# Coarse categorical data under epistemic and ontologic uncertainty

Julia Plaß

05th of July 2014

# Outline

# Epistemic vs. ontologic uncertainty

- Ontologic uncertainty

Coarse variable of interest

- coarse nature induced by indecision
- truth is represented by coarse variable

Precise observation of sth. coarse

# Epistemic vs. ontologic uncertainty

## Ontologic uncertainty

Coarse variable of interest

- coarse nature induced by indecision
- truth is represented by coarse variable

Precise observation of sth. coarse

## Epistemic uncertainty

Precise variable of interest

Coarsening mechanism

Coarse observation of sth. precise

# Why should data under ontologic uncertainty be collected?

e.g. GLES, 2005 (F13, Item: second vote):

"Assuming you voted at all, which party would you give your second vote to?"

☐ CDU/CSU ☐ SPD ... ☐ other party ☐ refusing to vote

⇒ Indecisive respondents are forced to an answer

⇒ In many cases a category "Don't know" is provided for indecisive respondents

⇓

Information loss

# Basic idea - The $^\star$-notation (random sets)

**General analysis**

Analysis on the power set

$\Rightarrow \Omega^\star = \mathcal{P}(\Omega) \setminus \emptyset$

$$\begin{array}{rcl} P^\star : \mathcal{P}(\Omega^\star) = \mathcal{P}(\mathcal{P}(\Omega)) & \rightarrow & [0,1] \\ E^\star & \rightarrow & P^\star(E^\star). \end{array}$$

**Example:**

$\Omega = \{A,\ B,\ C\}$

$$\begin{array}{rcl} \Omega^\star & = & \{\ \{A\},\ \{B\},\ \{C\},\ \{A,\ B\}, \\ & & \{A,\ C\}, \{B,\ C\}, \{A,\ B,\ C\}\ \} \end{array}$$

$E^\star$: "Being indecisive between at least two parties" $\Rightarrow P^\star(E^\star) = \frac{|E^\star|}{|\Omega^\star|} = \frac{4}{7}$

**Prediction**

Consider $F^\star : \Omega^\star \rightarrow [0,1]$

$$\Rightarrow F^\star(Q) = [\underline{F^\star}(Q), \overline{F^\star}(Q)]$$

$$\text{where} \quad \begin{array}{rcl} \underline{F^\star}(Q) & = & Bel(Q) \quad \text{and} \\ \overline{F^\star}(Q) & = & Pl(Q) \end{array}$$

**Example:**

observations:

$\{A\},\ \{C\},\ \{A,B\},\ \{A,B,C\},\ \{B\}$

$\Rightarrow F^\star(B) = [\frac{1}{5},\ \frac{3}{5}]$

# Model under ontologic uncertainty

## Data under ontologic uncertainty:

- $Y_i$: categorical random variable of nominal scale of measurement (precise and coarse categories)
- $\Omega^\star = \mathcal{P}(\Omega) \setminus \emptyset$: sample space
- $m = |\Omega^\star|$: number of categories of $Y_i$

## Model under ontologic uncertainty:

The probability of occurence for category $r = 1,\ 2,\ 3, ...,\ m-1$ can be calculated by

$$P(Y_i = r|\boldsymbol{x_i}) = \frac{\exp(\boldsymbol{x_i^T}\boldsymbol{\beta}_r)}{1 + \sum_{s=1}^{m-1} \exp(\boldsymbol{x_i^T}\boldsymbol{\beta}_s)}$$

and for category $m$ by

$$P(Y_i = m|\boldsymbol{x_i}) = \frac{1}{1 + \sum_{s=1}^{m-1} \exp(\boldsymbol{x_i^T}\boldsymbol{\beta}_s)}$$

# Epistemic vs. ontologic uncertainty

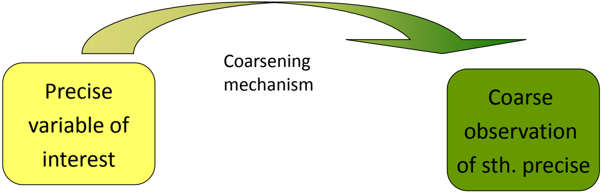
- Ontologic uncertainty

Coarse variable of interest

- coarse nature induced by indecision
- truth is represented by coarse variable

Precise observation of sth. coarse

- Epistemic uncertainty



# When do data under epistemic uncertainty occur?

## Reasons for coarse categorical data:

- Guarantee of anonymization, prevention of refusals

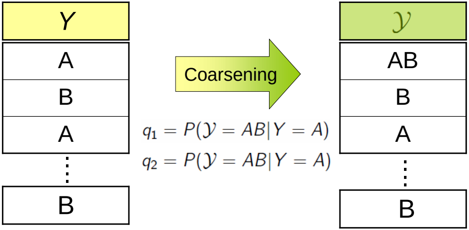
  **Example:**
  "Which kind of party did you elect?"
  ☐ rather left ☐ center ☐ rather right

- Different levels of reporting accuracy
  (lack of knowledge, vague question formulation)

  **Examples:**
  "To which electoral district do you belong to?"
  "Which car do you drive?"

# The general log-likelihood

**Addressed data situation :**




**log-Likelihood under the iid assumption :**

$$l(\pi_A, q_1, q_2) = \ln \Bigg( \prod_{i:\mathcal{Y}_i = A} \underbrace{P(\mathcal{Y} = A|Y = A)}_{(1-q_1)} \pi_{iA} \prod_{i:\mathcal{Y}_i = B} \underbrace{P(\mathcal{Y} = B|Y = B)}_{(1-q_2)} (1 - \pi_{iA})$$

$$\prod_{i:\mathcal{Y}_i = AB} \underbrace{P(\mathcal{Y} = AB|Y = A)}_{q_1} \pi_{iA} + \underbrace{P(\mathcal{Y} = AB|Y = B)}_{q_2} (1 - \pi_{iA}) \Bigg)$$

$$\overset{iid}{=} n_A \cdot [\ln(1 - q_1) + \ln(\pi_A)] + n_B \cdot [ln(1 - q_2) + ln(1 - \pi_A)]$$

$$n_{AB} \cdot [q_1 \pi_A + q_2 (1 - \pi_A))]$$

# The general log-likelihood

## FOC :

$$
\begin{aligned}
\text{I.)}\ \frac{\partial}{\partial \pi_A} &= \frac{n_{AB}}{q_1\pi_A + q_2(1-\pi_A)}(q_1 - q_2) + \frac{n_A}{\pi_A} - \frac{n_B}{1-\pi_A} \overset{!}{=} 0 \\
\text{II.)}\ \frac{\partial}{\partial q_1} &= \frac{n_{AB}}{q_1\pi_A + q_2(1-\pi_A)}\pi_A - \frac{n_A}{1-q_1} \overset{!}{=} 0 \\
\text{III.)}\ \frac{\partial}{\partial q_2} &= \frac{n_{AB}}{q_1\pi_A + q_2(1-\pi_A)}(1-\pi_A) - \frac{n_B}{1-q_2} \overset{!}{=} 0
\end{aligned}
$$

**Neccessary and sufficient solutions:**

Estimators $(\hat{\pi}_A, \hat{q}_1, \hat{q}_2)$ are solutions of the estimation problem if and only if

$$\frac{n_{AB}}{n} = \hat{q}_1 \cdot \hat{\pi}_A + \hat{q}_2 \cdot (1 - \hat{\pi}_A)$$

is fulfilled.

$\Rightarrow$ Contentual additional restriction: $\hat{\pi}_A$, $\hat{q}_1$ and $\hat{q}_1 > 0$ and $< 1$.

# Distinguishing different cases

## Estimation of parameter of interest ....

... implying point-identifying assumptions

- known coarsening mechanism
- $q_1 = q_2$: data are *coarsened at random* (CAR)

$$\hat{\pi}_A = \frac{n_A}{n_A + n_B}$$
$$\hat{q}_1 = \hat{q}_2 = \frac{n_{AB}}{n_A + n_B + n_{AB}}$$

- relation between coarsening parameters $R = \frac{q_1}{q_2}$ is known
  $\Rightarrow$ Generalization of CAR

... without any assumptions

$\Rightarrow$ Find lower and upper bounds of parameter estimators

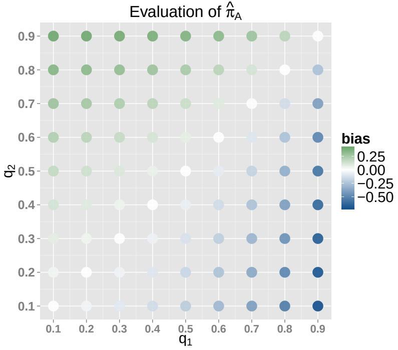
# Implying assumptions - some results

## Analysis by inclusion of CAR

Median relative bias of $\hat{\pi}_A$

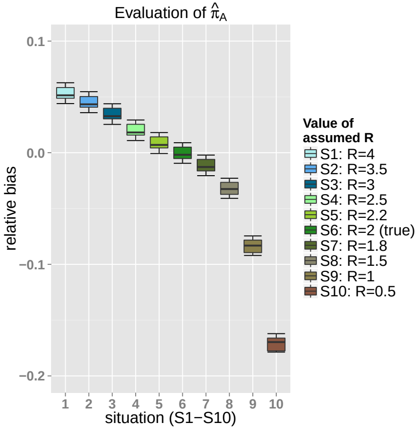
for different combinations of

true $q_1$ and $q_2$ values:




## Imposing an assumption about R

$$\left.\begin{array}{l} q_1 = 0.3 \\ q_2 = 0.15 \end{array}\right\} \quad R_{\text{true}} = \frac{q_1}{q_2} = 2$$



# Summary

- Important to distinguish between epistemic and ontologic uncertainty
- One can deal with ontologic uncertainty by redefining the sample space
- In case of iid variables under epistemic uncertainty
  - ... generally a set of estimators results characterized by a special condition
  - ... using correctly the assumptions of *CAR* leads to identified and nearly unbiased estimators