Inferences to a voluntary sample in a household survey

Soonpil Kwon[1,2]

Heeyoung Jung[1], Youngmi Kwon[1]

[1]Statistics Korea, [2]University of Seoul

Speed Talk Session 3 5 June, 2024

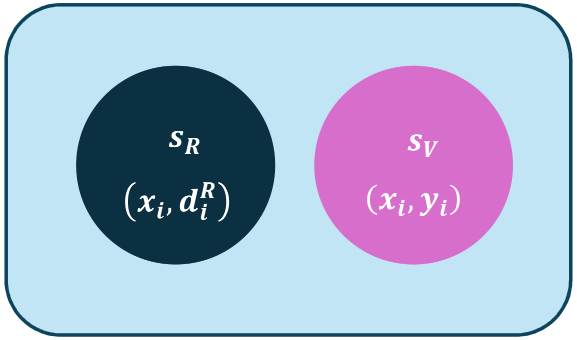
Instituto Nacional de Estatística
Statistics Portugal
eurostat
The conference is partly financed by the European Union

Objective

To estimate the population mean using the variable of interest obtained from a non-probability sample. → Reducing the selection bias of non-probability sample.

Assumptions

1. Existence of probability and non-probability samples represent the same population.
2. Variable of interest is only in the non-probability sample.
3. Two samples share useful covariate (auxiliary) variables.
4. Two samples are independent and don't have measurement error.



The conference is partly financed by the European Union

Notation

$U = \{1,2, \dots N\}$: the set of N units for the finite population

y_i : interest variable, $\boldsymbol{x}_i$: auxiliary variables, $i = 1,2, \cdots, N$

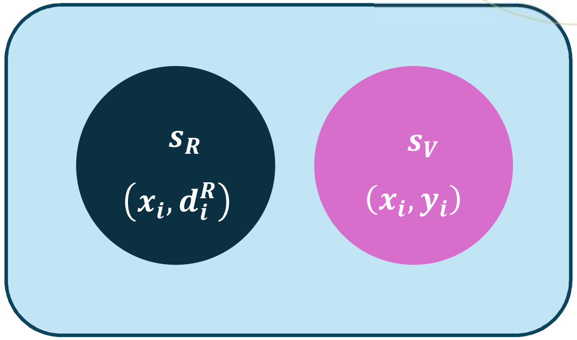
$\mu_y = N^{-1} \sum_{i=1}^{N} y_i$: population mean

$\boldsymbol{s_V}$: the size of n_V non-probability sample with $\{(\boldsymbol{x_i}, \boldsymbol{y_i}), \boldsymbol{i} \in \boldsymbol{s_V}\}$

$\boldsymbol{s_R}$: the size of n_R probability sample with $\{(\boldsymbol{x_i}, \boldsymbol{d_i^R}), \boldsymbol{i} \in \boldsymbol{s_R}\}$

$d_i^R = 1/\pi_i^R$, where π_i^R is ith unit's inclusion probability, $\boldsymbol{i} \in \boldsymbol{s_R}$

δ_i : indicator variable for unit i, $\begin{cases} \boldsymbol{\delta_i = 1, \ if \ i \in s_V} \\ \boldsymbol{\delta_i = 0, \ if \ i \notin s_V} \end{cases}$, $i = 1,2, \dots, N$



pop.

2021 survey of Household Finances and Living conditions (N=18,000HH)

pop.mean μ (household income per month)

Poisson sampling →

Non-prob. sample s_V (n_V)

	x			y	d^V
	age	gender	# of people in HH	HH income (₩1,000)	(wgt)
	30	F	1	2,500	?
	35	M	2	4,000	
	43	M	4	6,000	
	⋮	⋮	⋮	⋮	

SRS →

prob. sample s_R (n_R)

	x			y	d^R
	age	gen Der	# of people in HH	HH income (month)	(wgt)
	35	M	1	?	20
	45	F	3		18
	58	M	4		19
	⋮	⋮	⋮		⋮

① Estimate $\hat{d}_i^V$ the weights of a s_V by referencing the weights of a s_R

→ Estimate $\hat{\mu}_V = \frac{\sum \hat{w}_i y_i}{\sum \hat{w}_i}$

③ Combination ① & ②

→ Estimate $\hat{\mu}_{DR}$

② Estimate the $\hat{y}$ of the s_R by referencing the relationship between x_i and y_i in s_V

→ Estimate $\hat{\mu}_R = \frac{\sum w_i \hat{y}_i}{\sum w_i}$

Estimators of mean

naïve est. — $\hat{\mu}_{naive} = n_V^{-1} \sum_{i \in s_V} y_i$ — O Assume the s_V is SRS

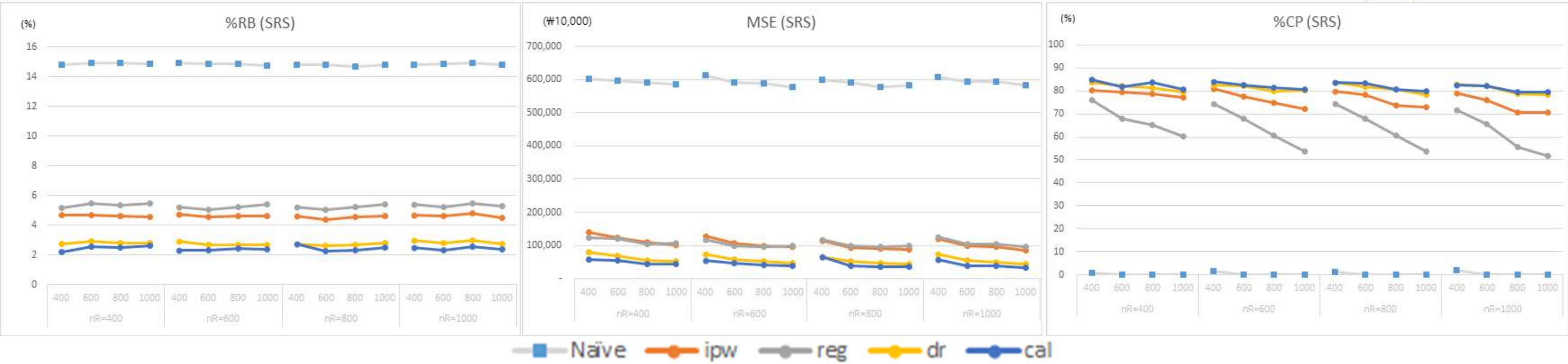
ipw est. — $\hat{\mu}_{ipw} = \hat{N}_V^{-1} \sum_{i \in s_V} \hat{d}_i^V y_i$ — ① Estimate propensity score $\hat{\pi}_i^V$ using logistic regression model → $\hat{d}_i^V = 1/\hat{\pi}_i^V$

reg est. — $\hat{\mu}_{reg} = \hat{N}_R^{-1} \sum_{i \in s_R} d_i^R \hat{y}_i$ — ② Estimate the $\hat{y}_i$ of the regression model in s_V

dr est. — $\hat{\mu}_{dr} = \hat{N}_V^{-1} \sum_{i \in s_V} \hat{d}_i^V \{y_i - m(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})\} + \hat{\mu}_{reg}$ — ③ Combination of ① & ②

cal est. — $\hat{\mu}_{cal} = \hat{N}_{V,cal}^{-1} \sum_{i \in s_V} \hat{d}_{i,cal}^V y_i$ — ① Estimate $\hat{d}_i^V$ using calibration (GREG)

		$n_R = 400$			$n_R = 600$			$n_R = 800$			$n_R = 1000$		
	est.	%RB	MSE	%CP	%RB	MSE	%CP	%RB	MSE	%CP	%RB	MSE	%CP
$n_V = 400$	$\hat{\mu}_{naive}$	14.79	602,644	1.0	14.95	613,127	1.6	14.81	599,321	1.4	14.83	607,148	1.9
	$\hat{\mu}_{ipw}$	4.67	139,610	80.2	4.75	129,690	81.1	4.62	115,825	80.0	4.70	120,514	79.1
	$\hat{\mu}_{reg}$	5.16	123,599	76.0	5.30	124,151	75.2	5.23	118,257	74.5	5.40	126,731	71.9
	$\hat{\mu}_{dr}$	2.74	79,440	83.5	2.92	74,791	82.5	2.74	67,398	83.7	2.95	74,539	82.9
	$\hat{\mu}_{cal}$	2.17	59,184	84.9	2.31	55,411	84.2	2.39	52,400	84.8	2.49	56,792	82.6
$n_V = 600$	$\hat{\mu}_{naive}$	14.92	597,689	0.1	14.85	591,982	0.1	14.83	589,879	-	14.87	594,942	-
	$\hat{\mu}_{ipw}$	4.69	122,992	79.5	4.58	106,804	77.6	4.35	93,838	78.5	4.62	99,294	75.9
	$\hat{\mu}_{reg}$	5.45	119,642	68.1	5.20	108,204	67.9	5.06	100,247	67.8	5.24	104,931	65.6
	$\hat{\mu}_{dr}$	2.93	68,488	82.1	2.69	58,659	82.4	2.60	52,069	81.9	2.77	54,818	82.2
	$\hat{\mu}_{cal}$	2.58	54,733	81.9	2.33	46,816	82.5	2.24	40,135	83.4	2.32	40,264	82.4
$n_V = 800$	$\hat{\mu}_{naive}$	14.93	592,182	-	14.88	588,438	-	14.72	577,205	-	14.96	594,295	-
	$\hat{\mu}_{ipw}$	4.60	108,707	78.5	4.61	99,814	74.8	4.53	91,215	73.7	4.79	96,241	70.7
	$\hat{\mu}_{reg}$	5.36	105,077	65.2	5.30	103,548	62.4	5.20	97,053	60.7	5.49	105,282	55.5
	$\hat{\mu}_{dr}$	2.77	54,963	81.5	2.68	51,479	80.0	2.67	47,609	80.7	2.96	50,808	78.8
	$\hat{\mu}_{cal}$	2.50	43,753	83.7	2.45	41,930	81.4	2.31	37,416	80.6	2.56	39,894	79.6
$n_V = 1000$	$\hat{\mu}_{naive}$	14.90	586,305	-	14.78	577,734	-	14.84	582,508	-	14.84	582,583	-
	$\hat{\mu}_{ipw}$	4.55	102,444	77.0	4.64	95,328	72.2	4.59	88,834	72.9	4.47	85,819	70.6
	$\hat{\mu}_{reg}$	5.44	106,097	60.4	5.28	97,674	55.1	5.40	99,206	53.6	5.30	95,420	51.7
	$\hat{\mu}_{dr}$	2.77	53,680	79.4	2.66	46,179	80.1	2.78	45,114	78.3	2.72	43,588	78.3
	$\hat{\mu}_{cal}$	2.64	45,399	80.7	2.39	37,894	80.7	2.47	36,289	80.0	2.34	33,236	79.5



- These estimators reduce the bias of non-probability samples.
- Treating a non-probability sample as a simple random sample can lead to a serious selection bias.

- The MSE decreases as the sample size increases.

- The Bigdata Paradox arises as the sample size increases, leading to a decrease in the probability that 95% confidence interval of the estimate including the population mean.

Additional notes

It’s important the good auxiliary variables.

- Highly explanatory auxiliary variable → reg, dr → single weight not available → not preferred by NSI
- Categorical auxiliary variable
  - → CAL estimator for existed data
  - → IPW estimator for survey designed data

Boostrap needs further development for variance estimation.

How to handle and interpret the remaining bias?

- If the bias can’t be completely eliminated, its utility can be assessed through trend analysis over time.
- Alternative methods are needed to quantify the risk of selection bias or non-coverage in bigdata or non-probability samples.

References

Castro-Martín et al. (2020). Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques, *Mathematics, June 2020, 8(6)*, 879.

Chen, Y., Li, P. and Wu, C. (2020). “Doubly robust inference with non-probability survey samples”, *Journal of the Americal Statistical Assocation* 115, 2011-2021.

Couper, M. P. (2013). “Is the sky falling? New technology changing media, and the future of surveys”. Surv. Res. Methods, 7, 145-156.

Deville, J.C. and Särndal, C.E. (1992). “Calibration estimators in survey sampling”, *Journal of the Americal Statistical Assocation 87*, 376-382.

Kim, J.K. (2022a),.“A gentle introduction to data integration in survey sampling”, *The Survey Statistician 85*, 19-29.

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

Thank You

pilsogood@korea.kr

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly financed by the European Union