# Coarse categorical data under ontologic and epistemic uncertainty

Julia Plaß

Ludwigs-Maximilians-University (LMU)

08th of September 2014

# Outline

# Epistemic vs. ontic/ontologic uncertainty (I. Couso, D. Dubois, 2014)

▶ Ontologic uncertainty

Coarse variable of interest

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

Precise observation of sth. coarse

# Epistemic vs. ontic/ontologic uncertainty (I. Couso, D. Dubois, 2014)

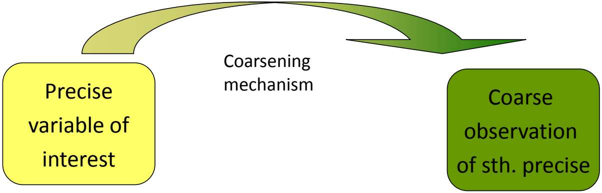
## ▶ Ontologic uncertainty

Coarse variable of interest

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

Precise observation of sth. coarse

## ▶ Epistemic uncertainty



# Why should data under ontologic uncertainty be collected?

**Example:**

***Which party will you give your vote?***

☐ *A* ☐ *B* ☐ *C* ☐ *Don‘t know*

# Why should data under ontologic uncertainty be collected?

**Example:**

***Which party will you give your vote?***

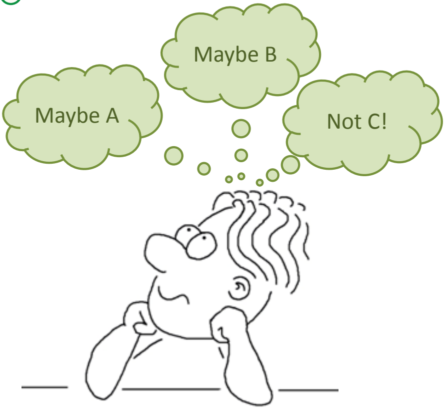
☐ *A* ☐ *B* ☐ *C* ☐ *Don‘t know*

# Why should data under ontologic uncertainty be collected?

**Example:**

*Which party will you give your vote?*

☐ *A* ☐ *B* ☐ *C* ☐ *Don't know*

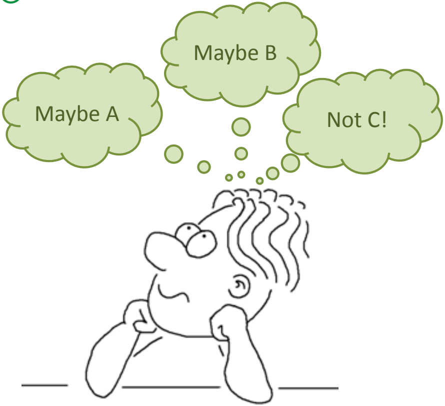


# Why should data under ontologic uncertainty be collected?

**Example:**

*Which party will you give your vote?*

☐ *A* ☐ *B* ☐ *C* ☒ *Don't know*

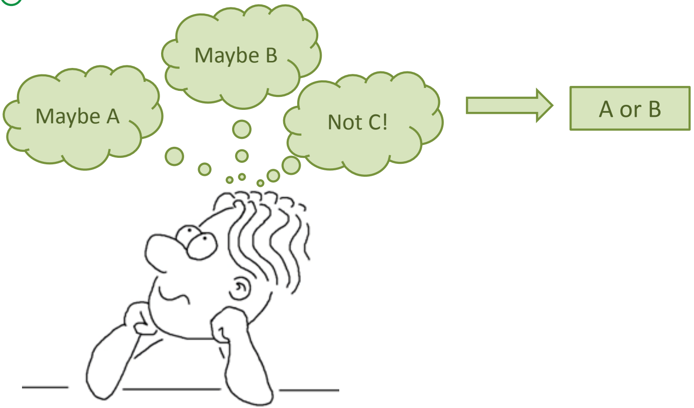


# Why should data under ontologic uncertainty be collected?

**Example:**

*Which party will you give your vote?*

☐ *A* ☐ *B* ☐ *C* ☒ *Don't know*

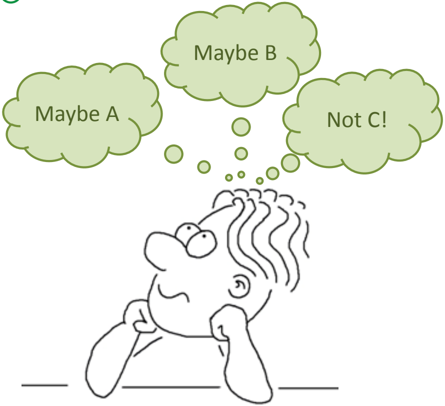


# Why should data under ontologic uncertainty be collected?

**Example:**

*Which party will you give your vote?*

☒ *A* ☒ *B* ☐ *C* ☐ *Don‘t know*




⇒ A or B

Dealing with ontologic uncertainty:

→ Allow multiple anwers

# Basic idea (illustrated by GLES 2013)

| certainty | vote | assesCD | assesSPD | assesGREEN | assesLEFT | ... |
|---|---|---|---|---|---|---|
| very certain | SPD | -1 | 5 | 2 | 1 | ... |
| certain | CD | 4 | 3 | 3 | 1 | ... |
| not that certain | GREEN | 3 | 4 | 4 | -1 | ... |
| not certain at all | CD | -3 | 2 | 2 | 2 | ... |

Exemplary extraction of the dataset

# Basic idea (illustrated by GLES 2013)

| certainty | vote | assesCD | assesSPD | assesGREEN | assesLEFT | ... |
|---|---|---|---|---|---|---|
| very certain | SPD | -1 | 5 | 2 | 1 | ... |
| certain | CD | 4 | 3 | 3 | 1 | ... |
| not that certain | GREEN | 3 | 4 | 4 | -1 | ... |
| not certain at all | CD | -3 | 2 | 2 | 2 | ... |

Exemplary extraction of the dataset

**Analysis:** *Prediction*

- Classical analysis - Vote of respondents who are certain:
  Prediction for "CD": $\frac{\#\text{respondents who will vote for "CD" with certainty}}{\#\text{respondents who are certain}} = 0.46$

# Basic idea (illustrated by GLES 2013)

| certainty | vote | assesCD | assesSPD | assesGREEN | assesLEFT | ... |
|---|---|---|---|---|---|---|
| very certain | SPD | -1 | 5 | 2 | 1 | ... |
| certain | CD | 4 | 3 | 3 | 1 | ... |
| not that certain | GREEN | 3 | 4 | 4 | -1 | ... |
| not certain at all | CD | -3 | 2 | 2 | 2 | ... |

Exemplary extraction of the dataset

**Analysis:** *Prediction*

- Classical analysis - Vote of respondents who are certain:
  Prediction for "CD": $\frac{\#\text{respondents who will vote for "CD" with certainty}}{\#\text{respondents who are certain}} = 0.46$
- Dealing with ontologic uncertainty:

| | | | | |
|---|---|---|---|---|
| **CD** 519 | **SPD** 287 | **GREEN** 105 | **LEFT** 101 | **OTHER** 62 |
| **SPD-CD** 34 | **GREEN-SPD** 34 | **CD-OTHER** 24 | **LEFT-SPD** 14 | **GREEN-LEFT** 13 |
| **GREEN-SPD-CD** 13 | **SPD-CD-OTHER** 13 | **LEFT-GREEN-SPD** 13 | **SPD-OTHER** 12 | **rare comb.** 77 |

$$\hat{Bel}(CD) = \frac{519}{1321} = 0.39$$

$$\hat{Pl}(CD) = \frac{519+34+24+13+13}{1321} = 0.45$$

$\Rightarrow$ Prediction for "CD": [0.39, 0.45]

# Basic idea (illustrated by GLES 2013)

| certainty | vote | assesCD | assesSPD | assesGREEN | assesLEFT | ... |
|---|---|---|---|---|---|---|
| very certain | SPD | -1 | 5 | 2 | 1 | ... |
| certain | CD | 4 | 3 | 3 | 1 | ... |
| not that certain | GREEN | 3 | 4 | 4 | -1 | ... |
| not certain at all | CD | -3 | 2 | 2 | 2 | ... |

Exemplary extraction of the dataset

**Analysis:** *Regression*

- For reasons of explanation interpretation of selected $\beta$ estimators only
- Reference category: "SPD"

- Classical analysis - Vote of respondents who are certain:

| | Intercept | sexFEM | InfoNEWSPAPER | InfoRADIO | InfoWEB |
|---|---|---|---|---|---|
| **CD** | 0.2914 | 0.2456 | -0.0037 | -0.1487 | -0.6774 |

# Model under ontologic uncertainty

## Data under ontologic uncertainty:

- $Y_i$: categorical random variable of nominal scale of measurement with $Y_i \subseteq \underbrace{\{a, b, \dots\}}_{\text{precise categories}}$
- $m = |\mathcal{P}(\Omega) \setminus \emptyset)|$: number of categories of $Y_i$

## Model under ontologic uncertainty:

⇒ classical multinomial logit model with different number of categories:

The probability of occurence for category $r = 1,\ 2,\ 3, \dots,\ m-1$ 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

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

# Basic idea (illustrated by GLES 2013)

| certainty | vote | assesCD | assesSPD | assesGREEN | assesLEFT | ... |
|---|---|---|---|---|---|---|
| very certain | SPD | -1 | 5 | 2 | 1 | ... |
| certain | CD | 4 | 3 | 3 | 1 | ... |
| not that certain | GREEN | 3 | 4 | 4 | -1 | ... |
| not certain at all | CD | -3 | 2 | 2 | 2 | ... |

Exemplary extraction of the dataset

**Analysis:** *Regression*

- For reasons of explanation interpretation of selected $\beta$ estimators only
- Reference category: "SPD"

- Classical analysis - Vote of respondents who are certain:

| | Intercept | sexFEM | InfoNEWSPAPER | InfoRADIO | InfoWEB |
|---|---|---|---|---|---|
| **CD** | 0.2914 | 0.2456 | -0.0037 | -0.1487 | -0.6774 |

- Dealing with ontologic uncertainty:

| | Intercept | sexFEM | InfoNEWSPAPER | InfoRADIO | InfoWEB |
|---|---|---|---|---|---|
| **CD** | 0.4706 | 0.3135 | -0.0784 | 0.1856 | -0.5025 |
| **SPD-CD** | -2.2007 | 0.4034 | -1.2556 | 0.9476 | 0.1886 |
| **GREEN-SPD** | -2.4432 | 0.9393 | -1.3053 | 0.2300 | 0.1844 |

# Basic idea (illustrated by GLES 2013)

| certainty | vote | assesCD | assesSPD | assesGREEN | assesLEFT | ... |
|---|---|---|---|---|---|---|
| very certain | SPD | -1 | 5 | 2 | 1 | ... |
| certain | CD | 4 | 3 | 3 | 1 | ... |
| not that certain | GREEN | 3 | 4 | 4 | -1 | ... |
| not certain at all | CD | -3 | 2 | 2 | 2 | ... |

Exemplary extraction of the dataset

**Analysis:** *Regression*

- For reasons of explanation interpretation of selected $\beta$ estimators only
- Reference category: "SPD"

- Classical analysis - Vote of respondents who are certain:

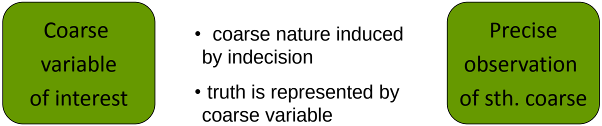
| | Intercept | sexFEM | InfoNEWSPAPER | InfoRADIO | InfoWEB |
|---|---|---|---|---|---|
| **CD** | 0.2914 | 0.2456 | -0.0037 | -0.1487 | -0.6774 |

- Dealing with ontologic uncertainty:

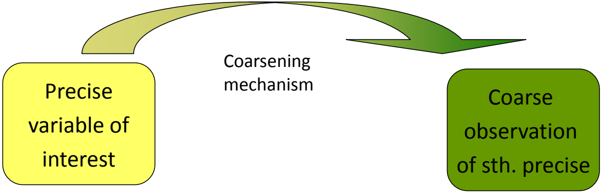
| | Intercept | sexFEM | InfoNEWSPAPER | InfoRADIO | InfoWEB |
|---|---|---|---|---|---|
| **CD** | 0.4706 | 0.3135 | -0.0784 | 0.1856 | -0.5025 |
| **SPD-CD** | -2.2007 | 0.4034 | -1.2556 | 0.9476 | 0.1886 |
| **GREEN-SPD** | -2.4432 | 0.9393 | -1.3053 | 0.2300 | 0.1844 |

# Epistemic vs. ontologic uncertainty

- Ontologic uncertainty




- 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

# When do data under epistemic uncertainty occur?

## Reasons for coarse categorical data:

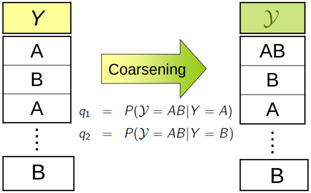
- Guarantee of anonymization, prevention of refusals

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

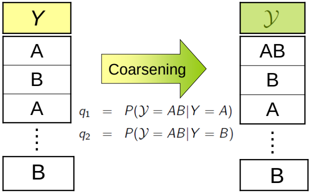
- Different levels of reporting accuracy
(lack of knowledge, vague question formulation)

**Examples:**
"Which car do you drive?"

# Addressed data situations



# Addressed data situations




Two different situations will be regarded:

## Case 1 - No covariates:

- IID-assumption

$$\pi_{iA} = \pi_A$$

- Constant coarsening mechanisms $q_1$ and $q_2$

# Addressed data situations

| $X$ | $Y$ | | $\mathcal{Y}$ |
|---|---|---|---|
| 1 | A | Coarsening → | AB |
| 1 | B | | B |
| 0 | A | | A |
| ⋮ | ⋮ | | ⋮ |
| 1 | B | | B |

$$q_1 = P(\mathcal{Y} = AB \mid Y = A)$$
$$q_2 = P(\mathcal{Y} = AB \mid Y = B)$$

Two different situations will be regarded:

## Case 1 - No covariates:

- IID-assumption

$$\pi_{iA} = \pi_A$$

- Constant coarsening mechanisms $q_1$ and $q_2$

## Case 2 - Binary covariates, no intercept

- $$\pi_{iA|X_i=1} = \frac{\exp(\beta_A)}{1+\exp(\beta_A)}$$
  $$\pi_{iA|X_i=0} = \frac{1}{2}$$

- Constant coarsening mechanisms $q_1$ and $q_2$

# Case 1 - No covariates + IID-assumption

## log-Likelihood under the iid assumption :

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

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

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

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

## FOC:

I.) $\frac{\partial}{\partial \pi_A} = \frac{n_{AB}}{q_1\pi_A + q_2(1-\pi_A)}(q_1 - q_2) + \frac{n_A}{\pi_A} - \frac{n_B}{1-\pi_A} \overset{!}{=} 0$

II.) $\frac{\partial}{\partial q_1} = \frac{n_{AB}}{q_1\pi_A + q_2(1-\pi_A)}\pi_A - \frac{n_A}{1-q_1} \overset{!}{=} 0$

III.) $\frac{\partial}{\partial q_2} = \frac{n_{AB}}{q_1\pi_A + q_2(1-\pi_A)}(1-\pi_A) - \frac{n_B}{1-q_2} \overset{!}{=} 0$

# Case 1 - No covariates + IID-assumption

## log-Likelihood under the iid assumption :

$$l(\pi_A, q_1, q_2) = \ln\Bigg( \prod_{i:\mathcal{Y}_i=A} \underbrace{P(\mathcal{Y}=A|Y=A)}_{(1-q_1)} \pi_{iA} \prod_{i:\mathcal{Y}_i=B} \underbrace{P(\mathcal{Y}=B|Y=B)}_{(1-q_2)} (1-\pi_{iA})$$
$$\prod_{i:\mathcal{Y}_i=AB} \underbrace{P(\mathcal{Y}=AB|Y=A)}_{q_1} \pi_{iA} + \underbrace{P(\mathcal{Y}=AB|Y=B)}_{q_2} (1-\pi_{iA}) \Bigg)$$
$$\stackrel{iid}{=} n_A \cdot [\ln(1-q_1) + \ln(\pi_A)] + n_B \cdot [ln(1-q_2) + ln(1-\pi_A)]$$
$$n_{AB} \cdot [q_1 \pi_A + q_2(1-\pi_A))]$$

## FOC:

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

$\Longrightarrow$

**Neccessary and sufficient condition for estimators $(\hat{\pi}_A, \hat{q}_1, \hat{q}_2)$**

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

# Case 2 - Binary covariates, no intercept

## Involved assumptions:

- *No intercept* $\beta_0$

  $\Rightarrow \pi_{iA|X_i=1} = \frac{\exp(\beta_A)}{1+\exp(\beta_A)}$ and $\pi_{iA|X_i=0} = \frac{1}{2}$

# Case 2 - Binary covariates, no intercept

## Involved assumptions:

- *No intercept* $\beta_0$

  $\Rightarrow \pi_{iA|X_i=1} = \frac{\exp(\beta_A)}{1+\exp(\beta_A)}$ and $\pi_{iA|X_i=0} = \frac{1}{2}$

- *Constant coarsening mechanisms* $q_1$ *and* $q_2$*:*
  **Example:** "Do you regularly steal candy ($\bowtie$) out of your mother's candy box?"
  Asked: girls ($g$) and boys ($b$)

| X | Y | $\mathcal{Y}$ |
|---|---|---|
| $g$ | $\bowtie$ | $\overline{\bowtie}$ or $\overline{\bowtie}$ |
| $g$ | $\overline{\bowtie}$ | $\overline{\bowtie}$ |
| $g$ | $\bowtie$ | $\bowtie$ |
| $b$ | $\bowtie$ | $\overline{\bowtie}$ or $\overline{\bowtie}$ |
| $b$ | $\bowtie$ | $\bowtie$ |
| $b$ | $\overline{\bowtie}$ | $\overline{\bowtie}$ |
| $b$ | $\overline{\bowtie}$ | $\overline{\bowtie}$ |

$$P(\mathcal{Y} = \overline{\bowtie} \text{ or } \overline{\bowtie} | Y = \bowtie, X = g) = P(\mathcal{Y} = \overline{\bowtie} \text{ or } \overline{\bowtie} | Y = \bowtie, X = b)$$
$$P(\mathcal{Y} = \overline{\bowtie} \text{ or } \overline{\bowtie} | Y = \overline{\bowtie}, X = g) = P(\mathcal{Y} = \overline{\bowtie} \text{ or } \overline{\bowtie} | Y = \overline{\bowtie}, X = b)$$

$$\Downarrow$$

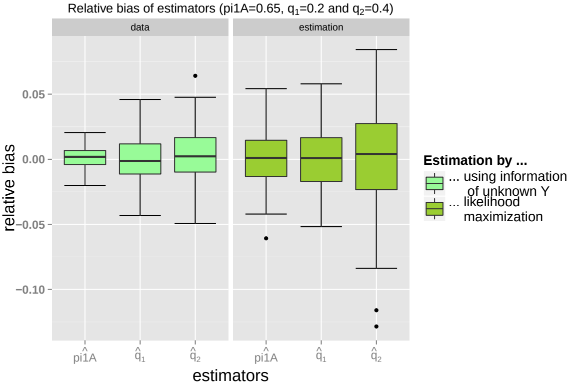
the coarsening mechanisms do not depend on subgroup $x$

# Case 2 - Binary covariates, no intercept

## log-Likelihood involving binary covariate $X$:

$$
\begin{aligned}
l(\pi_{0A}, \pi_{1A}, q_1, q_2) &= n_{1A} \cdot [\ln(1-q_1) + \ln(\pi_{1A})] + n_{0A} \cdot [\ln(1-q_1) + \ln(\pi_{0A})] \\
&+ n_{1B} \cdot [\ln(1-q_2) + ln(1-\pi_{1A})] + n_{0B} \cdot [\ln(1-q_2) + ln(1-\pi_{0A})] \\
&+ n_{1AB} \cdot [q_1\pi_{1A} + q_2(1-\pi_{1A}))] + n_{0AB} \cdot [q_1\pi_{0A} + q_2(1-\pi_{0A}))]
\end{aligned}
$$

with $\pi_{01} = \frac{1}{2}$



# Case 2 - Binary covariates, no intercept

**Questions / Discussion suggestions:**

- Is this result reasonable? (Remember: The coarsening mechanism does not depend on the values of $X$)

# Case 2 - Binary covariates, no intercept

## Questions / Discussion suggestions:

- Is this result reasonable? (Remember: The coarsening mechanism does not depend on the values of $X$)
- Is it possible to derive those estimators by means of equations like in the *iid* case, but now for each subgroup of $X$?

$$
\begin{aligned}
\frac{n_{1AB}}{n_1} &= \hat{q}_1 \cdot \hat{\pi}_{1A} + \hat{q}_2 \cdot (1 - \hat{\pi}_{1A}) \\
\frac{n_{0AB}}{n_0} &= \hat{q}_1 \cdot \hat{\pi}_{0A} + \hat{q}_2 \cdot (1 - \hat{\pi}_{0A}) \\
\frac{n_{1A}}{n_1} &= (1 - \hat{q}_1) \cdot \hat{\pi}_{1A} \\
\frac{n_{0A}}{n_0} &= (1 - \hat{q}_1) \cdot \hat{\pi}_{0A}
\end{aligned}
$$

# Case 2 - Binary covariates, no intercept

## Questions / Discussion suggestions:

- Is this result reasonable? (Remember: The coarsening mechanism does not depend on the values of $X$)
- Is it possible to derive those estimators by means of equations like in the *iid* case, but now for each subgroup of $X$?

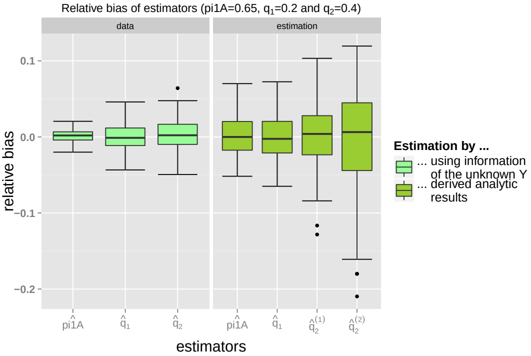
$$\begin{aligned} \frac{n_{1AB}}{n_1} &= \hat{q}_1 \cdot \hat{\pi}_{1A} + \hat{q}_2 \cdot (1 - \hat{\pi}_{1A}) \\ \frac{n_{0AB}}{n_0} &= \hat{q}_1 \cdot \hat{\pi}_{0A} + \hat{q}_2 \cdot (1 - \hat{\pi}_{0A}) \\ \frac{n_{1A}}{n_1} &= (1 - \hat{q}_1) \cdot \hat{\pi}_{1A} \\ \frac{n_{0A}}{n_0} &= (1 - \hat{q}_1) \cdot \hat{\pi}_{0A} \end{aligned} \quad \Longrightarrow$$

**Resulting estimators**

$$\begin{aligned} \hat{\pi}_{1A} &= \frac{n_{1A} \cdot n_0}{2 \cdot n_1 \cdot n_{0A}} \\ \hat{q}_1 &= 1 - 2 \cdot \frac{n_{0A}}{n_0} \end{aligned}$$

and $\hat{q}_2^{(1)}$ and $\hat{q}_2^{(2)}$

# Case 2 - Binary covariates, no intercept




$$\hat{q}_2^{(1)} = \frac{2 \cdot n_{0AB} - n_0 + 2 \cdot n_{0A}}{n_0}$$

$$\hat{q}_2^{(2)} = \frac{\frac{n_{1AB}}{n_1} - \frac{n_{1A}(n_0 - 2n_{0A})}{2n_1 n_{0A}}}{\frac{2n_1 n_{0A} - n_{1A} n_0}{2n_1 n_{0A}}}$$

# 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, a set of estimators results characterized by a special condition
  - ... being a binary covariate available, precise real valued point estimators seem to result

# References


Couso, I. and Dubois, D.

*Statistical reasoning with set-valued information: Ontic vs. epistemic views, IJAR, 2014.*

Heitjan, D. and Rubin, D.

*Ignorability and Coarse Data, Annals of Statistics, 1991.*

Matheron, G.

*Random Sets and Integral Geometry, Wiley New York, 1975.*

Plaß, J.

*Coarse categorical data under epistemic and ontologic uncertainty: Comparison and extension of some approaches, master's thesis, 2013*