# New methods for generating significance levels from multiply-imputed data

**Dissertation**

zur Erlangung des akademischen Grades

eines Doktors der Sozial- und Wirtschaftswissenschaften
(Dr. rer. pol.)

an der Fakultät Sozial- und Wirtschaftswissenschaften
der Otto-Friedrich-Universität Bamberg

vorgelegt von
Christine Licht
aus Apolda

Bamberg, Oktober 2010

Date of the defence: 2010-12-10

Prof. Dr. Susanne Raessler (1st referee)
Prof. Dr. Donald B. Rubin (2nd referee)

# Acknowledgments

First, I would like to thank my advisors Susanne Rässler and Donald B. Rubin for their support. Susanne Rässler introduced me to missing-data problems and invited me to join the world of multiple imputation. She attended my first steps in this field and prepared me meeting and finally doing research with Donald B. Rubin, the "father" of multiple imputation. He also is the "father" of this thesis, since his incredible previous and current ideas and the close co-operation with him are the fundament of this thesis. I would like to thank him for the uncountable lessons in multiple imputation theory, for his patience, when he answered all my more or less smart questions, for inviting me to do research at the Harvard Statistics department, and in general for the whole support of this thesis.

I am very greatful to Holger Aust, who supported me whenever it was needed and beyond. He motivated me in difficult times, when no solution of the tricky problems was in sight. He shared the great moments of success with me and he always believed in me. He inspired me in many precious discussions on the topic and he made a lot of very helpful comments and corrections concerning this thesis. He attended and supported me carringly the last three years in all areas of life, even when he was just cooking pasta, while I was writing on this thesis.

I am very thankful to my parents for their wonderful support and care. They were always interested in the progress of my work and helped me whenever they could.

Last but not least I would like to thank Julia Cielebak, who shared the office with me, for all the inspiring professional talks and especially for the wonderful "girls-topics" talks that always lighten up the long days in the office.

Bamberg, Oktober 2010 Christine Licht

# Contents

# 1

# Introduction

Missing data are an ubiquitous problem in statistical analyses that has become an important research field in applied statistics because missing values are frequently encountered in practice, especially in survey data. Many statistical methods have been developed to deal with this issue. Substantial advances in computing power, as well as in theory, in the last $30$ years enables the application of these methods for applied researchers. A highly useful technique to handle missing values in many settings is multiple imputation, which was first proposed by Rubin (1977, 1978) and extended in Rubin (1987). The key idea of multiple imputation is to replace the missing values with more than one, say $m$, sets of plausible values, thereby generating $m$ completed data sets. Each of these completed data sets is then analyzed using standard complete-data methods. These repeated analyses are combined to create one imputation inference, that takes correctly account into the uncertainty due to missing data. Multiple imputation retains the major advantages and simultaneously overcomes the major disadvantages inherent in single imputation techniques.

Due to the ongoing improvement in computer power in the last $10$ years, multiple imputation has become a well known and often used tool in statistical analyses. Multiple imputation routines are now implemented in many statistical software packages. However, there still exists a problem in generally obtaining significance levels from multiply-imputed data, because Rubin's combining rules (1978)

for the completed-data estimates require normally distributed or $t$-distributed complete-data estimators. Some procedures were offered in Rubin (1987), but they had limitations. Today there are basically three methods that extend the suggestions given in Rubin (1987). First, Li, Raghunathan, and Rubin (1991) proposed a procedure, where significance levels are created by computing a modified Wald-test statistic which is then referred to an $F$-distribution. This procedure is essentially calibrated and the loss of power due to a finite number of imputations is quite modest in cases likely to occur in practice. But this procedure requires access to the completed-data estimates and their variance-covariance matrices. The full variance-covariance matrix may not be available in practice with standard software, especially when the dimensionality of the estimand is high. This can easily occur, e.g., with partially classified multidimensional contingency tables. Second, Meng and Rubin (1992) proposed a complete-data two-stage-likelihood-ratio-test-based procedure, which was motivated by the well-known relationship between the Wald-test statistic and the likelihood-ratio-test statistic. In large samples this procedure is equivalent to the previous one and only requires the complete-data log-likelihood-ratio statistic for each multiply-imputed data set. However, common statistical software does not provide access to the code for the calculation of the log-likelihood-ratio statistics in their standard analyses routines. Third, Li, Meng, Raghunathan, and Rubin (1991) developed an improved version of a method in Rubin (1987) that only requires the $\chi^2_k$-statistics from a usual complete-data Wald-test. These statistics are provided by every statistical software. Unfortunately, this method is only approximately calibrated and has a substantial loss of power compared to the previous two.

To sum, there exist several relatively "easy" to use procedures to generate significance levels in general from multiply-imputed data, but none of them has satisfactory applicability due to the facts mentioned above. Since many statistical analyses are based on hypothesis tests, especially on the Wald-test in regression analyses, it is very important to find a method that retains the advantages and overcomes the disadvantages of the existing procedures, just as multiple imputation does with the existing techniques to handle missing data. Developing such a method was the aim of the present thesis, that results from a close co-operation

with my advisor Susanne Raessler and especially with my second advisor - the "father" of multiple imputation - Donald B. Rubin.

In Chapter 2 we briefly introduce the multiple imputation theory and give some important notations and definitions. In Chapter 3 we describe in detail the three existing procedures mentioned above that create significance levels from multiply-imputed data. In Chapter 4 we present a new procedure based on a $z$-transformation. First we examine this new $z$-transformation-based procedure for simple hypothesis tests like the $z$-test in Section 4.2 and the $t$-test in Section 4.3, before we consider the Wald-test in Section 4.4. Despite the success of this new $z$-transformation procedure in several practical settings, problems arise when two-sided tests are performed. Therefore we develop and discuss a possible solution in the first section of Chapter 5. Based on a comprehensive simulation study described in Section 5.2, in Section 5.3 we discover an interesting general statistical problem: Using a $\chi^2_k$-distribution rather than an $F_{k,n}$-distribution, can lead to a not negligible error for small sample sizes $n$, especially with larger $k$. This problem seems to be unnoticed until now. In addition, we show the influence of the sample size for generating accurate significance levels from multiply imputed data. Due to these problems described in Chapter 5, in Chapter 6 we present an adjusted procedure, the componentwise-moment-based method, to easily calculate correct significance levels from multiply-imputed data under some assumptions. In Chapter 7 we examine this new componentwise-moment-based method and the already existing procedures in detail by an extensive simulation study and compare them with each other. We also compare the results with former simulation studies of Li, Raghunathan, and Rubin (1991), and Li, Raghunathan, Meng, and Rubin (1991), where they simulated draws from the theoretically calculated distributions of the test statistics, because it was too computationally expensive to generate data sets and impute them several times due to the lack of computer power at that time. Our simulation study enables us to give some practical advices in Chapter 8 about how to calculate correct significance levels from multiply-imputed data. Finally in Chapter 9, an overview is given for addressing many challenging tasks left for future research.

# 2

# Multiple imputation

Multiple imputation is a general statistical technique for handling missing data. It was developed by Rubin (1978) and is described in detail in Rubin's book (1987) on multiple imputation. The key idea is to replace the set of missing values with $m \geq 2$ sets of draws from the posterior predictive distribution of the missing data. Each of these $m$ completed data sets can now be analyzed using standard complete-data techniques, thereby resulting in $m$ completed-data statistics. These are combined to form one multiple imputation inference, which takes account of the uncertainty due to nonresponse or in general missing data.

Let $\theta$ be the quantity of interest in the data set, for example a $k$-component regression coefficient vector from a simple least squares regression. If there were no missing data, we assume that

$$(\hat{\theta} - \theta) \sim N(0, U), \tag{2.1}$$

where $\hat{\theta}$ is the estimate of $\theta$ with associated variance-covariance matrix $U$ produced by using standard complete-data analysis. Suppose now that $m$ completed data sets were created by drawing $m$ repeated imputations. Let $\hat{\theta}_{*1}, \ldots, \hat{\theta}_{*m}$ denote the $m$ values for $\hat{\theta}$, $U_{*1}, \ldots, U_{*m}$ the $m$ associated variance-covariance matrices, and $S_m = \{(\hat{\theta}_{*l}, U_{*l}), l = 1, \ldots, m\}$ the collection of completed-data moments. The $m$ repeated completed-data estimates and associated completed-data

(within) variance-covariance matrices can be combined using Rubin's (1987) combining rules. Let

$$\overline{\theta}_m = \frac{1}{m}\sum_{l=1}^{m}\hat{\theta}_{*l} \tag{2.2}$$

be the average of the $m$ completed-data estimates, let

$$\overline{U}_m = \frac{1}{m}\sum_{l=1}^{m}U_{*l} \tag{2.3}$$

be the average of the $m$ completed-data variance-covariance matrices, and let

$$B_m = \frac{1}{m-1}\sum_{l=1}^{m}(\hat{\theta}_{*l} - \overline{\theta}_m)^t(\hat{\theta}_{*l} - \overline{\theta}_m) \tag{2.4}$$

be the between variance of the $m$ completed-data estimates. The total variance of $(\overline{\theta}_m - \theta)$ is defined to be

$$T_m = \overline{U}_m + (1 + m^{-1})B_m. \tag{2.5}$$

If $\theta$ is a scalar, Rubin (1987) showed that, approximately

$$(\overline{\theta}_m - \theta) \sim t_\nu(0, T_m), \tag{2.6}$$

with

$$\nu = (m-1)(1 + r_m^{-1})^2 \tag{2.7}$$

degrees of freedom, where $r_m$ is the relative increase in variance due to nonresponse:

$$r_m = (1 + m^{-1})B_m/\overline{U}_m. \tag{2.8}$$

If $\theta$ is a $k$-dimensional vector, (2.2)-(2.7) still hold approximately with $r_m$ in (2.8) generalized to be the average relative increase in variance due to nonresponse

$$r_m = (1 + m^{-1})\text{Tr}(B_m\overline{U}_m^{-1})/k, \tag{2.9}$$

where Tr$(A)$ denotes the trace of the matrix $A$.

The fraction of missing information is defined as

$$\gamma_m = [\overline{U}_m^{-1} - (\nu+1)(\nu+3)^{-1}T_m^{-1}] \cdot \overline{U}_m, \tag{2.10}$$

where for scalar $\theta$ we obtain

$$\gamma_m = \frac{r_m + 2/(\nu+3)}{r_m + 1}. \tag{2.11}$$

For calculating significance levels based on the combined estimates and variance-covariance matrices, when $m$ is modest relative to $k$ we use the statistic

$$\tilde{D}_m = (1+r_m)^{-1}(\overline{\theta}_m - \theta_0)\overline{U}_m^{-1}(\overline{\theta}_m - \theta_0)^t/k \tag{2.12}$$

where $\theta_0$ is the null value of $\theta$. In Rubin (1987) the statistic $\tilde{D}_m$ is referred to an $F$-distribution on $k$ and $(k+1)\nu/2$ degrees of freedom.

# 3

# Significance levels from multiply-imputed data

## 3.1 Significance levels from multiply-imputed data using moment-based statistics and an improved $F$-reference-distribution

Li, Raghunathan, and Rubin (1991) presented an improved procedure for creating significance levels based on the set of completed-data moments. To start with, we provide some further notation, which we need throughout this thesis.

Let $\theta_t$ be the true value of the $k$-dimensional parameter of interest and let $\hat{\theta}_{\text{obs}}$ be the maximum-likelihood estimate of $\theta$ based on the observed data. Let $U_t$ denote the true variance of the complete data, that is, $U_t = V(\hat{\theta}|\theta = \theta_t)$, and $U_t^{-1}$ is the complete-data information. $T_t = V(\hat{\theta}_{\text{obs}}|\theta = \theta_t)$ describes the true variance of $\hat{\theta}_{\text{obs}}$ and $T_t^{-1}$ is the observed information. The subscripts $t$ on $\theta$, $U$, and $T$ designate the true values of $\theta$, $U$, and $T$. Then

$$B_t = T_t - U_t$$

is the increase in variance due to nonresponse and the missing information is $U_t^{-1} - T_t^{-1}$. Thus the ratios of missing to observed information are given by the eigenvalues of $(U_t^{-1} - T_t^{-1})T_t$, or after some calculations, by the eigenvalues of $B_t$

relative to $U_t$, which we label by $(\lambda_1, \ldots, \lambda_k) \in [0, \infty)^k$, since each symmetric matrix can be characterized by their real eigenvalues. The ratios of complete to observed information are given by

$$\xi_i = (1 + \lambda_i), \qquad i = 1, \ldots, k, \tag{3.1}$$

and the ratios of missing to complete information, that is, the fractions of missing information, $\gamma_i$, based on the true variances are given by the eigenvalues of $(U_t^{-1} - T_t^{-1})U_t$. In addition $\xi_i = (1 + \lambda_i) = (1 - \gamma_i)^{-1}$. Furthermore, let $C_\xi$ be the coefficient of variation of the $\xi_i$ defined as

$$1 + C_\xi^2 = \frac{1}{k} \sum_{i=1}^{k} (\xi_i / \overline{\xi})^2, \tag{3.2}$$

where $\overline{\xi} = \frac{1}{k} \sum_{i=1}^{k} \xi_i$ denotes the average ratio of complete to observed information.

The procedure proposed by Li, Raghunathan, and Rubin (1991) is based on the test statistic $\tilde{D}_m$ from (2.12) with $\overline{\theta}_m$, $\overline{U}_m$ and $r_m$ defined in Chapter 2. They show (Li, Raghunathan, and Rubin (1991), page 1069) that $\tilde{D}_m$ in (2.12) can be written as

$$\tilde{D}_m = \frac{\sum_{i=1}^{k} \overline{\theta}_{m,i}^2 / k}{1 + r_m} \tag{3.3}$$

with

$$r_m = (1 + m^{-1}) \sum_{l=1}^{k} \sum_{l=1}^{m} (\hat{\theta}_{*l} - \overline{\theta}_m)^2 / k(m - 1) \tag{3.4}$$

under certain assumptions, especially if the sample size is large. They derive the distribution of $\tilde{D}_m$ as

$$\tilde{D}_m \sim \frac{\chi_k^2 / k}{(1 + a\chi_b^2 / b)/(1 + a)}, \tag{3.5}$$

where $b = k(m - 1)$ and $a = (1 + m^{-1})\overline{\lambda}$ under the equal eigenvalue assumption, that is, $\lambda_i = \overline{\lambda}$. Note, that the derivations of (3.1)-(3.5) are given in Appendix A. Li, Raghunathan, and Rubin (1991) improved a procedure in Rubin (1987) by

using a moment matching method to approximate the distribution of $\tilde{D}_m$ in (3.5) by a multiple of an $F$-distribution, $\delta F_{k,w}$. Calculating the Taylor-series expansion of (3.5) in $1/\chi_b^2$ around its expectation, $1/(b-2)$ and then matching the first two moments of that expansion with the first two moments of the $F$-distribution, gives

$$\delta = (1-2/w)[1+ab/(b-2)]/(1+a) = (1-2/w)\cdot\frac{b(1+a)-2}{b(1+a)-2a-2}, \tag{3.6}$$

and

$$w = 4+(b-4)[1+(1-2b^{-1})/a]^2 = 4+(b-4)\left[1+\frac{b-2}{ab}\right]^2. \tag{3.7}$$

Note that with our calculation, which is given in Appendix B, we get similar, but not identical degrees of freedom:

$$\delta' = (1-2/w)\cdot\frac{b(1+a)}{b(1+a)-2a} \quad \text{and} \quad w' = 4+(b-4)\left[1+\frac{b}{a(b-2)}\right]^2.$$

Unfortunately, we could not derive the degrees of freedom given in Li, Raghunathan, and Rubin (1991), and thus it is not possible to show where the difference comes from. Nevertheless, all our simulations described in the following chapters use the original degrees of freedom $\delta$ and $w$. First, the difference between $(\delta, w)$ and $(\delta', w')$ is not that important: also $\delta'$ is also approximately 1, and $w$ and $w'$ are often very large. Second, all their simulation studies and conclusions were based on their degrees of freedom and we want on the one hand to reproduce and on the other hand to compare their results in our simulation study (Chapter 7) with our new "componentwise-moment-based" procedure.

Based on the derivation of $\delta$ and $w$, they consider the behavior of $\tilde{D}_m$ also with unequal ratios of complete to observed information. Moreover, they examine the loss of power for finite $m$ as well as for infinite $m$. For $m \to \infty$ they showed that $\tilde{D}_m$ is essentially the same as the ideal procedure - the two-stage-likelihood-ratio-test based directly on the observed data. In addition to their analytical calculations, they run several simulation studies where they, due to the processing power of the computers at that time, use repeated draws from the $\chi^2$-distributions

in (3.5) and compare the associated levels with the nominal levels. In Chapter 7 we will calculate values of $\tilde{D}_m$ directly from generated multiply-imputed data.

They finally conclude that their procedure based on $\tilde{D}_m$ is essentially well calibrated and has no substantial loss of power except in relatively extreme circumstances, as for example with a large variation in the fractions of missing information.

The disadvantage of this procedure is that it requires access to the collection of completed-data moments $S_m = \{(\hat{\theta}_{*l}, U_{*l}), l = 1, \ldots, m\}$ and the inverse of the within variance-covariance matrix $\overline{U}_m$. Because of recent computer power and depending on the dimension $k$ of the estimand, it might not be an intractable problem in some settings to calculate the inverse of the within variance-covariance matrix, but standard analysis software does usually not provide the set of completed-data moments, $S_m$.

## 3.2 Significance levels from multiply-imputed data using parameter estimates and likelihood-ratio statistics

Motivated by the well-known relationship between the Wald-test statistic and the likelihood-ratio-test statistic, Meng and Rubin (1992) suggested a procedure that does not require the variance-covariance matrices, $U_{*l}$. Yet it needs access to the code for the complete-data log-likelihood-ratio statistic as a function of parameter estimates for each data set completed by multiple imputation. They assume that the complete-data analysis provides the $\chi^2_k$-distributed test statistic of a likelihood-ratio-test, that can be evaluated at new values.

As introduced in Chapter 2, $\theta$ denotes the parameter of interest. In addition, there usually will be nuisance parameters $\phi$, which include all other parameters of the analysis. For example, let $X$ be an $(n \times k)$-data matrix where $X_i$ $(i = 1, \ldots, k)$

denotes the $i$th column vector of $X$, and let $Y$ denote the outcome variable. When setting all of the $k$ regression coefficients, $\theta$, of the linear regression model

$$Y = \theta_0 + X\theta + \epsilon = \beta_0 + \theta_1 X_1 \ldots + \theta_k X_k + \epsilon,$$

where each component of $\epsilon$ is independent, identically distributed with zero mean and common variance $\sigma^2$, equal to zero, $\phi$ includes the estimates of the intercept and the residual variance of the null model. That is, the nuisance parameters $\phi$ are estimated by different values when $\theta = \hat{\theta}$ and $\theta = \theta_0$, respectively. Let $\hat{\phi}$ denote the complete-data estimate of $\phi$ when $\theta = \hat{\theta}$ and $\hat{\phi}_0$ be the complete-data estimate of $\phi$ when $\theta = \theta_0$. For the following procedure, Meng and Rubin (1992) suppose that the complete-data analysis of each of the $m$ imputed data sets produces the estimates $(\hat{\theta}, \hat{\phi})$, the null estimates $(\theta_0, \hat{\phi}_0)$, and the $\chi^2$-statistic of the likelihood-ratio-test, $d$. Consider this complete-data $\chi^2$-statistic as a function of $(\hat{\theta}, \hat{\phi})$, $(\theta_0, \hat{\phi}_0)$ and the data set, say $d(\hat{\theta}, \hat{\phi}, \theta_0, \hat{\phi}_0)$. In our regression example we have

$$d(\hat{\theta}, \hat{\phi}, \theta_0, \hat{\phi}_0) = d(\hat{\beta}, \hat{\sigma}^2_\epsilon, \hat{\beta}_0, \hat{\sigma}^2_{\epsilon_0}) = -2(LL_1 - LL_0),$$

with

$$\begin{aligned} LL_1(\hat{\beta}, \hat{\sigma}^2_\epsilon | Y, X) &= -\tfrac{n}{2} \cdot \ln(2\pi) - \tfrac{n}{2} \cdot \ln(\hat{\sigma}^2_\epsilon) - \tfrac{1}{2\hat{\sigma}^2_\epsilon} \cdot (Y - \hat{\beta}_0 - \hat{\beta}_k X_1 - \ldots - \hat{\beta}_k X_k)^2, \\ LL_0(\hat{\beta}_0, \hat{\sigma}^2_{\epsilon_0} | Y, X) &= -\tfrac{n}{2} \cdot \ln(2\pi) - \tfrac{n}{2} \cdot \ln(\hat{\sigma}^2_{\epsilon_0}) - \tfrac{1}{2\hat{\sigma}^2_{\epsilon_0}} \cdot (Y - \hat{\beta}_0)^2, \end{aligned}$$

where $\{Y, X\}$ denotes the given (completed) data set with $X_1, \ldots, X_k$ as the independent variables, on which $Y$ is regressed.

Let $\overline{\theta}, \overline{\phi}, \overline{\phi}_0$, and $\overline{d}$ denote the average values of $\hat{\theta}, \hat{\phi}, \hat{\phi}_0$, and $d$ across the $m$ imputations. Assume that the function $d$ can be evaluated at $\overline{\theta}, \overline{\phi}, \theta_0$, and $\overline{\phi}_0$ for each of the $m$ completed data sets to obtain $m$ values of $d(\overline{\theta}, \overline{\phi}, \theta_0, \overline{\phi}_0)$, whose average across the $m$ imputations is $\overline{d}_*$. Then the repeated-imputation p-value is

$$\text{p-value} \;=\; \text{Prob}(F_{k,w} > \breve{D}),$$

where

$$\breve{D} = \overline{d}_* / [k + (m+1)(\overline{d} - \overline{d}_*)/(m-1)], \tag{3.8}$$

and $F_{k,w}$ is an $F$-random variable on $k$ and $w$ degrees of freedom, where $k$ equals the number of components of $\theta$, and $w$ equals the denominator degree of freedom of the moment-based procedure given by (3.7).

Meng and Rubin (1992) show that for large samples, their two-stage-likelihood-ratio-test-based method is equivalent to the moment-based procedure for any number of multiple imputations. Instead of requiring the variance-covariance matrices and the inverse of the within variance-covariance matrix, that is a difficult problem especially when the dimensionality of the estimand is high, the two-stage-likelihood-ratio-test-based procedure requires only the point estimates and evaluations of the complete-data log-likelihood-ratio statistic as a function of these estimates and the completed data. The disadvantage of this procedure is, that none of the today's common statistical software packages provide access to the code for evaluating the complete-data log-likelihood at user-specified values of the parameters, although it is easy and fast to calculate and implement, and it does not involve the computation of any matrices.

## 3.3 Significance levels from repeated p-values with multiply-imputed data

Both of the above described procedures inherently have the problem that especially for practical problems with hundreds of variables the standard complete-data analysis provides the collection of completed-data $\chi^2_k$-statistics $S_d = \{d_{*1}, \ldots, d_{*m}\}$ with

$$d_{*l} = (\theta_0 - \hat{\theta}_{*l})^t U_{*l}^{-1} (\theta_0 - \hat{\theta}_{*l}) \tag{3.9}$$

but not the collection of completed-data moments $S_m$ or the likelihood-ratio-test statistic $d(\overline{\theta}, \overline{\phi}, \theta_0, \overline{\phi}_0)$. The problem of directly combining $\{d_{*l}, l = 1, \ldots, m\}$ according to (2.2) is difficult, because Rubin's combining rules require normally distributed or $t$-distributed estimators, but $d_{*l}$ is $\chi^2_k$-distributed. Disregarding that fact and combining $d_{*l}$ directly, leads to too significant p-values.

Li, Meng, Raghunathan, and Rubin (1991) proposed a procedure for creating significance levels based on $S_d$ rather than $S_m$. They use the fact that Rubin (1987) showed that (2.12) implies

$$\tilde{D}_m \approx \hat{D}_m = \frac{\overline{d}_m k^{-1} - (\frac{m-1}{m+1}) r_m}{1 + r_m}, \tag{3.10}$$

where $\overline{d}_m$ is the sample mean of $\{d_{*l}, l = 1, \ldots, m\}$ and $r_m$ is given by (2.9). Replacing $r_m$ in $\hat{D}_m$ with estimates obtained from the set $S_d$ rather than $S_m$ leads to procedures for calculating p-values when only $S_d$ is given. A suggestion of Rubin (1987) provides accurate levels if $m \geq k$, which in practice often is impossible, and a modest fraction of missing information. Therefore Li, Meng, Raghunathan, and Rubin (1991) proposed the following replacement of $r_m$ by the estimate $\hat{r}_d$ with

$$\hat{r}_d = (1 + m^{-1})[\frac{1}{m-1} \sum_{l=1}^{m} (\sqrt{d_{*l}} - \overline{\sqrt{d}})^2], \tag{3.11}$$

that is, $\hat{r}_d$ is the sample variance of $\sqrt{d_{*1}}, \ldots, \sqrt{d_{*m}}$ times $(1 - m^{-1})$. The corresponding test statistic $\hat{\tilde{D}}_d$ is of the form $\hat{D}_m$ from (3.10) with $r_m$ replaced by the estimate $\hat{r}_d$ from (3.11). As reference distribution they use an $F$-distribution on $k$ and $a_{k,m} w_s$ degrees of freedom, where

$$w_s = (m-1)(1 + \hat{r}_d^{-1})^2 \tag{3.12}$$

and

$$a_{k,m} = k^{-3/m}. \tag{3.13}$$

The obvious advantage of this procedure is that only the completed-data test statistics, $\{d_{*l}, l = 1, \ldots, m\}$, are needed for computing the p-value and it is simple to apply. But the procedure is only approximately calibrated and has a substantial loss of power compared to $\tilde{D}_m$. The problem with this method and other methods based on $S_d$, as shown for example in Li (1985) and Raghunathan (1987), is that the loss of information using $S_d$ instead of $S_m$ is too big.

# 4

# $z$-transformation procedure for combining repeated p-values

In 2009 Rubin came up with an idea for combining p-values from multiply-imputed data directly, using a simple transformation and his usual combining rules introduced in Chapter 2. In the following sections we will describe and explore behavior and possibilities of this new procedure, which we call the $z$-transformation procedure.

## 4.1 The new $z$-transformation procedure

Suppose the standard complete-data analysis of multiply-imputed data provides the collection of statistics $S_s = \{s_{*1}, \ldots, s_{*m}\}$ of an arbitrary hypothesis test and/or the set of the corresponding p-values $S_p = \{p_{*1}, \ldots, p_{*m}\}$, where

$$p_{*l} = \text{Prob}(\textit{ reference distribution } \geq s_{*l}), \qquad l = 1, \ldots, m. \tag{4.1}$$

As described in Section 3.3 we cannot combine these p-values directly to get valid inferences, because under the null hypothesis these p-values are uniformly distributed and Rubin's combining rules require a normal distribution or a $t$-distribution. The idea is to transform the p-values to a normal distribution using the quantile function $\Phi^{-1}$ of the standard normal distribution.

Let $S_z = \{z_{*l}, l = 1, \ldots, m\}$ be the collection of the transformed completed-data p-values, where

$$z_{*l} = \Phi^{-1}(1 - p_{*l}). \tag{4.2}$$

After this transformation we calculate the multiple imputation estimator $\overline{z}_m$ as average over the transformed test statistics $z_{*l}$ as

$$\overline{z}_m = \frac{1}{m} \sum_{l=1}^{m} z_{*l}, \tag{4.3}$$

and the between variance $B_m$ given in (2.4) as

$$B_m = \frac{1}{m-1} \sum_{l=1}^{m} (z_{*l} - \overline{z}_m)^t (z_{*l} - \overline{z}_m). \tag{4.4}$$

Because of the $z$-transformation, the within variance $\overline{U}_m$ given in (2.3) equals $1$. Thus, the total variance $T_m$ given in (2.5) is calculated as

$$T_m = \overline{U}_m + (1 + m^{-1}) B_m = 1 + (1 + m^{-1}) B_m. \tag{4.5}$$

It follows from (2.6) that the multiple imputation estimator $\overline{z}_m$ is $t_\nu(0, T)$-distributed with $\nu$ given in (2.7). The corresponding p-value

$$p_m = \text{Prob}(t_\nu(0, T) \geq \overline{z}_m) \tag{4.6}$$

is the intended p-value for multiply-imputed data, which is produced just by using this simple transformation and the set $S_d$ or $S_p$, respectively.

The interesting question is how well this simple procedure performs and for which settings it will be applicable.

## 4.2 $z$-test

First we consider a simple hypothesis test - a one-sided $z$-test, for example a two sample location test of the null hypothesis that the means of one normally dis-

tributed population with known variance is less than or equal to the mean of a second normally distributed population both with known variance. The corresponding test statistic is

$$S = \frac{\overline{X}_1 \quad \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1), \tag{4.7}$$

where $\overline{X}_1$ and $\overline{X}_2$ are the sample means and $n_1$ and $n_2$ are the sample sizes.

In addition we choose a simple model for the following calculations. Let $X$ be a random sample of size $n$ with each component of $X$ distributed as $N(0,1)$. $X^{(1)}$ denotes the first half of $X$ with size $n^{(1)} = n/2$ and $X^{(2)}$ the second half of $X$ with size $n^{(2)} = n^{(1)} = n/2$. Now we randomly delete the first $n^{(1)}_{\text{mis}} = \frac{n}{2} \cdot \gamma$ values of $X^{(1)}$, where $\gamma$ denotes the missingness-rate, which in this case equals the fraction of missing information defined in Chapter 2. Furthermore denote the observed part of $X$ by $X_{\text{obs}}$, the missing values by $X_{\text{mis}}$ and the observed part of $X^{(1)}$ by $X^{(1)}_{\text{obs}}$ and the missing values of $X^{(1)}$ by $X^{(1)}_{\text{mis}}$, respectively, as shown in Figure 4.1.

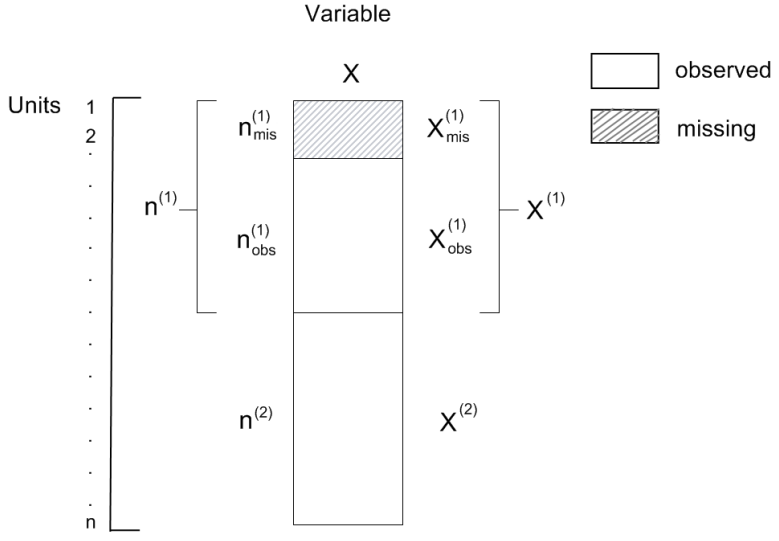


Figure 4.1: Example of an $(n \times 1)$-data vector seperated in two subsamples $X^{(1)}$ and $X^{(2)}$ with same sample size, where the first values of $X^{(1)}$ are missing: Solid = missing, white = observed

In addition we use $\overline{X}^{(1)}$, $\overline{X}^{(2)}$, $\overline{X}^{(1)}_{\text{mis}}$ and $\overline{X}^{(1)}_{\text{obs}}$ for the corresponding sample means. To impute the missing values, we apply the following proper imputation model:

$$\begin{aligned} \mu|X_{\text{obs}} &\sim N\left(\overline{X}^{(1)}_{\text{obs}}, \tfrac{1}{n^{(1)}_{\text{obs}}}\right) = N\left(\overline{X}^{(1)}_{\text{obs}}, \tfrac{2}{n\cdot(1-\gamma)}\right), \\ X^{(1)}_{\text{mis}}|X_{\text{obs}}, \mu &\sim N(\mu, 1). \end{aligned} \tag{4.8}$$

First of all we are interested in the distribution of the $z$-statistic given in (4.7) after one (single) imputation. From (4.8) we get:

$$\overline{X}^{(1)}_{\text{mis}}|X_{\text{obs}}, \mu \sim N\left(\mu, \tfrac{1}{n^{(1)}_{\text{mis}}}\right) = N\left(\mu, \tfrac{2}{n\gamma}\right),$$

$$\begin{aligned} \overline{X}^{(1)}|X_{\text{obs}}, \mu - \overline{X}^{(2)} &= \gamma \cdot \overline{X}^{(1)}_{\text{mis}}|X_{\text{obs}}, \mu + (1-\gamma)\cdot \overline{X}^{(1)}_{\text{obs}} - \overline{X}^{(2)}, \\ &\sim N\left(\gamma\cdot\mu + (1-\gamma)\overline{X}^{(1)}_{\text{obs}} - \overline{X}^{(2)}, \tfrac{2\gamma}{n}\right), \end{aligned}$$

$$\begin{aligned} s_{*l}|X_{\text{obs}}, \mu &= \frac{\overline{X}^{(1)}|X_{\text{obs}},\mu - \overline{X}^{(2)}|X_{\text{obs}},\mu}{\sqrt{\frac{\sigma^2_{(1)}}{n^{(1)}} + \frac{\sigma^2_{(2)}}{n^{(2)}}}} = \frac{\overline{X}^{(1)}|X_{\text{obs}},\mu - \overline{X}^{(2)}|X_{\text{obs}},\mu}{\sqrt{\frac{4}{n}}}, \\ &\sim N\left(\sqrt{\tfrac{n}{4}}\cdot\left(\gamma\mu + (1-\gamma)\overline{X}^{(1)}_{\text{obs}} - \overline{X}^{(2)}\right), \tfrac{n}{4}\cdot\tfrac{2\gamma}{n}\right), \\ &= N\left(\sqrt{\tfrac{n}{4}}\cdot\left(\gamma\mu + (1-\gamma)\overline{X}^{(1)}_{\text{obs}} - \overline{X}^{(2)}\right), \tfrac{\gamma}{2}\right), \end{aligned}$$

$$\begin{aligned} s_{*l}|X_{\text{obs}} &\sim N\left(\sqrt{\tfrac{n}{4}}\cdot\left(\gamma\cdot\overline{X}^{(1)}_{\text{obs}} + (1-\gamma)\overline{X}^{(1)}_{\text{obs}} - \overline{X}^{(2)}\right), \tfrac{\gamma}{2} + \tfrac{n}{4}\cdot\gamma^2\cdot\tfrac{2}{n(1-\gamma)}\right), \\ &= N\left(\sqrt{\tfrac{n}{4}}\cdot\left(\overline{X}^{(1)}_{\text{obs}} - \overline{X}^{(2)}\right), \tfrac{\gamma}{2} + \tfrac{\gamma^2}{2(1-\gamma)}\right), \\ &= N\left(\sqrt{\tfrac{n}{4}}\cdot\left(\overline{X}^{(1)}_{\text{obs}} - \overline{X}^{(2)}\right), \tfrac{\gamma}{2(1-\gamma)}\right). \end{aligned} \tag{4.9}$$

The corresponding completed-data p-values $p_{*l}$ are calculated using the distribution function $\Phi(.)$ of a standard normally distribution

$$p_{*l}|X_{\text{obs}} = 1 - \Phi(s_{*l}|X_{\text{obs}}), \tag{4.10}$$

because with complete data the test statistic $S$ given in (4.7) has a standard normal distribution as reference distribution. If we apply the $z$-transformation to the $p_{*l}$ given in (4.10) we get the transformed values $z_{*l}$

$$z_{*l}|X_{\text{obs}} = \Phi^{-1}(1 - p_{*l}|X_{\text{obs}}) = \Phi^{-1}(\Phi(s_{*l}|X_{\text{obs}})) = s_{*l}|X_{\text{obs}}. \tag{4.11}$$

Thus, for the $z$-test the test statistic, $s_{*l}$, and the test statistic after transformation, $z_{*l}$, are equal, because the reference distribution of the $z$-test (without missing data) is the standard normal distribution and for the $z$-transformation also a standard normal distribution is used.

We combine the $m$ values of $z_{*l}$ or $s_{*l}$, respectively given in (4.9) to

$$\overline{z}_m|X_{\text{obs}} \sim N\left(\sqrt{\frac{n}{4}} \cdot \left(\overline{X}^{(1)}_{\text{obs}} - \overline{X}^{(2)}\right), \frac{\gamma}{2m(1-\gamma)}\right). \tag{4.12}$$

Because $\overline{X}^{(1)}_{\text{obs}} \sim N\left(0, \frac{2}{n(1-\gamma)}\right)$ and $\overline{X}^{(2)} \sim N\left(0, \frac{2}{n}\right)$, it follows

$$\sqrt{\frac{n}{4}} \cdot \left(\overline{X}^{(1)}_{\text{obs}} - \overline{X}^{(2)}\right) \sim N\left(0, \frac{n}{4} \cdot \left(\frac{2}{n} + \frac{2}{n(1-\gamma)}\right)\right) = N\left(0, \frac{2-\gamma}{2(1-\gamma)}\right). \tag{4.13}$$

From (4.12) and (4.13) it follows:

$$\begin{aligned}
\overline{z}_m &\sim N\left(0, \tfrac{\gamma}{2m(1-\gamma)} + \tfrac{2-\gamma}{2(1-\gamma)}\right), \\
&= N\left(0, \tfrac{\gamma}{2m(1-\gamma)} + \tfrac{2-\gamma+\gamma-\gamma}{2(1-\gamma)}\right), \\
&= N\left(0, \tfrac{\gamma}{2m(1-\gamma)} + \tfrac{2-2\gamma+\gamma}{2(1-\gamma)}\right), \\
&= N\left(0, \tfrac{\gamma}{2m(1-\gamma)} + \tfrac{2(1-\gamma)+\gamma}{2(1-\gamma)}\right), \\
&= N\left(0, \tfrac{\gamma}{2m(1-\gamma)} + 1 + \tfrac{\gamma}{2(1-\gamma)}\right), \\
&= N\left(0, 1 + \tfrac{\gamma}{2(1-\gamma)}(1 + m^{-1})\right), \\
&= N(0, U_t + B_t(1 + m^{-1})),
\end{aligned} \tag{4.14}$$

where $U_t$ denotes the true variance of the complete data and $B_t$ denotes the true variance of the incomplete data. Usually we have to estimate these quantities, but in our example we can explicitly derive them and we see that $U_t = 1$ since we started with the complete data $X \sim N(0,1)$. With $U_t = 1$ und $B_t = \frac{\gamma}{2(1-\gamma)}$ from (4.14) it follows $T_t = U_t + B_t = \frac{2-\gamma}{2(1-\gamma)}$, which we have already calculated in (4.13) as the true variance of $\hat{\theta}_{\text{obs}}$. Because we know the true variances in our example we can calculate the distribution of $(\overline{z}_m - z_0)/\sqrt{T_t}$ from (4.14) as

$$\begin{aligned}
\tfrac{\overline{z}_m - z_0}{\sqrt{T_t}} &\sim N\left(0, \tfrac{\frac{\gamma}{2m(1-\gamma)}}{\frac{2-\gamma}{2(1-\gamma)}} + \tfrac{\frac{2-\gamma}{2(1-\gamma)}}{\frac{2-\gamma}{2(1-\gamma)}}\right), \\
&= N\left(0, \tfrac{2\gamma(1-\gamma)}{2m(1-\gamma)(2-\gamma)} + 1\right), \\
&= N\left(0, \tfrac{\gamma}{m(2-\gamma)} + 1\right), \\
&\xrightarrow{m\to\infty} N(0,1),
\end{aligned} \tag{4.15}$$

Thus, for infinite $m$ the distribution of the test statistic after applying multiple imputation and the $z$-transformation is $N(0,1)$, which means that our procedure is asymptotically correct for the $z$-test.

Note that the unconditioned $z$-statistic after one (single) imputation is distributed with

$$z_{*1} \sim N\Big(0, \underbrace{\frac{1}{1-\gamma}}_{>1}\Big). \tag{4.16}$$

This example also shows very nicely how single imputation leads to an underestimation of the variance of the test statistic, if there is no special correction for the variance estimate. If we use the standard normal distribution as the reference distribution, because the true distribution (4.16) has a higher variance than one, we would underestimate the variance thereby leading to a too significant p-value.

Note that the derivations above are also valid for any variable $X$ with its components distributed as $N(\mu, \sigma^2)$ with arbitrary $\mu$ and $\sigma^2$, and when $n^{(1)} \neq n^{(2)}$.

## 4.3 $t$-test

In this section we consider the same settings as in the section before, except that now the variance of $X$ is unknown, that is, each component of $X$ is $N(0, \sigma^2)$-distributed and therefore we use a (one-sided) two-sample $t$-test with a Student's $t$-distribution as reference. We have to take into account the unknown population variance in the imputation model as follows:

$$\begin{aligned} \sigma_*^2|X_{\text{obs}} &\sim S^2_{n^{(1)}_{\text{obs}}} \cdot (n^{(1)}_{\text{obs}} - 1) \cdot \chi^{-2}_{(n^{(1)}_{\text{obs}}-1)}, \\ \mu_*|\sigma_*^2 X_{\text{obs}} &\sim N\left(\overline{X}^{(1)}_{\text{obs}}, \tfrac{\sigma_*^2}{n^{(1)}_{\text{obs}}}\right), \\ X^{(1)}_{\text{mis}}|\mu_*\sigma_*^2, X_{\text{obs}} &\sim N(\mu_*, \sigma_*^2), \end{aligned} \tag{4.17}$$

where $S_n^2$ generally denotes the sample variance of a sample with size $n$ and $\chi^2_{(n^{(1)}_{\text{obs}}-1)}$ is a $\chi^2$-random variable with $(n^{(1)}_{\text{obs}} - 1)$ degrees of freedom.

Now we try to calculate the distribution of the test statistic after one (single) imputation. From (4.17) we get:

$$\overline{X}^{(1)}_{\text{mis}}|\mu_*, \sigma_*^2, X_{\text{obs}} \sim N\left(\mu_*, \tfrac{\sigma_*^2}{n^{(1)}_{\text{mis}}}\right),$$

$$\begin{aligned} \overline{X}^{(1)} - \overline{X}^{(2)}|\mu_*, \sigma_*^2, X_{\text{obs}} &= \tfrac{n^{(1)}_{\text{mis}}}{n^{(1)}}\overline{X}^{(1)}_{\text{mis}}|\mu_*, \sigma_*^2 X_{\text{obs}} + \tfrac{n^{(1)}_{\text{obs}}}{n^{(1)}}\overline{X}^{(1)}_{\text{obs}} - \overline{X}^{(2)}, \\ &\sim N\left(\tfrac{n^{(1)}_{\text{mis}}}{n^{(1)}}\mu_* + \tfrac{n^{(1)}_{\text{obs}}}{n^{(1)}}\overline{X}^{(1)}_{\text{obs}}, \tfrac{n^{(1)}_{\text{mis}}}{(n^{(1)})^2}\sigma_*^2\right). \end{aligned}$$

It is $t_{*l}|\mu_*, \sigma_*^2, X_{\text{obs}} = \frac{\overline{X}^{(1)} - \overline{X}^{(2)}}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n^{(1)}} + \frac{1}{n^{(2)}}\right)}}|\mu_*, \sigma_*^2, X_{\text{obs}}$, where

$$\begin{aligned} \hat{\sigma}^2|\mu_*, \sigma_*^2, X_{\text{obs}} &= \tfrac{\sigma_*^2 S^2_{n^{(1)}} + \sigma^2 S^2_{n^{(1)}}}{n^{(1)} + n^{(2)} - 2}|\mu_*, \sigma_*^2, X_{\text{obs}}, \\ &\sim \chi^2_{\nu_{\text{Welch}}(S^2, \sigma_*^2)}, \end{aligned} \tag{4.18}$$

and

$$\nu_{\text{Welch}}(S^2, \sigma_*^2) = \frac{\left(\sigma_*^2 \cdot S^2_{n^{(1)}} + \sigma^2 \cdot S^2_{n^{(2)}}\right)^2}{\frac{\left(\sigma_*^2 \cdot S^2_{n^{(1)}}\right)^2}{n^{(1)}-1} + \frac{\left(\sigma^2 \cdot S^2_{n^{(2)}}\right)^2}{n^{(2)}-1}}.$$

From (4.18) it follows

$$\begin{aligned}
t_{*l}|\mu_*, \sigma_*^2, X_{\text{obs}} \quad &\sim \quad \frac{N\left(\overbrace{\frac{n^{(1)}_{\text{mis}}}{n^{(1)}}}^{=:\,\zeta}\mu_* + \overbrace{\frac{n^{(1)}_{\text{obs}}}{n^{(1)}}}^{=:\,\eta}\overline{X}^{(1)}_{\text{obs}} - \overline{X}^{(2)}, \overbrace{\frac{n^{(1)}_{\text{mis}}}{(n^{(1)})^2}\sigma_*^2}^{=:\,b^2_{\sigma_*^2}}\right)}{\sqrt{\underbrace{\left(\frac{1}{n^{(1)}} + \frac{1}{n^{(2)}}\right)}_{=:\,\xi} \cdot \chi^2_{\nu_{\text{Welch}}(S^2,\sigma_*^2)}}}, \\
&= \quad \frac{b_{\sigma_*^2}}{\sqrt{\xi}} \cdot \frac{1}{\sqrt{\nu_{\text{Welch}}(S^2,\sigma_*^2)}} \cdot \frac{N\big(\overbrace{\zeta\mu_* + \eta\overline{X}^{(1)}_{\text{obs}} - \overline{X}^{(2)}}^{=:\,\epsilon}, 1\big)}{\sqrt{\frac{\chi^2_{\nu_{\text{Welch}}(S^2,\sigma_*^2)}}{\nu_{\text{Welch}}(S^2,\sigma_*^2)}}}, \\
&= \quad \frac{b_{\sigma_*^2}}{\sqrt{\xi \cdot \nu_{\text{Welch}}(S^2,\sigma_*^2)}} \cdot \frac{N(0,1) - \epsilon_{(\mu_*, X_{\text{obs}})}}{\sqrt{\frac{\chi^2_{\nu_{\text{Welch}}(S^2,\sigma_*^2)}}{\nu_{\text{Welch}}(S^2,\sigma_*^2)}}}, \\
&= \quad \frac{b_{\sigma_*^2}}{\sqrt{\xi \cdot \nu_{\text{Welch}}(S^2,\sigma_*^2)}} \cdot t_{\nu_{\text{Welch}}(S^2,\sigma_*^2), \epsilon_{(\mu_*, X_{\text{obs}})}},
\end{aligned} \tag{4.19}$$

where $t_{\nu_{\text{Welch}}(S^2,\sigma_*^2), \epsilon_{(\mu_*, X_{\text{obs}})}}$ denotes a non-central $t$-distribution with noncentrality parameter $\epsilon_{(\mu_*, X_{\text{obs}})}$ and $\nu_{\text{Welch}}(S^2, \sigma_*^2)$ degrees of freedom.

Unfortunately at that point we are unable to go on with our analytical calculation, in order to integrate over $\mu_*$, then over $\sigma_*^2$ and finally over $X_{\text{obs}}$, all the quantities $\epsilon_{(\mu_*, X_{\text{obs}})}$, $b_{\sigma_*^2}$, $\nu_{\text{Welch}}(S^2, \sigma_*^2)$, and at last $S^2$ have to be random variables, whose distributions we cannot handle analytically. But anyway, for a large sample size the unknown variance $\sigma^2$ can be considered as known and we can use the $z$-test. As we have shown in Section 4.2, the $z$-transformation procedure works in that case, so for sufficient large $n$, the procedure also works well for the one-sided

$t$-test too. In general, the $z$-transformation procedure works for every hypothesis test, that can be approximated by the $z$-test.

We run a little simulation study to illustrate that the $z$-transformation procedure works well for the $t$-test. Using the settings described at the beginning of Section 4.2 with sample size $n^{(1)} = n^{(2)} = 1000$, $\gamma = 0.4$, i.e., $40\%$ of the values of $X_1$ are randomly deleted. Note that with $\gamma = 0.4$ we choose a relatively high missing rate to see differences between the distribution of the p-values before and after the $z$-transformation more clearly. We apply the imputation model (4.17). Then, for each of the $m = 5$ imputed data sets we calculate the corresponding p-value using the complete-data test-statistic, which is $t$-distributed with $(n^{(1)} + n^{(2)} - 2)$ degrees of freedom. Afterwards we transform these p-values to the corresponding $z$-values using the quantile function of the standard normal distribution and use these $z$-values for the multiple-imputation inference as described in Section 4.1. If the $z$-transformation works correctly, the resulting p-values (from $N = 10,000$ replications in our study) have to be uniformly distributed, because the null hypothesis $H_0 : \ \mu^{(1)} \leq \mu^{(2)}$ is set to be true when we generated the data. Figure 4.2 shows the distribution of the p-values after the imputation from one of the $m = 5$ imputations, the distribution of the corresponding (transformed) $z$-values and finally the distribution of the combined multiple imputation p-values across the $10,000$ replications. In addition Figure 4.2 shows the corresponding Q-Q-plots, where the quantiles of the p-values after imputation (first panel) and the quantiles of the combined multiple imputation p-values (third panel) are plotted against the quantiles of a uniform distribution on $[0, 1]$ and the quantiles of the $z$-transformed values (second panel) are plotted against a standard normal distribution.

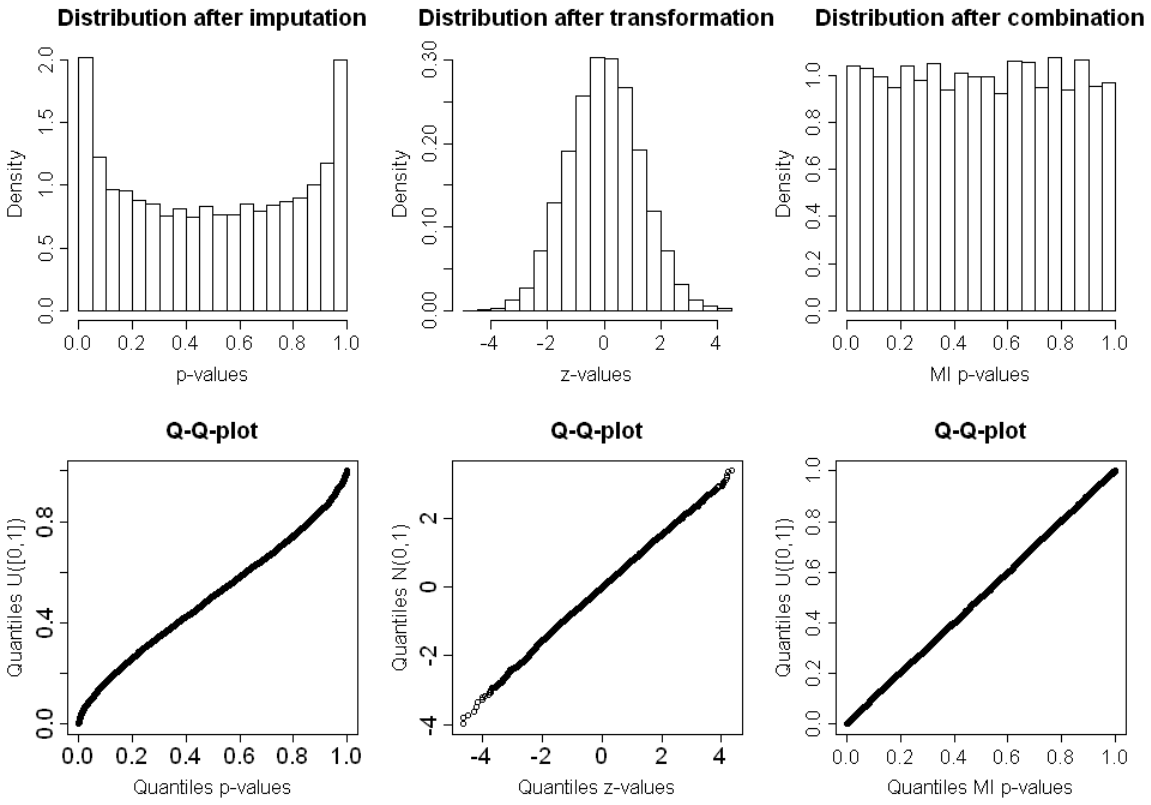


Figure 4.2: $z$-transformation for a one-sided t-test: 1st row: Histograms of the distribution of the p-values after one (single) imputation, the distribution of the transformed p-values (= $z$-values) and distribution of the combined MI p-values; 2nd row: Corresponding Q-Q-plots: 1st panel: Quantiles of the p-values after imputation plotted against the quantiles of $U[0,1]$; 2nd panel: Quantiles of the $z$-transformed values plotted against $N(0,1)$; 3rd panel: Quantiles of the $z$-values plotted against the quantiles of $U[0,1]$

From the left panel, we see that the p-values corresponding to one completed data set are not uniformly distributed, but bimodal at the tails, because the distribution of the test statistic changes as we have shown in the $z$-test calculations in Section 4.2; the variance is underestimated. After the application of the $z$-transformation to every completed data set and combining the transformed $m$ $z$-values, the p-values are uniformly distributed as they should be under the null hypothesis. Hence, the simulation study confirms that the $z$-transformation method is working for the one-sided $t$-test, too.

In addition to the $z$-test and the $t$-test we examine one-dimensional $F$-tests, e.g. to test equality of variances. The simulation study shows that the $z$-transformation holds for the $F$-test, too. Thus, our procedure is working well for all one-dimensional tests, because they can be linked to an $F$-test or to a $z$-test, respectively.

## 4.4 Wald-test

The Wald-test is a well known parametric statistical test, which is part of almost all standard statistical analyses and is implemented in nearly all statistical software packages. Given a statistical model, for example a linear regression model with parameters $\theta$ to be estimated from a sample, the Wald-test is often used to test if some or all of the parameters are equal to zero or generally equal to the true value $\theta_0$. In this case it tests whether the variables corresponding to these parameters have an influence on the dependent variable. Therefore, the Wald-test is often one of the first tests performed in a statistical analysis to specify the model chosen.

Let $X$ be an $(n \times k)$-data matrix where $X_i$ $(i = 1, \ldots, k)$ denotes the $i$th column vector of $X$. Here, the dimension $k$ equals the number of columns of $X$, because throughout this thesis we always consider the null hypothesis that all coefficients $\beta_1, \ldots, \beta_k$ of the linear regression model

$$Y = \beta_0 + X\beta + \epsilon = \beta_0 + \beta_1 X_1 \ldots + \beta_k X_k + \epsilon,$$

are equal to zero. $Y$ denotes the outcome variable and each component of $\epsilon$ is independent, identically distributed with zero mean and common variance $\sigma^2$. The Wald-test checks if the distance of the $k$-dimensional point-estimator $\hat{\theta} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)^t$ in the $k$-dimensional space is significantly different from the origin (or any arbitrary other null value) using the inverse of the variance-covariance matrix, $U = \text{Var}(\hat{\theta})$ of $\hat{\theta}$, as the metric. Thus, for the null hypothesis $H_0 : \beta_1 = \ldots = \beta_k = 0$ or in general $H_0 : \theta = \theta_0$ the Wald-statistic is of the form

$$D_w = (\hat{\theta} - \theta_0)^t U^{-1} (\hat{\theta} - \theta_0) \sim \chi^2_k. \tag{4.20}$$

As an alternative, the likelihood-ratio-test described in Section 3.2 can also be used. The likelihood-ratio-test can be more extensive in calculation but more precise for smaller samples. As mentioned in Section 3.2, the code for the likelihood-ratio-test is not provided by default in statistical software packages, whereas the Wald-test is the basic test in many statistical analyses.

To start with, we consider a one-dimensional Wald-test, that is we generate a $(n \times 1)$-data matrix $X$ with only $k = 1$ column vector $X = X_1$. Each component of $X_1$ is independent, identically standard normally distributed. We consider the underlying linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \epsilon, \tag{4.21}$$

where each component of the outcome variable $Y$ is independent, identically standard normally distributed and each component of $\epsilon$ is independent, identically distributed with zero mean and common variance $1$. We delete the last $\gamma = 40\%$ of the values of $X_1$. Throughout this thesis the outcome variable $Y$ is always fully observed. We impute $m = 5$ times the missing values of $X_1$ using a linear regression model based on the observed data as imputation model. Afterwards we perform a Wald-test testing if $\beta_1 = 0$ for each imputed data set, thereby producing $m = 5$ p-values in every replication. We apply the $z$-transformation procedure described in section 4.1 to get the multiple imputation p-value. Since there is no correlation between $X_1$ and $Y$, the null hypothesis $H_0 : \ \beta_1 = 0$ is true. For our simulation we use a sample size of $n = 1000$ and $N = 10,000$. Analogously to Figure 4.2, Figure 4.3 shows the distribution of the p-values after the imputation from one of the $m = 5$ imputations, the distribution of the corresponding (transformed) $z$-values, and finally the distribution of the combined multiple imputation p-values across the $10,000$ replications, as well as the corresponding Q-Q-plots.

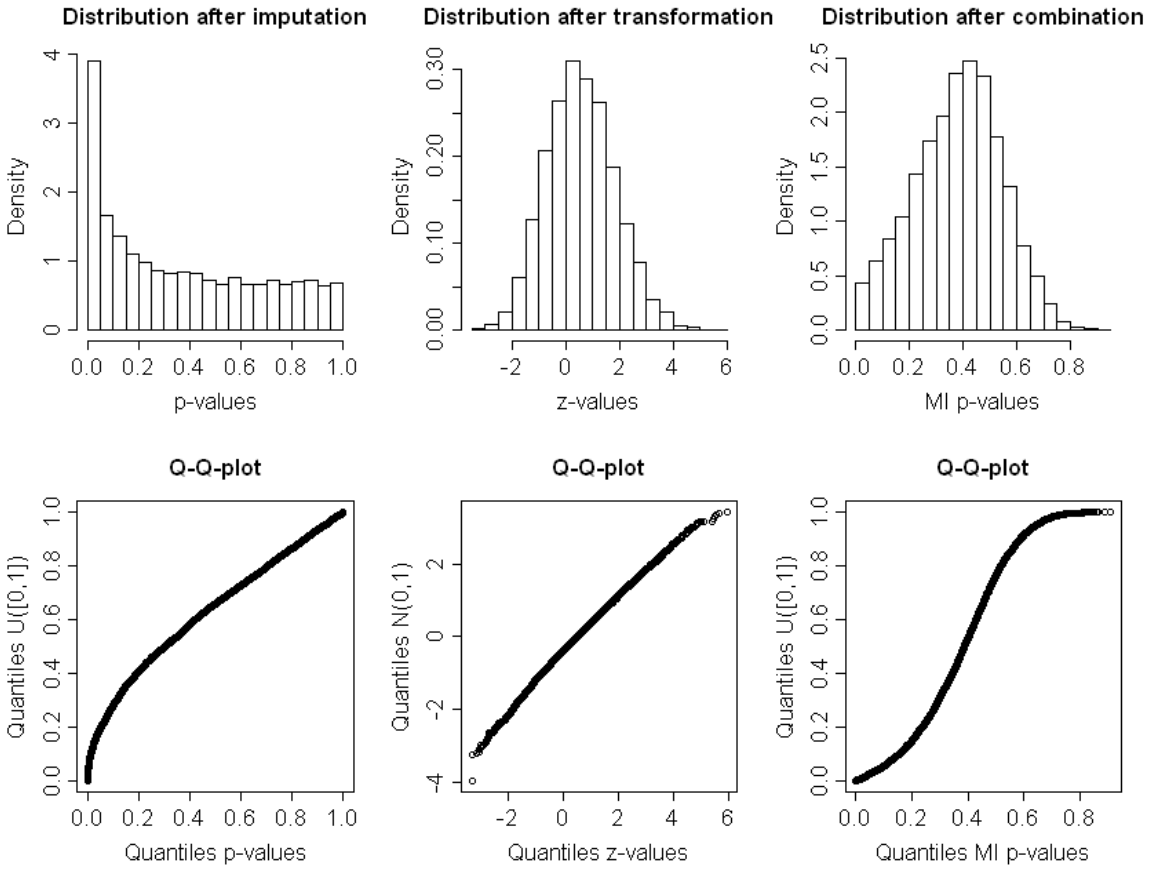


Figure 4.3: $z$-transformation for a one-dimensional Wald-test: 1st row: Histograms of the distribution of the p-values after one (single) imputation, the distribution of the transformed p-values (= $z$-values) and the distribution of the combined MI p-values; 2nd row: Corresponding Q-Q-plots: 1st panel: Quantiles of the p-values after one (single) imputation plotted against the quantiles of $U[0,1]$; 2nd panel: Quantiles of the $z$-transformed p-values plotted against $N(0,1)$; 3rd panel: Quantiles of the MI p-values plotted against the quantiles of $U[0,1]$

In the third panel of Figure 4.3 we see that the multiple imputation p-values are not uniformly distributed, as they should be under the null hypothesis. The reason lies in the skewness of the distribution of the p-values after the imputation (before transformation) as illustrated in the first panel of Figure 4.3. In contrast to the distribution of the p-values after the imputation (before transformation) from a one-sided $t$-test given in the first panel of Figure 4.2, where the distribution of the p-values after the imputation is bimodal at the tails, the distribution of

the p-values in here is skewed. This is due to the fact that the one-dimensional Wald-test is a two-sided test. Because of the skewness the $z$-transformation does not work, and does not lead to normally distributed $z$-values, which we need to use Rubin's (1987) usual combining rules.

Note, that our analytical calculations for the $z$-test and the $t$-test in sections 4.2 and 4.3 are based on the assumption that the hypotheses to be tested are one-sided. Similar non-symmetric skewed p-values as for the Wald-test also occur for two-sided $z$-tests and $t$-tests. We can solve this problem by separating a two-sided test into two one-sided tests: one with "less" and one with "greater" as alternative. For $k = 1$ the Wald-test has the same aim, but a different form like a two-sided $t$-test, that checks if the (only) coefficient $\beta_1$ of the regression model given in (4.21) equals $0$. The difference in the test statistic of the Wald-test is that the distance $(\hat{\beta}_1 - 0)$ is squared. Thus the reference distribution is an $F$-distribution (or a $\chi^2$-distribution for large $n$) instead of a $t$-distribution as for the $t$-test. To conduct a one-dimensional Wald-test, a two-sided $t$-test that is separated in two one-sided $t$-tests can be used. Then the $z$-transformation can be applied as shown before.

Because the $z$-transformation does not work for the one-dimensional Wald-test due to the skewed distribution of the p-values after the imputation, the $z$-transformation does not work for higher dimensional Wald-test with $k \geq 2$. The value of the Wald-statistic or the corresponding p-value, respectively, only provides the distance of the $k$-dimensional parameter vector estimate $\hat{\beta}$ to the origin of the $k$-dimensional sphere, but we have no further information. For example, if $k = 2$, all point-estimators $\hat{\beta}$ with the same distance to the origin are obviously on a circle around the origin with this distance as radius, whereas the corresponding p-value is the area outside of this circle as shown in Figure 4.4. Given only the value of the Wald-statistic or the corresponding p-value, respectively, we only know the radius of the circle, but we do not know in which quadrant of the two-dimensional vector space the point estimator $\hat{\beta}$ lies. That means that we do not have any information of the direction of the parameter

vector estimate. In the next chapter we suggest some solutions regarding this difficult problem.

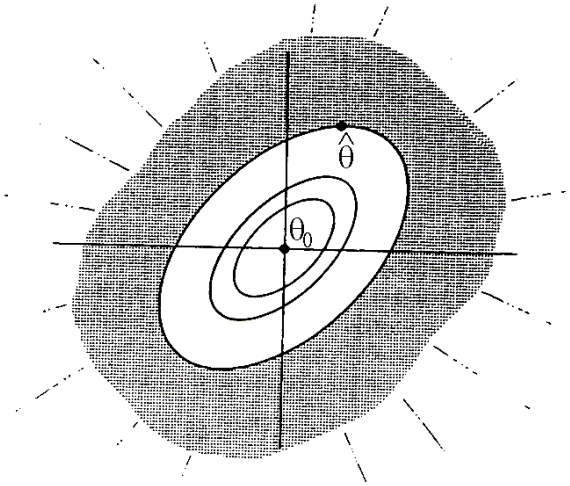


Figure 4.4: Contours of the distribution of $\hat{\theta}$ with the null value $\theta_0$ indicated. The corresponding p-value is the shaded area and beyond. (*Source: Rubin (1991), p.62, adapted to our notation*)

# 5

# How to handle the multi-dimensional test problem

## 5.1 Idea

To fix the two-sided test problem by using more information than just the set $S_d = \{d_{*1}, \ldots, d_{*m}\}$, the following procedure is proposed:

In addition to the collection of completed-data $\chi^2$-statistics, $S_d$, we need the p-values of $k$ one-sided $t$-tests for each completed data set, that is, we test if $\theta_i > 0$ for $i = 1, \ldots, k$. The $z$-transformation of these $(k \times m)$ p-values yields $(k \times m)$ corresponding $z$-values. Across the $m$ imputations, we calculate the between variances of these $z$-values yielding to the $(k \times k)$-diagonal between-variance-matrix $B_m$ assuming the sampling distributions of the $k$ estimated components are independent. We take the mean of the $k$ diagonal elements of $B_m$ times $(1 + m^{-1})$ as an estimator of the average relative increase in variance due to nonresponse, $r_m = (1+m^{-1})\text{Tr}(B_m\overline{U}_m^{-1})/k$, given in (2.9). Averaging over $B_m$ equates to calculating the trace of $B_m\overline{U}_m^{-1}$ divided by $k$, since the within variance-covariance matrix $U_m$ here is the identity matrix. This estimation of $r_m$ initially seems to be reasonable, because we use the additional information from the one-sided tests. For the test statistic we use $\hat{D}_m$, given by (3.10), which is based on the set $S_d$, and also used in Section 3.3. We replace $r_m$ by $(1+m^{-1})$ times the mean of the between variances

as described above. As reference distribution we take an $F$-distribution with $k$ and $w$ degrees of freedom, whereas $w$ given by (3.7) is the same denominator degree of freedom we use in the moment-based procedure described in Section 3.1.

We analyze the effectiveness of this procedure, which we call $\chi^2_k$-statistic-based method with additional information, by a simulation study described in the next section.

## 5.2 Simulation study

Let $X$ be an $(n \times k)$-data matrix where $X_i$ $(i = 1, \ldots, k)$ denotes the $i$th column vector of $X$. Each element of $X$ is independent, identically standard normally distributed. We consider the $k$-dimensional linear regression model

$$Y = \beta_0 + X\beta + \epsilon = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon,$$

where each component of the outcome variable $Y$ is independent, identically standard normally distributed. Each component of $\epsilon$ is independent, identically distributed with zero mean and common variance $1$. We delete in each of the column vectors $X_i$ the last $\gamma = 40\%$ of the values. All $X_i$ have the same fraction of missing information. Here, the fraction of missing information is equal to the missingness-rate, because the $X_i$ are all independent. That is, the data matrix $X$ is fully observed for the first $1 - \gamma = 60\%$ of the units. The remaining $40\%$ of the units are missing values as shown in Figure 5.1. The outcome variable $Y$ is always fully observed.

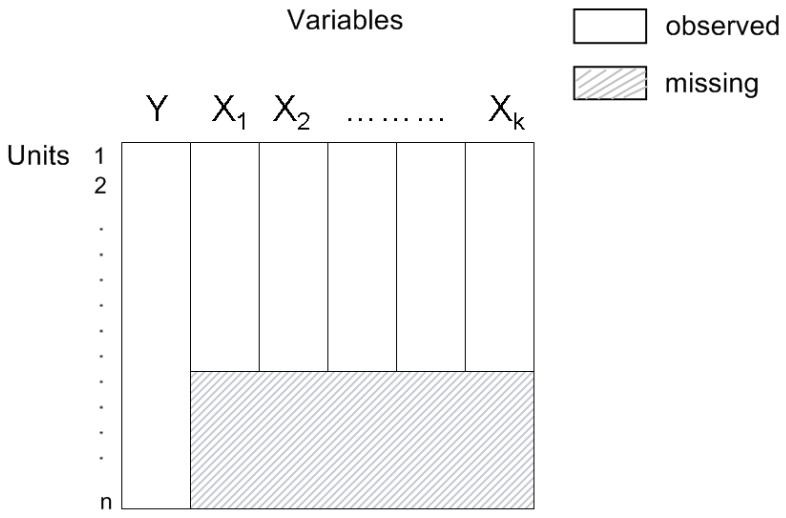


Figure 5.1: An example of a special monotone missingness pattern with $k$ variables, with all $X_i$ $(i = 1, \ldots, k)$ have the same last units missing : Solid = missing, White = observed

For the imputation model we use a linear regression model fitted for each variable $X_i$ $(i = 1, \ldots, k)$ with the previous variables as covariates. The missing values of $X_1$ are imputed from $Y$ ignoring the other components of $X$, then the missing values of $X_2$ are imputed from $(X_1, Y)$ ignoring the other components of $X$ and so on. After $m = 5$ imputations we perform a Wald-test on $H_0 : \beta_1 = \ldots = \beta_k = 0$ with each of the $m = 5$ imputed data sets yielding to the set $S_d$. Then we follow the $\chi^2_k$-statistic-based procedure with additional information described in Section 5.1, which produces the multiple imputation p-value of the suggested procedure. In addition we apply the moment-based procedure to the same multiple imputed data sets and get the multiple imputation p-value from this method. Across $5000$ replications we compare the p-values from our $\chi^2_k$-statistic-based procedure with additional information and the moment-based procedure, because Li, Raghunathan, and Rubin (1991) analytically have shown that the moment-based procedure works well in large samples. The results for different dimensions $k$ with $k = 2, 10, 20, 50, 75$ and a sample size of $n = 1000$ are illustrated in Figure 5.2, where the first row shows the distribution of the multiple imputation p-value with the $\chi^2_k$-statistic-based procedure with additional information and the second row

the results from the moment-based procedure. The third and fourth row show the corresponding Q-Q-plots where the distributions of the multiple imputation p-values are plotted against a uniform distribution on $[0, 1]$.

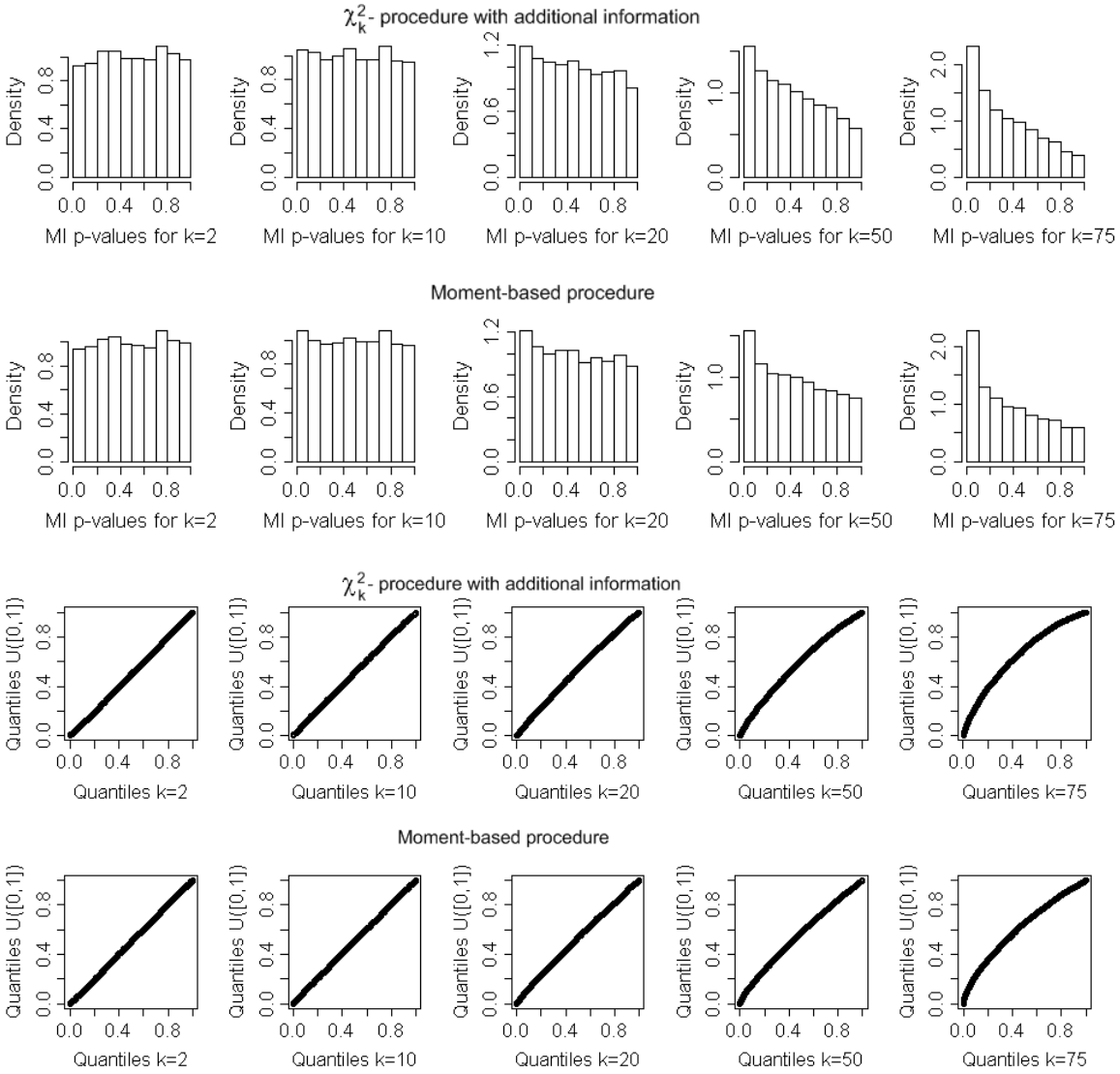


Figure 5.2: $k$-dimensional Wald-test with increasing $k$: 1st row: Histogram of the distribution of the MI p-values using $\chi^2_k$-statistic-based procedure with additional information; 2nd row: Histogram of the distribution of the MI p-values using moment-based procedure; 3rd row: Corresponding Q-Q-plot: Quantiles of the MI p-values plotted against $U[0, 1]$ for $\chi^2_k$-statistic-based procedure with additional information; 4rd row: Corresponding Q-Q-plot: Quantiles of the MI p-values plotted against $U[0, 1]$ for moment-based procedure

This study points out something interesting. From $k = 2$ until $k = 20$ both methods produce essentially uniformly distributed p-values, as we expected for the moment-based procedure and for the $\chi^2_k$-statistic-based procedure with additional information. For $k > 20$ both methods tend to be more and more too significant even with $m$ much larger than $5$, which is surprising, especially for the moment-based procedure. But all derivations in Li, Raghunathan, and Rubin (1991) were based on the assumption that the sample size $n$ is large. As we see in the simulation study, a sample size of $n = 1000$ is too small for larger $k$. Note that for large $k$, a small change in the radius of the sphere, determined by the estimated regression coefficients, has a major effect on the volume of that sphere, which corresponds to one minus the p-value. Due to the lack of computer power in the late eighties, when the moment-based procedure was developed, they did not run simulation studies with higher sample size and higher dimensions of $k$, as we can do today. We run the same simulation study with a sample size of $n = 5000$ and the same $k$. Now we get essentially uniformly distributed p-values for both procedures, even for $k = 75$.

This simulation study shows, that the moment-based procedure and our $\chi^2_k$-based method with additional information with higher dimension and too small sample sizes break down.

## 5.3 Further problems

The problem of a sample size being too small in relation to the dimension $k$ seems to be an unnoted general problem in multivariate statistics, especially in statistical test theory. In many hypothesis tests, as for example the often used Wald-test, a $\chi^2_k$-distribution is used as reference distribution, assuming that the sample size is sufficiently large. Otherwise, especially with higher dimension $k$, an $F$-distribution has to be used as reference. A reason for using the $\chi^2_k$-distribution rather than the $F$-distribution may be that it was much harder in times of poor computer performance to provide quantile-tables and probability-tables for an

$F$-distribution than for a $\chi^2_k$-distribution, because the $F$-distribution has two different degrees of freedom. To get an impression what it means to use a $\chi^2_k$-distribution rather than an $F$-distribution, we calculate the corresponding $\alpha$-levels for $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$ when we wrongly use a $\chi^2_k$-distribution. They are shown in the following Table 5.1:

| | $\boxed{\alpha = 0.01}$ | | | | $\boxed{\alpha = 0.05}$ | | | |
|---|---|---|---|---|---|---|---|---|
| $d_1$ \ $d_2$ | 100 | 500 | 1000 | 5000 | 100 | 500 | 1000 | 5000 |
| 2 | 0.012 | 0.010 | 0.010 | 0.010 | 0.054 | 0.051 | 0.050 | 0.050 |
| 4 | 0.014 | 0.010 | 0.010 | 0.010 | 0.057 | 0.052 | 0.051 | 0.050 |
| 10 | 0.017 | 0.011 | 0.011 | 0.010 | 0.065 | 0.053 | 0.052 | 0.050 |
| 20 | 0.022 | 0.012 | 0.011 | 0.010 | 0.075 | 0.055 | 0.053 | 0.051 |
| 35 | 0.030 | 0.013 | 0.012 | 0.010 | 0.090 | 0.058 | 0.054 | 0.051 |
| 50 | 0.038 | 0.015 | 0.012 | 0.010 | 0.103 | 0.061 | 0.056 | 0.051 |
| 75 | 0.051 | 0.017 | 0.013 | 0.011 | 0.122 | 0.065 | 0.058 | 0.052 |
| 100 | 0.064 | 0.019 | 0.014 | 0.011 | 0.139 | 0.070 | 0.061 | 0.052 |

$\boxed{\alpha = 0.1}$

| $d_1$ \ $d_2$ | 100 | 500 | 1000 | 5000 |
|---|---|---|---|---|
| 2 | 0.106 | 0.101 | 0.100 | 0.100 |
| 4 | 0.109 | 0.102 | 0.101 | 0.100 |
| 10 | 0.118 | 0.104 | 0.101 | 0.100 |
| 20 | 0.130 | 0.106 | 0.103 | 0.101 |
| 35 | 0.147 | 0.110 | 0.105 | 0.101 |
| 50 | 0.161 | 0.114 | 0.107 | 0.102 |
| 75 | 0.182 | 0.120 | 0.110 | 0.102 |
| 100 | 0.199 | 0.125 | 0.113 | 0.103 |

Table 5.1: Rejection rates from uniformly distributed p-values when using a $\chi^2_{d_1}$-distribution rather than an $F_{d_1,d_2}$-distribution, for $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$; $d_1 =$ numerator degrees of freedom, $d_2 =$ denominator degrees of freedom

With the true $F$-distribution our rejection rates are equal to the nominal levels $1\%$, $5\%$ and $10\%$ across $1,000,000$ replications we simulated. The table shows that especially for a larger numerator degree of freedom, $d_1$, which corresponds to the dimension $k$ in the Wald-test, and a relatively small denominator degree of freedom, $d_2$, which corresponds approximately to the sample size $n$, the rejection rates are really bad. We also see that with a larger $d_2$ or sample size, respectively, the rejection rates are getting better and for $d_2 = 5000$ we achieve the correct rejection rates for almost all considered $d_1$ and $\alpha$. Thus, especially in practical applications with hundreds of variables, one should use an $F$-distribution as the reference distribution, because usually a sufficiently large sample size is not provided by the data set to justify the $\chi^2_k$-approximation.

In all following simulation studies we use the $F$-distribution as reference and calculate the corresponding quantile in the $\chi^2_k$-distribution. Hence, we "convert" from the $F$-distribution into the $\chi^2_k$-distribution, because our procedure and the theoretical derivations, presented in Chapter 3 are based on the set $S_d$, i.e., on the $\chi^2_k$-test statistics.

The problem of a sample size being too small related to the dimension $k$ is especially important for the moment-based method, described in Chapter 3, because all derivations and calculations are based on the assumption of large sample size relative to the dimension $k$. Thus, the moment-based method does not work for a small sample size, as our simulation study in this chapter in Section 5.2 shows, even if the respective requirements like providing the variance-covariance matrices are fulfilled. It is difficult to fix the procedure analytically, because with small sample size, $(\hat{\theta} - \theta)$ is not normally distributed as required in (2.1), but rather $t$-distributed. All further calculations based on this assumption turn to be very complicated, because the $t$-distribution is not as manageable as is the normal distribution.

On the one hand, the requirements of the existing procedures are too demanding, like for the two-stage-likelihood-ratio-test-based procedure or the moment-based

method. On the other hand, the $\chi^2_k$-statistic-based approach, which has the weakest requirements, is not well calibrated. The suggested $\chi^2_k$-statistic-based procedure with additional information does not have strong requirements, but it turned out that this procedure does not work for a higher dimension $k$, when the sample size is small. This is also a problem for the moment-based method. Thus, finding a new procedure is an important and challenging task.

# 6

# Small-sample significance levels from repeated p-values using a componentwise-moment-based method

## 6.1 Small-sample degrees of freedom with multiple imputation

As we described in Chapter 2 (see (2.6) and (2.7)), Rubin (1987) showed that

$$(\overline{\theta}_m - \theta) \sim t_\nu(o, T_m), \tag{6.1}$$

with

$$\nu = (m-1)(1 + r_m^{-1})^2, \tag{6.2}$$

based on the large sample size assumption, which implies that $(\hat{\theta} - \theta)$ can be assumed normally distributed. Barnard and Rubin (1999) derived degrees of freedom, $\widetilde{\nu}$ for small sample sizes, making the assumption that $(\hat{\theta} - \theta)$ is $t$-distributed instead of normally distributed. The need for small-sample degrees of freedom and thus generating small-sample significance levels, arise because $\nu$ can be many times the degrees of freedom available if there were no missing data. This is often the case in small data sets, when due to the large sample assumption, the

complete-data degrees of freedom are set to be infinity and there is only a small fraction of missing information $\gamma$.

Let

$$\hat{\gamma}_m = (1 - m^{-1})Tr(B_m T_m^{-1})/k \tag{6.3}$$

be the approximate fraction of missing information, $\gamma$, given in (2.10).

Let $\nu = (m-1)(1 + r_m^{-1})^2$ the usual multiple imputation degrees of freedom given in (2.7) and let

$$\hat{\nu}_{\text{obs}} = \frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3} \nu_{\text{com}} (1 - \hat{\gamma}_m), \tag{6.4}$$

where $\nu_{\text{com}}$ denotes the complete-data degrees of freedom. Then the adjusted small-sample degrees of freedom suggested by Barnard and Rubin (1999) are defined as the harmonic total of $\nu$ and $\hat{\nu}_{\text{obs}}$ according to

$$\tilde{\nu}_m = \left( \frac{1}{\nu} + \frac{1}{\hat{\nu}_{\text{obs}}} \right)^{-1} = \nu_{\text{com}} \left[ \frac{\nu_{\text{com}} + 3}{(\nu_{\text{com}} + 1)(1 - \hat{\gamma}_m)} + \frac{\nu_{\text{com}}}{\nu} \right]^{-1}. \tag{6.5}$$

If $\nu_{\text{com}} = \infty$, that is, if the sample size is infinitely large, $\tilde{\nu}_m$ equals the multiple imputation degrees of freedom $\nu$ derived by Rubin (1987).

## 6.2 Significance levels from multiply imputed data with small sample size based on $\tilde{S}_d$

Due to the fact that the small-sample degrees of freedom from Barnard and Rubin (1999) are not valid in the $k$-dimensional case with $k > 1$, we suggest the following componentwise procedure for combining repeated p-values based on their small-sample degrees of freedom. In an analogous manner as the procedure proposed by Li, Raghunathan, and Rubin (1991), we also use the test statistic $\hat{D}_m$ given in (3.10) and estimate the relative increase in variance due to nonresponse $r_m$ as follows. Analyzing the $m$ completed data sets gives us the $m$ point estimates $\hat{\theta}_{*l}$, with

$l = 1, \ldots, m$. We assume that the components of each $\hat{\theta}_{*l}$ are independent, i.e., there is no correlation between them. Thus, the variance-covariance matrices $U_{*l}$ are diagonal matrices and they can be considered as $(k \times 1)$-vectors instead of $(k \times k)$-matrices. Thus, we need the vectors of variances, $u_{*l}$, of each $\hat{\theta}_{*l}$, instead of the whole variance-covariance matrices $U_{*l}$. From this it follows that the within variance $\overline{U}_m$ given by (2.3), also is specified by a $(k \times 1)$-vector instead of a $(k \times k)$-matrix. The between variance matrix $B_m$ given by (2.4), is a diagonal matrix, because of the simplifying assumption that the missing values are independent draws from their posterior distribution. This is the reason, why we use $B_m$ as a $(k \times 1)$-vector consisting of the diagonal elements of $B_m$. Therefore the approximate fraction of missing information $\hat{\gamma}_m$ given in (6.3), the multiple imputation degrees of freedom $\nu$ given in (2.7), the observed degrees of freedom $\nu_{\text{obs}}$ given in (6.4) and finally the small-sample degrees of freedom $\tilde{\nu}_m$ given in (6.5), are all calculated componentwise. Let

$$r = (1 + m^{-1}) \begin{pmatrix} b_1/\overline{u}_1 \\ \vdots \\ b_k/\overline{u}_k \end{pmatrix} = \begin{pmatrix} r_1 \\ \vdots \\ r_k \end{pmatrix} \tag{6.6}$$

be the componentwise calculated relative increase in variance based on (2.8), where $b_i$ and $\overline{u}_i$ with $i = 1, \ldots, k$ are the components of the between and within variance vectors $B_m$ and $\overline{U}_m$.

Our proposed estimation of $r_m$ in $\hat{D}_m$, $\ddot{r}_m$, is a function of $r$ (given in (6.6)), which is, for example, the mean, the median, the minimum or the maximum of the $r_i$. After testing several choices in a simulation study, it turns out that the mean of the $r_i$ is a very robust choice, because the maximum leads to conservative p-values and the minimum to liberal p-values and the mean seems to be a stable compromise. Important for the distribution of the p-values is the choice of the function $v = f(\tilde{\nu}_m)$. Finding the correct denominator degrees of freedom is one of the major tasks of this works. The simulation described in Chapter 7 shows that we have to take at least the maximum of the components of $\tilde{\nu}_m$, $\tilde{\nu}_{m,i}$ $(i = 1, \ldots, k)$ as an estimation of the denominator degrees of freedom for our

$F$-reference distribution. Taking the mean or the minimum, for example, leads to too conservative p-values. Thus, we choose $v$ as the maximum of the $\tilde{\nu}_{m,i}$:

$$v = \max_{i \in \{1,\ldots,k\}} \{\tilde{\nu}_{m,i}\}. \tag{6.7}$$

In summary, we suggest the test statistic $\ddot{D}_d$ based on the set $\ddot{S}_d = \{d^f_{*l}, u_{*l};\ l = 1, \ldots, m\}$

$$\ddot{D}_d = \frac{\overline{d} \cdot k^{-1} + \frac{m-1}{m+1} \cdot \ddot{r}_m}{1 + \ddot{r}_m} \sim F_{k,v}, \tag{6.8}$$

where $\ddot{r}_m$ is the mean of the componentwise calculated relative increases in variance, $v$ is the maximum of the componentwise calculated small-sample degrees of freedom, $\overline{d}$ is the mean of the $d^f_{*l}$ and the superscript $f$ denotes that the $\chi^2_k$-test statics actually are $F$-statistics which were transformed to the corresponding $\chi^2_k$-statistic due to the small sample size, which we mentioned in Section 5.3.

Indeed, the $\chi^2_k$-statistic-based procedure with the $z$-transformated $t$-tests as additional information, does not work for a higher dimension $k$, because the loss of information using just the one-sided test statistics is too big. But we suggest a procedure, that uses the Barnard and Rubin (1999) small-sample degrees of freedom and thus, will hopefully fix the problem of small sample sizes. The set $\ddot{S}_d$, on which our procedure is based, only needs the $m$ $\chi^2_k$-test statistics and the $m$ standard error tests of each component, which are standard in every statistical software package.

Note that the componentwise-moment-based procedure and the moment-based procedure are asymptotically equivalent, because for a large sample size $n \to \infty$ all variance-covariance matrices of the $m$ point estimators $\hat{\theta}_{*l}$, $U_{*l}$, are supposed to be the $(k \times k)$-identity matrix. Hence, the components of each $\hat{\theta}_{*l}$ are assumed to be independent. From this it follows, that our proposed estimation of the relative increase in variance, $\ddot{r}_m = \text{mean}(r_i)$ given in (6.6), is asymptotically equal to the average relative increase in variance due to nonresponse, $r_m$ given in (2.9). The within variance-covariance matrix $\overline{U}_m$ given in (2.3) is also supposed to

be a $(k \times k)$-identity matrix, when all $U_{*l}$ equal the $(k \times k)$-identity matrix. Thus the test statistic $\tilde{D}_m$ given in (2.12) used for the moment-based procedure and the test statistic $\hat{D}_m$ given in (3.10) used for the componentwise-moment-based procedure are asymptotically equivalent (for a proof see Rubin (1987), p.100). Hence, the componentwise-moment-based procedure and the moment-based procedure are asymptotically equivalent.

In the next chapter we describe a large-scale simulation study that compares all four methods to generate significance levels from multiply-imputed data using the existing and the small-sample degrees of freedom. We want to find out if the componentwise-moment-based procedure is well calibrated and give practical advice when to use which procedure.

# 7

# Comparing the four methods for generating significance levels from multiply-imputed data

To analyze the behavior of the three existing procedures (moment-based method, two-stage-likelihood-ratio-test-based method, and the $\chi_k^2$-statistic-based method) for combining repeated p-values described in Chapter 3, and especially the validity of the new componentwise-moment-based procedure described in Chapter 6, we perform an extensive simulation study using the statistical software package **R**. There are already several simulation studies for the three former procedures done by the authors in the early $90$'s, but due to the computer power at that time, their simulations use draws from analytically derived distributions (see e.g. Li, Raghunathan, and Rubin (1991), page 1067). They did not generate data sets and impute missing values. We study these earlier methods and the componentwise-moment-based procedure using today's computer power.

## 7.1 Simulation study

Our simulation can be described as a $3 \times 2 \times 6 \times 3 \times 4^3 \times 8$ factorial experiment with $55296$ different situations resulting from the following settings that are summarized in Table 7.1.

| Factor | Levels of factor |
|---|---|
| $\alpha$ (nominal level) | $\{1\%, 5\%, 10\%\}$ |
| $n$ (sample size) | $\{1000, 5000\}$ |
| $k$ (dimension) | $\{2, 5, 10, 20, 35, 50\}$ |
| $\overline{\xi}$ (average ratio of complete to observed information, (3.1)) | $\{1.2, 1.5, 2\}$ |
| $C_\xi$ (coefficient of variation of the $\xi_i$, (3.2)) | $\{0, 0.1, 0.2, 0.4\}$ |
| $m$ (number of imputations) | $\{5, 10, 20, 30\}$ |
| Method | {moment-based method, two-stage-likelihood-ratio-test-based method, $\chi^2_k$-statistic-based method, componentwise-moment-based procedure} |
| $\nu$ (degrees of freedom) | $\{w, a_{k,m}w_s, v, (k+1)v/2, v_{\text{mean}}, (k+1)v_{\text{mean}}/2, v_{\text{min}}, (k+1)v_{\text{min}}/2\}$ |

Table 7.1: Factorial design - simulation factors with their levels

We study three levels of the nominal level $\alpha$, $\{1\%, 5\%, 10\%\}$, two levels of the sample size $n$, $\{1000, 5000\}$, six levels of the dimension $k$, $\{2, 5, 10, 20, 35, 50\}$, three levels of the average ratio of complete to observed information $\overline{\xi} = \frac{1}{k}\sum_{i=1}^{k}\xi_i$ (given in (3.1)), $\{1.2, 1.5, 2\}$, four levels of variation among components of $\theta$ in the complete to observed information $C_\xi$ (given in (3.2)), $\{0, 0.1, 0.2, 0.4\}$, four levels of the number of imputations $m$, $\{5, 10, 20, 30\}$, four levels of the method used, {moment-based method, two-stage-likelihood-ratio-test-based method, $\chi^2_k$-statistic-based method, componentwise-moment-based procedure}, and eight levels of denominator degrees of freedom, $\{w, a_{k,m}w_s, v, (k+1)v/2, v_{\text{mean}}, (k+1)v_{\text{mean}}/2, v_{\text{min}}, (k+1)v_{\text{min}}/2\}$.

For each simulation situation, i.e. for each combination of factor levels, we generate a $(n \times 1)$-data vector $Y$ where each component of the outcome variable $Y$ is independent, identically standard normally distributed and a $(n \times k)$-data matrix $X$ where each element of $X$ also is independent, identically standard normally distributed. Let $X_i$ $(i = 1, \ldots, k)$ denote the $i$th column vector of $X$. We consider the $k$-dimensional linear regression model

$$Y = \beta_0 + X\beta + \epsilon = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon, \tag{7.1}$$

where each component of $\epsilon$ is independent, identically distributed with zero mean and common variance $1$. For each combination of $k$, $\overline{\xi}$ and $C_\xi$, the $k$ values of $\xi_i$ are selected to have mean $\overline{\xi}$ and a coefficient of variation $C_\xi$ by drawing $k$ values of $\lambda_i = \xi_i - 1$ from a gamma distribution with shape $((\overline{\xi} - 1)/C_\xi \cdot \overline{\xi})^2$ and scale $(C_\xi \cdot \overline{\xi})^2/(\overline{\xi} - 1)$. With the definition of the $\xi_i$ given in (3.1), it follows that the fractions of missing information $\gamma_i$ are determined as

$$\gamma_i = 1 - \frac{1}{1 + \lambda_i} \qquad \text{for } i = 1, \ldots, k.$$

Because in our simulation, the $k$ variables $X_1, \ldots, X_k$ are all independent, the $\gamma_i$ are equal to the missingness rate. Depending on the coefficient of variation, $C_\xi$, the last values of the $X_i$ are deleted with the increasing ordered fractions of missing information, $\gamma_i$, such that a monotone missingness pattern is generated. A data set $X$ with $k$ column vectors $X_i$ $(i = 1, \ldots, k)$ as independent variables has a monotone missingness pattern when a variable $X_i$ is missing for an individual implies that all variables $X_{i+1}, X_{i+2}, \ldots, X_k$ are all missing for that individual $p$ as shown in Figure 7.1.

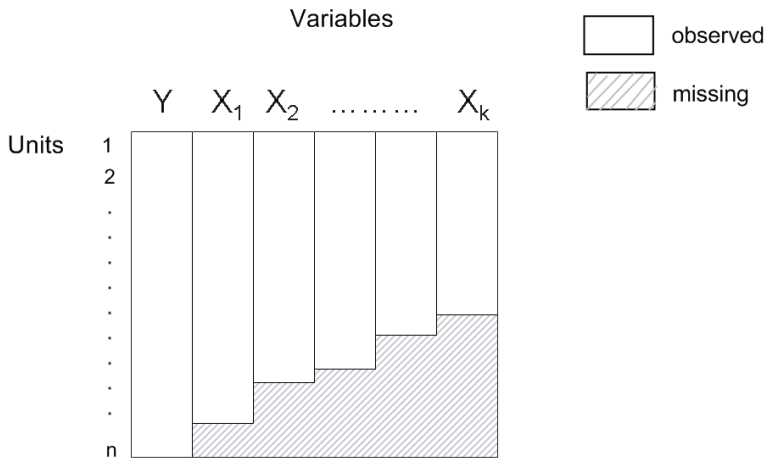


Figure 7.1: An example of a monotone missingness pattern with $k$ variables, with $X_i$ less missing than $X_{i+1}$: Solid = missing, White = observed

As a special monotone pattern we consider the case where all fractions of missing information, $\gamma_i$, are equal and thus the coefficient of variation, $C_\xi$, equals zero (see Figure 5.1 in Section 5.2). The outcome variable $Y$ is always fully observed.

For the imputation model we use a linear regression model fitted for each variable $X_i$ $(i = 1, \ldots, k)$ with the previous variables as predictors. The missing values of $X_1$ are imputed from $Y$ ignoring the other components of $X$, then the missing values of $X_2$ are imputed from $(X_1, Y)$ ignoring the other components of $X$ and so on. With the multiply-imputed data set, we create the required sets $S_m$, $S_d$ and $\ddot{S}_d$, and the function $d$, described in Chapters 3 and 6. Based on the set $S_m$ we calculate the test statistic $\tilde{D}_m$ defined by (2.12), and use the $F_{k,w}$-distribution as reference distribution with $w$ given by (3.7) to calculate the multiple imputation p-value from the moment-based method. Based on the function $d$ we calculate the test statistic $\breve{D}$ defined by (3.8), and use the same reference distribution $F_{k,w}$ to calculate the multiple imputation p-value from the two-stage-likelihood-ratio-test-based procedure. Based on the set $S_d$ we calculate the test statistic $\hat{\hat{D}}_d$ defined by (3.10), and (3.11) and use the reference distribution $F_{k,a_{k,m}w_s}$ given by (3.13) and (3.12) to calculate the multiple imputation p-value from the $\chi^2_k$-statistic procedure.

Finally, based on the set $\ddot{S}_d$ we calculate the test statistic $\ddot{D}_m$ defined by (6.8), and use the reference distribution $F_{k,v}$ given by (6.7) to calculate the multiple imputation p-value from the componentwise-moment-based procedure.

In the simulation study we do not only use these $F_{k,\cdot}$-distributions as reference distributions for the respective corresponding method, but also as reference distribution for each other method. In addition to these original reference distributions for each of the four methods, we use five further $F_{k,\cdot}$-reference distributions with $\{(k+1)v/2, v_{\text{mean}}, (k+1)v_{\text{mean}}/2, v_{\text{min}}, (k+1)v_{\text{min}}/2\}$ as particular denominator degrees of freedom. For "$v_{\text{min}}$" we take the minimum of the componentwise calculated degrees of freedom (6.7), each calculated according to Barnard and Rubin (1999). Analogously, for "$v_{\text{mean}}$" we take the mean of the componentwise calculated degrees of freedom, each calculated according to Barnard and Rubin (1999). The multiplication of $v$, $v_{\text{mean}}$, and $v_{\text{min}}$ with the factor $(k+1)/2$ is based on a suggestion from Rubin (1987, p.99). The denominator degrees of freedom $(k+1)v/2$ ($(k+1)v_{\text{mean}}/2$, $(k+1)v_{\text{min}}/2$, respectively) are chosen simply because it is halfway between the minimum degrees of freedom $v$ ($v_{\text{mean}}$, $v_{\text{min}}$, respectively) and the maximum degrees of freedom $kv$ ($kv_{\text{mean}}$, $kv_{\text{min}}$). We examine each of the four test statistics $\tilde{D}_m$, $\breve{D}$, $\hat{\hat{D}}_d$, and $\ddot{D}_m$ with the eight different reference distributions $F_{k,w}$, $F_{k,a_{k,m}\cdot w_s}$, $F_{k,v}$, $F_{k,(k+1)v/2}$, $F_{k,v_{\text{mean}}}$, $F_{k,(k+1)v_{\text{mean}}/2}$, $F_{k,v_{\text{min}}}$, and $F_{k,(k+1)v_{\text{min}}/2}$. The null hypothesis being tested is always $H_0 : \beta_1 = \ldots = \beta_k = 0$, i.e., we perform a $k$-dimensional Wald-test, where $\beta_1, \ldots, \beta_k$ denote the regression coefficients of the linear regression model given in (7.1). The corresponding multiple imputation p-values for each combination of $\alpha$, $n$, $k$, $\overline{\xi}$, $C_\xi$, and $m$ are calculated using the four different methods with the eight different $F_{k,\cdot}$-reference distributions, described above. We will see that the "right" choice of the denominator degrees of freedom is very important and is critical for their valid use.

In the construction of the simulation, we nest the sample size $n$, that is, from a sample of size $5000$, we create a subsample of size $1000$. Analogously, for the number of imputations, $m$, we first generate the maximal number of imputations, $m = 30$, and from this we take the imputation settings with smaller numbers of

imputations. Using these subsamples to generate and impute the data, reduces the variance between the different situations in each replication. Thus, these nestings increase the precision in the comparison of the rejection rates based on the simulation. In addition, the nesting of the replications saves computational time. The simulation study was done with $N = 10,000$ replications.

## 7.2 Results

### 7.2.1 ANOVA

The multiple imputation p-values from $N = 10,000$ replications for each situation of our factorial experiment lead to the rejection rate of the test at any nominal level $\alpha$. We examine the deviation of the simulated rejection rates to the three corresponding nominal levels $\alpha = \{1\%, 5\%, 10\%\}$, transform them to be more normally distributed using a logit-transformation suggested in Cangul, Chretien, Gutman, and Rubin (2009), and perform an ANOVA on these values to identify the most important factors for the distribution of the multiple imputation p-values. Since the results for the particular nominal levels $\alpha$ are very similar, we present only the table for $\alpha = 5\%$ here. Table 7.2 presents the ANOVA on the deviation of the simulated rejection rates from nominal level $\alpha = 5\%$ for the main effects and the two-way interactions denoted by "$*$" for seven factors. The factors are presented in descending order according to their Mean Squared Errors.

| Factors | Df | Mean Squared Error |
|---|---|---|
| *method* | 3 | 560900 |
| $\nu$ | 7 | 560432 |
| $m$ | 3 | 110183 |
| $m * \nu$ | 21 | 89966 |
| $k * \nu$ | 35 | 67756 |
| $\overline{\xi} *$ *method* | 6 | 65989 |
| $k$ | 5 | 51274 |
| $k *$ *method* | 15 | 42880 |
| $\overline{\xi} * \nu$ | 14 | 26713 |
| $m *$ *method* | 9 | 25813 |
| $k * m$ | 15 | 12782 |
| $\overline{\xi}$ | 2 | 7264 |
| $\nu *$ *method* | 21 | 7149 |
| $C_\xi$ | 3 | 6248 |
| $\overline{\xi} * m$ | 6 | 4682 |
| $n$ | 1 | 4602 |
| $n *$ *method* | 3 | 4256 |
| $k * C_\xi$ | 15 | 2266 |
| $C_\xi * \nu$ | 21 | 2049 |
| $k * \overline{\xi}$ | 10 | 1860 |
| $n * C_\xi$ | 3 | 1407 |
| $\overline{\xi} * \nu$ | 6 | 792 |
| $n * \overline{\xi}$ | 2 | 415 |
| $n * \nu$ | 7 | 397 |
| $C_\xi *$ *method* | 9 | 358 |
| $C_\xi * m$ | 9 | 175 |
| $n * k$ | 5 | 158 |
| Residuals | 18172 | 129 |
| $n * m$ | 3 | 5 |

Table 7.2: Deviation of rejection rates from nominal level $\alpha = 5\%$ from simulation: ANOVA for main effects and two-way interactions for seven factors

The three factors that most strongly affect the rejection rates are $(i)$ the method; $(ii)$ the degrees of freedom, $\nu$; and $(iii)$ the number of imputations, $m$. These three factors account for $74\%$ of the p-value variance at the $\alpha = 5\%$-level. To gain more insight into the behavior of the rejection rates, we examine the four methods given in the factorial design in Table 7.1 with their corresponding originally proposed degrees of freedom separately, more specifically the moment-based method with $w$ given in (3.7) as degrees of freedom, the two-stage-likelihood-ratio-test-based

procdure also with $w$ given in (3.7) as degrees of freedom, the componentwise-moment-based method with $v$ given in (6.7) as degrees of freedom, and the $\chi^2_k$-statistic-based procedure with $a_{k,m} \cdot w_s$ given in (3.13) and (3.12) as degrees of freedom. In Tables 7.3 to 7.6 are: The results of the two-way interaction ANOVA of the deviation of the rejection rates from nominal level $\alpha = 5\%$ for: the moment-based method; the two-stage-likelihood-ratio-test-based method; the $\chi^2_k$-statistic-based method, and the componentwise-moment-based method with two-way interaction.

Table 7.3 presents the two-way interaction ANOVA of the deviation of the rejection rates from nominal level $\alpha = 5\%$ for the moment-based method.

| Factors | Df | Mean Squared Error |
|---|---|---|
| $n$ | 1 | 1822.04 |
| $k$ | 5 | 782.25 |
| $\overline{\xi}$ | 2 | 270.64 |
| $n * k$ | 5 | 139.98 |
| $n * \overline{\xi}$ | 2 | 131.20 |
| $k * \overline{\xi}$ | 10 | 47.06 |
| $C_\xi$ | 3 | 35.39 |
| $n * C_\xi$ | 3 | 11.72 |
| $k * m$ | 15 | 7.49 |
| $k * C_\xi$ | 15 | 4.13 |
| $\overline{\xi} * C_\xi$ | 6 | 2.70 |
| $m$ | 3 | 2.04 |
| $\overline{\xi} * m$ | 6 | 1.44 |
| Residuals | 487 | 0.93 |
| $n * m$ | 3 | 0.93 |
| $C_\xi * m$ | 9 | 0.19 |

Table 7.3: Deviation of rejection rates from nominal level $\alpha = 5\%$ from simulation: ANOVA for main effects and two-way interactions for five factors for the moment-based method with $F_{k,w}$ given in (3.7) as reference distribution

For the moment-based method we see that the four most important factors are $(i)$ the sample size, $n$; $(ii)$ the dimension, $k$; $(iii)$ the average ratio of complete

to observed information, $\bar{\xi}$; and $(iv)$ the interaction of the sample size and the dimension, $n * k$. These four factors account for $93\%$ of the p-value variance at the $\alpha = 5\%$-level. On the contrary, the number of imputations $m$ is one of the four factors that have the least effect on the rejection rates (measured by the MSE). These results are consistent with the results from the simultation study done in Chapter 5. There we see that the p-values using the moment-based method are not uniformly distributed with increasing $k$, when the sample size is $n = 1000$. Increasing the number of imputations $m$ does not improve the results.

Table 7.4 presents the two-way interaction ANOVA of the deviation of the rejection rates from nominal level $\alpha = 5\%$ for the two-stage-likelihood-ratio-test-based procedure.

| Factors | Df | Mean Squared Error |
|---|---|---|
| $\bar{\xi}$ | 2 | 131.891 |
| $k$ | 5 | 75.628 |
| $n * C_\xi$ | 3 | 71.000 |
| $n * k$ | 5 | 57.601 |
| $n * \bar{\xi}$ | 2 | 54.148 |
| $k * \bar{\xi}$ | 10 | 32.050 |
| $k * C_\xi$ | 15 | 22.404 |
| $n$ | 1 | 19.269 |
| $k * m$ | 15 | 11.508 |
| $\bar{\xi} * C_\xi$ | 6 | 5.275 |
| $m$ | 3 | 1.867 |
| $C_\xi$ | 3 | 1.783 |
| Residuals | 487 | 1.324 |
| $n * m$ | 3 | 0.501 |
| $C_\xi * m$ | 9 | 0.187 |
| $\bar{\xi} * m$ | 6 | 0.183 |

Table 7.4: Deviation of rejection rates from nominal level $\alpha = 5\%$ from simulation: ANOVA for main effects and two-way interactions for five factors for the two-stage-likelihood-ratio-test-based procedure with $F_{k,w}$ given in (3.7) as reference distribution

For the two-stage-likelihood-ratio-test-based procedure, the four most influential factors are $(i)$ the average ratio of complete to observed information, $\bar{\xi}$; $(ii)$ the dimension, $k$; $(iii)$ the interaction between the sample size and the coefficient

of variation, $n * C_\xi$; and $(iv)$ the interaction between the sample size and the dimension, $n * k$. These four factors account for $69\%$ of the p-value variance at the $\alpha = 5\%$-level.

Table 7.5 shows the results of the two-way interaction ANOVA of the deviation of the rejection rates from nominal level $\alpha = 5\%$ for the componentwise-moment-based method.

| Factors | Df | Mean Squared Error |
|---|---|---|
| $n * \overline{\xi}$ | 2 | 520.77 |
| $m$ | 3 | 247.12 |
| $n$ | 1 | 216.72 |
| $k * \overline{\xi}$ | 10 | 151.50 |
| $n * C_\xi$ | 3 | 57.87 |
| $\overline{\xi} * m$ | 6 | 25.36 |
| $k * C_\xi$ | 15 | 22.66 |
| $k$ | 5 | 20.63 |
| $\overline{\xi}$ | 2 | 18.85 |
| $n * k$ | 5 | 17.71 |
| $k * m$ | 15 | 13.35 |
| Residuals | 487 | 4.76 |
| $C_\xi * m$ | 9 | 1.30 |
| $\overline{\xi} * C_\xi$ | 6 | 1.17 |
| $C_\xi$ | 3 | 0.64 |
| $n * m$ | 3 | 0.47 |

Table 7.5: Deviation of rejection rates from nominal level $\alpha = 5\%$ from simulation: ANOVA for main effects and two-way interactions for five factors for the componentwise-moment-based method with $F_{k,v}$ given in (6.7) as reference distribution

For the componentwise-moment-based method, the four factors that most strongly affect the simulated rejection rates are $(i)$ the interaction between the sample size and the average ratio of complete to observed information, $n * \overline{\xi}$; $(ii)$ the number of imputations, $m$; $(iii)$ the sample size, $n$; and $(iv)$ the interaction between the dimension and the average ratio of complete to observed information, $k * \overline{\xi}$. These four factors account for $86\%$ of the p-value variance at the $\alpha = 5\%$-level.

The number of imputations is important with the componentwise-moment-based method, because with today's computer power, it is easy to increase $m$ to improve results. The interaction between the sample size and the dimension, $n * k$, is one of the most important factors for the moment-based procedure, yet is one of the six factors that have the least effect on the p-value variance for the componentwise-moment-based method. For the moment-based method the main effect of sample size has the highest mean squared error, whereas for the componentwise-moment-based method, the interaction of the sample size with the average ratio of complete to observed information, $n * \overline{\xi}$, has the strongest effect. We will illustrate the importance of this difference in the next section, where we characterize the effects of the main factors and interactions on rejection rates in more detail.

Table 7.6 presents the results of the two-way interaction ANOVA of the deviation of the rejection rates from nominal level $\alpha = 5\%$ for the $\chi^2_k$-statistic-based method.

| Factors | Df | Mean Squared Error |
|---|---|---|
| $\overline{\xi}$ | 2 | 5263.7 |
| $k$ | 5 | 1534.2 |
| $m$ | 3 | 709.1 |
| $C_\xi$ | 3 | 528.3 |
| $k * \overline{\xi}$ | 10 | 305.6 |
| $n$ | 1 | 294.7 |
| $k * m$ | 15 | 157.8 |
| $\overline{\xi} * m$ | 6 | 100.7 |
| $n * \overline{\xi}$ | 2 | 96.9 |
| $k * C_\xi$ | 15 | 87.0 |
| $n * k$ | 5 | 32.7 |
| $\overline{\xi} * C_\xi$ | 6 | 30.3 |
| $n * C_\xi$ | 3 | 13.9 |
| $C_\xi * m$ | 9 | 9.1 |
| Residuals | 487 | 4.9 |
| $n * m$ | 3 | 0.3 |

Table 7.6: Deviation of rejection rates from nominal level $\alpha = 5\%$ from simulation: ANOVA for main effects and two-way interactions for five factors for the $\chi^2_k$-statistic-based method with $F_{k,a_{k,m}w_s}$ given in (3.12) and (3.13) as reference distribution

For the $\chi^2_k$-statistic-based method, the four most important factors are $(i)$ the average ratio of complete to observed information, $\bar{\xi}$; $(ii)$ the dimension, $k$; $(iii)$ the number of imputations, $m$; and $(iv)$ the coefficient of variation, $C_\xi$. These four factors account for $88\%$ of the p-value variance at the $\alpha = 5\%$-level. It is noticeable that only the main effects are the factors that most affect the p-value variance. The two-way interactions are the factors with the least effect on the p-value variance.

The ANOVAs analyzed above were done for the four methods using their originally proposed degrees of freedom as denominator degrees of freedom for their particular reference distributions. In the next subsections we examine the behavior of the four different methods using other "method and degrees of freedom"-combinations and analysis tools other than ANOVA to characterize the effects of the factors of the factorial simulation study on the rejection rates in more detail.

### 7.2.2 Combination of method and appropriate degrees of freedom

If we keep the method and the degrees of freedom fixed, $576$ combinations of the other $5$ factors ($n$, $k$, $\bar{\xi}$, $C_\xi$ and $m$) remain for every nominal level $\alpha$. Now, around $\alpha$ we build an $1\%$-interval $[\alpha - 0.005; \alpha + 0.005]$. Then, for each of the four combinations of one of the four methods and one of the eight degrees of freedom we count the number of situations out of $576$, where the rejection rate is not included in this interval. If one of these $4 \times 8$ combinations has many situations with rejection rates that are less than the lower bound of the interval, that combination tends to be more conservative, that is, the rejection rate is smaller than the nominal level, but still valid. A combination with many situations with rejection rates greater than the upper bound, tends to be too liberal, which is not good, because the null hypothesis will be rejected too often and thus the inference will be invalid. The following Table 7.7 ranks the different "method and degrees of freedom"-combinations based on the "rate of conservative and invalid situations",

that is, the percentage of situations with rejection rates not included in the interval $[\alpha - 0.005; \alpha + 0.005]$. Table 7.7 only shows the ranking of the plausible "method and degrees of freedom"-combinations. We say that a combination is plausible when the percentage of situations with rejection rates less than $\alpha - 0.005$ (conservative situations), is at most approximately $50\%$ of all possible situations, and the percentage of situations with rejection rates greater than $\alpha + 0.005$ (invalid situations), is at most approximately $20\%$ of all possible situations. The moment-based procedure with $w$ given in (3.7) as degrees of freedom and the $\chi^2_k$-statistic-based procedure with $a_{k,m}w_s$ given in (3.12) and (3.13) as degrees of freedom are not plausible combinations concerning our definition, but they are also included in the ranking and for simplicity we also name them "plausible". Table 7.7 presents these plausible "method and degrees of freedom "-combinations ordered by the "rate of conservative and invalid situations" only for $\alpha = 5\%$, because the ranking of the combinations is similar for each nominal level $\alpha$, only the rates of situations with rejection rates not included in the interval $[\alpha - 0.005; \alpha + 0.005]$ differ slightly, but we are primary interested in the ranking. The rejection rate tables in Subsection 7.2.3 will give us more information about how closely the rejection rates reach the nominal levels. Note that we use the following abbreviations in Table 7.7: mom = moment-based procedure, lik = two-stage-likelihood-ratio-test-based procedure, como = componentwise-moment-based procedure and $\chi^2_k$-stat = $\chi^2_k$-statistic-based procedure.

| Rank | Degrees of freedom | Method | $< 0.045$ | $> 0.055$ | Rate of conservative and invalid situations |
|---|---|---|---|---|---|
| 1 | $w$ | lik | 41 | 76 | 0.20 |
| 2 | $(k+1)v_{\text{mean}}/2$ | lik | 28 | 103 | 0.23 |
| 3 | $v$ | como | 93 | 65 | 0.27 |
| 4 | $v$ | lik | 170 | 28 | 0.34 |
| 5 | $(k+1)v_{\text{min}}/2$ | como | 151 | 65 | 0.38 |
| 6 | $(k+1)v_{\text{min}}/2$ | lik | 222 | 10 | 0.40 |
| 7 | $v_{\text{mean}}$ | mom | 295 | 4 | 0.52 |
| 8 | $w$ | mom | 10 | 305 | 0.55 |
| 9 | $a_{k,m}w_s$ | $\chi^2_k$-stat | 22 | 409 | 0.75 |

Table 7.7: Ranking of the plausible "method and degrees of freedom"-combinations based on the rate of situations with rejection rates not included in the interval $[0.05-0.005; 0.05+0.005]$

Figure 7.2 shows the plot of the number of conservative situations, that is, the number of situations with rejection rate $< 0.045$, against the number of invalid situations, that is, the number of situations with rejection rate $> 0.055$, for each plausible "method and degrees of freedom"-combination. The $x$-axis presents the conservativeness, the $y$-axis presents the invalidity. The numbers from $1$ to $9$ correspond to the rank of each plausible combination given in Table 7.7.

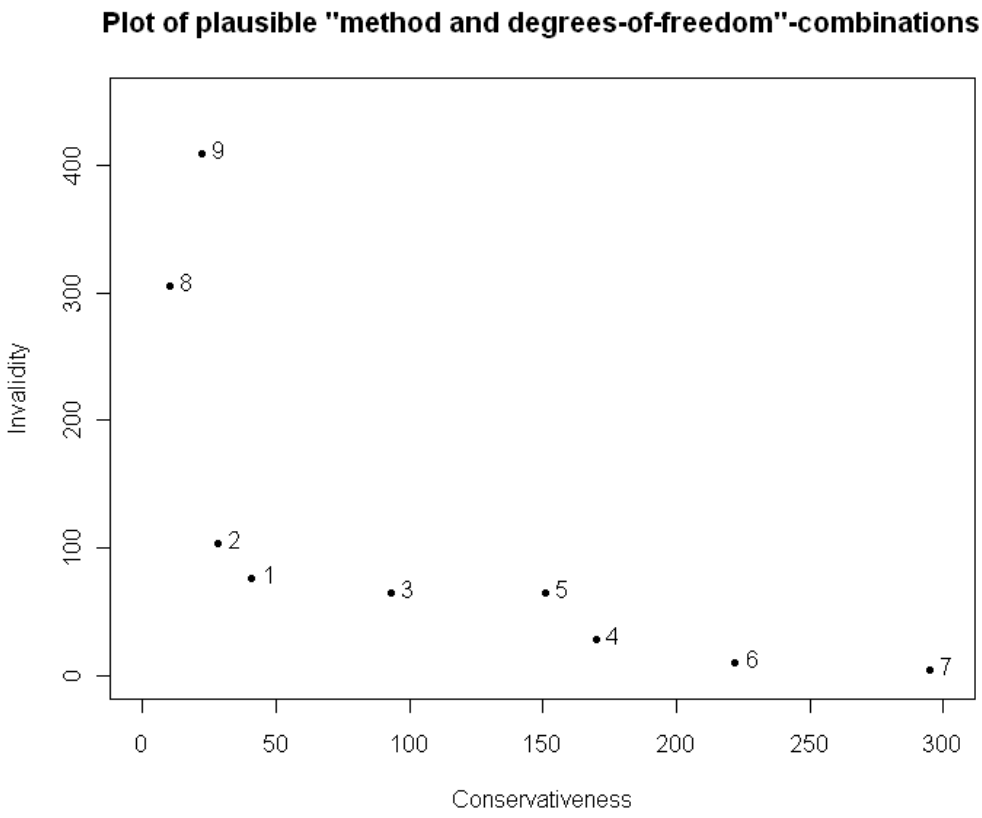


Figure 7.2: Plot of the number of conservative situations against the number of invalid situations for each plausible "method and degrees of freedom"-combination: $x$-axis $=$ conservativeness, $y$-axis $=$ invalidity; numbers $1-9$ correspond to the ranks given in Table 7.7

Table 7.7 and the corresponding Figure 7.2 help to identify which degrees of freedom fit best to the different methods. The procedure that has the most situations with rejection rates included in the interval, is the two-stage-likelihood-ratio-test-based method with the originally proposed degrees of freedom $w$ given in (3.7). This "method and degrees of freedom"-combination has a "rate of conservative and invalid situations" of $20\%$. Thus, $80\%$ of all $576$ situations for the two-stage-likelihood-ratio-test-based procedure are valid and not too conservative. This is not surprising, because the two-stage-likelihood-ratio-test-based method is the best procedure we can use, due to the asymptotically close relationship of the log-likelihood function and the Wald-test. On the second rank follows the

two-stage-likelihood-ratio-test-based method with $(k+1)v_{\text{mean}}/2$ as denominator degrees of freedom and a "rate of conservative and invalid situations" of $23\%$. It is interesting that the first two two-stage-likelihood-ratio-test-based method-combinations have more situations with rejection rates that are greater than the upper bound, which means that these procedures tend to be too liberal. Two further two-stage-likelihood-ratio-test-based method-combinations with $v$ and $(k+1)v/2$ as degrees of freedom have rank $4$ and $6$ with a "rate of conservative and invalid situations" of $34\%$ and $40\%$, respectively. Both procedures have more situations with rejection rates smaller than $0.045$, thus, they tend to be too conservative. Nevertheless, the originally proposed degrees of freedom $w$ seems to fit best for the two-stage-likelihood-ratio-test-based procedure, because of the smallest "rate of conservative and invalid situations".

However, the new componentwise-moment-based method using $v$ given in (6.7) as degrees of freedom is on the third rank. The "rate of conservative and invalid situations" is $27\%$, thus $73\%$ of the $576$ situations using the componentwise-moment-based method are valid and not too conservative. Contrary to the two-stage-likelihood-ratio-test-based method with $w$ and $(k+1)v_{\text{mean}}/2$ as degrees of freedom on rank $1$ and $2$, the componentwise-moment-based method using $v$ as degrees of freedom has more situations with rejection rates that are less than the lower bound, thus, it tends to be too conservative, but not invalid. The second plausible componentwise-moment-based method has $(k+1)v_{\text{min}}/2$ as degrees of freedom and has rank $5$. The "rate of conservative and invalid situations" is $38\%$, thus, this combination has more invalid and conservative situations than the componentwise-moment-based method with $v$ as degrees of freedom . In addition the combination with $(k+1)v_{\text{min}}/2$ as degrees of freedom has more situations with rejection rates that are less than $0.045$ than the componentwise-moment-based method using $v$ as degrees of freedom, thus, it is more conservative. Hence, the originally proposed degrees of freedom $v$ seems to fit best for the componentwise-moment-based procedure.

The moment-based procedure combined with $v_{\text{mean}}$ as degrees of freedom has rank $7$ with a "rate of conservative and invalid situations" of $52\%$, thus, less than $50\%$ of the situations are valid and not too conservative. From $299 = 295 + 4$ situations that are either less than $0.045$ or greater than $0.055$, $295$ are less than the lower bound. Thus this combination tends to be very conservative. The moment-based method with their originally proposed degrees of freedom, $w$, has rank $8$ and a "rate of conservative and invalid situations" of $55\%$, thus only $45\%$ of the situations are included in the interval. About $50\%$ of all situations with this combination have rejection rates that are greater than the upper bound of the interval. Hence, in contrary to the combination with $v_{\text{mean}}$ as degrees of freedom, the combination with the originally proposed degrees of freedom , $w$, tends to be too liberal, and thus invalid in many situations. It seems that neither $v_{\text{mean}}$ nor $w$ as degrees of freedom fit very well with the moment-based procedure, because of their high "rate of conservative and invalid situations" of more than $50\%$. Combinations with other degrees of freedom are not plausible concerning our definition of a "plausible combination" given above.

The $\chi^2_k$-statistic-based approach has the highest "rate of conservative and invalid situations" with $75\%$, thus only $25\%$ of the $576$ situations are included in the interval $[0.045; 0.055]$. Almost all of the invalid and conservative situations have rejection rates that are greater than the upper bound, thus, in almost every situation the $\chi^2_k$-statistic-based method is too liberal and the null hypothesis is rejected too often. Combinations with other degrees of freedom are not plausible concerning our definition of a "plausible combination" given above.

Table 7.7 and Figure 7.2 show which are the best combinations of method and degrees of freedom based on the rate of conservative and invalid situations. In the next section, we examine the rejection rates of six special combinations: $(i)$ the two-stage-likelihood-ratio-test-based method with the originally proposed degrees of freedom $w$. It is the best procedure we can use if the likelihood-function is correctly specified. The combination of this method with $w$ has the most situations with rejection rates that are included in the interval; $(ii)$ the componentwise-

moment-based procedure with $v$ as degrees of freedom, because $v$ seems to fit best with this method; $(iii)$ the moment-based procedure with their originally proposed degrees of freedom $w$; $(iv)$ the moment-based method with $v_{\text{mean}}$ as a choice of degrees of freedom that might improve results; $(v)$ the $\chi^2_k$-statistic-based approach with its originally proposed degrees of freedom $a_{k,m}w_s$. Additionally we examine $(vi)$ the moment-based method with the new degrees of freedom $v$ given in (6.7), because of the close relationship of the moment-based method to our new componentwise-moment-based procedure mentioned in Section 6.2. However, this "method and degrees of freedom"-combination is not plausible concerning our definition of a "plausible combination". More than $20\%$ of all situations of this combination have rejections rates that are greater than $0.055$, and thus this combination is not included in Table 7.7 and Figure 7.2.

### 7.2.3 Rejection rates

Having identified the main determinants of the rejection rates using the ANOVA described in Section 7.2.1 and having identified the best choices of degrees of freedom for each of the four methods, we now examine how the rejection rates depend on the levels of the other factors $n$, $k$, $\overline{\xi}$, $C_\xi$ and $m$.

#### 7.2.3.1 Moment-based method

First we consider the moment-based method with sample size $n = 1000$, number of imputations $m = 5$, and the original degrees of freedom $w$ given in (3.7). Table 7.8 presents the empirical rejection rates when the nominal levels $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$ are used, for each combination of $k$, $\overline{\xi}$ and $C_\xi$, obtained by the simulation described in Section 7.1. Note that all following tables are subdivided into $(i)$ the number of components being tested, $k$, $(ii)$ the average ratio of complete information to observed information, $\overline{\xi}$, $(iii)$ the coefficient of variation of the ratios of complete information to observed information, $C_\xi$, and $(iv)$ the nominal level $\alpha$.

| | | $\alpha = 1\%$ | | | $\alpha = 5\%$ | | | $\alpha = 10\%$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\overline{\xi}$ | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 |
| | | | | | $C_\xi = 0\%$ | | | | | |
| k | 2 | 1.0 | 0.9 | 1.0 | 4.8 | 4.6 | 4.4 | 9.5 | 9.1 | 9.1 |
| | 5 | 1.0 | 1.2 | 1.2 | 5.4 | 5.4 | 5.4 | 10.1 | 10.4 | 10.4 |
| | 10 | 1.1 | 1.2 | 1.4 | 5.1 | 5.7 | 6.4 | 9.9 | 11.1 | 11.3 |
| | 20 | 1.2 | 1.5 | 1.9 | 5.8 | 6.4 | 7.1 | 11.3 | 11.5 | 12.5 |
| | 35 | 1.4 | 1.9 | 2.3 | 6.3 | 7.2 | 8.4 | 11.7 | 12.9 | 15.2 |
| | 50 | 1.6 | 2.4 | 2.7 | 6.9 | 8.0 | 9.9 | 12.8 | 14.1 | 17.3 |
| | | | | | $C_\xi = 10\%$ | | | | | |
| k | 2 | 1.0 | 1.1 | 0.9 | 4.8 | 4.6 | 4.7 | 9.4 | 8.9 | 8.9 |
| | 5 | 1.2 | 1.3 | 1.1 | 5.4 | 5.5 | 5.4 | 10.3 | 10.5 | 10.5 |
| | 10 | 1.0 | 1.3 | 1.5 | 5.1 | 5.5 | 5.9 | 10.3 | 10.8 | 11.2 |
| | 20 | 1.2 | 1.6 | 1.7 | 5.9 | 6.1 | 6.6 | 11.3 | 11.3 | 12.1 |
| | 35 | 1.4 | 1.8 | 2.3 | 6.1 | 7.3 | 8.4 | 11.5 | 13.0 | 14.6 |
| | 50 | 1.7 | 2.1 | 3.0 | 6.8 | 7.8 | 9.6 | 12.3 | 13.8 | 17.2 |
| | | | | | $C_\xi = 20\%$ | | | | | |
| k | 2 | 1.1 | 0.8 | 0.9 | 4.8 | 4.8 | 4.4 | 9.6 | 9.2 | 9.2 |
| | 5 | 1.1 | 1.3 | 1.2 | 5.4 | 5.4 | 5.4 | 10.5 | 10.4 | 10.0 |
| | 10 | 1.1 | 1.5 | 1.4 | 5.5 | 5.9 | 5.9 | 10.7 | 11.2 | 11.3 |
| | 20 | 1.3 | 1.8 | 1.7 | 5.8 | 6.3 | 6.6 | 11.0 | 11.6 | 12.7 |
| | 35 | 1.6 | 2.0 | 2.3 | 6.1 | 7.1 | 8.5 | 11.7 | 12.7 | 14.5 |
| | 50 | 1.8 | 2.1 | 2.8 | 7.0 | 7.5 | 10.2 | 12.6 | 13.8 | 17.4 |
| | | | | | $C_\xi = 40\%$ | | | | | |
| k | 2 | 1.1 | 1.0 | 1.0 | 4.9 | 4.7 | 4.7 | 9.8 | 9.2 | 9.6 |
| | 5 | 1.6 | 1.4 | 1.5 | 5.5 | 5.6 | 5.6 | 10.2 | 10.6 | 10.4 |
| | 10 | 1.5 | 1.7 | 1.7 | 6.4 | 6.4 | 6.1 | 11.6 | 11.5 | 11.1 |
| | 20 | 1.8 | 2.0 | 1.8 | 6.1 | 6.7 | 6.9 | 11.6 | 11.8 | 12.5 |
| | 35 | 1.8 | 2.1 | 2.2 | 6.6 | 7.3 | 8.2 | 12.3 | 13.0 | 14.2 |
| | 50 | 1.5 | 1.8 | 2.6 | 6.5 | 7.5 | 9.2 | 11.7 | 13.5 | 16.2 |

Table 7.8: Rejection rates (in $\%$) of multiple imputation methodology from simulation using the moment-based procedure with $F_{k,w}$ given in (3.7) as reference distribution; $n = 1000$ and $m = 5$

From Table 7.8 we see that the rejection rate exceeded the particular nominal level when the dimension $k$ is greater than $2$ in almost every simulation setting with $n = 1000$ and $m = 5$. Thus, the corresponding p-values are too liberal and the null hypothesis will be rejected too often. The same is already be indicated by the ranking of the "method and degrees of freedom"-combinations based on the rate of situations with rejection rates not included in the interval $[0; \alpha + 0.005]$ as given in Table 7.7. The deviation of the rejection rate from the corresponding nominal

level $\alpha$ increases with increasing average ratio of complete to observed information $\overline{\xi}$ for all nominal levels $\alpha$, when $k$ is presumably greater than $5$. These results are consistent with the results of the ANOVA for the moment-based procedure given in Table 7.3 and described in detail in Section 7.2.1. Besides the sample size, $n$, which in this setting is fixed at $n = 1000$, the two factors that most affect the rejection rates are the dimension, $k$, and the average ratio of complete to observed information, $\overline{\xi}$. The worst situations are: $(i)$ $k = 50$, $\overline{\xi} = 2$ and $C_\xi = 10\%$, where the rejection rate reaches as high as $3.0\%$ for $\alpha = 1\%$; $(ii)$ $k = 50$, $\overline{\xi} = 2$ and $C_\xi = 20\%$, where the rejection rate reaches as high as $10.2\%$ for $\alpha = 5\%$, and $(iii)$ $k = 50$, $\overline{\xi} = 2$ and $C_\xi = 20\%$, where the rejection rate reaches as high as $17.4\%$ for $\alpha = 10\%$. The rejection rates are similar for $m = 30$. We have to raise the sample size to $n = 5000$ before we get rejection rates close to the nominal levels. With $n = 5000$, $m = 5$ imputations are sufficient to get well calibrated rejection rates. Table 7.9 shows the results for $n = 5000$ and $m = 5$.

| | | $\alpha = 1\%$ | | | $\alpha = 5\%$ | | | $\alpha = 10\%$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\overline{\xi}$ | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 |
| | | | | | $C_\xi = 0\%$ | | | | | |
| k | 2 | 1.0 | 0.8 | 0.7 | 4.7 | 4.2 | 4.1 | 9.6 | 9.0 | 8.4 |
| | 5 | 1.0 | 1.1 | 1.1 | 4.9 | 5.2 | 5.0 | 9.7 | 10.0 | 9.8 |
| | 10 | 0.9 | 1.1 | 1.1 | 4.8 | 5.3 | 5.1 | 9.7 | 10.3 | 10.1 |
| | 20 | 1.0 | 1.6 | 1.2 | 5.0 | 5.5 | 5.4 | 10.0 | 10.6 | 10.7 |
| | 35 | 1.1 | 1.6 | 1.4 | 5.1 | 6.3 | 5.9 | 10.1 | 11.4 | 11.3 |
| | 50 | 1.3 | 1.5 | 1.5 | 5.3 | 6.0 | 6.1 | 10.6 | 11.4 | 11.6 |
| | | | | | $C_\xi = 10\%$ | | | | | |
| k | 2 | 0.9 | 0.8 | 0.8 | 4.5 | 4.5 | 4.2 | 9.3 | 9.2 | 9.0 |
| | 5 | 1.1 | 1.3 | 1.1 | 5.1 | 5.4 | 5.2 | 10.1 | 10.4 | 10.2 |
| | 10 | 0.9 | 1.1 | 1.2 | 4.9 | 5.2 | 5.3 | 9.9 | 10.4 | 10.3 |
| | 20 | 1.0 | 1.2 | 1.2 | 5.0 | 5.7 | 5.3 | 9.8 | 10.6 | 10.4 |
| | 35 | 1.0 | 1.5 | 1.3 | 5.2 | 5.7 | 5.8 | 9.9 | 11.0 | 11.2 |
| | 50 | 1.1 | 1.5 | 1.4 | 5.3 | 6.0 | 6.1 | 10.7 | 11.3 | 11.1 |
| | | | | | $C_\xi = 20\%$ | | | | | |
| k | 2 | 0.9 | 0.9 | 0.8 | 4.5 | 4.6 | 4.1 | 8.9 | 9.3 | 8.5 |
| | 5 | 1.2 | 1.2 | 1.2 | 5.2 | 5.1 | 5.3 | 10.3 | 10.3 | 10.3 |
| | 10 | 1.1 | 1.2 | 1.2 | 4.9 | 5.0 | 5.2 | 9.9 | 10.3 | 10.3 |
| | 20 | 1.0 | 1.3 | 1.4 | 5.4 | 5.6 | 5.6 | 10.1 | 10.8 | 10.5 |
| | 35 | 1.3 | 1.5 | 1.4 | 5.6 | 6.4 | 5.6 | 10.7 | 11.3 | 11.0 |
| | 50 | 1.3 | 1.5 | 1.4 | 5.8 | 6.2 | 5.9 | 10.5 | 11.7 | 11.2 |
| | | | | | $C_\xi = 40\%$ | | | | | |
| k | 2 | 1.0 | 0.9 | 0.9 | 4.6 | 4.0 | 4.3 | 9.5 | 8.3 | 8.8 |
| | 5 | 1.4 | 1.5 | 1.1 | 5.6 | 5.2 | 5.3 | 10.6 | 10.4 | 10.4 |
| | 10 | 1.2 | 1.4 | 1.4 | 5.3 | 5.8 | 5.7 | 10.5 | 10.9 | 10.7 |
| | 20 | 1.5 | 1.6 | 1.5 | 5.9 | 6.0 | 6.0 | 10.9 | 11.0 | 11.3 |
| | 35 | 1.8 | 1.7 | 1.7 | 6.4 | 6.5 | 6.2 | 11.6 | 11.5 | 11.0 |
| | 50 | 1.6 | 1.5 | 1.7 | 6.0 | 6.7 | 6.0 | 11.4 | 12.0 | 11.7 |

Table 7.9: Rejection rates (in $\%$) of multiple imputation methodology from simulation using the moment-based procedure with $F_{k,w}$ given in (3.7) as reference distribution; $n = 5000$ and $m = 5$

The settings with $n = 5000$ and $m = 5$ for the moment-based procedure is approximately conform to the settings in the work from Li, Raghunathan, and Rubin (1991). They used draws from the theoretical large-sample distribution $F_{k,w}$ of $\tilde{D}_m$ given in (3.7) and (3.5) and choose $m = 3$. Table 7.10 shows the corresponding results of Li, Raghunathan, and Rubin (1991).

| | | $\alpha = 1\%$ | | | $\alpha = 5\%$ | | | $\alpha = 10\%$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\overline{\xi}$ | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 |
| | | | | | $C_\xi = 0\%$ | | | | | |
| k | 2 | 1.1 | 1.2 | 1.5 | 4.7 | 5.1 | 5.6 | 9.7 | 9.9 | 10.4 |
| | 5 | 1.0 | 0.9 | 0.8 | 4.4 | 4.5 | 4.1 | 9.4 | 9.4 | 9.0 |
| | 10 | 1.0 | 0.8 | 1.0 | 4.8 | 4.7 | 4.5 | 9.8 | 9.5 | 9.7 |
| | 20 | 1.1 | 1.1 | 1.0 | 5.2 | 5.2 | 4.9 | 9.7 | 9.7 | 10.1 |
| | 35 | 1.0 | 1.1 | 1.1 | 5.1 | 4.8 | 5.3 | 9.9 | 9.8 | 9.9 |
| | | | | | $C_\xi = 10\%$ | | | | | |
| k | 2 | 0.8 | 1.0 | 1.2 | 4.6 | 4.4 | 5.6 | 9.3 | 9.0 | 10.6 |
| | 5 | 1.0 | 0.8 | 1.2 | 4.7 | 4.4 | 5.2 | 9.3 | 9.2 | 10.1 |
| | 10 | 1.1 | 1.1 | 1.1 | 5.2 | 4.6 | 5.4 | 10.2 | 9.3 | 10.1 |
| | 20 | 1.2 | 1.1 | 1.1 | 5.0 | 5.2 | 5.1 | 10.2 | 10.3 | 10.0 |
| | 35 | 0.9 | 1.1 | 1.1 | 5.2 | 4.9 | 5.1 | 9.8 | 9.9 | 10.2 |
| | | | | | $C_\xi = 20\%$ | | | | | |
| k | 2 | 0.9 | 1.2 | 1.3 | 4.8 | 4.9 | 5.4 | 9.5 | 9.6 | 10.2 |
| | 5 | 1.2 | 1.0 | 0.8 | 5.3 | 5.1 | 4.5 | 10.3 | 9.6 | 9.5 |
| | 10 | 1.5 | 1.1 | 1.2 | 5.5 | 5.5 | 5.3 | 10.7 | 10.4 | 10.2 |
| | 20 | 1.3 | 1.3 | 1.2 | 5.9 | 5.4 | 5.4 | 11.2 | 10.6 | 10.7 |
| | 35 | 1.4 | 1.2 | 1.1 | 5.9 | 5.8 | 5.3 | 11.1 | 10.7 | 10.5 |
| | | | | | $C_\xi = 40\%$ | | | | | |
| k | 2 | 1.3 | 1.6 | 1.6 | 5.1 | 5.0 | 6.7 | 10.3 | 10.0 | 12.0 |
| | 5 | 1.7 | 1.7 | 1.3 | 6.1 | 5.5 | 5.5 | 10.9 | 10.0 | 11.1 |
| | 10 | 1.8 | 2.0 | 1.8 | 6.6 | 6.3 | 6.8 | 11.2 | 11.1 | 12.0 |
| | 20 | 2.1 | 2.1 | 1.9 | 7.4 | 7.7 | 6.9 | 12.9 | 13.0 | 12.1 |
| | 35 | 2.5 | 2.1 | 2.1 | 7.3 | 7.3 | 7.4 | 12.6 | 12.9 | 12.9 |

Table 7.10: Large-sample levels (in $\%$) of multiple imputation methodology using the moment-based method and draws from the theroretical $F_{k,w}$-distribution given in (3.7); $m = 3$

The rejection rates of the moment-based procedure from our simulation given in Table 7.9 with generated and imputed data are as good as the results of Li et al. (1991), which are shown in Table 7.10. The worst situation is $k = 50$, $\overline{\xi} = 2$ and $C_\xi = 40\%$, where the rejection rate from our simulation given in Table 7.9 reaches as high as $12.0\%$ for $\alpha = 10\%$. Thus, the largest deviation of the rejection rate from the nominal level is $2.0$-percent points. For a large sample size, here $n = 5000$, the moment-based procedure is well calibrated. Only $m = 5$ imputations are sufficient. The quantities of $\overline{\xi}$ and $C_\xi$ do not affcet the rejection rates strongly.

Based on the results of the ranking of the ″method and degrees of freedom″-combinations as given in Table 7.7, we also consider the moment-based procedure using the degrees of freedom $v_{\text{mean}}$ given in (6.7) where we take the mean instead of the maximum of the componentwise calculated degrees of freedom according to Barnard and Rubin (1999). The results for $n = 1000$ and $m = 30$ are presented in Table 7.11.

| | | $\alpha = 1\%$ | | | $\alpha = 5\%$ | | | $\alpha = 10\%$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\bar{\xi}$ | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 |
| | | | | | $C_\xi = 0\%$ | | | | | |
| k | 2 | 1.1 | 1.1 | 1.1 | 5.1 | 4.9 | 4.9 | 9.8 | 9.8 | 9.5 |
| | 5 | 1.0 | 1.1 | 1.0 | 5.0 | 5.0 | 4.8 | 10.2 | 9.9 | 9.7 |
| | 10 | 0.9 | 1.0 | 1.0 | 4.8 | 5.0 | 4.7 | 9.8 | 10.1 | 9.8 |
| | 20 | 1.0 | 1.0 | 0.6 | 5.3 | 4.9 | 4.4 | 10.8 | 9.9 | 9.4 |
| | 35 | 1.0 | 0.8 | 0.6 | 5.5 | 4.9 | 4.5 | 10.8 | 9.9 | 9.6 |
| | 50 | 1.1 | 0.8 | 0.5 | 5.5 | 4.8 | 4.3 | 11.0 | 10.0 | 9.8 |
| | | | | | $C_\xi = 10\%$ | | | | | |
| k | 2 | 1.0 | 1.1 | 1.2 | 5.2 | 5.0 | 4.7 | 9.8 | 9.8 | 9.4 |
| | 5 | 1.1 | 1.1 | 1.0 | 5.2 | 5.2 | 4.7 | 10.1 | 9.8 | 9.5 |
| | 10 | 0.8 | 1.1 | 0.8 | 4.9 | 5.1 | 4.6 | 9.9 | 10.0 | 9.6 |
| | 20 | 1.0 | 0.9 | 0.6 | 5.2 | 4.9 | 4.4 | 10.4 | 9.9 | 9.3 |
| | 35 | 1.1 | 0.7 | 0.6 | 5.0 | 4.7 | 4.2 | 10.4 | 9.8 | 9.6 |
| | 50 | 1.2 | 0.8 | 0.5 | 5.4 | 4.4 | 4.2 | 11.0 | 9.9 | 9.7 |
| | | | | | $C_\xi = 20\%$ | | | | | |
| k | 2 | 1.1 | 1.1 | 1.1 | 4.9 | 5.0 | 5.0 | 10.1 | 9.7 | 9.6 |
| | 5 | 1.0 | 0.9 | 0.8 | 5.2 | 5.2 | 4.8 | 10.1 | 10.1 | 9.5 |
| | 10 | 0.9 | 1.0 | 0.9 | 5.0 | 5.2 | 4.5 | 9.9 | 10.1 | 9.6 |
| | 20 | 1.0 | 0.9 | 0.5 | 5.1 | 4.8 | 4.4 | 10.4 | 9.8 | 9.2 |
| | 35 | 1.1 | 0.8 | 0.6 | 5.1 | 4.6 | 4.2 | 10.2 | 9.7 | 9.3 |
| | 50 | 1.0 | 0.6 | 0.4 | 5.2 | 4.4 | 4.1 | 10.7 | 9.6 | 10.0 |
| | | | | | $C_\xi = 40\%$ | | | | | |
| k | 2 | 1.2 | 1.1 | 1.1 | 5.0 | 4.9 | 5.1 | 10.1 | 9.9 | 10.3 |
| | 5 | 1.4 | 1.0 | 1.0 | 5.2 | 5.2 | 5.1 | 9.9 | 10.0 | 9.8 |
| | 10 | 1.2 | 1.1 | 0.9 | 5.4 | 5.5 | 4.9 | 10.4 | 10.0 | 9.7 |
| | 20 | 1.2 | 0.9 | 0.6 | 5.2 | 5.0 | 4.3 | 10.1 | 9.9 | 9.2 |
| | 35 | 0.9 | 0.8 | 0.6 | 5.0 | 4.4 | 4.1 | 9.7 | 9.3 | 8.8 |
| | 50 | 0.8 | 0.5 | 0.3 | 4.6 | 3.9 | 3.3 | 9.5 | 8.8 | 8.4 |

Table 7.11: Rejection rates (in %) of multiple imputation methodology from simulation using the moment-based procedure with $F_{k,v_{\text{mean}}}$ given in (6.7) as reference distribution; $n = 1000$ and $m = 30$

The rejection rates of the moment-based procedure given in Table 7.11 with $v_{\text{mean}}$ given in (6.7) (taking the mean instead of the maximum) as degrees of freedom, and $n = 1000$ and $m = 30$, are well calibrated and close to the particular nominal level $\alpha$ in almost every situation. The rejection rates have a slight tendency to be smaller than the particular nominal level $\alpha$, thus, they are a bit too conservative, but still valid, as the ranking of the "method and degrees of freedom"-combinations given in Table 7.7 already indicates. The rejection rates are particularly small when the coefficient of variation, $C_\xi$, is equal to $40\%$. The worst situations are: $(i)$ $k = 50$, $\overline{\xi} = 2$ and $C_\xi = 40\%$, where the rejection rate reaches as low as $3.3\%$ for $\alpha = 5\%$, and $(ii)$ $k = 50$, $\overline{\xi} = 2$ and $C_\xi = 40\%$, where the rejection rate reaches as low as $8.4\%$ for $\alpha = 10\%$. The rejection rates are similar for $n = 5000$ and $m = 30$, thus, increasing the sample size does not improve the results. It seems that the degrees of freedom, $v_{\text{mean}}$, from our suggested procedure work better with the moment-based method than with the originally porposed degrees of freedom, $w$. Using $v_{\text{mean}}$ yield slightly too conservative, but still valid, rejection rates in every situation.

Finally, we examine the moment-based method with the originally proposed degrees of freedom, $v$, from the componentwise-moment-based procedure given in (6.7). The rejection rates for $n = 1000$ and $m = 30$ are presented in Table 7.12.

| | | $\alpha = 1\%$ | | | $\alpha = 5\%$ | | | $\alpha = 10\%$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\overline{\xi}$ | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 |
| | | | | | $C_\xi = 0\%$ | | | | | |
| k | 2 | 1.1 | 1.1 | 1.1 | 5.1 | 5.0 | 5.0 | 9.8 | 9.8 | 9.5 |
| | 5 | 1.0 | 1.2 | 1.0 | 5.0 | 5.2 | 5.1 | 10.3 | 10.1 | 10.0 |
| | 10 | 1.0 | 1.2 | 1.1 | 4.9 | 5.3 | 5.2 | 9.8 | 10.5 | 10.4 |
| | 20 | 1.1 | 1.2 | 1.0 | 5.4 | 5.4 | 5.5 | 11.0 | 10.5 | 10.5 |
| | 35 | 1.1 | 1.2 | 1.4 | 5.7 | 5.7 | 6.0 | 11.2 | 11.2 | 11.6 |
| | 50 | 1.3 | 1.3 | 1.4 | 5.9 | 6.3 | 6.5 | 11.4 | 11.7 | 13.2 |
| | | | | | $C_\xi = 10\%$ | | | | | |
| k | 2 | 1.1 | 1.2 | 1.2 | 5.2 | 5.0 | 4.8 | 9.8 | 9.9 | 9.5 |
| | 5 | 1.1 | 1.2 | 1.1 | 5.3 | 5.3 | 5.0 | 10.1 | 10.0 | 9.9 |
| | 10 | 0.9 | 1.2 | 1.1 | 5.0 | 5.4 | 5.1 | 10.0 | 10.5 | 10.2 |
| | 20 | 1.1 | 1.1 | 0.9 | 5.4 | 5.5 | 5.2 | 10.6 | 10.5 | 10.4 |
| | 35 | 1.2 | 1.1 | 1.2 | 5.4 | 5.5 | 5.9 | 10.8 | 11.2 | 11.5 |
| | 50 | 1.3 | 1.2 | 1.2 | 5.8 | 5.9 | 6.6 | 11.6 | 11.7 | 13.0 |
| | | | | | $C_\xi = 20\%$ | | | | | |
| k | 2 | 1.1 | 1.2 | 1.1 | 4.9 | 5.1 | 5.1 | 10.1 | 9.7 | 9.7 |
| | 5 | 1.1 | 1.1 | 0.9 | 5.3 | 5.4 | 5.0 | 10.3 | 10.3 | 9.8 |
| | 10 | 1.0 | 1.2 | 1.1 | 5.1 | 5.5 | 4.9 | 10.1 | 10.5 | 10.5 |
| | 20 | 1.1 | 1.1 | 0.9 | 5.4 | 5.4 | 5.4 | 10.7 | 10.5 | 10.5 |
| | 35 | 1.2 | 1.2 | 1.1 | 5.6 | 5.6 | 6.0 | 10.8 | 10.9 | 11.4 |
| | 50 | 1.2 | 1.3 | 1.2 | 5.8 | 5.9 | 6.6 | 11.4 | 11.7 | 13.3 |
| | | | | | $C_\xi = 40\%$ | | | | | |
| k | 2 | 1.2 | 1.2 | 1.2 | 5.0 | 5.0 | 5.2 | 10.1 | 10.0 | 10.4 |
| | 5 | 1.4 | 1.1 | 1.2 | 5.3 | 5.4 | 5.4 | 10.0 | 10.3 | 10.1 |
| | 10 | 1.2 | 1.2 | 1.1 | 5.6 | 5.8 | 5.5 | 10.6 | 10.5 | 10.3 |
| | 20 | 1.3 | 1.2 | 1.0 | 5.4 | 5.6 | 5.2 | 10.5 | 10.9 | 10.6 |
| | 35 | 1.1 | 1.4 | 1.1 | 5.5 | 5.6 | 5.9 | 10.4 | 10.6 | 11.0 |
| | 50 | 1.2 | 1.1 | 0.9 | 5.3 | 5.5 | 5.6 | 10.5 | 10.9 | 12.0 |

Table 7.12: Rejection rates (in %) of multiple imputation methodology from simulation using the moment-based procedure with $F_{k,v}$ given in (6.7) as reference distribution; $n = 1000$ and $m = 30$

For the moment-based method with $v$ as degrees of freedom, $n = 1000$, and $m = 30$, we see in Table 7.12 that the deviations of the rejection rates from the corresponding nominal levels are similar to the deviations in the setting before given in Table 7.9, where we have a sample size of $n = 5000$ and $w$ as degrees of freedom. But in both cases (using $v$ and $w$) the rejection rates exceed the particular nominal levels. Thus they tend to be too liberal, and thus invalid. However, it seems that the moment-based procedure with $v$ as degrees of freedom from our

suggested componentwise-moment-based procedure performs at least as good as the moment-based procedure using the originally proposed degrees of freedom, and is not affected by the sample size when the number of imputations is at least $30$.

Using $w$ or $v$ as degrees of freedom yield rejection rates that tend to be too liberal. Using $v_{\text{mean}}$ as degrees of freedom yield rejection rates that tend to be too conservative. Maybe degrees of freedom between $v_{\text{mean}}$ and $v$ or $w$ might be a good choice. Of course, more detailed research is needed to verify this advice.

#### 7.2.3.2 Two-stage-likelihood-ratio-test-based method

We consider the two-stage-likelihood-ratio-test-based method described in detail in Section 3.2 with its original degrees of freedom $w$ given in (3.7). Note that the original proposed degrees of freedom are equal to the originally proposed degrees of freedom of the moment-based method. Table 7.13 shows the rejection rates using the two-stage-likelihood-ratio-test-based method with sample size $n = 1000$ and number of imputations $m = 5$.

| | | $\alpha = 1\%$ | | | $\alpha = 5\%$ | | | $\alpha = 10\%$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\overline{\xi}$ | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 |
| | | | | | $C_\xi = 0\%$ | | | | | |
| k | 2 | 1.0 | 0.8 | 0.9 | 4.7 | 4.5 | 4.2 | 9.5 | 9.0 | 8.9 |
| | 5 | 0.9 | 1.1 | 1.0 | 5.3 | 5.2 | 5.0 | 10.0 | 10.1 | 9.8 |
| | 10 | 1.1 | 1.1 | 1.0 | 4.9 | 5.4 | 5.6 | 9.7 | 10.6 | 10.4 |
| | 20 | 1.1 | 1.1 | 1.5 | 5.6 | 5.6 | 5.5 | 11.0 | 10.5 | 10.4 |
| | 35 | 1.2 | 1.2 | 1.3 | 5.9 | 5.5 | 5.3 | 11.3 | 10.8 | 10.5 |
| | 50 | 1.4 | 1.4 | 1.0 | 6.6 | 5.4 | 4.7 | 12.1 | 10.6 | 9.5 |
| | | | | | $C_\xi = 20\%$ | | | | | |
| k | 2 | 1.0 | 1.0 | 0.9 | 4.8 | 4.5 | 4.5 | 9.4 | 8.8 | 8.6 |
| | 5 | 1.2 | 1.2 | 1.1 | 5.3 | 5.3 | 5.0 | 10.2 | 10.2 | 10.2 |
| | 10 | 0.9 | 1.1 | 1.2 | 5.0 | 5.0 | 5.1 | 10.1 | 10.2 | 10.2 |
| | 20 | 1.1 | 1.3 | 1.2 | 5.6 | 5.3 | 5.2 | 10.9 | 10.3 | 10.0 |
| | 35 | 1.2 | 1.3 | 1.1 | 5.6 | 5.6 | 5.4 | 11.0 | 10.9 | 10.2 |
| | 50 | 1.4 | 1.2 | 1.1 | 6.1 | 5.2 | 4.7 | 11.4 | 10.1 | 9.2 |
| | | | | | $C_\xi = 20\%$ | | | | | |
| k | 2 | 1.1 | 0.8 | 0.8 | 4.8 | 4.7 | 4.1 | 9.5 | 9.0 | 8.8 |
| | 5 | 1.1 | 1.1 | 1.0 | 5.3 | 5.1 | 4.9 | 10.3 | 10.1 | 9.5 |
| | 10 | 1.1 | 1.3 | 1.2 | 5.4 | 5.4 | 5.1 | 10.5 | 10.7 | 10.2 |
| | 20 | 1.2 | 1.4 | 1.1 | 5.3 | 5.4 | 5.1 | 10.5 | 10.5 | 10.1 |
| | 35 | 1.3 | 1.3 | 1.0 | 5.6 | 5.5 | 5.1 | 10.9 | 10.2 | 9.8 |
| | 50 | 1.3 | 1.0 | 1.0 | 5.9 | 4.8 | 4.5 | 11.4 | 9.6 | 9.1 |
| | | | | | $C_\xi = 40\%$ | | | | | |
| k | 2 | 1.0 | 0.9 | 0.9 | 4.8 | 4.5 | 4.4 | 9.6 | 9.0 | 9.4 |
| | 5 | 1.4 | 1.3 | 1.3 | 5.3 | 5.3 | 5.0 | 9.9 | 10.2 | 9.9 |
| | 10 | 1.4 | 1.4 | 1.3 | 6.1 | 5.9 | 5.2 | 11.3 | 10.8 | 9.9 |
| | 20 | 1.4 | 1.4 | 1.1 | 5.6 | 5.6 | 5.0 | 10.7 | 10.4 | 9.7 |
| | 35 | 1.3 | 1.2 | 0.9 | 5.4 | 5.1 | 4.3 | 10.5 | 9.7 | 8.3 |
| | 50 | 0.9 | 0.8 | 0.6 | 4.5 | 3.9 | 3.1 | 9.2 | 8.2 | 6.6 |

Table 7.13: Rejection rates (in $\%$) of multiple imputation methodology from simulation using the two-stage-likelihood-ratio-test-based method with $F_{k,w}$ given in (3.7) as reference distribution; $n = 1000$ and $m = 5$

The rejection rates using the two-stage-likelihood-ratio-test-based method are essentially well calibrated. The number of imputations equal to $5$ is sufficient. Only for situations with $\overline{\xi} = 2$, $C_\xi = 40\%$ and $k = 50$, the rejection rates fall below the particular nominal level $\alpha$. For $\overline{\xi} = 1.2$, $C_\xi = 0\%$, $k = 50$ and $\alpha = 10\%$, the rejection rate reaches as high as $12.1\%$. In every other situation studied, the rejection rates are close to the corresponding nominal level $\alpha$. These results are consistent with the results from the corresponding ANOVA given in Table 7.4.

The average ratio of complete to observed information $\overline{\xi}$ and the dimension $k$ are the factors that most strongly affect the rejection rate. If we raise the sample size to $n = 5000$ and use $m = 5$ as number of imputations, we get almost perfect rejection rates in every situation, as shown in Table 7.14.

| | | $\alpha = 1\%$ | | | $\alpha = 5\%$ | | | $\alpha = 10\%$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\overline{\xi}$ | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 |
| | | | | | $C_\xi = 0\%$ | | | | | |
| k | 2 | 1.0 | 0.8 | 0.7 | 4.7 | 4.1 | 4.1 | 9.6 | 8.9 | 8.3 |
| | 5 | 1.0 | 1.1 | 1.0 | 4.8 | 5.1 | 5.0 | 9.7 | 9.9 | 9.6 |
| | 10 | 0.9 | 1.0 | 1.0 | 4.8 | 5.3 | 5.0 | 9.7 | 10.3 | 9.9 |
| | 20 | 1.0 | 1.5 | 1.1 | 4.9 | 5.5 | 5.1 | 10.0 | 10.4 | 10.3 |
| | 35 | 1.0 | 1.5 | 1.3 | 5.0 | 6.1 | 5.4 | 10.0 | 11.0 | 10.6 |
| | 50 | 1.2 | 1.3 | 1.2 | 5.2 | 5.6 | 5.3 | 10.6 | 10.8 | 10.4 |
| | | | | | $C_\xi = 10\%$ | | | | | |
| k | 2 | 0.9 | 0.8 | 0.8 | 4.5 | 4.5 | 4.1 | 9.3 | 9.2 | 8.9 |
| | 5 | 1.1 | 1.3 | 1.1 | 5.1 | 5.4 | 5.2 | 10.1 | 10.4 | 10.2 |
| | 10 | 0.9 | 1.0 | 1.2 | 4.9 | 5.1 | 5.2 | 9.9 | 10.3 | 10.2 |
| | 20 | 1.0 | 1.2 | 1.1 | 4.9 | 5.5 | 5.2 | 9.7 | 10.4 | 10.1 |
| | 35 | 1.0 | 1.4 | 1.1 | 5.1 | 5.4 | 5.2 | 9.8 | 10.6 | 10.4 |
| | 50 | 1.1 | 1.3 | 1.1 | 5.2 | 5.5 | 5.3 | 10.5 | 10.7 | 10.1 |
| | | | | | $C_\xi = 20\%$ | | | | | |
| k | 2 | 0.9 | 0.9 | 0.8 | 4.5 | 4.6 | 4.0 | 8.9 | 9.3 | 8.4 |
| | 5 | 1.1 | 1.2 | 1.2 | 5.2 | 5.0 | 5.2 | 10.2 | 10.2 | 10.2 |
| | 10 | 1.1 | 1.2 | 1.1 | 4.9 | 5.0 | 4.9 | 9.9 | 10.2 | 10.1 |
| | 20 | 1.0 | 1.2 | 1.3 | 5.3 | 5.3 | 5.3 | 10.0 | 10.5 | 10.2 |
| | 35 | 1.3 | 1.5 | 1.2 | 5.4 | 6.2 | 5.2 | 10.5 | 10.9 | 10.1 |
| | 50 | 1.2 | 1.3 | 1.1 | 5.7 | 5.8 | 4.9 | 10.2 | 11.0 | 9.8 |
| | | | | | $C_\xi = 40\%$ | | | | | |
| k | 2 | 1.0 | 0.8 | 0.9 | 4.6 | 4.0 | 4.3 | 9.6 | 8.2 | 8.7 |
| | 5 | 1.4 | 1.4 | 1.1 | 5.6 | 5.1 | 5.2 | 10.6 | 10.4 | 10.3 |
| | 10 | 1.2 | 1.4 | 1.3 | 5.2 | 5.7 | 5.5 | 10.4 | 10.8 | 10.5 |
| | 20 | 1.5 | 1.5 | 1.3 | 5.8 | 5.8 | 5.7 | 10.8 | 10.8 | 10.7 |
| | 35 | 1.7 | 1.5 | 1.5 | 6.2 | 6.1 | 5.5 | 11.3 | 11.1 | 10.0 |
| | 50 | 1.5 | 1.3 | 1.3 | 5.8 | 6.0 | 5.1 | 10.9 | 11.2 | 10.0 |

Table 7.14: Rejection rates (in $\%$) of multiple imputation methodology from simulation using the two-stage-likelihood-ratio-test-based method with $F_{k,w}$ given in (3.7) as reference distribution; $n = 5000$ and $m = 5$

Our simulation confirms that the two-stage-likelihood-ratio-test-based method is well calibrated, even for a relatively small sample size, here $n = 1000$ when the

average ratio of complete to observed information $\overline{\xi}$ is presumably less than $2$. If however, $\overline{\xi}$ is as large as $2$, then this method turns to be invalid.

#### 7.2.3.3 Componentwise-moment-based method

We now consider the componentwise-moment-based method with the suggested degrees of freedom, $v$, given in (6.7) for a setting with sample size $n = 1000$ and with number of imputations $m = 5$. Table 7.15 presents the corresponding rejection rates from our simulation.

| | | $\alpha = 1\%$ | | | $\alpha = 5\%$ | | | $\alpha = 10\%$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\overline{\xi}$ | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 |
| | | | | | $C_\xi = 0\%$ | | | | | |
| k | 2 | 1.1 | 1.3 | 1.3 | 5.0 | 5.2 | 5.2 | 9.9 | 10.0 | 9.9 |
| | 5 | 0.9 | 1.2 | 1.1 | 5.3 | 5.4 | 5.1 | 9.9 | 10.5 | 9.8 |
| | 10 | 1.0 | 1.1 | 1.2 | 4.8 | 5.4 | 5.3 | 9.6 | 10.7 | 10.3 |
| | 20 | 0.9 | 1.1 | 1.5 | 5.3 | 5.6 | 5.8 | 10.7 | 10.5 | 10.7 |
| | 35 | 1.0 | 1.2 | 2.1 | 5.5 | 5.5 | 7.2 | 10.9 | 10.7 | 13.0 |
| | 50 | 1.2 | 1.4 | 3.3 | 6.0 | 5.4 | 10.6 | 11.4 | 10.4 | 17.2 |
| | | | | | $C_\xi = 10\%$ | | | | | |
| k | 2 | 1.1 | 1.3 | 1.4 | 5.1 | 5.2 | 5.4 | 9.8 | 9.7 | 9.7 |
| | 5 | 1.2 | 1.3 | 1.0 | 5.2 | 5.5 | 5.1 | 10.3 | 10.4 | 10.1 |
| | 10 | 0.9 | 1.2 | 1.2 | 4.8 | 5.1 | 5.1 | 10.0 | 10.3 | 10.2 |
| | 20 | 1.0 | 1.3 | 1.3 | 5.3 | 5.3 | 5.3 | 10.7 | 10.1 | 10.3 |
| | 35 | 0.9 | 1.2 | 2.2 | 5.0 | 5.5 | 7.2 | 10.4 | 10.7 | 12.9 |
| | 50 | 1.0 | 1.2 | 3.6 | 5.4 | 5.3 | 10.8 | 10.6 | 9.7 | 17.8 |
| | | | | | $C_\xi = 20\%$ | | | | | |
| k | 2 | 1.2 | 1.2 | 1.2 | 5.3 | 5.5 | 5.1 | 10.0 | 10.1 | 10.1 |
| | 5 | 1.1 | 1.2 | 1.2 | 5.4 | 5.4 | 5.1 | 10.4 | 10.2 | 9.5 |
| | 10 | 1.0 | 1.2 | 1.2 | 5.3 | 5.5 | 5.1 | 10.4 | 10.5 | 10.1 |
| | 20 | 1.1 | 1.4 | 1.2 | 5.1 | 5.3 | 5.5 | 10.1 | 10.4 | 10.6 |
| | 35 | 1.0 | 1.2 | 2.1 | 5.0 | 5.3 | 7.4 | 10.0 | 10.2 | 12.9 |
| | 50 | 1.0 | 1.1 | 3.6 | 5.1 | 4.7 | 11.4 | 10.3 | 9.7 | 18.3 |
| | | | | | $C_\xi = 40\%$ | | | | | |
| k | 2 | 1.3 | 1.3 | 1.5 | 5.6 | 5.7 | 5.5 | 10.6 | 10.3 | 10.4 |
| | 5 | 1.6 | 1.4 | 1.3 | 5.6 | 5.8 | 5.0 | 10.3 | 10.6 | 9.9 |
| | 10 | 1.4 | 1.5 | 1.4 | 6.1 | 5.9 | 5.1 | 11.3 | 10.9 | 9.8 |
| | 20 | 1.3 | 1.4 | 1.4 | 5.3 | 5.4 | 5.4 | 10.4 | 10.2 | 10.5 |
| | 35 | 1.1 | 1.3 | 2.1 | 4.5 | 4.9 | 7.3 | 9.2 | 9.5 | 12.7 |
| | 50 | 0.7 | 1.0 | 2.7 | 3.3 | 4.1 | 9.4 | 7.2 | 8.4 | 16.5 |

Table 7.15: Rejection rates (in $\%$) of multiple imputation methodology from simulation using the componentwise-moment-based method with $F_{k,v}$ given in (6.7) as reference distribution; $n = 1000$ and $m = 5$

In Table 7.15 we see that the rejection rates for the componentwise-moment-based method with $v$ as degrees of freedom are accurate in almost all situations except when the average ratio of complete to observed information $\overline{\xi}$ equals $2$ and the dimension $k$ is equal to $35$ or equal to $50$. The worst situation is $k = 50$, $\overline{\xi} = 2$ and $C_\xi = 20\%$, where the rejection rate reaches as high as $18.3\%$ for $\alpha = 10\%$. The rejection rates in all situations with $\overline{\xi} \leq 1.5$ and $k \leq 20$ are very close to the corresponding nominal levels.

Increasing the number of imputations to $m = 30$ yields for $n = 1000$ very similar rejection rates, given in Table 7.16.

| | | $\alpha = 1\%$ | | | $\alpha = 5\%$ | | | $\alpha = 10\%$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\bar{\xi}$ | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 |
| | | | | | $C_\xi = 0\%$ | | | | | |
| k | 2 | 1.1 | 1.1 | 1.1 | 5.0 | 4.9 | 5.0 | 9.7 | 9.8 | 9.5 |
| | 5 | 0.9 | 1.1 | 1.0 | 5.0 | 5.1 | 4.9 | 10.2 | 10.0 | 9.9 |
| | 10 | 0.9 | 1.0 | 1.0 | 4.8 | 5.0 | 4.8 | 9.7 | 10.0 | 9.9 |
| | 20 | 1.0 | 1.0 | 0.9 | 5.1 | 4.8 | 4.9 | 10.7 | 9.7 | 9.9 |
| | 35 | 0.9 | 0.8 | 1.6 | 5.4 | 4.6 | 5.9 | 10.9 | 9.3 | 11.4 |
| | 50 | 1.1 | 0.8 | 2.4 | 5.5 | 4.3 | 8.2 | 11.0 | 8.7 | 15.0 |
| | | | | | $C_\xi = 10\%$ | | | | | |
| k | 2 | 1.0 | 1.2 | 1.2 | 5.2 | 5.0 | 4.7 | 9.8 | 9.9 | 9.5 |
| | 5 | 1.0 | 1.1 | 1.0 | 5.2 | 5.2 | 4.8 | 10.1 | 9.9 | 9.8 |
| | 10 | 0.8 | 1.1 | 0.9 | 4.8 | 5.0 | 4.8 | 9.8 | 10.1 | 9.8 |
| | 20 | 1.0 | 0.9 | 0.9 | 5.1 | 4.9 | 4.7 | 10.2 | 9.7 | 9.6 |
| | 35 | 1.0 | 0.7 | 1.5 | 4.8 | 4.5 | 6.1 | 10.2 | 9.1 | 11.4 |
| | 50 | 1.0 | 0.8 | 2.2 | 5.1 | 3.9 | 8.7 | 10.5 | 8.6 | 15.6 |
| | | | | | $C_\xi = 20\%$ | | | | | |
| k | 2 | 1.1 | 1.1 | 1.1 | 4.9 | 5.0 | 5.1 | 10.1 | 9.7 | 9.6 |
| | 5 | 1.0 | 1.0 | 0.8 | 5.2 | 5.3 | 4.9 | 10.1 | 10.2 | 9.6 |
| | 10 | 0.9 | 1.0 | 0.9 | 4.9 | 5.2 | 4.7 | 9.7 | 10.1 | 10.1 |
| | 20 | 0.9 | 0.8 | 0.8 | 4.9 | 4.7 | 4.9 | 10.0 | 9.4 | 9.7 |
| | 35 | 0.9 | 0.8 | 1.5 | 4.7 | 4.2 | 6.3 | 9.4 | 9.0 | 11.2 |
| | 50 | 0.7 | 0.7 | 2.2 | 4.4 | 3.7 | 9.5 | 9.5 | 8.2 | 16.6 |
| | | | | | $C_\xi = 40\%$ | | | | | |
| k | 2 | 1.2 | 1.1 | 1.2 | 4.9 | 4.9 | 5.2 | 10.1 | 9.9 | 10.4 |
| | 5 | 1.4 | 1.1 | 1.1 | 5.2 | 5.1 | 5.2 | 9.9 | 10.1 | 10.0 |
| | 10 | 1.1 | 1.1 | 1.0 | 5.3 | 5.4 | 5.1 | 10.2 | 9.9 | 9.8 |
| | 20 | 1.0 | 0.8 | 0.9 | 4.7 | 4.7 | 4.9 | 9.3 | 9.3 | 9.4 |
| | 35 | 0.7 | 0.8 | 1.3 | 3.8 | 3.9 | 6.1 | 8.0 | 7.8 | 10.9 |
| | 50 | 0.4 | 0.7 | 1.5 | 2.8 | 3.1 | 7.8 | 6.4 | 6.6 | 14.5 |

Table 7.16: Rejection rates (in %) of multiple imputation methodology from simulation using the componentwise-moment-based method with $F_{k,v}$ given in (6.7) as reference distribution; $n = 1000$ and $m = 30$

From Table 7.16 we see that compared with the results for $n = 1000$ and $m = 5$ given in Table 7.15 the rejection rates are similar, only the maximal deviations of

the rejection rates from the corresponding nominal levels are a bit smaller with $m = 30$. Increasing the sample size to $n = 5000$ improves the results. The rejection rates for $n = 5000$ and $m = 30$ are presented in the following Table 7.17.

| | | $\alpha = 1\%$ | | | $\alpha = 5\%$ | | | $\alpha = 10\%$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\overline{\xi}$ | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 |
| | | | | | $C_\xi = 0\%$ | | | | | |
| k | 2 | 1.1 | 1.1 | 0.8 | 5.0 | 4.7 | 4.5 | 9.9 | 9.6 | 9.2 |
| | 5 | 1.0 | 1.1 | 1.0 | 4.7 | 5.1 | 4.7 | 9.9 | 9.9 | 9.8 |
| | 10 | 1.0 | 1.0 | 0.9 | 4.6 | 4.6 | 4.6 | 9.7 | 9.8 | 9.1 |
| | 20 | 0.9 | 1.1 | 0.9 | 5.0 | 4.9 | 4.4 | 9.6 | 9.9 | 8.8 |
| | 35 | 1.1 | 1.1 | 0.8 | 4.8 | 5.0 | 4.2 | 9.8 | 10.1 | 8.9 |
| | 50 | 1.0 | 1.0 | 0.7 | 4.9 | 4.6 | 4.4 | 10.0 | 9.4 | 9.4 |
| | | | | | $C_\xi = 10\%$ | | | | | |
| k | 2 | 1.0 | 1.1 | 0.9 | 5.1 | 5.0 | 4.5 | 9.9 | 9.6 | 9.9 |
| | 5 | 1.1 | 1.2 | 0.8 | 4.9 | 5.0 | 4.7 | 10.2 | 10.0 | 9.9 |
| | 10 | 1.0 | 0.9 | 0.9 | 4.9 | 4.7 | 4.5 | 9.7 | 9.7 | 9.1 |
| | 20 | 0.8 | 1.0 | 0.9 | 4.7 | 4.9 | 4.2 | 9.6 | 9.5 | 9.1 |
| | 35 | 1.0 | 1.1 | 0.7 | 4.8 | 4.8 | 4.0 | 9.8 | 9.5 | 8.7 |
| | 50 | 1.0 | 1.0 | 0.7 | 5.1 | 4.5 | 4.5 | 10.0 | 9.1 | 9.2 |
| | | | | | $C_\xi = 20\%$ | | | | | |
| k | 2 | 1.1 | 1.1 | 0.9 | 5.1 | 5.2 | 4.5 | 9.8 | 10.0 | 9.5 |
| | 5 | 1.2 | 1.1 | 1.0 | 5.0 | 5.0 | 4.7 | 10.2 | 10.3 | 9.9 |
| | 10 | 1.1 | 1.0 | 0.9 | 4.9 | 4.7 | 4.5 | 9.8 | 9.7 | 9.0 |
| | 20 | 0.9 | 0.9 | 0.8 | 4.9 | 4.6 | 4.4 | 9.7 | 9.7 | 9.1 |
| | 35 | 1.1 | 1.1 | 0.7 | 4.9 | 4.8 | 4.1 | 9.9 | 9.6 | 8.7 |
| | 50 | 1.1 | 1.0 | 0.6 | 4.9 | 4.6 | 4.3 | 9.6 | 9.3 | 9.1 |
| | | | | | $C_\xi = 40\%$ | | | | | |
| k | 2 | 1.1 | 1.0 | 1.0 | 5.0 | 4.9 | 4.9 | 9.7 | 9.5 | 9.8 |
| | 5 | 1.3 | 1.3 | 1.0 | 5.4 | 5.5 | 5.1 | 10.5 | 10.2 | 10.0 |
| | 10 | 1.1 | 1.1 | 0.9 | 5.2 | 5.2 | 4.7 | 10.1 | 10.1 | 9.8 |
| | 20 | 1.1 | 1.2 | 0.7 | 5.4 | 4.9 | 4.4 | 10.2 | 9.8 | 9.2 |
| | 35 | 1.4 | 1.1 | 0.8 | 5.4 | 4.9 | 4.4 | 10.3 | 9.9 | 9.0 |
| | 50 | 1.2 | 1.1 | 0.7 | 5.0 | 4.6 | 4.2 | 9.7 | 9.2 | 9.0 |

Table 7.17: Rejection rates (in %) of multiple imputation methodology from simulation using the componentwise-moment-based method with $F_{k,v}$ given in (6.7) as reference distribution; $n = 5000$ and $m = 30$

Table 7.17 shows that with a sample size of $n = 5000$ and $m = 30$ imputations, all rejection rates are well calibrated with a slight tendency to be too conservative. The maximal deviation of the rejection rates from the nominal level is $1.2$-percent points for $\alpha = 10\%$ when $\overline{\xi} = 2$, $C_\xi = 0\%$ and $k = 20$.

From Table 7.15 and Table 7.15 it follows that for the componentwise-moment-based method with $v$ as degrees of freedom, the combination of the sample size and the average ratio of complete to observed information, $n * \overline{\xi}$, has a strong influence on the validity of the rejection rates. The corresponding ANOVA given in Table 7.5 pointed out that the two-way interaction between the $n$ and $\overline{\xi}$ is the factor that most strongly affects the rejection rate.

Our method is well calibrated for a small sample size, here $n = 1000$, when we have a dimension, presumably, less than $35$ or an average ratio of complete to observed information $\overline{\xi}$ less or equal than $1.5$, which corresponds to a missing rate of $33\%$ here. For $n = 5000$ and $m = 30$ the rejection rates are almost perfectly calibrated with a slight tendency to be too conservative as the ranking of the "method and degrees of freedom"-combinations given in Table 7.7 already indicates.

#### 7.2.3.4 $\chi^2_k$-statistic-based approach

Although the procedure based on the set $S_d$ proposed by Li, Meng, Raghunathan and Rubin (1991), and described in detail in Section 3.3, is only approximately calibrated and has a substantial loss of power, we analyze the results of this method for sake of completeness. If we consider the rejection rates corresponding to the best setting with sample size $n = 5000$ and number of imputations $m = 30$, which are presented in Table 7.18, we see that the results are by far not as good as the three other methods we have already onsidered.

| | | $\alpha = 1\%$ | | | $\alpha = 5\%$ | | | $\alpha = 10\%$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\overline{\xi}$ | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 | 1.2 | 1.5 | 2 |
| | | | | | $C_\xi = 0\%$ | | | | | |
| k | 2 | 1.1 | 1.1 | 1.1 | 5.0 | 4.9 | 5.3 | 9.9 | 10.4 | 10.6 |
| | 5 | 1.2 | 1.0 | 1.5 | 5.1 | 5.7 | 6.7 | 10.5 | 11.2 | 13.4 |
| | 10 | 1.1 | 1.2 | 1.9 | 5.2 | 6.2 | 8.6 | 10.5 | 12.4 | 16.7 |
| | 20 | 1.1 | 1.8 | 3.3 | 5.8 | 7.4 | 11.9 | 11.5 | 13.8 | 20.5 |
| | 35 | 1.5 | 2.5 | 4.8 | 6.8 | 9.2 | 14.9 | 12.3 | 16.1 | 23.9 |
| | 50 | 1.7 | 3.1 | 5.0 | 7.5 | 9.7 | 14.7 | 13.6 | 16.8 | 23.2 |
| | | | | | $C_\xi = 10\%$ | | | | | |
| k | 2 | 1.0 | 1.1 | 1.1 | 5.2 | 5.4 | 5.3 | 10.3 | 10.1 | 11.3 |
| | 5 | 1.1 | 1.1 | 1.3 | 5.2 | 5.8 | 6.7 | 10.7 | 11.3 | 13.6 |
| | 10 | 1.2 | 1.4 | 1.9 | 5.3 | 6.5 | 9.4 | 10.5 | 12.5 | 17.3 |
| | 20 | 1.1 | 1.6 | 3.6 | 6.1 | 7.8 | 12.9 | 11.5 | 14.3 | 22.1 |
| | 35 | 1.5 | 2.6 | 4.6 | 6.3 | 8.8 | 14.5 | 12.3 | 15.8 | 23.4 |
| | 50 | 2.0 | 2.8 | 5.1 | 7.7 | 10.2 | 15.5 | 13.6 | 17.3 | 24.1 |
| | | | | | $C_\xi = 20\%$ | | | | | |
| k | 2 | 1.1 | 1.1 | 1.0 | 5.0 | 5.4 | 5.5 | 10.1 | 10.4 | 11.2 |
| | 5 | 1.1 | 1.1 | 1.4 | 4.9 | 5.9 | 7.2 | 10.5 | 11.8 | 14.4 |
| | 10 | 1.2 | 1.2 | 2.1 | 5.1 | 6.1 | 9.2 | 10.0 | 11.9 | 17.2 |
| | 20 | 1.1 | 1.7 | 3.5 | 5.2 | 7.5 | 12.9 | 10.5 | 14.1 | 22.1 |
| | 35 | 1.4 | 2.7 | 4.5 | 6.3 | 9.5 | 14.7 | 11.4 | 16.4 | 24.1 |
| | 50 | 1.5 | 3.4 | 4.9 | 6.3 | 10.4 | 14.6 | 12.1 | 17.7 | 23.0 |
| | | | | | $C_\xi = 40\%$ | | | | | |
| k | 2 | 1.1 | 0.9 | 1.0 | 4.9 | 5.1 | 5.6 | 9.8 | 10.2 | 11.6 |
| | 5 | 1.0 | 1.0 | 1.3 | 4.8 | 5.6 | 6.9 | 9.8 | 11.5 | 13.6 |
| | 10 | 0.9 | 1.1 | 2.0 | 4.5 | 5.8 | 8.6 | 9.0 | 11.6 | 16.1 |
| | 20 | 0.9 | 1.6 | 3.0 | 4.2 | 6.3 | 10.8 | 8.2 | 12.5 | 19.3 |
| | 35 | 0.8 | 2.0 | 3.5 | 3.9 | 7.4 | 12.1 | 7.3 | 13.2 | 20.1 |
| | 50 | 0.9 | 2.2 | 3.8 | 3.7 | 7.8 | 11.6 | 7.1 | 13.3 | 18.7 |

Table 7.18: Rejection rates (in %) of multiple imputation methodology from simulation using the $\chi^2_k$-statistic-based approach with $F_{k,a_{k,m}w_s}$ given in (3.13) and (3.12) as reference distribution; $n = 5000$ and $m = 30$

In Table 7.18 we see that for the $\chi^2_k$-statistic-based approach, the rejection rates are wildly off the mark in some cases. For the situations with $\overline{\xi} = 2$, $C_\xi = 10\%$, $k = 50$ and $\overline{\xi} = 2$, $C_\xi = 20\%$, $k = 35$, the rejection rates reach as high as $24.1\%$ for $\alpha = 10\%$. Thus, the maximal deviation of the rejection rates to the nominal level $\alpha$ is $14.1$-percent points. Only for the smallest dimension, $k = 2$, the rejection rates are very close to the nominal levels. Taking other degrees of freedom for example, $v$ given in (6.7) or $(k+1)/2v_{\text{min}}$ introduced in Section 7.1., does not improve the results. We therefore do not present the correspoonding tables.

### 7.2.4 Conclusions

After analysing the results of our simulation study, we see that the two-stage-likelihood-ratio-test-based method is the best procedure that can be applied. This was to be expected, because it uses the best information the data provides and the relationship between the Wald-test statistic and the likelihood-ratio-test statistic. The only disadvantage of the two-stage-likelihood-ratio-test-based procedure is the need of access to the code for the likelihood-ratio-test statistic, that is not provided in standard statistical software packages. Hence, it is worth the effort deriving a new procedure that is easier to compute.

The componentwise-moment-based procedure derived in this thesis using the originally proposed degrees of freedom $v$ suffices the requirements and yields very good results even under difficult conditions such as a small sample size, a small number of imputations, and a high average of complete to observed information. Only when the sample size is small, here $n = 1000$, and simultaneously the average ratio of complete to observed information is presumably $\geq 2$, the rejection rates exceed the particular nominal levels, and thus the null hypothesis is rejected too often. Anyway, it is a practical advice to impute data sets only if the missing rate is less or equal than $20\%$, depending on the settings. A number of imputations $m = 5$ is sufficient, but increasing the number of imputations to $m = 30$ improves the results. Due to today's computer power, it is no problem to use $m = 30$ imputations to get improved results. Hence, the restriction $\bar{\xi} < 2$ of our method should be no problem in practical applications. Due to the restriction in the average ratio of complete to observed information and the dimension, when the sample size is $1000$, the componentwise-moment-based method is less accurate than the likelihood-ratio-based method, which is the best procedure that can be used. Our procedure performs better than the moment-based procedure for small sample size and is in particular much easier to apply, because providing covariance-variance matrices is not needed. We also used other degrees of freedom for the $F_{k,.}$-reference distribution. For example, the rejection rates with $(k+1)/2v_{\min}$ as degrees of freedom for the $F_{k,.}$-reference

distribution (not presented in this thesis) are well calibrated with $n = 5000$ and $m = 30$, but with sample size $n = 1000$ the deviations of the rejection rates to the corresponding nominal levels are higher when using $(k + 1)/2v_{\text{min}}$ instead of the originally proposed degrees of freedom $v$. Consequently, $v$ is the best choice for the componentwise-moment-based method. A remaining task is to analytically derive the degrees of freedom $v$ of our new componentwise-moment-based procedure.

In contrast to the componentwise-moment-based procedure, the moment-based method is fully analytically derived under the large-sample assumption and for $m \to \infty$ it is identical with the ideal procedure - the two-stage-likelihood-ratio-test based directly on the observed data. But the variance-covariance matrices needed for its calculation are not provided by statistical standard software packages and/or the calculation of the inverse of the within variance-covariance matrix might be not possible especially if the dimenison is high. This was the motivation of this thesis. Our simulation study showed that the moment-based method does not work for a small sample size, here $n = 1000$, relative to the dimension. Increasing the number of imputations to $m = 30$ does not improve the results. If the sample size is large enough, here $n = 5000$, the procedure is already well calibrated for a small number of imputations, e.g. $m = 5$. We can improve the results, if we use the degrees of freedom $v_{\text{mean}}$, that are based on the originally proposed degrees of freedom $v$ of the componentwise-moment-based procedure. If we use at least $m = 30$, the moment-based procedure using $v_{\text{mean}}$ performs well for a small sample size, $n = 1000$, with a slightly tendency to be too conservative. We also examined the moment-based method using $v$ as degrees of freedom. The rejection rates exceed the particular nominal levels $\alpha$ also when using $w$ as degrees of freedom, thus, the null hypothesis is rejected too often. Only the deviations of the rejection rates from the nominal levels are a bit smaller when using $v$ instead of $w$. It seems that degrees of freedom between $v_{\text{mean}}$ and $v$ might be a good choice to improve results. Of course, more detailed research is needed to verify this advice. Finally, the moment-based method with their originally proposed degrees

of freedom, $w$, is only suitable, if we have a large sample size and the necessary statistics are provided by statistical software packages.

The procedure with the fewest requirements, the $\chi^2_k$-statistic-based approach, is not suitable due to its wrong significance levels. An exception are situations where the dimension $k$ is really small, e.g. $k = 2$, but then also the componentwise-moment-based procedure is applicable and the effort of calculation is not much higher.

In the next chapter we will summarize our work and give practical advices to the reader how to calculate valid significance levels from multiply-imputed data.

# 8

# Summary and practical advices

Missing data are a pervasive problem in statistics. Multiple imputation, first proposed by Rubin (1977, 1978), is a general statistical technique to handle missing data. A difficult problem in the analysis of a multiply-imputed data set is how to combine repeated p-values to create valid inference and significance levels, respectively. Rubin (1987), Li, Raghunathan, and Rubin (1991), Li, Meng Raghunathan, and Rubin (1991), and Meng and Rubin (1992) suggested several methods, but none of them is fully satisfying. The procedure developed by Meng and Rubin (1992) is based on the likelihood-ratio-test of the completed-data. This procedure requires access to the code for calculating likelihood-ratio-test statistics that is not provided by statistical analysis software. If there is the possibility to get access to code, this method is the best that can be applied to get valid significance levels from repeated p-values when the likelihood function is specified correctly. In addition, our simulation study based on the simplest settings with multivariate standard normally distributed data, no correlation between the variables and a missing completely at random data drop out, points out that a sample size of $n = 1000$ and a number of imputations $m = 5$ are sufficient to reach rejection rates as high as the nominal level $\alpha$ with the two-stage-likelihood-ratio-based method. An exception in the extreme situation when the coefficient of variation $C_\xi$ is presumably $\geq 40\%$ and the dimension $k$ is presumably $\geq 50$. Otherwise this method is well calibrated and should be applied whenever possible.

Another method with similar problems is a procedure based on the set of completed-data moments and is proposed by Li, Raghunathan, and Rubin (1991). It is called the moment-based procedure and uses the complete-data variance-covariance matrices of the parameter estimates. However, providing the completed-data variance-covariance matrices is not standard in statistical analysis software and the necessary calculation of the inverse of the within variance-covariance matrix may become expensive with increasing dimension of the estimate. This method is analytically derived for large sample sizes. Our simulation study shows that even a large number of imputations $m = 30$ does not improve the results when the sample size is small, here $n = 1000$. Only raising the sample size corresponding to the dimension yields valid inferences, e.g., with dimension $k = 50$ a sample size of $n = 5000$ is needed. We can improve this method by taking $v_{\text{mean}}$ based on the new degrees of freedom $v$ given in (6.7) as degrees of freedom instead of the originally proposed degrees of freedom $w$ given in (3.7). If we are able to use at least $m = 30$ imputations, we attain accurate rejection rates, even with a small sample size of $n = 1000$, when we use $v_{\text{mean}}$ as degrees of freedom. The moment-based method with its originially proposed degrees of freedom, $w$, is not appropriate except with a large sample size. Otherwise, the new degrees of freedom, $v_{\text{mean}}$, can be used, but it is always necessary to provide the variance-covariance-matrices and the inverse of the within variance-covariance matrix. Also, $v_{\text{mean}}$ is not yet analytically derived and further research is needed.

The third method, called the $\chi^2_k$-statistic-based approach, presented by Li, Meng, Raghunathan, and Rubin (1991), requires only the collection of completed-data $\chi^2_k$-statistics (or distances), which is offered by every statistical software package. But due to the loss of information when going from the set of completed-data moments to the set of completed-data $\chi^2_k$-statistics, this procedure is only approximately calibrated and suffers from a substantial loss of power. In addition, our simulation shows that, even under the null hypothesis and ideal conditions like a large sample size, a large number of imputations, and a small dimension,

the corresponding rejection rates are far from their nominal levels. Thus, the authors suggested that this method should primarily be used as a screening test statistic to get a range of p-values.

The disadvantages of all these methods were the motivation for this thesis - finding a new procedure that performs as well as the two-stage-likelihood-ratio-test-based procedure and that requires only standard statistical software quantities. The first method we suggest, the $z$-transformation, is very easy and intuitive. It works very well but it is restricted to one-dimensional tests. Besides several ideas to use this $z$-transformation, we finally develop the componentwise-moment-based procedure using the small-sample degrees of freedom from Barnard and Rubin (1999), componentwise. Our simulation study shows, that with normally distributed and uncorrelated data this procedure works well except under extreme conditions, such as a high average ratio of complete to observed information $\overline{\xi}$ presumably $\geq 2$, and simultaneously a large coefficient of variation $C_\xi$ presumably $\geq 40\%$. In these cases a larger sample size and a larger number of imputations, e.g. $m \geq 30$, are necessary.

Finally we want to give the reader some practical advice concerning generating valid significance levels from multiply-imputed data.

**1. Advice for the data collector**

- In general, if the percentage of missing values in each variable is higher than $20\%$, try to get a new data set or more information, to be sure to provide a data set that enables the imputer to create valid imputations, and thus enables the data analyst to get valid inferences.
- Try to get a sample size of at least $n = 1000$ depending on the dimension, that is, the higher the dimension of the estimand, the data analyst might be interested in, the larger the sample size should be. If there is no information about the dimension of the estimand, try to get a sample size as large as possible, but at least $n = 1000$.

**2. Advice for the imputer**

- In general, if the percentage of missing values in each variable is higher than $20\%$, do not multiply impute the values to be sure to provide a multiply-imputed data set that enables the data analyst to get valid inferences.
- Generally use the largest number of imputations that is possible.

**3. Advice for the data analyst**

- If only one-dimensional tests are performed, use our new $z$-transformation to combine repeated p-values.
- If there is access to the code for the likelihood-ratio-test statistic, use the two-stage-likelihood-ratio-test-based method to generate significance levels that are valid if the likelihood function is correctly specified.
- If there is only access to standard statistical software quantities, use our new componentwise-moment-based method. A sample size of $n = 1000$ for a dimension of $k = 50$ and $m = 5$ imputations are sufficient. But pay attention to the average ratio of complete to observed information, $\overline{\xi}$, and the coefficient of variation, $C_\xi$. They should be presumably less than $2$ and $40\%$, respectively, to be sure to get valid inferences.

# 9

# Future tasks and outlook

As we point out at different places in this work, there are some very interesting open problems and future tasks. As a general problem in statistics detached from any imputation theory, first of all we revealed the problem of using a $\chi^2_k$-distribution devided by $k$ rather than the correct $F$-distribution for small sample sizes. It would be important to explicitly calculate the error that is done if the $\chi^2_k$-distribution is used as the reference distribution. We need an analytical approximation of the cumulative distribution function of an $F$-distribution and an analytical approximation of the quantile function of a $\chi^2_k$-distribution. Then the correct value of the $\chi^2_k$-statistic can be calculated as corresponding quantile of the p-value from the $F$-statistic. Beside this apparently unnoticed general problem, some other research-topics arised.

Another difficult problem is, for example, to derive the Barnard and Rubin degrees of freedom (1991) for dimensions $k > 1$. Then we would be able to create a procedure based on the multivariate degrees of freedom for small sample sizes. We also could derive new degrees of freedom for the moment-based method to generate small-sample significance levels and thus further improve this method. The difficulty of the generalisation of the Barnard and Rubin degrees of freedom consists in handling multivariate $t$-distributions instead of multivariate normal distributions, because with small sample sizes we have to use $t$-distributions instead of normal distributions.

Moreover, it seems to be desirable to analytically derive the degrees of freedom of our componentwise-moment-based procedure, which are at the moment mostly based on simulation-supported considerations. That is, we would like to justify our componentwise-moment-based method using a mathematical proof, and assess the results by adequate simulation studies, which is beyond the scope of this thesis. It is also important to analyse the different procedures, and especially the componentwise-moment-based procedure, under the alternative hypothesis and to consider and compare the methods for their power. In addition, more complex models with correlation between the independent variables, with a MAR-missing-mechanism, and multivariate $t$-distributed variables could be studied in general. Initially that can be done using simulation studies, but we wish to base some further analyses on a mathematical foundament. Finally, we hope that our procedure will still work as well as the two-stage-likelihood-ratio-test-based procedure, at least approximately. In the long run, we would like to see a procedure that only needs the standardly provided quantities of common statistical software and simultaneously fulfills the requirements of calculating valid significance levels from multiply-imputed data. It would raise the applicability of multiple imputation to a great extent and at the same time it would ease the work of applied statisticians.

# List of Figures

# List of Tables

# Appendix A

# Derivation of (3.1)-(3.5) from Section 3.1

Let $\theta_t$ be the true value of the $k$-component parameter of interest, $\theta$, and let $\hat{\theta}$ be the complete-data maximum likelihood estimate of $\theta$. In the following the subscript $t$ designates the true value. We assume that

$$(\hat{\theta}|\theta = \theta_t) \sim N(\theta_t, U_t), \tag{A.1}$$

where $U_t = V(\hat{\theta}|\theta = \theta_t)$ and $U_t^{-1}$ is the complete-data information.

Let $\hat{\theta}_{\text{obs}}$ be the maximum-likelihood estimate of $\theta$. Then we have

$$(\hat{\theta}_{\text{obs}}|\theta = \theta_t) \sim N(\theta_t, T_t), \tag{A.2}$$

where $T_t = V(\hat{\theta}_{\text{obs}}|\theta = \theta_t)$ and $T_t^{-1}$ is the observed information.

From (A.1) and (A.2) it follows, that the increase in variance dure to missing data is

$$B_t = T_t - U_t, \tag{A.3}$$

and that the missing information is $U_t^{-1} - T_t^{-1}$. The ratios of missing to observed information are given by the eigenvalues of $(U_t^{-1} - T_t^{-1})T_t$, which we label by

$(\lambda_1, \ldots, \lambda_k)$. Because of

$$(U_t^{-1} - T_t^{-1})T_t = U_t^{-1}T_t - I = U_t^{-1}(B_t + U_t) - I = U_t^{-1}B_t + I - I = U_t^{-1}B_t, \quad \text{(A.4)}$$

where $I$ denotes the $k$-dimensional identity-matrix, $(\lambda_1, \ldots, \lambda_k)$ are also the eigenvalues of $B_t$ relative to $U_t$. Since complete information is observed information plus missing information the ratios of complete to observed information

$$U_t^{-1}T_t = (T_t^{-1} + (U_t^{-1} - T_t^{-1}))T_t = I + (U_t^{-1} - T_t^{-1})T_t \quad \text{(A.5)}$$

are given by the eigenvalues $\xi_i$ $(i = 1, \ldots, k)$ of $U_t^{-1}T_t$. From (A.5) it follows that

$$\xi_i = 1 + \lambda_i. \quad \text{(A.6)}$$

Furthermore, the ratios of missing to complete information, that is the fractions of missing information, $\gamma_i$,

$$(U_t^{-1} - T_t^{-1})U_t = I - T_t^{-1}U_t = I - (U_t^{-1}T_t)^{-1} \quad \text{(A.7)}$$

are given by the eigenvalues of $(U_t^{-1} - T_t^{-1})U_t$. From (A.7) and (A.4) it follows that

$$\gamma_i = 1 - \frac{1}{1 + \lambda_i} = \frac{1 + \lambda_i - 1}{1 + \lambda_i} = \frac{\lambda_i}{1 + \lambda_i}, \quad \text{(A.8)}$$

and thus

$$\begin{aligned} & \lambda_i && = \gamma_i(1 + \lambda_i) \\ \Leftrightarrow\ & \lambda_i && = \gamma_i + \gamma_i\lambda_i, \\ \Leftrightarrow\ & \lambda_i - \gamma_i\lambda_i && = \gamma_i, \\ \Leftrightarrow\ & \lambda_i(1 - \gamma_i) && = \gamma_i, \\ \Leftrightarrow\ & \lambda_i && = \tfrac{\gamma_i}{1-\gamma_i}. \end{aligned} \quad \text{(A.9)}$$

From (A.6) and (A.9) it follows

$$\xi_i = 1 + \lambda_i = 1 + \frac{\gamma_i}{1 - \gamma_i} = \frac{1 - \gamma_i + \gamma_i}{1 - \gamma_i} = \frac{1}{1 - \gamma_i} = (1 - \gamma_i)^{-1}, \quad \text{(A.10)}$$

wherefrom (3.1) given in Section 3.1 follows.

Furthermore let $C_\xi$ be the coefficient of variation of the $\xi_i$ defined as

$$1 + C_\xi^2 = \frac{1}{k}\sum_{i=1}^{k}(\xi_i/\overline{\xi})^2, \tag{A.11}$$

where $\overline{\xi} = \frac{1}{k}\sum_{i=1}^{k}\xi_i$ denotes the average ratio of complete to observed information. Definition (A.11) equals (3.2) given in Section 3.1. To motivate this definition, one consider $R_i = \xi_i/\overline{\xi}$ as random variables with expectation $E(R_i) = \frac{1}{k\overline{\xi}}\sum_{i=1}^{k}\xi_i = \frac{\overline{\xi}}{\overline{\xi}} = 1$. Then the variance $V(R_i) =: C_\xi^2$ of the $R_i$ can be calculated as

$$V(R_i) = E(R_i^2) - (E(R_i))^2 = \frac{1}{k}\sum_{i=1}^{k}(\xi_i/\overline{\xi})^2 - 1,$$

wherefrom (A.11) immediately follows.

Let $\{X_{*l}, l = 1, \ldots, m\}$ denote the $m$ completed data sets and let $\hat{\theta}_{*1}, \ldots, \hat{\theta}_{*m}$ denote the $m$ repeated completed-data estimates with the associated variance-covariance matrices $U_{*1}, \ldots, U_{*m}$. From Rubin (1987, Chapter 4) it follows for large samples that

$$(\hat{\theta}_{*l}|\theta = \theta_t, X_{\text{obs}}) \overset{iid}{\sim} N(\hat{\theta}_{\text{obs}}, B_t). \tag{A.12}$$

Therefrom it follows

$$(\overline{\theta}_m|\theta = \theta_t, X_{\text{obs}}) \sim N(\hat{\theta}_{\text{obs}}, B_t/m). \tag{A.13}$$

Obviously it is

$$(B_t^{-1/2} B_m B_t^{-1/2}|\theta = \theta_t, X_{\text{obs}}) \sim \text{ Wishart with } k \text{ components and } k-1 \text{ dof}, \tag{A.14}$$

where $B_m = \frac{1}{m-1}\sum_{l=1}^{m}(\hat{\theta}_{*l} - \overline{\theta}_m)^t(\hat{\theta}_{*l} - \overline{\theta}_m)$ is the the between-variance. If the true variances $B_t$ and $T_t = B_t + U_t$ are known, then it follows from (A.2) and (A.13)

$$(\overline{\theta}_m|\theta = \theta_t) \sim N(\theta_t, (B_t + U_t) + B_t/m) = N(\theta_t, U_t + (1 + m^{-1})B_t). \tag{A.15}$$

We derive the distribution of the test statistic $\tilde{D}_m$ given in (2.12) in Chapter 2 as

$$\tilde{D}_m = (1 + r_m)^{-1}(\overline{\theta}_m - \theta_0)\overline{U}_m^{-1}(\overline{\theta}_m - \theta_0)^t/k \tag{A.16}$$

with $r_m$ given in (2.9) in Chapter 2 as

$$r_m = (1 + m^{-1})\text{Tr}(B_m\overline{U}_m^{-1})/k. \tag{A.17}$$

For a large sample size $n$ the within-variance $\overline{U}_m = \frac{1}{m}\sum_{l=1}^{m} U_{*l}$ given in (2.3) in Chapter 2 can be replaced by $U_t$. Because of the affine invariance of $\tilde{D}_m$, we can set $\theta_0 = 0$, $U_t = I$, and $B_t = \text{diag}(\lambda_1, \ldots, \lambda_k)$ with no loss of generality. From (A.16) and (A.17) it follows that $\tilde{D}_m$ can be written as

$$\tilde{D}_m = \frac{\sum_{i=1}^{k} \overline{\theta}_{m,i}^2/k}{1 + r_m} \tag{A.18}$$

with

$$\begin{aligned} r_m &= (1 + m^{-1})\text{Tr}(B_m\overline{U}_m^{-1})/k = (1 + m^{-1})\text{Tr}(B_m)/k, \\ &= (1 + m^{-1})\text{Tr}\left(\tfrac{1}{m-1}\sum_{l=1}^{m}(\hat{\theta}_{*l} - \overline{\theta}_m)^t(\hat{\theta}_{*l} - \overline{\theta}_m)\right)/k, \\ &= (1 + m^{-1})\left(\sum_{i=1}^{k}\tfrac{1}{m-1}\sum_{l=1}^{m}(\hat{\theta}_{*l} - \overline{\theta}_m)_i^2\right)/k, \\ &= (1 + m^{-1})\sum_{i=1}^{k}\sum_{l=1}^{m}(\hat{\theta}_{*l} - \overline{\theta}_m)_i^2/(k(m-1)), \end{aligned} \tag{A.19}$$

where the subscript $i$ indexes the $i$th component of $\overline{\theta}_m$ and $(\hat{\theta}_{*l} - \overline{\theta}_m)$. (A.18) and (A.19) are equal to (3.3) and (3.4). Because we set $\theta_{t,i} = 0$, $B_{t,i} = \lambda_i$, and $U_{t,i} = 1$ it follows from (A.15) and (A.14) that for $i = 1, \ldots, k$

$$(\overline{\theta}_{m,i}|\theta = \theta_t) \overset{ind}{\sim} N(0, 1 + (1 + m^{-1})\lambda_i) \tag{A.20}$$

and

$$\sum_{l=1}^{m}(\hat{\theta}_{*l} - \overline{\theta}_m)_i^2|\theta = \theta_t \overset{ind}{\sim} \lambda_i\chi_{m-1}^2. \tag{A.21}$$

When all $\theta_{t,i} = 0$ and all ratios of missing to observed information are equal, i.e., when all $\lambda_i = \overline{\lambda}$, (A.18)-(A.21) imply

$$\begin{aligned}
\tilde{D}_m &\sim \frac{(1+(1+m^{-1})\overline{\lambda})\chi^2_k/k}{1+((1+m^{-1})\overline{\lambda}\chi^2_{k(m-1)})/(k(m-1))}, \\
&= \frac{\chi^2_k/k}{\left[1+((1+m^{-1})\overline{\lambda}\chi^2_{k(m-1)})/(k(m-1))\right]/(1+(1+m^{-1})\overline{\lambda})}, \qquad \text{(A.22)} \\
&= \frac{\chi^2_k/k}{(1+a\chi^2_b/b)(1+a)},
\end{aligned}$$

where $b = k(m-1)$ and $a = (1+m^{-1})\overline{\lambda}$. (A.22) equals (3.5) in Section 3.1.

# Appendix B

# Derivation of the degrees of freedom $\delta$ and $w$ in the moment-based procedure described in Section 3.1

Li (1985), Rubin (1987) and Raghunathan (1987) derived the following distribution of the test statistic $D_m$ given in (3.5):

$$D_m \sim \frac{\chi_k^2/k}{(1+a\chi_b^2/b)/(1+a)}, \tag{B.1}$$

Li, Raghunathan and Rubin (1991) suggested a moment matching method to approximate the distribution of $D_m$ in (B.1) by a multiple of an $F$-distribution, $\delta F_{k,w}$. First we build the first-order Taylor-series expansion of (B.1) in $1/\chi_b^2$ around its expectation, $1/(b-2)$, that is, we consider $D_m$ as a function of $1/\chi_b^2$. Let

$$f(y) = c \cdot \frac{(1+a)}{1+\frac{a}{by}} \qquad \text{with} \quad y = \frac{1}{\chi_b^2} \quad \text{and} \quad c = \frac{\chi_k^2}{k}. \tag{B.2}$$

The first-order Taylor-expansion, $T_{f(y)}$, is

$$\begin{aligned} T_{f(y)} &= f(y)|_{y=1/(b-2)} + f'(y)|_{y=1/(b-2)} \cdot \left(y - \tfrac{1}{b-2}\right) \\ &= c \cdot \tfrac{1+a}{1+\frac{a}{b\cdot\frac{1}{b-2}}} + c \cdot \tfrac{ab(1+a)}{\left(\frac{b}{b-2}+a\right)^2} \cdot \left(y - \tfrac{1}{b-2}\right) \\ &= c \cdot \left(\tfrac{(1+a)\cdot b}{b+a\cdot(b-2)}\right) + c \cdot \left(\tfrac{ab(1+a)\cdot(b-2)^2}{(b+a\cdot(b-2))^2} \cdot \left(y - \tfrac{1}{b-2}\right)\right). \end{aligned} \tag{B.3}$$

To ascertain the mean and the variance of $T_{f(y)}$ we use the independency of $y = \frac{1}{\chi^2_b}$ and $c = \frac{\chi^2_k}{k}$ with $E(y) = 1/(b-2)$, $\mathrm{Var}(y) = 2/((b-2)^2(b-4))$, $E(c) = 1$ and $\mathrm{Var}(c) = 2/k$. Let $Z := T_{f(y)}$. Then it is

$$\begin{aligned} E(Z) &= E\left(c \cdot \tfrac{(1+a)\cdot b}{b+a\cdot(b-2)}\right) + \tfrac{ab(1+a)\cdot(b-2)^2}{(b+a\cdot(b-2))^2} \cdot E\left(c \cdot \left(y - \tfrac{1}{b-2}\right)\right) \\ &= \tfrac{(1+a)\cdot b}{b+a\cdot(b-2)} \cdot \underbrace{E(c)}_{=1} + \underbrace{\frac{ab(1+a)\cdot(b-2)^2}{(b+a\cdot(b-2))^2} \cdot E(c) \cdot \underbrace{\left(\underbrace{E(y)}_{=1/(b-2)} - \frac{1}{b-2}\right)}_{=0}}_{=0} \\ &= \tfrac{(1+a)b}{b+a(b-2)} = \tfrac{b+ab}{b+ab-2a}, \end{aligned} \tag{B.4}$$

and

$$\begin{aligned} E(Z^2) &= E(c^2) \cdot E\left(\left[\tfrac{(1+a)b}{b+a(b-2)} + \tfrac{ab(1+a)(b-2)^2}{(b+a(b-2))^2}\left(y - \tfrac{1}{b-2}\right)\right]^2\right) \\ &= E(c^2) \cdot \left(\left(\tfrac{(1+a)b}{b+a(b-2)}\right)^2 + \left(\tfrac{ab(1+a)(b-2)^2}{(b+a(b-2))^2}\right)^2 \cdot \left(E(y^2) - 2 \cdot \tfrac{1}{b-2}E(y) + \left(\tfrac{1}{b-2}\right)^2\right)\right) \\ &= \left(\tfrac{2}{k} + 1\right) \cdot \left[\left(\tfrac{(1+a)b}{b+a(b-2)}\right)^2 + \left(\tfrac{ab(1+a)(b-2)^2}{(b+a(b-2))^2}\right)^2 \cdot \left(\tfrac{1}{(b-2)(b-4)} - 2 \cdot \tfrac{1}{(b-2)^2} + \tfrac{1}{(b-2)^2}\right)\right] \\ &= \left(\tfrac{2}{k} + 1\right) \cdot \left[\tfrac{(1+a)^2\cdot b^2\cdot(b+a(b-2))^2\cdot(b-4)+2\cdot a^2b^2\cdot(1+a)^2\cdot(b-2)^2}{(b+a(b-2))^4(b-4)}\right]. \end{aligned} \tag{B.5}$$

Therefrom it follows

$$\begin{aligned} \mathrm{Var}(Z) &= E(Z^2) - (E(Z))^2 \\ &= \tfrac{(2+k)[(1+a)^2b^2(b+a(b-2))^2(b-4)+2a^2b^2(1+a)^2(b-2)^2]}{k\cdot(b+a(b-2))^4(b-4)} - \tfrac{b^2(1+a)^2\cdot k\cdot(b+a(b-2))^2(b-4)}{k\cdot(b+a(b-2))^4(b-4)} \\ &= \tfrac{2\cdot(1+a)^2b^2(b+a(b-2))^2(b-4)+4a^2b^2(1+a)^2(b-2)2+2ka^2b^2(1+a)^2(b-2)^2}{k\cdot(b+a(b-2))^4(b-4)} \\ &= \tfrac{2\cdot(1+a)^2b^2\cdot[(b+a(b-2))^2(b-4)+2a^2(b-2)^2+k\cdot a^2(b-2)^2]}{k\cdot(b+a(b-2))^4(b-4)} \\ &= \tfrac{2\cdot(1+a)^2b^2\cdot[(b+a(b-2))^2(b-4)+a^2(b-2)^2\cdot(2+k)]}{k\cdot(b+a(b-2))^4(b-4)}. \end{aligned} \tag{B.6}$$

From (B.4) and (B.6) the matching of the first two moments with the $\delta F_{k,w}$-distribution yields to the following system of equations with two equations and two variables of interest, $\delta$ and $w$:

$$E(Z) = \frac{b(1+a)}{b+a(b-2)} \stackrel{!}{=} \delta' \cdot \frac{w'}{w'-2}, \tag{B.7}$$

$$\begin{aligned} V(Z) &= \frac{2\cdot(1+a)^2b^2\cdot[(b+a(b-2))^2(b-4)+a^2(b-2)^2\cdot(2+k)]}{k\cdot(b+a(b-2))^4(b-4)} \\ &\stackrel{!}{=} (\delta')^2\cdot\frac{2(w')^2(k+w'-2}{k(w'-2)^2(w'-4)}. \end{aligned} \tag{B.8}$$

Obviously, solving (B.7) with respect to $\delta'$ gives

$$\delta' = \left(1-\frac{2}{w'}\right)\left(\frac{ab+b}{ab+b-2a}\right), \tag{B.9}$$

which is not equal to $\delta = (1-2/w)\cdot(ab+b-2)/(ab+b-2a-2)$ given in (3.6) by Li, Raghunathan, and Rubin (1991).

Applying $\delta'$ in (B.8) we get

$$\begin{aligned} \left(\tfrac{w'-2}{w'}\right)^2\cdot\left(\tfrac{b(1+a)}{b+a(b-2)}\right)^2\cdot\tfrac{2(w')^2(k+w'-2)}{k(w'-2)^2(w'-4)} &\stackrel{!}{=} \frac{2(1+a)^2b^2\cdot[\overbrace{(b+a(b-2))^2(b-4)}^{:=\zeta}+\overbrace{a^2(b-2)^2(k+2)}^{=:\eta}]}{k(b+a(b-2))^4(b-4)} \\ \Leftrightarrow\quad (k+w'-2)\zeta &= (\zeta+\eta)\cdot w' - 4\cdot(\zeta+\eta) \\ \Leftrightarrow\quad (k-2)\zeta + w'\cdot\zeta &= (\zeta+\eta)\cdot w' - 4\cdot(\zeta+\eta) \\ \Leftrightarrow\quad (k-2)\zeta + 4\cdot(\zeta+\eta) &= w'\cdot((\zeta+\eta)-\zeta) \end{aligned} \tag{B.10}$$

$$
\begin{aligned}
\Leftrightarrow \quad w' &= \tfrac{(k-2)\zeta+4\cdot(\zeta+\eta)}{(\zeta+\eta)-\zeta} \\
&= \tfrac{(k-2)(b+a(b-2))^2(b-4)+4\cdot(b+a(b-2))^2(b-4)+4a^2(b-2)^2(k+2)}{(b+a(b-2))^2(b-4)+a^2(b-2)^2(k+2)-(b+a(b-2))^2(b-4)} \\
&= \tfrac{(k+2)(b+a(b-2))^2(b-4)+4(k+2)a^2(b-2)^2}{(k+2)a^2(b-2)^2} \\
&= \tfrac{(b+a(b-2))^2(b-4)}{a^2(b-2)^2} + 4 \\
&= 4 + (b-4)\cdot\left(\tfrac{b^2}{a^2(b-2)^2} + \tfrac{2ab(b-2)}{a^2(b-2)^2} + \tfrac{a^2(b-2)^2}{a^2(b-2)^2}\right) \\
&= 4 + (b-4)\cdot\left(\tfrac{b^2}{a^2(b-2)^2} + \tfrac{2b}{a(b-2)} + 1\right) \\
&= 4 + (b-4)\cdot\left(1 + \tfrac{b}{a(b-2)}\right)^2,
\end{aligned}
$$

which also is not equal to $w = 4 + (b-4)\cdot\left(1+\frac{b-2}{ab}\right)^2$ given in (3.7) by Li, Raghunathan, and Rubin (1991).

# References


Barnard, J., Rubin, D.B. (1999), Small-sample degrees of freedom with Multiple Imputation, *Biometrika*, **86**, 948 - 955.

Box, G.E.P., Tiao, G.C. (1992), Bayesian inference in statistical analysis, *Wiley and Sons*, New York.

Cangul, M.Z., Chretien, Y.R., Gutman R., Rubin D.B. (2009), Testing treatment effects in unconfounded studies under model misspecification: Logistic regression, discretization, and their combination, *Statistics in Medicine*, **28**, 2531 - 2551.

Cox, D.R., Wermuth, N. (1996), Multivariate dependencies, *Chapman and Hall*, London.

Li, K.H., Meng, X.L., Raghunathan, T.E., Rubin, D.B. (1991), Significance levels from repeated p-values with multiply-imputed data, *Statistica Sinica*, **1**, 65 - 92.

Li, K.H., Raghunathan, T.E., Rubin, D.B. (1991), Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution, *Journal of the American Statistical Association*, **86**, 1065 - 1073.

Meng, X.L., Rubin, D.B. (1992), Performing likelihood ratio tests with multiply-imputed Data Sets, *Biometrika*, **79**, 103 - 111.

Rubin, D.B. (1977), Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys, *Journal of the American Statistical Association*, **72**, 538 - 543

Rubin, D.B. (1978), Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse, *American Statistical Association Proceedings of the Section on Survey Research Methods*, 20 - 40.

Rubin, D.B. (1987), Multiple imputation for nonresponse in surveys, *John Wiley and Son*, New York.

Rubin, D.B., Schenker, N. (1991), Multiple imputation in health-care databases: an overview and some applications, *Statistics in Medicine*, **10**, 585 - 598.