Bias-Corrected Crosswise Estimators for Sensitive Inquiries

Yuki Atsusaka (website, paper) and Randy Stevenson, Rice University

MOTIVATION

The conventional crosswise estimator for the population prevalence of sensitive traits is **biased** in the presence of **inattentive respondents**.

$\Rightarrow$ We propose a simple design-based bias correction and its extensions including sensitivity analysis, weighting strategy, and regression models.

CROSSWISE MODEL (AND ITS PITFALL)

> **Instruction: Please read the two statements below**
> Statement A: I would feel uncomfortable if an immigrant family moved in next door
> Statement B: My mother was born in January, February, or March
>
> **Crosswise Question: Which of the following fits your case?**
> (1) Both statements are TRUE, or both statements are FALSE
> (2) Otherwise

π: population prevalence of sensitive traits (A) (Quantity of interest)

p: **known** population proportion for (B)

λ: population proportion for choosing (1)

$\Rightarrow \mathbb{P}(\text{TRUE-TRUE} \cup \text{FALSE-FALSE}) = \lambda = \pi p + (1-\pi)(1-p)$ [Naïve model]

Naïve crosswise estimator: $\widehat{\pi}_{CM} = \frac{\widehat{\lambda}+p-1}{2p-1}$

In the presence of inattentive respondents (who don't follow the design):

γ: population proportion of attentive respondents

κ: probability with which inattentive respondents choose (1)

$\Rightarrow \lambda = \left\{\pi p + (1-\pi)(1-p)\right\}\gamma + \kappa(1-\gamma)$ [True model]

Bias in the naïve estimator: $\mathbb{E}[\widehat{\pi}_{CM}] - \pi = \left(\frac{1}{2} - \frac{1}{2\gamma}\right)\left(\frac{\lambda-\kappa}{p-\frac{1}{2}}\right)$

BIAS-CORRECTED ESTIMATOR

Our strategy: Estimate the proportion of inattentive respondents by adding an Anchor Question $\rightarrow$ estimate the bias $\rightarrow$ correct the bias.

> [*suppose that this survey is administered in the U.S.]
> **Instruction: Please read the two statements below**
> Statement C: I am taking this survey in France
> Statement D: My best friend was born in January, February, or March
>
> **Anchor Question: Which of the following fits your case?**
> (3) Both statements are TRUE, or both statements are FALSE
> (4) Otherwise

π': **known** population proportion for (C) [we choose $\pi' = 0$]

p': **known** population proportion for (D)

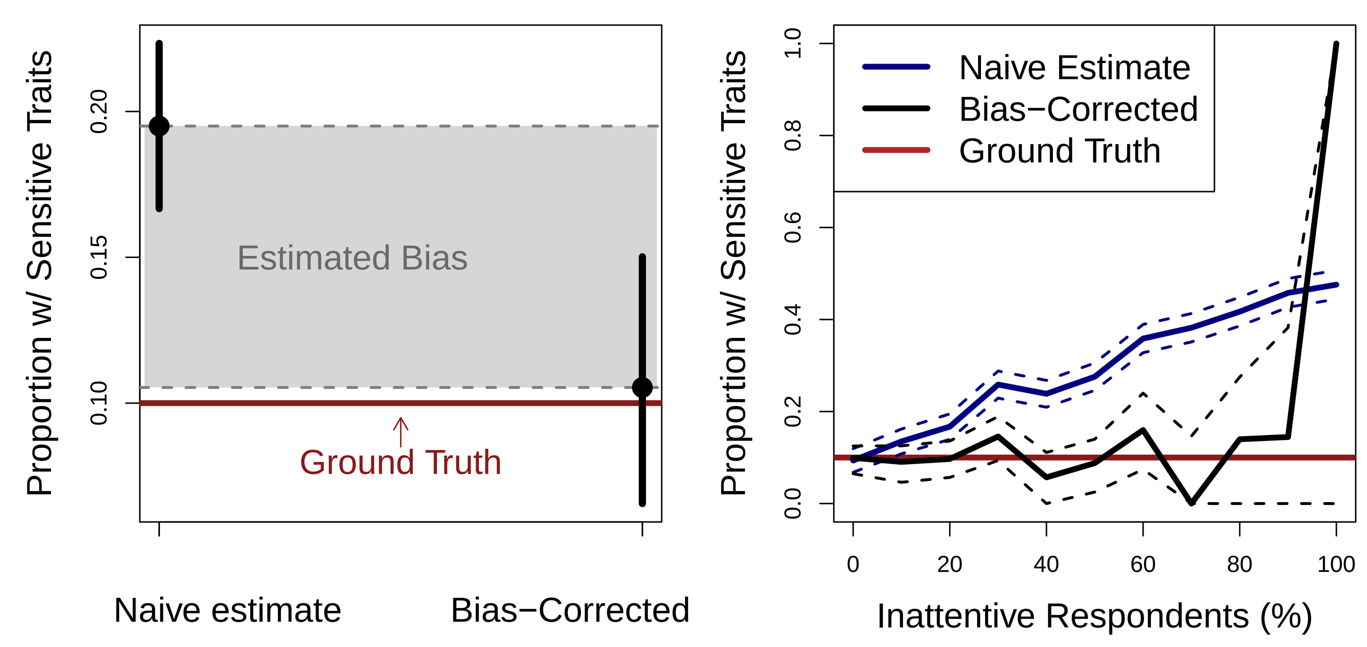
$\widehat{\lambda}'$: Observed proportion for choosing (3) [we set $\kappa = 0.5$]

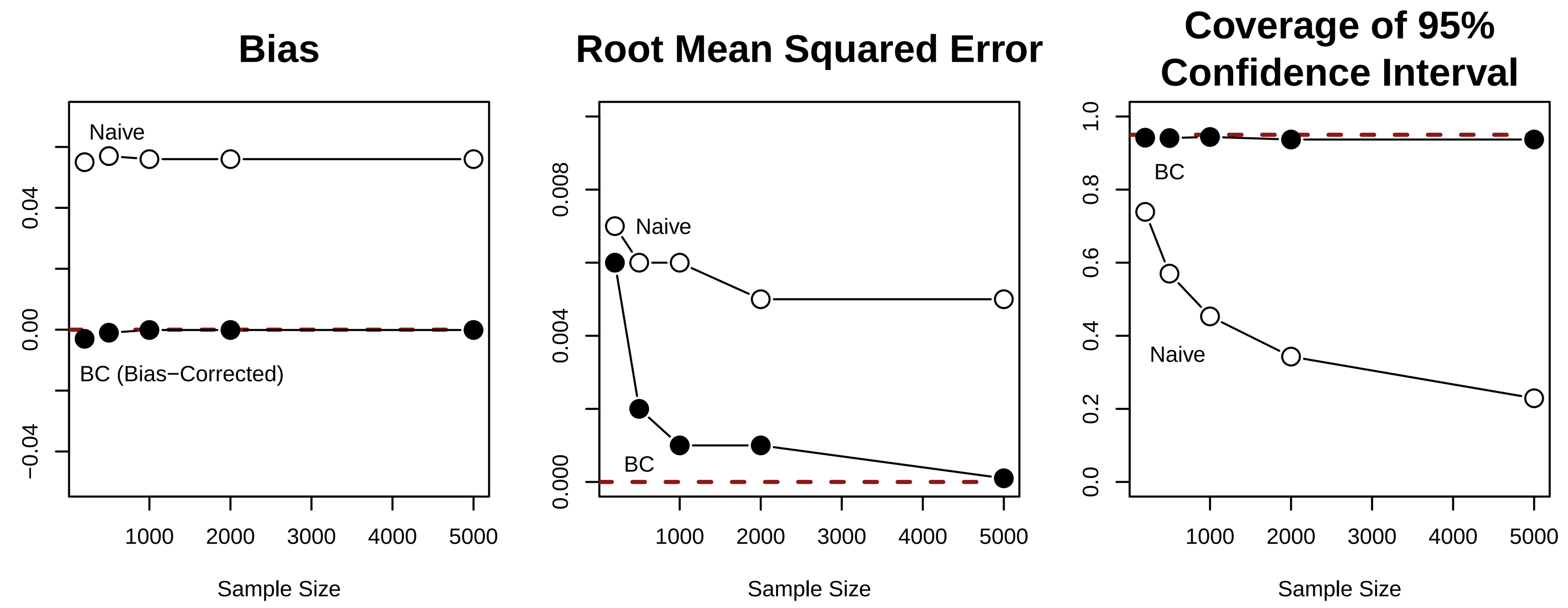
$\Rightarrow$ Estimated the proportion of attentive respondents: $\widehat{\gamma} = \frac{\widehat{\lambda}'-\frac{1}{2}}{\frac{1}{2}-p'}$

Bias-corrected estimator: $\widehat{\pi}_{BC} = \widehat{\pi}_{CM} - \left(\frac{1}{2} - \frac{1}{2\widehat{\gamma}}\right)\left(\frac{\widehat{\lambda}-\kappa}{p-\frac{1}{2}}\right)$

Estimator for sample variance: $\mathbb{V}(\widehat{\pi}_{BC}) = \widehat{\mathbb{V}}\left[\frac{\widehat{\lambda}}{\widehat{\lambda}'-\frac{1}{2}}\right]$ (bootstrapped in practice)

SIMULATION STUDIES

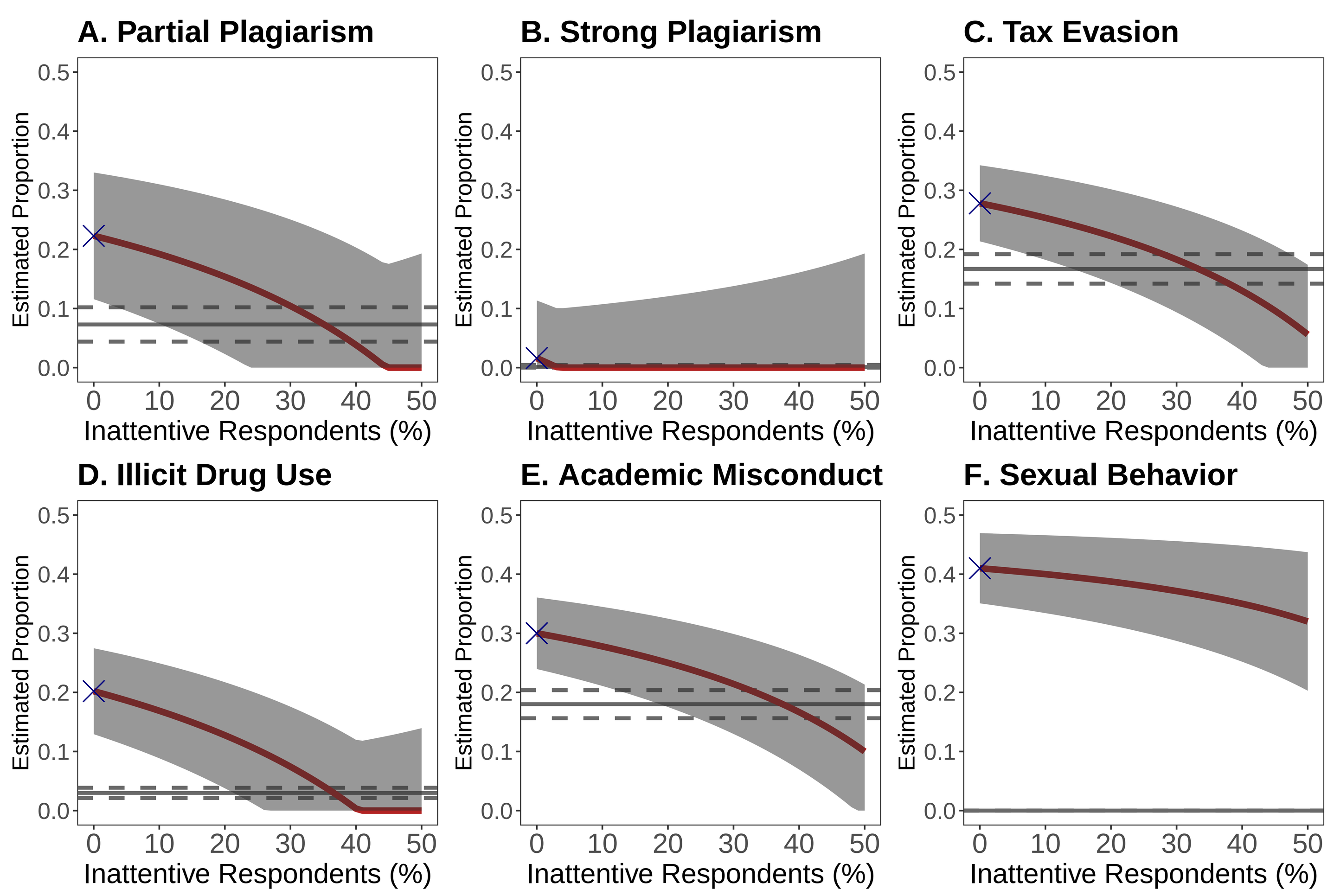




SENSITIVITY ANALYSIS

We can also apply our bias-correction to surveys where the anchor question is not available to ask: How sensitive is the original conclusion to potential inattentive respondents?

Sensitivity Analysis Applied to Six Published Studies



× Original Estimate — Bias-corrected Estimate — Direct Questioning

$\Rightarrow$ Allow researchers to make an assumption about inattentive respondents in order to keep their original claims.

WEIGHTING

Sensitive questions are usually asked in surveys with unrepresentative samples (e.g., online opt-in sample) $\rightarrow$ Weighting is essential to extend inferences.

Y_i: Binary indicator for choosing (1)

A_i: Binary indicator for choosing (3)

S_i: Binary indicator for being in the sample; $\mathbf{X_i}$: covariates.

Survey weight: $w_i = \mathbb{P}(S_i = 1|\mathbf{X_i})^{-1}$

Our strategy: Apply the Horvitz-Thompson estimator of the mean as:

$$\widehat{\lambda}_w = \frac{1}{n}\left(\sum_{i=1}^{n} \widehat{w}_i Y_i S_i\right), \quad \text{where } \mathbb{E}[\widehat{\lambda}_w] = \lambda$$
$$\widehat{\gamma}_w = \frac{\frac{1}{n}\left(\sum_{i=1}^{n} \widehat{w}_i A_i S_i\right) - \frac{1}{2}}{\frac{1}{2} - p'}, \quad \text{where } \mathbb{E}[\widehat{\gamma}_w] = \gamma$$

CROSSWISE REGRESSIONS (WITH BIAS CORRECTION)

Our strategy: Model the joint probability of the regression of interest and entire crosswise data $\rightarrow$ maximize the log-likelihood for estimation.

1. Using the Latent Sensitive Trait as an Outcome

Z_i: Latent indicator for having a sensitive trait

T_i: Latent indicator for being attentive (i.e., following the instruction)

Regression of interest: $\mathbb{E}[Z_i|\mathbf{X_i} = \mathbf{x}] = \mathbb{P}(Z_i = 1|\mathbf{X_i} = \mathbf{x}) = \pi_\beta(\mathbf{x})$

Auxiliary CEF: $\mathbb{E}[T_i|\mathbf{X_i} = \mathbf{x}] = \mathbb{P}(T_i = 1|\mathbf{X_i} = \mathbf{x}) = \gamma_\theta(\mathbf{x})$

Logit specification: $\pi_\beta(\mathbf{x}) = logit^{-1}(\beta\mathbf{X_i}); \quad \gamma_\theta(\mathbf{x}) = logit^{-1}(\theta\mathbf{X_i})$

The observed-data likelihood function (with bias-correction):

$$\mathcal{L}(\beta,\theta|\{\mathbf{X_i}, Y_i, A_i\}_{i=1}^{n}, p, p')$$
$$= \prod_{i=1}^{n}\left\{\lambda_\beta(\mathbf{X_i})\right\}^{Y_i}\left\{1-\lambda_\beta(\mathbf{X_i})\right\}^{1-Y_i}\left\{\lambda'_\theta(\mathbf{X_i})\right\}^{A_i}\left\{1-\lambda'_\theta(\mathbf{X_i})\right\}^{1-A_i}$$
$$= \prod_{i=1}^{n}\left\{\left((2p-1)\pi_\beta(\mathbf{X_i}) + \left(\frac{1}{2}-p\right)\right)\gamma_\theta(\mathbf{X_i}) + \frac{1}{2}\right\}^{Y_i}$$
$$\times \left\{1 - \left[\left((2p-1)\pi_\beta(\mathbf{X_i}) + \left(\frac{1}{2}-p\right)\right)\gamma_\theta(\mathbf{X_i}) + \frac{1}{2}\right]\right\}^{1-Y_i}$$
$$\times \left\{\left(\frac{1}{2}-p'\right)\gamma_\theta(\mathbf{X_i}) + \frac{1}{2}\right\}^{A_i} \times \left\{1 - \left[\left(\frac{1}{2}-p'\right)\gamma_\theta(\mathbf{X_i}) + \frac{1}{2}\right]\right\}^{1-A_i}$$

2. Using the Latent Sensitive Trait as a Predictor

V_i: continuous or discrete response variable

Regression of interest: $g_\Theta(V_i|\mathbf{X_i} = \mathbf{x}, Z_i = z)$

The observed-data likelihood function (with bias-correction):

$$\mathcal{L}(\beta,\theta,\Theta|\{V_i, \mathbf{X_i}, Y_i, A_i\}_{i=1}^{n}, p, p')$$
$$= \prod_{i=1}^{n} g_\Theta(V_i|\mathbf{X_i}, Z_i, T_i)\mathbb{P}(Y_i = 1, Z_i, T_i|\mathbf{X_i})\mathbb{P}(A_i = 1, Z_i, T_i|\mathbf{X_i})$$
$$= \prod_{i=1}^{n}\Big\{g_\Theta(V_i|\mathbf{X_i},1,1)p^{Y_i}(1-p)^{1-Y_i}\pi_\beta(\mathbf{X_i})(1-p')^{A_i}p'^{1-A_i}\gamma_\theta(\mathbf{X_i})$$
$$+ g_\Theta(V_i|\mathbf{X_i},0,1)(1-p)^{Y_i}p^{1-Y_i}(1-\pi_\beta(\mathbf{X_i}))(1-p')^{A_i}p'^{1-A_i}\gamma_\theta(\mathbf{X_i})$$
$$+ g_\Theta(V_i|\mathbf{X_i},1,0)\tfrac{1}{2}\pi_\beta(\mathbf{X_i})\tfrac{1}{2}(1-\gamma_\theta(\mathbf{X_i}))$$
$$+ g_\Theta(V_i|\mathbf{X_i},0,0)\tfrac{1}{2}(1-\pi_\beta(\mathbf{X_i}))\tfrac{1}{2}(1-\gamma_\theta(\mathbf{X_i}))\Big\}$$