Statistical Modelling in Surveys without Neglecting *The Undecided*:

Multinomial Logistic Regression Models and Imprecise Classification Trees under Ontic Data Imprecision

Julia Plass*, Paul Fink*, Nobert Schöning** and Thomas Augustin*

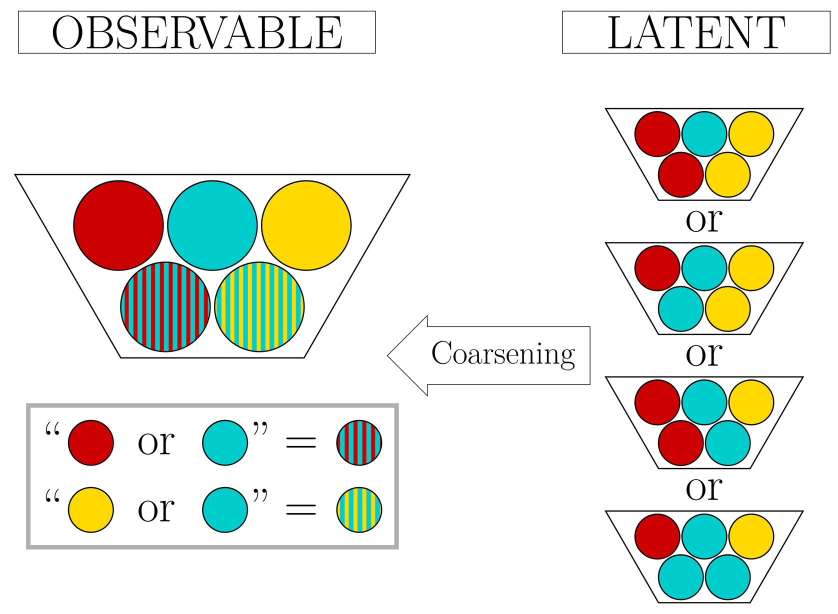
*Department of Statistics and **Geschwister Scholl Institute of Political Science, Ludwig-Maximilians-Universität Munich, Germany

Distinction between Epistemic and Ontic Interpretation [1]

Epistemic data imprecision:

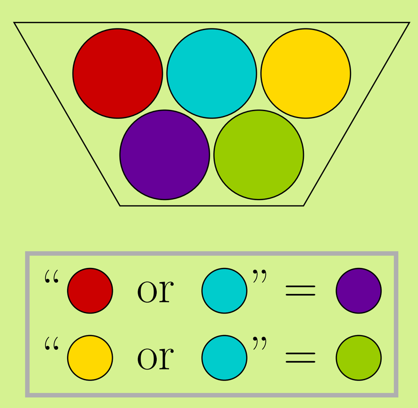
- Imprecise observation of something precise
- Actually precise values may only be observed in a coarse form, due to an underlying coarsening mechanism

Epistemic poster in session on Tuesday

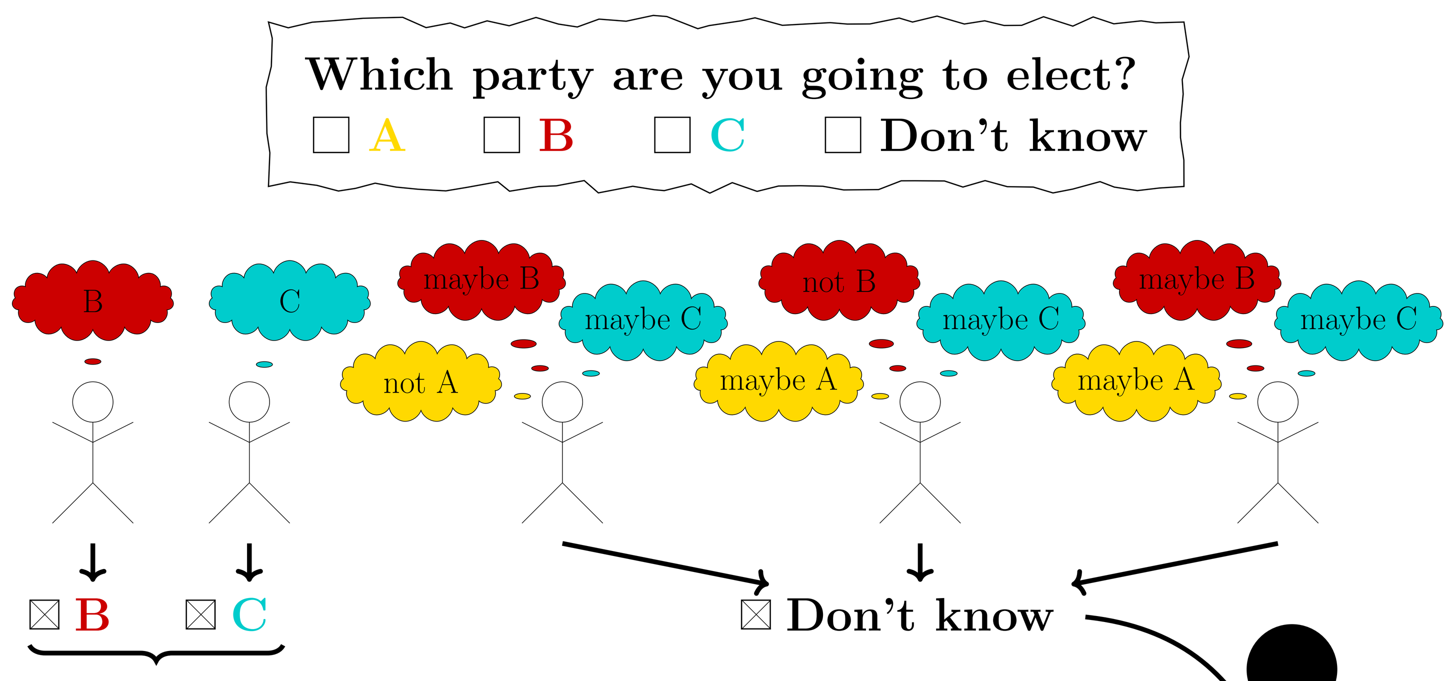


Ontic data imprecision:

- Precise observation of something imprecise
- Truth is represented by coarse observations
- Example: Answers of indecisive respondents (no unique preference)



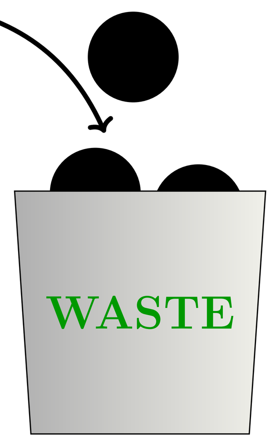
Classical Analyses ⇒ Neglect the Undecided



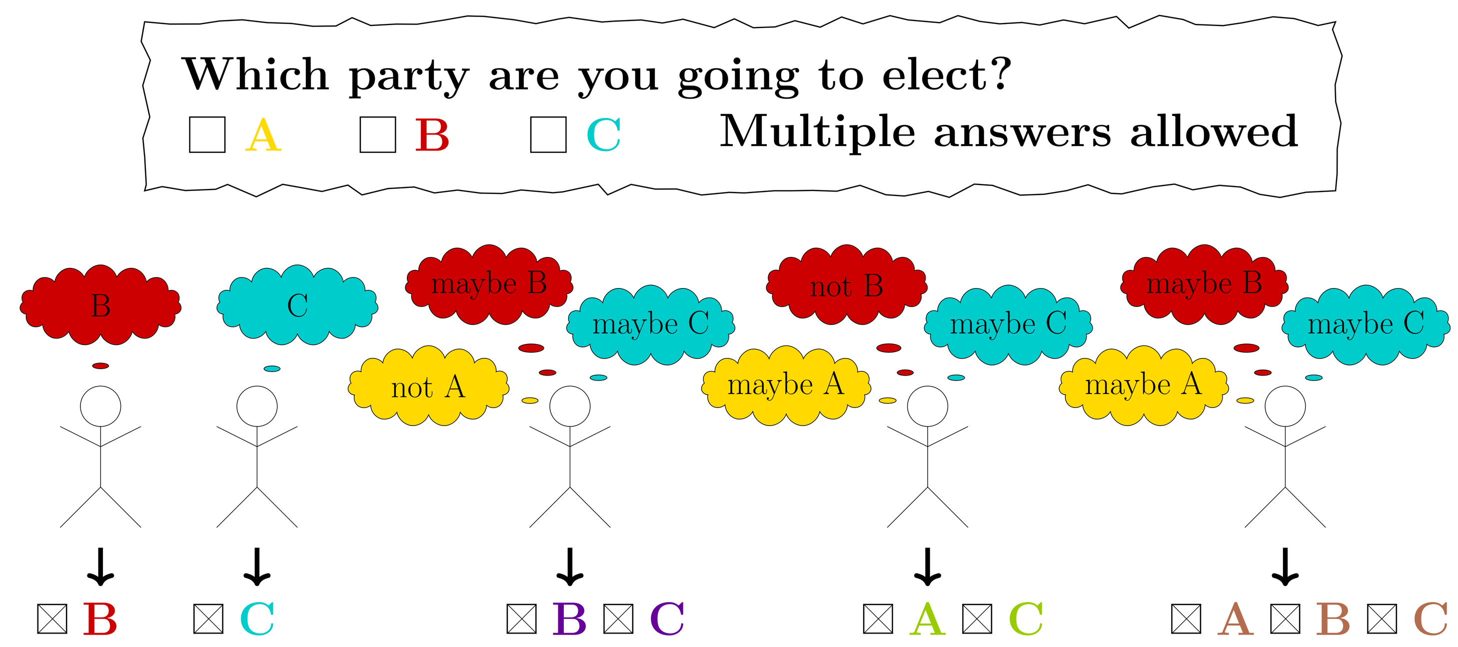
In most analyses:

Include only decisive respondents

⇒ **loss of information, biased results!**



Do not neglect the Undecided!



In this analysis:

Multiple answers as "AB" form own categories

⇒ **Opinions are reflected in the most informative way**

Construction of an ontic variable reflecting indecisiveness for GLES 2013 (German Longitudinal Election Study)

i	*certainty*	*vote*	*assessment*					*ontic*
			CD	*S*	*F*	*L*	*G*	
13	very certain	G	−1	+2	−4	−4	+4	G
126	fairly certain	S	−1	+1	0	+1	−3	L:S
1515	neither/nor	CD	+3	+3	0	−5	+2	CD:G:S

General Analysis

- Interpretation of coarse answers as ontic sets [1], i.e. as a mapping $Z^* : \Omega \to \mathcal{P}(S)$ such that for any $A \subseteq S$ holds: $Z^{*-1}(\{A\}) = \{\omega \in \Omega : Z^*(\omega) = A\} \in \mathcal{A}$
- Regard coarse answers like "A or B" as own categories ⇒ Extension of state space $S = \{1, \ldots, c\}$ to $S^* = \mathcal{P}(S) \setminus \emptyset$
  - ⇒ Basing (precise) analyses on the power set: $Y_i^* \subset \{1, \ldots, c\}$

Examples

Multinomial Regression [2]

For each $Y_i^* \subset \{1, \ldots, c\}$ probabilities $\pi_{i1}^*, \ldots, \pi_{im}^*$ ($m = |S^*|$) are modelled individually for each category $s \in \{1, \ldots, m-1\}$ by:

$$P^*(Y_i^* = s \,|\, \mathbf{x}_i) = \pi_{is}^* = \frac{\exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s^*)}{1 + \Sigma_{r=1}^{m-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r^*)}$$

and for reference category m by

$$P^*(Y_i^* = m \,|\, \mathbf{x}_i) = \pi_{im}^* = \left(1 + \Sigma_{r=1}^{m-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r^*)\right)^{-1}.$$

⇒ One obtains own regression coefficients for each coarse category, which exactly reflects the underlying idea that different types of indecisive respondents are regarded as own group.

Illustration by the GLES data:

- Dependent variable Y: first vote (reference category S)
- Covariates: religious denomination, most important information source
- Comparing the ontic to the classical approach, remarkable differences partly associated with a change in sign are obtained

Coefficient	ontic		classical
	CD	G:S	CD
intercept	0.33	−1.41 **	−0.12
rel.christ	0.37 **	−0.25	0.52 ***
info.tv	−0.02	−0.32	0.25
info.np	−0.12	−1.69 **	0.13

Classification Trees [3]

For classification trees requiring class probabilities in nodes, estimate them over the extended state space S^*.

Estimation of conditional class probabilities for extended state space with Nonparametric Predicitve Inference:

$$P^*(Y^* = y_i^*) \in \left[\max\left(0, \frac{n_i - 1}{n}\right), \min\left(\frac{n_i + 1}{n}, 1\right)\right],$$

⇒ General idea as technique already applied in multiclass classification algorithms

Illustration by the GLES data:

- Class variable Y: second vote
- Feature variables scenario 1: religious denomination, most important information source
- Feature variables scenario 2: additionally to scenario 1, stratum, sex, party identification, interest in politics, economic situation
- NPI based imprecise classification trees
- Correct classification rate and its standard deviation estimated by 10-fold cross-validation

Scen.	ontic		classical	
	mean	sd	mean	sd
1	0.407	0.040	0.446	0.041
2	0.704	0.026	0.817	0.042

Conclusion and Outlook

- Incorporate different types of "The Undecided" into statistical analyses
- Only the state space changes, the statistical methods remain the same
- In the data example including indecisive respondents does make a difference even as we were forced to assess indecisiveness indirectly
- Adaption of this idea to coarse response variables of ordinal scale
- Application of the power-set based idea for coarse categorical covariates

References

[1] cp. Couso, Dubois & Sánchez, 2014, Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables

[2] e.g. Tutz, 2011, Regression for Categorical Data

[3] e.g. Abellán & Moral, 2003, INT J INTELL SYST

[4] Rattinger, Roßteutscher, Schmitt-Beck, Weßels & Wolf, 2013, GLES 2013