Sense and sensitivity when intended data are missing

Els Goetghebeur[1] Geert Molenberghs[2] Michael G. Kenward[3] *

Abstract

Classical inferential procedures infer conclusions from a set of data to a population of interest, accounting for the imprecision implied by the designed sampling frame. Less attention is devoted to uncertainty arising from unintended incompleteness in the data. Through the choice of an identifiable model for (non-ignorable) non-response, one narrows the possible data generating mechanisms to the point where inference only suffers from imprecision, which typically reduces to zero as the sample size tends to infinity. Some proposals have been made for assessment of sensitivity to these modeling assumptions; many are based on fitting several plausible but competing models. We develop an alternative approach which identifies and incorporates explicitly both sources of *uncertainty* in inference: *imprecision* due to finite sampling and *ignorance* due to incompleteness. The introduction of sensitivity parameters helps inspection of the whole set of estimators compatible with the observed data, in function of more or less plausible assumptions about which the data carry no information. The developments in this paper focus on contingency tables, and are illustrated using data from an HIV prevalence study and data from a Slovenian plebiscite.

Key words: Contingency Table; Imprecision, Missing At Random; Overspecified Model; Saturated Model

1 Introduction

The problem of missing intended data in a well designed study is a common one. The reasons for data being missing are many and varied. There is therefore no straightforward approach to statistical inference that accommodates the unknown behavior of unintentionally unobserved data. Rubin (1976) provided one of the first systematic studies of this issue, and we use his terminology for classifying different classes of processes that give rise to missing values. The abbreviations MCAR, MAR and MNAR will refer to missingness processes that are Completely At Random, At Random and Not At Random (i.e. nonignorable) respectively. This paper is concerned with how one might approach inference when the possibility of a non-random missingness process cannot be ruled out on *a priori* grounds.

*[1] Universiteit Gent, TWI, Krijgslaan 281-S9, B-9000 Gent, Belgium; [2] Limburgs Universitair Centrum, Universitaire Campus Building D, B-3590 Diepenbeek, Belgium; University of Kent, Canterbury, U.K. The authors gratefully acknowledge support from the FWO-Vlaanderen Research Project "Sensitivity Analysis for Incomplete and Coarse Data"

One approach is to formulate models that describe the complete data and additionally model the probability of a specific response pattern conditional on the possibly unobserved value of the complete outcome. For such models, the likelihood of the observed data is obtained by integrating the complete data likelihood over the distribution of the missing data. Provided the model is identified, conventional inference proceeds from there (see Little, 1995, for a review). Theoretical identifiability alone does not protect against practical problems however. Likelihoods can be very flat or multimodal indicating little information on specific parameter(s). More alarmingly perhaps, models with an entirely different interpretation at the complete data level can show the same or similar deviance. At the same time, the intrinsic nature of missing information implies that modeling assumptions cannot fully be examined using the data alone. Thus a characteristic form of ambiguity remains. A natural way to proceed is to acknowledge the range of inferences that are consistent with what is observed, to perform a sensitivity analysis.

While there is an established formal framework for imprecision (variance, standard errors, sampling distributions, confidence intervals, hypothesis tests and so on) most implementations of statistical sensitivity analysis have remained *ad hoc.* In this paper we propose a framework in which general sensitivity concepts can be formalized and develop it from the frequentist perspective, focusing on categorical data. To this end, we describe *ignorance* (due to incompleteness of the data) in addition to the familiar *imprecision* (due to finite sampling) and combine both into *uncertainty.*

In the next section two sets of data are presented that will serve as examples throughout. In Section 3 the concepts of *imprecision*, *ignorance* and *uncertainty* are introduced. In Section 4 we find that besides enumeration of all possible completed data sets, sensitivity parameters can bring added insight into the uncertainty and allow for a natural expression of maximum likelihood estimates from overspecified likelihoods. In Section 5, the data on the Slovenian plebiscite are analyzed using the new tools.

2 Examples

2.1 HIV Prevalence estimation in Kenya

Example 1 *In the context of disease monitoring, HIV prevalence is estimated in a population of pregnant women in Kenya.*

To this end a sample of 787 Kenyan women were tested, with the following result:

Known HIV +	=	52
Known HIV -	=	699
Unknown HIV Status	=	36

What we learn is that between 52 and 52+36 out of the 787 observed women had a HIV + test result. Hence we can produce a best (and worst) case estimator L=6.6% (R=11.2%) for the probability of a positive HIV test result. This is the basic form of sensitivity analysis. We explore it in more depth in the following sections.

2.2 The Slovenian Plebiscite

Example 2 *Rubin, Stern, and Vehovar (1995) studied data from a plebiscite organized in Slovenia on secession from the former Yugoslavia, in which the Slovenians overwhelmingly voted for independence.*

It was deemed useful to anticipate results of the plebiscite from additional questions in the Slovenian Public Opinion (SPO) Survey, carried out four weeks prior. The main questions were: (1) Are you in favor of Slovenian independence? (2) Are you in favor of Slovenia's secession from Yugoslavia? (3) Will you attend the plebiscite? Questions (1) and (2) are different since independence would have been possible in confederal form as well. Question (3) is highly relevant since not attending was treated as an effective NO to question (1). The data are presented in Table 1. Full details on the study are provided in Rubin, Stern, and Vehovar (1995), who considered a number of identified ignorable and non-ignorable models with varying conclusions. The ignorable models turned out to outperform the non-ignorable one in the sense that they were much closer to the results of the (true) plebiscite. Rubin, Stern and Vehovar saw this as an argument to be more generally in favor of MAR models. To gain more insight into the available information we feel a sensitivity analysis is called for (Kenward 1998).

3 Imprecision, Ignorance and Uncertainty

To distinguish formally between *statistical imprecision* which is due to sampling variation, and *statistical ignorance*, which is due to the incompleteness of the observations, we consider the simple

Table 1: Data form the Slovenian Public Opinion Survey.

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

Example 1 first. Let Y be the binary outcome (1 for a positive test result, 2 for a negative one), and R indicate observed response (1) or not (0). The complete data model consists of the prevalence $\pi = \text{Prob}(Y = 1)$ and two conditional probabilities η_i of being observed, given infected ($i = 1$) or not ($i = 2$).

Consider N_α 'observed cases' ($(Y, R) = (1, 1)$) and N_β observed controls ($(Y, R) = (2, 1)$) along with N_γ observations with missing outcomes. The theoretical probability of falling into each of those three respective categories is denoted $\alpha, \beta, \gamma = 1 - \alpha - \beta$ respectively. Besides the 3 observable outcomes, we consider 4 completed (full data) outcome categories for R and Y as shown in Table 2. Under the sampling scheme that generated the data but with an infinite sample size, we can observe α, β and γ but no more. In the limit we thus learn that $\pi \in [\alpha, \alpha + \gamma]$, but no further specific localization of π. The limiting interval for π will be called the *interval of ignorance (II)* on π. in a similar fashion, regions of ignorance would appear for multidimensional parameters, like for (π, η_1) for example.

A finite sample data set thus provides information on this limiting interval rather than on the exact position of π. Generally, we consider an *estimated region of ignorance* to be the set of estimators derived from all possible completed tables collapsing back to the observed incomplete

Table 2: Incomplete and Complete Data of Example 1.

	Observed		Latent	
	Incomplete		Full	
	Success	Failure	Success	Failure
Observed	N_α	N_β	N_α	N_β
Unobserved	N_γ		$N_{\gamma\alpha}$	$N_\gamma - N_{\gamma\alpha}$
Observed	$\alpha = \pi\eta_1$	$\beta = (1-\pi)\eta_2$	$\pi\eta_1$	$(1-\pi)\eta_2$
Unobserved	$\gamma = 1 - \pi\eta_1 - (1-\pi)\eta_2$		$\pi(1-\eta_1)$	$(1-\pi)(1-\eta_2)$

table. Enumeration of all such tables can be done in terms of a single parameter for this example. Let the number of successes out of the N_γ missing observations be denoted by $N_{\gamma\alpha}$, then the corresponding number of failures is $N_\gamma - N_{\gamma\alpha}$, where $N_{\gamma\alpha}$ can take values in the interval $[0, N_\gamma]$. The compatible full data tables are depicted in Table 2.

The estimated interval of ignorance on π is the set of estimators $\hat{\pi}(N_{\gamma\alpha})$ generated by each of the completed tables, and is given by the interval $[\frac{N_\alpha}{N}, \frac{N_\alpha+N_\gamma}{N}]$. To acknowledge imprecision on the estimated interval, we consider the left and right hand limit of the 95% confidence interval on $\hat{\pi}(N_{\gamma\alpha} = 0)$ and $\hat{\pi}(N_{\gamma\alpha} = N_\gamma)$, respectively. This yields an uncertainty interval approximated by $\left[\frac{N_\alpha}{N} - 1.96\sqrt{\frac{N_\alpha(N_\beta+N_\gamma)}{N^3}}, \frac{N_\alpha+N_\gamma}{N} + 1.96\sqrt{\frac{(N_\alpha+N_\gamma)N_\beta}{N^3}}\right]$. For finite N the uncertainty interval is larger than the estimated interval of ignorance. In the limit, the interval of uncertainty and the estimated ignorance interval both converge to the true interval of ignorance. In our approach the former intervals take the place of the traditional point estimator and confidence interval, respectively.

Another way of retrieving an estimated interval of ignorance, is by considering the observed data likelihood in terms of the most general model one is prepared to consider for complete data + missingness mechanism. For instance, Table 3 shows in the second column the saturated model, called M_{SAT}, in terms of parameters (π, η_1, η_2). In the first column it presents two parameters (π, η) for an identified model that satisfies the missing at random constraint (MAR=MCAR in this simple situation). In both cases it gives the complete data maximum likelihood estimate. The log observed-likelihood for Model M_{SAT} in terms of (π, η_1, η_2) is:

$$\ell = N_\alpha \ln(\pi\eta_1) + N_\beta \ln[(1-\pi)\eta_2] + N_\gamma \ln[\pi(1-\eta_1) + (1-\pi)(1-\eta_2)]. \qquad (3.1)$$

Table 3: Two Reparameterizations of the Complete-Data Likelihood.

Model M_0: MAR≡MCAR	**Model M_{sat}: MNAR**
Parameterization:	
$\alpha = \pi\eta$	$\alpha = \pi\eta_1$
$\beta = (1-\pi)\eta$	$\beta = (1-\pi)\eta_2$
$\gamma_1 = \pi(1-\eta)$	$\gamma_1 = \pi(1-\eta_1)$
$\gamma_2 = (1-\pi)(1-\eta)$	$\gamma_2 = (1-\pi)(1-\eta_2)$
Solution:	
$\hat{\pi}_{\gamma\alpha} = \frac{N_\alpha + N_{\gamma\alpha}}{N}$	$\hat{\pi}_{\gamma\alpha} = \frac{N_\alpha + N_{\gamma\alpha}}{N}$
$\hat{\eta}_{\gamma\alpha} = \frac{N_\alpha + N_\beta}{N}$	$\hat{\eta}_{1\gamma\alpha} = \frac{N_\alpha}{N_\alpha + N_{\gamma\alpha}}$
	$\hat{\eta}_{2\gamma\alpha} = \frac{N_\beta}{N_\beta + N_\gamma - N_{\gamma\alpha}}$
$N_{\gamma\alpha} \in [0, N_\gamma]$	$N_{\gamma\alpha} \in [0, N_\gamma]$

Typically, maximum likelihood estimators are found by solving the score equations:

$$\frac{N_\alpha}{\pi} - \frac{N_\beta}{1-\pi} = \frac{N_\gamma(\eta_1 - \eta_2)}{\pi(1-\eta_1) + (1-\pi)(1-\eta_2)}, \tag{3.2}$$

$$\frac{N_\alpha}{\eta_1} = \frac{N_\gamma \pi}{\pi(1-\eta_1) + (1-\pi)(1-\eta_2)}, \tag{3.3}$$

$$\frac{N_\beta}{\eta_2} = \frac{N_\gamma(1-\pi)}{\pi(1-\eta_1) + (1-\pi)(1-\eta_2)}. \tag{3.4}$$

Here, these equations are linearly dependent as $(3.3) \times \eta_1/\pi - (3.4) \times \eta_2/(1-\pi) = (3.2)$, and have an infinite set of solutions.

Generally, one can remove the overspecification from a likelihood expressed in terms of a parameter vector $\boldsymbol{\theta}$ by possibly reparameterizing first, and considering a minimal set of new parameters $\boldsymbol{\lambda}$, conditional upon which the remaining ones, $\boldsymbol{\mu}$, are identified. We term $\boldsymbol{\lambda}$ a sensitivity parameter and $\boldsymbol{\mu}$ the estimable parameter. Thus each value of $\boldsymbol{\lambda}$ produces an estimate $\hat{\boldsymbol{\mu}}(\boldsymbol{\lambda})$, and the union of these produces the estimated region of ignorance. A natural estimator for the region of uncertainty is then the union of confidence regions around each $\hat{\boldsymbol{\mu}}(\boldsymbol{\lambda})$ whilst $\boldsymbol{\lambda}$ varies over the allowable range.

The choice of sensitivity parameter is non-unique and deserves some thought. It may be based on mathematical convenience, ease of interpretation of $\boldsymbol{\lambda}$ or availability of external sources of

Table 4: Two Sensitivity Parameterizations for the Observed Data Likelihood.

Parameterization III	**Parameterization IV**
Parameterization:	
$\alpha = p\eta$	$\alpha = p\eta$
$\beta = (1-p)(\eta+\theta)$	$\beta = (1-p)\eta\lambda$
$\gamma = 1 - p\eta - (1-p)(\eta+\theta)$	$\gamma = 1 - p\eta - (1-p)\eta\lambda$

information on $\boldsymbol{\lambda}$. Sometimes one may choose to let the sensitivity set overlap with the parameter of direct scientific interest (see White and Goetghebeur 1997). The particular choice of sensitivity parameter does not affect the estimated region of ignorance on the complete set of parameters, when the sensitivity parameter is varied over the 'allowable range'. However, in our approach there is no direct estimate of imprecision available for the sensitivity parameter. The region of uncertainty is built from confidence regions conditional on a particular value of the sensitivity parameter, it will typically vary with the choice made. For instance, in model IV of Table 4 one may consider as sensitivity parameter $\lambda = \frac{\eta_2}{\eta_1}$, the relative risk of responding for nondiseased versus diseased subjects. The value $\lambda = 1$ corresponds to MAR whilst $\lambda = 2$ expresses that sick people are twice as likely as healthy people to have an outcome observed. λ is theoretically bounded by α, β and γ. Scientists in the field may well be able to formulate more stringent plausible bounds on λ based on their experience. The sensitivity analysis then consists of interpreting results conditionally on any of those λ values. Alternatively, a more detailed prior distribution of beliefs could be elicited on the relative response rates. Those help to attach differential weights to the different $\lambda-$based estimators. A Bayesian analysis might want to average the estimated π over the prior $\lambda-$distribution. We would prefer not to do that but rather consider the possible range of estimators keeping in mind their plausibility.

4 The Sensitivity Approach to Ignorance and Uncertainty

Consider two additional parameterizations of (3.1), as in Table 4. In Models III and IV we view θ and λ as sensitivity parameters respectively. The maximum likelihood estimators for π and η, given a value of the sensitivity parameter, will be subscripted by the sensitivity parameter. With some algebra, Model III is seen to imply $\hat{\pi}_\theta = N_\alpha/(N\hat{\eta}_\theta)$ and $\hat{\eta}_\theta$ is found as the valid solution

$$\hat{\eta}_\theta = \frac{1}{2}\left[\frac{N_\gamma}{N} - \theta \pm \sqrt{\left(\theta - \frac{N_\gamma}{N}\right)^2 - 4\frac{N_\alpha}{N}}\right].$$

Calculations are quickly getting cumbersome and therefore, rather than pursuing this approach, we turn attention to Model IV. Some algebra yields simple expressions for $\hat{\pi}_\lambda$ and $\hat{\eta}_\lambda$:

$$\hat{\pi}_\lambda = \frac{\hat{\alpha}\lambda}{\hat{\beta} + \hat{\alpha}\lambda} = \frac{\lambda N_\alpha}{N_\beta + N_\alpha\lambda}, \tag{4.5}$$

$$\hat{\eta}_\lambda = \frac{\hat{\beta} + \hat{\alpha}\lambda}{\lambda} = \frac{N_\beta + N_\alpha\lambda}{N\lambda}. \tag{4.6}$$

Using the delta method, an asymptotic variance-covariance matrix of $(\hat{\pi}_\lambda\ , \hat{\eta}_\lambda\ )$ can be found, for instance:

$$\widehat{\text{Var}}(\widehat{\pi_\lambda}\ ) = \frac{\hat{\pi}_\lambda\ (1 - \hat{\pi}_\lambda\ )}{N\lambda\hat{\eta}_\lambda}\left\{1 + \frac{1-\lambda}{\lambda}(1 - \hat{\pi}_\lambda\ )[1 - \hat{\pi}_\lambda\ \hat{\eta}_\lambda\ (1-\lambda)]\right\}. \tag{4.7}$$

The remaining elements of that matrix and similar expressions for the estimated variance of the prevalence odds and logit are given in an appendix. The parameter estimates are asymptotically correlated, except when $\lambda = 1$, i.e., under the MAR assumption, or under boundary values ($\pi_\lambda\ = 0, 1; \eta_\lambda\ = 0$). This is in line with the ignorable nature of the MAR model (Rubin, 1976).

The constraints $0 \leq \hat{\pi}_\lambda, \hat{\eta}_\lambda, \lambda\hat{\eta}_\lambda \leq 1$ imply a set of allowable values for λ :

$$\lambda \in \left[\frac{N_\beta}{N_\beta + N_\gamma}, \frac{N - N_\beta}{N_\alpha}\right].$$

Clearly, $\lambda = 1$ is always valid. For the HIV example the range equals $\lambda \in [0.951; 1.692]$.

Table 5 presents estimates for limiting cases. The interval of ignorance for the success probability is thus seen to be as in Table 3. It is interesting to observe that the estimated success odds is linear in the sensitivity parameter and its estimated interval of ignorance states:

$$\text{odds}(\hat{\pi}) \in \left[\frac{N_\alpha}{N_\beta + N_\gamma}, \frac{N_\gamma + N_\alpha}{N_\beta}\right].$$

For each chosen λ, a confidence interval C_λ can be constructed for either π_λ , its odds, or its logit, the union of the C_λ forms then the *interval of uncertainty* on the corresponding parameter. For the prevalence data set, these intervals along with point estimates in function of the sensitivity

Table 5: Limiting Cases for the Sensitivity Parameter Analysis.

Estimator	λ	$\lambda = \frac{N_\beta}{N-N_\alpha}$	$\lambda = 1$	$\lambda = \frac{N-N_\beta}{N_\alpha}$
$\hat{\pi}_\lambda$	$\frac{\lambda N_\alpha}{N-N_\gamma-N_\alpha(1-\lambda)}$	$\frac{N_\alpha}{N}$	$\frac{N_\alpha}{N-N_\gamma}$	$\frac{N_\gamma+N_\alpha}{N}$
$\hat{\eta}_\lambda$	$\frac{N-N_\gamma-N_\alpha(1-\lambda)}{N\lambda}$	1	$\frac{N_\alpha+N_\beta}{N}$	$\frac{N_\alpha}{N-N_\beta}$
$\hat{\eta}_\lambda\ \lambda$	$\frac{N-N_\gamma-N_\alpha(1-\lambda)}{N}$	$\frac{N_\beta}{N-N_\alpha}$	$\frac{N_\alpha+N_\beta}{N}$	1
$\frac{\hat{\pi}_\lambda}{1-\hat{\pi}_\lambda}$	$\lambda\frac{N_\alpha}{N_\beta}$	$\frac{N_\alpha}{N-N_\alpha}$	$\frac{N_\alpha}{N_\beta}$	$\frac{N-N_\beta}{N_\beta}$

Table 6: Estimates of the proportion θ attending the plebiscite and voting for independence, as presented in Rubin, Stern, and Vehovar (1995).

		NO via
Estimation method	θ	nonattendance
Conservative	0.694	0.192
Complete cases	0.928	0.020
Available cases	0.929	0.021
MAR (2 questions)	0.892	0.042
MAR (3 questions)	0.883	0.043
Non-ignorable	0.782	0.122
Plebiscite	0.885	0.065

parameter λ are shown in Figure 1. Remember that $\lambda = 1$ refers to MAR. Larger values than 1 barely shift the point estimate for π but a larger probability of missingness for the sick people can have a drastic effect on our prevalence estimation. Alternatively, one may chose to plot logit(λ) on the x-axis. If desired, a confidence ellipsoid could be built around $(\pi, \eta_\lambda\ )$.

5 Analysis of the Slovenian Plebiscite

Rubin, Stern, and Vehovar (1995) conducted several analyses of this data set. Their main emphasis was on determining the proportion θ of the population that would attend the plebiscite and vote for independence. The three other combinations of both binary outcomes would be treated as voting "no". Their estimates are reproduced in Table 6. The conservative method is the ratio of the (yes,

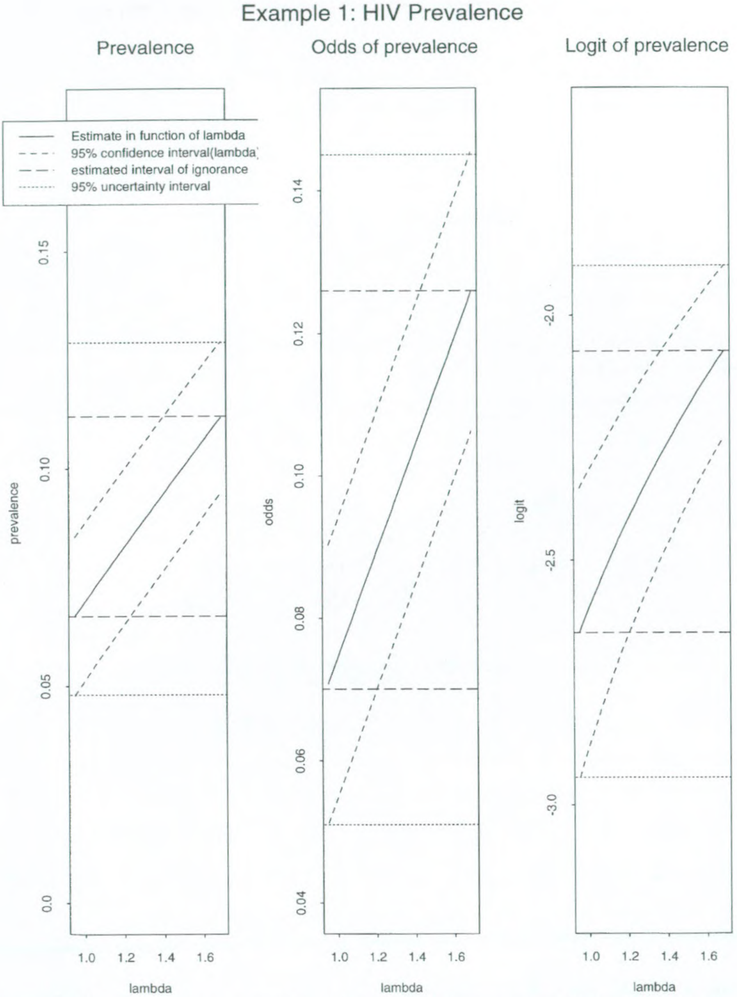

Figure 1: Graphical Representation of Interval of Ignorance and Interval of Uncertainty

yes) answers to the (attendance, independence) pair and the total sample, i.e., 1439/2074. This is the most pessimistic scenario. At the opposite end of the spectrum, we can add the most optimistic estimate that replaces the numerator by all who are no definite "no":

$$\frac{1439 + 159 + 144 + 136}{2074} = \frac{1878}{2074} = 0.905.$$

Both estimates together yield the range $II = [0.694; 0.905]$. The complete case estimate is based on the subjects answering all three questions and the available case estimate is based on the subjects answering the two questions of interest here. It is noteworthy that both estimates fall outside the interval of ignorance and should be disregarded as they fall outside the interval produced by all of the completed data sets.

There are two MAR models, the first one based on two questions only, the second one using all three. The non-ignorable model is based on the assumption that missingness on a question depends on the answer to that question but not on the other questions. The authors argue that the MAR results are very close to the true response, unlike the non-ignorable model, and suggest that the MAR assumption is generally more plausible in carefully designed surveys. To illustrate a number of sensitivity issues, we examine a range of fitted models of Baker, Rosenberger, and DerSimonian (1992). Results are presented in Table 7, we introduce the models and the issues below.

5.1 General Missingness Patterns in the 2x2 table

When 4 possible patterns of missingness occur, the full data comprise 15 degrees of freedom, while there are only 8 observed degrees of freedom. An interesting class of models has been proposed by Baker, Rosenberger, and DerSimonian (1992). It is based on log-linear models for the four-way classification of both outcomes, together with their respective missingness indicators. Denote the counts by $Y_{r_1 r_2 jk}$ where $r_1, r_2 = 0, 1$ indicate whether a measurement is missing or taken for variables 1 and 2 respectively, and $j, k = 1, 2$ indicate the response categories for both outcomes. The models are written as:

$$\begin{aligned} E(Y_{11jk}) &= m_{jk}, & E(Y_{01jk}) &= m_{jk}\alpha_{jk}, \\ E(Y_{10jk}) &= m_{jk}\beta_{jk}, & E(Y_{00jk}) &= m_{jk}\alpha_{jk}\beta_{jk}\gamma, \end{aligned}$$

with $m_{jk} = Y_{++++}\pi_{11jk}$, and

$$\alpha_{jk} = \frac{q_{01|jk}}{q_{11|jk}}, \qquad \beta_{jk} = \frac{q_{10|jk}}{q_{11|jk}}, \qquad \gamma = \frac{q_{11|jk}q_{00|jk}}{q_{10|jk}q_{01|jk}}.$$

The subscripts are missing from γ since Baker *et al* have shown that this quantity is independent of j and k. These authors consider nine identifiable models, based on setting α_{jk} and β_{jk} constant in one or more indices:

$$\begin{array}{lll} \text{BRD1} : (\alpha,\beta) & \text{BRD4} : (\alpha,\beta_k) & \text{BRD7} : (\alpha_k,\beta_k) \\ \text{BRD2} : (\alpha,\beta_j) & \text{BRD5} : (\alpha_j,\beta) & \text{BRD8} : (\alpha_j,\beta_k) \\ \text{BRD3} : (\alpha_k,\beta) & \text{BRD6} : (\alpha_j,\beta_j) & \text{BRD9} : (\alpha_k,\beta_j). \end{array}$$

Interpretation follows from there. For example, BRD1 is MCAR, in BRD4 missingness in the first variable is constant, while missingness in the second variable depends on its value. Two of the main advantages of this family is ease of computation in general, and the existence of closed-form solution for several of its members (BRD2–BRD9). Molenberghs, Goetghebeur, Lipsitz and Kenward (1997) used these models in an informal sensitivity analysis of repeated binary measures in a psychiatric study. We consider a slightly different but equivalent parameterization

$$\pi_{r_1 r_1 jk} = p_{jk} \frac{\exp[\beta_{jk}(1-r_2) + \alpha_{jk}(1-r_1) + \gamma(1-r_1)(1-r_2)]}{1+\exp(\beta_{jk}) + \exp(\alpha_{jk}) + \exp(\beta_{jk}+\alpha_{jk}+\gamma)}, \tag{5.8}$$

which contains the marginal success probabilities p_{jk} and forces the missingness probabilities to obey their range restrictions. Model 10 will be defined as (α_k, β_{jk}) with

$$\beta_{jk} = \beta_0 + \beta_j + \beta_k. \tag{5.9}$$

Since one parameter is not indentified, we propose to use β_k as the sensitivity parameter.

Observe that BRD1, being MAR, is equivalent to MAR (2 questions) in Table 6. Model BRD2 produces an estimate for θ which is extremely close to the results of the plebiscite. It assumes that missingness on the independence question depends on the attendance question. Note that BRD8 assumes that missingness on either question depends on the question itself and therefore is very similar to the non-ignorable model of Rubin, Stern, and Vehovar (1995).

Next, we present three estimated Intervals of Ignorance, the result of considering in turn 3 different 'horizons', that is, we consider ignorance under the constraint that models 10, 11 and 12 respectively hold. Model 10 in Table 7 is based on 1 sensitivity parameter, as in (5.9). Similarly, Model 11 uses

$$\alpha_{jk} = \alpha_0 + \alpha_j + \alpha_k, \tag{5.10}$$

while Model 12 combines both (5.9) and (5.10). The estimated II for Models 10 and 11 are very similar and the true plebiscite value is marginal within these II. Note that Model 11, and hence

Table 7: Estimates of the proportion θ (confidence interval) attending the plebiscite and voting for independence, following from fitting the Baker, Rosenberger, and DerSimonian models (1992).

Model	d.f.	loglik	θ
BRD1	6	-2503.06	0.891[0.877;0.906]
BRD2	7	-2476.38	0.884[0.868;0.899]
BRD3	7	-2471.59	0.881[0.865;0.896]
BRD4	7	-2476.38	0.779[0.702;0.857]
BRD5	7	-2471.59	0.848[0.814;0.882]
BRD6	8	-2440.67	0.822[0.792;0.850]
BRD7	8	-2440.67	0.774[0.719;0.828]
BRD8	8	-2440.67	0.753[0.691;0.815]
BRD9	8	-2440.67	0.866[0.849;0.884]
Model 10	9	-2440.67	[0.762;0.893][0.744;0.907]
Model 11	9	-2440.67	[0.766;0.883][0.715;0.920]
Model 12	10	-2440.67	[0.694;0.904]

also Model 12, does contain a number of boundary solutions for the model parameters, which could be seen as evidence against these models.

Another quantity which Rubin, Stern, and Vehovar (1995) reported is the proportion of NO's via nonattendance (see Table 6). Observe that most estimates are way below the plebiscite value. We can gain some insight in this phenomenon by plotting the estimated joint region of ignorance for θ and the proportion of NO's via nonattendance. Since Models 10 and 11 are based on a single sensitivity parameter, the regions of ignorance are curves, while a planar region is obtained for Model 12. The regions are shown twice in Figure 2, on different scales, where Models 10 and 11 are represented by curves and Model 12 by the result of sampled points from the bivariate sensitivity parameter. Model 10 incorporates relatively little ignorance about NO via nonattendance, but at the same time fails to include the plebiscite value. Models 11 and 12 on the contrary allow a relatively large II for this quantity. A black square marks the plebiscite values for both quantities. It is clear that the plebiscite result is *outside* of the range produced by Model 12. Note that a saturated model would incorporate 5 extra sensitivity parameters ! It becomes easier then to abandon the representation in terms of sensitivity parameters and just show the estimated global

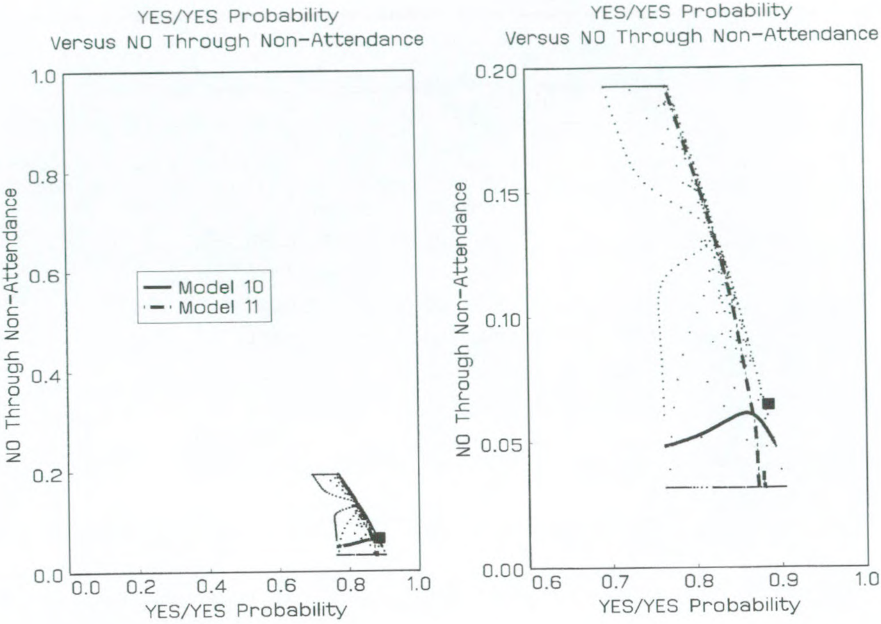

Figure 2: *Graphical Representation of Regions of Ignorance for the Slovenian Plebiscite. Proportion of YES Votes Versus Proportion of NO via Nonattendance. Models 10, 11, and 12.*

bounds, [0.694; 0.905] for θ for instance.

6 Discussion

In this paper we have defined the concept of *ignorance* and incorporated this into a frequentist framework by combining it with the familiar idea of statistical imprecision, producing a measure of *uncertainty*. As an extension of the concept of confidence, uncertainty is expressed as an interval for scalar unknowns (parameters) and a region for vectors. These reduce to conventional confidence intervals and regions when it is assumed that there is no ignorance about the statistical model underlying the data. In the special case of simple categorical data settings with missing values, we have seen that intervals of ignorance can be constructed in a relatively straightforward way, and these reflect in a natural way ignorance about underlying relationships involving unobserved data. The construction of the intervals of uncertainty in function of sensitivity parameters is seen to add useful information in the examples about the problems of interest. In particular, we see that earlier conclusions about the selection and behavior of classes of models for the Slovenian Plebiscite are not strictly justified.

We can approach the calculation of the interval of ignorance in several ways, but we found that a (possibly) overspecified model and associated likelihood are natural concepts to use. This approach will be important, if not essential when we extend the simple categorical setting to more complex problems with continuous and finely stratified covariates and, very importantly, to continuous responses. For these the complete table approach leads to considering missing continuous values within a finite horizon, as members of a bounded set. In these settings careful consideration needs to be given to the family of models within which we are assumed to be ignorant.

Formal tools to assess validity of the new concepts are clearly needed. In a separate paper we suggest consistency definitions for the region of ignorance and coverage for the region of uncertainty. They extend familiar concepts used when there is no ignorance and they might provide a reasonable starting point for further exploration of the notions introduced.

Returning to the specifics of social science practice, we acknowledge that many studies suffer from a high percentage of missing data. Whereas this may tempt one to turn to identified models and ignore the ignorance, it may be argued that the opposite reflex should guide the scientist. Confronting one with the existing level of ignorance, the sensitivity approach encourages more conscious gathering of the external information necessary to narrow uncertainty intervals to useful proportions. It is our sincere hope that the tools provided here will be helpful in this process.

References

Baker, S.G., Rosenberger, W.F., and DerSimonian, R. (1992) Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, **11**, 643–657.

Kenward, M.G. (1998) Selection models for repeated measurements with non-random dropout:

an illustration of sensitivity. *Statistics in Medicine*, **17**, 2723–2732.

Little, R.J.A. (1995) Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 1112–1121.

Molenberghs, G., Goetghebeur, E., Lipsitz, S.R. and Kenward (1998) Nonrandom missingness in categorical data: strengths and limitations. *The American Statistician*, In press.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

Rubin, D.B., Stern, H.S., and Vehovar, V. (1995) Handling "Don't Know" Survey Responses: The Case of the Slovenian Plebiscite. *Journal of the American Statistical Association*, **90**, 822–828.

White I.R. en Goetghebeur E.J.T., (1998). "Clinical trials comparing two treatment arm policies: which aspects of the treatment policies make a difference?" *Statistics in Medicine*, **17**, 319-340, 1998.

Appendix: Some variance expressions for model IV, in Table IV

For the prevalence example, the variance of $\hat{\pi}_\lambda$ is given by (4.7). Following the delta method one obtains in a similar fashion:

$$\begin{aligned}
\widehat{\mathrm{Cov}}((\widehat{\pi_\lambda}, \widehat{\eta_\lambda})) &= -\frac{1}{N}\hat{\pi}_\lambda\,(1-\hat{\pi}_\lambda\,)\frac{1-\lambda}{\lambda}\hat{\eta}_\lambda\,, \\
\widehat{\mathrm{Var}}(\widehat{\eta_\lambda}) &= \frac{\hat{\eta}_\lambda\,(1-\hat{\eta}_\lambda\,)}{N}\left\{1+\frac{1-\hat{\pi}_\lambda}{1-\hat{\eta}_\lambda}\frac{1-\lambda}{\lambda}\right\}.
\end{aligned}$$

For the prevalence odds:

$$\widehat{\mathrm{Var}}(\widehat{\mathrm{odds}(\hat{\pi}_\lambda}\,)) = \frac{1}{N\lambda\hat{\eta}_\lambda}\frac{\hat{\pi}_\lambda}{1-\hat{\pi}_\lambda}\left\{1+\frac{1-\lambda}{\lambda}(1-\hat{\pi}_\lambda\,)[1-\hat{\pi}_\lambda\,\hat{\eta}_\lambda\,(1-\lambda)]\right\}$$

and for the prevalence logit:

$$\widehat{\mathrm{Var}}(\widehat{\mathrm{logit}(\pi_\lambda}\,)) = \frac{1}{N\lambda\hat{\eta}_\lambda}\frac{1}{\hat{\pi}_\lambda\,(1-\hat{\pi}_\lambda\,)}\left\{1+\frac{1-\lambda}{\lambda}(1-\hat{\pi}_\lambda\,)[1-\hat{\pi}_\lambda\,\hat{\eta}_\lambda\,(1-\lambda)]\right\}.$$

Ontvangen: 22-01-1999
Geaccepteerd: 25-07-1999