Identifying Missing Data Mechanisms in (2×2)–Contingency Tables

T. Nittner, H. Toutenburg

February 3, 2004

Abstract

Consider the sample of two binary variables X and Y with some missing structure within X or Y. The knowledge about the corresponding values of the observed covariate allows to play through all possible 'originally' complete data sets. After defining the notation, including some theoretical work, a test for non–MCAR within the complete case table is presented. Simulating all possible tables enables some testing on non–MAR. A simulation experiment is used to illustrate this context.

KEY WORDS: missing data mechanism, odds–ratio, simulation experiment, testing non–MAR.

1 Introduction

Binary variables are of large interest in many surveys and studies, e.g., when exploring the relation between smoking and lung cancer. For binary variables, the odds–ratio is an important measure to analyze the risk for response when one variable is supposed to be dependent of the other. Especially awkward questions concerning private habits or private status are affected by missing values, at least not missing completely at random (MCAR, see p. 4). The aim of this work is to show how the additional information of the incomplete cases can be used to run over all possible situations, to take a look at the change of the odds–ratio, to show whether the complete case table gives suspect for non–MCAR and to test for non–MAR within all possible completed data tables.

Consider a bivariate sample $(X_i, Y_i), i = 1, \ldots, n$, where both, X and Y, are of binary outcome '0' or '1'. Visualize the sample of size n in a (2×2)–contingency table, see Table 1.1. Using the well–known notation $n_{ij}, i, j = 1, 2$, the marginal frequencies are denoted according to

$$n_{i\cdot} = \sum_{j=1}^{2} n_{ij} \quad \text{and} \quad n_{\cdot j} = \sum_{i=1}^{2} n_{ij}\,. \tag{1.1}$$

Given the joint distribution $\{n_{ij}\}$ or $\{n_{ij}/n\}$, the odds–ratio θ can be estimated

	X	1	0	
	1	n_{11}	n_{12}	$n_{1\cdot}$
Y	0	n_{21}	n_{22}	$n_{2\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	n

Table 1.1: (2×2)–contingency table.

by

$$\hat{\theta} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}} \tag{1.2}$$

having values within $[0; \infty)$. The two variables are said to be independent for $\hat{\theta} = 1$. For $\hat{\theta} < 1$ response in the second row is more likely than in the first row. Of course, for $\hat{\theta} > 1$ response in the first row is more likely.

Instead of $\hat{\theta}$, often $\hat{\theta}_0 = \ln \hat{\theta}$ is considered in order to have a measure being symmetric with respect to zero. Following e.g. Agresti (1996), $\hat{\theta}_0$ is asymptotically normal distributed with mean θ_0 and standard deviation σ_{θ_0} with their estimates

$$\begin{aligned} \hat{\theta}_0 &= \ln\left(\frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}\right) & (1.3) \\ \hat{\sigma}_{\hat{\theta}_0} &= \left(\frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}}\right)^{\frac{1}{2}} . & (1.4) \end{aligned}$$

In case of independence of X and Y we have $\hat{\theta} = 1$ and, therefore, $\hat{\theta}_0 = \ln \hat{\theta} = 0$; for $-\infty < \hat{\theta}_0 < 0$ and $0 < \hat{\theta}_0 < \infty$ we have negative and positive correlation, respectively. The hypothesis H_0 : 'X and Y are independent' versus H_1 : 'X and Y are not independent' can be tested by computing the test–statistic z which is standard normally distributed under H_0 according to

$$z = \frac{\hat{\theta}_0}{\hat{\sigma}_{\hat{\theta}_0}} \sim N(0,1) . \tag{1.5}$$

Reject $\mathrm{H}_0 : \hat{\theta}_0 = 0$ if $\mid z \mid > z_{1-\alpha/2}$ (two–sided). This test decision corresponds to 0 not covered by the confidence interval

$$\left[\hat{\theta}_0 - z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{\theta}_0}; \hat{\theta} + z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{\theta}_0}\right] = [I_l; I_u] . \tag{1.6}$$

Interval (1.6) corresponds to a confidence interval for $\hat{\theta}$ itself according to

$$[\exp(I_l); \exp(I_u)] . \tag{1.7}$$

The introduction of the odds–ratio will be extended to the case of having missing data within the next section.

2 An Extension to Missing Data

2.1 The (2×2)–Contingency Table with Missing Data

First of all, we want to extend Table 1.1 to the situation of missing data in general. However, the situation is restricted to samples where at least cases of one variable are observed. Assume that the number of incomplete cases is known for each marginal frequency and denote them by l, o, m and k. The new situation can be illustrated as shown in Table 2.1.

	X	1	0	observed	additionally observed	
Y	1	n_{11}	n_{12}	$n_{1\cdot}$	l	$\tilde{n}_{1\cdot}$
	0	n_{21}	n_{22}	$n_{2\cdot}$	o	$\tilde{n}_{2\cdot}$
observed		$n_{\cdot 1}$	$n_{\cdot 2}$	n		
additionally observed		m	k			
		$\tilde{n}_{\cdot 1}$	$\tilde{n}_{\cdot 2}$			$\tilde{n}$

Table 2.1: (2×2)–contingency table with missing data.

A value of $l = 4$ and $o = 7$, for example, means that X is missing for 11 values, four cases with $Y = 1$ and seven cases with $Y = 0$. Define the sample size of all the observed values by

$$\tilde{n} = n + l + o + m + k\,.$$

Assuming further that $l, o, m, k \in \mathbb{N}_0$ allows to illustrate each of all missing data situations where at least X (or Y) is observed. One of the simplest cases refers to $l = o = k = 0$ and $m > 0$ which corresponds to the case where Y is missing for m observed cases having $X = 1$, see Figure 2.1. If Y is missing for any values of X, both, $m > 0$ and $k > 0$ and $l = o = 0$; this case is shown in Figure 2.2.

The more complicated case where X and Y both are incomplete, but not for the same case indices, corresponds to $l, o, m, k > 0$ and is illustrated in Figure 2.3.

Figures 2.1–2.3 are called missing data pattern (MDP), see e.g. Little and Rubin (1987), (2002), and may give a first impression of the extent of incompleteness. Another important tool to characterize the problem of missing data is the so–called missing data mechanism (MDM) also going back to Little and Rubin (1987), (2002). The MDMs are defined within the next section.

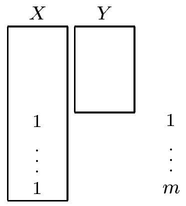

Figure 2.1: $m > 0$

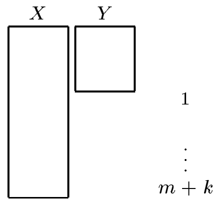

Figure 2.2: $m, k > 0$

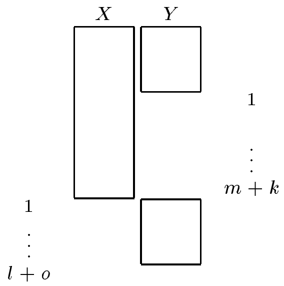

Figure 2.3: $l, o, m, k > 0, X$ and Y mutually observed.

2.2 The Missing Data Mechanism (MDM)

As denoted before, the MDM can be used to characterize dependencies between observed and missing data. In general one usually differs between missing completely at random (MCAR), missing at random (MAR) and non–missing at random (non–MAR). Let $Z = (X, Y) = (Z_{\text{obs}}, Z_{\text{mis}}), Z_{\text{obs}} = (X_{\text{obs}}, Y_{\text{obs}}), Z_{\text{mis}} = (X_{\text{mis}}, Y_{\text{mis}})$ denote the data matrix that would have occurred without any missing data. Further let $R = (r_{ij})$ be a matrix defined by

$$r_{ij} = \begin{cases} 1 & \text{if} \quad z_{ij} \quad \text{observed} \\ 0 & \text{if} \quad z_{ij} \quad \text{missing} \end{cases} \quad \forall\, i = 1, \ldots, n,\ j = 1, 2. \tag{2.1}$$

The so–called indicator matrix R indicates whether a value is observed or missing. In the figurative sense, the problem of missing data here is a random experiment because R and its elements, respectively, are random variables. Considering the conditional density $f(R \mid Z_{\text{obs}}, Z_{\text{mis}}, \Phi)$ with Φ being the parameter of the missing mechanism allows the distinction between the three missing data mechanisms according to

1. MCAR (missing completely at random), if

$$f(R \mid Z, \Phi) = f(R \mid \Phi) \quad \forall Z \tag{2.2}$$

2. MAR (missing at random), if

$$f(R \mid Z, \Phi) = f(R \mid Z_{\text{obs}}, \Phi) \quad \forall Z_{\text{mis}} \qquad \text{and} \tag{2.3}$$

3. non–MAR (non–missing at random) if

$$f(R \mid Z, \Phi) = f(R \mid Z_{\text{obs}}, Z_{\text{mis}}, \Phi) \tag{2.4}$$

whereas (2.4) cannot be simplified, i.e. R has to depend at least on Z_{mis}. If, for example, X is affected by missing data,

1. MCAR means that the missingness within X is a random subsample of X,
2. MAR means that $f(R \mid Z, \Phi) = f(R \mid Y, \Phi)$, i.e. missingness within X depends on the values of Y, and,
3. non–MAR means that $f(R \mid Z, \Phi) = f(R \mid X, Y, \Phi)$, i.e. missingness within X at least depends on the values of X itself.

In case of $(2{\times}2)$–contingency tables of course R is of dimension $(n \times 2)$ and the patterns 2.1 and 2.3 correspond to

$$R^{(2.1)} = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ & \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ & \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix} \qquad R^{(2.3)} = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} .$$

2.3 Parameterizing the Missing Data

In order to be able to include the additional information, the parameters of the complete case table have to be combined with the four parameters characterizing the missing data situation. Altogether, $(m+1)\cdot(k+1)\cdot(l+1)\cdot(o+1)$ possibilities exist to 'distribute' the missing values to the corresponding cells yielding the same number of 'completed' tables, denoted by their frequencies $\tilde{n}_{ij}$. Let us further denote the indices by z, w, i and j,

$$\begin{aligned} z &= 0, \ldots, l, \\ w &= 0, \ldots, o, \\ i &= 0, \ldots, m, \\ j &= 0, \ldots, k, . \end{aligned}$$

The main idea of including the missing data information consists of using growth factors for each of the complete case frequencies. Define the growth factors according to

$$\begin{aligned} a &= \frac{n_{11} + z + i}{n_{11}} & c &= \frac{n_{12} + (l-z) + (k-j)}{n_{12}} \\ b &= \frac{n_{21} + (o-w) + (m-i)}{n_{21}} & d &= \frac{n_{22} + w + j}{n_{22}} \quad . \end{aligned} \tag{2.5}$$

These growth factors could also be written according to

$$a = \frac{\tilde{n}_{11}}{n_{11}}, c = \frac{\tilde{n}_{12}}{n_{12}}, b = \frac{\tilde{n}_{21}}{n_{21}}, d = \frac{\tilde{n}_{22}}{n_{22}} . \tag{2.6}$$

With the help of the four growth factors a, b, c and d the odds–ratio of the filled–up table can easily be denoted by

$$\begin{aligned}
\widehat{\tilde{\theta}} &= \frac{\tilde{n}_{11} \cdot \tilde{n}_{22}}{\tilde{n}_{12} \cdot \tilde{n}_{21}} \\
&= \frac{a \cdot n_{11} \cdot d \cdot n_{22}}{c \cdot n_{12} \cdot b \cdot n_{21}} \\
&= \frac{a \cdot d}{c \cdot b} \cdot \hat{\theta} \,. \qquad (2.7)
\end{aligned}$$

Both, from a practical as well as from a theoretical point of view, it may be of interest to consider the norm of the relative change in $\hat{\theta}$, i.e.,

$$\begin{aligned}
\frac{\mid \widehat{\tilde{\theta}} - \hat{\theta} \mid}{\hat{\theta}} &= \frac{\mid \hat{\theta}(\delta - 1) \mid}{\hat{\theta}} \\
&= \frac{\hat{\theta} \mid \delta - 1 \mid}{\hat{\theta}} \\
&= \mid \delta - 1 \mid, \qquad (2.8)
\end{aligned}$$

with $\delta = (a \cdot d)/(c \cdot b)$. That is, the norm of the relative change of $\hat{\theta}$ can be restricted by considering the difference between the ratio of the growth factors and 1. The norm of the change in $\hat{\theta}$ simply follows

$$\mid \widehat{\tilde{\theta}} - \hat{\theta} \mid = \hat{\theta} \mid \delta - 1 \mid \ . \qquad (2.9)$$

Therefore, we are especially interested in the minimum and the maximum of (2.9). Let us denote these as $\Delta_{\min}$ and $\Delta_{\max}$, i.e.,

$$\hat{\theta} \mid \delta - 1 \mid \in \ [\Delta_{\min}; \Delta_{\max}] \,. \qquad (2.10)$$

Of course, the maximum of $\widehat{\tilde{\theta}}$ corresponds to ‘putting’ all values on the main diagonal and the minimum to ‘putting’ all values on the secondary diagonal. However, in the following, let’s consider the three possible cases for (2.9) in general.

Case 1: $\widehat{\tilde{\theta}} - \hat{\theta} \geq 0$ holds $\forall \, z, w, i, j$.

This corresponds to

$$\widehat{\tilde{\theta}} \geq \hat{\theta} \quad \forall \, z, w, i, j \,.$$

Therefore, also the minimum and the maximum value of $\widehat{\tilde{\theta}}$ is larger than $\hat{\theta}$, i.e.,

$$\begin{aligned}
\min_{z,w,i,j} \widehat{\tilde{\theta}} \geq \hat{\theta} \quad &\Longrightarrow \quad \Delta_{\min} = \min_{z,w,i,j} \widehat{\tilde{\theta}}, \\
\max_{z,w,i,j} \widehat{\tilde{\theta}} \geq \hat{\theta} \quad &\Longrightarrow \quad \Delta_{\max} = \max_{z,w,i,j} \widehat{\tilde{\theta}} \,.
\end{aligned}$$

Case 2: $\hat{\tilde{\theta}} - \hat{\theta} < 0$ holds $\forall\, z, w, i, j$.

This corresponds to
$$\hat{\tilde{\theta}} < \hat{\theta} \quad \forall\, z, w, i, j\,.$$

As before, we conclude that

$$\begin{aligned} \min_{z,w,i,j} \hat{\tilde{\theta}} < \hat{\theta} \quad &\Longrightarrow \quad \Delta_{\min} = \max_{z,w,i,j} \hat{\tilde{\theta}}, \\ \max_{z,w,i,j} \hat{\tilde{\theta}} < \hat{\theta} \quad &\Longrightarrow \quad \Delta_{\max} = \min_{z,w,i,j} \hat{\tilde{\theta}}\,. \end{aligned}$$

As $\delta \in (0; \infty)$ and $a \geq 1, b \geq 1, c \geq 1, d \geq 1$, δ gets minimum for $a = d = 1$ and maximum for $c = b = 1$. For the minimum of δ, meaning $\delta < 1$ we have

$$\min_{z,w,i,j} \hat{\tilde{\theta}} < \hat{\theta} \quad \text{for at least one } z, w, i, j\,, \tag{2.11}$$

and for the maximum of δ, i.e. $\delta > 1$,

$$\max_{z,w,i,j} \hat{\tilde{\theta}} \geq \hat{\theta} \quad \text{for at least one } z, w, i, j\,. \tag{2.12}$$

Equations (2.11) and (2.12) mean that

$$\min_{z,w,i,j} \hat{\tilde{\theta}} < \hat{\theta} < \max_{z,w,i,j} \hat{\tilde{\theta}}\,, \tag{2.13}$$

i.e. there is no missing structure for which each simulated table yields odds–ratios fulfilling Case 1 or Case 2. Cases 1 and 2 therefore never hold and the main focus will be on Case 3.

Case 3: There is at least one setting $\{z, w, i, j\}$ for $\hat{\tilde{\theta}} - \hat{\theta} < 0$ and there is at least one different setting $\{\tilde{z}, \tilde{w}, \tilde{i}, \tilde{j}\} \neq \{z, w, i, j\}$ for which $\hat{\tilde{\theta}} - \hat{\theta} > 0$ holds.

Thus, $\min_{z,w,i,j} \hat{\tilde{\theta}} < \hat{\theta}$ and $\max_{z,w,i,j} \hat{\tilde{\theta}} > \hat{\theta}$. As δ in (2.8) can't obtain $\delta = 1$ for every simulated setting, the minimum is the δ nearest to 1. The maximum is $\min_{z,w,i,j} \hat{\tilde{\theta}}$ or $\max_{z,w,i,j} \hat{\tilde{\theta}}$—depending on the absolute distance with respect to $\hat{\tilde{\theta}}$.

Let us consider the simplest case illustrated in Figure 2.1 where Y is missing for $X = 1$, meaning $k = l = o = 0$, and, therefore, $c = d = 1$. We see that

$$\hat{\tilde{\theta}} = \frac{a \cdot n_{11} \cdot n_{22}}{n_{12} \cdot b \cdot n_{21}} = \frac{a}{b} \cdot \hat{\theta}\,.$$

For a and b we have for $i = 0, \ldots, m$,

$$a = \frac{n_{11} + i}{n_{11}} \qquad\qquad b = \frac{n_{21} + m - i}{n_{21}}\,. \tag{2.14}$$

$i=0$			$i=1$		
$\lvert \widehat{\tilde{\theta}} - \hat{\theta} \rvert$	$=$	$\frac{n_{11}\cdot n_{22}}{n_{12}\cdot n_{21}} \cdot \left\lvert \frac{n_{21}}{n_{21}+m} - 1 \right\rvert$	$\lvert \widehat{\tilde{\theta}} - \hat{\theta} \rvert$	$=$	$\frac{n_{11}\cdot n_{22}}{n_{12}\cdot n_{21}} \cdot \left\lvert \frac{n_{11}+m}{n_{11}} - 1 \right\rvert$
	$=$	$\frac{n_{11}\cdot n_{22}}{n_{12}\cdot n_{21}} \cdot \left\lvert \frac{n_{21}-n_{21}-m}{n_{21}+m} \right\rvert$		$=$	$\frac{n_{11}\cdot n_{22}}{n_{12}\cdot n_{21}} \cdot \left\lvert \frac{n_{11}+m-n_{11}}{n_{11}} \right\rvert$
	$=$	$\frac{m\cdot n_{11}\cdot n_{22}}{n_{21}\cdot n_{12}\cdot (n_{21}+m)}$		$=$	$\frac{m\cdot n_{22}}{n_{12}\cdot n_{21}}$
	$:=$	γ_1		$:=$	γ_2 .

Table 2.2: Minimum and maximum vakues of δ.

Of course, $\delta = a/b$ is minimum for $i = 0$ and maximum for $i = m$. Let us deduce a condition for the cells of Case 3, i.e. whether the minimum or the maximum of $\widehat{\tilde{\theta}}$ is farther from the complete case estimate. See Table 2.2 for preparing the solution.

The question is for which values γ_1 is larger or equal to γ_2, i.e. when $\min_{\hat{\theta}} - \hat{\theta}$ is larger than $\max_{\hat{\theta}} - \hat{\theta}$ and vice versa. Thus, consider its ratio and compute

$$
\begin{aligned}
\frac{\gamma_1}{\gamma_2} &= \frac{m \cdot n_{11} \cdot n_{22}}{n_{21} \cdot n_{12} \cdot (n_{21}+m)} \cdot \frac{n_{12} \cdot n_{21}}{m \cdot n_{22}} \\
&= \frac{n_{11}}{n_{21}+m} \, . \qquad (2.15)
\end{aligned}
$$

Using (2.15) we get the following conditions for the cell frequencies,

$$
\gamma_1 \geq \gamma_2 \quad \Leftrightarrow \quad n_{11} \geq n_{21} + m \, , \qquad (2.16)
$$

$$
\gamma_1 < \gamma_2 \quad \Leftrightarrow \quad n_{11} < n_{21} + m \, . \qquad (2.17)
$$

Therefore, $\min_{\hat{\theta}} > \max_{\hat{\theta}}$ for $n_{11} \geq n_{21} + m$; the common distribution of the variables of the imputed table is equivalent or nearly equivalent to the common distribution of the variables of the observed table when $\delta = 1$ or δ is near to 1, i.e. when $a = b$ or a is near b. For checking where $\hat{\theta}$ is nearer to—the maximum or the minimum of the imputed tables—it is just necessary to look at condition (2.16).

For a more complex missing data structure, i.e., $l, o, k \neq 0$ such a condition is not easy to deduce because of the larger number of parameters. But note that it is sufficient to look at the δ nearest to 1 to get the minimum of (2.9).

Within the next section it is shown how to test for non–MCAR. Further, some testing for non–MAR is introduced.

3 Testing on the Missing Data

The main idea of testing on the missing data is based on Cohen and Cohen (1983): The cases of a complete co–variable are split into an observed part ($R = 1$) and a missing part ($R = 0$) according to the incomplete variable, building one population and one sample for a standard testing procedure.

3.1 Testing for non–MCAR

Before introducing the test procedure let's denote the situation where X is missing for $l + o$ cases. Thus, the indicator variable R_X is given by

$$R_X = (\underbrace{1, \ldots, 1}_{1,\ldots,n-(l+o)}, \underbrace{0, \ldots, 0}_{n-(l+o)+1,\ldots,n})' . \tag{3.1}$$

Based on this indicator variable we are able to separate Y into two groups. Suppose that the data can be rearranged according to Figure 2.2, i.e., the pairs of values of X and Y are not lost and one index of both of them is an identifier for the case. Before joining the theory of the binomial–test with diagnosis on non–MCAR some remarks on the binomial test.

Assume a random variable Z having the two values '1' and '0', representing incidence and non–incidence of an event. The probability p may denote the probability for incidence in the population. From a sample

$$Z = (Z_1, \ldots, Z_n) \quad \text{with} \quad Z_i \overset{\text{iid}}{\sim} B(1; p) ,$$

we are able to estimate $p = P(Z = 1)$ by the unbiased estimate $\hat{p} = \sum_{i=1}^{n} Z_i$ (see Toutenburg (2000)). The object is to test $\mathrm{H}_0 : p = p_0$ versus the alternative $\mathrm{H}_1 : p \neq p_0$. Under H_0 it holds that

$$\mathrm{Var}(\hat{p}) = \frac{p_0 \cdot (1 - p_0)}{n} . \tag{3.2}$$

Standardizing gives the test statistic

$$T(Z) = \frac{\hat{p} - p_0}{\sqrt{p_0 \cdot (1 - p_0)}} \cdot \sqrt{n} , \tag{3.3}$$

where the binomial distribution can be approximated by the normal distribution if $n \cdot p_0 \cdot (1 - p_0) > 9$ holds. For a fixed level of significance, α,

$$P_{p_0}(T(X) < k_l) \leq \alpha/2 \qquad\qquad P_{p_0}(T(X) \geq k_u) \leq \alpha/2 , \tag{3.4}$$

with a lower bound k_l and an upper bound k_u, has to hold for H_0. Let's go back to the actual problem here: X is missing for $l + o$ values where Y is observed. If the probability for $Y = 1$ differs for $R_X = 1$ and $R_X = 0$ there would be suspect for X missing depending on the values of Y. Considering $P(Y = 1 \mid R_X = 1)$ as the probability of the population and denoting $\hat{p} = P(Y = 1 \mid R_X = 0)$ yields

$$\begin{aligned} T(Y) &= \frac{P(Y = 1 \mid R_X = 0) - P(Y = 1 \mid R_X = 1)}{\sqrt{P(Y = 1 \mid R_X = 1) \cdot (1 - P(Y = 1 \mid R_X = 1)}} \cdot \sqrt{l + o} \\ &= \frac{\frac{l}{l+o} - \frac{n_{1\cdot}}{n}}{\sqrt{\frac{n_{1\cdot}}{n} \cdot (1 - \frac{n_{1\cdot}}{n})}} \cdot \sqrt{l + o} . \end{aligned} \tag{3.5}$$

It is called $T(Y)$ because it depends on the distribution of Y. Further, the test statistic just depends on given values and therefore can easily be computed. Considering the case where Y is incomplete the corresponding test statistic $T(X)$ can be computed by using $\hat{p} = P(X = 1 \mid R_Y = 0) = m/(m + k)$ and the probability to test for, $p_0 = P(X = 1 \mid R_Y = 1) = n_{\cdot 1}/n$; the corresponding sample size is $m + k$.

In case of rejecting H_0 there might be suspect for the incomplete variable not to be affected by missing completely at random. At least, the missingness depends on the values of the covariate.

Apart from some diagnosis on the complete case table some more diagnostics concerning all filled–up tables may be of interest.

3.2 Some Testing for non–MAR

In fact, it cannot be tested for non–MAR that's why the title of this section is called 'some testing'. However, simulating all complete data sets—which is possible when considering binary or categorical data—enables to test for non–MAR.

Again, the idea is based on the binomial–test which here is based on the (marginal) distribution of the incomplete variable itself. The test compares the probability for $Y = 1$ ($X = 1$) based on the complete cases, i.e. $R_X = 1$ ($R_Y = 1$), with the probability for $X = 1$ ($Y = 1$) for all filled–up data, i.e., $R_X = 0$ ($R_Y = 0$). If the probabilities differ too much—in the sense of some confidence—, there's suspect for the missing values to miss at least depending on the values of X (Y) itself.

Based on the situation described in Section 3.1 with its basic test statistic (3.3) the question whether X is missing according to non–MAR can be answered statistically. Here, the probabilities $P(X = 1 \mid R_X = 1)$ and $P(X = 1 \mid R_X = 0)$ have to be compared. Use

$$\begin{aligned} \hat{p} &= P(X = 1 \mid R_X = 0) = \frac{\tilde{n}_{\cdot 1} - n_{\cdot 1}}{l + o + m}, \quad \text{and,} && (3.6) \\ p_0 &= P(X = 1 \mid R_X = 1) = \frac{n_{\cdot 1}}{n} && (3.7) \end{aligned}$$

to compute the test statistic

$$T(X) = \frac{\frac{\tilde{n}_{\cdot 1} - n_{\cdot 1}}{l+o+m} - \frac{n_{\cdot 1}}{n}}{\sqrt{\frac{n_{\cdot 1}}{n} \cdot (1 - \frac{n_{\cdot 1}}{n})}} \cdot \sqrt{l + o + m}\,. \qquad (3.8)$$

When testing for non–MAR within Y (3.6) and (3.7) have to be modified acoording to

$$\begin{aligned} \hat{p} &= P(Y = 1 \mid R_Y = 0) = \frac{\tilde{n}_{1 \cdot} - n_{1 \cdot}}{l + m + k}, \quad \text{and,} && (3.9) \\ p_0 &= P(Y = 1 \mid R_Y = 1) = \frac{n_{1 \cdot}}{n}, && (3.10) \end{aligned}$$

and the corresponding test statistic follows

$$T(Y) = \frac{\frac{\tilde{n}_{1\cdot} - n_{1\cdot}}{l+m+k} - \frac{n_{1\cdot}}{n}}{\sqrt{\frac{n_{1\cdot}}{n} \cdot (1 - \frac{n_{1\cdot}}{n})}} \cdot \sqrt{l+m+k}\,. \tag{3.11}$$

Though the two test statistics (3.8) and (3.11) may also easily be computed the probabilities $P(X = 1 \mid R_X = 0)$ and $P(Y = 1 \mid R_Y = 0)$ are not known; they can be computed by simulating all common distributions of X and Y, based on their frequencies $\{\tilde{n}_{ij}\}$.

4 A Simulation Experiment

Within this section the simulation experiment used to deal with the test problems is described. The program was realized by using R programming language, version 1.8.0.

4.1 Some Details

When the program is started via command line the user has to indicate the frequencies of the complete case table, i.e. n_{11}, n_{12}, n_{21} and n_{22} as well as the additional information about the incomplete cases, i.e. l, o, m and k. After some consistency checks, e.g. $0 < n_{ij} > 5000 \; \forall \; i, j, n < 10000$ (to reduce the sample size and the computing time) or $n_{1\cdot} < l$ (don't try to copy the oracle of Delphi), the simulation is started. First of all, the MCAR–diagnosis is achieved by testing the complete case table. Some descriptive statistics are also computed for the complete case table, e.g., the estimated odds–ratio, its estimated variance, the z–test for independency and the confidence interval for the odds–ratio. After the analysis of the complete case table, the actual simulation starts. Four loops are used to compute the 'virtual' frequencies $\tilde{n}_{ij}$. For each of these tables again the descriptive statistics are computed as well as the p–values for the tests on non–MAR. The results contain

- a missing data pattern for the indicated frequencies of the complete case table,
- the p–values of the tests for non–MCAR,
- descriptive statistics (minimum and maximum sample odds–ratio, minimum and maximum sample standard deviation, shortest and longest confidence interval for the estimated odds–ratio) for all distribution $\{\tilde{n}_{ij}\}$, and,
- the p–values of the tests for non–MAR, illustrated by graphics (in case of normal approximation).

The next section describes the simulation experiment and discusses some results.

4.2 An Example

Within this section a simple example is used to illustrate the tests and the main aspects mentioned so far. Table 4.1 resumes the complete case table of the experiment and contains the frequencies as they were introduced in Table 2.1.

		X = 1	X = 0		
Y	1	257	181	438	23
	0	245	337	582	15
		502	518	1020	
		31	8		1097

Table 4.1: Table for X and Y including information about the incomplete cases.

Altogether, 77 cases are incomplete which corresponds to a missing percentage of 7,02%. See Figure 4.1 where the corresponding missing data pattern is illustrated.

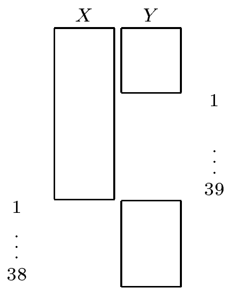

Figure 4.1: $l = 23, o = 15, m = 31, k = 8$; X and Y mutually observed.

As X and Y both are incomplete, one main interest is the question whether X and Y are missing according to MCAR. We have $l + o = 38 > 0$, so we are able to test for non–MCAR for the incomplete variable X. Referring to the actual problem here, $T(Y)$ follows

$$T(Y) = \frac{\frac{l}{l+o} - \frac{n_{1\cdot}}{n}}{\sqrt{\frac{n_{1\cdot}}{n} \cdot (1 - \frac{n_{1\cdot}}{n})}} \cdot \sqrt{l+o} \tag{4.1}$$

$$= \frac{0.605 - 0.429}{\sqrt{0.429 \cdot (1 - 0.429)}} \cdot \sqrt{38} = 2.19\,. \tag{4.2}$$

The corresponding p–value for H_0 : 'X is missing completely at random' for the given data is 0.0285 and therefore the hypothesis that X is missing according to MCAR has to be rejected for $\alpha = 0.05$. Note that $n \cdot p_0 \cdot (1 - p_0) = (l + o) \cdot \frac{n_{1\cdot}}{n} \cdot (1 - \frac{n_{1\cdot}}{n}) = 9.3$, so we used the normal approximation.

The test statistic for non–MCAR for the incomplete variable Y follows

$$
\begin{aligned}
T(X) &= \frac{P(X = 1 \mid R_Y = 0) - P(X = 1 \mid R_Y = 1)}{\sqrt{P(X = 1 \mid R_Y = 1) \cdot (1 - P(X = 1 \mid R_Y = 1))}} \cdot \sqrt{m + k} \\
&= \frac{\frac{m}{m+k} - \frac{n_{\cdot 1}}{n}}{\sqrt{\frac{n_{\cdot 1}}{n} \cdot (1 - \frac{n_{\cdot 1}}{n})}} \cdot \sqrt{m + k} \qquad (4.3)
\end{aligned}
$$

Here, the test statistic is computed according to

$$
\begin{aligned}
T(X) &= \frac{\frac{m}{m+k} - \frac{n_{\cdot 1}}{n}}{\sqrt{\frac{n_{\cdot 1}}{n} \cdot (1 - \frac{n_{\cdot 1}}{n})}} \cdot \sqrt{m + k} \qquad (4.4) \\
&= \frac{0.795 - 0.492}{\sqrt{0.492 \cdot (1 - 0.492)}} \cdot \sqrt{39} = 3.78\,. \qquad (4.5)
\end{aligned}
$$

The corresponding p–value for H_0 : 'Y is missing completely at random' for the given data is 0.00016 so also Y is not supposed to be MCAR. Note that $n \cdot p_0 \cdot (1 - p_0) = (m + k) \cdot \frac{n_{\cdot 1}}{n} \cdot (1 - \frac{n_{\cdot 1}}{n}) = 9.7$, so we here also used the normal approximation.

Before analyzing the simulated tables with respect to non–MAR, the descriptive statistics are resumed in Table 4.2.

	$\hat{\tilde{\theta}}$	$\log(\hat{\tilde{\theta}})$	$\hat{\sigma}(\log(\hat{\tilde{\theta}}))$	p–value	I_l	I_u	$\tilde{n}_{11}$	$\tilde{n}_{12}$	$\tilde{n}_{21}$	$\tilde{n}_{22}$	$\tilde{n}$
$\hat{\tilde{\theta}}_{\min}$	**1.4**	0.34	0.12	0.0056	1.1	1.8	257	212	291	337	1097
$\hat{\tilde{\theta}}_{\max}$	**2.5**	0.93	0.12	1.2e-13	2	3.2	311	181	245	360	1097
$\hat{\sigma}_{\min}(\log(\hat{\tilde{\theta}}))$	1.6	0.46	**0.12**	0.00015	1.3	2	274	212	274	337	1097
$\hat{\sigma}_{\max}(\log(\hat{\tilde{\theta}}))$	2.5	0.93	**0.12**	1.2e-13	2	3.2	311	181	245	360	1097
min CI	1.4	0.34	0.12	0.0056	**1.1**	**1.8**	257	212	291	337	1097
max CI	2.5	0.93	0.12	1.2e-13	**2**	**3.2**	311	181	245	360	1097

Table 4.2: Minimum/maximum values of essential terms for the simulated data.

The odds–ratio of the complete case table is about 1.95. For the minimum and for the maximum odds–ratio of the simulated tables, the hypothesis for independency of X and Y has to be rejected. The estimated standard deviations of the sample odds-ratios are nearly constant so the z–statistic just varies depending on the estimates of θ itself. Therefore, independency has to be rejected for all tables because it is rejected for the maximum estimate of θ.

But let's take a look at the tests for non–MAR now. Instead of using descriptive statistics to analyze the different simulation experiments a graphic is used to illustrate the different results.

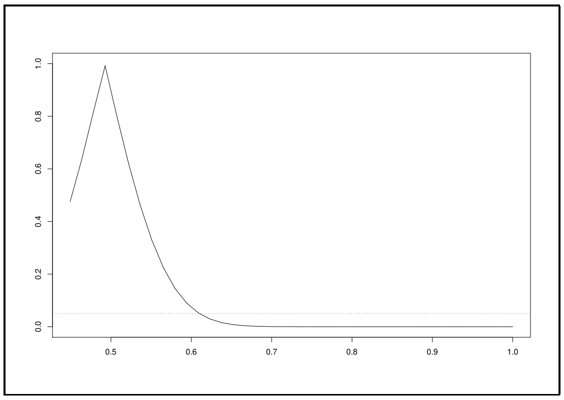

Figure 4.2: Plot of $P(X = 1 \mid R_X = 0)$ (x–axis) against the p–values (y–axis) for testing non–MAR, i.e. $\mathrm{H}_0 : P(X = 1 \mid R_X = 0) = P(X = 1 \mid R_X = 1)$.

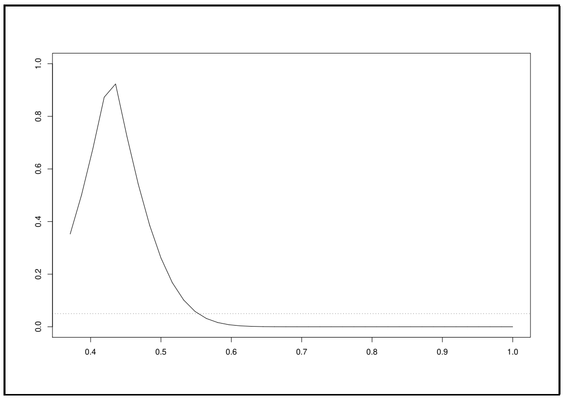

Figure 4.3: Plot of $P(Y = 1 \mid R_Y = 0)$ (x–axis) against the p–values (y–axis) for testing non–MAR, i.e. $\mathrm{H}_0 : P(Y = 1 \mid R_Y = 0) = P(Y = 1 \mid R_Y = 1)$.

Figure 4.2 plots the p–value for the test of non–MAR within X against $P(X = 1 \mid R_X = 0)$. We see that we have to reject MAR when $P(X = 1 \mid R_X = 0)$ is about to be larger than 0.6; for about $P(X = 1 \mid R_X = 0) \leq 0.6$, X seems to miss depending on the values of Y. The p–value gets maximum for about 0.5—the probability for $P(X = 1 \mid R_X = 1) = 0.49$. A similar situation could be seen by considering Figure 4.3 where the p–value for H_0 : 'Y is missing according to MAR' is plotted against $P(Y = 1 \mid R_Y = 0)$. Here, we have to reject MAR when $P(Y = 1 \mid R_Y = 0)$ is larger than about 0.55. The maximum p–value of course is near or equal to $P(Y = 1 \mid R_Y = 1) = 0.43$.

4.3 Concluding Remarks

It is planned to post the program on the internet, so please watch

`www.stat.uni-muenchen.de/~nittner/program`

When the link exists, the program and all its modules could be downloaded. Note that automatically a LaTeX–report will be generated containing basic information about the study you made. The report is about 7–8 pages and contains the complete case table, the table including the missing data information, the missing data pattern, the test statistics and results for testing on non–MCAR as well as the descriptive statistics and the test results for independency for the possible distributions $\{\tilde{n}_{ij}\}$. For studies using the normal approximation within the binomial test also the corresponding graphics are included. Note that this report will correspond to the missing data problem you had. For printing, it is just necessary to compile the source code and to generate a postscript or a pdf–file. Additionally a logfile is generated and could be expanded by setting the debugger. An exact description will be on the homepage.

The program is thought to be useful for researchers analyzing (2×2)–contingency tables affected by missing data. Additionally, reserachers being capable of R programming language easily can extend the existing program. The logfile contains some more information (when setting the debugger) which also could additionally be used. Tools for doing diagnostics on missing data are rare and some more work has to be done on this area.

References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, Wiley, New York.

Cohen, J. and Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Lawrence Erlbaum, Hillsdale, NJ.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2 edn, Wiley, New York.

Rüger, B. (1996). *Induktive Statistik: Einführung für Wirtschafts- und Sozialwissenschaftler*, Oldenbourg, München.

Toutenburg, H. (1992). *Moderne nichtparametrische Verfahren der Risikoanalyse*, Physica, Heidelberg.

Toutenburg, H. (2000). *Induktive Statistik*, 2 edn, Springer–Verlag, Heidelberg.