Combining probability and nonprobability samples on an aggregated Level

Ton de Waal

Summer school 2023

Combining probability and nonprobability samples

- Based on work by former master student Sofia Villalobos-Aliste (and my colleague Sander Scholtus)

Nonprobability samples

- Nonprobability samples are increasingly popular due to their convenience and low costs
- Unfortunately, nonprobability samples are often selective and estimators based on such data are generally biased

Situation

- Estimates for proportions of a categorical target variable y per category of a categorical background variable x are available from
  - relatively small probability sample PS (size $n^{(P)}$), and
  - large nonprobability sample NPS (size $n^{(NP)}$)
- Estimator based on PS is usually unbiased, but its sampling variance is usually large
- Estimator based on NPS is likely to be biased, but its variance is generally small

Approaches

- Tailor-made approaches
- Mass imputation
  - Impute all population units
- Bayesian approach where information from NPS is used to construct prior
  - Prior based on NPS and data from PS are used to calculate posterior distribution
- Frequentist approach where estimates from NPS and PS are combined by weighting them

Bayesian approach: model setup

- Assume model for target variable, for instance multivariate normal model

$$y \sim N(\boldsymbol{\beta x}, \sigma^2)$$

  where $\boldsymbol{x} = \left(x_1, \dots, x_p\right)$
- Assume model for response R (depending on $\boldsymbol{x}$ and $\boldsymbol{\beta}$)

Bayesian approach: general prior

- Conjugate prior distribution for β_j $(j = 1, \ldots, p)$

$$\beta_j \sim N\left(\beta_{j0}, \sigma_{j0}^2\right)$$

β_{j0} and σ_{j0}^2 are hyper parameters

Bayesian approach: prior 1

- Weakly or non-informative prior

$$\beta_{j0} = 0$$
$$\sigma^2_{\beta_{j0}} = C$$

  where C is large number
- Posterior distribution basically only depends on data from probability sample

Bayesian approach: prior 2

- Based probability and nonprobability sample

$$\beta_j \sim N\left[\hat{\beta}_j^{NP}, \left(\hat{\beta}_j^{P} - \hat{\beta}_j^{NP}\right)^2\right]$$

  - $\hat{\beta}_j^{P}$ and $\hat{\beta}_j^{NP}$: ordinary least estimates from probability and nonprobability sample, respectively

Bayesian approach: prior 3

- Based on nonprobability sample

$$\beta_j \sim N\left[\hat{\beta}_j^{NP}, \left(\sigma_{\beta_{j0}}^{BNP}\right)^2\right]$$

  - $\hat{\beta}_j^{NP}$: ordinary least estimate from non-probability sample
  - Nonprobability data are bootstrapped to produce uncertainty measure
  - $\sigma_{\beta_{j0}}^{BNP}$ is bootstrap standard deviation of $\hat{\beta}_j^{NP}$

Approaches

- Tailor-made approaches
- Mass imputation
  - Impute all population units
- Bayesian approach where information from NPS is used to construct prior
  - Prior based on NPS and data from PS are used to calculate posterior distribution
- Frequentist approach where estimates from NPS and PS are combined by weighting them

Our approach

- We propose a method that combines estimates from probability and nonprobability sample on aggregated level
- Our method does not require any unit level data

Notation and assumptions

- Categories of target variable y are denoted as c $(c =$

Combined estimator

- We construct combined estimator of the form
$$\widehat{D}_{kc} = W_{kc}\hat{Z}_{kc}^{(P)} + (1 - W_{kc})\hat{Z}_{kc}^{(NP)}$$
where W_{kc} is weight between zero and one.
- If Mean Square Errors (MSEs) for $\hat{Z}_{kc}^{(P)}$ and $\hat{Z}_{kc}^{(NP)}$ were known, we could find weight W_{kc} for which MSE of $\widehat{D}_{kc}$ is minimum
- Optimal weight would be given by
$$W_{kc} = \mathrm{MSE}\left(\hat{Z}_{kc}^{(NP)}\right) \Big/ \left(\mathrm{MSE}\left(\hat{Z}_{kc}^{(P)}\right) + \mathrm{MSE}\left(\hat{Z}_{kc}^{(NP)}\right)\right)$$

Combined estimator

- Elliott & Haviland (2007) also use a combined estimator of the form

$$\widehat{D}_{kc} = W_{kc}\hat{Z}_{kc}^{(P)} + (1 - W_{kc})\hat{Z}_{kc}^{(NP)}$$

  where W_{kc} is weight between zero and one.
- They assumed that bias of $\hat{Z}_{kc}^{(NP)}$ is known
- We estimate bias of $\hat{Z}_{kc}^{(NP)}$ from PS and NPS

Estimating MSEs

- MSE = Variance + Bias^2
- Variance estimate of $\hat{Z}_{kc}^{(P)}$ is $\hat{Z}_{kc}^{(P)}\left(1-\hat{Z}_{kc}^{(P)}\right)/\left(n^{(P)}-1\right)$
- Estimator $\hat{Z}_{kc}^{(P)}$ is unbiased
- Design-based sampling variance of $\hat{Z}_{kc}^{(NP)}$ is unknown/undefined, but – assuming that nonprobability sample is large – reasonable estimate might be

$$\hat{Z}_{kc}^{(NP)}\left(1-\hat{Z}_{kc}^{(NP)}\right)/\left(n^{(NP)}-1\right)$$

- The big problem: bias of $\hat{Z}_{kc}^{(NP)}$ cannot be estimated from nonprobability sample only

Estimating MSEs: the plan

- We assume simple model for bias of $\hat{Z}_{kc}^{(NP)}$
- This allows us to estimate expected MSEs (EMSEs) under this model
- We then use

$$\hat{D}_{kc} = W_{kc}\hat{Z}_{kc}^{(P)} + (1 - W_{kc})\hat{Z}_{kc}^{(NP)}$$

with

$$W_{kc} = \frac{\text{EMSE}\left(\hat{Z}_{kc}^{(NP)}\right)}{\text{EMSE}\left(\hat{Z}_{kc}^{(P)}\right) + \text{EMSE}\left(\hat{Z}_{kc}^{(NP)}\right)}$$

Estimating MSEs: assumptions (main model)

- We introduce $b_{kc} = \mathrm{E}_d\left(\hat{Z}_{kc}^{(NP)}\right) - Z_{kc}$, where E_d denotes expectation under (unknown) “sampling design” of nonprobability sample
- Note that within each domain k we have $\sum_{c=1}^{C} b_{kc} = 0$ since estimated proportions in each domain add up to one

Estimating MSEs: assumptions (main model)

- We assume model such that b_{kc} is distributed as random variable with
  - $\mathrm{E}_b(b_{kc}) = \beta_c$, i.e. expected bias in category c is assumed to be constant across domains, with $\sum_{c=1}^{C} \beta_c = 0$
  - $\mathrm{Var}_b(b_{kc}) = \mathrm{E}_b\left((b_{kc} - \beta_c)^2\right) = \sigma^2$

Estimating MSEs: simple model

- In our study we also studied simpler model where we assumed that b_{kc} is distributed as random variable with
  - $\mathrm{E}_b(b_{kc}) = 0$
  - $\mathrm{Var}_b(b_{kc}) = \sigma^2$

Flavour of computations

- We define $\tilde{Z}_{kc} = \mathrm{E}_d\left(\hat{Z}_{kc}^{(NP)}\right)$, so $b_{kc} = \tilde{Z}_{kc} - Z_{kc}$
- We find $\mathrm{E}_b(Z_{kc}) = \mathrm{E}_b\left(\tilde{Z}_{kc} - b_{kc}\right) = \mathrm{E}_b\left(\tilde{Z}_{kc}\right)$
- $\mathrm{E}_b\left(Z_{kc}^2\right) = \mathrm{E}_b\left[\left(\tilde{Z}_{kc} - b_{kc}\right)^2\right] =$
  $\mathrm{E}_b\left(\tilde{Z}_{kc}^2\right) + \mathrm{E}_b\left(b_{kc}^2\right) - 2\mathrm{E}_b\left(\tilde{Z}_{kc}b_{kc}\right) = \mathrm{E}_b\left(\tilde{Z}_{kc}^2\right) + \sigma^2$
  where we assume that b_{kc} is not correlated to $\tilde{Z}_{kc}$
- So, $\mathrm{E}_b[Z_{kc}(1 - Z_{kc})] = \mathrm{E}_b\left[\tilde{Z}_{kc}\left(1 - \tilde{Z}_{kc}\right)\right] - \sigma^2$

Flavour of computations

- Note that we could estimate $\mathrm{E}_b[Z_{kc}(1 - Z_{kc})]$ by means of $\hat{Z}_{kc}^{(P)}\left(1 - \hat{Z}_{kc}^{(P)}\right)$
- However, since size of PS is rather small, this is likely to be an inaccurate estimate
- We therefore base our estimate for $E_b[Z_{kc}(1 - Z_{kc})]$ on NPS in combination with model for b_{kc}

Expressions for EMSEs

- $\mathrm{EMSE}\left(\hat{Z}_{kc}^{(NP)}\right) = \beta_c^2 + \sigma^2 + \frac{v_{kc}}{n_k^{(NP)}-1}$
- $\mathrm{EMSE}\left(\hat{Z}_{kc}^{(P)}\right) =$
$$\frac{1}{n_k^{(P)}}\left(\frac{n_k^{(NP)}}{n_k^{(NP)}-1} v_{kc} + \beta_c\left[2\mathrm{E}_b\mathrm{E}_d\left(\hat{Z}_{kc}^{(NP)}\right) - 1\right] - \beta_c^2 - \sigma^2\right)$$
- $n_k^{(P)}$ and $n_k^{(NP)}$ are sizes of PS, respectively NPS in domain k and

$$v_{kc} = \mathrm{E}_b\mathrm{E}_d\left(\hat{Z}_{kc}^{(NP)}\left(1 - \hat{Z}_{kc}^{(NP)}\right)\right)$$

Estimating EMSEs

- Unbiased estimator for v_{kc} is $\hat{Z}_{kc}^{(NP)}\left(1-\hat{Z}_{kc}^{(NP)}\right)$
- $\mathrm{E}_b\mathrm{E}_d\left(\hat{Z}_{kc}^{(NP)}\right)$ is estimated by $\hat{Z}_{kc}^{(NP)}$
- Ordinary least squares estimate for β_c is
  - $\hat{\beta}_c = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{Z}_{kc}^{(NP)} - \hat{Z}_{kc}^{(P)}\right)$
- Ordinary least squares estimate for σ^2 is
  - $\hat{\sigma}^2 = \frac{1}{(K-1)C}\sum_{k=1}^{K}\sum_{c=1}^{C}\left(\hat{Z}_{kc}^{(NP)} - \hat{Z}_{kc}^{(P)}\right)^2 - \frac{K}{(K-1)C}\sum_{c=1}^{C}\hat{\beta}_c^2$
- This leads to estimates $\widehat{\mathrm{EMSE}}\left(\hat{Z}_{kc}^{(P)}\right)$ and $\widehat{\mathrm{EMSE}}\left(\hat{Z}_{kc}^{(NP)}\right)$
- In turn this leads to W_{kc} and hence to our combined estimator

Simulation study (part 1)

- We simulated population of 100,000 units and repeatedly drew two datasets (PS and NPS), and applied our estimator
- We considered
  - one, four, ten, and 15 domains
  - three, five, eight, and 15 categories
  - a first scenario where all categories are of equal size in each domain, and a second scenario where categories have unequal sizes

Simulation study (part 2)

- We also considered
  - sample sizes per domain for probability sample $n^{(P)} \in \{10, 100, 400, 900\}$
  - sample sizes per domain for nonprobability sample $n^{(NP)} \in \{100, 1000, 2000, 6000\}$
  - two levels of selectivity for nonprobability sample: weak selectivity and severe selectivity
- We used full factorial design (1024 scenarios)
  - we drew R = 1000 simulations for each scenario
  - for each simulation we calculated $\hat{Z}_{kc}^{(P)}$, $\hat{Z}_{kc}^{(NP)}$ and $\widehat{D}_{kc}$

Evaluation criteria

- We compute root mean squared error (RMSE) per domain k and category c over the R simulations
  - $RMSE_{kc} = \sqrt{\sum_{r=1}^{R}\left(Z_{kc} - \hat{Q}_{kc,r}\right)^2 / R}$

  with $\hat{Q}_{kc,r}$ estimate for domain k and category c in simulation r ($\hat{Q}_{kc,r}$ is $\hat{Z}_{kc}^{(P)}$, $\hat{Z}_{kc}^{(NP)}$ or $\widehat{D}_{kc}$)
- We compute $ARMSE_k = \sum_{c=1}^{C} RMSE_{kc}/C$
- Finally, we compute $MARMSE = \sum_{k=1}^{K} ARMSE_k/K$
- We also assess bias of $\hat{Q}_{kc,r}$ ($\hat{Q}_{kc,r}$ is $\hat{Z}_{kc,r}^{(NP)}$ and $\widehat{D}_{kc,r}$) by means of $MAB = \sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{c=1}^{C}\left|Z_{kc} - \hat{Q}_{kc,r}\right|/RKC$

Results: is combined estimator better?

Selectivity	Size of categories	better than $\hat{Z}_{kc}^{(P)}$	better than $\hat{Z}_{kc}^{(NP)}$	better than both
Weak	Equal	94	64	58
	Unequal	92	64	57
Severe	Equal	83	84	67
	Unequal	80	85	65

Percentage of times out of 256 simulation conditions (differing with respect to numbers of domains and categories, and sample sizes) that combined estimator outperforms direct estimators in terms of MARMSE

Results: differences in MARMSE

Selectivity	Size of categories	Difference C–PS	Difference C–NPS
Weak	Equal	-0.0128	-0.0035
	Unequal	-0.0125	-0.0037
Severe	Equal	-0.0079	-0.0546
	Unequal	-0.0073	-0.0555

Average difference between MARMSE of the combined estimator (C) and direct estimators for probability (PS) and nonprobability (NPS) sample

Results

Selectivity	Size of categories	Bias reduced
Weak	Equal	99
	Unequal	98
Severe	Equal	100
	Unequal	100

Proportion (× 100%) of combined estimators with lower MAB than direct estimator for nonprobability sample

Results: is combined estimator better?

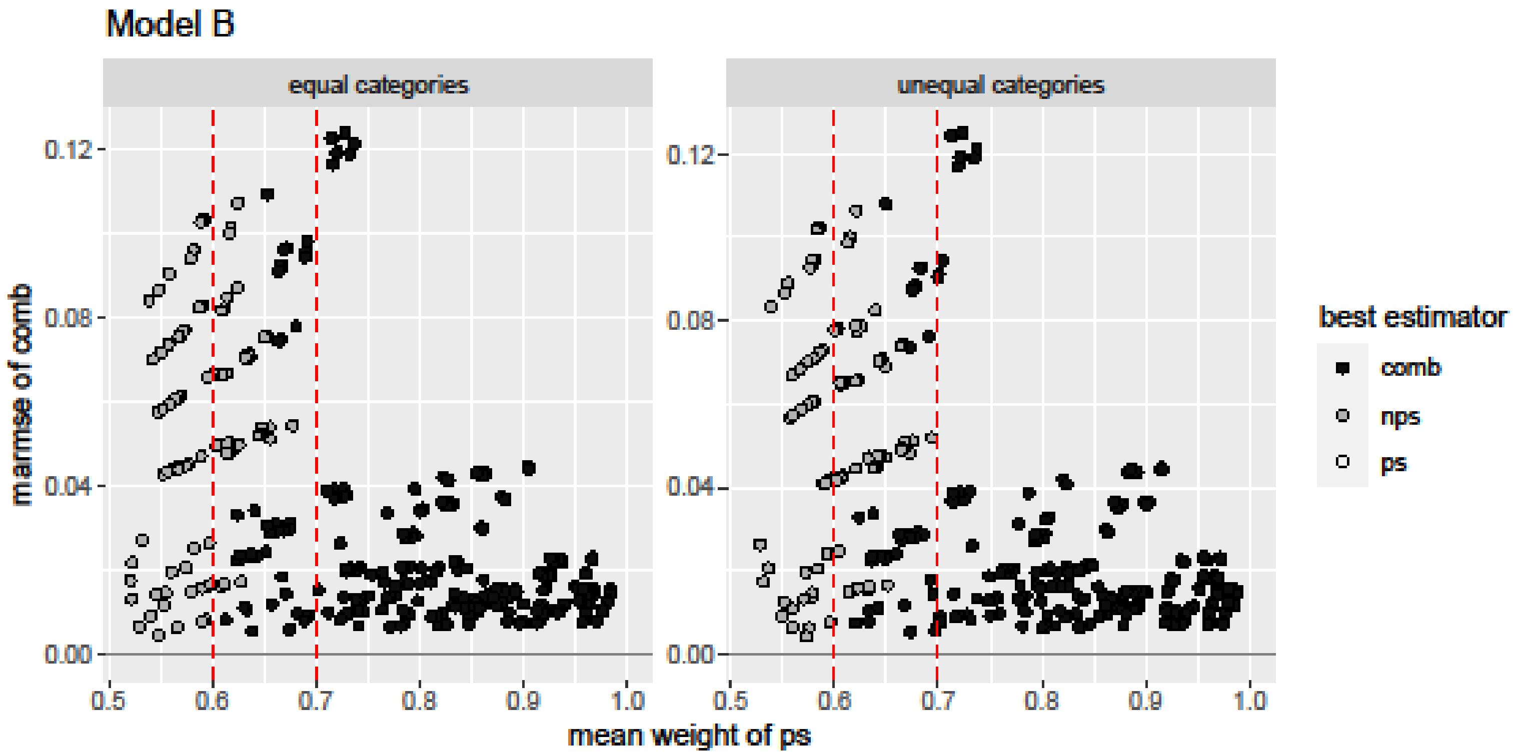
C	3				5			
$n^{(P)}$								
10	0.12	0.12	0.12	0.09	0.09	0.09	0.09	0.07
100	0.04	0.04	0.04	0.03	0.04	0.04	0.04	0.03
400	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01
900	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
$n^{(NP)}$	100	100	1000	6000	100	100	1000	6000

MARMSE of equal-size categories and severe selectivity

Results: relation with weights

- In our simulation we found that
  - when $W_{kc} \geq 0.7$, combined estimator always had lowest MARMSE of the three estimators
  - When $W_{kc} \leq 0.6$, estimator based on nonprobability sample always performed best
  - When $0.6 < W_{kc} < 0.7$, it depended on specific scenario which of the three estimators performed best

Relation between weight and MARMSE



Conclusions

- Advantage of method is that it is not necessary to link two samples at level of individual observations
- It is also not important whether NPS and PS overlap or not
- Proposed method is quite robust
  - EMSE of combined estimator is always less than or equal to lowest EMSE of estimators for the two samples
  - Actual MSE of combined estimator is never higher than highest MSE of estimators for the two samples
- Proposed method is very easy to implement in R

Possible extensions

- Current version of method is only suitable for PS that is drawn by means of simple random sampling: this could be extended to other sampling designs
- If microdata are available, one might consider correcting for selection error in NPS first and then apply method for combining estimates

When can we apply method?

- Hard to say
  - One needs reasonable estimate of variance of NPS
  - Model for bias needs to be reasonable
    - Weights seem to provide useful information about the latter

References

- Tailor-made approaches

  - Kuijvenhoven, L. & S. Scholtus (2010), *Estimating Accuracy for Statistics based on Register and Survey Data*. CBS discussion paper, https://www.cbs.nl/nl-nl/achtergrond/2010/11/estimating-accuracy-for-statistics-based-on-register-and-survey-data.

- Mass imputation

  - Kim, J.K., S. Park, Y. Chen & C. Wu (2021), Combining Non-Probability and Probability Survey Samples Through Mass Imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society* **184**(3), 941-963, https://doi.org/10.1111/rssa.12696.

References

- Bayesian approach
  - Sakshaug, J.W., A. Wiśniowski, D.A. Perez Ruiz & A.G. Blom (2019), Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach. *Journal of Official Statistics* **35**(3), 653-681, http://dx.doi.org/10.2478/JOS-2019-0027.
  - Wiśniowski, A., J.W. Sakshaug, D.A. Perez Ruiz & A.G. Blom (2020), Integrating Probability and Nonprobability Samples for Survey Inference. *Journal of Survey Statistics and Methodology* **8**, 120-147.

References

- Frequentist approach

  - Elliott, M.N. & A. Haviland (2007), Use of a Webbased Convenience Sample to Supplement a Probability Sample. *Survey Methodology* **33**(2), 211-215.
  - Villalobos Aliste, S. (2022), *Combining Probability and Non-Probability Samples for Estimation*. Master thesis, Utrecht University.
