Survey sampling with the R ‘sampling’ package

A. Matei and Yves Tillé

Institute of Statistics,
University of Neuchâtel, Switzerland

Rencontres de statistique appliquée, INED
April 2010

Overview

- Using R and the 'sampling' package for teaching survey sampling theory and for training in this area;
- The sampling package is no used in public statistics. Especially in developing countries.
- Short overview on sampling theory;
- Specifical features of our package.
- Some exercises.
- Conclusions.

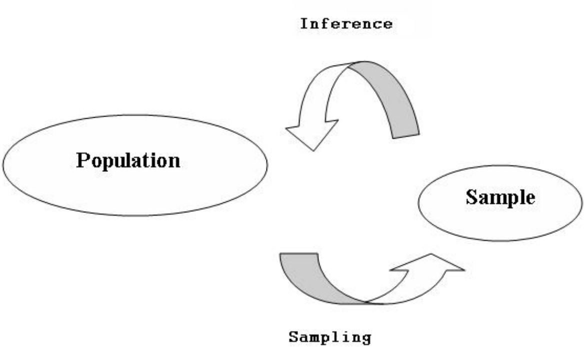
The R language

- R is a language and environment for statistical computing and graphics.
- Shareware available on the Comprehensive R Archive Network web page (http://cran.r-project.org/)
- Installation: 5 minutes; versions for Windows, MacOS X, Linux.
- Everyone can write an additional package (currently, the CRAN package repository features 2449 available packages)
- Packages are loaded directly from CRAN.
- Our package (called ‘sampling’) for survey sampling.
- The package manual is available online and in pdf.

The history of the package 'sampling'

- EFTA (European Free Trade Association) - training course for national office employers, organized by Eurostat and Swiss Federal Statistical Office, in April 2005 at Neuchâtel, Switzerland.
- The main goal: to study the sampling theory using R as statistical software.
- Writing of a large set of functions.
- The decision to submit the package to CRAN.

Inference



Population and sample

- let U be a finite population of size N;
- let $s \subseteq U$ be a sample with the probability to be selected $p(s)$;
- let $k \in U$ be the reference unit;
- $\pi_k = Pr(k \in s) = \sum_{s;s \ni k} p(s)$ is the inclusion probability for unit k;
- $\pi_{k\ell} = Pr(k, \ell \in s) = \sum_{s;k,\ell \ni s} p(s)$.

Estimation

- let $y = (y_1, \ldots, y_N)$ be the variable of interest, which is known only for the units in sample s;
- the goal in survey sampling is to estimate totals, means etc;
- a very popular unbiased estimator for the total $\sum_{k \in U} y_k$ is the Horvitz-Thompson estimator

$$\widehat{t}_{HT} = \sum_{k \in s} y_k / \pi_k .$$

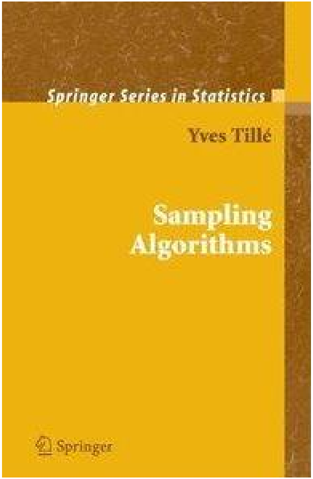
- an usual variance estimator for the HT estimator is the Sean-Yates-Grundy estimator

$$\widehat{var}_{\mathrm{SYG}}[\widehat{t}_{HT}] = \frac{1}{2} \sum_{k \in s} \sum_{\substack{\ell \in s \\ \ell \neq k}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \frac{\pi_k \pi_\ell - \pi_{k\ell}}{\pi_{k\ell}} .$$

Sampling

- Sampling with or without replacement.
- Equal or unequal probability sampling.
- Complex sampling: stratified sampling, multistage sampling, cluster sampling, balanced sampling and combination of them.
- `srswor(n,N), srswor1(n,N), srswr(n,N).`
- `strata(data, stratanames=NULL, size, method=c("srswor","srswr","poisson", "systematic"), pik, description=FALSE)`

Balanced sampling



Unequal probability sampling

- often an auxiliary information x_k is available for all $k \in U$;
- $$\pi_k = \frac{n x_k}{\sum_{\ell \in U} x_\ell}$$

  (πps sampling with sample size or expected sample size equal to n).
- `inclusionprobabilities(x,n)`,
- `UPmaxentropy(pik)`, `UPmidzuno(pik)`,`UPmidzunopi2(pik)`, `UPminimalsupport(pik)`, `UPmultinomial(pik)`, `UPopips(pik)`, `UPpivotal(pik)`, `UPpoisson(pik)`, `UPrandompivotal(pik)`, `UPrandomsystematic(pik)`, `UPsampford(pik)`, `UPsampfordpi2(pik)`, `UPsystematic(pik)`, `UPsystematicpi2(pik)`, `UPtille(pik)`,`UPtillepi2(pik)`.

Balanced sampling

- Design that satisfies the balancing equations

$$\sum_{k\in s} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in U} \mathbf{x}_k,$$

  where $\mathbf{x}_k$ is a vector of auxiliary variables.
- Cube algorithm: flight phase and landing phase.
- `samplecube(X,pik)`.
- `fastflightcube(X,pik)`.
- `landingcube(X,pik)`.

Calibration estimators

- A calibration estimator for the population total $\sum_{k\in U} y_k$ is defined as

$$\widehat{t}_{CAL} = \sum_{k\in s} w_k y_k,$$

where

$$\sum_{k\in s} w_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k = \mathbf{t}_x = \text{ known}, \tag{1}$$

for a vector of auxiliary variables $\mathbf{x}_k$.

- `calib(Xs, d, total, q=rep(1,length(d)), method=c("linear","raking","truncated", "logit"), bounds=c(low=0,upp=10), description=FALSE, max_iter=500)`
- `gencalib(Xs, Zs, d, total, q=rep(1,length(d)), method=c("linear","raking","truncated","logit"), bounds=c(low=0,upp=10), description=FALSE, max_iter=500)`

Other functions

1. *Estimation*: Horvitz-Thompson, Hájek, calibration, general calibration, regression, ratio, poststratified estimator,
2. *Tools*: computation of inclusion probabilities for UP from an auxiliary variable variable, crossing strata, response homogeneity groups, propensity scores,
3. *Data bases*: MU284, Swiss municipalities, Belgian municipalities.

Package manual and vignettes

- the package manual (in pdf format or HTML)
- a set of three vignettes.

Exercise 1

Exercise

Compute the inclusion probabilities, for a sample of size 200 drawn from the Belgian municipalities population, proportional to the population in 2004.

Exercise 2

Exercise

Use the Belgian database. Select a sample of 200 municipalities with unequal probabilities proportional to the number of inhabitants in 2004.

- *with Poisson sampling*
- *with a method of unequal probabilities and fixed sample size*
- *with simple random sampling.*

Conduct Monte-Carlo simulations, compute the Horvitz-Thompson (HT) estimator of the taxable income variable for 10'000 samples and finally draw a boxplot for each previous method to compare empirically the estimated variance of the HT estimator.

Exercise 3

Exercise

Use the MU284 population. Compute a vector of inclusion probabilités proportional to variable P75 for a sample size $n = 50$*. Construct a matrix with the balancing variables P75,CS82,SS82,S82,ME84,REV84. Next select a balanced sample on these variables and a sample of fixed sample size* $n = 50$ *with the same vector of inclusion probabilities. Run a set of simulations in order to compare the Horvitz-Thompson estimators of these two sampling designs.*

Exercise 3

Exercise

Conduct Monte-Carlo simulations to compare the accuracy of the Horvitz-Thompson estimator and Hájek estimator in terms of MSE. Four cases are considered:

1. *the variable of interest y_k is randomly generated using the $N(3,4)$ distribution;*
2. *Poisson sampling is used to draw a sample s and y_k is constant for $k = 1, \ldots, N$;*
3. *y_k is generated using the following model:* $x_k = k, \pi_k = nx_k / \sum_{i=1}^{N} x_i, y_k = 1/\pi_k$;
4. *y_k are generated using the following model:* $x_k = k, y_k = 5(x_k + \epsilon_k), \epsilon_k \sim N(0, 1/3)$;

For cases 1, 3 and 4 use Tillé sampling. In all cases, the population size is 100 and the sample size (or the expected sample size) is 20.

Exercise 4

Exercise

Use the database of Swiss municipalities, and select a stratified balanced sample. A balanced sample is first selected in each strata. Next the results of the flight phase are merged and a flight phase is applied again on the whole population. Finally, a landing phase is applied on all the population. Use the following balancing variables: HApoly, Surfacesbois, P00BMTOT, P00BWTOT, POPTOT, Pop020, Pop2040, Pop4065, Pop65P, H00PTOT. The sample size is 400 *and the municipalities must be selected with inclusion probabilities proportional to POPTOT. The stratification variable is REG (swiss regions). Next, print the names of the selected municipalities.*

Exercise 5

Exercise

Use the Belgian database. Select a sample of 200 municipalities with unequal probabilities proportional to the number of inhabitants in 2004 with Poisson sampling design. Next calibrate the sample by means of the raking ratio estimator on the variables:
Men03/mean(Men03), Women03/ mean(Women03), Diffmen, Diffwom, TaxableIncome, Totaltaxation, averageincome, medianincome.
The division by the means is necessary to avoid too large numbers. Compute the Horvitz-Thompson estimator and the calibrated estimators for the calibration variables. Limit the variation of the g-weights between 0.5 and 1.5.

Conclusions

- the R 'sampling' package is a tool to teach survey sampling theory, to do training and research in this area;
- it can be used for training in official statistics, for university courses in survey sampling and biostatistics, on graduate or post-graduate level.
- it is also a valuable reference for practicing statisticians who are involved in the design of sample surveys and estimations.