Adjusting Sample Weights to Integers: Effects on Mean and Variance Estimators in Simple Random Sampling

Alberto M. Padilla Terán

Formerly Statistical Researcher at Banco de México

1 Background

- In probability sampling, **sampling weights** are used to construct estimates of averages and variances, among other quantities.
- Such **weights** may include.
  - Non-response adjustments.
  - Adjustments to known population totals.
  - Calibration with information from auxiliary variables positively correlated with the variables of interest in the survey.
  - Rounding sampling weights to Integers.
- **Rounding** sampling weights is carried out in surveys relevant to decision-making such as:
  - ENOE (National Survey of Occupation and Employment), conducted by INEGI-quarterly.
  - The monthly survey of IMEF (Mexican Institute of Finance Executives) from where the manufacturing and non-manufacturing IMEF Indicators are constructed.
    - These indicators are used to anticipate the direction of economic activity, see Heath and Domínguez.
- The methodological documents for these two surveys do not describe the rounding method used or the reasons for its use.

2 Objective

In this poster, it will be shown that rounding to integers the sampling weights biased the estimator of the sample average under simple random sampling. Some examples will exemplify this effect.

- The effect on systematic sampling with equal probabilities of selection will also be discussed.

3 Sample Designs Used

Two sample designs widely used in practice will be analysed, see Cochran or Särndal et al.:

- Simple random sampling without replacement, ***srswr***.
- Systematic sampling with equal probability of unit selection, ***sys***.
  - It will be exemplified using circular systematic sampling.

4 Notation

- **N** = total of elements in population
- **n** = total of elements in sample
- ω = N/n = sampling weight **without rounding**
- $\omega_{r,inf}$ = [ω] - 1 = sampling weight rounded down, [·] it's the integer part
- $\omega_{r,sup}$ = [ω] = sampling weight rounded up
- **r** = remainder of the división of N by n
- **c** = quotient of the división of N by n

5 Formula for Evaluating the Effects of Rounding on Sampling Weights

Because handling the integer function [ ] is not useful for the desired development, note that N can be expressed as :

$$N = n\,c + r$$

From this expression and the notation it can be seen that $c = \omega_{r,inf}$, therefore:

$$N = n\,\omega_{r,inf} + r$$

From this, the values of the rounded sampling weights are written as :

$$\omega_{r,inf} = (N-r)/n \quad \text{and} \quad \omega_{r,sup} = (N-r)/n+1$$

With these definitions, all that remains is to **determine the number of elements in the sample** that are rounded up and those that are rounded down, so that their sum is **n**.

- n_{inf}, will denote the number of ielements in sample rounded down
- n_{sup}, number of elements in sample rounded up.

The sum of the two must equal **n**:

$$n = n_{inf} + n_{sup}$$

7 Determining Rounded Sample Sizes

- The values of n_{inf} and n_{sup} can be obtained by solving the following linear system of equations :

$$\begin{aligned} n_{inf} + n_{sup} &= n \\ \omega_{r,inf}\, n_{inf} + \omega_{r,sup}\, n_{sup} &= N \end{aligned}$$

solution:

$$n_{inf} = (N - n\,\omega_{r,sup}) / (\omega_{r,inf} - \omega_{r,sup})$$

$$n_{sup} = n - n_{inf}$$

8 Estimations of the population total under srswr with rounding effects

- The Horvitz-Thompson estimator of the total is used, see Särndal et al. (1992) for a variable y_k.
- The **sampling weights** for the **srswr** are N/n.
- Population total estimator: $\hat{y} = \sum_k \frac{N}{n} y_k$
- Index *k* runs over all the elements in the sample.
- The total estimator, using the **rounded sampling weights**, is constructed with the weights rounded down (indices in *r1*) and up (indices in *r2*) as:
- $\hat{y} = \sum_{k\in r1} \omega_{k,r,inf}\, y_k + \sum_{k\in r2} \omega_{k,r,sup} y_k$
- To obtain the **estimated total rounded** using ***msrswr***, $\hat{y}_{red}$, the values of $\omega_{r,inf}$ and $\omega_{r,sup}$ are substituted in $\hat{y}$

$$\hat{y}_{red} = \frac{N-r}{n}\sum_k y_k + \sum_{k\in r2} y_k$$

- Index *k* runs over all the elements in the sample..
- Note that the second part of the estimator only depends on the values of the variable y_k in the part of the sample that was rounded up. The first part has an effect of the residual of division N/n.
- The expected value using ***srswr*** is:

$$E(\hat{y}_{red}) = \left(1 - \frac{r}{N}\right) y_U + E\left(\sum_{k\in r2} y_k\right)$$

- This value is **biased** and the second term depends on the sampling weights that are rounded up.

9 Examples

Example 1: Population with ***N=8*** elements and values y_k equal to {3,34,29,36,43,31,20,17} From this population all possible 56 samples of size ***n=3*** are drawn using ***srswr***.

In this example: *f=N/n=2.67*, *r=2*, $\omega_{r,inf}$ = 2, $\omega_{r,inf}$ = 3, n_{inf} = 1, n_{sup} = 2. For all possible samples, the possible roundings are calculated; that is; rounding down on the first item in the sample (denoted by red_1); rounding the second item in sample (red_2) and the third (red_3) and are compared with the non-rounding estimator (red_0), which uses *f=N/n*.

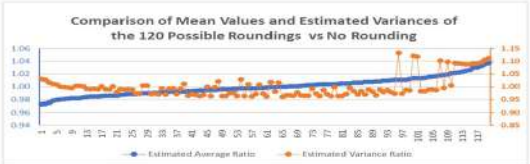
- *The population parameters are:*
  - ***Mean*** *= 26.625*
  - S^2 *= 161.411*
  - ***Variance*** *of the* ***mean*** *estimator = 33.627*
- The table on the right shows that rounding has slight effects on the average and large effects on the variance.

	red_0	red_1	red_2	red_3
Coefficient of variation (mean)	22%	19%	24%	23%
Estimated mean	26.625	27.188	25.775	26.913
Variance between means	33.627	27.730	38.762	38.311
Comparison of Menas and Variances of Rounding vs No Rounding weights				
mean red_i/red_0		2%	-3%	1%
variance red_i/red_1		-18%	15%	14%

Note: this exercise was carried out for circular systematic sampling in which there are eight possible samples and with the possible roundings, the deviations of population variances were: -1%, 28% and 1%. The estimated means were not changed due to the balance of the units sampled in the subsets that are rounded up and down..

Example 2, *Estimate the total number of conservative legislators in Sweden's city councils with a small sample from the book by Särndal et al. (1992), example 4.2.1, page 129. It's a simple random sample of one-stage clusters of size n=16 from a population with N=50. The estimated value of the total is 2,347 legislators and the variance estimator is 62.312, see page 130 of Särndal et al. (1992).*

- *In this example we have f=N/n= 3.125, r=2, $\omega_{r,inf}$ = 3, $\omega_{r,inf}$ = 4, n_{inf} = 14, n_{sup} = 2. These estimators will be compared to those computed with all sampling weights rounding combinations, which are all size 2 subsets of 16, which gives 120 possible types of rounding.*
- *As there are 120 possible values, two graphs of the mean and variance estimators are shown compared with those obtained by Särndal et al. (1992). Values were sorted by estimated average in an increasing manner.*
- *The graph shows that the estimated values of the average show deviations of up to 4% compared to the average without using rounding. The estimated variances show deviations of up to 15% with respect to the estimated variance without rounding to integers the sampling weights.*



10 Conclusions

- An expression was constructed showing that the mean estimator using sampling weights rounded to integers is biased and exemplified with two small populations (they could be part of a stratum in a larger population).
- In the examples, it can be appreciated a medium to large effect of deviation in the estimated variance using sampling weights rounded to integers compared to the variance without rounding in **srswr**.
- In the case of circular systematic sampling, there are *effects in the estimation of the variance of the total*; however, there is no impact on the mean estimator because the sampled items appear the same numbers.

References

[1] Cochran, W.G. (1977), Sampling Techniques, 3rd edn. New York: Wiley.
[2] Heath, J. y Domínguez, L., Marco Conceptual y Metodológico del Indicador del Entorno Empresarial Mexicano IMEF, Instituto Mexicano de Ejecutivos de Finanzas, México, (sin fecha en la nota).
[3] INEGI, (National Survey of Occupation and Employment, ENOE. (2007), Cómo se hace la ENOE. Métodos y Procedimientos.
[4] Murthy, M.N. & Rao, T.J. (1988) Systematic Sampling, Chapter 7 in Handbook of Statistics 6: Sampling, ed. by C.R. Rao, Amsterdam, North Holland.
[5] Padilla, A. (2009) "An Unbiased Estimator of the Variance of Simple Random Sampling using Mixed Random-Systematic Sampling", Banco de México, Documento de Investigación No. 2009-13.
[6] Särndal, C.E., Swensson, B. & Wretman, J.H., Model Assisted Survey Sampling, Springer-Verlag, New York, 1992.

Contact

betoampfq@gmail.com

Note:

This article was presented at the *XV Semana Internacional de la Estadística y la Probabilidad 2022 13 al 17 de junio FCFM-BUAP, Puebla, México* (XV International Week of Statistics and Probability 2022 June 13 to 17 FCFM-BUAP, Puebla, Mexico.)