Appl. Statist. (2001)
50, *Part* 1, *pp.* 15–29

Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case

Geert Molenberghs,

Limburgs Universitair Centrum, Diepenbeek, Belgium

Michael G. Kenward

London School of Hygiene and Tropical Medicine, UK

and Els Goetghebeur

University of Ghent, Belgium

[Received February 1999. Final revision July 2000]

Summary. Classical inferential procedures induce conclusions from a set of data to a population of interest, accounting for the imprecision resulting from the stochastic component of the model. Less attention is devoted to the uncertainty arising from (unplanned) 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. Some proposals have been made for assessing the sensitivity to these modelling assumptions; many are based on fitting several plausible but competing models. For example, we could assume that the missing data are missing at random in one model, and then fit an additional model where non-random missingness is assumed. On the basis of data from a Slovenian plebiscite, conducted in 1991, to prepare for independence, it is shown that such an *ad hoc* procedure may be misleading. We propose an approach which identifies and incorporates both sources of *uncertainty* in inference: *imprecision* due to finite sampling and *ignorance* due to incompleteness. A simple sensitivity analysis considers a finite set of plausible models. We take this idea one step further by considering more degrees of freedom than the data support. This produces sets of estimates (regions of ignorance) and sets of confidence regions (combined into regions of uncertainty).

Keywords: Contingency table; Missing at random; Non-ignorable missingness; Overspecified model; Saturated model; Sensitivity parameter

1. Introduction

In 1991 Slovenians voted for independence from former Yugoslavia in a plebiscite. To prepare for this result, the Slovenian Government collected data in the Slovenian public opinion (SPO) survey, a month before the plebiscite. Rubin *et al.* (1995) studied the three fundamental questions added to the SPO survey and, in comparing it with the outcome of the plebiscite, drew conclusions about the missing data process.

The three questions added were as follows.

(a) Are you in favour of Slovenian independence?

Address for correspondence: Geert Molenberghs, Department of Biostatistics, Limburgs Universitair Centrum, 3590 Diepenbeek, Belgium.
E-mail: geert.molenberghs@luc.ac.be

 0035–9254/01/50015

(b) Are you in favour of Slovenia's secession from Yugoslavia?
(c) Will you attend the plebiscite?

In spite of their apparent equivalence, questions (a) and (b) are different since independence would have been possible in confederal form as well and therefore the secession question was added. Question (c) is highly relevant since the political decision was taken that not attending was treated as an effective answer *no* to question (a).

The data are presented in Table 1. Rubin *et al.* (1995) considered, apart from simple models such as complete and available case analyses, both ignorable models and a non-ignorable (NI) model. The ignorable models outperformed the NI model in that they were much closer to the results of the plebiscite. Even though they were obtained 4 weeks after the SPO survey, the plebiscite data do provide an important bench-mark.

From these observations Rubin *et al.* (1995) inferred that ignorable models are often simpler, can provide realistic results and can therefore be a good starting-point. We add that a deeper look into this problem is necessary, and that the analyses done should be embedded in a, preferably formal, sensitivity analysis.

The problem of analysing data sets from which observations are missing is a common one, and the reasons for data being missing are many and varied. In this setting two main problems need to be addressed. The first, accommodating the lack of balance induced, has to a great extent been answered. Modern statistical tools are not as dependent on simple data structures as was the case before the computer became ubiquitous. The second problem is more fundamental in nature and is far from admitting a straightforward solution. How should we approach statistical inference that accommodates the possible, but unknown, behaviour of the 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. A process is said to be *missing completely at random* (MCAR) if the probability of an observation being missing is independent of both unobserved and observed random variables and *missing at random* (MAR) if, conditionally on the observed data, the probability is independent of the unobserved variables. A process that is neither MCAR nor MAR is termed *non-random*. For likelihood inference MCAR and MAR missing value processes are said to be *ignorable* when the parameters governing the measurement and missing value processes are functionally independent, whereas a non-random process is NI. The importance

Table 1. Data from the SPO survey†

Secession	*Attendance*	*Independence*		
		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

†The 'don't know' category is indicated by *.

of the MCAR and MAR processes is that, given all the observed data, there remains no dependence of the likelihood on unobserved variables and, broadly, inferences can be made that *ignore* the missing value process. This begs the question of whether we can reasonably make assumptions of MCAR and MAR. Sometimes a study 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 value process is non-random. This paper is concerned with how we might approach inference when the possibility of a non-random missingness process cannot be ruled out on *a priori* grounds.

At the technical level, it is not difficult to formulate models for the NI setting, i.e. models in which the probability of an outcome being missing depends on unobserved values. The observed data likelihood is then obtained by integrating over the distribution of the missing data. Little (1995) has provided a review of such approaches. However, there is a fundamental interpretational problem. Molenberghs *et al.* (1999) provided examples, in the contingency table setting, where different NI models that produce the same fit to the observed data are different in their prediction of the unobserved counts. This implies that such models cannot be examined by using data alone. Indeed, even if two models fit the observed data equally well, we still need to reflect on the plausibility of the assumptions made. Several issues are listed in Molenberghs *et al.* (1999). Similar problems manifest themselves in the continuous setting. For example, the distributional form assumed for the unobserved outcomes may determine whether a missingness process is found to be MAR or NI (Little and Rubin (1987), section 11.4, and Kenward (1998)).

Such problems with NI models do not imply, however, that they are of no value. Firstly, many of these issues apply equally well to MAR models which have no *a priori* justification: an MAR model can usually be formulated as a special member of a general family of NI models, although it may be easier to fit. It might be argued, then, that one role of NI models is to supplement information obtained from the MAR model. The concept of fitting a single model is then replaced by that of *sensitivity analysis*, where several plausible NI models are contrasted. This route has been advocated by Vach and Blettner (1995).

Thus, a natural way to proceed is to acknowledge the inherent ambiguity and to explore the range of inferences that are consistent with the gap in our knowledge. Essentially this is a form of sensitivity analysis. Kenward *et al.* (2000) have attempted to formalize this idea. Indeed, although there is a formal mathematical statistical framework for imprecision (variance, standard errors, sampling distributions, confidence intervals, hypothesis tests and so on) most implementations of sensitivity analysis have remained *ad hoc*. The goal of this paper is to develop a simple framework in which general sensitivity concepts can be formalized and further developed, by means of the Slovenian plebiscite case-study. For this, a language will be proposed to describe *ignorance* (due to incompleteness of the data) in addition to the familiar *imprecision* (due to finite sampling) and to combine both into *uncertainty*. We focus here on the multinomial contingency table setting, for which interesting and practically useful results can be obtained in a reasonably direct fashion.

The original analysis of the Slovenian plebiscite will be discussed in Section 2. An extended family of non-random models, due to Baker *et al.* (1992), is considered in Section 3. Section 4 presents a formalization of sensitivity analysis by using the fundamental concept of an overspecified likelihood and the plebiscite data are reanalysed in Section 5.

The programs that were used to analyse the data can be obtained from

```
http://www.blackwellpublishers.co.uk/rss/
```

Table 2. Estimates of the proportion θ attending the plebiscite and voting for independence, as presented in Rubin *et al.* (1995)

Estimation method	*For independence, θ*	*No via non-attendance, ν*
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
NI	0.782	0.122
Plebiscite	0.885	0.065

2. The original analysis

Rubin *et al.* (1995) conducted several analyses of the data. Their main emphasis was in determining the proportion θ of the population who would attend the plebiscite and vote for independence. The three other combinations of these two binary outcomes would be treated as voting 'no'. Their estimates are reproduced in Table 2, which also shows the proportion ν of no via non-attendance (i.e. the proportion of the population who would not attend the plebiscite). The conservative method is the ratio of the (yes, 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 to their analysis the most optimistic estimate that replaces the numerator by all who would not definitely vote no:

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

These figures are obtained by first collapsing over the secession variable and then summing the counts of the (yes, yes), (yes, $*$), ($*$, yes) and ($*$, $*$) categories.

Both estimates together yield the interval

$$\theta \in [0.694; 0.905]. \tag{1}$$

The corresponding interval for ν (no through non-attendance) is

$$\nu \in [0.031; 0.192]. \tag{2}$$

The complete-case estimate for θ is based on the subjects answering all three questions,

$$\hat{\theta} = \frac{1191 + 158}{1454} = 0.928,$$

and the available case estimate is based on the subjects answering the two questions of interest here

$$\hat{\theta} = \frac{1191 + 158 + 90}{1549} = 0.929.$$

It is noteworthy that both estimates fall outside the pessimistic–optimistic interval and should be disregarded, since these seemingly straightforward estimators do not take the decision to treat absences as no votes into account and thus discard available information.

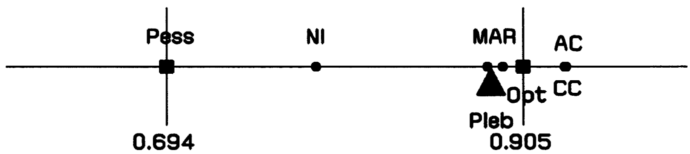

Fig. 1. Slovenian plebiscite: relative positions for the estimates of the proportion of yes votes, based on the models considered in Rubin *et al.* (1995); the vertical lines indicate the nonparametric pessimistic–optimistic bounds (Pess, pessimistic boundary; Opt, optimistic boundary; MAR, the MAR model of Rubin *et al.* (1995); NI, the NI model of Rubin *et al.* (1995); AC, available cases; CC, complete cases; Pleb, outcome of the plebiscite)

Rubin *et al.* (1995) considered two MAR models, the first based on the two questions of direct interest only and the second using all three, yielding $\hat{\theta} = 0.892$ and $\hat{\theta} = 0.883$ respectively. Finally, they considered a single NI model, based on the assumption that missingness on a question depends on the answer to that question but not on the other questions. They argued that this is a plausible assumption. The corresponding estimator is $\hat{\theta} = 0.782$.

Fig. 1 sketches the relative position of the estimates of Rubin *et al.* (1995). In summary, we see that

(a) the available case and complete-case estimates are outside the pessimistic–optimistic range (1) (indicated by vertical bars) and
(b) both MAR estimates are very close to and the NI estimate is very far from the 'truth', which is the proportion who voted yes at the actual plebiscite: $\theta = 0.885$.

On the basis of these findings, and those from other carefully designed surveys, Rubin *et al.* (1995) concluded that the MAR assumption can be plausible, when there is limited non-response and good covariate information. Although we agree with the closeness of the MAR analyses in this case, it is of course unclear whether the MAR mechanism will always be the preferred non-response mechanism. In addition, we aim to place the MAR analysis within a whole family of NI models, to shed additional light on these data.

In the next sections we shall first consider a (finite) parametric family of NI models. This can be seen as an informal sensitivity analysis, which will then be formalized by considering continuous intervals (or regions) of NI models.

3. A parametric family of non-ignorable models

We focus on the case of two binary outcomes with arbitrary patterns of incompleteness, such as for the first and the third question in the SPO survey. An interesting class of models has been proposed by Baker *et al.* (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 $r_1, r_2 = 0, 1$ indicates whether a measurement is either missing or taken at occasions 1 and 2 respectively, and $j, k = 1, 2$ indicates the response categories for both outcomes. The models are written as

$$E(Y_{11jk}) = m_{jk}, \qquad E(Y_{01jk}) = m_{jk}\alpha_{jk},$$
$$E(Y_{10jk}) = m_{jk}\beta_{jk}, \qquad E(Y_{00jk}) = m_{jk}\alpha_{jk}\beta_{jk}\gamma,$$

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

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

Here, $q_{r_1 r_2|jk}$ is the probability associated with response pattern (r_1, r_2), given outcomes j and k. The subscripts are missing from γ since Baker *et al.* (1992) have shown that this quantity is independent of j and k. They considered nine identifiable models, based on setting α_{jk} and β_{jk} constant in one or more indices:

$$\text{BRD1}, (\alpha, \beta); \quad \text{BRD2}, (\alpha, \beta_j); \quad \text{BRD3}, (\alpha_k, \beta);$$
$$\text{BRD4}, (\alpha, \beta_k); \quad \text{BRD5}, (\alpha_j, \beta); \quad \text{BRD6}, (\alpha_j, \beta_j);$$
$$\text{BRD7}, (\alpha_k, \beta_k); \quad \text{BRD8}, (\alpha_j, \beta_k); \quad \text{BRD9}, (\alpha_k, \beta_j).$$

The interpretation is similar to that for log-linear models. For example, model BRD1 is MCAR, whereas in models BRD2 and BRD4 missingness in the first variable is constant, and missingness in the second variable depends on the, possibly unobserved, value of the first (BRD2) or second (BRD4) variable. A log-linear representation for the latter two models would be $(Y_1 Y_2, R_1 R_2, Y_1 R_2)$ and $(Y_1 Y_2, R_1 R_2, Y_2 R_2)$ respectively. Two of the main advantages of this family are ease of computation in general and the existence of closed form solutions for several of its members (models BRD2–BRD9). The result of fitting these models is presented in Table 3. Observe that BRD1, being MCAR, is equivalent to MAR (two questions) in Table 2. 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 attendence question, a mechanism that is *different* from the mechanism assumed by Rubin *et al.* (1995). Note that model BRD8 assumes that missingness on either question depends on the question itself and therefore is very similar to their NI model. Fig. 1, supplemented with the BRD estimates, is shown as Fig. 2. The range covered by models BRD1–BRD9 is

$$\theta \in [0.753; 0.891], \tag{3}$$

which is a considerable part of the nonparametric pessimistic–optimistic range (1). We conclude that the conclusion about NI models, as presented by Rubin *et al.* (1995), is at best premature, since considering a set of models shows that, depending on the (unverifiable) assumptions made, NI models range from relatively remote from the plebiscite data to very close.

Table 3. Estimates of the proportion θ (confidence interval) attending the plebiscite and voting for independence, following from fitting the models of Baker *et al.* (1992)†

Model	*Degrees of freedom*	*Log-likelihood*	θ	ν
BRD1	6	−2503.06	0.891 [0.877; 0.906]	0.044 [0.034; 0.054]
BRD2	7	−2476.38	0.884 [0.868; 0.899]	0.048 [0.037; 0.060]
BRD3	7	−2471.59	0.881 [0.865; 0.896]	0.047 [0.036; 0.058]
BRD4	7	−2476.38	0.779 [0.702; 0.857]	0.048 [0.037; 0.060]
BRD5	7	−2471.59	0.848 [0.814; 0.882]	0.106 [0.071; 0.141]
BRD6	8	−2440.67	0.822 [0.792; 0.850]	0.134 [0.105; 0.163]
BRD7	8	−2440.67	0.774 [0.719; 0.828]	0.050 [0.037; 0.062]
BRD8	8	−2440.67	0.753 [0.691; 0.815]	0.134 [0.105; 0.163]
BRD9	8	−2440.67	0.866 [0.849; 0.884]	0.061 [0.047; 0.076]
10	9	−2440.67	[0.762; 0.893] [0.744; 0.907]	[0.037; 0.044] [0.029; 0.055]
11	9	−2440.67	[0.766; 0.883] [0.715; 0.920]	[0.032; 0.193] [−0.037; 0.242]
12	10	−2440.67	[0.694; 0.905]	[0.032; 0.193]

†BRD1–BRD9 are identified models (Section 3) whereas models 10–12 are overspecified (Section 5).

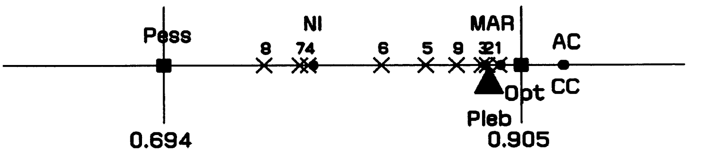

Fig. 2. Slovenian plebiscite: relative positions for the estimates of the proportion of yes votes, based on the models considered in Rubin *et al.* (1995) and on the models of Baker *et al.* (1992); the vertical lines indicate the nonparametric pessimistic–optimistic bounds (Pess, pessimistic boundary; Opt, optimistic boundary; MAR, the MAR model of Rubin *et al.* (1995); NI, the NI model of Rubin *et al.* (1995); AC, available cases; CC, complete cases; Pleb, outcome of the plebiscite; the numbers refer to the BRD models)

Although the conduct of such an informal sensitivity analysis is enlightening, it does not remove all concerns. Indeed, there is no guarantee that a family of intervals will provide a good coverage of all (NI) models within a class of plausible models. A formal sensitivity analysis strategy that addresses this issue is discussed next.

4. A formal sensitivity analysis

It is useful to distinguish between two types of *statistical uncertainty*. The first, *statistical imprecision*, is due to finite sampling. The SPO survey included not all Slovenians but only 2074 respondents. However, even if all had been included, there would have been residual uncertainty because some fail to report at least one answer. This second source of uncertainty, due to incompleteness, will be called *statistical ignorance*.

Statistical imprecision is classically quantified by means of estimators (standard error and variance, confidence regions etc.) and properties of estimators (consistency, asymptotic distribution, efficiency etc.). To quantify statistical ignorance, it is useful to distinguish between complete and observed data. Let us focus on two binary questions, such as the independence and attendance questions in the SPO survey. Using indices as in Section 3, the 16 theoretical complete-cell probabilities are as in Table 4, thus producing 15 complete-data degrees of freedom. Similarly, the nine observed cells can be represented as in Table 5, which is directly comparable with the observed data structure. In the SPO case, for example, these nine counts are obtained from collapsing Table 1 over the secession question, hence producing Table 6.

A sample from Table 5 produces empirical proportions representing the πs with error. This imprecision disappears as the sample size tends to ∞. What remains is ignorance regarding the redistribution of all except the first four πs over the missing outcomes value. This leaves ignorance regarding any probability in which at least one of the first or second indices is equal to 0, and hence regarding any derived parameter of scientific interest. For such a parameter, θ say, a region of possible values which is consistent with Table 5 is called a region of ignorance. Analogously an observed incomplete table leaves ignorance regarding the

Table 4. Theoretical distribution over completed cells

$\pi_{11,11}$	$\pi_{11,12}$	$\pi_{10,11}$	$\pi_{10,12}$	$\pi_{01,11}$	$\pi_{01,12}$	$\pi_{00,11}$	$\pi_{00,12}$
$\pi_{11,21}$	$\pi_{11,22}$	$\pi_{10,21}$	$\pi_{10,22}$	$\pi_{01,21}$	$\pi_{01,22}$	$\pi_{00,21}$	$\pi_{00,22}$

Table 5. Theoretical distribution over observed cells

$\pi_{11,11}$	$\pi_{11,12}$
$\pi_{11,21}$	$\pi_{11,22}$

$\pi_{10,1+}$
$\pi_{10,2+}$

$\pi_{01,+1}$	$\pi_{01,+2}$

$\pi_{00,++}$

Table 6. Observed cells for the SPO survey, collapsed over the secession question

1439	78
16	18

159
32

144	54

136

would-be observed complete table, which in turn leaves imprecision regarding the true complete probabilities. The region of estimators for θ consistent with the observed data provides an estimated region of ignorance. The $100(1-\alpha)\%$ *region of uncertainty* is a larger region in the spirit of a confidence region, designed to capture the combined effects of imprecision and ignorance. Various ways of constructing regions of ignorance and regions of uncertainty are conceivable.

4.1. Overspecified models

In standard statistical practice, ignorance is avoided by considering a single identified model, such as models BRD1–BRD9. Among those, models BRD6–BRD9 are said to saturate the degrees of freedom. To be precise, they saturate the *observed data* degrees of freedom. A model that would saturate the *complete*-*data* degrees of freedom would need 15 rather than eight parameters. From a classical observed data perspective, such a model would be overspecified, as would be any model with nine or more parameters. (Note that it is possible to construct an overspecified model with degrees of freedom less than those in an identifiable saturated model at the observed level.)

We shall construct three such overspecified models, which will be used later for the analysis of the SPO data. To proceed, we first consider a slightly different but equivalent parameterization in terms of the joint probabilities:

$$\pi_{r_1 r_1 jk} = p_{jk} \frac{\exp\{\beta_{jk}(1-r_2) + \alpha_{jk}(1-r_1) + \gamma(1-r_1)(1-r_2)\}}{1+\exp(\beta_{jk})+\exp(\alpha_{jk})+\exp(\beta_{jk}+\alpha_{jk}+\gamma)}, \tag{4}$$

which contains the marginal success probabilities p_{jk} and forces the missingness probabilities to obey their range restrictions.

We shall consider two models (models 10 and 11) with a single sensitivity parameter, whereas model 12 will include two sensitivity parameters. Model 10 is defined as (α_k, β_{jk}) with

$$\beta_{jk} = \beta_0 + \beta_j + \beta_k, \tag{5}$$

an additive decomposition for missingness on the independence question. In log-linear representation, we could write $(Y_1Y_2,\ R_1R_2,\ Y_2R_1,\ Y_1R_2,\ Y_2R_2)$.

Similarly, model 11, (α_{jk}, β_j), uses

$$\alpha_{jk} = \alpha_0 + \alpha_j + \alpha_k, \tag{6}$$

an additive decomposition of the missingness parameter on the attendance question. An alternative representation is $(Y_1Y_2,\ R_1R_2,\ Y_1R_1,\ Y_2R_1,\ Y_1R_2)$.

Finally, we define model 12, $(\alpha_{jk},\ \beta_{jk})$, as a combination of both equation (5) and equation (6).

We shall now outline the general principle behind considering such overspecified models and then focus on the sensitivity parameter approach to proceed with model fitting.

4.2. General principle

We start from the classical approach of fitting a single identifiable model M_0 to incomplete data (e.g. a particular BRD model). Maximum likelihood estimation produces a parameter estimate $\hat{\pi}$ along with measures of imprecision (estimated standard errors). From $\hat{\pi}$ four predicted contingency tables can be derived as in Table 4.

The fitted complete tables collapse back to fitted values for the incomplete Table 5. Contrasting the latter with the observed data shows the goodness of fit of model M_0. If there is a substantial lack of fit, the original model M_0 needs to be reconsidered. A lack of fit has strong bearings on imprecision and, since we want to focus on ignorance, we shall assume that the fit is acceptable. In what follows, models with poor fit (or boundary solutions) will be dropped.

We can now range through *all possible* complete tables, which collapse back to the M_0 predicted incomplete table. We call the tables 'M_0 compatible' and we denote the set by $\mathcal{S}(M_0)$. The general principle is that to each table in $\mathcal{S}(M_0)$ an extended model M^* will be fitted. This implies that each table produces an estimated parameter vector and a confidence region. The unions of those are termed the *region of ignorance* and *region of uncertainty* respectively. For scalar parameters the terms interval of ignorance and interval of uncertainty will be used.

Apart from explicitly constructing the (real-valued) set of complete tables, we can proceed in an alternative way. This is done by fitting the model M^* directly to the observed data. This implies that the general principle translates to fitting an overspecified model to the observed data, which will produce a *range* of parameters maximizing the observed data likelihood. This range is then the region of ignorance. If this route is followed, there are technically several ways to find the region. One method is described in Section 4.3.

4.3. Sensitivity parameter approach

The overspecification can be removed by considering a minimal set of parameters $\boldsymbol{\eta}$, conditional on which the others, $\boldsymbol{\mu}$, are identified. We term $\boldsymbol{\eta}$ the sensitivity parameter and $\boldsymbol{\mu}$ the estimable parameter. Such a technique has been proposed for specific examples by Nordheim (1984) and Vach and Blettner (1995). Each value of $\boldsymbol{\eta}$ will produce an estimate $\hat{\mu}(\boldsymbol{\eta})$. The union of these yields the region of ignorance. It is important to realize that in general there will not be a unique choice for $\boldsymbol{\eta}$ and hence for $\boldsymbol{\mu}$. Changing the partitioning will produce the same region for $\boldsymbol{\theta} = (\boldsymbol{\eta}',\ \boldsymbol{\mu}')'$. Models 10 and 11 have a single sensitivity parameter. We chose $\eta = \beta_k$ and $\eta = \alpha_k$ from equations (5) and (6) respectively. In model 12, both these parameters $\boldsymbol{\eta} = (\beta_k,\ \alpha_k)'$ are treated as sensitivity parameters. In practice, an easy computation scheme is to consider a grid in the sensitivity parameter space, at each value of which the estimable parameter is maximized.

A natural estimate of the region of uncertainty is the union of confidence regions for each

$\hat{\mu}(\eta)$. We must ensure that η is within the allowable range. Since the choice of sensitivity parameter is non-unique a proper choice can greatly simplify the treatment. Another issue is whether the parameters of direct scientific interest can overlap with the sensitivity set or not (see White and Goetghebeur (1998)). For example, if the scientific question is a sensitivity analysis for treatment effect, then we should consider the implications of including the treatment effect parameters in the sensitivity set. No direct estimate of imprecision will be available for the sensitivity parameter. Clearly, the particular choice of sensitivity parameter will not affect the estimate of the region of ignorance. However, the region of uncertainty is built from confidence regions which are conditional on a particular value of the sensitivity parameter and hence will typically vary with the choice made.

5. Artificial examples

To study the behaviour of the intervals of ignorance and uncertainty, we consider eight artificial sets of data, as presented in Table 7. The complete-data counts are presented. It is easy to derive the observed data. For example, set (a) produces

200	100
100	200

300
300

300	300

600

The eight sets are chosen to study the effect of three factors. First, sets (a)–(d) are MCAR (model BRD1), whereas sets (e)–(h) are NI (model BRD9). Second, for sets (a), (b), (e) and (f), the proportions in the four response patterns are (25%, 25%, 25%, 25%), whereas for the other sets they are (50%, 20%, 20%, 10%). The α and β in equation (4) were chosen to approximate these proportions. For sets (e) and (f), we set $\alpha_0 = 0$, $\alpha_1 = 0.75$, $\beta_0 = 0$ and $\beta_1 = -1$. For sets (g) and (h), we set $\alpha_0 = -1.65$, $\alpha_1 = -0.5$, $\beta_0 = -0.5$ and $\beta_1 = -1.65$. In both cases, $\gamma = 0$. Owing to rounding, this is only approximately so for the NI models. Third, the sample size in the even patterns (b), (d), (f), (h) is 10 times the size of the odd-numbered patterns (a), (c), (e), (g).

Table 7. Artificial sets of data

Set	*(1, 1)*		*(1, 0)*		*(0, 1)*		*(0, 0)*	
(a)	200	100	200	100	200	100	200	100
	100	200	100	200	100	200	100	200
(b)	2000	1000	2000	1000	2000	1000	2000	1000
	1000	2000	1000	2000	1000	2000	1000	2000
(c)	400	200	160	80	160	80	80	40
	200	400	80	160	80	160	40	80
(d)	4000	2000	1600	800	1600	800	800	400
	2000	4000	800	1600	800	1600	400	800
(e)	200	64	200	64	200	136	200	136
	146	188	54	69	146	397	54	146
(f)	2000	640	2000	640	2000	1360	2000	1360
	1460	1880	540	690	1460	3970	540	1460
(g)	417	154	254	94	80	94	49	58
	281	417	54	80	54	254	11	49
(h)	4170	1540	2540	940	800	940	490	580
	2810	4170	540	800	540	2540	110	490

Table 8. Artificial data: estimates of the proportion θ (confidence interval) attending the plebiscite and voting for independence, based on the pessimistic–optimistic range and on overspecified model BRD10 (interval of ignorance and interval of uncertainty)

Set	*Mechanism*	*Incompleteness*	*Sample*	*Pessimistic–optimistic interval*	*Interval of ignorance*	*Interval of uncertainty*
(a)	MCAR	Large	Small	[0.083; 0.583]	[0.167; 0.417]	[0.140; 0.443]
(b)	MCAR	Large	Large	[0.083; 0.583]	[0.167; 0.417]	[0.161; 0.425]
(c)	MCAR	Small	Small	[0.167; 0.467]	[0.233; 0.383]	[0.215; 0.405]
(d)	MCAR	Small	Large	[0.167; 0.467]	[0.233; 0.383]	[0.228; 0.390]
(e)	NI	Large	Small	[0.083; 0.561]	[0.167; 0.429]	[0.148; 0.455]
(f)	NI	Large	Large	[0.083; 0.561]	[0.167; 0.429]	[0.161; 0.437]
(g)	NI	Small	Small	[0.174; 0.444]	[0.208; 0.402]	[0.191; 0.423]
(h)	NI	Small	Large	[0.174; 0.444]	[0.208; 0.402]	[0.203; 0.409]

The results of fitting overspecified model 10 to the artificial data are summarized in Table 8. There are three intervals. First, we present the nonparametric bounds on the probability of voting yes–yes, as was done in interval (1). Next, we present intervals of ignorance and intervals of uncertainty. Recall that the interval of ignorance reflects uncertainty due to missingness, whereas the interval of uncertainty combines both sources of uncertainty, i.e. it also reflects sampling variability.

We can make several observations. First, the effect of the sample size is seen only in the interval of uncertainty. Indeed, each pair of sets yields the same nonparametric bounds and the same interval of ignorance, but the even sets produce sharper intervals of uncertainty than do the odd ones. Second, the proportion of incompleteness is larger in sets (a), (b), (e) and (f), as opposed to in sets (c), (d), (g) and (h) respectively. This is reflected in larger ignorance, i.e. in larger nonparametric ranges and larger intervals of ignorance.

Third, comparing the nonparametric range and the interval of ignorance shows the effect of the modelling strategy. Indeed, the interval of ignorance is neither based on an identifiable model nor fully nonparametric. Rather, it compromises between both by allowing overspecification, but in a controlled fashion. In this case, we chose to illustrate the potential by including one extra parameter (one sensitivity parameter), by means of model 10. Of course, in a real situation we must reflect on the plausibility of such model assumptions. It then provides a compromise between the acknowledgement of ignorance due to incompleteness (abandoning a single point estimate) and useful lengths of the corresponding intervals (avoiding the nonparametric bounds).

Finally, there is a striking symmetry between the results for the MCAR models (a)–(d) and their NI counterparts in sets (e)–(h). This implies that, other things being equal, the precise form of the NI mechanism seems to be less relevant. This feature distinguishes our sensitivity analysis from fitting a single identified model. Let us expand on this point. If we consider models BRD1–BRD9 for set (a), then all models produce $\hat{\theta} = 0.333$, the true value. This follows from the fact that all the models are extensions of the MCAR model (BRD1), which fits the data exactly. However, model 10 intrinsically includes deviations from the MCAR mechanism by considering a whole range for the sensitivity parameter. In contrast, if we fit models BRD1–BRD9 to set (e), then we obtain

BRD1, 0.317;	BRD2, 0.342;	BRD3, 0.321;
BRD4, 0.426;	BRD5, 0.167;	BRD6, 0.167;
BRD7, 0.426;	BRD8, 0.193;	BRD9, 0.333.

In other words, models BRD1–BRD9 span a wide variety of estimates. This also holds for the subset BRD6–BRD9 of the saturated models. These almost reproduce the entire interval of ignorance. Thus, the formal sensitivity analysis removes the *ad hoc* nature of intervals, computed from fitting a number of identified models. In this case, we have shown that intervals can be everything from a single point to almost the entire interval of ignorance.

6. Application to the plebiscite data

We shall apply these formal sensitivity concepts to the plebiscite data, based on the introduction of sensitivity parameters in the model family of Baker *et al.* (1992), as in models 10–12, as introduced in Section 4.

The estimated intervals of ignorance and intervals of uncertainty are shown in Table 3, whereas a graphical representation of the yes votes is given in Fig. 3. A representation for the proportion of no votes via non-attendance is given in Fig. 4. Let us first discuss the proportion θ of yes votes. Model 10 shows an interval of ignorance which is very close to interval (3), the range produced by the models BRD1–BRD9, whereas model 11 is somewhat sharper and just fails to cover the plebiscite value. However, it should be noted that the corresponding intervals of uncertainty convincingly cover the truth.

Interestingly, model 12 virtually coincides with the nonparametric range (1), even though it does not saturate the complete-data degrees of freedom. To do so, not two but in fact seven

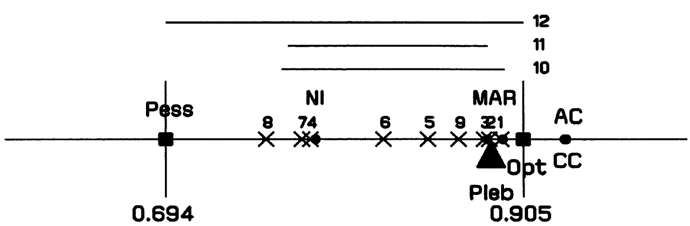

Fig. 3. Slovenian plebiscite: relative positions for the estimates of the proportion of yes votes, based on the models considered in Rubin *et al.* (1995) and on the models of Baker *et al.* (1992); the vertical lines indicate the nonparametric pessimistic–optimistic bounds (Pess, pessimistic boundary; Opt, optimistic boundary; MAR, the MAR model of Rubin *et al.* (1995); NI, the NI model of Rubin *et al.* (1995); AC, available cases; CC, complete cases; Pleb, outcome of the plebiscite; numbers refer to the BRD models; intervals of ignorance (models 10–12) are represented by horizontal bars)

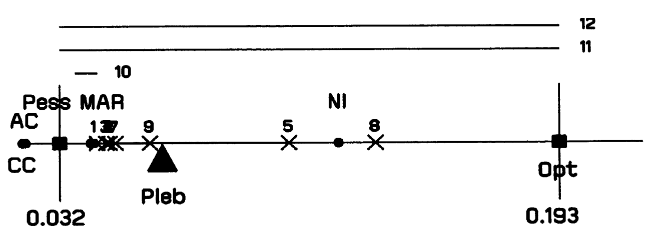

Fig. 4. Slovenian plebiscite: relative positions for the estimates of the proportion of no votes via non-attendance, based on the models considered in Rubin *et al.* (1995) and on the models of Baker *et al.* (1992); the vertical lines indicate the nonparametric pessimistic–optimistic bounds (Pess, pessimistic boundary; Opt, optimistic boundary; MAR, the MAR model of Rubin *et al.* (1995); NI, the NI model of Rubin *et al.* (1995); AC, available cases; CC, complete cases; Pleb, outcome of the plebiscite; numbers refer to the BRD models; horizontal lines refer to overspecified BRD models; intervals of ignorance (models 10–12) are represented by horizontal bars)

sensitivity parameters would have to be included. Thus, it appears that a relatively simple sensitivity analysis is sufficient to increase the insight in the information provided by the incomplete data about the proportion of valid yes votes. This simplicity may not hold in all cases, as will be illustrated next.

Let us now turn to ν, the proportion of no votes via non-attendance. In some aspects, a similar picture holds in the sense that model 10 just fails to cover the plebiscite value, whereas models 11 and 12 produce intervals of ignorance which virtually coincide with the non-parametric range. A major difference between θ and ν is that in the first case the MAR models of Rubin *et al.* (1995) are very close to the plebiscite value, whereas in the second case the MAR models are relatively far out. The plebiscite value of the proportion of no votes via non-attendance is best reproduced by model BRD9. Thus, a specific model, such as MAR, can be acceptable for one estimand but not necessarily for another.

We can enhance insight by studying the *pair* (θ, ν). For this, let us plot the region of ignorance for both θ and ν. Since models 10 and 11 are based on a single sensitivity parameter, the regions of ignorance are one-dimensional curves, whereas a two-dimensional planar region is obtained for model 12. A graphical sketch is given in Fig. 5. Figs 5(a) and 5(b) contain the same information. Fig. 5(a) is useful to answer the substantive questions. Indeed, the conclusion from it is that, even when ignorance is taken into account, a convincing majority will vote for independence and only a very small proportion will provide a no vote through non-attendance. Fig. 5(b) zooms in on the region of ignorance to distinguish its features better.

In fact, Fig. 5 combines the univariate intervals from Figs 3 and 4. Models 10 and 11 are represented by curves. To obtain a representation for model 12, points of the sensitivity parameter are sampled from a bivariate uniform distribution. For each of those pairs, the model is fitted and the corresponding $(\hat{\theta}, \hat{\nu})$ determined. These points are then plotted. A full square marks the plebiscite values for both quantities. The MAR analysis is represented by a full circle. It is clear that the plebiscite result is *on the boundary* of the range produced by model 12, whereas it is not on the boundary of the optimistic–pessimistic range (represented by means of a broken box). Thus, whereas the univariate intervals of ignorance convincingly include the plebiscite value, this is less so for the bivariate region, indicating that it enhances understanding. Note that a saturated model would incorporate five extra sensitivity parameters! Such an extended analysis will increase the region of ignorance into the direction of the optimistic–pessimistic box, thereby relaxing the boundary location of the plebiscite value.

7. Discussion

In this paper we have defined the concept of *ignorance* and combined 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. The construction of the intervals of uncertainty in the examples was seen to convey useful information about the problems concerned, providing information that has not previously been appreciated. In particular, we see that earlier conclusions about the selection and behaviour of classes of models for the Slovenian plebiscite are not strictly justified.

We have introduced three paths to sensitivity analysis. 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

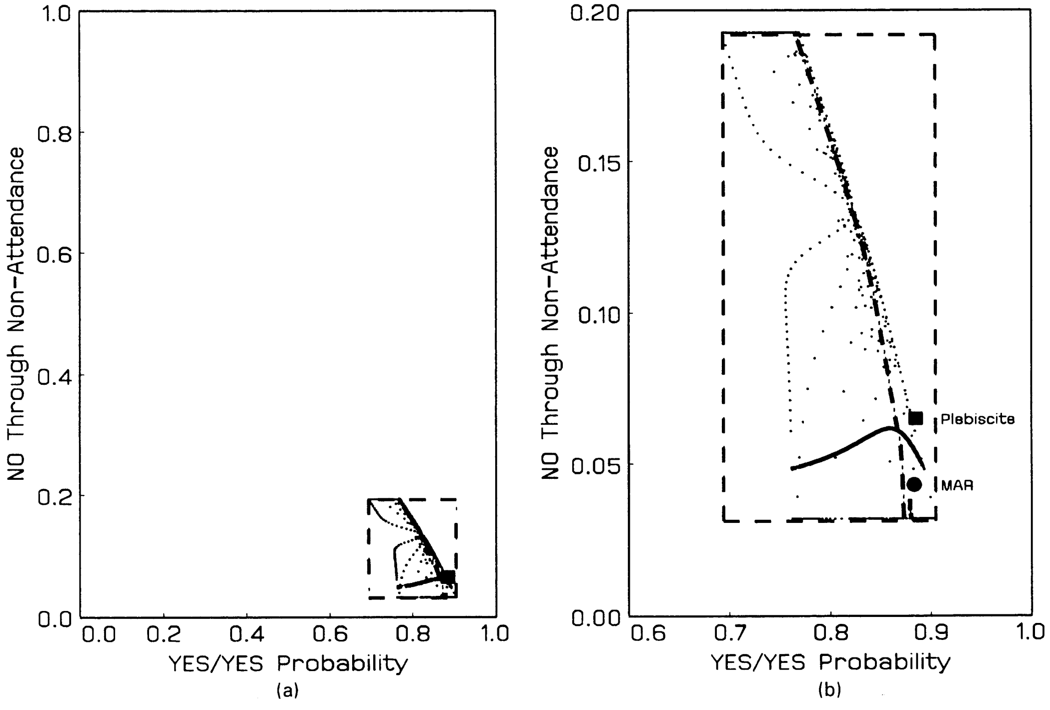

Fig. 5. Graphical representation of regions of ignorance for the Slovenian plebiscite, proportion of yes votes *versus* proportion of no votes via non-attendance (the interval of ignorance is the envelope of the points so obtained; ———, model 10; - - - - -, model 11; •, model 12; ⬚, optimistic–pessimistic bounds): (a) absolute impression of ignorance in the unit square; (b) focus on the relative position of the models by zooming in on the relevant region

favour of independence. Second, a range of plausible models can be considered, such as those proposed by Baker *et al.* (1992). Here, their range is qualitatively not much different from the range obtained by the bounds but enables further distinction between

(a) well fitting and poorly fitting models and
(b) model formulations (drop-out mechanisms) that are deemed plausible, in contrast with models where the drop-out mechanism is not tenable on substantive grounds.

This is necessarily subjective, but with incomplete data subjectivity should be controlled rather than avoided. Third, plausible but overspecified models can be considered. More overspecification will yield models that produce intervals of ignorance that are closer to the bounds, whereas models that are too parsimonious or not plausible may miss the true value. Of course, in many studies the true (plebiscite) value will not be known and such an ultimate check cannot follow. However, the strategies presented here enable a consideration of *classes* of models, and the amount of parsimony can be controlled. It is also possible to 'average' over models, e.g. using priors, such as in Forster and Smith (1998).

We can approach the calculation of the interval of ignorance in several ways, but it is seen that a (possibly) overspecified model and associated likelihood are the more natural concepts to use. We have focused on the use of a sensitivity parameter to determine the set of maxima of this saturated likelihood.

Formal tools to assess the validity of the new concepts are clearly needed. In a separate paper we shall suggest consistency definitions for the region of ignorance and coverage for the region of uncertainty.

References

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

Forster, J. J. and Smith, P. W. F. (1998) Model-based inferences for categorical survey data subject to non-ignorable non-response. *J. R. Statist. Soc.* B, **60**, 57–70.

Kenward, M. G. (1998) Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statist. Med.*, **17**, 2723–2732.

Kenward, M. G., Goetghebeur, E. and Molenberghs, G. (2000) Sensitivity analysis of incomplete categorical data. To be published.

Little, R. J. A. (1995) Modelling the drop-out mechanism in repeated-measures studies. *J. Am. Statist. Ass.*, **90**, 1112–1121.

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

Molenberghs, G., Goetghebeur, E., Lipsitz, S. R. and Kenward, M. G. (1999) Non-random missingness in categorical data: strengths and limitations. *Am. Statistn*, **53**, 110–118.

Murray, G. D. and Findlay, J. G. (1988) Correcting for the bias caused by drop-outs in hypertension trials. *Statist. Med.*, **7**, 941–946.

Nordheim, E. V. (1984) Inference from nonrandomly missing categorical data: an example from a genetic study on Turner's syndrome. *J. Am. Statist. Ass.*, **79**, 772–780.

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

——— (1977) Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Am. Statist. Ass.*, **72**, 538–543.

Rubin, D. B., Stern, H. S. and Vehovar, V. (1995) Handling "Don't know" survey responses: the case of the Slovenian plebiscite. *J. Am. Statist. Ass.*, **90**, 822–828.

Vach, W. and Blettner, M. (1995) Logistic regression with incompletely observed categorical covariates—investigating the sensitivity against violation of the missing at random assumption. *Statist. Med.*, **14**, 1315–1329.

White, I. R. and Goetghebeur, E. J. T. (1998) Clinical trials comparing two treatment arm policies: which aspects of the treatment policies make a difference? *Statist. Med.*, **17**, 319–340.