statistics 29 (3), p. 329–353.

Brick, J. Michael; Williams, Douglas (2013): Explaining Rising Nonresponse Rates in Cross-Sectional Surveys. In: The Annals of the American Academy of

Political and Social Science 645 (1), p. 36–59.

Campanelli, Pamela (2009): Testing Survey Questions . In: Leeuw, Edith Desirée de; Hox, Joop J.; Dillman, Don A. (Hg.): International handbook of survey methodology. Repr. New York, NY: Psychology Press. Wiley Series in Probability and Statistics, p. 176-200.

Chen, Sixia; Haziza, David (2019): Recent Developments in Dealing with Item Non-response in Surveys: A Critical Review. In: International Statistical Review 87, p. 192-218.

Demirtas, H., S. A. Freels, and R. M. Yucel. 2008. "Plausibility of Multivariate Normality Assumption When Multiply Imputing Non-Gaussian Continuous Outcomes: A Simulation Assessment." Journal of Statistical Computation and Simulation 78 (1): 69–84.

Earp, Morgan; Mitchell, Melissa; McCarthy, Jaki; Kreuter, Frauke (2014): Modeling Nonresponse in Establishment Surveys: Using an Ensemble Tree Model to Create Nonresponse Propensity Scores and Detect Potential Bias in an Agricultural Survey. In: Journal of Official Statistics 30 (4), p. 701–719.

Graham, John W. (2009): Missing data analysis: making it work in the real world. In: Annual review of psychology 60, p. 549–576.

Groves, Robert M. (2006): Nonresponse Rates and Nonresponse Bias in Household Surveys. In: Public Opinion Quarterly 70 (5), p. 646–675.

Gustavson, Kristin; Røysamb, Espen; Borren, Ingrid (2019): Preventing bias from selective non-response in population-based survey studies: findings from a Monte Carlo simulation study. In: BMC medical research methodology 19 (1), p. 120.

Horvitz, D. G.; Thompson, D. J. (1952): A Generalization of Sampling Without Replacement From a Finite Universe. In: Journal of the American Statistical Association 47 (260), p. 663-685.

Hox, Joop J. (2009): Accommodating measurement errors In: Leeuw, Edith Desirée de; Hox, Joop J.; Dillman, Don A. (Hg.): International handbook of survey methodology. Repr. New York, NY: Psychology Press. Wiley Series in Probability and Statistics, p. 387-402.

Kalton, Graham; Flores-Cervantes Ismael (2003): Weighting Methods. In: Journal of Ofcial Statistics 19 (2), p. 81–97.

Kreuter, Frauke (2013): Facing the Nonresponse Challenge. In: The Annals of the American Academy of Political and Social Science, p. 23–35.

Leeuw, Edith D. de (2001): Reducing Missing Data in Surveys: An overview of methods. In: Quality and Quantity 35 (2), p. 147–160.

Li, Lingling; Shen, Changyu; Li, Xiaochun; Robins, James M. (2013): On weighting approaches for missing data. In: Statistical methods in medical research 22 (1), S. 14–30

Little, Roderick J. A. (1988). Missing-data adjustments in large surveys. J. Bus. Econom. Statist. 6 287–296.

Little, Roderick J. A.; Rubin, Donald B. (2002): Statistical analysis with missing data. 2nd ed. Hoboken: Wiley (Wiley Series in Probability and Statistics).

Lohr, Sharon, L. (2009): Coverage and Sampling. In: Leeuw, Edith Desirée de; Hox, Joop J.; Dillman, Don A. (Hg.): International handbook of survey methodology. Repr. New York, NY: Psychology Press. Wiley Series in Probability and Statistics, p. 97-112.

Matei, Alina, Ranalli, Giovanna M. (2015): Dealing with non-ignorable nonresponse in survey sampling: a latent modeling approach. In: Survey Methodology (41) 1, p. 145-162.

Murray, Jared S. (2018): Multiple Imputation: A Review of Practical and Theoretical Findings. In: Statist. Sci. 33 (2), p. 142–159.

Münnich, Ralf; Schürle, Josef (2003): On the Simulation of Complex Universes in the Case of Applying the German Microcensus. In: DACSEIS Research Paper Series No. 4.

Nishimura, Raphael; Wagner, James; Elliott, Michael R. (2016): Alternative indicators for the risk of non-response bias: a simulation study. In: International statistical review = Revue internationale de statistique 84 (1), p. 43–62.

R Development Core Team (2019): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rubin, Donald B. (1976): Inference and Missing Data. In: Biometrika 63 (3), p. 581-592.

Rubin, Donald B. (2009): Multiple Imputation for Nonresponse in Surveys. 99th ed. Hoboken: John Wiley  Sons Inc.

Särndal, Carl-Erik; Thomsen, Ib; Hoem, Jan. M.; Lindley, D. V.; Barndorff-Nielsen, O.; Dalenius, Tore (1978): Design-Based and Model-Based Inference in Survey Sampling [with Discussion and Reply]. In: Scandinavian Journal of Statistics 5 (1), p. 27–52.

Särndal, Carl-Erik; Lundström, Sixten (2006): Estimation in surveys with nonresponse. Repr. Chichester: Wiley (Wiley series in survey methodology).

Schafer, Joseph L.; Graham, John W. (2002): Missing data: Our view of the state of the art. In: Psychological Methods 7 (2), p. 147–177.

Schouten, Rianne Margaretha; Lugtig, Peter; Vink, Gerko (2018): Generating missing values for simulation purposes: a multivariate amputation procedure. In: Journal of Statistical Computation and Simulation 88 (15), p. 2909–2930.

Sullivan, Danielle; Andridge, Rebecca (2015): A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck. In: Computational Statistics  Data Analysis 82, p. 173–185.

Templ, Matthias; Meindl, Bernhard; Kowarik, Alexander; Dupriez, Olivier (2017): Simulation of Synthetic Complex Data: The R Package simPop. In: J. Stat. Soft. 79 (10).

Van Buuren, Stef; Groothuis-Oudshoorn, Karin (2011): mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL https://www.jstatsoft.org/v45/i03/.

Van Buuren, Stef (2018): Flexible imputation of missing data. Second edition. Boca Raton, London, New York: Chapman and Hall/CRC (Chapman and Hall/CRC Interdisciplinary statistics series).

Vink, G. (2016). Towards a standardized evaluation of multiple imputation routines.

Venables, W. N.; Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York.

Yan, T.; Curtin, R. (2010): The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective. In: International Journal of Public Opinion Research 22 (4), p. 535–551.

# ERKLÄRUNG ZUR BACHELORARBEIT / MASTERARBEIT

Hiermit erkläre ich, dass ich die Bachelorarbeit / Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die aus fremden Quellen direkt oder indirekt übernommenen Gedanken als solche kenntlich gemacht habe.

Die Arbeit habe ich bisher keinem anderen Prüfungsamt in gleicher oder vergleichbarer Form vorgelegt. Sie wurde bisher nicht veröffentlicht.

______________________________ ______________________________

Datum Unterschrift