Statistical Modelling under Epistemic Data Imprecision: Some Results on Estimating Multinomial Distributions and Logistic Regression for Coarse Categorical Data

Julia Plass
Department of Statistics, LMU Munich
julia.plass@stat.uni-muenchen.de

Thomas Augustin
Department of Statistics, LMU Munich
augustin@stat.uni-muenchen.de

Marco E. G. V. Cattaneo
Department of Mathematics, University of Hull
m.cattaneo@hull.ac.uk

Georg Schollmeyer
Department of Statistics, LMU Munich
georg.schollmeyer@stat.uni-muenchen.de

Abstract

The paper deals with parameter estimation for categorical data under epistemic data imprecision, where for a part of the data only coarse(ned) versions of the true values are observable. For different observation models formalizing the information available on the coarsening process, we derive the (typically set-valued) maximum likelihood estimators of the underlying distributions. We discuss the homogeneous case of independent and identically distributed variables as well as logistic regression under a categorical covariate. We start with the imprecise point estimator under an observation model describing the coarsening process without any further assumptions. Then we determine several sensitivity parameters that allow the refinement of the estimators in the presence of auxiliary information.

Keywords. Coarse data, missing data, epistemic data imprecision, sensitivity analysis, partial identification, categorical data, multinomial logit model, coarsening at random (CAR), likelihood.

1 The Problem and its Background

A frequent challenge in statistical modelling is *data imprecision*, where some data are *coarse*, i.e. they are not observed in the resolution originally intended in the subject matter context. Throughout this paper, we focus on the case where the coarse observations are data under *epistemic data imprecision.* For categorical data as considered here this means that there exists a true precise value y of a generic variable Y taking values in a finite sample space $\Omega_Y = \{1, \dots, K\}$, but we may only observe a non-singleton set $\mathscr{Y}$ containing y. It is important to distinguish epistemic from *ontic* data imprecision, where data are coarse by nature and thus have to be interpreted as indivisible entities of their own (see, in particular, [7, 8]; [24] for an application in a multinomial logit model and classification.)

Epistemic data imprecision emerges most naturally in a huge variety of applications. Missing data, interpreted as the prominent special case where the whole sample space is observed only, arise, for instance directly by design in observational studies on treatment effects, see, e.g., [27], and unit non-response is quite frequent in surveys, in particular as refusals to answer sensitive questions. Typical instances of not missing but still coarse data include the numerous data sets where coarsening is deliberately applied as an anonymization technique (see, e.g., [10]), matched data sets with not completely identical categories, secondary data where the originally coded categories turn out to be not fine enough and, as a technical example, reliability analysis of a system whose components are tested separately prior to assembly [30].

Trapped in the framework of precise probabilities, traditional statistical methods are forced to neglect data imprecision or to impose quite strong, empirically untestable assumptions on the underlying coarsening process. Thus, except the very rare cases where the external information on the subject matter problem is rich enough to justify such an extent of precision of the modelling of the coarsening process, the price of the (seemingly) precise result is a substantial debilitation of the reliability of the conclusions drawn.

Against this background, set-valued approaches, aiming at a proper reflection of the available information, have been gathering momentum, also becoming a popular topic at the ISIPTA symposia ([5, 26, 17, 32, 33], to name just a few contributions). In different areas of application concepts of cautious data completion emerged, where a classical procedure is extended by considering the set of all virtual precise observations in accordance with the coarse data (see, e.g., the exposition in [2], and the references therein). General investigations of coarse data from an imprecise-probability-based Bayesian point of view include [6, 36]; random set-based perspectives are developed for instance in

[8, 21]. Linear regression under metrical coarse data (interval data) is vividly discussed in the partial identification literature in the spirit of [19] (see also, e.g, [26], and the references therein). Mainly focusing on missing data, [34] suggests a framework for a systematic sensitivity analysis for statistical modelling under epistemic data imprecision. [5] introduces a profile likelihood approach for coarse data (for missing data see also [37]) and derive from it a uniform framework for robust regression analysis with imprecise data.

This paper will develop another likelihood-based (see, e.g., [4, § 6.3, 7.2.2] for a general introduction) approach and we will in addition briefly sketch Bayesian approaches in Section 3. Our work is strongly influenced by the methodology of partial identification, dealing with the trade-off between information and credibility by first using the empirical evidence only, i.e. using information implied by the data and including only those assumptions about which there exists a common consensus concerning their validity (e.g., [19, 28, 20]). Sensitivity analysis pursues the same goal, but proceeds in a different direction. While partial identification starts from total uncertainty and gradually adds assumptions, in the framework of sensitivity analysis the collection of all precise results from successively relaxed assumptions is considered. Thereby, the analysis is framed by a sensitivity parameter, which is not identified but suffices to identify the parameter of interest, (e.g., [34]).

Our paper is structured as follows. In the next section we fix the notation and formulate the problem setting more exactly for the cases considered in this paper: independent and identically distributed (i.i.d.) variables and logistic regression with a categorical covariate. The crucial technical argument underlying our paper (developed in general terms in Section 3) is to introduce an observation model and utilize invariance properties of the likelihood. In Section 4 we derive and discuss the set-valued estimators arising from a fully non-committal observation model, and we then turn to settings where this interval is narrowed when we benefit from the presence of additional auxiliary information. For technically handling this by sensitivity parameters, it is helpful to go to the other extreme, investigating point identifying additional assumptions in some special cases. For the homogeneous situation, after studying known coarsening in Section 5.1, we focus on the coarsening at random (CAR) assumption and illustrate the disastrous behaviour of the resulting point estimator when CAR is inappropriate (Section 5.2). Then in Section 5.3 we consider an extension of CAR and determine the corresponding ratio of coarsening probabilities as a sensitivity parameter. For the logistic regression case in Section 5.4 we work out that there is, as an alternative to CAR and its extensions, a further assumption refining the initial set of estimators to a precise result. This assumption is called subgroup independent coarsening and its generalization again can serve as a sensitivity parameter (Section 5.5). These sensitivity parameters frame a systematic sensitivity analysis, resulting in imprecise point estimators reflecting justifiable auxiliary information.

2 The Basic Setting

Let $Y_1, \dots, Y_n$ be a random sample of a categorical response variable of interest Y with realizations $y_1, \dots, y_n$ in sample space $\Omega_Y = \{1, \dots, j, \dots, K\}$. Problematically, some of those realizations are not known in a precise form, and thus only realizations $\mathscr{y}_1, \dots, \mathscr{y}_n$ of a sample $\mathcal{Y}_1, \dots, \mathcal{Y}_n$ of a random variable $\mathcal{Y}$ within sample space $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \emptyset$ can be observed, where $\mathcal{P}$ denotes the power set. The possible categories of $\mathcal{Y}$ constitute the singletons of $(\Omega_{\mathcal{Y}},\ \mathcal{P}(\Omega_{\mathcal{Y}}))$, with corresponding probability mass functions $p_{\mathscr{y}_i} = P(\mathcal{Y}_i = \mathscr{y}_i)$ $(i = 1, \dots, n)$. But as we are interested in the random variables $Y_1, \dots, Y_n$, our basic goal consists of gathering information about the individual probabilities $\pi_{i1} = P(Y_i = 1), \dots, \pi_{iK} = P(Y_i = K)$. Thereby, we assume throughout the paper that the coarsening process is error-free, in the sense that $\mathscr{y}_i \ni y_i$, $i = 1, \dots, n$.

We discuss the homogeneous case (i.i.d. case), in biometrical terms *prevalence* estimation, as well as situations with one precise categorical covariate X, in biometrical terms called *treatment*, with sample space Ω_X, being available. Both situations will be illustrated by means of the following example.

Running Example: *We refer to the data from the German panel study "Labor Market and Social Security" (PASS, wave 1, 2006/2007, [29]). As asking for the income may be regarded as a sensitive question and thus the response rate is expected to be low, in this study non-responders are required to report their income in classes starting from rather large classes that are narrowed by following questions. By proceeding in this way, anonymization is guaranteed in the level that is requested by the respondents and answers of different degrees of coarseness are obtained. Keeping things simple, here we refer to the data from question "HEK0700", where respondents are asked to report if their income Y is < 1000€ or ≥ 1000€ ($y_i \in \{<, \geq\}$; "$<$" and "$\geq$" abbreviating these classes, respectively) and our main goal is the estimation of $\pi_<$. As some respondents gave no suitable answer ("na") and cannot be allocated to one of the classes, partly only coarsened values of the variable $\mathcal{Y}$ are observed ($\mathscr{y}_i \in \{<, \geq, na\}$).*

Example, version 1: In order to illustrate the i.i.d. case, we only consider the reported answers of the income question, where 238, 835 and 338 respondents reported "<", "≥" and "na", respectively ($n_{<} = 238,\ n_{\geq} = 835,\ n_{\text{na}} = 338$).

In the case with categorical covariates, we here confine ourselves to one categorical covariate only, as this is technically equivalent to any finite set of categorical covariates. While in the i.i.d. case probabilities $\pi_{i1} = \pi_1, \ldots,\ \pi_{iK} = \pi_K$ are assumed to be independent of individual i, in the case with one covariate the probabilities $\pi_{i1} = P(Y_i = 1|X_i = x_i) = \pi_{x_i 1}, \ldots,\ \pi_{iK} = P(Y_i = K|X_i = x_i) = \pi_{x_i K}$ are influenced by individual i through the corresponding value of the covariate X_i. One of most generally applied models is the *multinomial logit model.* It describes the dependence of a categorical dependent variable Y of nominal scale on covariates X by

$$\pi_{ij} = P(Y_i = j|\mathbf{x}_i) = \frac{\exp(\beta_{j0} + \mathbf{x}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{s=1}^{K-1} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)} \tag{1}$$

$i = 1, \ldots, n$ for categories $j = 1, \ldots, K-1$ and by

$$\pi_{iK} = \big(1 + \sum\nolimits_{s=1}^{K-1} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)\big)^{-1} \tag{2}$$

with category specific regression coefficients, that is $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm})^T$ referring to m covariates and intercept β_{j0}. As we here address the case of one categorical covariate $X_i \in \{1, \ldots, c\}$, dummy coded variables $X_{i1}, \ldots, X_{im}$ with $m = c - 1$ are included into the model.[1]

It is common to summarize categorical data in contingency tables by reporting the counts for possible outcomes, where the covariates X are supposed to be in the rows (e.g., [31]). Thus, in our case the contingency table in Table 1 will be addressed. The number of observations with $\mathcal{Y} = \mathscr{y}$ and treatment group $X = x$ is denoted by $n_{x\mathscr{y}}$, where $n_0 = n_{0A} + n_{0B} + n_{0AB}$, $n_1 = n_{1A} + n_{1B} + n_{1AB}$, $n_A = n_{0A} + n_{1A}$, $n_B = n_{0B} + n_{1B}$ and $n_{AB} = n_{0AB} + n_{1AB}$.

Example, version 2: Illustrating the case with a categorical covariate, apart from the partial income knowledge, the receipt of the so-called Unemployment Benefit II (variable alg2abez; here denoted by UBII) is considered and serves in the model in Expressions (1) and (2) as covariate X_i, $i, \ldots, n$. The data are summarized in Table 2.

		$\mathcal{Y}$			
		A	B	AB	total
X	0	n_{0A}	n_{0B}	n_{0AB}	n_0
	1	n_{1A}	n_{1B}	n_{1AB}	n_1
	total	n_A	n_B	n_{AB}	n

Table 1: Contingency table that introduces used notation.

		income			
		<	≥	na	total
UBII	yes (0)	130	114	75	319
	no (1)	108	721	263	1092
	total	238	835	338	1411

Table 2: Contingency table to illustrate some results by means of the PASS data.

3 Sketch of the Basic Argument

This paper, similarly to [5, 37], relies on the likelihood as the fundamental concept to derive parameter estimators under epistemic data imprecision, but looks at it from a different angle. In order to support the appropriate incorporation of the available information provided by the data and the background knowledge, we explicitly formulate, and utilize, an *observation model* relating the observable level and the ideal level. The observation model is a set $\mathcal{Q}$ of (precise) coarsening probabilities,[2] and thus the medium to specify carefully and flexibly the available information about the coarsening process.

By virtue of the theorem of total probability, the elements of $\mathcal{Q}$ relate the probability distribution of the imprecise observation $\mathcal{Y}$ to the distribution of the underlying latent variable Y (and, if present, certain covariates).

Parametrizing the distributions, again possibly after splitting with respect to certain covariate values, let ϑ (the various p's in the following sections) and η (the various π's below) be the parameters determining the distribution of $\mathcal{Y}$ and Y, respectively, and let ζ be the parameter characterising the elements of $\mathcal{Q}$ (the various q's, possibly constrained by the specified constraints: $\big(q_{\mathscr{y}|y} := P(\mathcal{Y} = \mathscr{y}|Y = y)\big)_{(\mathscr{y} \in \Omega_{\mathcal{Y}}, y \in \Omega_Y)}$ in the i.i.d. case, while in the regression context the coarsening mechanisms generally also depend on the values of X_i, i.e., $(q_{\mathscr{y}|xy} := P(\mathcal{Y} = \mathscr{y}|X = x, Y = y)_{(\mathscr{y} \in \Omega_{\mathcal{Y}}, y \in \Omega_Y, x \in \Omega_X)}$ has to be considered).

Then we can describe the relationship between $\gamma := (\eta^T, \zeta^T)^T \in \Gamma$ and $\vartheta \in \Theta$ via the mapping $\Phi : \Gamma \to \Theta$, $\gamma \mapsto \vartheta$. Figure 1 and the running example illustrate

[1] Dummy variable X_{il} ($l = 1, \cdots, m$) attains value 1 if the l-th category is chosen by individual i, otherwise it is 0. In this way, reference category c is represented by all dummy variables being 0.

[2] More precisely, $\mathcal{Q}$ is a generalized transition kernel, consisting of credal sets indexed by the values of Y.

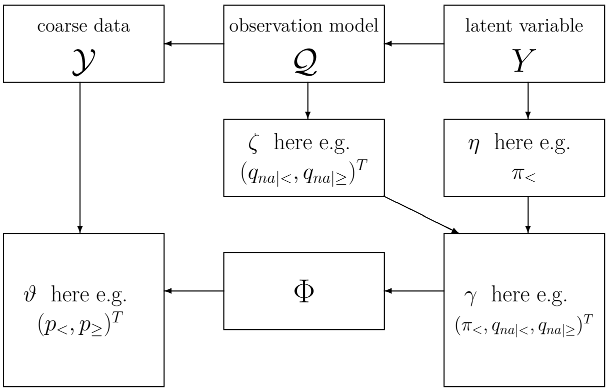

Figure 1: Observable and latent variable and the corresponding parameters.

this mapping $\Phi(\cdot)$ and all parameters involved.

Example, version 1 (cont.): The mapping $\Phi(\cdot)$ with arguments $\zeta = (q_{\text{na}|<}, q_{\text{na}|\geq})^T$ and $\eta = \pi_<$ establishes a connection to the parameters determining the probabilities of the observable income variable $\mathcal{Y}$, namely $\vartheta = (p_<, p_\geq)^T$.

In a first step (Section 4), we will only assume that the coarsening process is error-free and therefore take $\mathcal{Q}$ as the set of all coarsening mechanisms compatible with error-freeness. Then (Section 5), by using auxiliary information, we sharpen this set $\mathcal{Q}$. Note that we do neither assume anything about the plausibility of different elements ζ of $\mathcal{Q}$ nor do we treat different $y \in \mathscr{Y}$ as differently plausible. To derive the estimators, the invariance of the likelihood under parameter transformations is crucial: evaluating the likelihood in terms of γ and in terms of $\vartheta = \Phi(\gamma)$ is equivalent here. Our random set modelling will allow us to determine the ML-estimator $\hat{\vartheta}$ of ϑ, which moreover, apart from trivial extreme cases, can be shown to be single-valued. Then the possibly set-valued maximum-likelihood estimator for γ is obtained as

$$\hat{\Gamma} = \left\{ \gamma \,\middle|\, \Phi(\gamma) = \hat{\vartheta} \right\} \qquad (3)$$

(see also [5, Section 2]). Thus, adapting the concept of maximum likelihood (ML) estimators to a persistent set-based perspective and to random set-based situations, we achieve a general and powerful framework for handling coarse categorical data via the mapping $\Phi(\cdot)$. If $\Phi(\cdot)$ is injective, then $\hat{\Gamma}$ is a singleton as well, and γ so-to-say empirically point identified; otherwise $\hat{\Gamma}$ is set-valued in the literal sense and γ empirically partially identified.

This compares to other approaches: A classical Bayesian analysis would put some prior on ζ and on η (cf., e.g., [23, 14]) while a generalized Bayesian analysis would replace one or both priors by a set of priors. This can be seen as imposing imprecise priors on ζ and on η. The non-committal analysis would start with a near-ignorance prior, for instance based on Dirichlet distributions adapting [35]'s imprecise Dirichlet model, and auxiliary information can be expressed by smaller credal sets; compare also the general Bayesian treatment of incomplete information in [6, 36]. Partially differently, in [3, Section 4.4.] an approach is presented that puts a precise prior on η and no prior on ζ and models the coarsening process with a multivalued mapping. This may be seen as imposing a vacuous imprecise probability on ζ. In another direction, one could impose some prior knowledge w.r.t. the imprecise data point $\mathscr{Y}$ by assuming different $y \in \mathscr{Y}$ as differently plausible. This can be done for example by imposing a possibility distribution on y (cf., e.g., [9, Section 3.2.]) or constructing observations directly by data augmentation (cf., e.g., [18]).

The dimension of the parameter vectors η and ζ increases substantially with the cardinality of Ω_Y and Ω_X. In the i.i.d. case $m = (\sum_{z=1}^{|\Omega_Y|} \binom{|\Omega_Y|}{z} \cdot z) - 1$ or equivalently $m = K \cdot 2^{K-1} - 1$ parameters have to be estimated, where in the case with one covariate this number even increases to $|\Omega_X| \cdot m$. Thus, for reasons of conciseness of presentation, we confine detailed explanations and derivations on the special, yet still representative cases of a binary response variable Y with sample space $\Omega_Y = \{A,\ B\}$ and observations within $\Omega_{\mathcal{Y}} = \{A,\ B,\ AB\}$, as well as a binary precise categorical covariate X with values 0 and 1. Then the underlying model expressed in Expression (1) and (2) is called *logit model*. As the inclusion of more than one dummy variable simply leads to an increase of the number of subgroups, all results can be transferred straightforwardly to more general cases, namely cases with more than one non-binary covariates. Furthermore, the main results not only will be shown for the situation of a binary Y, where coarsening corresponds to missingness, but also in a general way.

4 Maximum Likelihood Estimation without Additional Information

In this section we derive the maximum likelihood estimators for the case where no additional information on the coarsening process is available, i.e. there are no constraints on the elements of $\mathcal{Q}$. A crucial step is to rely on the random set view that treats data imprecision as a change of the sample space with corresponding random variables $\mathcal{Y}_i$, $i = 1, \ldots, n$, which then lead to multinomially distributed variables with parameter ϑ for the counts based on the new sample space. According to the argumentation in Section 3, the resulting likelihood in ϑ, and the estimator derived from maximizing it, will then be related to the parameters of the distribution of the latent variable (and the observation model). As just discussed, we explain the construction in some detail for the representative special cases with $\Omega_Y = \{A,\ B\}$ (and $\Omega_X = \{0,1\}$) and then report the general results.

4.1 Estimation in the i.i.d. Case

Considering categorical i.i.d. random variables $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$ with realizations $\mathscr{y}_1, \ldots, \mathscr{y}_n$ in the sample space $\Omega_{\mathcal{Y}} = \{A,\ B,\ AB\}$, we obtain the following likelihood function for the parameter $\vartheta = (p_A, p_B)^T$ given the data, summarized by the counts n_A, n_B and n_{AB} (with $p_{AB} = 1 - p_A - p_B$):[3]

$$\begin{aligned} L(\vartheta) &= L(p_A,\ p_B) = L(p_A,\ p_B||\mathscr{y}_1, \ldots, \mathscr{y}_n) \quad (4)\\ &= P(\mathscr{y}_1, \ldots, \mathscr{y}_n||p_A,\ p_B) \propto p_A^{n_A} \cdot p_B^{n_B} \cdot p_{AB}^{n_{AB}} . \end{aligned}$$

For $n = n_A + n_B + n_{AB} > 0$ this likelihood is uniquely maximized by the relative frequencies (see [25]),

$$\hat{p}_A^{(MLE)} = \frac{n_A}{n}, \qquad \hat{p}_B^{(MLE)} = \frac{n_B}{n}, \tag{5}$$

and thus $\hat{p}_{AB}^{(MLE)} = 1 - \hat{p}_A^{(MLE)} - \hat{p}_B^{(MLE)} = \frac{n_{AB}}{n}$.

Essentially, we are interested in the parameter $\eta = \pi_A$ determining the probabilities of the true, but unobserved variable Y being equal to particular categories and the associated maximum likelihood estimator. Those probabilities of interest, in our case π_A and $\pi_B = 1 - \pi_A$, can be related with probabilities p_A, p_B and p_{AB} corresponding to the observable variables by

$$\begin{aligned} p_A &= (1 - q_{AB|A}) \cdot \pi_A\,, \\ p_B &= (1 - q_{AB|B}) \cdot (1 - \pi_A)\,, \end{aligned} \tag{6}$$

where $p_{AB} = q_{AB|A} \cdot \pi_A + q_{AB|B} \cdot (1 - \pi_A)$ results from the law of total probability.

This means that the likelihood in terms of $\vartheta = (p_A, p_B)^T$ in Expression (4) and in terms of $\gamma = (\pi_A, q_{AB|A}, q_{AB|B})^T$, coincide, indeed.

By the invariance of the likelihood under parameter transformations, Expressions (5) and (6) can be combined, resulting in the following system of equations:

$$\begin{aligned} (1 - \hat{q}_{AB|A}) \cdot \hat{\pi}_A &= \frac{n_A}{n} = \hat{p}_A^{(MLE)}, \\ (1 - \hat{q}_{AB|B}) \cdot (1 - \hat{\pi}_A) &= \frac{n_B}{n} = \hat{p}_B^{(MLE)}, \\ \hat{q}_{AB|A} \cdot \hat{\pi}_A + \hat{q}_{AB|B} \cdot (1 - \hat{\pi}_A) &= \frac{n_{AB}}{n} = \hat{p}_{AB}^{(MLE)}. \end{aligned} \tag{7}$$

For reasons of redundancy we can leave the third equation out of consideration. As there typically are multiple triples $\hat{\gamma} = (\hat{\pi}_A,\ \hat{q}_{AB|A},\ \hat{q}_{AB|B})^T$ that lead to the same values of $\hat{\vartheta} = (\hat{p}_A^{(MLE)},\ \hat{p}_B^{(MLE)})^T$, the mapping $\Phi:\ [0,\ 1]^3 \to [0,\ 1]^2$ with

$$\Phi\begin{pmatrix} \pi_A \\ q_{AB|A} \\ q_{AB|B} \end{pmatrix} = \begin{pmatrix} \pi_A \cdot (1 - q_{AB|A}) \\ (1 - \pi_A) \cdot (1 - q_{AB|B}) \end{pmatrix} = \begin{pmatrix} p_A \\ p_B \end{pmatrix} \tag{8}$$

(cf. Figure 1 for the case of the running example) connecting both parametrizations in general is not injective. Thus the maximum likelihood estimate $\hat{\Gamma}$ from Expression (3) is set-valued in the literal sense. Points in this set are constrained through the relationships in (7), and thus $\hat{\Gamma}$ is not a cuboid in $[0,\ 1]^3$. Building the one dimensional projections, set-valued estimators of the single components of γ are obtained via

$$\begin{aligned} \hat{\pi}_A &\in \left[\frac{n_A}{n},\ \frac{n_A + n_{AB}}{n}\right], \\ \hat{q}_{AB|A} &\in \left[0,\ \frac{n_{AB}}{n_A + n_{AB}}\right], \end{aligned} \tag{9}$$

and analogously for $\hat{q}_{AB|B}$, where $\frac{0}{0} := 1$.

Extending the discussion here to the general case of $\Omega_Y = \{1, \ldots, K\}$ and the corresponding $\Omega_{\mathcal{Y}}$, the estimators in Expression (9) generalize to

$$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n},\ \frac{\sum_{\mathscr{y} \ni y} n_{\mathscr{y}}}{n}\right] \quad \hat{q}_{\mathscr{y}|y} \in \left[0,\ \frac{n_{\mathscr{y}}}{n_{\{y\}} + n_{\mathscr{y}}}\right], \tag{10}$$

(where as above $\frac{0}{0} := 1$) for all $y \in \Omega_y = \{1, \ldots, K\}$ and all $\mathscr{y} \in \Omega_{\mathcal{Y}}$ such that $\{y\} \subset \mathscr{y}$.[4]

Example, version 1 (cont.): Applying Expression (10) to our example, one obtains

$$\hat{\pi}_< \in \left[\frac{238}{1411},\ \frac{238 + 338}{1411}\right] = [0.17,\ 0.41]\ .$$

4.2 Logistic Regression with a Categorical Covariate

Now we consider the heterogeneous situation expressed by a discrete covariate X, which also has been depicted in Table 1. Again we can derive set-valued estimators of the parameters of interest $\eta = (\pi_{0A}, \pi_{1A})^T$ (and the auxiliary parameter ζ characterizing the coarsening mechanisms) by taking the random set perspective, setting up the corresponding likelihood function and

[3] In the following, we will use the abbreviated notation of the likelihood without referring to the data.

[4] The estimators of the probability components of the distribution of Y_i prove to be the same as arising from a belief functions like construction of empirical probabilities and also coincide with the estimator obtained from cautious data completion, plugging in all potential precise sample outcome compatible with the observations $\mathscr{y}_1, \ldots, \mathscr{y}_n$ (see, e.g., [2])

applying the appropriate parameter transformations. Proceeding in this way, for fixed treatment group x the cell counts $(n_{xA}, n_{xB},\ n_{xAB})$ follow a multinomial distribution, i.e. $(n_{xA}, n_{xB}, n_{xAB}) \sim M(n_x, (p_{xA}, p_{xB}, p_{xAB}))$ with conditional probabilities $p_{x\mathscr{y}} = P(\mathcal{Y} = \mathscr{y} | X = x)$ (see [31, 1]).[5] Therefore, the corresponding likelihood function is given by

$$\begin{aligned} L(\vartheta) &= L(p_{0A},\ p_{1A},\ p_{0B},\ p_{1B}) \\ &\propto p_{0A}^{n_{0A}} \cdot p_{0B}^{n_{0B}} \cdot p_{0AB}^{n_{0AB}} \cdot p_{1A}^{n_{1A}} \cdot p_{1B}^{n_{1B}} \cdot p_{1AB}^{n_{1AB}}. \end{aligned} \tag{11}$$

For $n_x > 0$ the maximum likelihood estimators for the parameters are unique and given by (see [25])

$$\hat{p}_{x\mathscr{y}}^{(MLE)} = \frac{n_{x\mathscr{y}}}{n_x}, \text{ for } x \in \{0, 1\}.$$

Analogously to Section 4.1, we consider the mapping, which connects both parametrizations, $\Phi :\ [0,\ 1]^6 \to [0,\ 1]^4$ with

$$\Phi \begin{pmatrix} \pi_{0A} \\ \pi_{1A} \\ q_{AB|0A} \\ q_{AB|1A} \\ q_{AB|0B} \\ q_{AB|1B} \end{pmatrix} = \begin{pmatrix} \pi_{0A} \cdot (1 - q_{AB|0A}) \\ \pi_{1A} \cdot (1 - q_{AB|1A}) \\ (1 - \pi_{0A}) \cdot (1 - q_{AB|0B}) \\ (1 - \pi_{1A}) \cdot (1 - q_{AB|1B}) \end{pmatrix} = \begin{pmatrix} p_{0A} \\ p_{1A} \\ p_{0B} \\ p_{1B} \end{pmatrix} \tag{12}$$

(cf. Figure 1) and observe that in this case it is also not injective and thus $\hat{\Gamma}$, constructed along the line of (3), is strictly set-valued, too. Illustrating $\hat{\Gamma}$ again by the corresponding projections along the axes, we obtain for given value $x \in \{0, 1\}$ in the general case with more than two categories in Y, i.e. $y \in \Omega_Y = \{1, \ldots, K\}$ and $\mathscr{y} \in \Omega_{\mathcal{Y}}$ with $\{y\} \subset \mathscr{y}$,

$$\hat{\pi}_{xy} \in \left[\frac{n_{x\{y\}}}{n_x}, \frac{\sum\limits_{\mathscr{y} \ni y} n_{x\mathscr{y}}}{n_x} \right],\ \hat{q}_{\mathscr{y}|xy} \in \left[0,\ \frac{n_{x\mathscr{y}}}{n_{x\{y\}} + n_{x\mathscr{y}}} \right], \tag{13}$$

where again $\frac{0}{0} := 1$.[6]

Example, version 2 (cont.): Applying Expression (13) to our example, one obtains

$$\begin{aligned} \hat{\pi}_{0<} &\in \left[\frac{130}{319}, \frac{130 + 75}{319} \right] = [0.41,\ 0.64], \\ \hat{\pi}_{1<} &\in \left[\frac{108}{1092}, \frac{108 + 263}{1092} \right] = [0.10,\ 0.34]. \end{aligned}$$

By recurring on the relation defined in Expression (1) and (2), and utilizing the injectivity of the logistic function, the likelihood function considered here can also be uniquely expressed in terms of the regression coefficients. In this way, instead of the estimators $\hat{\pi}_{0A}$ and $\hat{\pi}_{1A}$ determined by Expression (13), equivalently one can consider the estimators

$$\begin{aligned} \hat{\beta}_{A0} &\in \left[\log\left(\frac{n_{0A}}{n_{0B} + n_{0AB}} \right), \log\left(\frac{n_{0A} + n_{0AB}}{n_{0B}} \right) \right] \\ \hat{\beta}_{A} &\in \left[\log\left(\frac{n_{1A} \cdot (n_{0B} + n_{0AB})}{n_{0A} \cdot (n_{1B} + n_{1AB})} \right), \right. \\ &\qquad \left. \log\left(\frac{n_{0B} \cdot (n_{1A} + n_{1AB})}{n_{1B} \cdot (n_{0A} + n_{0AB})} \right) \right], \end{aligned} \tag{14}$$

assuming all expressions to be well-defined.

Example, version 2 (cont.): In terms of the regression coefficients, we obtain the estimates $\hat{\beta}_{<0} \in [-0.37,\ 0.59]$ and $\hat{\beta}_{<} \in [-1.83,\ -1.25]$.

Interpreting the indeterminate sign of intercept $\beta_{<0}$, one notes that for the group of persons that receives UBII (i.e. $X = 0$) the chance of being in the lower income group (< 1000€) in comparison to being in the higher income group (≥ 1000€) varies between $\exp(-0.37) = 0.69$ and $\exp(0.59) = 1.89$. In this way, one cannot judge the impact of the UBII on the dependent variable income without implying further assumptions about the coarsening. Unjustifiably ignoring the coarsening (see Section 5.2) pretends a particular sign of the regression coefficients. This corroborates the importance of including all imaginable coarsening mechanisms for obtaining a trustworthy result, which will be discussed now more in detail.

5 Reliable Incorporation of Auxiliary Information: Sensitivity Parameters and Partial Identification

The set-valued estimators from Expression (9) (and analogously from Expression (13)) are a typical application of the methodology of partial identification, emphasizing that only justified assumptions should be made which do not have to induce point identified parameters, but at least identify the parameter of interest in parts compared to the set of parameters that seemed to be possible in the beginning of the analysis (e.g., [19]). In this way, the trivial bounds [0, 1] on the probabilities have been refined substantially. In the spirit of partial identification and sensitivity analysis we can further refine the analysis if, and also only if, auxiliary information beyond the empirical evidence is available. Vansteelandt et al. [34] suggests to determine a sensitivity parameter δ in some range Δ under which the problem is identified and then to calculate the parameter of interest η for different values of the sensitivity parameter, where the whole region of the

[5] This corresponds to a product-multinomial sampling scheme (e.g. [31, 1]).

[6] Reminiscing about the derivation given here, we see that the categorical covariate case for the logistic model – in strict contrast to the continuous case (see Section 6) – in essence consists of a subgroup-specific consideration of the i.i.d. case.

resulting parameters of interest is called Ignorance Region $ir(\eta, \Delta)$ and the corresponding region of estimates Honestly Estimated Ignorance Region (HEIR) $\hat{ir}_n(\eta, \Delta)$. In order to account for statistical uncertainty due to finite sample size as well, in context of sensitivity analysis uncertainty regions are addressed that either can be constructed as covering the parameter of interest or the whole ignorance region with a probability of at least $(1 - \alpha)$ [13, 34].

To handle the inclusion of reliable information technically, we start with distinguishing and investigating point identifying additional assumptions, in order to utilize them as a technical means to derive sensitivity parameters, governing the incorporation of additional information.

Due to the fact that the imprecise point estimators in Expression (13) directly result from considering Expression (9) in a subgroup specific way, in Section 5.1 to Section 5.3 the detailed presentation is confined on the i.i.d. case. In Section 5.4, considering explicitly the regression model, another point-identifying assumption is suggested, where again the corresponding generalization may be used as a sensitivity parameter which allows the inclusion of partial knowledge.

5.1 Known Coarsening

If one or both coarsening parameters $q_{AB|A}$ and $q_{AB|B}$ are known (and different from 1), one can conclude directly that the corresponding mapping $\Phi(\cdot)$ from (8) is injective as in this case the parameter π_A can be uniquely related to the parameter p_A. Therefore, the set-valued estimator for π_A specified in Expression (9) can be shrunk to a single-valued estimator. The exact values of the coarsening parameters are most often unknown, but in case there is material information available that allows to bound them in non-trivial intervals, the consideration here gives a first way to perform a systematic sensitivity analysis. In most situations however such direct bounds will not be available. Therefore we look for alternative ways to introduce auxiliary knowledge.

5.2 Coarsening at Random (CAR)

If the coarsening is non-stochastic, the underlying degree of coarsening is predetermined and known. For instance, if respondents are requested to give their answer in a grouped way and we assume that all respondents answer correctly, then the coarsening is predefined in the sense that there is a unique coarsened outcome for every true answer. In the context of distinguishing between non-stochastic and stochastic coarsening mechanisms, Heitjan and Rubin [12] investigated under which properties the corresponding likelihood can be simplified to the so-called grouped likelihood and introduced the concept of *coarsening at random (CAR)*. This is a simplifying property requesting that the probability $q_{\mathcal{Y}|y}$ is constant, no matter which true value y is underlying as long as it fits to the observed value $\mathcal{Y}$. Illustrated by the running example, CAR postulates that the probability of giving no suitable answer should not depend on the true income category, which contradicts practical experiences (e.g., [16]). In the dichotomous situation of this example we are then actually concerned with the assumption of missing at random (MAR) [18], which can be regarded as a special case of CAR.

Focusing again on the i.i.d. case, incorporating the CAR assumption of $q_{AB|A} = q_{AB|B}$ into the likelihood and in the observation model specifying $\Phi(\cdot)$, the situation simplifies substantially. Indeed, Φ is (almost) injective now, and we get the empirically point identified estimators, corresponding to having simply ignored the units with coarse values:

$$\begin{aligned} \hat{\pi}_A &= \frac{n_A}{n_A + n_B} \\ \hat{q}_{AB|A} &= \hat{q}_{AB|B} = \frac{n_{AB}}{n_A + n_B + n_{AB}} . \end{aligned}$$

There are ideal-type situations in which CAR can be justified indeed.[7] Nevertheless, this assumption must be treated with greatest care. Deviating from such an ideal-type situation and wrongly assuming CAR can lead to a bias of an extent that for sure destroys the relevance of the analysis, as is also illustrated in Figure 2. There the estimation of π_A under obstinately assumed CAR but varying coarsening probabilities is evaluated by the median relative empirical bias $\frac{\hat{\pi}_A - \pi_A}{\pi_A}$ based on 100 simulated datasets (here with $\pi_A = 0.6$).[8] The absolute value of the relative median bias increases the more one deviates from the case of CAR, indeed, up to a median relative bias of almost 80%.

5.3 Ratio of Coarsening Parameters

In our context the paper by Nordheim [22] obtains new importance. He considers the ratio between different mechanisms in the context of non-randomly missing and misclassified data. By fixing the ratio between the coarsening probabilities the corresponding maximum likelihood problem leads to quadratic equations, where

[7] For instance, rounding, type I censoring, which is present if the censoring times are fixed, and progressive type II censoring, which investigates censoring after the fixed d-th failure, in their pure form are CAR [15, 11].

[8] Thereby, in all addressed situations characterized by different true underlying coarsening mechanisms ($q_{AB|A}$ and $q_{AB|B}$ varying between 0.1 and 0.9 in equidistant breaks of 0.1, respectively), the assumption of CAR is involved into the estimation by plugging $q_{AB|A} = q_{AB|B}$ into the likelihood that is maximized.

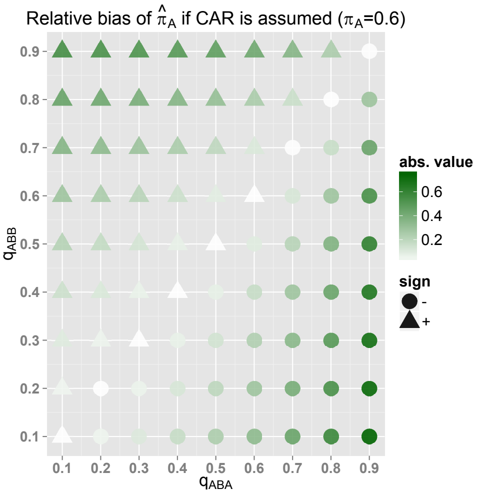

Figure 2: Consequences for the median relative bias of $\hat{\pi}_A$ if there is a deviation from assumed CAR.

one solution is contained in the interval of $\hat{\pi}_A$ from Expression (9), while the other solution lies outside of [0, 1] (cf. [22, p. 774]). Here we set $R = \frac{q_{B|B}}{q_{A|A}} = \frac{1-q_{AB|B}}{1-q_{AB|A}}$, slightly modifying the ratio of Nordheim by referring to the probabilities of the complementary events. Treating this ratio between the probabilities of precise observation fixed and including it into the likelihood in Section 4.1, unique, empirically point identified estimators are obtained as

$$\begin{aligned}\hat{\pi}_A &= \frac{n_A \cdot R}{n_B + n_A \cdot R}, \qquad (15)\\ \hat{q}_{AB|A} &= \frac{n_B \cdot (R-1) + n_{AB} \cdot R}{n \cdot R}\end{aligned}$$

containing CAR as the special case $R = 1$. As in the case of CAR, the impact of assuming a wrong value of R has been investigated (results are available on request, see also [22]), where again a substantial bias can occur. The fact that there a similar variance of the estimators is obtained independently of the amount of deviation from the true value of R shows drastically that such deviations do not increase statistical uncertainty in the traditional sense and thus cannot be discovered by a traditional statistical analysis.

Because the parameter of interest π_A is identified given the typically unknown value of R, the ratio R can be used as a sensitivity parameter. In many cases it might be difficult to gain information about the exact value of R, but it seems quite realistic that a rough evaluation of the magnitude of R can be derived from material considerations, former studies or experiments. Thus, it is interesting to investigate the gain of information resulting from implying a factor R that is roughly known only, compared to the situation without any additional assumptions.[9] Considering the ratio R as a sensitivity parameter leads to the HEIRs.[10]

5.4 Subgroup Independent Coarsening

In the situation with covariates, there is apart from CAR, i.e. $\hat{q}_{AB|xA} = \hat{q}_{AB|xB}$, an alternative kind of uninformative coarsening, namely the independence of the underlying covariate value. Illustrated by the running example, imposing this kind of assumption means that answering in a coarse form, i.e., giving no suitable answer, does not depend on the receipt of unemployment benefit. As the receipt of unemployment benefit depends on the income, and the value of the income may influence the non-response to the income question (cf. Section 5.2), this assumption should be treated with particular caution here.

We will establish injectivity of the corresponding mapping $\Phi(\cdot)$ under an intuitive regularity condition and then, analogously to the procedure in Sections 5.2 and 5.3, this idea will be generalized in Section 5.5 by again considering the corresponding fraction as a sensitivity parameter. Imposing such *subgroup independent coarsening*

$$\begin{aligned} q_{AB|0A} &= q_{AB|1A} =: q_{AB|A} \qquad (16)\\ q_{AB|0B} &= q_{AB|1B} =: q_{AB|B},\end{aligned}$$

in the estimation problem of Section 4.2, the mapping $\Phi(\cdot)$ from Expression (12) is now injective[11] if restricted to the arguments $(\pi_{0A}, \pi_{1A}, q_{AB|A}, q_{AB|B})^T \in (0,1)^4$ such that

$$\pi_{0A} \notin \{0,1\},\ \pi_{1A} \notin \{0,1\}\ \ \underline{\text{and}}\ \ \pi_{0A} \neq \pi_{1A}. \qquad (17)$$

One obtains the following unique estimators

$$\begin{aligned}\hat{\pi}_{0A} &= \frac{n_{0A}}{n_0}\,\frac{n_{1B}n_0 - n_1 n_{0B}}{n_{0A}n_{1B} - n_{0B}n_{1A}}, \qquad (18)\\ \hat{\pi}_{1A} &= \frac{n_{1A}}{n_1}\,\frac{n_{1B}n_0 - n_1 n_{0B}}{n_{0A}n_{1B} - n_{0B}n_{1A}},\\ \hat{q}_{AB|A} &= 1 - \frac{n_{0A}n_{1B} - n_{0B}n_{1A}}{n_{1B}n_0 - n_1 n_{0B}},\\ \hat{q}_{AB|B} &= 1 - \frac{n_{0A}n_{1B} - n_{0B}n_{1A}}{n_{0A}n_1 - n_{1A}n_0},\end{aligned}$$

[9] An example is given in the preliminary version of a technical report available at `http://www.statistik.lmu.de/~jplass/forschung.html`

[10] In more general cases of $|\Omega_Y| > 2$, the relations between the precise observation probabilities are not sufficient and relations concerning different coarsening mechanisms have to be known in order to obtain point identified estimators. More detailed information can be found in the preliminary version of a technical report cited in footnote 9.

[11] A proof of the injectivity of Φ in this situation is given in the preliminary version of a technical report cited in footnote 9. The case of $\pi_{0A} = \pi_{1A}$ reproduces the i.i.d. case, where there are multiple solutions.

when these are well-defined and inside the interval [0, 1]. Otherwise the maximum likelihood estimation is more challenging, but it can be shown that asymptotically ($n \to \infty$) the estimators of Expression (18) typically for all cases satisfying Expression (17) will be in [0, 1]. It has to be re-emphasized that in practical applications one must carefully reflect the plausibility of the subgroup independent coarsening assumption of Expression (16). In addition, the restrictions

$$p_{0A} \leq \frac{P(X=0)\cdot p_{1B} - p_{0B}\cdot P(X=1)}{p_{1B} - p_{0B}\cdot \frac{p_{1A}}{p_{0A}}} \leq 1 - p_{0B}$$

offer, at least under large sample sizes, a possibility to check whether the subgroup independent coarsening is appropriate at all.

5.5 A Generalization of Subgroup Independent Coarsening

There are situations in which one might have an idea about the relative magnitude of the probabilities of precise observations in both subgroups. For instance, knowledge from former studies could be available concerning the question whether respondents who do receive Unemployment Benefit II rather report their income class in a precise or a coarse way compared to the respondents that do not receive this benefit.

Analogously to the generalization of CAR in Section 5.3, we now generalize the assumption of subgroup independent coarsening by considering the ratio between the subgroup specific probabilities of precise observation, i.e., $R_1 = \frac{q_{A|1A}}{q_{A|0A}}$ and $R_2 = \frac{q_{B|1B}}{q_{B|0B}}$, where the case of $R_1 = R_2 = 1$ corresponds to assuming subgroup independent coarsening. As in Section 5.4, the mapping $\Phi(\cdot)$ from Expression (12) is injective for all cases in Expression (17) and thus unique estimators result.[12] Again, inclusion of partial knowledge is possible by regarding R_1 and R_2 as sensitivity parameters and considering all estimators resulting from incorporating a region of plausible values R_1 and R_2.

6 Concluding Remarks

We presented a maximum likelihood analysis of categorical data under epistemic data imprecision. Our approach working with possibly set-valued maximum likelihood estimators overcomes the dilemma of the precise probability based approaches, often damned to debilitate conclusions by the need to incorporate unjustified formal assumptions to ensure identifiability of parameters. The explicit reliance on an observation model specifying the coarsening process allows us to incorporate properly auxiliary information whenever it is present, in order to refine appropriately estimates derived from the empirical evidence alone.

The crucial arguments were developed, mutatis mutandis, for the i.i.d. case as well as a logistic regression based on one (or more) categorical covariates. From the applied point of view, an extension to metrical covariates is highly desirable. Although then a subgroup specific investigation is not possible any more, appropriate generalizations seem achievable in further work, especially when sensitivity parameters can be determined. However, to allow estimation of the underlying distribution from the data and to maintain the metric character, (partially) parametric modelling is needed. This implicitly restricts the set of distributions considered and in particular raises further issues in the understanding of statistical models as discussed, e.g., in [26, Sec. 3.1] for linear regression modelling.

In addition to this, the invariance property of the likelihood under different parametrizations, which is the technical basis of our results, offers two further directions of generalization. Further work may utilize these relationships beyond maximum likelihood estimation, in order to derive likelihood-based hypotheses tests and regions taking finite sample variability into account explicitly. These estimators also should be compared to confidence intervals derived along the lines of [34] when an appropriate sensitivity parameter could be determined.

Other areas of further research include a deeper investigation of the alternative generalized Bayesian (and possibilistic) approaches briefly mentioned in Section 3 as well as the consideration of other "deficiency" processes, most notably misclassification, which can be formalized in a very similar way. Our methodology thus also offers an alternative to, and a generalization to logistic regression of, recent work on misclassification from a partial identification perspective [20, 17].

Acknowledgements We are very grateful for the helpful remarks of three anonymous reviewers and appreciate their open sharing of stimulating ideas very much.

References

[1] A. Agresti. *Categorical Data Analysis.* 3rd edn., Wiley, 2013.

[2] T. Augustin, G. Walter, F. Coolen. Statistical inference. In: T. Augustin, F. Coolen, G. de Cooman, M. Troffaes (eds.): *Introduction to Imprecise Probabilities*, Wiley, 2014, pp. 135–189.

[3] A. Benavoli. Belief function and multivalued mapping robustness in statistical estimation. *Int. J. Approx. Reasoning*, 55:311–329, 2014.

[12] They are given in the preliminary version of the technical report cited in footnote 9.

[4] G. Casella, R. Berger. *Statistical Inference.* 2nd edn., Duxbury, 2002.

[5] M. Cattaneo, A. Wiencierz. Likelihood-based imprecise regression. *Int. J. Approx. Reasoning*, 53:1137–1154, 2012. [based on an ISIPTA '11 paper]

[6] G. de Cooman, M. Zaffalon. Updating beliefs with incomplete observations. *Artif. Intell.*, 159:75–125, 2004.

[7] I. Couso, D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *Int. J. Approx. Reasoning*, 55:1502–1518, 2014.

[8] I. Couso, D. Dubois, L. Sánchez. *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables.* Springer, Cham, 2014.

[9] T. Denoeux. Likelihood-based belief function: justification and some extensions to low-quality data. *Int. J. Approx. Reasoning*, 55:1535–1547, 2014.

[10] A. Dobra, S. Fienberg. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *P. Natl. Acad. Sci. USA*, 97: 11885–11892, 2000.

[11] D. Heitjan. Ignorability and coarse data: Some biomedical examples. *Biometrics*, 49:1099–1109, 1993.

[12] D. Heitjan, D. Rubin. Ignorability and coarse data. *Ann. Stat.*, 19:2244–2253, 1991.

[13] G. Imbens, C. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72:1845–1857, 2004.

[14] T. Jiang, J. Dickey, Bayesian methods for categorical data under informative censoring, *Bayesian Anal.*, 3:541–553, 2008.

[15] J. Kalbfleisch, R. Prentice. *The Statistical Analysis of Failure Time Data.* 2nd edn., Wiley, 2002.

[16] A. Korinek, J. Mistiaen, M. Ravallion. Survey nonresponse and the distribution of income. *J. Econ. Inequal.*, 4:33–55, 2006.

[17] H. Küchenhoff, T. Augustin, A. Kunz. Partially identified prevalence estimation under misclassification using the kappa coefficient. *Int. J. Approx. Reasoning* 53:1168–1182, 2012. [based on an ISIPTA '11 paper]

[18] R. Little, D. Rubin, *Statistical Analysis with Missing Data.* 2nd edn., Wiley, 2002.

[19] C. Manski. *Partial Identification of Probability Distributions.* Springer, 2003.

[20] F. Molinari. Partial identification of probability distributions with misclassified data. *J. Econom.*, 144:81–117, 2008.

[21] H. Nguyen, B. Wu. Random and fuzzy sets in coarse data analysis. *Comput. Stat. Data. An.*, 51:70–85, 2006.

[22] E. Nordheim. Inference from nonrandomly missing categorical data: An example from a genetic study on Turner's syndrome. *J. Am. Stat. Assoc.*, 79:772–780, 1984.

[23] D. Paulino, C. De B. Pereira. Bayesian analysis of categorical data informatively censored. *Commun. Stat., Theory Methods*, 21:2689–2705, 1992.

[24] J. Plass, P. Fink, N. Schöning, T. Augustin. Statistical modelling in surveys without neglecting "the undecided": Multinomial logistic regression models and imprecise classification trees under ontic data imprecision. *under revision for ISIPTA '15.* See also: *Techn. Rep., 179, Dep. Statistics, LMU Munich,* 2015 (url: www.epub.ub.uni-muenchen.de/23816).

[25] C. Rao. Maximum likelihood estimation for the multinomial distribution. *Indian J. Stat.*, 18:139–148, 1957.

[26] G. Schollmeyer, T. Augustin. Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *Int. J. Approx. Reasoning*, 56:224–248, 2015. [based on an ISIPTA '13 paper]

[27] J. Stoye. Partial identification and robust treatment choice: An application to young offenders. *J. Statistical Theory and Practice*, 3:239–254, 2009.

[28] E. Tamer. Partial identification in econometrics. *Annu. Rev. Econ.*, 2:167–195, 2010.

[29] M. Trappmann, S. Gundert, C. Wenzig, D. Gebhardt. PASS: a household panel survey for research on unemployment and poverty. *Schmollers Jahrbuch*, 130:609–623, 2010.

[30] M. Troffaes, F. Coolen. Applying the imprecise Dirichlet model in cases with partial observations and dependencies in failure data. *Int. J. Approx. Reasoning*, 50:257–268, 2009.

[31] G. Tutz. *Regression for Categorical Data.* Cambridge University Press, 2011.

[32] L. Utkin, T. Augustin. Decision making under imperfect measurement using the imprecise Dirichlet model. *Int. J. Approx. Reasoning*, 44: 322–338, 2007. [based on an ISIPTA '05 paper]

[33] L. Utkin, F. Coolen. Interval-valued regression and classification models in the framework of machine learning. In: F. Coolen, G. de Cooman, T. Fetz, M. Oberguggenberger (eds.), *ISIPTA '11*, pp. 371–380, 2011.

[34] S. Vansteelandt, E. Goetghebeur, M. Kenward, G. Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Stat. Sin.*, 16:953–979, 2006.

[35] P. Walley. Inferences from multinomial data: Learning about a bag of marbles (with discussion). *J. R. Stat. Soc. B*, 58:3–57, 1996.

[36] M. Zaffalon, E. Miranda. Conservative inference rule for uncertain reasoning under incompleteness. *J. Artif. Intell. Res.*, 34:757–821, 2009.

[37] Z. Zhang. Profile likelihood and incomplete data. *Int. Stat. Rev.*, 78:102–116, 2010.