Non-probability sample integration in the survey of Lithuanian census

Ieva Burakauskaitė, Statistics Lithuania, ieva.burakauskaite@stat.gov.lt

(joint work with *Andrius Čiginas, Statistics Lithuania, andrius.ciginas@stat.gov.lt*)

2022-04-28

Session 38

Outline

- **Objects of interest**
  - The use of administrative data in the Census 2021
  - Combination of voluntary and probability samples
  - Imputation of missing values: historical, deductive and k-nearest neighbors methods
  - Sampling design
- **Calibration (generalized regression) estimator**
  - Sampling weight calibration
  - Estimation of variance
- **Propensity scores**
  - Propensity score model
  - Estimation of propensity scores
- **Inverse probability weighted (IPW) estimator**
  - Estimation of asymptotic variance
- **Composite estimator**
- **Summary**
- **Literature**

Objects of interest

- **The Statistical survey on population by ethnicity, native language and religion 2021** aimed to evaluate population proportions of:
  - *religion professed* (16 categories),
  - *mother tongue* (more than 12 categories),
  - *knowledge of other languages* (16 languages),
  - *ethnicity* (mass imputation was used).
- Let us further consider **binary variables**, where y denotes one of the above mentioned categories of a corresponding variable with the fixed values $y_1, \dots, y_N$ in a finite census population of N units $\mathcal{U} = \{1, \dots, k, \dots, N\}$.

The use of administrative data in the Census 2021

- Variables of interest were completely observed in previous **population and housing censuses**.

  Based on the data of the last census carried out in 2011:
  - Population of Lithuania comprised people of ***154 ethnicities***;
  - ***One in three*** residents indicated that they spoke ***two foreign languages***;
  - The residents belonged to ***59 different religious communities***.
- The main part of **Census 2021** was based on **administrative data**.
- Additional variables were collected through **the Statistical survey on population by ethnicity, native language and religion 2021**.

Combination of voluntary and probability samples

Survey sample

(A) voluntary sample

- An **online survey** was carried out from 15 January to 28 February, 2021.
- Approximately **2%** of census population filled in the given questionnaire.

(B) not in the sampling frame

- After the end of the online survey, a sampling frame for probability sampling was constructed. It **excluded households if**: at least one individual from the household participated in the online survey, it was an institution, more than 15 individuals were its permanent residents, etc.

(C) probability sample

- Around **40 thousand households** were sampled from the Population Register.
- Approximately **6%** of census population was interviewed.

Imputation of missing values: *historical*, *deductive* and *k-nearest neighbors* methods

- Missing values in the sample were **historically filled in** using **information from censuses 2011 and 2001** consecutively, as variables of interest are fully known for populations of previous censuses.
- **Additional sociodemographic characteristics** of previous and current censuses (such as age, gender, marital status, household structure, country of birth, citizenship, education, employment status, etc.) were used for **deductive imputation**.
- The remaining missing values in the sample were then filled in using ***k*-nearest neighbors imputation**.

Sampling design

- Sampling frame was divided into $H = 113$ strata:

  municipality × area of residence (i.e., urban / rural).

- The sample $s \subset \mathcal{U}$ of size $n < N$ was drawn according to the sampling design $p(\cdot)$ with **inclusion into the sample probabilities** $\pi_k = \mathrm{P}_p\{k \in s\} > 0,\ k \in \mathcal{U}$:
  - Inclusion into the sample probability for unit k in stratum h equals to

$$\pi_k \approx \frac{m_k n'_h}{N'_h},$$

    where N'_h denotes the size of the hth stratum, n'_h is the number of households selected, m_k is the number of individuals in the corresponding household;
  - $\pi_k = 1$ for voluntary sample respondents and households not in the sampling frame.
- **The primary sampling weights** then equal to $d_k = 1/\pi_k$.

Statistics Poland IAOS IMPROVING OFFICIAL STATISTICS Statistical Office in Kraków

Calibration (generalized regression) estimator

- We aim to estimate the population proportion

$$\theta = \frac{1}{N} \sum_{k \in \mathcal{U}} y_k$$

for every binary variable y.

- **The generalized regression estimator** with calibrated weights w_k is used to evaluate the proportion θ:

$$\hat{\theta}^{GR} = \frac{1}{\hat{N}} \sum_{k \in s} w_k y_k \,,$$

where $\hat{N} = \sum_{k \in s} w_k$.

Sampling weight calibration

- Weights d_k, $k \in s$, are calibrated according to Deville and Särndal (1992):

$$\sum_{k \in s} \frac{(w_k - d_k)^2}{d_k} \to \min$$

subject to

$$\sum_{k \in s} w_k x_k^{(1)} = \sum_{k \in \mathcal{U}} x_k^{(1)}, \dots, \sum_{k \in s} w_k x_k^{(P)} = \sum_{k \in \mathcal{U}} x_k^{(P)}$$

for P auxiliary variables $x^{(1)}, \dots, x^{(P)}$ with values known for the entire population.

- In our case, $x_k^{(1)} = 1$, $k \in \mathcal{U}$. The rest auxiliary information includes binary variables on **age groups**, **gender** and **religions professed in 2011 intersected with counties**.

Estimation of variance

- **Variance of** $\hat{\theta}^{GR}$ is estimated according to Deville and Särndal (1992):

$$\hat{\psi}^{GR} = \frac{1}{\hat{N}^2} \sum_{k \in s} \sum_{l \in s} \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right) \frac{\left(y_k - \mathbf{x}_k' \hat{\mathbf{B}}\right)\left(y_l - \mathbf{x}_l' \hat{\mathbf{B}}\right)}{\pi_k \pi_l},$$

where

$$\hat{\mathbf{B}} = \left(\sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k}\right)^{-1} \left(\sum_{k \in s} \frac{\mathbf{x}_k y_k}{\pi_k}\right),$$

with $\mathbf{x}_k = \left(x_k^{(1)}, \dots, x_k^{(P)}\right)'$ and $\pi_{kl} = \mathrm{P}_p\{k, l \in s\} > 0$.

Table 1: Comparison of proportions of some additional sociodemographic characteristics in the voluntary sample vs. the whole population.

		Voluntary sample	Population
County	Vilnius	0.64	0.29
Ethnicity	Lithuanian	0.56	0.85
	Pole	0.35	0.07
Education	higher	0.48	0.20
	(lower) secondary	0.24	0.37
	primary	0.09	0.20
Employment	employed	0.63	0.45
Marital status	married	0.52	0.42
Age group	$\geq 30, < 50$	0.37	0.27
Gender	male	0.41	0.46

Table 2: Comparison of *religion* proportions in the voluntary sample vs. the whole population.

	Voluntary sample	Population	Difference in %
Karaites	0.00130	0.00009	1307
New Apostolic Church	0.00161	0.00014	1049
Evangelical Reformed Believers	0.00833	0.00207	302
Other	0.01596	0.00514	211
Pentecostalists	0.00198	0.00067	194
Greek Catholics (Uniats)	0.00048	0.00021	131
Evangelical Lutherans	0.01311	0.00585	124
Judaists	0.00074	0.00035	112
Baptists and Free Churches	0.00083	0.00048	74
Sunni Muslims	0.00130	0.00085	52
Not indicated	0.07621	0.10090	-24
Seventh-Day Adventist Church	0.00026	0.00032	-20
None	0.07580	0.06424	18
Old Believers	0.00615	0.00683	-10
Orthodox	0.04047	0.03787	7
Roman Catholics	0.75548	0.77398	-2

Propensity scores

- Consider a non-probability sample s_A consisting of n_A units from the finite census population $\mathcal{U}$. Let $R_k = \mathbb{I}(k \in s_A)$ be the indicator variable for unit $k \in \mathcal{U}$ being included in the sample s_A.
- The **propensity scores** (Rosenbaum and Rubin, 1983) are given by

$$\pi_k^A = \mathrm{E}_q(R_k|\mathbf{x}_k, y_k) = \mathrm{P}_q(R_k = 1|\mathbf{x}_k, y_k), \qquad k \in \mathcal{U},$$

where the subscript q refers to the model for the selection mechanism for the sample s_A – the propensity score model.

Propensity score model

- **Model assumptions**:
  1. The selection indicator R_k and the response variable y_k are independent given the set of covariates $\mathbf{x}_k$.
  2. All units have a nonzero propensity score: $\pi_k^A > 0$ for all $k \in \mathcal{U}$.
  3. The indicator variables R_k and R_l are independent given $\mathbf{x}_k$ and $\mathbf{x}_l$ for $k \neq l$.

- Propensity scores $\pi_k^A = \mathrm{P}_q(R_k = 1|\mathbf{x}_k)$ can be modelled **parametrically** as

$$\pi_k^A = \pi(\mathbf{x}_k, \boldsymbol{\theta}_0) = \frac{\exp(\mathbf{x}_k' \boldsymbol{\theta}_0)}{1 + \exp(\mathbf{x}_k' \boldsymbol{\theta}_0)},$$

where $\boldsymbol{\theta}_0$ is the true value of the unknown model parameters.

Estimation of propensity scores

- **The maximum likelihood estimator** for π_k^A is computed as $\boxed{\hat{\pi}_k^A = \pi\left(\mathbf{x}_k, \widehat{\boldsymbol{\theta}}\right),}$ where $\widehat{\boldsymbol{\theta}}$ maximizes the log-likelihood function

$$l(\boldsymbol{\theta}) = \sum_{k \in s_A} \log\left\{\frac{\pi(\mathbf{x}_k, \boldsymbol{\theta})}{1 - \pi(\mathbf{x}_k, \boldsymbol{\theta})}\right\} + \sum_{k \in \mathcal{U}} \log\{1 - \pi(\mathbf{x}_k, \boldsymbol{\theta})\}$$

$$= \sum_{k \in s_A} \mathbf{x}_k' \boldsymbol{\theta} - \sum_{k \in \mathcal{U}} \log\{1 + \exp(\mathbf{x}_k' \boldsymbol{\theta})\}.$$

- The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ can be obtained by solving the score equations

$$U(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) = \sum_{k \in \mathcal{U}} \{R_k - \pi(\mathbf{x}_k, \boldsymbol{\theta})\} \mathbf{x}_k = \mathbf{0}.$$

Inverse probability weighted (IPW) estimator

- The estimated propensity scores $\hat{\pi}_k^A = \pi(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$, $k \in s_A$, can be used to compute **the IPW estimator** for the proportion θ (Chen et al., 2020):

$$\hat{\theta}^{IPW} = \frac{1}{\hat{N}^A} \sum_{k \in s_A} \frac{y_k}{\hat{\pi}_k^A},$$

where $\hat{N}^A = \sum_{k \in s_A} 1/\hat{\pi}_k^A$.

Estimation of asymptotic variance

- Under certain regularity conditions and assuming the logistic regression model for the propensity scores, we have $\hat{\theta}^{IPW} - \theta = O_p\left(n_A^{-1/2}\right)$, and **asymptotic variance of** $\hat{\theta}^{IPW}$ can be derived as

$$\hat{V}^{IPW} = \frac{1}{\left(\hat{N}^A\right)^2} \sum_{k \in s_A} \left(1 - \hat{\pi}_k^A\right) \left(\frac{y_k - \hat{\theta}^{IPW}}{\hat{\pi}_k^A} - \hat{\mathbf{b}}' \mathbf{x}_k \right)^2 ,$$

where

$$\hat{\mathbf{b}}' = \left\{ \sum_{k \in s_A} \left(\frac{1}{\hat{\pi}_k^A} - 1 \right) \left(y_k - \hat{\theta}^{IPW} \right) \mathbf{x}_k' \right\} \left\{ \sum_{k \in \mathcal{U}} \hat{\pi}_k^A \left(1 - \hat{\pi}_k^A\right) \mathbf{x}_k \mathbf{x}_k' \right\}^{-1} .$$

Composite estimator

- **Estimates of population proportions** θ (e.g., *religion* proportions) equal to

$$\hat{\theta}^c = \hat{\lambda}\hat{\theta}^{GR} + \left(1 - \hat{\lambda}\right)\hat{\theta}^{IPW},$$

where $\hat{\lambda} = \hat{V}^{IPW}/\left\{\hat{\psi}^s + \hat{V}^{IPW}\right\}$, and $\hat{\psi}^s$ is a smoothed version of the variance $\hat{\psi}^{GR}$.

For the smoothing of variance, we assume that $\mathrm{var}_p\left(\hat{\theta}_1^{GR}\right) \approx K\tilde{N}^{\gamma}$, with $\tilde{N}$ as a size of 2011 *religion* in the population of Census 2021 (Dick, 1995). Parameters $K > 0$ and $\gamma \in \mathbb{R}$ are evaluated through regression with all categories of the variable of interest as auxiliary information.

- **Variance estimator for the composition** $\hat{\theta}^c$ is then set as

$$\hat{V}^c = \hat{\lambda}\hat{\psi}^s.$$

- Estimates $\hat{\theta}^c$ are **benchmarked** according to the variance estimates $\hat{V}^c$.

Table 3: *Religion* proportions in 2001, 2011 and 2021 Census populations.

	$\theta^{(2001)}$	$\theta^{(2011)}$	$\hat{\theta}^{GR}$	$\hat{\theta}^{c}$
Roman Catholics	0.78391	0.77233	*0.73664*	0.74191
Not indicated	0.05671	0.10112	*0.15701*	0.13665
None	0.09696	0.06146	*0.05408*	0.06113
Orthodox	0.04150	0.04113	*0.03433*	0.03747
Old Believers	0.00806	0.00767	*0.00434*	0.00647
Evangelical Lutherans	0.00565	0.00604	*0.00389*	0.00560
Other	0.00282	0.00493	*0.00566*	0.00546
Evangelical Reformed Believers	0.00208	0.00221	*0.00122*	0.00197
Pentecostalists	0.00037	0.00061	*0.00117*	0.00108
Sunni Muslims	0.00075	0.00089	*0.00058*	0.00077
Baptists and Free Churches	0.00034	0.00044	*0.00017*	0.00039
Judaists	0.00039	0.00040	*0.00025*	0.00032
Greek Catholics (Uniats)	0.00010	0.00023	*0.00030*	0.00028
Seventh-Day Adventist Church	0.00016	0.00030	*0.00014*	0.00026
New Apostolic Church	0.00012	0.00014	*0.00015*	0.00015
Karaites	0.00008	0.00010	*0.00008*	0.00009

Table 4: Comparison of variance of *religion* proportion estimates $\hat{\theta}^{GR}$ and $\hat{\theta}^{c}$ ($\hat{\psi}^{s}$ and $\hat{V}^{c}$ accordingly).

	$\hat{\psi}^{s} \times 10^{6}$	$\hat{V}^{c} \times 10^{6}$	Difference in %
Old Believers	0.0517	0.0381	36
Orthodox	0.3126	0.2413	30
Baptists and Free Churches	0.0032	0.0029	9
Sunni Muslims	0.0058	0.0055	6
Judaists	0.0023	0.0022	5
Seventh-Day Adventist Church	0.0021	0.0020	4
Karaites	0.0006	0.0005	4
Evangelical Lutherans	0.0440	0.0426	3
Greek Catholics (Uniats)	0.0013	0.0013	2
Evangelical Reformed Believers	0.0148	0.0145	2
Other	0.0383	0.0377	2
Pentecostalists	0.0045	0.0045	1
New Apostolic Church	0.0009	0.0009	0
None	0.5445	0.5445	0
Not indicated	0.8748	0.8748	0
Roman Catholics	7.4356	7.4356	0

Summary

- The main part of Lithuanian census 2021 was based on administrative data.

  Some variables of interest (i.e., religion, native language, knowledge of other languages) were estimated using both voluntary (non-probability) and probability samples.

- The inverse probability weighted estimator was used in order to properly integrate the non-probability sample, as the generalized regression estimator was not able to accurately estimate small proportions of interest.

Literature

Chen, Y., Li, P., Wu, C. (2020). Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association*, 115(532):1–25.

Deville, J. C., Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Dick, P. (1995). Modelling net undercoverage in the 1991 Canadian census. *Survey Methodology*, 21(1):45–54.

Rosenbaum, P. R., Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55.