Reliable analysis of categorical data under epistemic imprecision

CFE-CMStatistics 2015

Julia Plass*, Thomas Augustin*,
Marco Cattaneo**, Georg Schollmeyer*

*Department of Statistics, Ludwigs-Maximilians University and
**Department of Mathematics, University of Hull

&

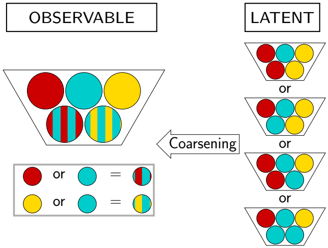
13th of December 2015

Coarse data

Data are not observed in the resolution originally intended

(epistemic vs. ontic interpretation)

- Here: coarse data "=" data under epistemic imprecision
- Imprecise observation of something precise:



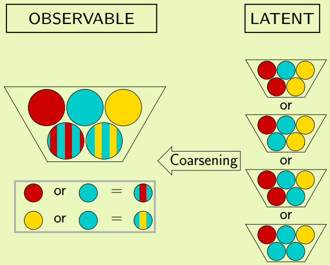
Outline

1. Where do data under epistemic imprecision typically arise?
2. How to deal with data under epistemic imprecision?
3. How to incorporate auxiliary information?
4. Are there possibilities to test on point identifying assumptions (coarsening at random, subgroup independence)?

Examples of data under epistemic imprecision

Epistemic imprecision:

"Imprecise observation of something precise"



$\Rightarrow$ Truth is hidden due to the underlying coarsening mechanism

Examples:

- Matched data sets with partially overlapping variables
- Coarsening as anonymization technique
- Missing data as special case

Here: PASS-data
$\mathcal{Y}$: income, X: UBII

$$\Omega_{\mathcal{Y}} = \{<, \geq, \text{na}\}$$
$$\Omega_X = \{0 \text{ (no)}, 1 \text{ (yes)}\}$$

Already existing approaches

- Still common to **enforce precise results**
- Variety of **set-valued approaches**
  - via random sets (e.g. Nguyen, 2006, An Introduction to Random Sets)
  - using Bayesian approaches (e.g. de Cooman, Zaffalon, 2004, Artif. Intell.)
  - via likelihood-based belief function (Denœux, 2014, IJAR)
  - via profile likelihood (Cattaneo, Wiencierz, 2012, IJAR)

Here: Likelihood-based approach
influenced by methodology of partial identification
(Manski, 2003, Partial Identification of Probability Distributions)
coarse categorical data

Basic problem

OBSERVABLE

coarse data

$\mathcal{Y}$

$$p_{\mathscr{Y}} = P(\mathcal{Y}_i = \mathscr{Y})$$

LATENT

latent variable

Y

$$\pi_{ij} = P(Y_i = j)$$

Basic problem (iid case)

OBSERVABLE | **LATENT**

coarse data

$\mathcal{Y}$

$p_{\mathscr{y}} = P(\mathcal{Y}_i = \mathscr{Y})$

Observation model $\mathcal{Q}$

error-freeness

$q_{\mathscr{y}|y} =$
$P(\mathcal{Y} = \mathscr{Y} | Y = y)$

latent variable

Y

$\pi_{ij} = P(Y_i = j)$

Main goal:

Maximum-Likelihood estimation of $\boldsymbol{\gamma} = (\boldsymbol{q}_{\mathscr{y}|\boldsymbol{y}}^T, \ \boldsymbol{\pi}_{\boldsymbol{y}}^T)^T$

Basic problem (regression case)

OBSERVABLE

LATENT

coarse data

$\mathcal{Y}$

$p_{x\mathscr{Y}} = P(\mathcal{Y} = \mathscr{Y} | X = x)$

Observation model $\mathcal{Q}$

error-freeness

$q_{\mathscr{Y}|xy} =$

$P(\mathcal{Y} = \mathscr{Y} | Y = y, X = x)$

latent variable

Y

for j=1,. . . ,K-1

$\pi_{ij} = P(Y_i = j | \mathbf{x}_i)$

$= \frac{\exp(\beta_{j0} + \mathbf{x}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{s=1}^{K-1} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)}$

for reference category K

$\pi_{iK} = \frac{1}{1 + \sum_{s=1}^{K-1} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)}$

(multinomial logit model)

Main goal:

Maximum-Likelihood estimation of $\boldsymbol{\gamma} = (\boldsymbol{q}^T_{\mathscr{Y}|\boldsymbol{xy}}, \ \boldsymbol{\pi}^T_{\boldsymbol{xy}})^T$

OBSERVABLE

LATENT

Use random-set perspective and determine ML estimator

$\hat{p}_{\mathscr{y}} = \hat{P}(\mathcal{Y} = \mathscr{y})$

$\longrightarrow$ $\hat{p}_{\mathscr{y}} = \frac{n_{\mathscr{y}}}{n}$

Use the **connection between** p and γ

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

$$\boldsymbol{\gamma} = (\boldsymbol{q}^T_{\mathscr{y}|\boldsymbol{y}},\ \boldsymbol{\pi}^T_{\boldsymbol{y}})^T$$

and the **invariance of the likelihood** under parameter transformations:

$$\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}$$

$$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n},\ \frac{\sum_{\mathfrak{y} \ni y} n_{\mathfrak{y}}}{n}\right]$$

$$\hat{q}_{\mathfrak{y}|y} \in \left[0,\ \frac{n_{\mathfrak{y}}}{n_{\{y\}} + n_{\mathfrak{y}}}\right]$$

OBSERVABLE

LATENT

Use random-set perspective and determine ML estimator $\hat{p}_{\mathscr{y}} = \hat{P}(\mathcal{Y} = \mathscr{y})$

$\longrightarrow$ $\hat{p}_{\mathscr{y}} = \frac{n_{\mathscr{y}}}{n}$

Use the **connection between** p and γ

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

$$\boldsymbol{\gamma} = (\boldsymbol{q}^T_{\mathscr{y}|\boldsymbol{y}},\ \boldsymbol{\pi}^T_{\boldsymbol{y}})^T$$

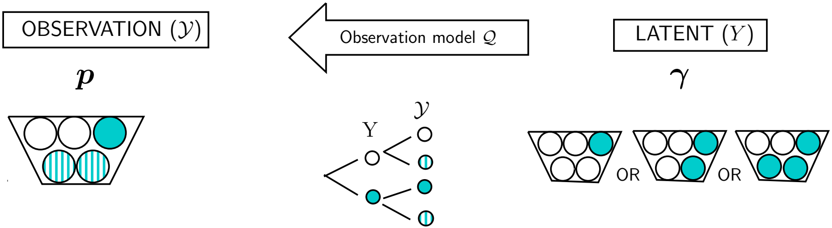
and the **invariance of the likelihood** under parameter transformations:

$$\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}$$

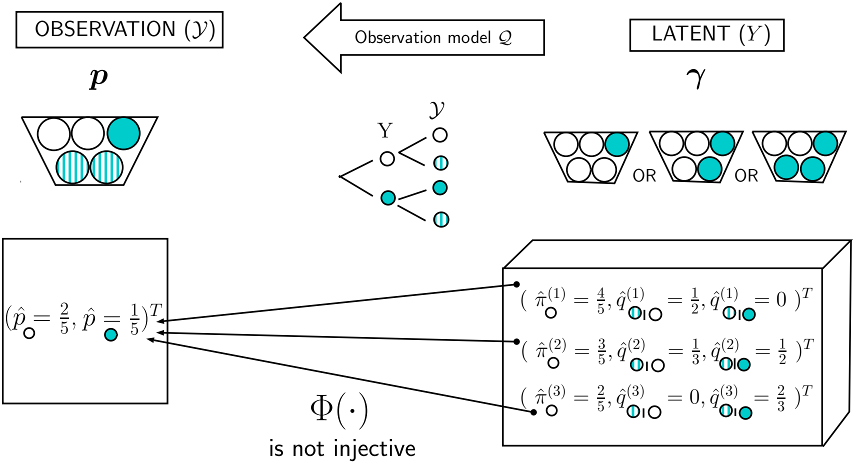
$$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n}, \frac{\sum_{\mathfrak{y} \ni y} n_{\mathfrak{y}}}{n}\right]$$

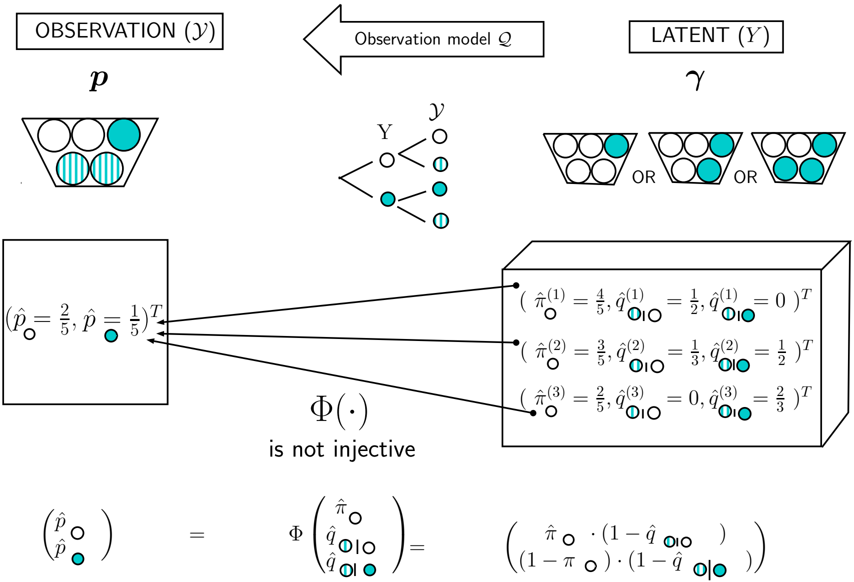
$$\hat{q}_{\mathfrak{y}|y} \in \left[0, \frac{n_{\mathfrak{y}}}{n_{\{y\}} + n_{\mathfrak{y}}}\right]$$

Basic idea (Plass, Augustin, Cattaneo, Schollmeyer, ISIPTA '15)



Basic idea (Plass, Augustin, Cattaneo, Schollmeyer, ISIPTA '15)




OBSERVATION ($\mathcal{Y}$)
Observation model $\mathcal{Q}$
LATENT (Y)
$\boldsymbol{p}$
$\boldsymbol{\gamma}$
$\mathcal{Y}$
Y
OR
OR
$\Phi(\cdot)$
is not injective

OBSERVABLE

LATENT

Use random-set perspective and determine ML estimator

$\hat{p}_{\mathscr{y}} = \hat{P}(\mathcal{Y} = \mathscr{y})$

$\longrightarrow$ $\boxed{\hat{p}_{\mathscr{y}} = \frac{n_{\mathscr{y}}}{n}}$

Use the **connection between** p and γ

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

$$\boldsymbol{\gamma} = (\boldsymbol{q}^T_{\mathscr{y}|\boldsymbol{y}},\ \boldsymbol{\pi}^T_{\boldsymbol{y}})^T$$

and the **invariance of the likelihood** under parameter transformations:

$$\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}$$

$$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n},\ \frac{\sum_{\boldsymbol{\mathfrak{y}} \ni y} n_{\boldsymbol{\mathfrak{y}}}}{n}\right]$$

$$\hat{q}_{\boldsymbol{\mathfrak{y}}|y} \in \left[0,\ \frac{n_{\boldsymbol{\mathfrak{y}}}}{n_{\{y\}} + n_{\boldsymbol{\mathfrak{y}}}}\right]$$

OBSERVABLE

LATENT

Use random-set perspective and determine ML estimator

$\hat{p}_{\mathscr{y}} = \hat{P}(\mathcal{Y} = \mathscr{y})$

$\longrightarrow$ $\hat{p}_{\mathscr{y}} = \frac{n_{\mathscr{y}}}{n}$

Use the **connection between** p and γ

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

$$\boldsymbol{\gamma} = (\boldsymbol{q}^T_{\mathscr{y}|\boldsymbol{y}},\ \boldsymbol{\pi}^T_{\boldsymbol{y}})^T$$

and the **invariance of the likelihood** under parameter transformations:

$$\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}$$

$$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n},\ \frac{\sum_{\mathfrak{y} \ni y} n_{\mathfrak{y}}}{n}\right]$$

$$\hat{q}_{\mathfrak{y}|y} \in \left[0,\ \frac{n_{\mathfrak{y}}}{n_{\{y\}} + n_{\mathfrak{y}}}\right]$$

Basic idea (Plass, Augustin, Cattaneo, Schollmeyer, ISIPTA '15)

OBSERVABLE

LATENT

Use random-set perspective and determine ML estimator

$\hat{p}_{\mathscr{y}} = \hat{P}(\mathcal{Y} = \mathscr{y})$

$\longrightarrow$ $\hat{p}_{\mathscr{y}} = \frac{n_{\mathscr{y}}}{n}$

Use the **connection between** p and γ

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

$$\boldsymbol{\gamma} = (\boldsymbol{q}^T_{\mathscr{y}|\boldsymbol{y}},\ \boldsymbol{\pi}^T_{\boldsymbol{y}})^T$$

and the **invariance of the likelihood** under parameter transformations:

$$\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}$$

$$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n}, \frac{\sum_{\mathfrak{y} \ni y} n_{\mathfrak{y}}}{n}\right]$$

$$\hat{q}_{\mathfrak{y}|y} \in \left[0, \frac{n_{\mathfrak{y}}}{n_{\{y\}} + n_{\mathfrak{y}}}\right]$$

$\downarrow$

Illustration (PASS data, wave 1)

$n_{<} = 238, \quad n_{\geq} = 835, \quad n_{\text{na}} = 338$

$\hat{\pi}_{<} \in [\frac{238}{1411}, \frac{238+338}{1411}]$

OBSERVABLE

LATENT

Use random-set perspective and determine ML estimator

$\hat{p}_{x\mathscr{y}} = \hat{P}(\mathcal{Y} = \mathscr{y} | X = x)$

$\longrightarrow$ $\hat{p}_{x\mathscr{y}} = \frac{n_{x\mathscr{y}}}{n_x}$

Use the **connection between** p and γ

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

$$\boldsymbol{\gamma} = (\boldsymbol{q}^T_{\mathscr{y}|\boldsymbol{xy}}, \ \boldsymbol{\pi}^T_{\boldsymbol{xy}})^T$$

and the **invariance of the likelihood** under parameter transformations:

$$\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}$$

$$\hat{\pi}_{xy} \in \left[\frac{n_{x\{y\}}}{n_x}, \ \frac{\sum_{\mathfrak{y} \ni y} n_{x\mathfrak{y}}}{n_x}\right]$$

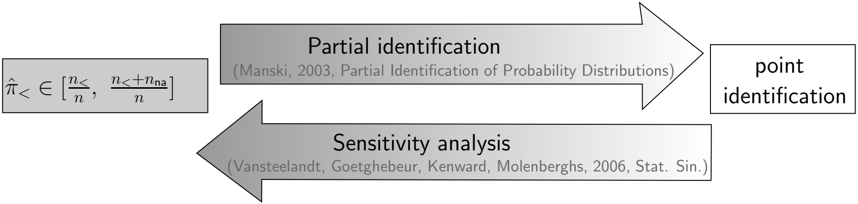
$$\hat{q}_{\mathfrak{y}|xy} \in \left[0, \ \frac{n_{x\mathfrak{y}}}{n_{x\{y\}} + n_{x\mathfrak{y}}}\right]$$

Illustration (PASS data, wave 1)

$\hat{\pi}_{0<} \in [0.41, \ 0.64]$ $\quad$ $\hat{\pi}_{1<} \in [0.10, \ 0.34]$

$\hat{\beta}_{<0} \in [-0.37, \ 0.59]$ $\quad$ $\hat{\beta}_{<} \in [-1.83, \ -1.25]$

Reliable incorporation of auxiliary information



Reliable incorporation of auxiliary information

$\hat{\pi}_< \in \left[\frac{n_<}{n}, \ \frac{n_< + n_{\text{na}}}{n}\right]$

Partial identification
(Manski, 2003, Partial Identification of Probability Distributions)

point identification

Sensitivity analysis
(Vansteelandt, Goetghebeur, Kenward, Molenberghs, 2006, Stat. Sin.)

↓ e.g.

Auxiliary information:

$\text{R} = \frac{q_{\text{na}|\geq}}{q_{\text{na}|<}}$

(Nordheim, 1984, J. Am. Stat. Assoc.)

CAR: R=1

↓ R=1

Refined estimators:

$\hat{\pi}_< = \frac{n_<}{n_< + n_\geq}$

Reliable incorporation of auxiliary information

$\hat{\pi}_< \in [\frac{n_<}{n}, \ \frac{n_< + n_{\text{na}}}{n}]$

Partial identification →
(Manski, 2003, Partial Identification of Probability Distributions)

point identification

← **Sensitivity analysis**
(Vansteelandt, Goetghebeur, Kenward, Molenberghs, 2006, Stat. Sin.)

↓ e.g.

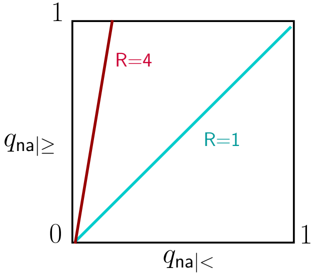
$\text{R} = \frac{q_{\text{na}|\geq}}{q_{\text{na}|<}}$

(Nordheim, 1984, J. Am. Stat. Assoc.)

CAR: R=1

↓ R=1

$\hat{\pi}_< = \frac{n_<}{n_< + n_\geq}$



Reliable incorporation of auxiliary information

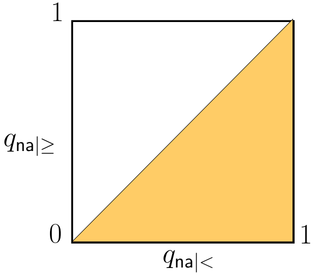
$\hat{\pi}_< \in [\frac{n_<}{n}, \frac{n_< + n_{\text{na}}}{n}]$

Partial identification
(Manski, 2003, Partial Identification of Probability Distributions)

point identification

Sensitivity analysis
(Vansteelandt, Goetghebeur, Kenward, Molenberghs, 2006, Stat. Sin.)

↓ e.g.



$R = \frac{q_{\text{na}|\geq}}{q_{\text{na}|<}}$

(Nordheim, 1984, J. Am. Stat. Assoc.)

CAR: R=1

↓ R=1

$\hat{\pi}_< = \frac{n_<}{n_< + n_\geq}$

Reliable incorporation of auxiliary information

$\hat{\pi}_< \in \left[\frac{n_<}{n}, \frac{n_< + n_{\mathsf{na}}}{n}\right]$

Partial identification →
(Manski, 2003, Partial Identification of Probability Distributions)

point identification

← **Sensitivity analysis**
(Vansteelandt, Goetghebeur, Kenward, Molenberghs, 2006, Stat. Sin.)

↓ e.g.

$R = \frac{q_{\mathsf{na}|\geq}}{q_{\mathsf{na}|<}}$
(Nordheim, 1984, J. Am. Stat. Assoc.)

CAR: R=1

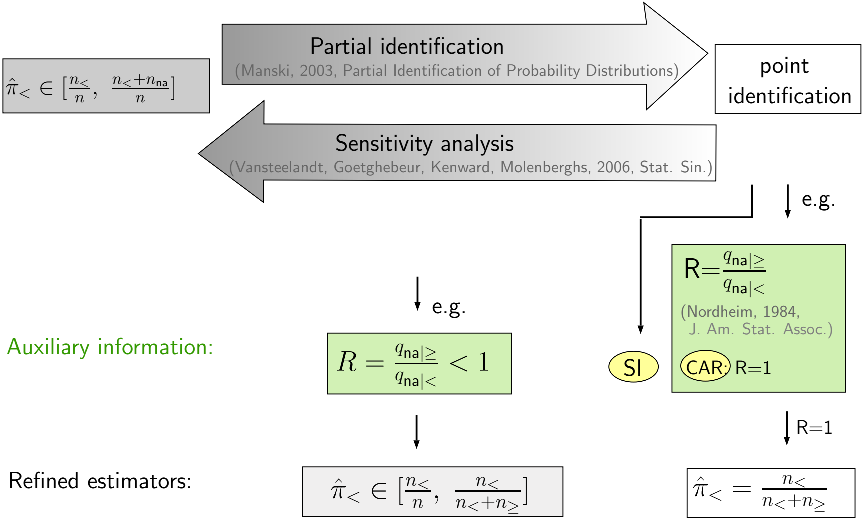
↓ R=1

↓ e.g.

Auxiliary information: $R = \frac{q_{\mathsf{na}|\geq}}{q_{\mathsf{na}|<}} < 1$

↓

Refined estimators: $\hat{\pi}_< \in \left[\frac{n_<}{n}, \frac{n_<}{n_< + n_\geq}\right]$ $\quad$ $\hat{\pi}_< = \frac{n_<}{n_< + n_\geq}$


$\hat{\pi}_< \in [\frac{n_<}{n}, \frac{n_< + n_{na}}{n}]$
Partial identification
(Manski, 2003, Partial Identification of Probability Distributions)
point identification
Sensitivity analysis
(Vansteelandt, Goetghebeur, Kenward, Molenberghs, 2006, Stat. Sin.)
e.g.
$R = \frac{q_{na|\geq}}{q_{na|<}}$
(Nordheim, 1984, J. Am. Stat. Assoc.)
CAR: R=1
SI
e.g.
Auxiliary information:
$R = \frac{q_{na|\geq}}{q_{na|<}} < 1$
R=1
Refined estimators:
$\hat{\pi}_< \in [\frac{n_<}{n}, \frac{n_<}{n_< + n_\geq}]$
$\hat{\pi}_< = \frac{n_<}{n_< + n_\geq}$

Coarsening at random & subgroup independence

	coarsening at random (CAR) (Heitjan, Rubin, 1991, Ann. Stat.)	subgroup independence (SI)
Generally	For each fixed $\mathscr{Y}$, $q_{\mathscr{Y}\|y}$ takes the same values for all y that are consistent with $\mathscr{Y}$	Coarsening does not depend on the value of the covariate
Example	$q_{\text{na}\|<} = q_{\text{na}\|\geq}$	$q_{\text{na}\|0<} = q_{\text{na}\|1<}$ and $q_{\text{na}\|0\geq} = q_{\text{na}\|1\geq}$

Coarsening at random & subgroup independence

	$<$	$\geq$	na	total
0	38	385	95	518
1	36	42	9	87

Table: PASS data, wave 5

$$\hat{\pi}_{0<} \in [0.07; 0.26] \qquad \hat{\pi}_{1<} \in [0.41; 0.52]$$
$$\hat{q}_{\text{na}|0<} \in [0; 0.71] \qquad \hat{q}_{\text{na}|1<} \in [0; 0.2]$$
$$\hat{q}_{\text{na}|0\geq} \in [0; 0.20] \qquad \hat{q}_{\text{na}|1\geq} \in [0; 0.18]$$

X=0

$\underline{\hat{\pi}}_{0<} = 0.07$
$\underline{\hat{q}}_{\text{na}|0<} = 0$
$\overline{\hat{q}}_{\text{na}|0\geq} = 0.20$

...

$\hat{\pi}^{(\text{CAR})}_{0<} = 0.09$
$\hat{q}^{(\text{CAR})}_{\text{na}|0<} = 0.18$
$\hat{q}^{(\text{CAR})}_{\text{na}|0\geq} = 0.18$

...

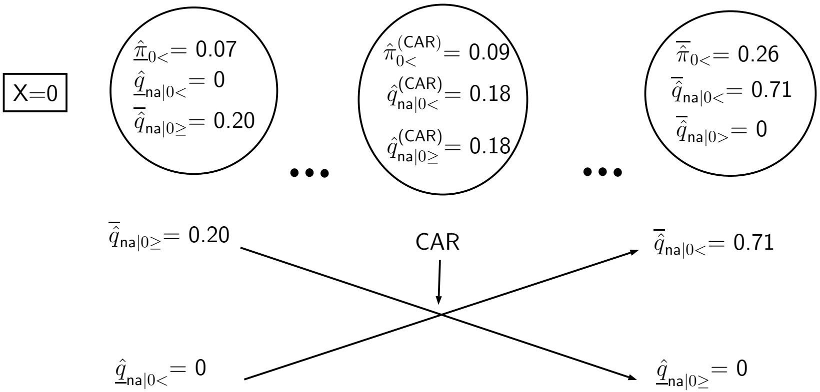
$\overline{\hat{\pi}}_{0<} = 0.26$
$\overline{\hat{q}}_{\text{na}|0<} = 0.71$
$\overline{\hat{q}}_{\text{na}|0>} = 0$

Coarsening at random & subgroup independence

	$<$	$\geq$	na	total
0	38	385	95	518
1	36	42	9	87

Table: PASS data, wave 5

$$\hat{\pi}_{0<} \in [0.07; 0.26] \qquad \hat{\pi}_{1<} \in [0.41; 0.52]$$
$$\hat{q}_{\text{na}|0<} \in [0; 0.71] \qquad \hat{q}_{\text{na}|1<} \in [0; 0.2]$$
$$\hat{q}_{\text{na}|0\geq} \in [0; 0.20] \qquad \hat{q}_{\text{na}|1\geq} \in [0; 0.18]$$



	$<$	$\geq$	na	total
0	38	385	95	518
1	36	42	9	87

Table: PASS data, wave 5

$$\hat{\pi}_{0<} \in [0.07; 0.26] \qquad \hat{\pi}_{1<} \in [0.41; 0.52]$$
$$\hat{q}_{na|0<} \in [0; 0.71] \qquad \hat{q}_{na|1<} \in [0; 0.20]$$
$$\hat{q}_{na|0\geq} \in [0; 0.20] \qquad \hat{q}_{na|1\geq} \in [0; 0.18]$$

X=0

$\underline{\hat{\pi}}_{0<} = 0.07$
$\underline{\hat{q}}_{na|0<} = 0$
$\overline{\hat{q}}_{na|0\geq} = 0.20$
...

$\overline{\hat{\pi}}_{0<} = 0.26$
$\overline{\hat{q}}_{na|0<} = 0.71$
$\underline{\hat{q}}_{na|0>} = 0$
...

SI
$q_{na|0<} = q_{na|1<}$
and
$q_{na|0\geq} = q_{na|1\geq}$

same subgroups:
$\Phi(\cdot)$ is not injective

otherwise:
$\Phi(\cdot)$ is injective

But:
Estimators under SI potentially do not maximize the likelihood

X=1

$\underline{\hat{\pi}}_{1<} = 0.41$
$\underline{\hat{q}}_{na|1\geq} = 0$
$\overline{\hat{q}}_{na|0\geq} = 0.18$
...

$\overline{\hat{\pi}}_{1<} = 0.52$
$\overline{\hat{q}}_{na|1\geq} = 0.20$
$\underline{\hat{q}}_{na|1\geq} = 0$
...

Coarsening at random & subgroup independence

	$<$	$\geq$	na	total
0	38	385	95	518
1	36	42	9	87

Table: PASS data, wave 5

$$
\begin{array}{lll lll}
\hat{\pi}_{0<} & \in & [0.07; 0.26] & \hat{\pi}_{1<} & \in & [0.41; 0.52] \\
\hat{q}_{na|0<} & \in & [0; 0.71] & \hat{q}_{na|1<} & \in & [0; 0.20] \\
\hat{q}_{na|0\geq} & \in & [0; 0.20] & \hat{q}_{na|1\geq} & \in & [0; 0.18]
\end{array}
$$

CAR

SI

One can never reject CAR

$$
\begin{aligned}
\hat{\pi}_{1<}^{(SI)} &= \frac{n_{1<}}{n_1} \frac{n_{1\geq} n_0 - n_1 n_{0\geq}}{n_{0<} n_{1\geq} - n_{0\geq} n_{1<}} \\
&= 0.39 \quad (\text{cf. } \hat{\pi}_{1<} \in [0.41; 0.52])
\end{aligned}
$$

$\Rightarrow$ Construction of hypothesis test with

H_0 : $q_{na|0<} = q_{na|1<}$ & $q_{na|0\geq} = q_{na|1\geq}$

H_1 : no restrictions on the coarsening parameters

Summary and outlook

- Via the observation model $\mathcal{Q}$ maximum-likelihood estimators referring to the latent variable may be obtained for both cases
  - ... the homogeneous case
  - ... the case with categorical covariates
- Proper inclusion of auxiliary information via further restrictions on $\mathcal{Q}$

Next steps:

- Likelihood-based hypothesis tests and uncertainty regions
- Comparison to Bayesian approaches
- Applying the observation model to coarse ordinal data

References

Couso, Dubois, Sánchez.
Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables, Springer, 2014.

Heitjan, Rubin.
Ignorability and Coarse Data, Annals of Statistics, 1991.

Manski.
Partial Identification of Probability Distributions, Springer, 2003.

E. Nordheim.
Inference from nonrandomly missing categorical data:An example from a genetic study on Turner's syndrome, J. Am. Stat. Assoc., 1984.

Plass, Augustin, Cattaneo, Schollmeyer.
Statistical modelling under epistemic data imprecision: some results on estimating multinomial distributions and logistic regression for coarse categorical data, ISIPTA, 2015.

Vansteelandt, Goetghebeur, Kenward, Molenberghs.
Ignorance and uncertainty regions as inferential tools in a sensitivity analysis, Stat. Sin., 2006.