A Sensitivity Analysis of a Nonignorable Nonresponse Model Via EM Algorithm and Bootstrap

By

Yujie Zong

A Master Thesis

Submitted to the Faculty

Of

Worcester Polytechnic Institute

In partial fulfillment of the requirement for the

Degree of Master of Science

In

Applied Statistics

May 2011

APPROVED:

Dr. Balgobin Nandram, Thesis Advisor

Acknowledgement

I would like to gratefully and sincerely thank Professor Balgobin Nandram for his guidance, understanding, patience, and most importantly, his friendship during my graduate studies at WPI. His mentorship provided me with a well rounded experience consistent with my long-term development. He encouraged me to develop myself not only as a statistician but also as an independent thinker. Thanks for everything, Professor Nandram.

My thanks also go to Dr. Dhiman Bhadra for reading previous drafts of this thesis and providing many valuable comments that improved the presentation and contents of this thesis.

Moreover, I would also like to thank my wife, Haolian Zong for her understanding and love during the past few years. Her support and encouragement was in the end what made this thesis possible. I cannot imagine life without her.

Contents

Abstract

The Slovenian Public Opinion survey (SPOS), which carried out in 1990, was used by the government of Slovenia as a benchmark to prepare for an upcoming plebiscite, which asked the respondents whether they support independence from Yugoslavia. However, the sample size was large and it is quite likely that the respondents and nonrespondents had divergent viewpoints. We first develop an ignorable nonresponse model which is an extension of a bivariate binomial model. In order to accommodate the nonrespondents, we then develop a nonignorable nonresponse model which is an extension of the ignorable model. Our methodology uses an EM algorithm to fit both the ignorable and nonignorable nonresponse models, and estimation is carried out using the bootstrap mechanism. We also perform sensitivity analysis to study different degrees of departures of the nonignorable nonresponse model from the ignorable nonresponse model. We found that the nonignorable nonresponse model is mildly sensitive to departures from the ignorable nonresponse model. In fact, our finding based on the nonignorable model is better than an earlier conclusion about another nonignorable nonresponse model fitted to these data.

Keywords: *Bivariate binomial distribution; Bootstrap; EM algorithm; Missing not at random; Multinomial model; 2X2 categorical tables.*

Chapter 1. Introduction

1.1 Brief Review of Dataset

The Republic of Slovenia is a country in Central Europe and its capital is Ljubljana. In 1991, Slovenians voted for independence from former Yugoslavia in a plebiscite. To prepare for this plebiscite, the Slovenian Government collected data through the Slovenian Public Opinion survey (SPOS), a month before the plebiscite.

The three main questions that were asked in the plebiscite are as follows:

(a) Will you attend the plebiscite?

(b) Are you in favor of Slovenia's secession from Yugoslavia?

(c) Are you in favor of Slovenian independence?

The answers were recorded as Yes, No, or Don't Know (DK). The results of the survey relating to these three questions are recorded in Table 1. The plebiscite counts as "Yes voters" only those voters who will attend the plebiscite and vote for independence (A subject is not counted as an independence supporter if he or she does not attend the plebiscite). "Don't Know" responses can be thought of as missing data – the true intention of the voter is unknown but must be either "Yes" or "No" for that response.

Table 1. Data from SPOS

		Independence		
Secession	Attendance	Yes	No	DK
Yes	Yes	1191	8	21
	No	8	0	4
	DK	107	3	9
No	Yes	158	68	29
	No	7	14	3
	DK	18	43	31
DK	Yes	90	2	109
	No	1	2	25
	DK	19	8	96

According to Table 1, there are 1191 people who will attend the plebiscite and vote for independence and secession. And another 248 people (158+90) will attend the plebiscite and vote for independence, regardless of the secession response. From this sample, we have a rough idea that most of voters would be in favor of the country's independence.

1.2 Missing Data Mechanism

Rubin (1976) performed one of the first systematic studies of nonresponse mechanisms. His terminology has since become a standard for classifying different types of situations that give rise to missing values. Broadly, there are three types of missing data mechanisms:

1) Missing completely at random (MCAR)
2) Missing at random (MAR)
3) Missing not at random (MNAR)

1.2.1 Missing Completely At Random (MCAR)

A missing data mechanism is MCAR if the probability of an observation being missing (r) does not depend on unobserved (y_m) and observed responses (y_o). Mathematically, it is expressed as

$$P(r|y_o, y_m) = P(r) \tag{1}$$

For a MCAR set up, the analysis using the complete data gives valid inferences. For example, if a participant's data were missing because he was stopped for a car accident and missed the SPOS data collection session, his data would presumably be missing completely at random. Another way to think of MCAR is to note that any piece of data is just as likely to be missing as any other piece of data.

1.2.2 Missing At Random (MAR)

Usually, data are not missing completely at random. A missing data mechanism is considered missing at random (MAR) if, given the observed data, the missing mechanism does not depend on the unobserved data. This can be expressed as

$$P(r|y_o, y_m) = P(r|y_o) \tag{2}$$

Note that under MAR the probability of a value being missing will generally depend on observed values, so it does not correspond to the intuitive notion of 'random'. The important idea is that the missing data mechanism can be expressed solely in terms of observed data. For example, people with prior conviction might be less inclined to report their income, and thus the reported income would be related to crime record. In general, people who had criminal record might have a lower income, so when we have a high rate of missing data among those people with criminal record, the mean income might be different from the one without missing data. However, if the probability of reported income being missing is independent of prior conviction for persons with prior criminal record, then the data could be considered as MAR.

1.2.3 Missing Not At Random (MNAR)

If a process is neither MCAR nor MAR, it is termed as missing not at random (MNAR). In practice, it is very common that the reason for observations being missing still depends on some unobserved data, even after accounting for the observed information. For example, people with low income are less likely to report their income on a data collection form. Clearly, the mean income for the available data will be a biased estimate of the mean income that we would have obtained with complete data. When missing data mechanism is MNAR, the only way to obtain an unbiased estimate of parameters is to model missingness. In other words, we would need to develop a modeling framework that accounts for the missing data. That model could then be incorporated into a more complex model for estimating missing values. Meanwhile, it is also difficult to give the appropriate model for the missing data mechanism.

Generally, it is hard to tell from the data at hand whether the missing data mechanism is MCAR, MAR or MNAR. Sometimes, a research design provides the justification (Murray and Findlay, 1988), but typically this is not so and the incomplete data under analysis can never alone answer the question of whether or not a missing data mechanism is MNAR. We need to explore how our inferences vary under assumptions of MAR, MNAR, and under various models. Kenward et al (2000) performed sensitivity analysis for the Slovenian plebiscite case. This paper develops an ignorable nonresponse model based on bivariate binomial distribution to calculate the parameter of interest under MAR assumption. Under MNAR assumption, we consider a plausible and much specified nonignorable nonresponse model. In this framework we introduce an additional centering model which assigns some common characteristic to the parameters. The nonignorable model's statistical result is as good as that of the ignorable model. Sensitivity analysis proves that model selection is essential for the MNAR case and it is fairly robust, regardless of the variability of the parameters.

1.3 Preliminary Analysis

Rubin et al (1987) indicated there are two methods, “complete cases” and “available cases”, probably among the most common ad hoc methods for dealing with missing data. We treat the DK’s as missing data because eventually every Slovenian would vote “Yes” or “No”. The left column in Table 2 is based on the 1,454 survey participants who answered all three questions “Yes” or “No” - they are the "complete" cases. The right column is based on the 1,549 participants who answered the independence and attendance questions “Yes” or “No” – they are the "available" cases.

Let θ be the population proportion of voters who plan to attend and vote for independence. Because only “Yes” votes for attendance count in the plebiscite, a “Yes” response to independence coupled with a “No” response to attendance effectively counts as a “No” vote in the plebiscite.

The estimate of θ based on the complete cases is 1,349/1,454 = .928, and the corresponding estimate based on the available cases is 1,439/1,549 = .929. Based on the confidence interval formula ($\hat{\theta} \pm Z_{\alpha/2}\sqrt{\hat{\theta}(1-\hat{\theta})/n}$), the 95% confidence intervals for complete cases and available cases are (.9147, .9413) and (.9162, .9418) respectively. Generally, a conservative estimate of θ is based on the assumption that by giving DK responses, subjects avoid revealing an unpopular opinion - a No response. The corresponding estimate is the proportion answering Yes to questions 1 and 3 among the original respondents (i.e., all DK responses are treated as “No”); this proportion is 1,439/2,074 = .694. This is the most pessimistic scenario. On the contrary, we can calculate the most optimistic estimate of θ by treating the DK responses for either independence or attendance questions as “Yes”. This estimate is (1439+21+29+109+107+18+19+9+31+96)/2074 = .905

We should notice that both complete and available case estimates fall outside the pessimistic – optimistic interval (.694, .905) and thus should be discarded. This is because, these two estimators do not incorporate missing data as No votes and thus disregard the available information.

Table 2. Survey Results for Attendance and Independence Questions

	Complete Cases (n = 1454)		Available Cases (n = 1549)	
	Independence		Independence	
Attendance	Yes	No	Yes	No
Yes	1349	76	1439	78
No	15	14	16	16

Another method which is also practical is to estimate θ under a non – parametric framework. Accordingly, we construct Table 3 and consider all the responses which contain DK's as missing values. Our purpose is to assign all the missing data to each blank cell based on Table 3-1 since it is a complete table.

Table 3. Four patterns in Slovenian Plebiscite case

	Yes	No
Yes	1439	78
No	16	16

3-1

	Yes	No	DK
Yes			159
No			32

3-2

	Yes	No
Yes		
No		
DK	144	54

3-3

	Yes	No	DK
Yes			
No			
DK			136

3-4

After filling in the blank cells, we can easily obtain the non – parametric estimate of θ as

$$\hat{\theta} = (1439 + 151 + 142 + 126)/2074 = .896$$

In this non – parametric method, we actually use the MAR assumption. All the estimates for missingness are based on the proportion of the observed data in Table 3-1. This method is easy to perform and it shows that MAR is a good assumption as a starting point of our research.

Table 4. Non – parametric estimation

	Yes	No
Yes	1439	78
No	16	16

4-1

	Yes	No	DK
Yes	151	8	159
No	16	16	32

4-2

	Yes	No
Yes	142	50
No	2	4
DK	144	54

4-3

	Yes	No	DK
Yes	126	7	
No	1.5	1.5	
DK			136

4-4

The official proportion of eligible Slovenian residents who attended the plebiscite and voted in favor of independence was 0.885. We notice that it is close to the non – parametric estimate given above. It also lies within the pessimistic – optimistic interval. Although the true value of θ is not too far from the estimates obtained using the complete and available cases, it is outside their 95% confidence limits.

These findings motivate us to develop more precise methods to estimate θ. The remaining sections are organized as follows: we will discuss the MAR assumption in section 2 and perform estimation using EM algorithm and bootstrap. MNAR will be introduced in section 3. Sensitivity analysis will be carried out in both sections 2 and 3 to check how sample size and missingness rate affect the estimate of θ. Additional sensitivity analysis for checking robustness of the nonignorable model is shown in section 3.

Chapter 2. Ignorable Nonresponse Model

We assume that responses on the attendance and independence questions can be thought of as data that are missing at random (MAR). Responses are said to be MAR if the occurrence of DK is conditionally independent of the actual answer that would have been observed given the observed responses to one or both of the other questions; that is, the probability of the occurrence of DK can depend on the observed answers to other questions, but given these, it is independent of the missing value itself. If the data are MAR and the parameters of the probability model for the missingness are different from the parameters of the probability model for the data, then the missingness model is called ignorable (Rubin 1976), because it does not affect likelihood-based inferences, such as MLE’s.

2.1 Theoretical Framework

We consider a 2X2 table with cell counts {z, x – z, y – z, n – x – y + z} and corresponding probabilities {θ, p – θ, q – θ, 1 – p – q + θ}. The table below shows all the cells and the respective probabilities (in brackets). We will develop a new model based on a bivariate binomial distribution.

	Yes	No	Margin
Yes	z (θ)	x – z (p – θ)	x (p)
No	y – z (q – θ)	n – x – y + z (1 – p – q + θ)	n – x (1 - p)
Margin	y (q)	n – y (1 - q)	n

The joint probability mass function of x, y, z is given by

$$f(x,y,z) = \frac{n!\theta^{z}(p-\theta)^{x-z}(q-\theta)^{y-z}(1-p-q+\theta)^{n-x-y+z}}{z!(x-z)!(y-z)!(n-x-y+z)!}$$, 0 ≤ x ≤ n, 0 ≤ y≤ n, 0 ≤ z ≤ min{x, y}, 0 ≤ θ ≤ p, q ≤ 1.

Thus we do not use a standard multinomial model for modeling (x, y, z). This is because of the fact that we want to model the random margins. And this setting has additional constraints for the variables and parameters – this will prove to be useful for the sensitivity analysis to be done later.

Hamdan (1969) and Kocherlakota (1989) discuss a bivariate binomial probability mass function (PMF) for the same situation. We have generalized their PMF by including three variables. They integrated out z since their variables of interest were x and y. Moreover, it is very difficult to integrate out each parameter from the mass function. However, our model is based on a three-parameter PMF with three variables and it gives us more flexibility during calculation (Hamdan's model had three parameters with two variables).

First, we obtain the marginal probability mass function of x, y, and z. Then, we get the conditional PMF of (y, z | x) and (x, z | y). The calculations are shown below.

(1) f(x)

$$
\begin{aligned}
f(x) &= \sum_{z=0}^{y} \sum_{y=z}^{n-x+z} \frac{n!\,\theta^z (p-\theta)^{x-z} (q-\theta)^{y-z} (1-p-q+\theta)^{n-x-y+z}}{z!\,(x-z)!\,(y-z)!\,(n-x-y+z)!} \\
&= \sum_{z=0}^{y} \frac{n!\,\theta^z (p-\theta)^{x-z}}{z!\,(x-z)!} \sum_{y-z=0}^{n-x} \frac{(q-\theta)^{y-z} (1-p-q+\theta)^{n-x-y+z}}{(y-z)!\,(n-x-y+z)!} \\
&= \frac{n! p^x (1-p)^{n-x}}{x!(n-x)!} \sum\nolimits_{z=0}^{y} \frac{x!\left(\frac{\theta}{p}\right)^z \left(1-\frac{\theta}{p}\right)^{x-z}}{z!(x-z)!} \sum\nolimits_{y-z=0}^{n-x} \frac{(n-x)!\left(\frac{q-\theta}{1-p}\right)^{y-z} \left(1-\frac{q-\theta}{1-p}\right)^{n-x-y+z}}{(y-z)!(n-x-y+z)!}.
\end{aligned}
$$

Thus,

$$
f(x) = \frac{n!\,p^x (1-p)^{n-x}}{x!\,(n-x)!} \tag{3}
$$

ie $x \sim Binomial(n, p)$

(2) f(y)

$$
\begin{aligned}
f(y) &= \sum_{z=0}^{x} \sum_{x=z}^{n-y+z} \frac{n!\,\theta^z (p-\theta)^{x-z} (q-\theta)^{y-z} (1-p-q+\theta)^{n-x-y+z}}{z!\,(x-z)!\,(y-z)!\,(n-x-y+z)!} \\
&= \sum_{z=0}^{x} \frac{n!\,\theta^z (q-\theta)^{y-z}}{z!\,(y-z)!} \sum_{x-z=0}^{n-y} \frac{(p-\theta)^{x-z} (1-p-q+\theta)^{n-x-y+z}}{(x-z)!\,(n-x-y+z)!} \\
&= \frac{n! q^x (1-q)^{n-y}}{y!(n-y)!} \sum\nolimits_{z=0}^{x} \frac{x!\left(\frac{\theta}{q}\right)^z \left(1-\frac{\theta}{q}\right)^{x-z}}{z!(y-z)!} \sum\nolimits_{x-z=0}^{n-y} \frac{(n-y)!\left(\frac{p-\theta}{1-q}\right)^{x-z} \left(1-\frac{p-\theta}{1-q}\right)^{n-x-y+z}}{(x-z)!(n-x-y+z)!}.
\end{aligned}
$$

Thus,

$$
f(y) = \frac{n!\,q^x (1-q)^{n-y}}{y!\,(n-y)!} \tag{4}
$$

ie $y \sim Binomial(n, q)$

(3) f(z)

$$f(z) = \sum_{y=z}^{n-x+z} \sum_{x=z}^{n-y+z} \frac{n!\,\theta^z(p-\theta)^{x-z}(q-\theta)^{y-z}(1-p-q+\theta)^{n-x-y+z}}{z!\,(x-z)!\,(y-z)!\,(n-x-y+z)!}$$

$$= \frac{n!\,\theta^z}{z!} \sum_{y-z=0}^{n-x} \sum_{x-z=0}^{n-y} \frac{(p-\theta)^{x-z}(q-\theta)^{y-z}(1-p-q+\theta)^{n-x-y+z}}{(x-z)!\,(y-z)!\,(n-x-y+z)!}$$

$$= \frac{n!\,\theta^z(1-\theta)^{n-z}}{z!\,(n-z)!} \sum_{y-z=0}^{n-x} \sum_{x-z=0}^{n-y} \left[\frac{(x+y-2z)!\left(\frac{p-\theta}{p+q-2\theta}\right)^{x-z}\left(\frac{q-\theta}{p+q-2\theta}\right)^{y-z}}{(x-z)!\,(y-z)!}\right]$$

$$\left[\frac{(n-z)!\,(\frac{p+q-2\theta}{1-\theta})^{x+y-2z}(\frac{1-p-q+\theta}{1-\theta})^{n-x-y+z}}{(x+y-2z)!\,(n-x-y+z)!}\right]$$

Thus,

$$f(z) = \frac{n!\,\theta^z(1-\theta)^{n-z}}{z!\,(n-z)!} \qquad (5)$$

ie $z \sim Binomial(n, \theta)$

(4) f(y, z|x)

$$f(y,z|x) = \frac{f(x,y,z)}{f(x)} = \frac{x!\,(n-x)!\,\theta^z(p-\theta)^{x-z}(q-\theta)^{y-z}(1-p-q+\theta)^{n-x-y+z}}{z!\,(x-z)!\,(y-z)!\,(n-x-y+z)!\,p^x(1-p)^{n-x}}$$

$$= \frac{x!\frac{\theta^z}{p}\left(\frac{p-\theta}{p}\right)^{x-z}}{z!\,(x-z)!} \frac{(n-x)!\left(\frac{q-\theta}{1-p}\right)^{y-z}\left(\frac{1-p-q+\theta}{1-p}\right)^{n-x-y+z}}{(y-z)!\,(n-x-y+z)!}$$

Thus,

$$f(y,z|x) = f(z|x)f(y-z|x,z) \qquad (6)$$

$$where\ f(z|x) = Binomial(x, \theta/p), f(y-z|x,z) = Binomial(n-x, (q-\theta)/(1-p))$$

(5) f(x, z|y)

$$f(x,z|y) = \frac{f(x,y,z)}{f(y)} = \frac{y!\,(n-y)!\,\theta^z(p-\theta)^{x-z}(q-\theta)^{y-z}(1-p-q+\theta)^{n-x-y+z}}{z!\,(x-z)!\,(y-z)!\,(n-x-y+z)!\,q^y(1-q)^{n-y}}$$

$$= \frac{y!\frac{\theta^z}{q}\left(\frac{q-\theta}{q}\right)^{y-z}}{z!\,(y-z)!} \frac{(n-y)!\left(\frac{p-\theta}{1-q}\right)^{x-z}\left(\frac{1-p-q+\theta}{1-q}\right)^{n-x-y+z}}{(x-z)!\,(n-x-y+z)!}$$

Thus,

$$f(x,z|y) = f(z|y)f(x-z|y,z) \qquad (7)$$

$$\begin{aligned} where\ f(z|y) = Binomial(y, \theta/q), f(x-z|y,z) \\ = Binomial(n-y, (p-\theta)/(1-q)) \end{aligned}$$

2.2 EM algorithm

Under the ignorability assumption, the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) can be used to obtain the MLE's of θ and the other response probabilities from the incomplete contingency table (as in Fuchs 1982). EM algorithm is a method for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. EM is an iterative method which alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood function evaluated using the current estimate for the latent variables, and the maximization (M) step, which computes parameters maximizing the expected log-likelihood function evaluated in the E step. These parameter estimates are then used to determine the distribution of the latent variables in the next E step. This process in repeated until the estimates convergence. The resulting estimates are the optimal ones.

Mathematically, we can write the log likelihood function as $g(y_o, y_m|\theta) = \log p(y_o, y_m|\theta)$. For the "E step", we replace the missing values with their expectations. As for the "M step", we need to obtain $\max_\theta E(g(y_o, y_m|\theta)$. We use ${y_m}^{(r+1)} = E(y_m|y_o, \theta^{(r)})$ (y_o and y_m are observed and missing data. They can be numbers or vectors, $\theta^{(r)}$ being the parameter value at the rth step) at the rth step of the EM algorithm. We would maximize the log likelihood when θ reaches convergence. The EM algorithm progresses monotonically to the maximum likelihood estimate.

For the plebiscite case, we initially consider drawing inferences using only the results of questions 1 and 3, which are directly relevant to θ, and make the MAR assumption for this two-variable data set. The relevant two-way margin of the data set is given below in Table 5.

Table 5. Analysis for Attendance and Independence Questions

Attendance	Independence		
	Yes	No	DK
Yes	1439	78	159
No	16	16	32
DK	144	54	136

The purpose of missing data analysis is to get the best (or optimal) parameter estimate based on the data. Missing data can have four patterns which are shown in Table 6. The (Yes, Yes) cell is z, (Yes, +) is x and (+, Yes) is y. The corresponding proportions are θ, p and q, respectively. In this case, we can obtain nine counts in Table 6 which are based on the cells in Table 5. In Table 6 we have four patterns: for the first one we have all the data; we have the vertical and horizontal margins in the second and third patterns; only the total is known in the last pattern.

Table 6. Observed Cells for the SPOS

z_1=1439		x_1 = 1517
y_1=1455		n_1 = 1549

		x_2=159
		n_2=191

y_3=144		n_3 = 198

		n_4=136

It is clear that each cell in the 2X2 table follow the multinomial distribution with four possible outcomes ((Yes, Yes), (Yes, No), (No, Yes) and (No, No)).

2.2.1 Implementation

Here we show the details of the EM algorithm.

The log likelihood function is given by:

$$\mathrm{L}(\mathrm{p},\mathrm{q},\theta|\mathrm{x}_t,\mathrm{y}_t,\mathrm{z}_t) = \log\prod_{t=1}^{4}\frac{\mathrm{n}_t!}{\mathrm{z}_t!(\mathrm{x}_t-\mathrm{z}_t)!(\mathrm{y}_t-\mathrm{z}_t)!(\mathrm{n}_t-\mathrm{x}_t-\mathrm{y}_t+\mathrm{z}_t)!} + \log\theta\sum_{t=1}^{4}\mathrm{z}_t + \log(\mathrm{p}-$$

$$\theta)\sum_{t=1}^{4}(\mathrm{x}_t-\mathrm{z}_t) + \log(\mathrm{q}-\theta)\sum_{t=1}^{4}(\mathrm{y}_t-\mathrm{z}_t) + \log(1-\mathrm{p}-\mathrm{q}+\theta)\sum_{t=1}^{4}(\mathrm{n}_t-\mathrm{x}_t-\mathrm{y}_t+\mathrm{z}_t),$$

$0 \le x_t \le n_t$, $0 \le y_t \le n_t$, $0 \le z_t \le \min\{x_t, y_t\}$.

E step:

$\mathrm{E}_{\mathrm{y_2,z_2,x_3,z_3,x_4,y_4,z_4|x_1,y_1,z_1,x_2,y_3,p^{(r)},q^{(r)},\theta^{(r)}}}\big(\mathrm{L(p,q,\theta|x_t,y_t,z_t)}\big) =$

$\mathrm{C} + \mathrm{E}_{\mathrm{y_2,z_2,x_3,z_3,x_4,y_4,z_4|x_2,y_3,p^{(r)},q^{(r)},\theta^{(r)}}}(\log \prod_{t=2}^{4} \frac{n_t!}{z_t!(x_t-z_t)!(y_t-z_t)!(n_t-x_t-y_t+z_t)!} + \log\theta^{(r)} \sum_{t=2}^{4} z_t +$

$\log\big(p^{(r)} - \theta^{(r)}\big) \sum_{t=2}^{4}(x_t - z_t) + \log\big(q^{(r)} - \theta^{(r)}\big) \sum_{t=2}^{4}(y_t - z_t) + \log\big(1 - p^{(r)} - q^{(r)} +$

$\theta^{(r)}\big) \sum_{t=1}^{4}(n_t - x_t - y_t + z_t))$, $0 \le x_t \le n_t$, $0 \le y_t \le n_t$, $0 \le z_t \le \min\{x_t, y_t\}$.

Here, C is the observed data which represent constant values.

Now,

$$\mathrm{E(z_2|x_2)} = x_2 \frac{\theta}{p}, \mathrm{E(y_2|x_2)} = x_2 \frac{\theta}{p} + (n_2 - x_2)\frac{q-\theta}{1-p},$$

$$\mathrm{E(z_3|y_3)} = y_3 \frac{\theta}{q}, \mathrm{E(x_3|y_3)} = y_3 \frac{\theta}{q} + (n_3 - y_3)\frac{p-\theta}{1-q},$$

$$E(x_4) = n_4 p, E(y_4) = n_4 q, E(z_4) = n_4 \theta$$

Let us show how we obtain the expectations of all the missing data.

When t = 1, it is a complete table (observed data) and we can get $\hat{p}$, $\hat{q}$ and $\hat{\theta}$ easily as

$$\hat{p} = \frac{1517}{1549} = 0.979$$

$$\hat{q} = \frac{1455}{1549} = 0.939$$

$$\hat{\theta} = \frac{1439}{1549} = 0.929$$

When t = 2, x_2 is known. We need to find out the distribution of x_2 first and then deduce the distribution of y_2, z_2 given x_2. In 2.1 we have already shown that x_2 follow Binomial distribution. Then we can deduce the distribution of y_2, z_2 given x_2.

From section 2.1,

$$z_2|x_2 \sim Binomial\left(x_2, \frac{\theta}{p}\right)$$

$$\mathrm{y_2 - z_2}|x_2, \mathrm{z_2} \sim \mathrm{Binomial}(\mathrm{n_2 - x_2}, \frac{\mathrm{q-\theta}}{\mathrm{1-p}})$$

Now we know the distributions of z_2 given x_2 and $y_2 - z_2$ given z_2 and x_2. Hence we can easily develop the expectations of z_2 and y_2 given x_2 as follows:

$$E(z_2|x_2) = x_2\frac{\theta}{p} = \frac{159\theta}{p}, E(y_2 - z_2|z_2, x_2) = (n_2 - x_2)\frac{q-\theta}{1-p} = 32\frac{q-\theta}{1-p}$$

$$E(y_2 - z_2|z_2, x_2) = E_{z_2}\big(E(y_2|z_2, x_2)\big) - E_{z_2}\big(E(z_2|z_2, x_2)\big) = E(y_2|x_2) - E(z_2|x_2)$$

$$E(y_2|x_2) = E(z_2|x_2) + E(y_2 - z_2|z_2, x_2) = \frac{159\theta}{p} + 32\frac{q-\theta}{1-p}$$

Now, we can fill up the missing cells in "t = 2" part in Table 6 by using their expected values. Same method is used in the situation when t = 3 (y_3 is known).

$$y_3 \sim Binomial(n_3, q)$$

$$z_3|y_3 \sim Binomial\left(y_3, \frac{\theta}{q}\right)$$

$$y_3 - z_3|y_3, z_3 \sim Binomial(n_3 - y_3, \frac{p-\theta}{1-q})$$

$$E(z_3|y_3) = \frac{144\theta}{q}$$

$$E(x_3|y_3) = \frac{144\theta}{q} + 54\frac{p-\theta}{1-q}$$

As for t = 4, only n_4 is known. The expectations of multinomial distribution should be filled in the table as:

$$E(x_4) = n_4 p$$

$$E(y_4) = n_4 q$$

$$E(z_4) = n_4 \theta$$

All of the results are shown in Table 7.

M step:

Now we have the new MLE's of the multinomial cell probabilities in the 2X2 table based on the current values of the expected complete – data sufficient statistics and we can maximize $\mathrm{E}\big(\mathrm{L}(\mathrm{p}, \mathrm{q}, \theta | \mathrm{x_t}, \mathrm{y_t}, \mathrm{z_t})\big)$ by using the expectations of all the missing values.

$$\theta^{(r+1)} = \frac{z_1 + \mathrm{x}_2 \frac{\theta^{(\mathrm{r})}}{\mathrm{p}^{(\mathrm{r})}} + \mathrm{y}_3 \frac{\theta^{(\mathrm{r})}}{\mathrm{q}^{(\mathrm{r})}} + n_4 \theta^{(r)}}{n} \qquad (8)$$

$$p^{(r+1)} = \frac{x_1 + \mathrm{x}_2 + \mathrm{y}_3 \frac{\theta^{(\mathrm{r})}}{\mathrm{q}^{(\mathrm{r})}} + (\mathrm{n}_3 - \mathrm{y}_3) \frac{\mathrm{p}^{(\mathrm{r})} - \theta^{(\mathrm{r})}}{1 - \mathrm{q}^{(\mathrm{r})}} + n_4 p^{(r)}}{n} \qquad (9)$$

$$q^{(r+1)} = \frac{y_1 + \mathrm{x}_2 \frac{\theta^{(\mathrm{r})}}{\mathrm{p}^{(\mathrm{r})}} + (\mathrm{n}_2 - \mathrm{x}_2) \frac{\mathrm{q}^{(\mathrm{r})} - \theta^{(\mathrm{r})}}{1 - \mathrm{p}^{(\mathrm{r})}} + \mathrm{y}_3 + n_4 q^{(r)}}{n} \qquad (10)$$

After convergence, the MLE of θ = 0.892 and its 95% confidence interval is (0.8768, 0.9080). Corresponding results for p and q are obtained in the same way. Detailed calculation and program outputs are given below:

Table 7. Missing Cells are filled with Expected Values

z_1=1439	

x_1 = 1517

y_1=1455

n_1 = 1549

t = 1

z_2=$\frac{159\theta}{p}$	

x_2=159

y_2=$\frac{159\theta}{p} + 32\,\frac{q-\theta}{1-p}$

n_2=191

t = 2

z_3=$\frac{144\theta}{q}$	

X_3 = $\frac{144\theta}{q} + 54\,\frac{p-\theta}{1-q}$

y_3=144

n_3 = 198

t = 3

z_4=136θ	

X_4=136p

y_4=136q

n_4=136

t = 4

$$\hat{p} = \frac{\sum_t x_t}{\sum_t n_t}, \hat{q} = \frac{\sum_t y_t}{\sum_t n_t}, \hat{\theta} = \frac{\sum_t z_t}{\sum_t n_t}$$

Using the values of $\widehat{p},\ \hat{q}\ and\ \hat{\theta}$ (when t = 1) as initial values for p, q and θ, we run EM algorithm in R and we attain quick convergence of all the parameters.

Table 8. R Output for EM Algorithm

iteration	θ	p	q
1	0.896186	0.96233	0.912643
2	0.892535	0.943979	0.921041
3	0.892293	0.947393	0.917218
4	0.892313	0.945974	0.917948
5	0.892356	0.946141	0.917618
6	0.892381	0.945969	0.917642
7	0.892396	0.945944	0.917597
8	0.892405	0.94591	0.917588
9	0.89241	0.945896	0.917577
10	0.892413	0.945886	0.917573
11	0.892415	0.945881	0.91757
12	0.892416	0.945877	0.917568
13	0.892416	0.945876	0.917567
14	0.892416	0.945875	0.917567
15	0.892417	0.945874	0.917567
16	0.892417	0.945874	0.917566
17	0.892417	0.945874	0.917566
18	0.892417	0.945873	0.917566
19	0.892417	0.945873	0.917566
20	0.892417	0.945873	0.917566

In order to get the 95% confidence interval of θ, we need to calculate the Hessian Matrix given by:

$$\mathrm{H} = \begin{bmatrix} \frac{\partial^2 L}{\partial \theta^2} & \frac{\partial^2 L}{\partial \theta \partial q} & \frac{\partial^2 L}{\partial \theta \partial p} \\ \frac{\partial^2 L}{\partial q \partial \theta} & \frac{\partial^2 L}{\partial q^2} & \frac{\partial^2 L}{\partial q \partial p} \\ \frac{\partial^2 L}{\partial p \partial \theta} & \frac{\partial^2 L}{\partial p \partial q} & \frac{\partial^2 L}{\partial p^2} \end{bmatrix}$$

Here,

$$\frac{\partial^2 L}{\partial \theta^2} = -\frac{\sum_{t=1}^{4} z_t}{\theta^2} - \frac{\sum_{t=1}^{4}(x_t - z_t)}{(p-\theta)^2} - \frac{\sum_{t=1}^{4}(y_t - z_t)}{(q-\theta)^2} - \frac{\sum_{t=1}^{4}(n_t - x_t - y_t + z_t)}{(1-p-q+\theta)^2}$$

$$\frac{\partial^2 L}{\partial q^2} = -\frac{\sum_{t=1}^{4}(y_t - z_t)}{(q-\theta)^2} - \frac{\sum_{t=1}^{4}(n_t - x_t - y_t + z_t)}{(1-p-q+\theta)^2}$$

$$\frac{\partial^2 L}{\partial p^2} = -\frac{\sum_{t=1}^{4}(x_t - z_t)}{(p-\theta)^2} - \frac{\sum_{t=1}^{4}(n_t - x_t - y_t + z_t)}{(1-p-q+\theta)^2}$$

$$\frac{\partial^2 L}{\partial \theta \partial q} = \frac{\partial^2 L}{\partial q \partial \theta} = \frac{\sum_{t=1}^{4}(y_t - z_t)}{(q-\theta)^2} + \frac{\sum_{t=1}^{4}(n_t - x_t - y_t + z_t)}{(1-p-q+\theta)^2}$$

$$\frac{\partial^2 L}{\partial \theta \partial p} = \frac{\partial^2 L}{\partial p \partial \theta} = \frac{\sum_{t=1}^{4}(x_t - z_t)}{(p-\theta)^2} + \frac{\sum_{t=1}^{4}(n_t - x_t - y_t + z_t)}{(1-p-q+\theta)^2}$$

$$\frac{\partial^2 L}{\partial p \partial q} = \frac{\partial^2 L}{\partial q \partial p} = -\frac{\sum_{t=1}^{4}(n_t - x_t - y_t + z_t)}{(1-p-q+\theta)^2}$$

Plugging into the EM result of θ, p and q (.892, .946, .918) and calculate the Hessian Matrix:

$$H = \begin{bmatrix} -267442 & 89824.22 & 246864.4 \\ 89824.22 & -89824.2 & -71571.1 \\ 246864.4 & -71571.1 & -246864 \end{bmatrix}$$

The negative inverse of the Hessian Matrix is given by:

$$-\mathrm{H}^{-1} = \begin{bmatrix} 6.34771\mathrm{E}-05 & 1.67741\mathrm{E}-05 & 5.86139\mathrm{E}-05 \\ 1.67741\mathrm{E}-05 & 1.89098\mathrm{E}-05 & 1.12917\mathrm{E}-05 \\ 5.86139\mathrm{E}-05 & 1.12917\mathrm{E}-05 & 5.9391\mathrm{E}-05 \end{bmatrix}$$

The (1, 1) element of this matrix is the variance of $\hat{\theta}$. Thus, using normal approximation, we can easily obtain the 95% confidence interval for θ, p and q as:

θ ϵ (0.8768, 0.9080)

p ϵ (0.9374, 0.9544)

q ϵ (0.9025, 0.9327)

Here we recall that the true value of θ is 0.885 which is included in the above interval.

2.3 Bootstrap distributions for θ, p and q

Bootstrapping is a resampling technique used to obtain estimates of sampling distributions of statistics. Bootstrapping is the practice of estimating properties of an estimator by sampling from an approximating distribution. One standard choice for an approximating distribution is the empirical distribution of the observed data. In the case where a set of observations can be assumed to be from an independent and identically distributed population, this can be implemented by constructing a number of resamples of the observed dataset (and of equal size to the observed dataset), each of which is obtained by random sampling with replacement from the original dataset.

The advantage of bootstrapping over analytical methods is its simplicity - it is straightforward to apply the bootstrap to derive estimates of standard errors and confidence intervals for estimators of complex parameters of a distribution, such as percentile points, proportions, odds ratio, and correlation coefficients. For instance, we have 10 values of weights. It would not be precise for us to estimate the population mean weight and its confidence interval based on normality assumption. For bootstrap, first of all, we draw a sample of size 10 with replacement and obtain the average weight. Then we repeat this procedure 100 times and we can get a new sample of 100 average weights. Now we can obtain a more precise estimate and a better distribution of the average weight.

In section 2, we have already implemented EM algorithm for the SPOS data. If all the counts except (n_1, n_2, n_3 and n_4) in Table 6 are random numbers, it would be difficult for us to obtain the distribution of θ. A bootstrapping procedure will be very useful here. Moreover bootstrap is a good method to test the normality assumption which we use in previous section, although the sample size of the SPOS data is fairly large.

In the plebiscite case, based on section 2.1's proof, we can set the following random cells:

When $t = 1$, $(z_{11}, z_{12}, z_{13}, z_{14}) \sim$ Multinomial $(n_1, \theta, p-\theta, q-\theta, 1-p-q+\theta)$ with $x_1 = z_{11}+z_{12}$, $y_1=z_{11}+z_{13.}$

When $t = 2$, $x_2 \sim$ Binomial (n_2, p)

When $t = 3$, $y_3 \sim$ Binomial (n_3, q)

Following section 2.2's procedure, we can know all the random numbers and fill up all the missing cells by using their expected values.

Table 9. Missing Cells are filled with Expected Values under random variables

z_{11}~Multinomial	z_{12} ~Multinomial
z_{13} ~Multinomial	z_{14}~Multinomial

$x_1 = z_{11}+z_{12}$

$y_1= z_{11}+z_{13}$ $n_1 = 1549$

t = 1

z_2	

x_2~Binomial

y_2 n_2=191

t = 2

z_3	

X_3

y_3=Binomial n_3 = 198

t = 3

z_4	

X_4

y_4 n_4=136

t = 4

Now, each time we obtain a sample of z_{11}, x_2 and y_3 and perform EM algorithm to get the estimates of θ, p and q (using the EM algorithm results as starting values for θ, p, and q). After repeating 1,000 times, we can get the bootstrap confidence intervals for them. Based on 1,000 repetitions, we find that 1,000 of the 1,000 bootstrapping confidence intervals contain the "true" value of p, q and θ (p = 0.946, q = 0.918 and θ = 0.892, which are the results of section 2. 2).

Here is one set of bootstrap confidence limits for θ, p, and q:

$\theta \in (0.8774068, 0.9059108)$

$p \in 0.9351495, 0.9559075)$

$q \in (0.903989, 0.9294458)$

Smoothing the data distribution with a kernel density estimate can be more effective than using a histogram to identify features. A kernel density estimate can also be more effective than a parametric curve fit when the process distribution is multi-modal. In Figure 1, we obtain both kernel curve and normal approximating curve to identify θ, p, and q.

The general form of the kernel density estimator is given as:

$$\hat{f}_{\varphi}(x) = \frac{1}{n\rho}\sum_{i=1}^{n} K_0(\frac{x - x_i}{\rho})$$

Where,

K_0 is the kernel function;

ρ is the bandwidth;

n is the sample size;

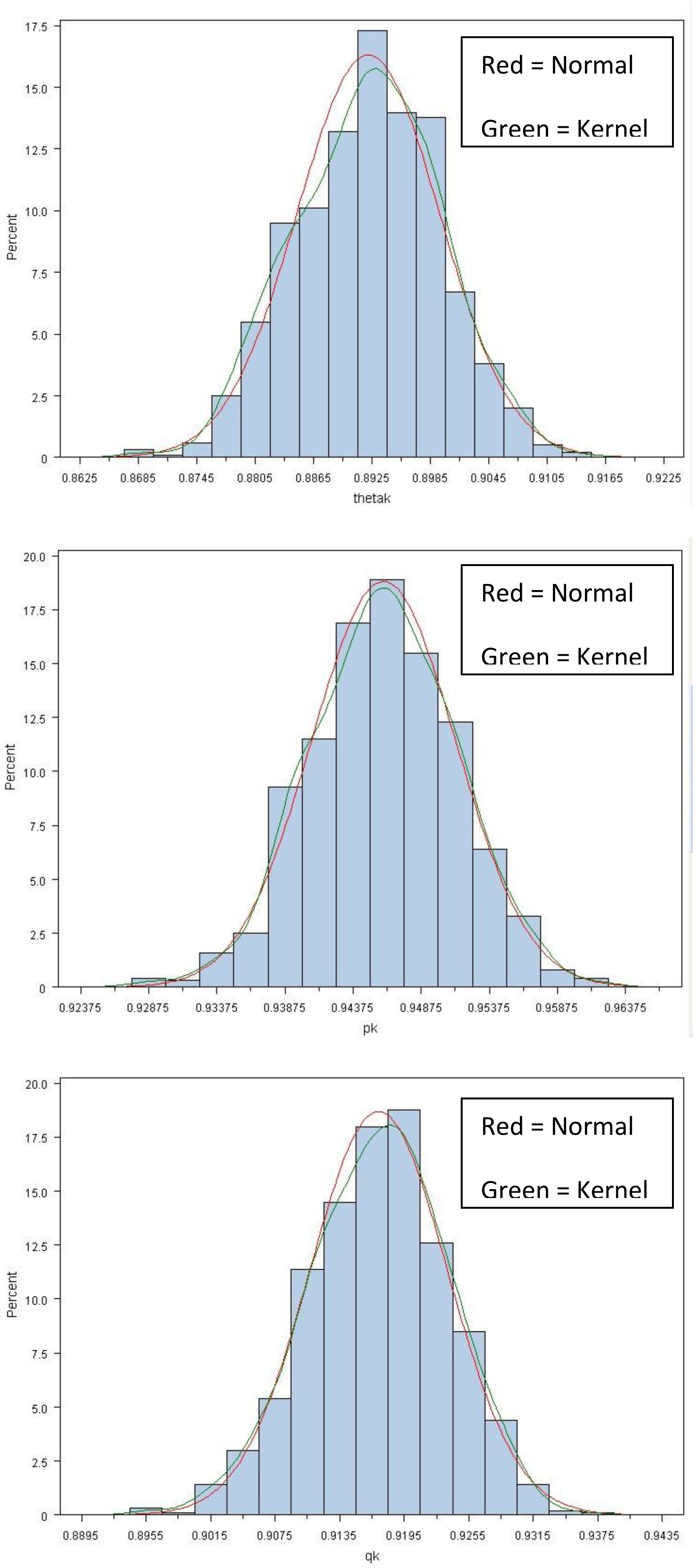
x_i is the ith observation;

Usually, there are three kernel functions: normal, quadratic, and triangular. We use a normal kernel here, which has the form

$$K_0(t) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}t^2\right)$$

Figure 1 shows the SAS output of the bootstrap density estimates of θ, p and q.

Figure 1. Bootstrap Densities of θ, p and q (MAR)



From Figure 1 we can see that the two curves are close and the kernel curves are more precise than the normal approximate curves. Moreover, the bootstrap confidence intervals for θ, p and q are also close to the 95% confidence intervals based on the normality assumption.

2.4 Sensitivity Analysis

Sensitivity analysis is a technique used to determine how different values of an independent variable will impact a particular dependent variable under a given set of scenarios. This technique is used within specific boundaries that will depend on one or more input variables, such as the effect that changes in interest rates will have on a bond's price.

The main goal of sensitivity analysis is to gain insight into which assumptions are critical, i.e., which assumptions affect decision. The process involves various ways of changing input values of the model to see the effect on the output value. In some decision situations you can use a single model to investigate several alternatives. In other cases, you may use different model for each alternative.

In the SPOS data, the sample size is 2074 which is fairly large and most of participants answered both attendance and independence questions. In order to test how the parameter changes with different settings of the ignorable model, we can perform a sensitivity analysis for θ, which is our parameter of interest.

Table 10. Sensitivity Analysis for Bootstrap θ

	n	Completeness	95% Lower Band for θ	95% Upper Band for θ
1	2,000	75%	0.8778373	0.9060658
2		50%	0.8743162	0.9070934
3		25%	0.8663464	0.9189638
4	1,000	75%	0.8716063	0.913232
5		50%	0.8687155	0.912698
6		25%	0.8641333	0.9189125
7	500	75%	0.8609234	0.9196522
8		50%	0.8603767	0.9240693
9		25%	0.856592	0.9317466
10	100	75%	0.8253963	0.9538733
11		50%	0.8144843	0.9552296
12		25%	0.8048008	0.9602226

According to Table 10, we can find out that a wider bootstrapping confidence limit would be obtained if the completeness (proportion of non-DK answers) or the sample size decreases. When we have large sample (n = 2,000 or 1,000), the parameter is not changed much. However, if we do not collect enough sample (say n = 100), the bootstrap confidence limits would be much wider for large samples. Hence, we can conclude that under MAR assumption, the surveyor needs to collect a large enough sample and try to get the respondents' complete questionnaires.

Chapter 3. Nonignorable Nonresponse Model

When data are missing not at random (MNAR), the results in section 2 does not hold anymore and maximum likelihood estimation of the model parameters based only on the observed likelihood will be biased. To obtain the correct maximum likelihood estimates of the model parameters, we need to extend the model in section 2, since we have different θ's, p's and q's under different patterns (each pattern has different reason to be complete / missing). Moreover, we allow them to have a common effect by assigning a distribution, or we cannot estimate those parameters. Thus we have a nonignorable nonresponse model. In most cases the parameters of the centering model will also be unknown. Fortunately, the parameters of the combined data and centering model can be estimated simultaneously by maximizing the full data log likelihood using the standard EM algorithm.

3.1 EM Algorithm

The data follow a multinomial distribution, which is same as what we used in section 2. The data distribution is given by:

$$f(x_t, y_t, z_t | p_t, q_t, \theta_t) = \frac{n_t! {\theta_t}^{z_t} (p_t-\theta_t)^{x_t-z_t} (q_t-\theta_t)^{y_t-z_t} (1-p_t-q_t+\theta_t)^{n_t-x_t-y_t+z_t}}{z_t!(x_t-z_t)!(y_t-z_t)!(n_t-x_t-y_t+z_t)!}, \ 0 \le x_t \le n_t, \ 0 \le y_t \le n_t,$$

and $0 \le z_t \le \min\{x_t, y_t\}$, t = 1 to 4.

Here we choose a Dirichlet distribution as a centering model because it is conjugate to the multinomial distribution. So we will get the expanded distribution easily. If we do not use a centering model, it would be impossible for us to estimate the last three patterns (t = 2, 3, 4).

The centering distribution is given by:

$$f(p_t, q_t, \theta_t) = \frac{{\theta_t}^{\mu_1\tau-1} (p_t-\theta_t)^{(\mu_2-\mu_1)\tau-1} (q_t-\theta_t)^{(\mu_3-\mu_1)\tau-1} (1-p_t-q_t+\theta_t)^{(1-\mu_2-\mu_3+\mu_1)\tau-1}}{D(\mu_1\tau, (\mu_2-\mu_1)\tau, (\mu_3-\mu_1)\tau, (1-\mu_2-\mu_3+\mu_1)\tau)}, \ 0 < \mu_1 < \mu_2, \ \mu_3 < 1$$

and $\tau > 0$.

$$\mathit{where}\ D(\mu_1\tau, (\mu_2-\mu_1)\tau, (\mu_3-\mu_1)\tau, (1-\mu_2-\mu_3+\mu_1)\tau) = \frac{\boldsymbol{\Gamma}(\mu_1\tau)\boldsymbol{\Gamma}(\mu_2-\mu_1\tau)\boldsymbol{\Gamma}(\mu_3-\mu_1\tau)\boldsymbol{\Gamma}(1-\mu_2-\mu_3+\mu_1\tau)}{\Gamma(\tau)}$$

We define the parameters as τ, μ₁, μ₂ and μ₃ because they are equivalent to n, θ, p and q in the data distribution.

The expanded distribution is given by:

$$f\left(\theta_t, p_t, q_t, x_t, y_t, z_t\right) = \frac{\theta_t^{z_t+\mu_1\tau-1}\left(p_t-\theta_t\right)^{x_t-z_t+\left(\mu_2-\mu_1\right)\tau-1}\left(q_t-\theta_t\right)^{y_t-z_t+\left(\mu_3-\mu_1\right)\tau-1}(1-p_t-q_t+\theta_t)^{n_t-x_t-y_t+z_t+\left(1-\mu_2-\mu_3+\mu_1\right)\tau-1}}{D\left(z_t+\mu_1\tau, x_t-z_t+(\mu_2-\mu_1)\tau, y_t-z_t+\left(\mu_3-\mu_1\right)\tau, n_t-x_t-y_t+z_t+(1-\mu_2-\mu_3+\mu_1)\tau\right)}, \theta < p, q < 1.$$

The likelihood function is given by:

$$f(\theta_t, p_t, q_t, x_t, y_t, z_t) = \prod_{t=1}^{4} \frac{\theta_t^{z_t+\mu_1\tau-1}(p_t-\theta_t)^{x_t-z_t+(\mu_2-\mu_1)\tau-1}(q_t-\theta_t)^{y_t-z_t+(\mu_3-\mu_1)\tau-1}(1-p_t-q_t+\theta_t)^{n_t-x_t-y_t+z_t+(1-\mu_2-\mu_3+\mu_1)\tau-1}}{D(z_t+\mu_1\tau, x_t-z_t+(\mu_2-\mu_1)\tau, y_t-z_t+(\mu_3-\mu_1)\tau, n_t-x_t-y_t+z_t+(1-\mu_2-\mu_3+\mu_1)\tau)} (3).$$

In section 2, we have random samples of θ, p and q. We use them to fit the centering model by R program package "dirichlet ()" and obtain μ₁, μ₂, μ₃, and τ.

Let y = (θ, p – θ, q – θ, 1 – p – q + θ) and a = (μ₁ τ, (μ₂ – μ₁) τ, (μ₃ – μ₁) τ, (1 - μ₂ – μ₃ + μ₁) τ). In R program, one has E (y_i) = a_i/ τ (i = 1 to 4), which are returned as the fitted values. For this distribution, Fisher scoring corresponds to Newton-Raphson algorithm. After running the program, we have:

$$\mu_1 = 0.892$$

$$\mu_2 = 0.941$$

$$\mu_3 = 0.921$$

$$\tau = 1950$$

Those estimates are close to n, θ, p and q in the ignorable model because the estimates are based on the bootstrap samples in section 2.

E step:

Unobserved parts contain x_3, x_4, y_2, y_4, z_2, z_3 and z_4 and observed parts contain x_1, x_2, y_1, y_3 and z_1 in formula (3). Let $w_{ti} = \left(z_t + \mu_1\tau, x_t - z_t + (\mu_2 - \mu_1)\tau, y_t - z_t + (\mu_3 - \mu_1)\tau, n_t - x_t - y_t + z_t + (1 - \mu_2 - \mu_3 + \mu_1)\tau\right)$ and $v_{ti} = (z_t + \mu_1\tau, x_t - z_t + (\mu_2 - \mu_1)\tau, y_t - z_t + (\mu_3 - \mu_1)\tau, n_t - x_t - y_t + z_t + (1 - \mu_2 - \mu_3 + \mu_1)\tau)$ (t, i = 1 to 4)

Thus we have,

$$E\big(L(\theta_t, p_t, q_t | x_t, y_t, z_t)\big) = E\left(\sum_{i=2}^{4} L(w_{ti}, v_{ti})\right) + L(w_{t1}, v_{t1})$$

$$\mathrm{E}(\mathrm{z}_2|\mathrm{x}_2) = \mathrm{x}_2\frac{\theta}{\mathrm{p}}, \mathrm{E}(\mathrm{y}_2|\mathrm{x}_2) = \mathrm{x}_2\frac{\theta}{\mathrm{p}} + (\mathrm{n}_2 - \mathrm{x}_2)\frac{\mathrm{q}-\theta}{1-\mathrm{p}},$$

$$\mathrm{E}(\mathrm{z}_3|\mathrm{y}_3) = \mathrm{y}_3\frac{\theta}{\mathrm{q}}, \mathrm{E}(\mathrm{x}_3|\mathrm{y}_3) = \mathrm{y}_3\frac{\theta}{\mathrm{q}} + (\mathrm{n}_3 - \mathrm{y}_3)\frac{\mathrm{p}-\theta}{1-\mathrm{q}},$$

$$E(x_4) = n_4 p, E(y_4) = n_4 q, E(z_4) = n_4\theta$$

A discussion on the four different tables is necessary to figure out the missing values. In this case, it would be very difficult and tedious for us to take the log likelihood function and calculate the MLE for the three parameters because they contain unobserved data and are not easy to obtain. However, the nonignorable model is an extension of the ignorable model in section 2. Moreover, because of the conjugacy of the data model and centering model, we can directly use the expectations of the missing values in the data model to estimate the parameters. In the data model we already know the MLE's of the parameters. And we also know the expectations of all the unobserved values. Thus, we use the same method as in the expanded model:

$$\hat{\theta}_t = \frac{z_t + \mu_1\tau - 1}{n_t + \tau - 4}, \text{ t = 1 to 4.}$$

$$\hat{p}_t = \frac{x_t - z_t + (\mu_2 - \mu_1)\tau - 1}{n_t + \tau - 4} + \hat{\theta}_t, \text{ t = 1 to 4.}$$

$$\hat{q}_t = \frac{y_t - z_t + (\mu_3 - \mu_1)\tau - 1}{n_t + \tau - 4} + \hat{\theta}_t, \text{ t = 1 to 4.}$$

When t = 1, it is a complete table and there is no unobserved data. Thus, we can directly calculate the estimates for the parameters. When t = 2, y_2 and z_2 are unknown. But in section 2, we have already deduced their expectations for them. The same idea is used when t = 3 and 4.

M step:

$$\theta_2{}^{(r+1)} = \frac{z_2{}^{(r)} + \mu_1\tau - 1}{n_2 + \tau - 4} \tag{11}$$

$$q_2{}^{(r+1)} = \frac{y_2{}^{(r)} - z_2{}^{(r)} + (\mu_3 - \mu_1)\tau - 1}{n_2 + \tau - 4} + \theta^{(r)} \tag{12}$$

$$\theta_3{}^{(r+1)} = \frac{z_3{}^{(r)} + \mu_1\tau - 1}{n_3 + \tau - 4} \tag{13}$$

$$p_3{}^{(r+1)} = \frac{x_3{}^{(r)} - z_3{}^{(r)} + (\mu_2 - \mu_1)\tau - 1}{n_3 + \tau - 4} + \theta^{(r)} \tag{14}$$

$$\theta_4{}^{(r+1)} = \frac{z_4{}^{(r)} + \mu_1\tau - 1}{n_4 + \tau - 4} \tag{15}$$

$$p_4{}^{(r+1)} = \frac{x_4{}^{(r)} - z_4{}^{(r)} + (\mu_2 - \mu_1)\tau - 1}{n_4 + \tau - 4} + \theta^{(r)} \tag{16}$$

$$q_4{}^{(r+1)} = \frac{y_4{}^{(r)} - z_4{}^{(r)} + (\mu_3 - \mu_1)\tau - 1}{n_4 + \tau - 4} + \theta^{(r)} \tag{17}$$

By performing EM algorithm, we can get convergence of the estimates for θ_i, p_i and q_i (i = 2 to 4) as shown below:

Table 11. R Output for EM Algorithm

Iteration	θ_2	p_2	q_2	θ_3	p_3	q_3	θ_4	p_4	q_4
1	0.88598	0.93220	0.92226	0.87428	0.92998	0.90456	0.89337	0.94036	0.92413
2	0.88432	0.93220	0.91886	0.87587	0.93461	0.90456	0.89346	0.94188	0.92270
3	0.88419	0.93220	0.91835	0.87599	0.93553	0.90456	0.89346	0.94198	0.92260
4	0.88418	0.93220	0.91825	0.87599	0.93575	0.90456	0.89346	0.94199	0.92260
5	0.88418	0.93220	0.91823	0.87599	0.93581	0.90456	0.89347	0.94199	0.92260
6	0.88418	0.93220	0.91823	0.87599	0.93582	0.90456	0.89347	0.94199	0.92260
7	0.88418	0.93220	0.91823	0.87599	0.93582	0.90456	0.89347	0.94199	0.92260
8	0.88418	0.93220	0.91823	0.87599	0.93583	0.90456	0.89347	0.94199	0.92260
9	0.88418	0.93220	0.91823	0.87599	0.93583	0.90456	0.89347	0.94199	0.92260
10	0.88418	0.93220	0.91823	0.87599	0.93583	0.90456	0.89347	0.94199	0.92260

From the table above, we notice that p_2 and q_3 are fixed numbers. Since the only unobserved data for estimating p_2 is z_2, it is replaced by its expectation that only contain the observed data x_2. Similar reasoning holds for q_3.

After completing EM algorithm, we can obtain all the three estimates under the assumption of MNAR. They are as follows,

$$\theta = \frac{\sum_t z_t}{\sum_t n_t} = 0.8923693$$

$$p = \frac{\sum_t x_t}{\sum_t n_t} = 0.9534305$$

$$q = \frac{\sum_t y_t}{\sum_t n_t} = 0.9119347$$

Clearly, the estimates are very close to the results obtained in section 2. Recall that the true value of θ is 0.885. The MNAR result is also close to the true value.

3.2 Bootstrap Distributions of θ, p and q

As in section 2, we perform bootstrapping to obtain the distribution of θ, p and q.

We use the same settings as in Table 9,

When t = 1, $(z_{11}, z_{12}, z_{13}, z_{14})$ ~ Multinomial (n_1, θ, p-θ, q-θ, 1-p-q+θ) with $x_1 = z_{11}+z_{12}$, $y_1=z_{11}+z_{13}$.

When t = 2, x_2 ~ Binomial (n_2, p)

When t = 3, x_3 ~ Binomial (n_3, q)

By performing EM algorithm in section 3.1, we can calculate θ, p and q. After repeating 1,000 times, we obtain the 95% bootstrap confidence intervals for θ, p and q as follows:

$\theta \in (0.8807033, 0.904615)$

$p \in (0.9431182, 0.9590651)$

$q \in (0.903332, 0.9247622)$

The interval of θ contains the true value of θ and the length of the interval is almost same as that under MAR. Hence, the result from the nonignorable nonresponse model is similar to what we got based on the ignorable nonresponse model.

Figure 2-1 to 2-3 depicts the bootstrap distributions of θ, p and q.

Figure 2-1 the Bootstrap Density of θ (sizes of 2074, 500 and 100 respectively)

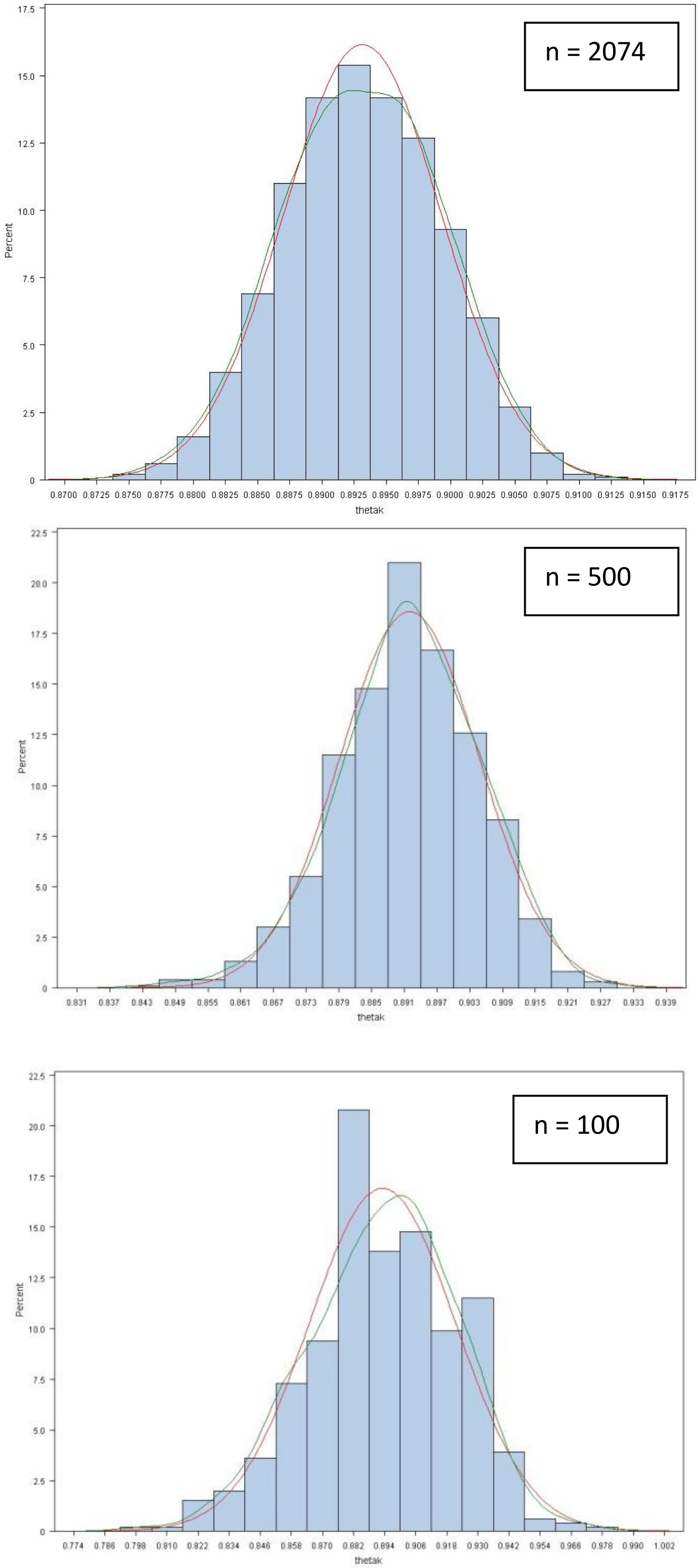


Figure 2-2 the Bootstrap Density of p (sizes of 2074, 500 and 100 respectively)

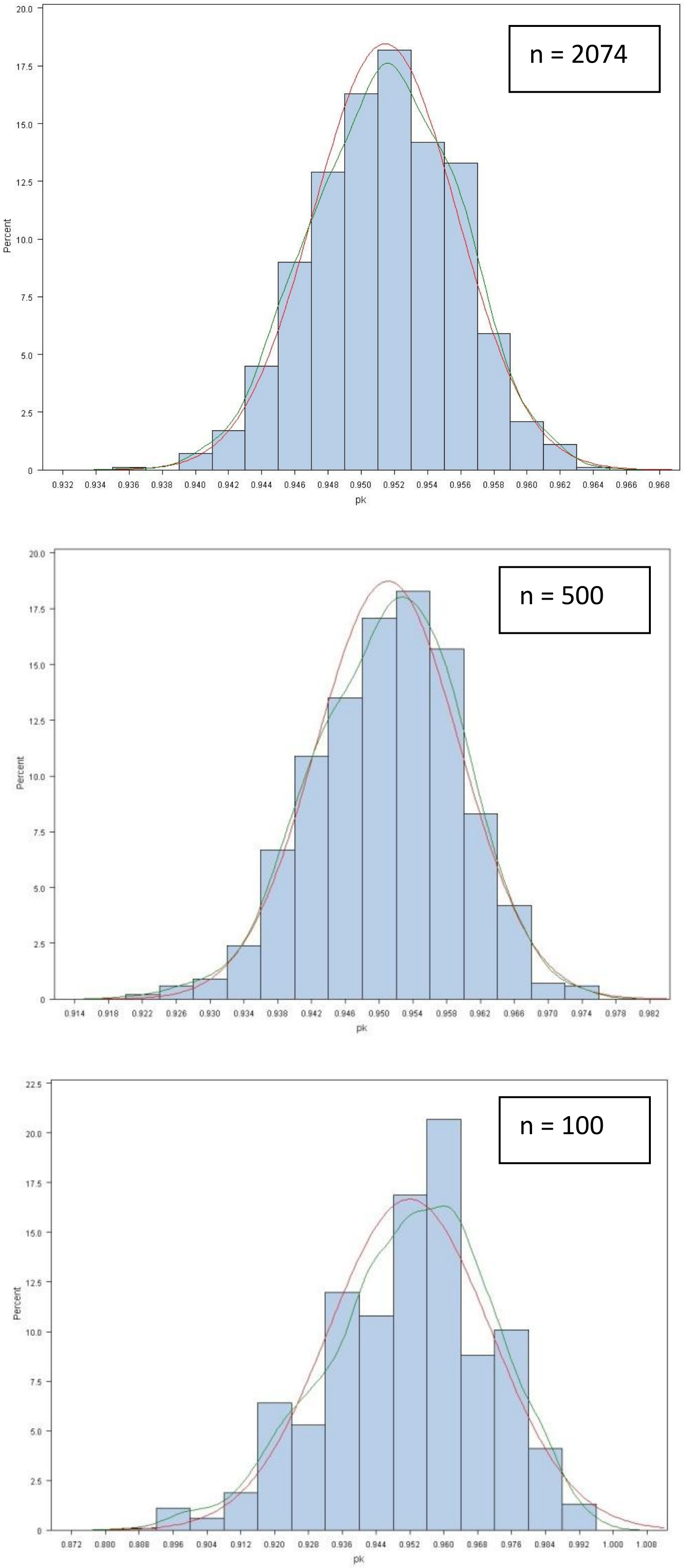
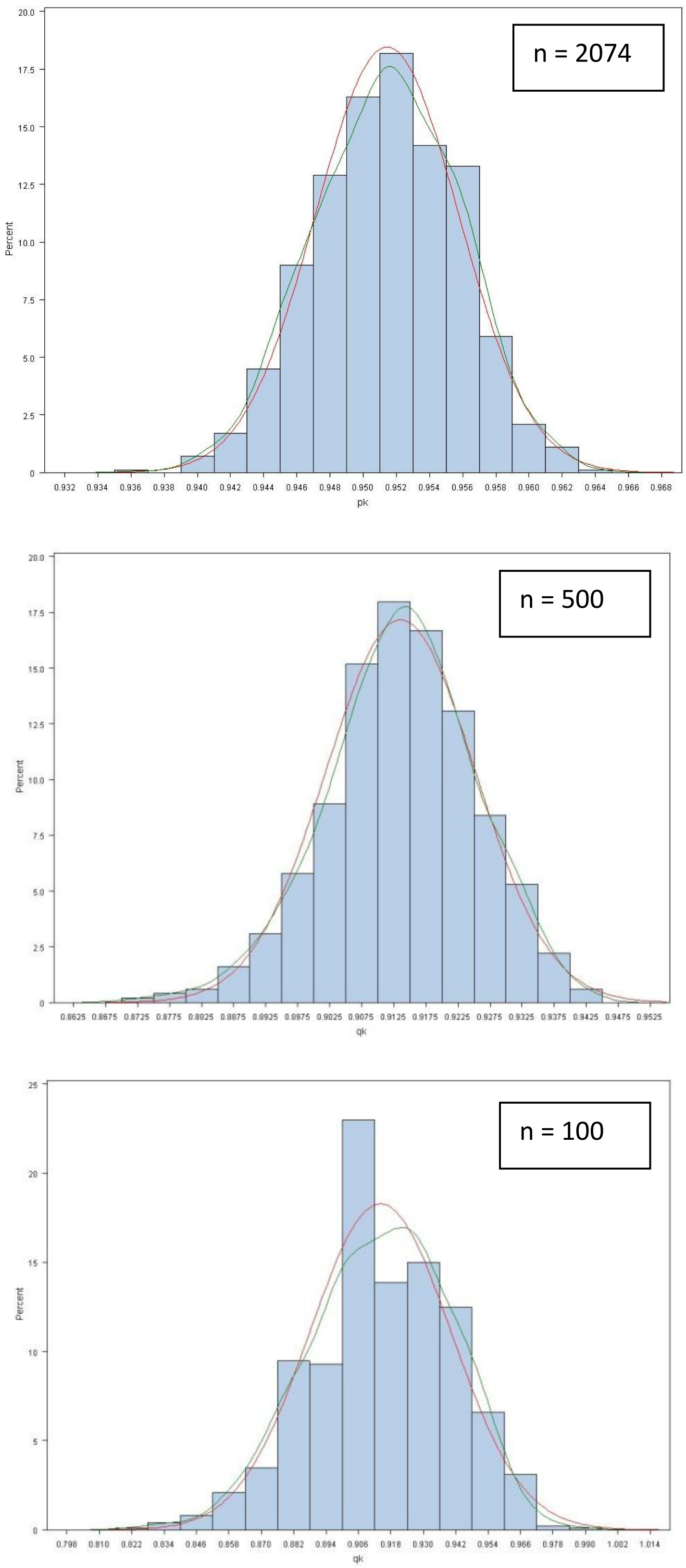


Figure 2-3 the Bootstrap Density of q (sizes of 2074, 500 and 100 respectively)



Comparing with what we got previously, the difference between kernel curves and the approximate normal curves is a little more obvious than in section 2. Once the sample size decreases, the distributions of θ, p and q are apparently different from those corresponding to a large sample size. This shows that the normality assumption may not be true when the sample size is small. Based on our research, we can thus conclude that there is no obvious difference between the nonignorable nonresponse model and the ignorable nonresponse model.

3.3 Sensitivity Analysis

As before, we perform a sensitivity analysis for the nonignorable model. In doing so, we can have a decent understanding of how sample size and completeness rate influence our parameter of interest, θ. (Note: by varying sample size, the parameter τ in the combined model should be proportional to the sample size.)

Table 12. Sensitivity Analysis for Bootstrap θ

	n	Completeness	95% Lower Band for θ	95% Upper Band for θ
1	2,000	75%	0.879864	0.904512
2		50%	0.882353	0.904638
3		25%	0.882851	0.903648
4	1,000	75%	0.874613	0.909539
5		50%	0.875523	0.908961
6		25%	0.87844	0.907566
7	500	75%	0.8689893	0.9183479
8		50%	0.8697652	0.9154328
9		25%	0.8739867	0.913455
10	100	75%	0.877563	0.94408
11		50%	0.83997	0.939344
12		25%	0.845835	0.933226

According to the table above, it is clear that sample size can indeed influence the parameter estimates. When we have large sample (n = 2,000 or 1,000), the confidence limits does not change much (the confidence limits are wider when sample size change from 2,000 to 1,000). However, for small sample size (n = 100), the confidence limits are much wider. As for the completeness rate, it changes our estimate slightly. Thus, we can infer that under MNAR assumption, sample size can significantly influence the precision of our estimate. No matter how many responders complete the survey, we can still obtain precise estimation if we have a big sample.

Another sensitivity analysis focuses on how θ change under different values of τ, μ_1, μ_2, μ_3. The parameters of the centering model are estimated by the samples from the ignorable model. The estimates of μ_1, μ_2, μ_3 are close to the estimates of θ, p and q, and we need to vary them to test the sensitivity of the model. Based on the constraints $0 < \mu_1 < \mu_2, \mu_3 <1$ and $\tau >0$, let us set the lower band of μ_1 as "a". We assume:

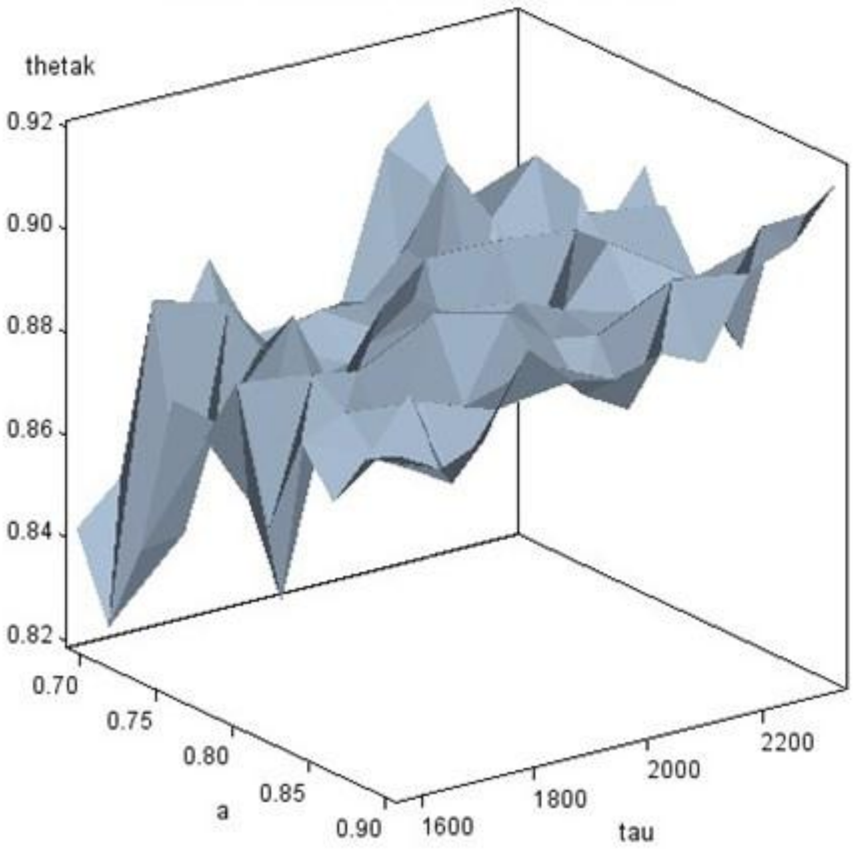
$$\mu_1 \sim Uniform(a, 1)$$

$$\mu_2 \sim Uniform(\mu_1, 1)$$

$$\mu_3 \sim Uniform(\mu_1, 1)$$

Thus, we generate samples of θ and define τ is between 1560 and 2340 (± 20% of 1950, recall τ is 1950 in section 3), "a" varies in the pessimistic – optimistic interval (.694, .905). The corresponding 3D surface plot can be developed in Figure 3.

Figure 3. 3D surface plot for θ



According to Figure 3, θ has an increasing trend as τ and “a” increases.

Table 13. Sensitivity Analysis for θ

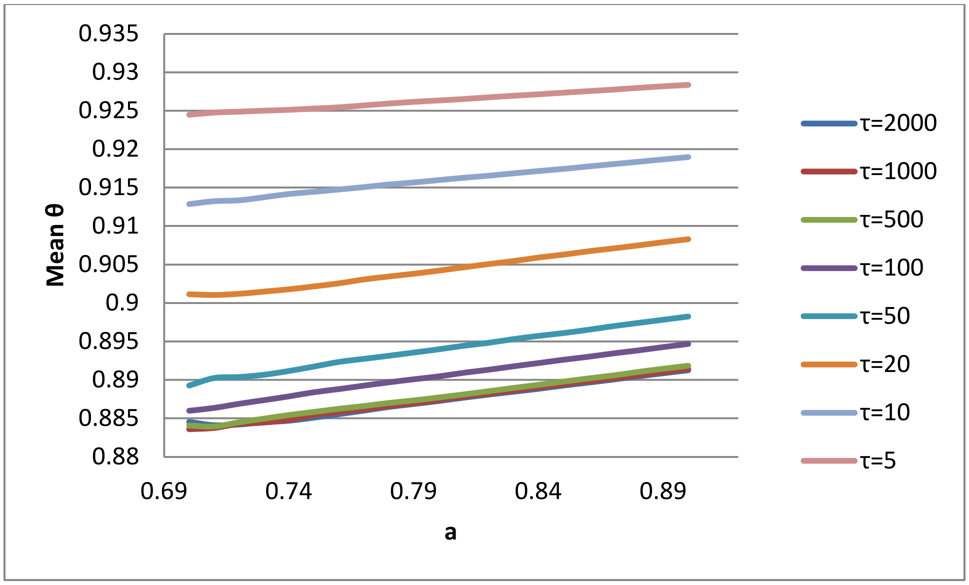
a = 0.7									
	τ	Mean	Std	θ lower	θ upper	Q1	Q2	Q3	Confidence Length
1	5	0.9240	0.0102	0.9042	0.9425	0.9168	0.9242	0.9315	0.0383
2	10	0.9126	0.0132	0.8858	0.9355	0.9039	0.9137	0.9230	0.0497
3	20	0.8990	0.0165	0.8669	0.9272	0.8872	0.9004	0.9115	0.0603
4	50	0.8904	0.0173	0.8547	0.9177	0.8781	0.8923	0.9041	0.0630
5	100	0.8862	0.0163	0.8528	0.9140	0.8745	0.8875	0.8988	0.0612
6	500	0.8836	0.0165	0.8500	0.9112	0.8718	0.8852	0.8963	0.0611
7	1000	0.8825	0.0167	0.8476	0.9104	0.8698	0.8838	0.8958	0.0628
8	2000	0.8830	0.0169	0.8497	0.9096	0.8712	0.8843	0.8962	0.0599
a = 0.75									
	τ	Mean	Std	θ lower	θ upper	Q1	Q2	Q3	Confidence Length
1	5	0.9264	0.0089	0.9088	0.9426	0.9198	0.9271	0.9327	0.0338
2	10	0.9152	0.0116	0.8918	0.9353	0.9071	0.9154	0.9239	0.0435
3	20	0.9045	0.0127	0.8781	0.9274	0.8964	0.9050	0.9136	0.0493
4	50	0.8942	0.0147	0.8659	0.9190	0.8836	0.8941	0.9057	0.0532
5	100	0.8909	0.0146	0.8621	0.9156	0.8803	0.8914	0.9024	0.0536
6	500	0.8888	0.0140	0.8590	0.9125	0.8790	0.8897	0.8996	0.0535
7	1000	0.8878	0.0142	0.8585	0.9113	0.8787	0.8887	0.8981	0.0527
8	2000	0.8866	0.0142	0.8575	0.9110	0.8765	0.8881	0.8979	0.0535
a = 0.8									
	τ	Mean	Std	θ lower	θ upper	Q1	Q2	Q3	Confidence Length
1	5	0.9289	0.0080	0.9120	0.9436	0.9236	0.9293	0.9344	0.0316
2	10	0.9187	0.0101	0.8996	0.9377	0.9118	0.9192	0.9258	0.0381
3	20	0.9081	0.0116	0.8849	0.9278	0.9001	0.9088	0.9167	0.0429
4	50	0.8985	0.0122	0.8740	0.9192	0.8895	0.8991	0.9075	0.0452
5	100	0.8950	0.0119	0.8707	0.9157	0.8863	0.8957	0.9040	0.0450
6	500	0.8920	0.0117	0.8670	0.9132	0.8844	0.8926	0.9002	0.0462
7	1000	0.8915	0.0126	0.8672	0.9135	0.8825	0.8925	0.9006	0.0463
8	2000	0.8910	0.0119	0.8674	0.9126	0.8828	0.8917	0.8995	0.0452

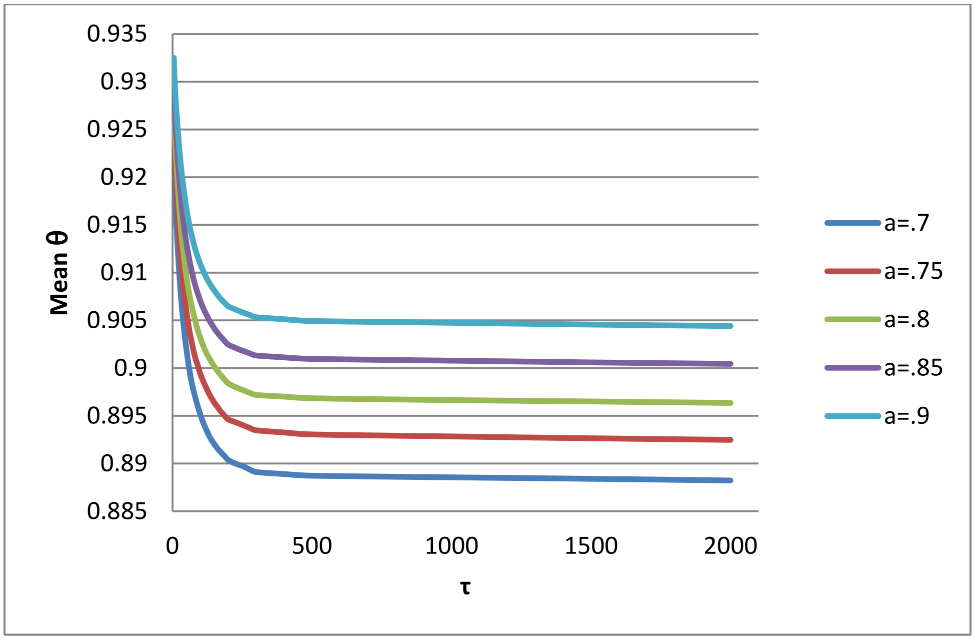
Table 14 (cont). Sensitivity Analysis for θ

a = 0.85									
	τ	Mean	Std	θ lower	θ upper	Q1	Q2	Q3	Confidence Length
1	5	0.9303	0.0073	0.9148	0.9436	0.9257	0.9304	0.9351	0.0287
2	10	0.9221	0.0086	0.9039	0.9369	0.9168	0.9225	0.9285	0.0331
3	20	0.9127	0.0098	0.8930	0.9304	0.9059	0.9132	0.9198	0.0374
4	50	0.9027	0.0097	0.8843	0.9202	0.8961	0.9030	0.9096	0.0360
5	100	0.8990	0.0098	0.8796	0.9172	0.8922	0.8989	0.9059	0.0376
6	500	0.8961	0.0098	0.8767	0.9141	0.8893	0.8963	0.9030	0.0374
7	1000	0.8954	0.0097	0.8754	0.9126	0.8887	0.8958	0.9020	0.0372
8	2000	0.8952	0.0099	0.8754	0.9129	0.8880	0.8956	0.9022	0.0375
a = 0.9									
	τ	Mean	Std	θ lower	θ upper	Q1	Q2	Q3	Confidence Length
1	5	0.9323	0.0071	0.9182	0.9464	0.9277	0.9324	0.9369	0.0282
2	10	0.9251	0.0075	0.9104	0.9390	0.9201	0.9253	0.9303	0.0287
3	20	0.9158	0.0077	0.9009	0.9302	0.9104	0.9162	0.9209	0.0293
4	50	0.9067	0.0083	0.8896	0.9223	0.9011	0.9069	0.9126	0.0327
5	100	0.9027	0.0083	0.8859	0.9178	0.8969	0.9030	0.9087	0.0318
6	500	0.8999	0.0080	0.8841	0.9145	0.8945	0.9003	0.9054	0.0304
7	1000	0.8994	0.0077	0.8846	0.9135	0.8939	0.8997	0.9052	0.0288
8	2000	0.8989	0.0078	0.8819	0.9130	0.8939	0.8990	0.9043	0.0311

According to Table 13, we find 1) when “a” increases, the mean of θ slightly increases and the standard deviation decreases. When τ is small (less than 50), the mean and standard deviation of θ are sensible. 2) “a” can negatively influence the confidence length. However, τ does not influence θ much except for small values. Since the variability of $\mu_{1,}$ μ_2 and μ_3 decreases as “a” increases, we will obtain more concentrated values of θ based on the combined model. 3) The quartiles of θ increase when “a” increases and τ decreases.

Figure 4. 2D plots for mean θ





In order to get a clear idea for how θ is influenced by τ and a, we make two – dimensional plots for θ. All the data points in Figure 4 are based on the mean of 1,000 bootstrap for θ under different sets of value for τ and a.

According to the first plot in Figure 4, we can infer that 1) a can positively influence θ under a fix value of τ. 2) There is a negative relationship between τ and θ and θ is sensitive when τ is less than 100.

According to the second plot in Figure 4, we can conclude that 1) τ can negatively influence θ under a fixed value of a. Also θ is very sensible when τ is small. 2) There is a positive relationship between θ and a. 3) θ is very sensitive when τ is small.

The two plots in Figure 4 give same conclusion. By performing sensitivity analysis in Table 13 and Figure4, we conclude that there is a positive relationship between a and θ and a negative relationship between τ and θ. Moreover, θ is fairly stable no matter how τ changes in large number. This means there is no difference between using τ = 500 and τ = 2,000. We can save cost of prior information.

Chapter 4. Discussion

In this paper we have discussed the concepts of MAR and MNAR and applied them with the familiar idea of statistical imprecision, producing the measure for θ. As an extension of the concept of confidence, our measures are expressed as the intervals for scalar parameters (θ, p and q). These reduce to conventional confidence intervals when it is assumed that there is MNAR about the statistical model underlying the data. The construction of the intervals in the plebiscite case is seen to convey useful information about the problem concerned.

We have introduced three paths to analyze this problem. The first is to look at the bounds produced by the most pessimistic and most optimistic scenarios. In the case of the Slovenian plebiscite, we learn that even the most pessimistic scenario translates into a clear majority in support of independence. Secondly, under MAR assumption we produce an innovative way to calculate our parameter of interest. We successfully model one cell and two margins rather than a standard multinomial model. Meanwhile, based on Hamdan (1969) and Kocherlakota (1989), we develop the model from a bivariate binomial which has three parameters and two variables to a more suitable model which has three parameters and three variables and we deduce its useful properties. EM algorithm and bootstrap prove that we obtain precise estimation which is close to the official result. Thirdly, the plausible and flexible nonignorable nonresponse model can be considered under MNAR assumption. We introduce an additional centering model based on ignorable model. The results are almost the same as MAR's results. It is true that the final value will not be known and such a confirmatory check is not possible. However, the method presented here enables a consideration of model selection, and the amount of parsimony can be controlled. The sensitivity analysis proves that model selection is essential for MNAR case and it gives fairly robust results, regardless of the variability of the parameters.

On the other hand, we cannot tell from the data at hand whether the missing observations are MCAR, NMAR or MAR (although sometimes we can distinguish between MCAR and MAR). In the MNAR setting it is very difficult to accurately pinpoint the appropriate model for the missingness mechanism. In this paper we have defined the concepts of missing data and use them in different models. Any of the three assumptions is difficult to verify in any situation, but it is quite popular for statisticians to begin with simplifying assumptions and analyze whether inferences are likely to change substantially as the assumptions are modified. In fact, for large well–performed surveys, MAR is often considered a reasonable starting point for statistical analysis (Little, 1988).

In our work, the added work involved the construction of a conjugate and plausible MNAR model which also gave accurate inferences. Moreover, because of the specified prior information and the robustness of the MNAR model, it can reduce the cost of collecting prior information and doing additional follow up. In many surveys the cost of additional field work can be enormous; in complex surveys the cost can be in millions.

For example, if you plan to launch a marketing research project, the general procedure would be as follows:

1) Define the business problem;
2) Prior research such as key personnel interview and focus group;
3) Questionnaire design and test;
4) Data collection;
5) Analysis;

Before running the project, it is a good idea to develop a well–designed nonignorable model and you would have a clear idea about how much prior information you need before starting your survey. You may save a lot of cost in the first three steps above if the nonignorable model fit well.

Of course, in many studies the true value will not be known and such a checking cannot be done. However, the strategies in this paper enable consideration of a general nonignorable nonresponse model and the variability of parameters, sample size and even completeness rate can be controlled.

References

Baker, S. G., and Laird, N. M. (1988), "Regression Analysis for Categorical Variables With outcome subject to nonignorable nonresponse," Journal of the American Statistical Association, 83, 62-69.

Biswas, A. and Hwang, J.S. (2002) “A new bivariate binomial distribution”, vol. 60, issue 2, pages 231-240.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), Journal of the Royal Statistical Society, 39, 1-38.

Efron, B. & Tibshirani, R. (1993), “An Introduction to the Bootstrap”. Boca Raton, FL: Chapman & Hall/CRC.

Fuchs, C. (1982), "Maximum Likelihood Estimation and Model Selection in Contingency Tables With Missing Data," Journal of the American Statistical Association, 77, 270-278.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), Bayesian Data Analysis (second edition), CRC Press.

Hamdan, M. A. and Martinson, E. O. (1971), “Maximum Likelihood Estimation in the Bivariate Binomial (0, 1) Distribution: Application to 2×2 Tables”, Australian Journal of Statistics, 13: 154–158.

Hamdan, M. A. (1972), “Canonical Expansion of the Bivariate Binomial Distribution with Unequal Marginal Indices”, International Statistical Review / Revue Internationale de Statistique, Vol. 40, No. 3, 277-280.

Kocherlakota, S. (1989), “A note on the bivariate binomial distribution”, Statistics & Probability Letters, Vol. 8, Issue 1, Pages 21-24.

Little, R. J. A. and Rubin, D. B. (2002), Statistical Analysis with Missing Data, Second Edition, New York: John Wiley & Sons.

Molenberghs, G., Kenward, M., and Goetghebeur, E. (2001), “Sensitivity Analysis for Incomplete Contingency Tables: the Slovenian Plebiscite Case”, Applied Statistics 50, Part 1, 15 – 29.

Madow, W. G., Nisselson, H., and Olkin,I . (eds.)(1983), Incomplete Data in Sample Surveys, Vol. 1, London: Academic Press.

Rubin, D. B. (1976), "Inference and Missing Data", Biometrika, 63, 581- 590.

Rubin, D. B., Stern, H. S. and Vehovar, V. (1995) “Handing ‘Don’t know’ survey responses: the case of the Slovenian plebiscite”, Journal of the American Statistical Association, 90, 822 – 828.