Integrating probability and non-probability samples to improve analytic inference and reduce costs

Camilla Salvatore[a],
Biffignandi S.[b], Sakshaug J.W.[c], Wiśniowski A. [d], Struminskaya B.[a]

a: Utrecht University, b: University of Bergamo, c: German Institute for Employment Research, d: University of Manchester

September 1, 2023

Introduction

Probability samples (PS)

Allow inferences to the general population

- Rely on sampling theory
- Design/Model based inference
- **Falling response rate, time-consuming, expensive**

Non-Probability samples (NPS)

Drawing inference is hard or not possible

- **More affordable, timely, conv.**
- No unified inferential framework
- Unknown selection mechanism: Self-selection $\rightarrow$ selection bias (SB)

Comparing PS and NPS estimates (Pasek, 2016):

- **Finite population** estimates tend to be more dissimilar than **correlations** and **regression coefficients**
- **No consensus** about whether and in which cases **differences** will be notable

The context

Problem

A researcher is interested in making inferences from a PS survey but cannot afford a large sample size

The context

Problem

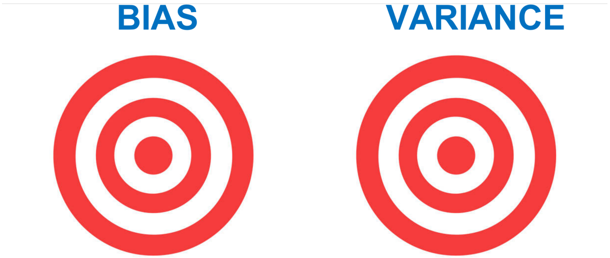
A researcher is interested in making inferences from a PS survey but cannot afford a large sample size

Alternatives

1. Reduce the sample size: small PS $\rightarrow$ large variance but theoretically unbiased estimates
2. Opt for a NPS: bias but low variance

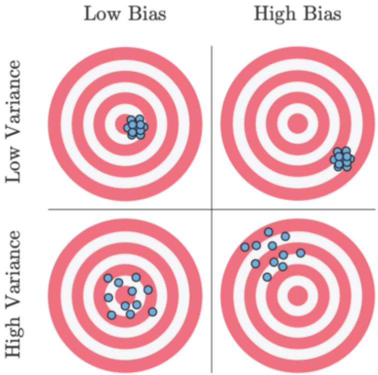
The context

Bias-variance trade off



The context

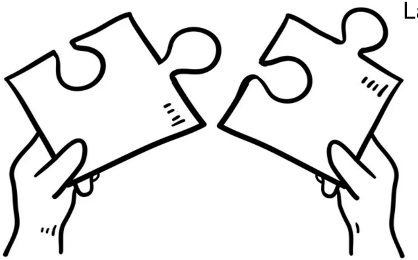
Bias-variance trade off



Our Proposal

The Data Integration Puzzle

PS (high quality)
Unbiased
Large variance



NPS (lower quality)
Selection Bias
Lower variance

Our Proposal

The Data Integration Perspective

- **Integrate** small PS + larger NPS
- to improve inference on **logistic regression coefficients**
- under the **Bayesian** framework
- **reducing survey costs**

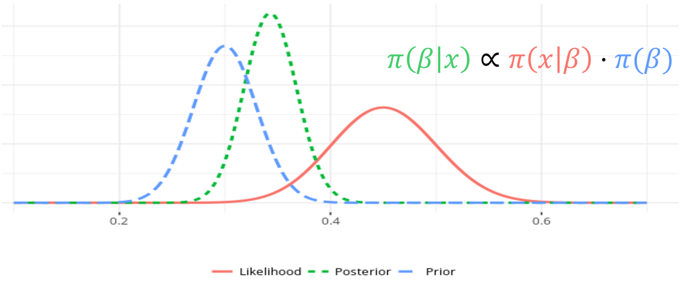
Inference

- Based on **small PS data** (unbiased, high var.)
- **Incorporation** of **biased NPS data** into the estimation process (low var.)
- Posterior estimates are likely to have more bias than PS estimates but possibly less variance (**bias/var trade-off**)

Why Bayesian? (Kruschke, 2014; Gelman et al., 2013)

- **Natural choice** to integrate data with varying levels of quality
- Its structure can be exploited in order to **incentivize high-quality** data

Posterior $\propto$ **Likelihood (PS)** $\cdot$ **Prior(NPS)**



Research structure

- **Background**: Sakshaug et al. (2019) and Wiśniowski et al. (2020) papers (**Continuous** outcome variable)
- **Part I** - Simulation study (100 repetitions):
  - **Different selection scenarios**, **prior** specifications, PS and NPS sizes
  - Evaluate the **performance** of several informative priors against a PS-ONLY one in terms of **MSE**
- **Part II** - Real data analysis:
  - American Trend Panel + 9 parallel NPS surveys
  - Shiny app with interactive cost analysis

Priors

PS-ONLY (No data integration):

- A weakly informative prior proposed by Gelman et al. (2008)
- **Control prior** against which compare data integration results

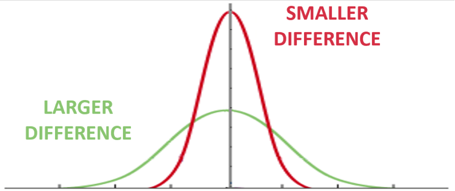
$$\beta_j \sim \mathit{Student}\,(\nu = 3, \mu = 0, s = 2.5) \quad \text{for} \quad j = 0, 1, 2$$

Informative priors: integrating PS and NPS data

Distances priors: The influence of the prior depends on the difference between ML estimates. Example:

- *Distance prior*

$$\beta_j \sim \mathcal{N}\left(\hat{\beta}_{NP}, |\hat{\beta}_P - \hat{\beta}_{NP}|\right) \quad \text{for} \quad j = 0, 1, 2$$



Mixed distance priors: Reference prior for β_0 and distances priors for other coefficients

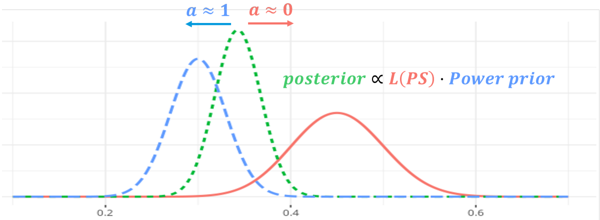
Informative priors: integrating PS and NPS data

Power prior (Ibrahim et al., 2000):

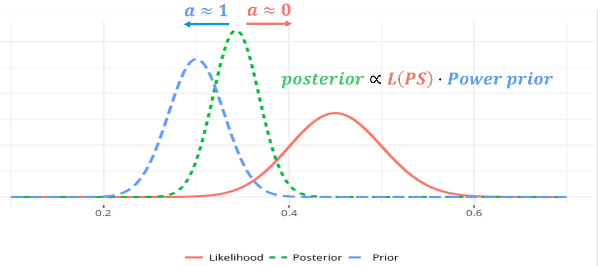
$$\pi(\boldsymbol{\beta}, a|D_{NP}) \propto L(\boldsymbol{\beta}|D_{NP})^{a}\pi_0(\boldsymbol{\beta})$$

and the posterior is:

$$\pi(\boldsymbol{\beta}|D_P, D_{NP}, a) \propto L(\boldsymbol{\beta}|D_P)L(\boldsymbol{\beta}|D_{NP})^{a}\pi_0(\boldsymbol{\beta})$$



Informative priors: Power prior

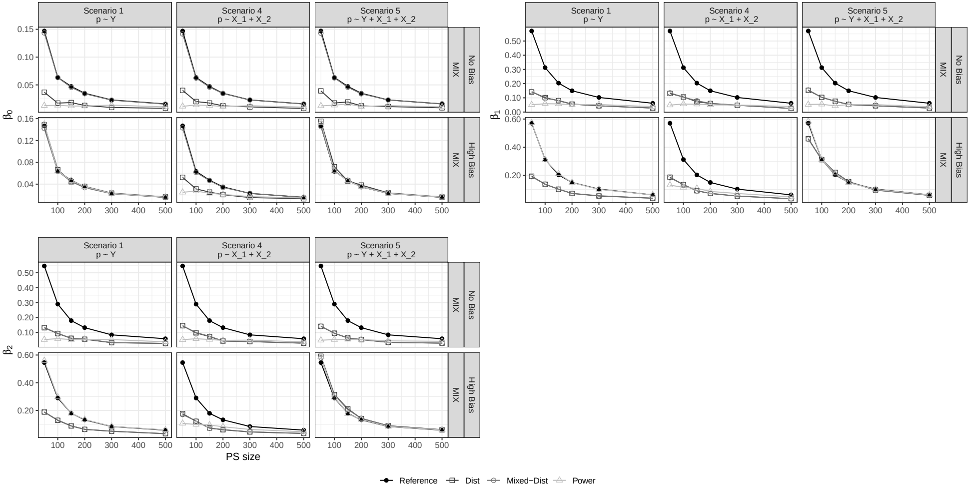


Influence of NPS data on the PS data is given by $0 \leq a \leq 1$:

- $a = 0$ - no borrowing
- $a = 1$ - full borrowing

We set a equal to the p-value resulting from Hotelling's T test for the difference between two vectors, $\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_{NP}$. We set the prior $\pi_0(\boldsymbol{\beta})$ as the reference prior

Results: selected cases



Results

- **In general: INF priors reduce MSE** especially for PS smaller than 200 obs
- **Worst-case scenario:** INF priors perform similarly to PS-only prior
- **Reduction in MSE** is driven by a **reduction** in the **variability**
- **High SB**: no substantial improvements in MSE
- **Best prior in MAR case**: Power Prior → **no a priori scenario knowledge**
- **Best prior overall**: Mixed-Distance

Application: the Data

PS data - American Trends Panel (ATP)

- Pew Research Center's nationally representative online survey panel
- Sample size: 3000 units $\rightarrow PS \in (50, 100, 150, 200, 500)$

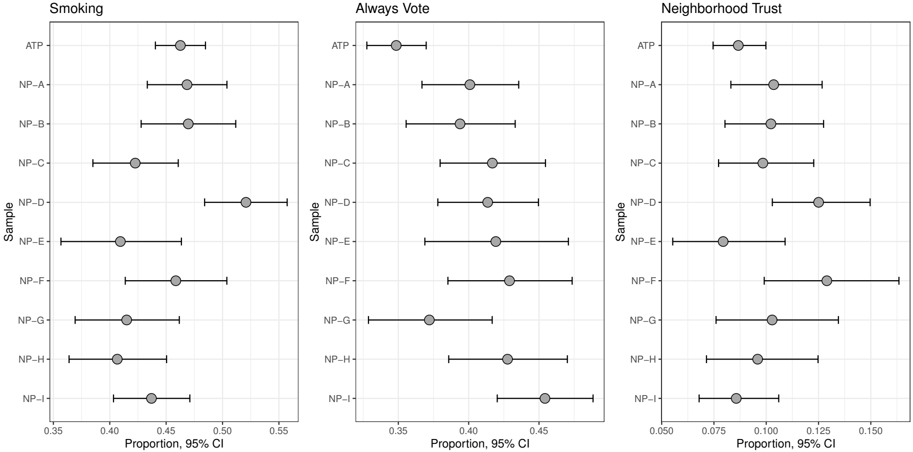
NPS data - 9 parallel online NPS from different vendors

- Vendors implemented quota sampling with different quota variables (demographic vs webographic)
- Sample size of about 1000 respondents

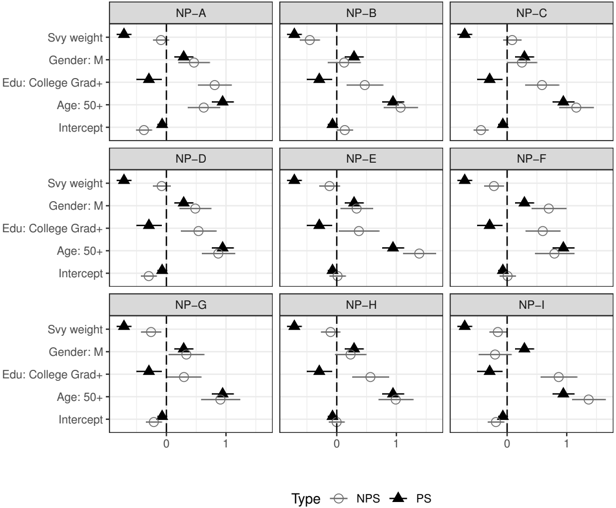
Outcome variables: Smoking, Always vote, Neighborhood Trust, Neighborhood Safety, Healthcare coverage, Volunteering

Covariates: Age, gender, education, survey weight

Comparing proportions



Comparing coefficients: an example with Always vote



Results

Similarly to the simulation study:

- **INF priors reduce MSE** and in the worst-case the perform similarly to PS-ONLY prior
- Reduction in MSE is driven by a **reduction** in the **variability**
- Results vary according to which NPS survey is used
- For low bias (neighborhood trust, healthcare coverage), all priors perform well

Largest reductions in MSEs:

- **Power prior** for very small PS sizes (**50-100** observations)
- **Distance-log prior** (and its mixed version) for sample sizes up to **200** observations

Interactive Cost Analysis: Shiny App

Three steps:

- We **assume** PS and NPS **costs**
- **Estimate** the expected cost of fielding a PS-only survey with the control prior that would achieve the same MSE as fielding parallel PS and NPS surveys with informative priors
- **Compare** it to the cost of fielding the parallel surveys

Interactive Cost Analysis

Take-aways:

- PS costs at least 3 times larger than NPS costs: best performing INF priors yield significant cost savings $\approx 70\%$
- PS costs twice NPS costs: cost savings are marginal or negative

Interactive Analysis: Shiny App

Main contributions

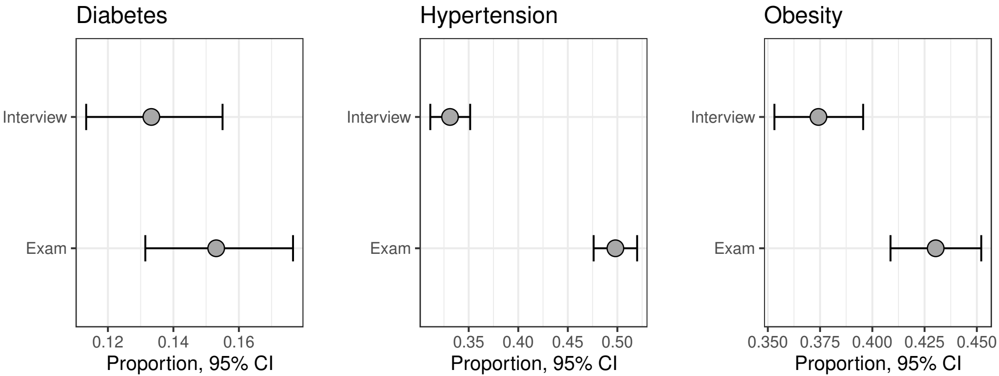
- Survey researchers face **budgetary** and **time constraints** $\rightarrow$ fielding large size PS is difficult
- **Small PS** yield **large variances** for survey estimates
- Our approach offers a **practical solution** to improve analytic inference (reduced variances and MSEs) while lowering survey costs
- **Shiny App:** facilitate researchers interested in designing and integrating parallel PS and NPS

Work in progress

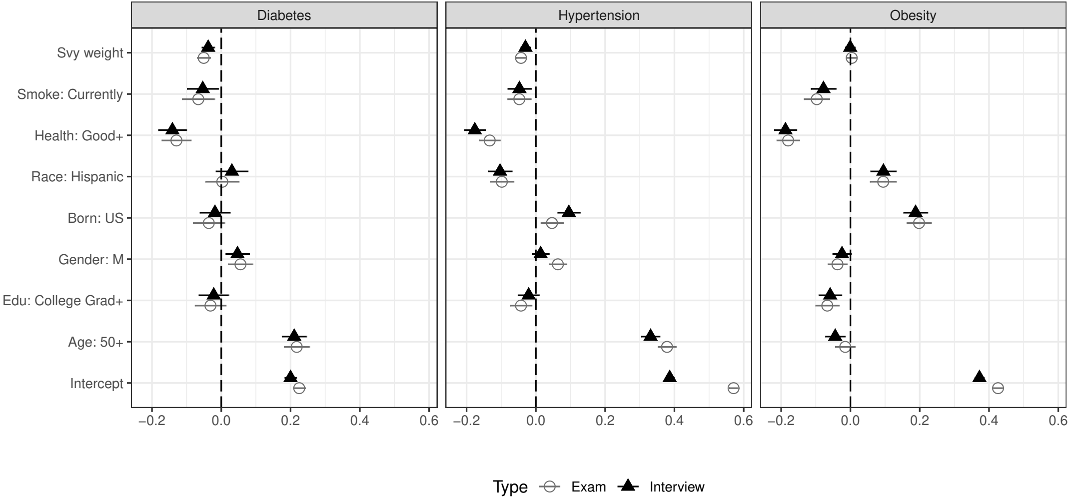
Integration of objective and subjective health measurements for survey research

- **Objective health measurements (PS):**
  - Gold standard in health survey research
  - Expensive and challenging to administer to large samples
- **Subjective health measurements (PS):**
  - Prone to misreporting
  - Less expensive and simpler to collect
- **Our aim:** Apply the Bayesian survey integration framework to this case where the difference is only in measurement (respondent from the same survey sample)

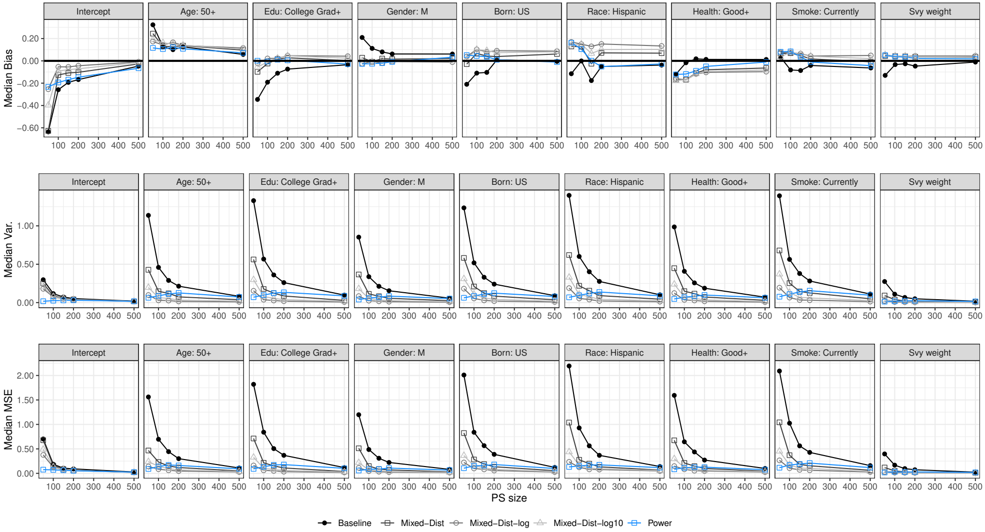
Comparing proportions



Comparing coefficients



Results: an example with Diabetes



References I

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

Ibrahim, J. G., Chen, M.-H., et al. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1):46–60.

Kruschke, J. (2014). Doing bayesian data analysis: A tutorial with r, jags, and stan.

Sakshaug, J. W., Wiśniowski, A., Ruiz, D. A. P., and Blom, A. G. (2019). Supplementing small probability samples with nonprobability samples: A bayesian approach. *Journal of Official Statistics*, 35(3):653–681.

References II

Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., and Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8(1):120–147.

Appendix

Simulation Framework

Details

- The **population** is generated from a logistic model with two binary predictors $x_1 \sim Ber(0.5)$, $x_2 \sim Ber(0.5)$
- Different values for the coefficients in order to test the **stability of the results**: $\beta_{NEG} \in (0.5, -1.3, -0.9)$, $\beta_{MIX} \in (0.5, -1.3, 0.9)$, $\beta_{POS} \in (0.5, 1.3, 0.9)$
- **Five selection mechanisms**, both **MAR** and **NMAR** scenarios with different selection probabilities and selection variables
- $NPS \in (1'000, 5'000)$ and $PS \in (50, 100, 150, 200, 500, 750, 1'000)$
- Several **informative** (INF) priors
- We compare the median MSE over 100 repetitions obtained using INF priors against the reference one

Simulation Framework Back

Five selection mechanism where the probability of participation p:

1. depends on Y (NMAR)
2. depends on Y and X_1 (NMAR)
3. depends on Y and X_2 (NMAR)
4. depends on X_1 and X_2 (MAR)
5. depends on Y, X_1 and X_2 (NMAR)

To introduce bias we consider different values of p for specific subgroups defined by the value of the selection variables:

$$p = \begin{cases} \{0.10, 0.20, 0.50, 0.90\} & \textit{if the value of selection variable(s) is 1} \\ 0.10 & \textit{otherwise} \end{cases}$$

Then, the probability of participation p is used to generate the participation indicator $P_i \sim Ber(p_i)$ for $i \in \{1, \ldots, N\}$ for each individual in the population.

Evaluation

- For each case, we repeat the simulation **100 times** using R and Stan
- We compute the **MSE**

$$\begin{aligned} MSE(\pi(\beta|Y,X)) &= Bias^2(\pi(\beta|Y,X)) + Var(\pi(\beta|Y,X)) \\ &= [\bar{\pi}(\beta|Y,X) - \beta^*]^2 + Var(\pi(\beta|Y,X)) \end{aligned}$$

where $\bar{\pi}(\beta|Y,X)$ is the mean of the posterior distribution for a given coefficient and $Var(\pi(\beta|Y,X)$ is the posterior variance

- We take the **median** over the 100 repetitions and we **compare** the values obtained using INF and PS-ONLY priors

Priors

- **Distance-log**: potentially more bias but lower posterior variance

$$\beta_j \sim \mathcal{N}\left(\hat{\beta}_{j_{NP}}, \sqrt{\frac{1}{log(n_{NP})} \cdot \max\left((\hat{\beta}_{j_P} - \hat{\beta}_{j_{NP}})^2, \hat{\sigma}^2_{\beta_{j_{NP}}}\right)}\right) \quad \text{for} \quad j = 0, 1, 2$$

- **Distance-log10**: potentially more bias but lower posterior variance

$$\beta_j \sim \mathcal{N}\left(\hat{\beta}_{j_{NP}}, \sqrt{\frac{1}{log_{10}(n_{NP})} \cdot \max\left((\hat{\beta}_{j_P} - \hat{\beta}_{j_{NP}})^2, \hat{\sigma}_{\beta}{}^2_{jNP}\right)}\right) \quad \text{for} \quad j = 0, 1, 2$$

- **Mixed-priors**: GJPY prior for β_0 and distance priors for other coefficients