Sampling schemes providing unbiased mean-of-the-ratios estimates: a review

Mariano Ruiz Espejo
Departamento de Matemáticas Fundamentales
UNED. Madrid

Abstract

We consider the mean-of-the-ratios estimate and propose two sampling schemes for which a modification of such statistic results unbiased for estimating the finite population mean. This study completes the ones by Hartley and Ross (1954) and by Ruiz Espejo and Santos Peñas (1989) which provided unbiased mean-of-the-ratios estimates for simple random sampling without replacement design.

Keywords: mean-of-the-ratios estimate, population mean, unbiased estimation, sampling schemes.

AMS classification: 62D05.

Esquemas muestrales que proporcionan estimaciones media-de-razones insesgadas: una revisión

Resumen

Consideramos la estimación media-de-razones y proponemos dos esquemas muestrales para los cuales una modificación de tal estadístico resulta insesgado para estimar la media de una población finita. Este estudio completa los de Hartley y Ross (1954) y de Ruiz Espejo y Santos Peñas (1989) que proporcionaban estimaciones media-de-razones insesgadas para diseño de muestreo aleatorio simple sin reemplazamiento.

Palabras clave: estimación media-de-razones, media poblacional, estimación insesgada, esquemas de muestreo.

Clasificación AMS: 62D05.

1. Introduction

The mean-of-the-ratios estimator is the sample statistic

$$\bar{r}_s = \frac{1}{n}\sum_{i\in s}\frac{y_i}{x_i},$$

where $\boldsymbol{s}$ is an ordered sample of size n selected from an identified finite population U of size N. Here the two variables y (interest one) and x (auxiliary one) are defined for each unit i of the population and are denoted for such unit y_i and x_i. We assume positive all the values of the auxiliary variable x. The value $r_i = y_i/x_i$ is the variable ratio for the ith unit of the finite population. We will study three sampling schemes for which this statistic $\bar{r}_s$ or a direct transformation of it provide unbiased estimates for the finite population mean $\bar{y}$ of the interest variable y.

2. Srswor design

If we like to estimate the population mean of the interest variable

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

using the mean-of-the-ratios statistic, Hartley and Ross (1954) provided an unbiased estimate of $\bar{y}$, namely

$$t_{HR} = \bar{r}_s\bar{x} + \frac{n(N-1)}{(n-1)N}(\bar{y}_s - \bar{r}_s\bar{x}_s),$$

where $\bar{y}_s$ and $\bar{x}_s$ are the sample means for the respective variables, and the sample $\boldsymbol{s}$ is drawn using "simple random sampling without replacement" (srswor) design. Ruiz Espejo and Santos Peñas (1989) provided another unbiased estimator of the population mean $\bar{y}$ using the mean-of-the-ratios statistic for the same sampling design,

$$t_{RS} = \frac{n(N-1)}{N-n}\bar{r}_s\bar{x}_s - \frac{(n-1)N}{N-n}\bar{r}_s\bar{x}$$

among many others based on the Ruiz Espejo statistic

$$z = (N-n)\bar{y}_s + (n-1)N\bar{r}_s\bar{x} - n(N-1)\bar{r}_s\bar{x}_s$$

which has expectation zero for srswor design.

In the following of this article, we propose other two ordered sampling schemes providing unbiased mean-of-the-ratios estimates. The first one was proposed by Hansen and Hurwitz (1943) namely "probability proportional to size with replacement" (ppswr) sampling design, and the other one was suggested by Sánchez-Crespo (1980) namely "probability gradually variable without replacement" (pgvwor) sampling design.

ESTADÍSTICA ESPAÑOLA Vol. 57 Núm. 187 / 2015

3. Hansen-Hurwitz ppswr sampling scheme

The sampling scheme proposed by these authors is the traditional ordered ppswr sampling design which consists in selecting a unit i of the finite population in each draw of the n ordered extractions that contains the vector sample $\boldsymbol{s}$ with probability proportional to the size $x_i > 0$ and with replacement. Thus, in each jth extraction the mathematical expectation of r_{i_j} is

$$E\left(r_{i_j}\right) = \sum_{i=1}^{N} r_i \frac{x_i}{N\bar{x}} = \frac{\bar{y}}{\bar{x}}.$$

The mean-of-the-ratios estimator of the population mean $\bar{y}$ is

$$t_{MR} = \frac{\bar{x}}{n}\sum_{i\in s}\frac{y_i}{x_i} = \frac{\bar{x}}{n}\sum_{j=1}^{n}\frac{y_{i_j}}{x_{i_j}} = \frac{\bar{x}}{n}\sum_{i\in U}\frac{y_i}{x_i}e_i$$

where e_i is the ith component of the vector $(e_1, e_2, \dots, e_N)$ which follows a multinomial distribution of parameters n and $p_i = x_i/N\bar{x}$ $(i = 1, 2, \dots, N)$. With this threefold expression we can give other respective proofs of the unbiasedness of the estimator t_{MR}. One of them is the following for clarity,

$$E(t_{MR}) = \frac{\bar{x}}{n}\sum_{i\in U}\frac{y_i}{x_i}E(e_i) = \frac{1}{N}\sum_{i\in U} y_i = \bar{y},$$

because $E(e_i) = nx_i/N\bar{x}$.

The variance of this estimator t_{MR} for ppswr sampling design uses the value of the variance $V(e_i) = nx_i(N\bar{x} - x_i)/N^2\bar{x}^2$ and the covariance for $i \neq j$,

$$Cov\left(e_i, e_j\right) = -n\frac{x_i x_j}{N^2\bar{x}^2}.$$

The variance of t_{MR} for the Hansen-Hurwitz sampling scheme is

$$V_{HH}(t_{MR}) = \frac{\bar{x}^2}{n^2}\left[\sum_{i\in U}\frac{y_i^2}{x_i^2}V(e_i) + \sum_{i\in U}\sum_{j\neq i\in U}\frac{y_i y_j}{x_i x_j}Cov\left(e_i, e_j\right)\right]$$

$$= \frac{1}{nN^2}\left[\sum_{i\in U}\frac{y_i^2}{x_i}(N\bar{x} - x_i) + \sum_{i\in U}\sum_{j\neq i\in U}\left(-y_i y_j\right)\right]$$

$$= \frac{1}{nN^2}\left[N^2\left(A_{2-1}A_{01} - \frac{A_{20}}{N}\right) - \sum_{i\in U} y_i(N\bar{y} - y_i)\right]$$

$$= \frac{1}{n}\left(A_{2-1}A_{01} - \frac{A_{20}}{N} - A_{10}^2 + \frac{A_{20}}{N}\right)$$

$$= \frac{1}{n}(A_{2-1}A_{01} - A_{10}^2),$$

where

$$A_{kj} = \frac{1}{N}\sum_{i\in U} y_i^k x_i^j$$

is the population noncentral moment of orders k and j.

An unbiased variance estimator of t_{MR} is obtained from the relation

$$\hat{V}(t_{MR}) = \frac{1}{n}\left(\widehat{A_{2-1}}A_{01} - \widehat{A_{10}^2}\right) = \frac{1}{n}\left[\widehat{A_{2-1}}A_{01} - t_{MR}^2 + \hat{V}(t_{MR})\right],$$

or also,

$$\hat{V}(t_{MR}) = \frac{1}{n-1}\left(\frac{\bar{x}^2}{n}\sum_{i\in U}\frac{y_i^2}{x_i^2}e_i - t_{MR}^2\right).$$

In this relation we have considered as unbiased estimator of A_{2-1} to

$$\widehat{A_{2-1}} = \frac{\bar{x}}{n}\sum_{i\in U}\frac{y_i^2}{x_i^2}e_i.$$

And obviously, an unbiased estimator of $\bar{y}^2$ is

$$\widehat{\bar{y}^2} = t_{MR}^2 - \hat{V}(t_{MR}).$$

This reasoning can be obtained directly from the results given by Hansen and Hurwitz (1943) in their article. This is possible substituting x_i by $N\bar{x}p_i$ and simplifying. Other unbiased estimation for the variance of t_{MR} is

$$\hat{V}(t_{MR}) = \frac{\sum_{i\in U}\left(\bar{x}\frac{y_i}{x_i} - t_{MR}\right)^2 e_i}{n(n-1)}.$$

4. Sánchez-Crespo pgvwor sampling scheme

The Sánchez-Crespo (1980) sampling scheme consists in the selection of n extractions without replacement of an urn which contains M numbered balls where there are $M_i \geq 1$ balls with the identification number i ($i = 1, 2, \dots, N$), and

$$M = \sum_{i=1}^{N} M_i.$$

Then, if we select n balls with equal probabilities without replacement of such urn, and e_i is the number of drawn balls in the sample of size n, the random vector $(e_1, e_2, \dots, e_N)$ follows a generalized hypergeometric distribution of parameters n, N, and M_1, M_2, …, M_N. Here, M_i is the minimum natural value proportional to $x_i > 0$, for all $i = 1, 2, \dots, N$. For this reason, 1 is the highest common factor of $\{M_i : i = 1, 2, \dots, N\}$ and $M_i \propto x_i$ $(i = 1, 2, \dots, N)$.

In these conditions, we can calculate the expectation, variance and covariances of the random numbers e_i as it follows. For all $i = 1, 2, \dots, N$ we have

$$E(e_i) = \frac{n x_i}{N \bar{x}},$$

$$V(e_i) = \frac{M-n}{M-1} \frac{n x_i (N\bar{x} - x_i)}{N^2 \bar{x}^2}.$$

And for all $i \neq j$,

$$Cov(e_i, e_j) = -\frac{M-n}{M-1} \frac{n x_i x_j}{N^2 \bar{x}^2}.$$

With a similar reasoning to the Hansen-Hurwitz scheme, in this case we can obtain the unbiasedness of the new estimator mean-of-the-ratios for estimating the finite population mean $\bar{y}$,

$$t_{MR} = \frac{\bar{x}}{n} \sum_{i \in U} \frac{y_i}{x_i} e_i,$$

where now the e_i follows the explained generalized hypergeometric distribution in the Sánchez-Crespo sampling scheme, a difference with respect to the e_i of the mean-of-the-ratios estimator with the Hansen-Hurwitz sampling scheme which was multinomial distribution.

With other similar reasoning, the variance of this estimator t_{MR} with the Sánchez-Crespo sampling scheme is

$$V_{SC}(t_{MR}) = \frac{M-n}{M-1} V_{HH}(t_{MR}) = \frac{M-n}{M-1} \frac{1}{n} (A_{2-1} A_{01} - A_{10}^2) \leq V_{HH}(t_{MR}).$$

Other two expressions of an unbiased estimator of the variance of t_{MR} for Sánchez-Crespo sampling design are (according to Ruiz Espejo, 1995, 2013)

$$\hat{V}(t_{MR}) = \frac{M-n}{M}\frac{1}{n-1}\left(\frac{\bar{x}^2}{n}\sum_{i\in U}\frac{y_i^2}{x_i^2}e_i - t_{MR}^2\right)$$

and

$$\hat{V}(t_{MR}) = \frac{M-n}{M}\frac{1}{n(n-1)}\sum_{i\in U}\left(\bar{x}\frac{y_i}{x_i} - t_{MR}\right)^2 e_i$$

where now the random values e_i correspond to the Sánchez-Crespo pgvwor sampling scheme.

5. Conclusions

The mean-of-the-ratios estimate is an interesting statistic when we dispose of an auxiliary variable approximately proportional to the variable of interest in each finite population unit. Some examples of this are ($y =$ "land yield", $x =$ "cultivated surface") and ($y =$ "family consumption", $x =$ "family income").

In such conditions, we have proposed three sampling schemes for which there are unbiased mean-of-the-ratios estimates for estimating the finite population mean of the variable of interest. These are the traditional srswor design studied by Hartley and Ross (1954) and by Ruiz Espejo and Santos Peñas (1989), the ppswr sampling design proposed by Hansen and Hurwitz (1943), and the pgvwor sampling design proposed by Sánchez-Crespo (1980). The admissibility of the Hartley-Ross estimator was studied by Sengupta (1983), and the Sánchez-Crespo sampling scheme is more precise than the Hansen-Hurwitz sampling scheme when both ones are realizable. For these last two sampling schemes we have provided unbiased variance estimators based on the corresponding sampling scheme, the auxiliary variable and the data of the interest variable obtained in the sample according to the considered sampling scheme.

As a consequence, we have reviewed the three most important sampling strategies for estimating without bias the finite population mean $\bar{y}$ using an auxiliary variable x when there is an approximate proportionality between the interest variable and the auxiliary variable for the units of the finite population.

References

HANSEN, M. H. AND HURWITZ, W. N. (1943). «On the theory of sampling from finite populations». *Annals of Mathematical Statistics* 14, 333-362.

HARTLEY, H. O. AND ROSS, A. (1954). «Unbiased ratio estimators». *Nature* 174, 270-271.

RUIZ ESPEJO, MARIANO (1995). «A relation between sample and population quasicovariances» (Spanish). *Revista de la Academia de Ciencias de Zaragoza* 50, 51-53.

RUIZ ESPEJO, MARIANO (2013). «Exactness of Inference in Finite Populations» (Spanish), 1st edn. Madrid. Bubok.

RUIZ ESPEJO, MARIANO AND SANTOS PEÑAS, JULIÁN (1989). «Unbiased mean-of-the-ratios estimators». *Statistica* 49, 617-622.

SÁNCHEZ-CRESPO RODRÍGUEZ, JOSÉ LUIS (1980). «Intensive Course of Sampling in Finite Populations» (Spanish), 2nd edn. Madrid. Instituto Nacional de Estadística.

SENGUPTA, S. (1983). «On the admissibility of Hartley-Ross estimator». *Sankhya Series B* 45, 471-474.