Survey Samplings with R

European Indicators of poverty and vulnerabilities for Sustainable Development Goal and Seminars

JMC - SAMPIEU

Gaia Bertarelli[1]

[1]Department of Economics & Management, University of Pisa

18-19-22 October 2018

The presentation at a Glance

Outline of my lectures

- ★ Sample selection
- ★ Estimation of totals and means:
  - the Horvitz-Thompson estimator
  - the Variance estimator
- ★ Improving the Horvitz-Thompson estimator

Introduction

- Large expansion of R packages dedicated to survey sampling over the last 10 years: from few packages to more than eighty;
- Comprehensive list of all packages dedicated to survey sampling techniques and official statistics at

  `https://cran.r-project.org/web/views/OfficialStatistics.html`

  maintained by Matthias Templ.

Introduction (2)

- ★ Broadly speaking, there are two main R-packages dedicated to survey sampling techniques: sampling and survey
  - Package sampling was suggested by Alina Matei and Yves Tillé from the University of Neuchatel and is concerned mainly by performing sample selection according to various with or without replacement sampling designs. Some estimation issues are also treated.
  - Package survey was suggested by Thomas Lumley from the University of Auckland and is concerned with design-based estimation of finite population interest parameters and of their variance.
- ★ The other existing packages are dedicated to a specific issue from survey sampling.

Sample selection

- Sample selection with package sampling is concerned with the selection of sample according to many designs:
  - simple random sampling without replacement
  - unequal probability sampling designs
  - stratified sampling design
  - multistage sampling
- Estimation and variance estimation with package survey, for which methods are called from pakage srvyr using the dplyr sintax.

Estimation of totals

- Consider a finite population $U = \{1, \ldots, N\}$ and a sample s, selected from U according to a sampling design $p(\cdot)$;
- Denote by $\pi_k = P(k \in s)$ the first-order inclusion probabilities suppose to be positive $\forall k \in U$.
- Let Y be the interest variable and we want to estimate the finite population total of Y on U:

$$t_Y = \sum_{uk \in U} y_k$$

or the population mean $\bar{Y}_U = \frac{t_Y}{N}$.

- Weighted estimators are built to estimate t_Y: $\hat{t}_w = \sum_s \omega_k y_k$.

Horvitz-Thompson estimator

- The Horvitz-Thompson estimator $\hat{t}_\pi$ is obtained for

$$\omega_k = d_k = \frac{1}{\pi_k}$$

with $\pi_k = P(k \in s) > 0, \forall k \in U$ are the first inclusion probabilities and $\omega_k = d_k$ are the sampling weights.

- Supposing that all $\pi_{kl} > 0$ (the second inclusion probabilities), the variance of the Horvitz-Thompson estimator may be estimated unbiasedly by the Horvitz-Thompson variance estimator given by

$$\hat{V}_{HT}(\hat{t}_d) = \sum_{k \in U} \sum_{l \in U} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

Horvitz-Thompson estimator (2)

- For without replacement designs of equal size, the variance may be estimated unbiasedly also by the Yates-Grundy-Sen variance estimator:

$$\hat{V}_{YGS}(\hat{t}_d) = -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2$$

- Remark that these two variance estimators may be different. Both packages sampling and survey provide functions for computing these two variance estimators. Besides, variance estimators based on replicates weights are implemented in package survey.

Properties regarding the sampling design

- **Unbiasedness**: An estimator is said to be unbiased if in the long run it takes on the value of the population parameter. That is, if you were to draw a sample, compute the statistic, repeat this many, many times, then the average over all of the sample statistics would equal the population parameter.
- **Efficency**: An estimator is said to be efficient if in the class of unbiased estimators it has minimum variance.
- **Consistency**: (depends on the sampling desing) A sequence of estimators is said to be consistent if it converges in probability to the true value of the parameter

Example with R (starting from Tillé (2010))

- ★ Data: Belgian municipalities'
- ★ This is an example of unequal probability (UP) sampling functions: selection of samples using the Belgian municipalities' data set, with equal or unequal probabilities, and comparison of the Horvitz-Thompson estimator accuracy using boxplots.
  - The following sampling schemes are used:
    - Poisson,
    - systematic,
    - simple random sampling without replacement.
  - Monte Carlo simulations are used to study the accuracy of the Horvitz-Thompson estimator of a population total
  - The sample size is 200 in each simulation
- ★ View file .R for code and comments

More complex parameters of interest

- If we want to estimate more complex parameters of interest θ, then most of the times θ may be obtained as the unique solution (explicite or implicite) of a population estimation equation:

$$S(\theta) = \sum_U S_k(\theta) = 0$$

  Coefficients of linear and logistic regression, ratio and calibration estimators may be obtained in this way.
- The estimator of θ is $\hat{\theta}_d$, i.e. the solution of

$$\hat{S}(\hat{\theta}_d) = \sum_s \frac{S_k(\hat{\theta}_d)}{\pi_k} = 0$$

More complex parameters of interest (2)

- By Taylor linearization, the variance estimator is given by Binder (1983).
- Package survey implements this Taylor variance estimator for many non-linear estimators such as the ratio or calibration estimator.
- Designbased estimation of quantiles is treated by function svyquantile from survey.
- Multivariate quantile with survey data is proposed by Gmedian.
- Packages laeken and convey are dedicated to the estimation of income inequality indicators such as Gini or Theil index.
- The ratio estimator : the svyratio function

The Calibration Approach

- The Horvitz-Thompson estimator may be improved by using auxiliary information. One of the most popular method is the calibration estimator (Deville and Sarndal, 1992):
- Consider p auxiliary variables $X_1, \ldots, X_p$ and let $\mathbf{x}_k^{'} = (X_{1k}, \ldots, X_{pk})$ with $k = 1, \ldots, N$, $k \in U$
- Suppose that the total of auxiliary variables $t_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$ are known.
- The calibration methods consists in finding weights $\mathbf{w}_s^{cal}$ such that they are as close as possible in the sense of a distance to the sampling design $\mathbf{d} = (d_k)_{k \in S}$ and subject to the calibration constraints $\sum_{k \in s} w_{ks}^{cal} \mathbf{x}_k = t_{\mathbf{x}}$

The Calibration Approach (2)

- The calibrated estimator is

$$\hat{t}_y^{cal} = \sum_{k \in s} w_{ks}^{cal} y_k$$

- Several distance functions have been considered in Deville and Sarndal (1992) : the chi-squared, raking, logit distances

The Calibration Approach with R

- ★ functions calib, gencalib in package sampling compute the calibration weights and the g- weights (quite slow);
- ★ the calibWeights function in package laeken and calibSample function in package are simPop faster (depending on the example) implementation of parts of calib;
- ★ functions svyratio, poststratified, calibrate in package survey compute the calibration estimators and variance estimators obtained by Taylor linearization
- ★ package icarus gives the calibration weights.

Main References I

- ★ Goga, C. (2018). Brief overview of survey sampling techniques with R. Romanian Statistical Review, (1).
- ★ Lumley, T. (2011). Complex surveys: a guide to analysis using R (Vol. 565). John Wiley & Sons.
- ★ Lumley, T. (2018). Estimates in subpopulations. `https://cran.r-project.org/web/packages/survey/vignettes/domain.pdf`
- ★ Särndal, C. E. (2010). The calibration approach in survey theory and practice. Survey Methodology, 33(2), 99-119.
- ★ Tillé, Y., & Matei, A. (2005). The R package sampling. The comprehensive R archive network.
- ★ Tillé, Y., & Matei, A. (2010). TEACHING SURVEY SAMPLING WITH THE 'SAMPLING'R PACKAGE. ICOTS8 proceedings.