HT and Sampling Design

Gaia Bertarelli [1]

[1] Dept.of Economics & Management , University of Pisa

Lab. 2, 7th Oct. 2019

The presentation at a Glance

1. Introduction to Probabilistic Sampling
2. The Horvitz-Thompson Estimator
3. Sampling Designs
4. Survey Sampling using R

1. Introduction to Probabilistic Sampling

2. The Horvitz-Thompson Estimator

3. Sampling Designs

4. Survey Sampling using R

Introduction to Probabilistic Sampling

- ★ In survey samples it is specified a population, whose data values are unknown but are regarded as **fixed**, not random. Although, the observed sample is **random** because it depends the random selection of individuals from this fixed population.

- ★ Properties of a sampling method
  - Every individual in the population must have a known and a nonzero probability of belonging to the sample ($\pi_i > 0$ for individual i). And π_i must be known for every individual who ends up in the sample.
  - Every pair of individuals in the sample must have a known and a nonzero probability of belonging to the sample ($\pi_{ij} > 0$ for the pair of individuals $(i; j)$). And π_{ij} must be known for every pair that ends up in the sample.

Sampling weights

- ★ If we take a simple random sample of 3500 people from Neverland (with total population 35 million) then any person in Neverland has a chance of being sampled equal to $\pi_i = \frac{3500}{35000000} = \frac{1}{10000}$ for every i.

- ★ Then, each of the people we sample represents 10000 Neverland inhabitants.

- ★ If 100 people of our sample are unemployed, we would expect then $100 \times 10000 = 1million$ unemployed in Neverland.

- ★ An individual sampled with a sampling probability of π_i represents $\frac{1}{\pi_i}$ individuals in the population.

- ★ $\frac{1}{\pi_i}$ is called the sampling weight.

Sampling weights (2)

- **Example**: Measure the income on a sample of one individual from a population of N individuals, where π_i might be different for each individual.
  - The estimate ($\hat{T}_{income}$) of the total income of the population (T_{income}) would be the income for that individual multiplied by the sampling weight

$$\hat{T}_{income} = \frac{1}{\pi_i} \times income_i$$

  - Not a good estimate, it is based on only one person, but it is will be **unbiased**: the expected value of the estimate will equal the true population total:

$$E(\hat{T}_{income}) = \sum_{i=1}^{N} \frac{1}{\pi_i} \times income_i \cdot \pi_i = income_i$$

1. Introduction to Probabilistic Sampling

2. The Horvitz-Thompson Estimator

3. Sampling Designs

4. Survey Sampling using R

The Horvitz-Thompson Estimator

- The so called Horvitz-Thompson estimator of the population total is the foundation for many complex analysis.
- If Y_i is a measurement of variable Y on person i , we write

$$\tilde{Y}_i = \frac{1}{\pi_i} Y_i$$

.

- Given a sample of size n the Horvitz-Thompson estimator

$$\hat{T}_Y = \sum_{i=1}^{n} \frac{1}{\pi_i} Y_i = \sum_{i=1}^{n} \tilde{Y}_i$$

The Horvitz-Thompson Estimator (2)

- The variance estimator is

$$V\hat{a}r(\hat{T}_Y) = \sum_{i,j} \left(\frac{Y_i Y_j}{\pi_{ij}} - \frac{X_i}{\pi_i} \frac{X_j}{\pi_j} \right)$$

- The formula applies to any design, however complicated, where π_i and π_{ij} are known for the sampled observations.
- The formula depends on the pairwise sampling probabilities π_{ij} , not just on the sampling weights: so the correlations in the sampling design enter the computations.
- For without replacement designs of equal size, the variance may be estimated unbiasedly also by the **Yates-Grundy-Sen** variance estimator.
- Remark that these two variance estimators may be different.

1. Introduction to Probabilistic Sampling
2. The Horvitz-Thompson Estimator
3. Sampling Designs
4. Survey Sampling using R

Simple Random Sampling (SRS)

- ★ SRS is the simplest way to sample a population. Its simplicity arises from the way that the sample is selected.
- ★ In this design, all possible samples have the same probability to be chosen.
- ★ In SRS the population to be sampled is considered as a simple collection of elements, where no subgroups are considered.
- ★ The probability π_i of selecting an element i from a finite population of size N, is $\pi_i = \frac{1}{N}$.
- ★ There are two ways to take a simple random sample:
  - **with replacement**: the elements are selected with replacement of the element into the population after each extraction,
  - **without replacement**.
- ★ SRS without replacement from a finite population: $f = \frac{n}{N}$ is the sampling fraction.
- ★ SRS with replacement (or infinite population): the sampling fraction is zero

Stratified random sampling

- ★ When the population is heterogeneous, dividing the whole population into sub-populations, called strata, can increase the precision of the estimates.
- ★ The strata should not overlap and each stratum should be sampled following some design.
- ★ All strata must be sampled.
- ★ The strata are sampled separately and the estimates from each stratum combined into one estimate for the whole population.
- ★ The theory of stratified sampling deals with the properties of the sampling distribution of the estimators and with different types of allocation of the sample sizes to obtain the maximum precision.
- ★ The principle of stratification is the partition of the population in such a way that the elements within a stratum are as similar as possible and the means of the strata are as different as possible.

1 Introduction to Probabilistic Sampling

2 The Horvitz-Thompson Estimator

3 Sampling Designs

4 Survey Sampling using R

Survey Sampling using R

- ★ 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.

Survey Sampling using R (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.

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