

UNIVERSITY OF TARTU
FACULTY OF SCIENCE AND TECHNOLOGY
Institute of Mathematics and Statistics
Mathematical Statistics Curriculum

Birgit Kadastik

Missing data in clinical trials

Bachelor's Thesis (9 ECTS)

Supervisors: Marju Valge, *MSc*

Pasi Korhonen, *PhD*

Tartu 2016

Missing data in clinical trials

Abstract:

The aim of this Bachelor's Thesis is to explain what missing data means and give some ways to deal with it in clinical trials. Firstly, an overview of different types of missing data is given and the reasons for their occurrence. Second part of the thesis explains which analytical approaches can be used to conduct an unbiased analysis. Further, missing data are simulated for a data set to show how the approaches described are used in practice with SAS software.

Keywords:

Clinical trials, complete case analysis, missing at random, missing completely at random, missing data, multiple imputation, SAS

P160 Statistics, operation research, programming, actuarial mathematics

Puuduvad andmed kliinilistes uuringutes

Lühikokkuvõte:

Käesoleva bakalaureusetöö eesmärgiks on kirjeldada puuduvaid andmeid ja nendega tegelemise meetodeid kliiniliste uuringute kontekstis. Esimeses peatükis antakse ülevaade erinevatest puudumise struktuuridest ja põhjustest. Töö teises osas seletatakse analüütilisi meetodeid, millega on võimalik teostada nihketa analüüsi. Viimases peatükis genereeritakse olemasolevasse andmestikku puuduvaid väärtusi, et näidata, kuidas eespool kirjeldatud meetodeid rakendustarkvaras SAS kasutada.

Võtmesõnad:

Kliinilised uuringud, täielike andmetega analüüs, juhuslik puudumine, täiesti juhuslik puudumine, puuduvad andmed, mitmene asendamine, SAS

P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Contents

- Introduction 5
- 1. Background of missing data 7
 - 1.1 Reasons for missing data 7
 - 1.2 Consequences of missing data 7
 - 1.3 Avoiding missing data 8
 - 1.4 Notation 9
 - 1.5 Types of missing data 10
- 2. Approaches for dealing with missing data 12
 - 2.1 Approaches for data MCAR 12
 - 2.1.1 Complete case analysis (listwise deletion)..... 12
 - 2.1.2 Available case analysis (pairwise deletion)..... 13
 - 2.1.3 Single imputation 15
 - 2.1.3.1 Unconditional mean imputation 15
 - 2.1.3.2 Conditional mean imputation or Buck’s method (regression)..... 16
 - 2.1.3.3 Last observation carried forward 18
 - 2.1.4 Hot deck 19
 - 2.2 Approaches for data MAR..... 21
 - 2.2.1 Inverse probability weighting 21
 - 2.2.2 Multiple imputation..... 22
 - 2.2.3 Likelihood-based analysis 24
- 3. Practical 25
 - 3.1 Overview of the original data 25
 - 3.2 Missing completely at random 26
 - 3.2.1 SAS programs 26
 - 3.2.1.1 Complete case analysis 27

3.2.1.2	Available case analysis	27
3.2.1.3	Unconditional mean imputation	28
3.2.1.4	Conditional mean imputation	29
3.2.1.5	Last observation carried forward	29
3.2.1.6	Hot deck.....	29
3.2.2	Results	30
3.2.2.1	10% missing data.....	30
3.2.2.2	25% missing data.....	31
3.2.2.3	50% missing data.....	33
3.3	Missing at random	35
3.3.1	SAS codes	35
3.3.1.1	Inverse probability weighting.....	35
3.3.1.2	Multiple imputation	36
3.3.1.3	Likelihood-based analysis	37
3.3.2	Results	38
3.3.2.1	10% missing data.....	38
3.3.2.2	25% missing data.....	39
3.3.2.3	50% missing data.....	40
3.4	MCAR analysis methods with MAR data set.....	41
	Conclusion.....	44
	References	45
	Appendices	47

Introduction

This bachelor thesis is written as a guide for a company named StatFinn Oy. The aim of the thesis is to give instructions on how to deal with missing data in clinical trials and explain how each specific missing data method can be implemented using SAS software.

Clinical trials are investigations in human subjects (participants of a clinical trial) to discover or verify effects of experimental treatments. Clinical trial's rationale, background, objectives, design, methodology and statistical considerations are described in a document called protocol. Subjects are usually divided into treatment group in which they receive experimental treatment and control group where they receive no treatment (placebo) or standard (previously available) treatment. The main goal is to prove efficacy (maximum response achievable from the treatment) and to estimate treatment effect which is usually obtained from a comparison of a specific outcome variable between two or more treatments. [1]

In clinical trials, it is important to get all the necessary information about subjects to conduct a thorough and unbiased analysis. But often, when working with human subjects, the data sets are incomplete and include missing data which are defined as values that are not available and would be meaningful for analysis if they were observed [2]. The degree of data incompleteness can be different, e.g. only baseline measurements can be available, or missingness can occur at baseline, measurements may be missing for one, several or all follow-up evaluations [3].

This bachelor thesis consists of three chapters. The first chapter gives general information about the nature of missing data. The author explains the reasons for missing data occurrence, why it is a problem that needs to be dealt with and how to avoid it. Different types of missing data mechanisms are also presented with their definitions and examples.

The second chapter explains which analytical approaches can be used to conduct an unbiased analysis. Explanation and general idea of each method is given, the strengths and weaknesses are also emphasized. In addition, when possible, it is shown how to use these methods on simplified examples without any programs.

In the last chapter, chapter 3, theory is put into practise. For each missing data analytical method SAS codes are presented and their use is explained based on a real data set. Missing data were generated by the author and results are given with proportion of missingness set at 10%, 25% and 50% with each different missing data mechanisms. In the last section it is also shown what are the results if an incorrect assumption about the missingness mechanism is made.

Thesis is written in Microsoft Office Word 2016 and statistical analyses are conducted in SAS software (version 9.4).

The author would like to thank supervisors Marju Valge and Pasi Korhonen for advice and suggestions.

1. Background of missing data

1.1 Reasons for missing data

Data might be missing for several reasons. In clinical trials, one of the reasons for missing data is a protocol violation (serious non-compliance with the protocol), for example subjects do not meet the inclusion/exclusion criteria or they use another medication that is prohibited in the protocol. Subjects can also drop out because of adverse events (an untoward medical occurrence that might or might not be related to treatment), lack of efficacy or illness that is not related to the study medication. [1] [2] [3]

In addition, data can be incomplete due to the lack of competence of the researcher or other study team members, e.g. the study nurse, lab personnel. There might be mistakes made in the data collection or in data entry. Researchers can also violate the protocol by mishandling the samples.

1.2 Consequences of missing data

This chapter is based on [3] if not mentioned otherwise.

The amount of missing data can affect the validity (Estonian *valiidsus*) of the clinical trial. If the losses to follow-up are less than 5% then the impact is likely not to be substantial, concerns about the validity rise when the losses are greater than 20% [4]. When the proportion of missing data is significant then it can affect the conclusions about the different treatments being studied, i.e. it might be impossible to conclude that evidence of efficacy has been established.

Missing values also serve as potential source of bias in clinical trials. The exclusion of subjects may influence comparability of the treatment groups which, in turn, leads to bias in the estimation of the treatment effect. It might also have an impact on the external validity that is the representativeness of the study sample in connection to the target population. The danger of bias relies upon the relationship between missingness, treatment and outcome. Those relationships can affect the bias differently:

- If the missing values are not connected to the actual value of the unobserved measurement then they will not be anticipated to lead to bias (for example, poor and good outcomes have the same likelihood of being missing).

- The estimate of the treatment can be biased if the unavailable observation is connected to the real value of the outcome (for example, mostly poor outcomes are missing), even if the missing values are not related to the treatment.
- If the missing observations are associated to both treatment and the unobserved outcome variable (e.g. missingness occurs more often in one of the treatment arms), then ignoring them will lead to bias.

One way of dealing with missing data is to completely exclude subjects who have values that are absent, therefore decreasing the sample size which in return will influence the statistical power (Estonian (*testi*) *võimsus*). The power of the trial will increase if the variability of the missing outcomes is reduced or if the sample size is increased. Consequently, the greater the number of missing values, the greater is the reduction in power.

Mishandling the missing data can also impact the confidence intervals. Excluding non-completers with extreme values (e.g. noticeably good or bad response before loss to follow-up) may lead to underestimate of variability which therefore narrows the confidence interval for the treatment effect.

1.3 Avoiding missing data

Although there are several approaches to deal with missing data (given later), the best way is to prevent it in the study design and conduct period. It can be useful to predict the anticipated proportions of missing data in the design phase because it can influence the variability and required sample size and also it might be helpful for managing the range of sensitivity analyses that are necessary. [4]

Clearly defined target population, along with efficacy and safety outcomes, and the analysis of the likely effects of missing data are attributes of a good clinical trial design. Researchers should target a population that has an incentive to stay in the study, for example because it is not sufficiently served by current treatment. [2] More, the study design should limit the burden of unnecessary data collection for the study participants. This can be accomplished by:

- reducing the number of follow-up visits;
- gathering only vital information at each visit;
- making case-report forms (document that records all protocol required information on each trial subject [1]) user-friendly;
- if attainable, using data capture that does not require clinic visits;

- shortening the follow-up period for the primary outcome as appropriate. [2][4][5]

The approaches to minimize the missing data in trial planning and conduct are aimed at the participant, the data collection process and the study team [4]. Firstly, incentives can be offered to participants. These can include payment for the number of finished visits rather than payment for each subject. In fact, monetary incentives for voluntary participation in a clinical trial are considered ethical. Secondly, it is important to engage participants to make them feel included and appreciated for their exertion, especially those who are at higher risk for dropout, for example by including study-branded gifts, constant expressions of gratitude and enjoyable experience at study visits. In addition, the trial conduct phase may be facilitated by a reminder system, which helps subjects to understand the commitment to the trial and record the reasons for withdrawal to help in the interpretation of the results. [5]

Approaches concerning the data collection process involve careful selection of study sites, training of the site personnel to ensure they know the importance of complete data collection, and structure for proficient communication among the study teams. Also, databases where data are inserted by the site personnel can have regulations, e.g. system gives a warning when a field is empty or inaccurate (for example, height is 1500 m). In addition, mandatory fields can be added. [4]

Furthermore, regular team gatherings or web-based discussion boards allow a chance to find a solution to a possible missing data issue [4].

1.4 Notation

Let the intended data be denoted by a $n \times p$ matrix \mathbf{Y} which is partitioned into $\mathbf{Y} = \{\mathbf{Y}_o, \mathbf{Y}_m\}$ where \mathbf{Y}_o represents observed and \mathbf{Y}_m represents missing part of the data matrix. Missing value indicator matrix \mathbf{R} ($n \times p$) that is corresponding to \mathbf{Y} is defined as

$$r_{i,j} = \begin{cases} 1 & \text{if } y_{i,j} \text{ is observed} \\ 0 & \text{if } y_{i,j} \text{ is missing} \end{cases}$$

where $i = 1, \dots, n$ and $j = 1, \dots, p$. [6]

1.5 Types of missing data

Before making any decisions about suitable approaches to deal with missing data, it is necessary to evaluate how the missing data may have occurred. There are three different categories of how missingness has developed.

Data are missing completely at random (MCAR, Estonian *täiesti juhuslik puudumine*) when the probability of an observation being missing is unrelated to any unobserved or observed variables. In mathematical terms it is written as

$$P(\mathbf{R} | \mathbf{Y}_o, \mathbf{Y}_m) = P(\mathbf{R}) [7].$$

It means that the probability of missing data is the same for all subjects, regardless of treatment received, treatment response or any other observed or unobserved aspect in \mathbf{Y} [4]. The assumption of MCAR assumes that data from participants with missing data can be disregarded without bias because their outcomes would be anticipated to be similar to outcomes of subjects whose data were completely observed [2]. Examples for mechanisms yielding to MCAR include migration, random failure of instruments (or laboratory sample is dropped), termination of follow-up due to administrative end of study and more [4].

Data are missing at random (MAR, Estonian *juhuslik puudumine*) when the likelihood of missing data depends on observed variables but not on unobserved variables [4]. Mathematically,

$$P(\mathbf{R} | \mathbf{Y}_o, \mathbf{Y}_m) = P(\mathbf{R} | \mathbf{Y}_o) [7].$$

In other words, if subjects share similar observed values, the statistical behaviour on their other observations would be similar, whether observed or not [2]. MAR assumption indicates that if the baseline characteristics and intermediate measures are similar for dropouts and completers, then the outcomes would be expected to be similar for both, therefore the missing outcomes can be modelled on the basis of completers' outcomes [2]. Subjects can drop out due to recorded side effects or known baseline features or absence of efficacy [4].

Data are considered missing not at random (MNAR, Estonian *mittejuhuslik puudumine*) when the missing data depends on the unobserved data [4]. It can be written as

$$P(\mathbf{R} | \mathbf{Y}_o, \mathbf{Y}_m) = P(\mathbf{R} | \mathbf{Y}_o, \mathbf{Y}_m) [7].$$

This missingness mechanism is also called non-ignorable because results will be biased if the process that leads to missing data is ignored. The assumption for MNAR implies that the decision to drop out can be based on events that were not observed, so outcomes for dropouts

are different from participants who have similar characteristics. [2] Examples of MNAR are dropout based on the unobserved response (if a person is not responding to treatment) and missed visits due to the fact that subjects have had an outcome (e.g. hospitalisation, significant improvements in the state of disease) already [4].

In this thesis, the author only explains approaches which deal with MCAR and MAR.

2. Approaches for dealing with missing data

2.1 Approaches for data MCAR

2.1.1 Complete case analysis (listwise deletion)

Complete case (CC, Estonian *täielike andmetega analüüs*) analysis includes only those participants who have all the measurements recorded [7]. Subjects who have missing observations are excluded from the analysis and standard methods are used on the remaining set of subjects. This approach is valid only when the missing data are MCAR, otherwise it may lead to biased results. [4]

CC method is simple to describe and use, since common statistical tests are applied. Additionally, it gives a common basis for conclusions (despite the type of the analysis) because the estimates are calculated on the same subset of completers. [7]

The main disadvantage of CC method is that it causes severe bias if the missingness mechanism is MAR or MNAR instead of MCAR because completely recorded cases are not usually representative of the whole sample. For example, in trials conducted to examine prevention of drug abuse, users are more likely to drop out than non-users, therefore completers do not represent the original sample, leading to bias in the parameters. Also, because of erasing some subjects and their information, the estimators might be inefficient. In addition, this method impacts the precision and power by reducing them. [7]

Although complete case analysis is easy to use, it is not a recommended approach due to the disadvantages described above.

Example

Systolic blood pressure was measured for five subjects; the results are presented in Table 2.1.

Table 2.1 Measured systolic blood pressure with missing values

Subject	Systolic blood pressure (mm Hg)
1	115
2	150
3	?
4	125
5	?

In complete case analysis, subjects who have missing observations are removed. Therefore, the final data set to be analysed would consist of subjects 1, 2 and 4 (displayed in Table 2.2).

Table 2.2 Systolic blood pressure measurements for complete case analysis

Subject	Systolic blood pressure (mm Hg)
1	115
2	150
4	125

2.1.2 Available case analysis (pairwise deletion)

Available case analysis (Estonian *tunnuspaari analüüis*) or pairwise deletion is an approach to deal with missing data that attempts to minimize the loss that usually occurs in complete case scenario. It mainly focuses on the covariance (or correlation) matrix. For each pair of variables which have valid data, the correlation is calculated. For the variable that has no missing data, denoted by X , all cases are used to calculate the mean and standard deviation. Mean (\bar{y}) and standard deviation (s_y) of variable with missing observations, denoted by Y , are calculated based on complete cases. The correlation between X and Y is then calculated as

$$r_{xy}^2 = \frac{1}{m-1} \frac{\sum_{i=1}^m (x_i - \bar{x}_{(m)})(y_i - \bar{y})}{s_{x(m)}s_y}$$

where $\bar{x}_{(m)}$ and $s_{x(m)}$ are the mean and standard deviation of X calculated from the m complete cases. Estimated correlation (or covariance) matrix is used as an input for methods like regression. [8]

Like in complete case analysis, estimated parameters will be unbiased only if the missingness is MCAR. Because this method uses all the data available, it does not decrease power as much as complete case analysis. Unfortunately, there is no apparent way to specify the sample size for this method, therefore making it hard to estimate the standard errors. [6]

Example

Weight and height were measured for five patients (in Table 2.3).

Table 2.3 Measured weight and height

Subject	Weight (kg)	Height (cm)
1	65	170
2	55	165
3	90	?
4	69	173
5	100	?

Firstly, the means are calculated. The mean weight is $\frac{65+55+90+69+100}{5} = 75.8$ kg and the mean height is $\frac{170+165+173}{3} = 169.33$ cm.

Secondly, standard deviations are found. The standard deviations for weight and height are

$$\sqrt{\frac{(65-75.8)^2+(55-75.8)^2+(90-75.8)^2+(69-75.8)^2+(100-75.8)^2}{5-1}} = \sqrt{345.7} = 18.59 \text{ kg}$$

and $\sqrt{\frac{(170-169.33)^2+(165-169.33)^2+(173-169.33)^2}{3-1}} = \sqrt{16.33} = 4.04$ cm, respectively. The mean

and standard deviation of weight from full data are calculated to summarize weight but they are not used for correlation calculations.

To calculate the correlation between weight and height the means and standard deviations over complete cases are calculated. The mean weight is then $\frac{65+55+69}{3} = 63$ kg and weight's standard

deviation is $\sqrt{\frac{(65-63)^2+(55-63)^2+(69-63)^2}{3-1}} = \sqrt{52} = 7.21$ kg. Then the correlation is

$$\frac{1}{3-1} \frac{(65-63)(170-169.33)+(55-63)(165-169.33)+(69-63)(173-169.33)}{7.21*4.04} = 0.995.$$

2.1.3 Single imputation

2.1.3.1 Unconditional mean imputation

In unconditional mean imputation (Estonian *keskväärtusega asendamine*) method, missing values are replaced with the average of the observed values on the same variable over other subjects. The method is called *unconditional* because it does not use other information that the subject with missing data has. [7] This method results in underestimation of variability which is proportional to the fraction of missing data because a constant is imputed for all of the subjects with missing data, regardless of their personal characteristics [4][6]. The bias in variability is proportional to $(n_o - 1)/(n_o + n_m - 1)$ if the missingness mechanism is MCAR, where n_o is the number of subjects having the value of a specific variable observed and n_m is the number of subjects having the value of a specific variable missing. The covariances, which are biased by similar factor, and variances will hence be underestimated because the unconditional mean imputation for missing cases has a variance of 0. [6].

Example

Five subjects were measured to find out their height and weight. The resulting measurements are presented in Table 2.4 below.

Table 2.4 Measured height and weight with missing values

Subject	Height (cm)	Weight (kg)
1	185	90
2	170	60
3	156	?
4	198	120
5	?	55

As can be seen from Table 2.4, one subject (subject number 3) is missing his/her weight and another one (subject number 5) his/her height. For unconditional mean imputation method, the mean for height ($\frac{185+170+156+198}{4} = 177.25$ cm) and the mean for weight (81.25 kg) are calculated based on the available data. Imputing the means for missing values leads to the following data set (presented in Table 2.5).

Table 2.5 Height and weight after unconditional mean imputation

Subject	Height (cm)	Weight (kg)
1	185	90
2	170	60
3	156	81.25
4	198	120
5	177.25	55

The mean height in the final data is $\frac{185+170+156+198+177.25}{5} = 177.25$ cm and the mean weight is $\frac{90+60+81.25+120+55}{5} = 81.25$ kg. In case of unconditional mean imputation, the means do not change, as can be seen also from the example.

2.1.3.2 Conditional mean imputation or Buck's method (regression)

This section is based on [7].

Conditional mean imputation (Estonian *lineaarsete prognoosidega asendamine*), known also as Buck's method or regression-based imputation, uses available information about the subject with missing data when imputing missing values. The method first estimates the mean μ and covariance matrix Σ based on the complete cases. Then these estimates are used to calculate the linear regression of the incomplete variable on the other variables. In the second step the conditional mean is calculated and the missing value is replaced.

With this method, it is vital that the regression of the missing components on the observed ones is constant across missingness patterns. Like the other single imputation methods, conditional mean imputation also overestimates the precision.

Example

Three females who suffered from anorexia were weighed before and after the study period. Results are show in Table 2.6.

Table 2.6 Measured pre-weight and post-weight

Subject	Pre-weight (kg)	Post-weight (kg)
1	36.6	36.4
2	40.6	?
3	41.6	39.2
4	33.6	39.1

Firstly, means and covariance matrix are found for both variables using complete cases. The mean of pre-weight is $\frac{36.6+41.6+33.6}{2} = 37.27$ kg and the mean of post-weight is $\frac{36.4+39.2+39.1}{2} = 38.23$ kg. Covariance matrix based on complete cases is (calculation not shown here)

$$\Sigma = \begin{pmatrix} 16.33 & 1.12 \\ 1.12 & 2.52 \end{pmatrix}.$$

The model of incomplete variable (post-weight) on other variable (pre-weight) that is used to find the estimates is

$$postweight = \hat{\beta}_0 + \hat{\beta}_1 \cdot preweight + \varepsilon.$$

Parameter estimates are found using least squares method. In this example

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

and the least square estimate is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

where \mathbf{X} is the model matrix and \mathbf{y} is vector of post-weight results.

Then (derivation is out of scope) $\hat{\beta}_1 = \frac{cov(X,Y)}{cov(X,X)} = \frac{1.12}{16.33} = 0.07$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = 38.23 - 0.07 \cdot 37.27 = 35.62$. Imputed post-weight value for subject 2 is $\hat{\beta}_0 + \hat{\beta}_1 \cdot preweight_2 = 35.62 + 0.07 \cdot 40.6 = 38.46$ kg. The data with imputed value is presented in Table 2.7.

Table 2.7 Pre-weight and post-weight after conditional mean imputation

Subject	Pre-weight (kg)	Post-weight (kg)
1	36.6	36.4
2	40.6	38.46
3	41.6	39.2
4	33.6	39.1

2.1.3.3 Last observation carried forward

In the approach of last observation carried forward (LOCF, Estonian *viimase vaatluse edasi kandmine*), missing values are replaced with the last observed value for the same subject, hence LOCF approach can only be used when the data has repeated structure. This technique can be used for monotone (when all observations are missing after dropout) and non-monotone (when a subject has missed some visits in between) missing data. [7]

Even though LOCF is one of the most used approaches for dealing with missing data, it could be risky for several reasons. Firstly, to guarantee the validity of this method, often unrealistic assumptions are made. Belief that subjects stay at the same level after dropout or during their unobserved period is required. Secondly, due to the fact that LOCF handles imputed and actually observed values on equal basis, it often overestimates the precision. [7] Furthermore, low p-values and underestimated variability are the results of attributing identical values for the same subject [4].

Example

Haemoglobin (g/dL) was measured for five male subjects during five visits. The data is presented in Table 2.8 below.

Table 2.8 Measured haemoglobin (g/dL) during five visits with missing values

Subject	Visit				
	1	2	3	4	5
1	13.3	13.4	14.0	?	?
2	16.5	16.5	16.7	17.0	17.0
3	12.5	?	13.0	13.5	?
4	14.5	14.6	14.6	?	?
5	14.0	14.0	14.2	14.2	14.3

The data set obtained after applying last observation carried forward method is presented in Table 2.9 below.

Table 2.9 Haemoglobin (g/dL) results during five visits after LOCF imputation

Subject	Visit				
	1	2	3	4	5
1	13.3	13.4	14.0	14.0	14.0
2	16.5	16.5	16.7	17.0	17.0
3	12.5	12.5	13.0	13.5	13.5
4	14.5	14.6	14.6	14.6	14.6
5	14.0	14.0	14.2	14.2	14.3

2.1.4 Hot deck

This section is based on [9].

Hot deck method handles missing data by replacing all the missing values with an observed response from a unit with similar characteristics. The non-respondent is called the recipient and the respondent the donor. There are two different kinds of hot deck methods: random hot deck methods and deterministic hot deck methods. For the first one, the donor is selected randomly from the donor pool, which is a set of potential donors. For the second method, only one donor is selected and used for the imputation.

Adjustment cell method is one of the approaches for identifying donors. Adjustment cells, also known as imputation classes or donor pools, are based on covariate information. Continuous covariates are categorized to create cells. For example, adjustment cell for weight uses variables like height, physical activity, consumption of alcohol, etc. So that subjects with similar height, physical activity and alcohol consumption are put into the same cell.

After creating the adjustment cells, randomly picked donor is used for replacing the missing value for non-respondent within each cell in the random hot deck method. In case of sparseness of donors some hot decks limit the number of times one donor can be used for imputations to avoid over-usage.

Another way for matching donors and recipients is to use some distance metrics. Let $\mathbf{x}_i = (x_{i1}, \dots, x_{iq})$ be the values for subject i of q covariates that are used to create adjustment cells, and let $C(\mathbf{x}_i)$ denote the cell in the cross-classification in which subject i falls. Then matching the recipients i to donors j can be done based on the metric

$$d(i, j) = \begin{cases} 0 & j \in C(\mathbf{x}_i) \\ 1 & j \notin C(\mathbf{x}_i) \end{cases}$$

which is same as matching in the same adjustment cell.

Other metrics are defined so that they do not need to categorize continuous variables. These are maximum deviation

$$d(i, j) = \max_{k \leq q} |x_{ik} - x_{jk}|$$

where \mathbf{x}_k have been suitably scaled to make differences comparable (using ranks and then standardizing), the Mahalanobis distance

$$d(i, j) = (\mathbf{x}_i - \mathbf{x}_j)^T \widehat{\mathbf{Var}}(\mathbf{x}_i)^{-1} (\mathbf{x}_i - \mathbf{x}_j)$$

where $\widehat{\mathbf{Var}}(\mathbf{x}_i)$ is an estimate of the covariance matrix of \mathbf{x}_i , and the predictive mean metric

$$d(i, j) = \left(\hat{Y}(\mathbf{x}_i) - \hat{Y}(\mathbf{x}_j) \right)^2$$

where $\hat{Y}(\mathbf{x}_i) = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ is the predicted value of Y for non-respondent i from the regression of Y on x using only the respondents' data.

The easiest distance to use is the predictive mean metric because it merely requires conversion to set of dummy variables for inclusion in the regression model. Its advantage is also that the metric will be dominated by variables that are predictive of Y , while the variables with little

predictive power may excessively influence the Mahalanobis metric. Predictive mean can be used for discrete and continuous outcomes if generalized linear models (e.g. logistic regression for binary responses) are used for modelling the metric.

After choosing the metric, set of donors are defined for each recipient. One possibility is to specify a maximum distance δ and then define a donor set with $d(i, j) < \delta$. Then the donor is randomly selected from the set (random hot deck). The alternative is to choose the nearest respondent and then the method is called a deterministic or nearest neighbour hot deck.

Hot deck methods are popular because they enable analysts to use complete-data methods for secondary analysis. These methods use values that come from observed responses in the donor pool, therefore only plausible values are imputed. On the other hand, if missing values were extreme and they were replaced with common value then the variability is reduced. Furthermore, finding good matches for non-respondents might be difficult, especially in a smaller sample.

2.2 Approaches for data MAR

2.2.1 Inverse probability weighting

This section is based on [4].

Inverse probability weighting (IPW, Estonian *pöördtõenäosusega kaalumine*) method is approach used to deal with missing data when the missingness is MAR. It is based on sample survey weights which are the inverse of participant's probability of being selected to the survey sample. In case of missing data, statisticians estimate the probability of data being observed and then the observed values are weighted by the inverse of these probabilities. Therefore, those who have lower probability of being observed will have bigger weight. The probability of a variable being observed may depend, for example, on which treatment group the subject is from, what are the previous outcomes of interest and other variables that might predict the observation. All of these are included into the model (for example logistic regression) from where the weights are acquired.

Unfortunately, inverse probability weighting method includes only participants with complete data in the final weighted model, consequently reducing the power.

Example

There are two groups of subjects in a trial of chronic back pain where one group receives placebo and the other group gets active medication. In a couple of weeks, the subjects had to answer from scale 1 to 5 how strong their pain was. The data is presented in Table 2.10 below.

Table 2.10 Level of back pain within two treatment groups

Group	Placebo					Active medication				
Response (actual):	5	3	4	4	5	3	4	2	2	1
Response (observed):	5	?	4	?	?	?	4	2	2	1

The average response for full data is 3.3. The mean calculated from the observed data is 3, which is biased. The probabilities of response for placebo and medication groups are 2/5 and 4/5, respectively. Weights for the groups are the inverse of their probability, therefore being 5/2 and 5/4. Hence the estimate for the response using IPW is

$$\frac{(5 + 4) * \frac{5}{2} + (4 + 2 + 2 + 1) * \frac{5}{4}}{2 * \frac{5}{2} + 4 * \frac{5}{4}} = 3.375$$

which is an unbiased estimate under the assumption that the probability model for the missing data mechanism was correctly specified, i.e. the missingness only depended on the treatment arm.

2.2.2 Multiple imputation

This section is based on [7] if not mentioned otherwise.

Multiple imputation (MI, Estonian *mitmene asendamine*) method is similar to single imputation methods (section 2.1.3) but instead of imputing one value for the missing observation, set of M plausible values are inserted. Firstly, it is important to look at the relationship between missing observations and observed ones to see what the conditional distribution of the missing observations given the observed data - $(\mathbf{Y}_m | \mathbf{Y}_o)$ - is. Secondly, missing values are replaced with the Bayesian value draw (it is not explained in this thesis; more thorough explanation is given in [7]) from the conditional distribution, and that imputation is done M (usually 5-20) times, therefore producing M complete data sets. Each of those data sets is then analysed using appropriate complete data analysis method that would have been used in the absence of

nonresponse, and all of those results are combined into one inference by finding the average of estimates. Imputations are generated from the imputation model, while the later analysis' model is called the substantive model.

It is of interest to make inferences about parameter β from the substantive model. Imputation model is used to make appropriate Bayesian posterior draws. M complete data sets are formulated by replacing the missing data with corresponding imputation samples. Let $\hat{\beta}_m$ and $\widehat{\sigma}_m^2$ denote the estimate of β and its variance from the m th complete data set ($m = 1, \dots, M$). The MI estimate of β is calculated as an average of these estimates,

$$\hat{\beta}_m^* = \frac{1}{M} \sum_{m=1}^M \hat{\beta}_m$$

To estimate the expected uncertainty in the imputations, between-imputation variability is calculated. It is defined as

$$B = \frac{1}{M-1} \sum_{m=1}^M (\hat{\beta}_m - \hat{\beta}^*)^2$$

The formula for calculating the estimation variability due to missing information, known as the within-imputation variability, is the following:

$$W = \frac{1}{M} \sum_{m=1}^M \widehat{\sigma}_m^2$$

The total variance is given by

$$V = W + \left(\frac{M+1}{M}\right) B$$

The advantages of MI are unbiased estimates and correct p-values if the missingness is MAR. In addition, this method is relatively easy to implement and gives opportunities to also handle the missing covariate information. On the other hand, the imputation model and substantive model need to be comparable which means that analysis model has to have the same variables as the imputation model. [4]

2.2.3 Likelihood-based analysis

This section is based on [7] if not mentioned otherwise.

Likelihood-based analysis, like maximum likelihood estimation (MLE, Estonian *suurima tõepära meetod*) methods which use expectation-maximization algorithm (EM, Estonian *EM algortim*), is another method for dealing with missing data when the mechanism is MAR.

With this MLE method likelihood of the observed data is found which is then maximized. When missing data occurs then the likelihood of observed data is more complex and maximizing the likelihood is complicated. An iterative method, the EM algorithm, is the solution. [8] The EM algorithm calculates maximum likelihood estimates in parametric models. There are two steps for each iteration that are repeated until convergence. E step that is expectation step and M step, i.e. the maximization step. The E step uses observed data and a set of parameter estimates to calculate the conditional expectation of the complete data log-likelihood. The M step computes parameters maximizing the expected log-likelihood from the E step.

The advantages of MLE-based imputation are that it produces unbiased estimates of the treatment effect and correct p-values if the missingness is MAR. For MLE-based imputation, there is only one estimate of treatment effect and since there is no imputation model, comparability of imputation model and analysis model is not needed (unlike with multiple imputation). Unfortunately, parametric assumptions (e.g. normality) have to be taken into consideration, but it is only fitting for missing outcome data (i.e. it is not capable of accommodating missing covariate data). [4]

3. Practical

3.1 Overview of the original data

The data used for this chapter was originally captured in Hand, D. J., Daly, F., McConway, K., Lunn, D. and Ostrowski, E. eds (1993) *A Handbook of Small Data Sets*. Chapman & Hall, Data set 285 (p. 229) which is available in [10] under “Anorexia”. The SAS code for reading in data and making necessary adjustment is located in Appendix 1.

The original data with no missing values had 72 female anorexia patients participating in a trial where their weights were measured before and after the study period. During the study period they got either cognitive behavioural treatment, family treatment or no treatment at all (control group). For the simplicity of the analysis, cognitive behavioural treatment and family treatment were combined into one group denoted by treatment 1 and control group was denoted by treatment 0, in this thesis. Basic statistical indicators and frequencies are presented in Table 3.1 and Table 3.2 which is a standard way of summarising data in clinical trials.

Table 3.1 Characteristics of pre-weight and post-weight

Statistic	Pre-weight (kg)	Post-weight (kg)
N	72	72
Mean	37.38	38.63
Standard deviation	2.351	3.645
Minimum	31.8	32.3
Median	37.33	38.12
Maximum	43.0	47.0

Table 3.2 Disposition of subjects

Treatment	Frequency
Control group	26/72 (36.11%)
Medication group	46/72 (63.89%)

In order to investigate how treatment group and pre-weight influenced the post-weight, a linear regression model was fitted. For the original data the regression model produced the following fit:

$$postweight = 20.20 + 2.61 \cdot treatment + 0.45 \cdot preweight.$$

All the variables were statistically significant with p-values lower than 0.05 (see Appendix 2 for details). The Root MSE of the model was 3.25 which means that with probability of 68% the real value of post-weight is ± 3.25 kg from the prognosis. The model accounted for 22.9% of the total variance of the post-weight (Appendix 2).

When fitting a regression model, it is also important to check if the assumptions are valid. However, as the aim of this thesis is to show what are the results generated by different missing data methods, then the validity of assumptions is not described here.

3.2 Missing completely at random

3.2.1 SAS programs

Program code that was used for generating missing values completely at random can be found from Appendix 3. The new data set was named “anorexia_mcar”. Three new variables were created: postwgt1, where 10% of post-weights were missing, postwgt2, where 25% of data was missing and postwgt3 with 50% of missing observations (Figure 3.1).

	Identifier	Treatment	Preweight	Postweight 10%	Postweight 25%	Postweight 50%
1	1	0	36.6		36.4	36.4
2	2	0	40.6	36.3	36.3	
3	3	0	41.6	39.2	39.2	39.2
4	4	0	33.6	39.1	39.1	39.1
5	5	0	35.4	34.5		
6	6	0	40.1	35.4		
7	7	0	39.6	34.1		
8	8	0	34.1	39.3		39.3
9	9	0	36.6			33.3
10	10	0	35.6	38.4	38.4	

Figure 3.1 First 10 observations of data set with missing values

3.2.1.1 Complete case analysis

Firstly, new data set was created where subjects with missing observations were excluded, that was done with command `if nmiss(postwgt1)=0`. If several numeric variables have missing values then all of them are removed with command `if nmiss(of _numeric_)=0`.

After creating the new data set, linear regression model was fitted with `proc reg` procedure.

The SAS code for anorexia trial example was following:

```
data cc_1;
    set anorexia_mcar;
    if nmiss(postwgt1)=0;
run;

proc reg data=cc_1;
    model postwgt1= treatm prewgt;
run;
```

3.2.1.2 Available case analysis

The first step of available case analysis is to find covariances and output these into new data set (seen in Figure 3.2) which is then used for fitting the regression model. With this method only `proc reg` can be used, so categorical variables have to be converted into numeric.

```
proc corr data=anorexia_mcar cov outp=ac_1;
    var postwgt1 treatm prewgt;
run;
```

	TYPE	_NAME_	postwgt1	treatm	prewgt
1	COV	postwgt1	13.774054296	0.6775536027	2.7290926044
2	COV	treatm	0.6775536027	0.2339593114	0.141295088
3	COV	prewgt	2.7290926044	0.141295088	5.5259183984
4	MEAN		38.782147635	0.6388888889	37.379791224
5	STD		3.7113413068	0.4836934064	2.3507272063
6	N		65	72	72
7	CORR	postwgt1	1	0.3828369615	0.3060742862
8	CORR	treatm	0.3828369615	1	0.1242666708
9	CORR	prewgt	0.3060742862	0.1242666708	1

Figure 3.2 Outputted data set by `proc corr` command

```
proc reg data=ac_1;
    model postwgt1=treatm prewgt;
run;
```

3.2.1.3 Unconditional mean imputation

For unconditional mean imputation approach means of variables with missing information are put into new data set.

```
proc means data=anorexia_mcar mean;
    var postwgt1 postwgt2 postwgt3;
    output out=mean1;
run;
```

Only means are taken from the data set outputted from `proc means` procedure. Statement `do ID=1 to 72` makes 72 rows with ID numbers in order to merge with data set “anorexia_mcar”.

```
data means;
    set mean1;
    where _STAT_='MEAN';
    drop _TYPE_ _FREQ_ _STAT_;

    do ID=1 to 72;
        m_postwgt1=postwgt1;
        m_postwgt2=postwgt2;
        m_postwgt3=postwgt3;
    output;
    end;
    drop postwgt1 postwgt2 postwgt3;
run;
```

In the next step means are imputed to missing values and then used for fitting a regression model.

```
data unconditional;
    merge anorexia_mcar means;
        by ID;

    format unpostwgt1 unpostwgt2 unpostwgt3 6.1;

    if postwgt1=. then unpostwgt1=m_postwgt1;
        else unpostwgt1=postwgt1;

    if postwgt2=. then unpostwgt2=m_postwgt2;
        else unpostwgt2=postwgt2;

    if postwgt3=. then unpostwgt3=m_postwgt3;
        else unpostwgt3=postwgt3;

    drop m_postwgt1 m_postwgt2 m_postwgt3;
run;

proc reg data=unconditional;
    model unpostwgt1=treatm prewgt;
run;
```

3.2.1.4 Conditional mean imputation

`Proc mi` procedure can be used for conditional mean imputation. `Nimpute` is the number of imputations and `nbiter` is the number of burn in iterations, they both should be set to one for the conditional mean imputation method to ensure that only one set of imputed data sets is generated. `Seed` is put into the command so every time the code is run the imputed values stay the same. Statement `fcs` uses stochastic regression for imputing data. Values for imputing are put into new data set which is then used for fitting the regression model.

```
proc mi data=anorexia_mcar nimpute=1 seed=37887 out=cond_1;
    fcs nbiter=1;
    var postwgt1 treatm prewgt;
run;

proc reg data=cond_1;
    model postwgt1= treatm prewgt;
run;
```

3.2.1.5 Last observation carried forward

For last observation carried forward method, if post-weight was missing for a patient then her pre-weight was imputed for the missing value. Afterwards, regression model was fitted.

```
data locf;
    set anorexia_mcar;

    format postw1 postw2 postw3 6.1;

    if postwgt1=. then postw1=prewgt;
        else postw1=postwgt1;
    if postwgt2=. then postw2=prewgt;
        else postw2=postwgt2;
    if postwgt3=. then postw3=prewgt;
        else postw3=postwgt3;
run;

proc reg data=locf;
    model postwgt1=treatm prewgt;
run;
```

3.2.1.6 Hot deck

Hot deck methods are done with procedures `proc hotdeck` and `proc surveyimpute` but since they are not available in base SAS available for the author of the thesis, it is not shown here.

3.2.2 Results

3.2.2.1 10% missing data

In Table 3.3 parameter estimates, standard errors and p-values for variables pre-weight and treatment obtained with different approaches are presented, for intercept p-values are not presented. Intercept's parameter estimates were mostly within the ± 2 of the original result and the standard errors' change was minimal. While parameter estimates and standard errors for pre-weight and treatment were either larger or smaller depending on the method, then p-values were larger with all the methods except for conditional mean imputation for pre-weight. Both variables stayed statistically significant for all approaches. Standard errors for pre-weight were the same as the original with complete case analysis and conditional mean imputation, with other methods they were ± 0.01 of the original result.

Table 3.3 Summary of results for different approaches with 10% missing data

Method	Intercept	Pre-weight		Treatment	
	Parameter estimate (s.e)	Parameter estimate (s.e)	p-value	Parameter estimate (s.e)	p-value
Original	20.20 (6.14)	0.45 (0.17)	0.0084	2.61 (0.80)	0.0017
Complete case analysis	21.63 (6.51)	0.41 (0.17)	0.0221	2.74 (0.88)	0.0028
Available case analysis	21.57 (6.66)	0.41 (0.18)	0.0240	2.69 (0.87)	0.0030
Unconditional mean imputation	22.90 (6.08)	0.38 (0.16)	0.0215	2.38 (0.79)	0.0038
Conditional mean imputation	18.76 (6.46)	0.49 (0.17)	0.0060	2.47 (0.84)	0.0046
Last observation carried forward	20.71 (6.06)	0.44 (0.16)	0.0093	2.57 (0.79)	0.0018

The Root MSE calculated with unconditional mean imputation and last observation carried forward were the closest to the original, all the other methods overestimated it. The most similar coefficient of determination with original data was with last observation carried forward method as can be seen from Table 3.4.

Table 3.4 Root MSE and coefficient of determination for different approaches with 10% missing data

Method	Root MSE	Coefficient of determination
Original	3.25	22.91%
Complete case analysis	3.34	21.63%
Available case analysis	3.34	21.44%
Unconditional mean imputation	3.21	19.31%
Conditional mean imputation	3.41	21.55%
Last observation carried forward	3.20	22.60%

3.2.2.2 25% missing data

With 25% missing data the pre-weight stayed statistically significant while treatment became insignificant with conditional mean imputation and unconditional mean imputation (Table 3.5). For treatment the parameter estimate changed considerably, especially with conditional mean imputation - the estimate being almost three times smaller. The most different (with 0.26 change) pre-weight estimate occurred with last observation carried forward method. LOCF method also had two times smaller parameter estimate for intercept.

Table 3.5 Summary of results for different approaches with 25% missing data

Method	Intercept	Pre-weight		Treatment	
	Parameter estimate (s.e)	Parameter estimate (s.e)	p-value	Parameter estimate (s.e)	p-value
Original	20.20 (6.14)	0.45 (0.17)	0.0084	2.61 (0.80)	0.0017
Complete case analysis	15.97 (7.60)	0.57 (0.20)	0.0071	2.06 (0.96)	0.0366
Available case analysis	17.79 (7.14)	0.53 (0.19)	0.0082	1.94 (0.93)	0.0429
Unconditional mean imputation	24.11 (5.53)	0.37 (0.15)	0.0156	1.41 (0.72)	0.0550
Conditional mean imputation	14.44 (6.05)	0.63 (0.16)	0.0002	0.96 (0.79)	0.2299
Last observation carried forward	10.60 (5.50)	0.71 (0.15)	<.0001	1.67 (0.72)	0.0227

The biggest underestimation of Root MSE happened with LOCF and unconditional mean imputation. For coefficient of determination underestimation was biggest with unconditional mean imputation and the overestimation was biggest with LOCF. With other approaches the change was minimal. Results are presented in Table 3.6.

Table 3.6 Root MSE and coefficient of determination for different approaches with 25% missing data

Method	Root MSE	Coefficient of determination
Original	3.25	22.91%
Complete case analysis	3.25	21.37%
Available case analysis	3.26	20.95%
Unconditional mean imputation	2.92	14.11%
Conditional mean imputation	3.20	20.80%
Last observation carried forward	2.90	31.77%

3.2.2.3 50% missing data

The effect of missing data on analysis results is best seen with 50% missing data (Table 3.7). Pre-weight became statistically insignificant with complete case analysis, with p-value almost 9 times larger than with original data set. Treatment became insignificant with LOCF, p-value also increased with complete case analysis, but otherwise it decreased. The most precise pre-weight estimate was found with conditional mean imputation (being 0.44), the estimate furthest from the original was observed with LOCF. For treatment, estimates differed considerably. While unconditional mean imputation and LOCF methods' estimates were smaller (by 0.64 and 1.57, respectively), then other methods overestimated treatment effect remarkably. With LOCF parameter estimate for intercept decreased two times and with available case analysis 1.4 times.

Table 3.7 Summary of results for different approaches with 50% missing data

Method	Intercept	Pre-weight		Treatment	
	Parameter estimate (s.e)	Parameter estimate (s.e)	p-value	Parameter estimate (s.e)	p-value
Original	20.20 (6.14)	0.45 (0.17)	0.0084	2.61 (0.80)	0.0017
Complete case analysis	23.67 (7.37)	0.37 (0.20)	0.0754	3.40 (1.12)	0.0046
Available case analysis	14.48 (7.79)	0.60 (0.21)	0.0069	3.94 (1.02)	0.0005
Unconditional mean imputation	24.85 (4.36)	0.36 (0.12)	0.0031	1.97 (0.57)	0.0009
Conditional mean imputation	20.61 (6.73)	0.44 (0.18)	0.0175	3.38 (0.88)	0.0003
Last observation carried forward	10.65 (4.97)	0.72 (0.13)	<.0001	1.04 (0.65)	0.1143

The only method that overestimated Root MSE was conditional mean imputation, all the other methods underestimated it. With available case analysis coefficient of determination was almost two times bigger than that of original analysis (Table 3.8).

Table 3.8 Root MSE and coefficient of determination for different approaches with 50% missing data

Method	Root MSE	Coefficient of determination
Original	3.25	22.91%
Complete case analysis	3.08	36.96%
Available case analysis	2.89	44.48%
Unconditional mean imputation	2.31	26.14%
Conditional mean imputation	3.56	25.32%
Last observation carried forward	2.63	33.43%

3.3 Missing at random

3.3.1 SAS codes

Data set “anorexia_mar” was used for MAR experiments. This data set included three variables postwgt1, postwgt2 and postwgt3 that had 10%, 25% and 50% missing data with MAR mechanism (missingness generated by the author of this thesis). For MAR generation, the following assumptions were used: females who weighed less than average or more than 40 kg were unlikely to respond because they were afraid to reveal their weight or they thought they were too heavy. It was a deterministic removal so everyone in the assumption category was removed. SAS code is presented in Appendix 4.

3.3.1.1 Inverse probability weighting

The first step of inverse probability weighting is to find a number of subjects in different treatment groups, for which `by` statement is used within `proc means`.

```
proc means data=anorexia_mar NMISS N;  
  var postwgt1 postwgt2 postwgt3;  
  by treatm;  
  output out=nmissing;  
run;
```

The probability of response is calculated by number of persons with complete data (postwgt1, postwgt2, postwgt3) divided by the number of persons in treatment group (`_freq_`). Probabilities are found for both treatment groups.

```
data weights;  
  set nmissing;  
  where _STAT_='N';  
  resp_w1=postwgt1/_freq_;  
  resp_w2=postwgt2/_freq_;  
  resp_w3=postwgt3/_freq_;  
  keep treatm resp_w1 -- resp_w3;  
run;
```

Data set with missing values is then merged with data set with probabilities of response and weights are found for treatment and control group.

```
data ipw;  
  merge anorexia_mar (in=a) weights;  
  by treatm;  
  if a;  
  
  format w1 w2 w3 6.2;  
  
  w1=1/resp_w1;
```

```

w2=1/resp_w2;
w3=1/resp_w3;

drop resp_w1 resp_w2 resp_w3;
run;

```

Then the regression model is fitted using the `weight` statement within `proc reg` to specify the pre-calculated inverse probability weights.

```

proc reg data=ipw;
  model postwgt1=treatm prewgt;
  weight w1;
run;

```

3.3.1.2 Multiple imputation

Multiple imputation has three phases in SAS: imputation phase, analysis phase and pooling phase. In imputation phase number of imputations is specified in `proc mi` procedure with `nimpute` command. The imputed data sets are outputted into new data set that is later used for analysis phase. `Proc mi` procedure creates indicator variable imputation to number each imputed data set [11].

```

/*imputation phase with M=10 imputations*/
proc mi data= anorexia_mar nimpute=10 out=mi_trial1 seed=54321;
  var postwgt1 treatm prewgt;
run;

```

Model is found in second – analysis – phase for every imputed data set individually with `by` statement. Parameter estimates from the regression model are outputted into data set that is used for the last phase – pooling.

```

/*analysis phase*/
proc reg data = mi_trial1;
  model postwgt1=treatm prewgt;
  by _imputation_;
  ods output ParameterEstimates=est_1;
run;

```

Procedure `proc mianalyze` used for pooling phase combines all the estimates across imputations. Coefficients are calculated as mean of individual coefficients for every imputed data set [11].

```

/*pooling phase*/
proc mianalyze parms=est_1;
  modeleffects intercept treatm prewgt;
run;

```

The output of pooling phase is presented in Figure 3.3. From it within-imputation and between imputations variances (explained in section 2.2.2) can be seen (Variance Information table). The procedure also releases 95% confidence limits for parameter estimates (Parameter Estimates table).

Model Information									
PARMS Data Set					WORK.EST_1				
Number of Imputations					10				
Variance Information									
Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency		
	Between	Within	Total						
intercept	4.806933	36.662873	41.950500	566.49	0.144223	0.129114	0.987253		
treatm	0.089911	0.625951	0.724853	483.43	0.158003	0.139995	0.986194		
prewgt	0.004157	0.026502	0.031075	415.63	0.172544	0.151228	0.985103		

Parameter Estimates										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > t
intercept	20.617997	6.476921	7.896286	33.33971	566.49	17.434761	24.773486	0	3.18	0.0015
treatm	2.449677	0.851383	0.776810	4.12255	483.43	1.991475	3.061904	0	2.88	0.0042
prewgt	0.449654	0.176280	0.103143	0.79617	415.63	0.324499	0.545342	0	2.55	0.0111

Figure 3.3 Output of `proc mianalyze` procedure

3.3.1.3 Likelihood-based analysis

For maximum-likelihood estimation procedure `proc mi` is used with added statement `EM`, which requires EM algorithm to be used. Outputted data set of `proc mi` procedure is then used for fitting a regression model.

```
proc mi data = anorexia_mar seed=45678;
    EM out = mle1;
    var postwgt1 treatm prewgt;
run;

proc reg data = mle1;
    model postwgt1=treatm prewgt;
run;
```

3.3.2 Results

3.3.2.1 10% missing data

The change in the parameter estimate and standard errors for intercept was minimal, so was it with pre-weight estimate and standard errors. The p-value for pre-weight decreased only with likelihood-based imputation, otherwise it increased but not as much to become insignificant. The same applied for p-value for treatment. The results are presented in Table 3.9.

Table 3.9 Summary of results for different approaches with 10% missing data

Method	Intercept	Pre-weight		Treatment	
	Parameter estimate (s.e)	Parameter estimate (s.e)	p-value	Parameter estimate (s.e)	p-value
Original	20.20 (6.14)	0.45 (0.17)	0.0084	2.61 (0.80)	0.0017
Inverse probability weighting	21.68 (6.62)	0.42 (0.18)	0.0237	2.53 (0.86)	0.0046
Multiple imputation	20.62 (6.48)	0.45 (0.18)	0.0111	2.45 (0.85)	0.0042
Likelihood-based imputation	20.98 (5.75)	0.44 (0.15)	0.0059	2.50 (0.75)	0.0014

Because Root MSE and coefficient of determination are not released in multiple imputation procedure, they are calculated as the average of each imputation. Root MSE and coefficient of determination were closest to the original with multiple imputation method (see Table 3.10).

Table 3.10 Root MSE and coefficient of determination for different approaches with 10% missing data

Method	Root MSE	Coefficient of determination
Original	3.25	22.91%
Inverse probability weighting	3.35	24.15%
Multiple imputation	3.20	22.38%
Likelihood-based imputation	3.04	24.05%

3.3.2.2 25% missing data

With 25% missing data, intercept's parameter estimate came negative with every method most probably due to the deterministic removal. The parameter estimates for pre-weight were considerably larger, multiple imputation generated 2.7 times bigger estimate. P-values decreased noticeably. The parameter estimates for treatment decreased with each method and p-values increased but not with likelihood-based imputation that had the closest value to the original. Standard errors for all parameters became larger with the exception of likelihood-based imputation. Results are shown in Table 3.11.

Table 3.11 Summary of results for different approaches with 25% missing data

Method	Intercept	Pre-weight		Treatment	
	Parameter estimate (s.e)	Parameter estimate (s.e)	p-value	Parameter estimate (s.e)	p-value
Original	20.20 (6.14)	0.45 (0.17)	0.0084	2.61 (0.80)	0.0017
Inverse probability weighting	-4.22 (9.81)	1.10 (0.27)	0.0001	2.14 (0.87)	0.0174
Multiple imputation	-8.25 (9.05)	1.21 (0.24)	<.0001	2.24 (0.84)	0.0091
Likelihood-based imputation	-6.66 (4.78)	1.17 (0.13)	<.0001	2.07 (0.62)	0.0015

Root MSE and coefficient of determination that were furthest from the original were generated with likelihood-based imputation as shown in Table 3.12. Coefficient of determination increased greatly, especially with multiple and likelihood-based imputation, being two times bigger than the original estimate.

Table 3.12 Root MSE and coefficient of determination for different approaches with 25% missing data

Method	Root MSE	Coefficient of determination
Original	3.25	22.91%
Inverse probability weighting	3.37	39.04%
Multiple imputation	2.86	55.38%
Likelihood-based imputation	2.53	59.76%

3.3.2.3 50% missing data

The most accurate intercept estimate was generated with likelihood-based imputation which was also the only one that had a decreased standard error (Table 3.13). Multiple imputation had the same parameter estimate for pre-weight, other methods' estimates were also close. Pre-weight was statistically insignificant with multiple imputation. Treatment effect was insignificant with inverse probability weighting and also with multiple imputation. Treatment estimate was smaller with each method and standard error was smaller only with likelihood-based imputation.

Table 3.13 Summary of results for different approaches data with 50% missing data

Method	Intercept	Pre-weight		Treatment	
	Parameter estimate (s.e)	Parameter estimate (s.e)	p-value	Parameter estimate (s.e)	p-value
Original	20.20 (6.14)	0.45 (0.17)	0.0084	2.61 (0.80)	0.0017
Inverse probability weighting	18.86 (8.75)	0.51 (0.24)	0.0448	1.74 (1.31)	0.1923
Multiple imputation	21.07 (8.91)	0.45 (0.25)	0.0769	1.94 (1.02)	0.0588
Likelihood-based imputation	19.92 (4.65)	0.48 (0.13)	0.0003	1.78 (0.61)	0.0046

In Table 3.14 it is shown that inverse probability weighting method overestimated Root MSE almost by two times. On the other hand, the coefficient of determination was most accurate with inverse probability method. While Root MSE decreased only with likelihood-based imputation, it was also the only method that had increased coefficient of determination.

Table 3.14 Root MSE and coefficient of determination with 50% missing data

Method	Root MSE	Coefficient of determination
Original	3.25	22.91%
Inverse probability weighting	5.13	19.78%
Multiple imputation	3.51	16.50%
Likelihood-based imputation	2.46	27.60%

3.4 MCAR analysis methods with MAR data set

This section shows that it is important to determine the missingness mechanism before deciding on a method. The analysis was conducted with 25% missing data with MAR mechanism but MCAR methods were used. Results are presented in Table 3.15.

Table 3.15 Summary of results with MCAR methods with 25% MAR missing data

Method	Intercept	Pre-weight		Treatment	
	Parameter estimate (s.e)	Parameter estimate (s.e)	p-value	Parameter estimate (s.e)	p-value
Original	20.20 (6.14)	0.45 (0.17)	0.0084	2.61 (0.80)	0.0017
Complete case analysis	-6.67 (9.88)	1.17 (0.27)	<.0001	2.07 (0.90)	0.0259
Available case analysis	5.41 (6.14)	0.84 (0.17)	<.0001	2.65 (0.80)	0.0018
Unconditional mean imputation	22.14 (5.48)	0.41 (0.15)	0.0076	2.04 (0.72)	0.0058
Conditional mean imputation	-5.06 (5.42)	1.14 (0.15)	<.0001	1.08 (0.71)	0.1313
Last observation carried forward	-4.07 (4.91)	1.10 (0.13)	<.0001	1.67 (0.64)	0.0113

For pre-weight the parameter estimate was the most accurate with unconditional mean imputation. It was also with the closest p-value to the original, while the others' p-value decreased to <.0001. Results furthest from the original was generated with complete case analysis, the parameter estimate 2.6 times bigger and standard error bigger by 0.10.

The most inaccurate result for treatment came with conditional mean imputation. With this approach p-value of treatment became statistically insignificant and parameter was underestimated by two times. Available case analysis method created the most similar result with the original, with p-value and parameter estimate bigger by 0.1 and 0.0001, respectively, and with the exact same standard error.

Three methods (complete case analysis, conditional mean imputation and LOCF) produced intercept with negative sign. Unconditional method was the only one that had even a bit similar intercept estimate to the original.

All the methods underestimated the Root MSE, the closest to the original was the one generated by complete case analysis and the most different with last observation carried forward method. Coefficient of determination increased two times with available case analysis and conditional mean imputation and even more with last observation carried forward approach. The most accurate coefficient of determination was obtained with unconditional mean imputation (presented in Table 3.16).

Table 3.16 Root MSE and coefficient of determination with 25% missing at random data

Method	Root MSE	Coefficient of determination
Original	3.25	22.91%
Complete case analysis	2.94	39.53%
Available case analysis	2.80	45.03%
Unconditional mean imputation	2.90	20.59%
Conditional mean imputation	2.86	49.62%
Last observation carried forward	2.59	54.66%

Conclusion

The aim of this thesis was to give information about missing data and explain approaches for dealing with it in clinical trials.

In the first chapter, reasons, consequences, such as biased estimates and decreased power, and avoidance of missing data were described. Mechanisms like missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) were defined and examples given.

Second chapter focused on approaches for missing completely at random and missing at random mechanisms. Complete case analysis, available case analysis, single imputation methods and hot deck method were explained under MCAR section. General ideas of inverse probability weighting, multiple imputation and likelihood-based imputation for MAR mechanism were described. To understand how each method works, written examples were given.

SAS programs with instructions were presented in chapter 3. Data set that had 72 female anorexia patients were divided into treatment and control group and weight before study and after study were measured. Regression model was found with every method for 10%, 25% and 50% of missing post-weight variable. Parameter estimate, standard error, p-values, Root MSE and coefficient of determination were shown and the results compared to the original. It was also shown how results change when using MCAR methods with MAR mechanism.

References

[1] ICH, 1996. *Guideline for Good Clinical Practice E6(R1)*.

[2] Little, R.J., D'Agostino, R., Cohen, M.L., Dickersin, K., 2012. The Prevention and Treatment of Missing Data in Clinical Trials, *The NEW ENGLAND JOURNAL of MEDICINE* [online].

Available at: <http://www.nejm.org/doi/pdf/10.1056/nejmsr1203730> [Accessed 21.04.2016].

[3] European Medicines Agency, 2010. *Guidelines on Missing Data in Confirmatory Clinical Trials* [PDF].

http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf [Accessed 18.04.2016].

[4] Dziura, J.D, Post L.A, Zhao Q, Fu Z, Peduzzi P., 2013. Strategies for dealing with missing data in clinical trials: from design to analysis, *Yale Journal of Biology and Medicine* [online].

Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3767219/> [Accessed 21.04.2016]

[5] Scharfstein D.O, Hogan J., Herman A., 2012. On the prevention and analysis of missing data in randomized clinical trials: the state of art, *The Journal of Bone and Joint Surgery* [online].

Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3393113/> [Accessed 28.04.2016].

[6] Fichman, M., Cummings, J.M., 2003. Multiple Imputation for Missing Data: Making Most of What you Know, *Organizational Research Methods*.

Available at: <http://repository.cmu.edu/cgi/viewcontent.cgi?article=1114&context=tepper> [Accessed 28.04.2016]

[7] Molenberghs, G., Kenward, M.G., 2007. *Missing Data in Clinical Studies*. England: John Wiley & Sons Ltd.

[8] Pigott, T.D., 2001. A Review of Methods for Missing Data, *Educational Research and Evaluation* [PDF].

Available at: <http://galton.uchicago.edu/~eichler/stat24600/Admin/MissingDataReview.pdf> [Accessed 21.04.2016].

[9] Andridge, R.R., Little, R.J.A., 2011. A review of Hot Deck Imputation for Survey Non-response, *International Statistical Review* [online].

Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130338/> [Accessed 28.04.2016].

[10] Rdatasets.

Available at: <https://vincentarelbundock.github.io/Rdatasets/datasets.html>

[11] UCLA: Statistical Computing Seminars, *Missing Data in SAS Part 1*.

Available at: http://www.ats.ucla.edu/stat/sas/seminars/missing_data/mi_new_1.htm [Accessed 19.04.2016].

Appendices

Appendix 1. Reading in the data set, making changes and finding regression model

```
PROC IMPORT OUT= WORK.anorexia
            DATAFILE= "C:\Users\birgit.kadastik\Downloads\anorexia.csv"
            DBMS=CSV REPLACE;
GETNAMES=YES;
DATAROW=2;
RUN;

data anorexia_trial;
    set anorexia;

    ID=input(var1,BEST12.);

    /*for simplicity only two treatments*/
    if treat in ('FT' 'CBT') then treatm=1;
        else if treat='Cont' then treatm=0;
    else treatm='';

    format prewgt 6.1 postwgt 6.1;
    prewgt=prewt*0.45359237; /*converting lbs into kg*/
    postwgt=postwt*0.45359237;

    label ID='Identifier'
           treatm='Treatment' /*0 is control, 1 is medication*/
           postwgt='Post weight'
           prewgt='Preweight'
    ;
    drop var1 treat prewt postwt;
run;

/*general statistics*/

proc means data=anorexia_trial N mean min max std median;
    var prewgt postwgt;
run;

proc freq data=anorexia_trial;
    tables treatm;
run;

/*regression model*/
proc reg data=anorexia_trial;
    model postwgt=treatm prewgt;
run;
```

Appendix 2. Goodness of fit and regression model estimates for original data

Root MSE	3.24621	R-Square	0.2291
Dependent Mean	38.63347	Adj R-Sq	0.2067
Coeff Var	8.40260		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	20.19728	6.14328	3.29	0.0016
treatm	Treatment	1	2.61482	0.80271	3.26	0.0017
prewgt	Preweight	1	0.44852	0.16517	2.72	0.0084

Appendix 3. Generating missing values with MCAR mechanism

```
data help;
call streaminit(114);
do ID = 1 to 72;
    miss1=rand('uniform',0,1);
    miss2=rand('uniform',0,1);
    miss3=rand('uniform',0,1);
    output;
end;
run;

proc sort data=help;
    by miss1;
run;

data help1;
    set help;
    if _N_ LT 8 then miss1=1;
    else miss1=0;
run;

proc sort data=help1;
    by miss2;
run;

data help2;
    set help1;
    if _N_ LT 19 then miss2=1;
    else miss2=0;
run;

proc sort data=help2;
    by miss3;
run;

data help3;
    set help2;
    if _N_ LT 37 then miss3=1;
```



```

        else miss3=0;

    proc sort;
        by ID;
run;

data anorexia_mcar;
    merge anorexia_trial help3;
        by ID;

    format postwgt1 postwgt2 postwgt3 6.1;

    if miss1=1 then postwgt1=.;
        else postwgt1=postwgt;

    if miss2=1 then postwgt2=.;
        else postwgt2=postwgt;

    if miss3=1 then postwgt3=.;
        else postwgt3=postwgt;

    label postwgt1='Postweight 10%'
           postwgt2='Postweight 25%'
           postwgt3='Postweight 50%';
    drop miss1 miss2 miss3 postwgt;
run;

```

Appendix 4. Generating missing values with MAR mechanism

```

data anorexia_mar1;
    set anorexia_trial;
    if prewgt le 35 or prewgt ge 40 then miss1=1; /*for 10% and 25%*/
        else miss1=0;
    if prewgt le 37 or prewgt ge 40 then miss2=1; /*for 50%*/
        else miss2=0;
run;

proc sort data=anorexia_mar1;
    by descending miss1 postwgt;
run;

/*10% and 25% missing*/

data anorexia_mar2;
    set anorexia_mar1;
    if _N_ LT 8 then postwgt1=.;
        else postwgt1=postwgt;
    if _N_ LT 19 then postwgt2=.;
        else postwgt2=postwgt;
run;

/*50% missing*/
proc sort data=anorexia_mar2;
    by miss2;
run;

data anorexia_mar;
    set anorexia_mar2;
    if _N_ LT 37 then postwgt3=.;
        else postwgt3=postwgt;

```

```
label postwgt1='Postweight 10%'
      postwgt2='Postweight 25%'
      postwgt3='Postweight 50%';

drop miss1 miss2 postwgt;

proc sort;
  by ID;
run;
```

Non-exclusive licence to reproduce thesis and make thesis public

I, Birgit Kadastik,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

- 1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and
- 1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Missing data in clinical trials,

supervised by Marju Valge and Pasi Korhonen,

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **29.04.2016**