Discussion of invited talks in Session 3 Data, data science and official statistics

Li-Chun Zhang

I. Uncertainty due to unknown selection

«Uncertainty-based analysis for non-probability samples» — Pier Luigi Conti, Daniela Marella

We observe (δ, x) and $y \mid x, \delta = 1$, but not $y \mid x, \delta = 0$, where

$$\begin{aligned} p(y \mid x, \delta = 0) &= p(\delta = 0, y, x)/p(x, \delta = 0) \\ &= \big[p(y \mid x) - p(\delta = 1, y \mid x)\big] p(x)/p(x, \delta = 0) \end{aligned}$$

However, $$\textstyle\sum_y \big[\cdots\big] = 1 - p(\delta = 1 \mid x)$$

Measure of uncertainty space:

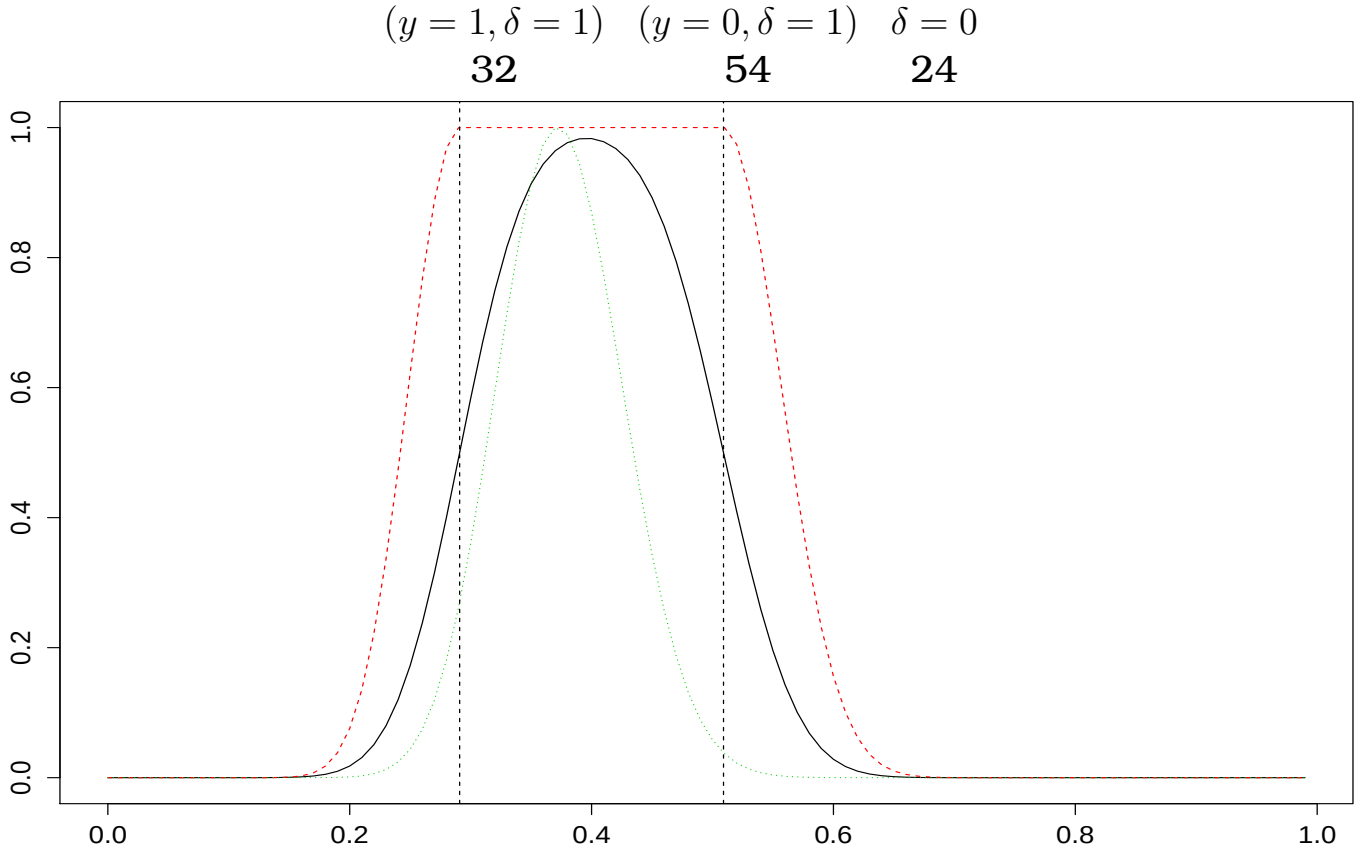
$$\textit{conditional} \;\; U^x\big[p(y \mid x)\big] = p(\delta = 0 \mid x)$$

$$\textit{marginal} \;\; U\big[p(y)\big] = E\{U^x\big[p(y \mid x)\big]\} = p(\delta = 0)$$

(e.g. Manski, 1993). Moreover, extra-sample information can reduce the uncertainty (due to unknown selection)

Discussion: related approaches to inference of $p(y)$ which do not require extra assumptions for point identification

Example: missing binary observations



Green: likelihood of $\theta = p(y=1)$ assuming MCAR (pointwise identification)

Red: profile lik. of θ from trinomial sampling $\{(y=1,\delta=1),(y=0,\delta=1),(\delta=0)\}$

NB. observations $x=(32,54,24) \mapsto \widehat{\Theta}$ of max. lik. θ-values (flat top)

Solid: ***corroboration*** of θ given as $\hat{c}(\theta) = \Pr(\theta \in \widehat{\Theta}; \hat{\psi})$

NB. probability with respect to observable trinomial sampling

Minimal inference (Zhang & Chambers, 2019)

Only based on the (trinomial) sampling distribution $f(x;\psi)$

which is agreeable to all, true ψ_0 identifiable, MLE $\hat{\psi}(x) \xrightarrow{P} \psi_0$

Consistent estimation of true corroboration

$$c_0(\theta) = c(\theta;\psi_0) = \Pr(\theta \in \widehat{\Theta};\psi_0)$$

or true level-α corroboration set

$$A_\alpha(\psi_0) = \{\theta : c(\theta;\psi_0) \geq \alpha\}$$

Define *observed maximum corroboration set*

$$\hat{A}^{max} = A^{max}(\hat{\psi})$$

θ-values in $\hat{A}^{max}$ are the *hardest to refute* given x (the data)
In contrast, values in $\widehat{\Theta}$ are *most likely to be true*

Corroboration test $H_A : \theta^* \in \Theta_0$ vs. $H_B : \theta^* \notin \Theta_0$ is strongly Chernoff-consistent, can reject θ^* with power $1 - \hat{c}(\theta^*)$

$\widehat{\Theta}$	$\theta^* = 0.2$	$\theta^* = 0.3$	$\theta^* = 0.4$	$\theta^* = 0.5$	$\theta^* = 0.6$
$[0.29, 0.51]$	$\hat{c} = 0.018$	$\hat{c} = 0.583$	$\hat{c} = 0.985$	$\hat{c} = 0.576$	$\hat{c} = 0.028$

II. Finite-sample conformal intervals

«Conformal Methods in Official Statistics: Perspectives and Challenges» — Nina Deliu, Brunero Liseo
Finite sample of size n, *level*-α prediction interval $\mathcal{C}_{n,\alpha}$ s.t.

$$\Pr\{y \in \mathcal{C}_{n,\alpha}(x)\} \geq \alpha$$

- *Full-conformal* by a lemma of sample quantile-q,

$$\Pr(z \leq q_{n+1,\alpha}) \geq \alpha \quad \text{given IID } \{z_1, ..., z_n\} \cup \{z\}$$

  with $z_i = |y_i - f_{n+1}(x_i)|$, $z = |y - f_{n+1}(x_{n+1})|$, arbitrary f_{n+1} from *augmented sample* $\{(y_i, x_i) : i = 1, ..., n\} \cup \{(y, x_{n+1})\}$
  NB. repeatedly evaluating a grid of stipulated y-values

- *Split-conformal* using random partition $s_1 \cup s_2 = \{1, ..., n\}$
  - obtain $f(x, s_1)$ from $\{(y_i, x_i) : i \in s_1\}$, training set s_1
  - $\Pr(z \leq q_{n_2,\alpha}) \geq \alpha$, $n_2 = |s_2|$, $z_i = |y_i - f(x_i, s_1)|$ for $i \in s_2$
    NB. $f(x, s_1)$ is constant of $j \notin s_1$ conditional on s_1

Discussion: split-conformal under ***design-based predictive inference*** *framework, in contrast to IID/WR sampling — Zhang, Sande-Sanguiao & Lee (2024), available at JOS*

What is design-based prediction?

Given a sample of observations, e.g. $\{(y_i, x_i) : i = 1, ..., n\}$

- obtain a function $\mu(s)$, which varies with $s = \{1, ..., n\}$
- prediction if $\mu(s)$ targets something random, estimation if fixed
- design-based prediction if only $s \sim p(s)$, but all (y_i, x_i) are fixed

Example: Fixed population U and associated values $\{y_i : i \in U\}$

Element	i_1	i_2	i_3	i_4
Value y_i	1	2	3	6

s by simple random sampling without replacement, $|s| = 2$

Sample s	(i_1, i_2)	(i_1, i_3)	(i_1, i_4)	(i_2, i_3)	(i_2, i_4)	(i_3, i_4)
Sample mean $\mu(s) = \bar{y}_s$	1.5	2	3.5	2.5	4	4.5
Out-of-sample mean $\bar{y}_R$	4.5	4	2.5	3.5	2	1.5
Unknown $\{y_k : k \notin s\}$	$y_3 = 3$ $y_4 = 6$	$y_2 = 2$ $y_4 = 6$	$y_2 = 2$ $y_3 = 3$	$y_1 = 1$ $y_4 = 6$	$y_1 = 1$ $y_3 = 3$	$y_1 = 1$ $y_2 = 2$
$\{(y_k - \mu(s))^2 : k \notin s\}$	2.25 20.25	0 16	2.25 0.25	2.25 12.25	9 1	12.25 6.25
$D_R = \sum_{k \notin s} (y_k - \mu(s))^2$	22.5	16	2.5	14.5	10	18.5

Random $R = U \setminus s$, $\bar{y}_R$, $\{y_k : k \notin s\}$ or D_R as sample s varies

$E_p(\bar{y}_s - \bar{y}_R) = 0$, i.e. unbiased **prediction** of $\bar{y}_R$ w.r.t. $p(s)$

*Predict D_R as total squared error of **unit-level** prediction ?*

Design-based predictive inference

Target y by $\mu(x)$ given x, *arbitrary* model or algorithm μ

- **pq-design**: $s \sim p(s)$, $s_1 \sim q(s_1 \mid s)$, $s_2 = s \setminus s_1$
  e.g. training set s_1 by K-fold partition, SRS or bootstrap given s
- **subsampling Rao-Blackwellisation (SRB)** of $\mu(x, s_1)$

$$\bar{\mu}(x, s) = E_q\big[\mu(x, s_1) \mid s\big]$$

- **finite-sample design-unbiased prediction** of

$$\text{STE} = \Big(\sum_{i \notin s} y_i - \bar{\mu}(x_i, s)\Big)^2 \text{ or } \text{TSE} = \sum_{i \notin s} \big(y_i - \bar{\mu}(x_i, s)\big)^2$$

Now, apply SRB under pq-design for <u>intervals</u> instead

- Conditional on s_1, treat s_2 as a sample from $U \setminus s_1$ under pq-design

$$q(s_1 \mid s)p(s) = f(s_1)f(s \mid s_1) \quad \Rightarrow \quad \pi_{2i} = \Pr(i \in s_2 \mid s_1)$$

- obtain π_{2i}-unbiased estimation of the distribution of error of $\mu(x, s_1)$ in $U \setminus s_1$ based on sample $\{y_i - \mu(x_i, s_1) : i \in s_2\}$ — thereby **predict the coverage** of $\big[\mu(x_i, s_1) - z_L,\ \mu(x_i, s_1) + z_U\big]$ over *all* $i \notin s$
  NB. split-conformal intervals $\mu(x_i, s_1) \pm \cdots$ biased if WOR-sampling
- Apply SRB to recover loss of efficiency due to single split