On studying the relationship between the ratio and the regression estimators for a population proportion

Muñoz, Juan F.
University of Granada, Department of Quantitative Methods in Economic and Business
Facultad de Ciencias Económicas y Empresariales, Campus Cartuja s/n
Granada (18071),Spain
E-mail: jfmunoz@ugr.es

Arcos, Antonio
University of Granada, Department of Statistics and O.R.
Facultad de Farmacia, Campus Cartuja s/n
Granada (18071),Spain
E-mail: arcos@ugr.es

Álvarez-Verdejo, Encarnación
University of Granada, Department of Quantitative Methods in Economic and Business
Facultad de Ciencias Económicas y Empresariales, Campus Cartuja s/n
Granada (18071),Spain
E-mail: encarniav@ugr.es

Abstract

Estimation of a proportion is commonly used in many areas and disciplines. Traditional estimators and confidence intervals for a population proportion does not assume auxiliary information at the estimation stage. Assuming auxiliary information, the ratio and regression methods are known techniques for the purpose of estimation of a population mean. They can be adapted to the problem of the estimation of the population proportion, although some differences can be observed. For example, a ratio estimator for a population proportion can be defined via a ratio estimator for the complementary of the population proportion. We define an optimum ratio estimator for a population proportion and show that this estimator coincides with the regression estimator. Results derived for the population proportion can be easily extended to the problem of estimating the distribution function.

Keywords: Auxiliary information, sample survey, variance, inclusion probability.

Introduction

Traditional estimators and confidence intervals for a population proportion (Blyth and Still, 1983; Cohen and Yang, 1994; Fleiss et al., 2003; Newcombe, 1998) do not assume auxiliary information at the estimation stage. Assuming auxiliary information, the ratio and regression methods are known techniques for the purpose of estimation of a population mean. They can be adapted to the problem of the estimation of the population proportion (see, for example, Rueda et al., 2011a), although some differences can be observed. For example, a ratio estimator for a population proportion can be defined via a ratio estimator for the complementary of the population proportion. Furthermore, one should be aware of the risks when confidence intervals are constructed for a population proportion, since limits outside $[0, 1]$ could be achieved.

We consider the scenario of a finite population $U = \{1, \ldots, N\}$ containing N units. Let $A_1, \ldots, A_N$ denote the values of a attribute of interest A, where $A_i = 1$ if ith unit possesses the attribute A and $A_i = 0$ otherwise. Let B denote an auxiliary attribute associated with A and values given by $B_1, \ldots, B_N$. We also assume that a sample s, of size n, is selected from U according to an

arbitrary sampling design with first and second order inclusion probabilities given by π_i and π_{ij}.

The aim is to estimate the population proportion of individuals that posses the attribute A, i.e. $P_A = N^{-1}\sum_{i=1}^{N} A_i$. The naive estimator of P_A, which makes no use of the auxiliary information, is given by $\widehat{p}_A = N^{-1}\sum_{i\in s} d_i A_i$, where $d_i = \pi_i^{-1}$. Note that $\widehat{p}_A$ can take values larger than 1. An alternative estimator that avoids values larger than 1 is the Hájek type estimator, which is given by $\widehat{p}_{A.H} = \widehat{N}^{-1}\sum_{i\in s} d_i A_i$, where $\widehat{N} = \sum_{i\in s} d_i$.

Note that results related to the population proportion can be easily extended to the problem of estimating the distribution function $F(t)$ of a variable of interest y, since $F(t)$ is defined as the proportion of individuals in the population which take values of y less or equal than a given argument t.

Assuming a general sampling design, Rueda et al. (2011a) defined the ratio type estimators $\widehat{p}_r = \widehat{R}P_B$ and $\widehat{p}_{r.q} = 1 - \widehat{q}_r = \widehat{R}_c Q_B$, where $\widehat{R} = \widehat{p}_A/\widehat{p}_B$ is the design-based estimator of the population ratio $R = P_A/P_B$, $\widehat{p}_B = N^{-1}\sum_{i\in s} d_i B_i$, $P_B = N^{-1}\sum_{i=1}^{N} B_i$ is the population proportion of individuals that posses the attribute B, $\widehat{R}_c = \widehat{q}_A/\widehat{q}_B$, and the complementary proportions are defined as $\widehat{q}_A = 1 - \widehat{p}_A$, $\widehat{q}_B = 1 - \widehat{p}_B$ and $Q_B = 1 - P_B$. It is assumed that P_B is known from a census or estimated without error. Rueda et al. (2011b) also defined a regression type estimator for the population proportion under a general sampling design.

Assuming a general sampling design, an optimum ratio estimator for the population proportion P_A is defined. This estimator is based on a lineal combination between the ratio estimators $\widehat{p}_r$ and $\widehat{p}_{r.q}$. We also show that the proposed optimum estimator coincides with the regression type estimator.

The optimum ratio estimator

Assuming a general sampling design and the ratio estimators $\widehat{p}_r$ and $\widehat{p}_{r.q}$ previously defined, the following class of estimator can be defined

$$\widehat{p}_{r.w} = w\widehat{p}_r + (1-w)\widehat{p}_{r.q}. \tag{1}$$

It can be easily seen that the optimum value for w in the sense of minimum variance into the class of estimators $\widehat{p}_{r.w}$ is

$$w_{opt} = \frac{AV(\widehat{p}_{r.q}) - cov(\widehat{p}_r, \widehat{p}_{r.q})}{AV(\widehat{p}_r) + AV(\widehat{p}_{r.q}) - 2cov(\widehat{p}_r, \widehat{p}_{r.q})}, \tag{2}$$

where the asymptotic variance of $\widehat{p}_r$ is given by

$$AV(\widehat{p}_r) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{A_i - RB_i}{\pi_i}\frac{A_j - RB_j}{\pi_j}.$$

Analogously, the asymptotic variance of $\widehat{p}_{r.q}$ is given by

$$AV(\widehat{p}_{r.q}) = AV(\widehat{q}_r) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{A_i^c - R_c B_i^c}{\pi_i}\frac{A_j^c - R_c B_j^c}{\pi_j}.$$

A^c and B^c denote the complementary attribute of A and B. The covariance between the ratio estimators $\widehat{p}_r$ and $\widehat{p}_{r.q}$ is

$$cov(\widehat{p}_r, \widehat{p}_{r.q}) = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{A_i - RB_i}{\pi_i}\frac{A_j^c - R_c B_j^c}{\pi_j}.$$

Proof

Using Taylor series (see Särndal et al., 1992, pg 178) under a general sampling design, the ratios $\widehat{R}$ and $\widehat{R}_c$ can be approximated as

$$\widehat{R} \cong R + \frac{1}{NP_B}\sum_{i\in s}\frac{A_i - RB_i}{\pi_i}$$

and

$$\widehat{R}_c \cong R_c + \frac{1}{NQ_B}\sum_{i\in s}\frac{A_i^c - R_cB_i^c}{\pi_i}.$$

From previous expressions we obtain

$$\widehat{p}_r \cong P_A + \frac{1}{N}\sum_{i\in s}\frac{A_i - RB_i}{\pi_i}$$

and

$$\widehat{q}_r \cong Q_A + \frac{1}{N}\sum_{i\in s}\frac{A_i^c - R_cB_i^c}{\pi_i},$$

which can be used to obtain the covariance between $\widehat{p}_r$ and $\widehat{p}_{r.q}$:

$$cov(\widehat{p}_r, \widehat{p}_{r.q}) = cov(\widehat{p}_r, 1-\widehat{q}_r) = -cov(\widehat{p}_r, \widehat{q}_r) =$$

$$= -cov\left(P_A + \frac{1}{N}\sum_{i\in s}\frac{A_i - RB_i}{\pi_i}, Q_A + \frac{1}{N}\sum_{i\in s}\frac{A_i^c - R_cB_i^c}{\pi_i}\right) =$$

$$= -\frac{1}{N^2}cov\left(\sum_{i\in s}\frac{A_i - RB_i}{\pi_i}, \sum_{i\in s}\frac{A_i^c - R_cB_i^c}{\pi_i}\right) =$$

$$= -\frac{1}{N^2}cov\left(\sum_{i=1}^{N}\frac{A_i - RB_i}{\pi_i}I_i, \sum_{i=1}^{N}\frac{A_i^c - R_cB_i^c}{\pi_i}I_i\right) =$$

$$= -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}cov\left(\frac{A_i - RB_i}{\pi_i}I_i, \frac{A_j^c - R_cB_j^c}{\pi_j}I_j\right) =$$

$$= -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{A_i - RB_i}{\pi_i}\frac{A_j^c - R_cB_j^c}{\pi_j}cov(I_i, I_j) =$$

$$= -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{A_i - RB_i}{\pi_i}\frac{A_j^c - R_cB_j^c}{\pi_j},$$

where I_i is the sample indicator, i.e., $I_i = 1$ if $i \in s$ and $I_i = 0$ otherwise.

We observe that w_{opt} depends on unknown quantities. An estimator of w_{opt} is given by

$$\widehat{w}_{opt} = \frac{\widehat{V}(\widehat{p}_{r.q}) - \widehat{cov}(\widehat{p}_r, \widehat{p}_{r.q})}{\widehat{V}(\widehat{p}_r) + \widehat{V}(\widehat{p}_{r.q}) - 2\widehat{cov}(\widehat{p}_r, \widehat{p}_{r.q})},$$

where

$$\widehat{V}(\widehat{p}_r) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{A_i - RB_i}{\pi_i}\frac{A_j - RB_j}{\pi_j},$$

$$\widehat{V}(\widehat{p}_{r.q}) = AV(\widehat{q}_r) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{A_i^c - R_cB_i^c}{\pi_i}\frac{A_j^c - R_cB_j^c}{\pi_j}$$

and

$$\widehat{cov}(\widehat{p}_r, \widehat{p}_{r.q}) = -\frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{A_i - RB_i}{\pi_i}\frac{A_j^c - R_cB_j^c}{\pi_j}.$$

The estimator of the optimum weight w_{opt} can be used to define an optimum estimator for P_A, which is given by

$$\widehat{p}_{r.opt} = \widehat{w}_{opt}\widehat{p}_r + (1 - \widehat{w}_{opt})\widehat{p}_{r.q}.$$

An alternative expression of the optimum weight w_{opt} is

$$w_{opt} = \frac{R_c - \beta}{R_c - R}, \tag{3}$$

where $R_c = P_A/P_B$,

$$\beta = \frac{cov(\widehat{p}_A, \widehat{p}_B)}{V(\widehat{p}_B)},$$

$$cov(\widehat{p}_A, \widehat{p}_B) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{A_i}{\pi_i}\frac{B_j}{\pi_j}$$

and

$$V(\widehat{p}_B) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{A_i}{\pi_i}\frac{A_j}{\pi_j}.$$

Proof

Variances and the covariance in (2) can be expressed as

$$V_1 = AV(\widehat{p}_r) = V(\widehat{p}_A) + R^2V(\widehat{p}_B) - 2Rcov(\widehat{p}_A, \widehat{p}_B)],$$

$$V_2 = AV(\widehat{p}_{r.q}) = V(\widehat{q}_A) + R_c^2V(\widehat{q}_B) - 2R_ccov(\widehat{q}_A, \widehat{q}_B) =$$

$$= V(\widehat{p}_A) + R_c^2V(\widehat{p}_B) - 2R_ccov(\widehat{p}_A, \widehat{p}_B)$$

and

$$C = cov(\widehat{p}_r, \widehat{p}_{r.q}) = V(\widehat{p}_A) + RR_cV(\widehat{p}_B) - (R + R_c)cov(\widehat{p}_A, \widehat{p}_B).$$

The numerator and the denominator given by (2) can be expressed as

$$V_2 - C = V(\widehat{p}_B)(R_c^2 - RR_c) - cov(\widehat{p}_A, \widehat{p}_B)[2R_c - (R - R_c)] =$$

$$= V(\widehat{p}_B)R_c(R_c - R) - cov(\widehat{p}_A, \widehat{p}_B)(R_c - R) =$$

$$= (R_c - R)[V(\widehat{p}_B)R_c - cov(\widehat{p}_A, \widehat{p}_B)]$$

and

$$V_1 + V_2 - 2C = V(\widehat{p}_B)(R^2 + R_c^2 - 2RR_c) - cov(\widehat{p}_A, \widehat{p}_B)[2R + 2R_c - 2(R + R_c)] =$$

$$= V(\widehat{p}_B)(R_c - R)^2$$

By substituting these expressions into (2) we obtain

$$w_{opt} = \frac{V_2 - C}{V_1 + V_2 - 2C} = \frac{(R_c - R)[V(\widehat{p}_B)R_c - cov(\widehat{p}_A, \widehat{p}_B)]}{V(\widehat{p}_B)(R_c - R)^2} =$$

$$= \frac{V(\widehat{p}_B)R_c - cov(\widehat{p}_A, \widehat{p}_B)}{(R_c - R)V(\widehat{p}_B)} = \frac{R_c - \beta}{R_c - R}.$$

An estimator of expression (3) is

$$\widehat{w}_{opt} = \frac{\widehat{R}_c - \widehat{\beta}}{\widehat{R}_c - \widehat{R}},$$

where

$$\widehat{\beta} = \frac{\widehat{cov}(\widehat{p}_A, \widehat{p}_B)}{\widehat{V}(\widehat{p}_B)}.$$

An alternative and simpler expression for the optimum ratio estimator $\widehat{p}_{r.opt}$ is

$$\widehat{p}_{r.opt} = \widehat{p}_A + \widehat{\beta}(P_B - \widehat{p}_B). \tag{4}$$

Proof

$$\widehat{p}_{r.opt} = \widehat{w}_{opt}\widehat{p}_r + (1-\widehat{w}_{opt})\widehat{p}_{r.q} = \frac{\widehat{R}_c - \widehat{\beta}}{\widehat{R}_c - \widehat{R}}\widehat{R}P_B + \left(1 - \frac{\widehat{R}_c - \widehat{\beta}}{\widehat{R}_c - \widehat{R}}\right)(1 - \widehat{R}_c Q_B) =$$

$$= \frac{\widehat{R}_c\widehat{R}P_B - \widehat{\beta}\widehat{R}P_B}{\widehat{R}_c - \widehat{R}} + \frac{\widehat{R}_c - \widehat{R} - \widehat{R}_c + \widehat{\beta}}{\widehat{R}_c - \widehat{R}}\left(1 - \widehat{R}_c(1-P_B)\right) =$$

$$= \frac{\widehat{R}_c\widehat{R}P_B - \widehat{\beta}\widehat{R}P_B - \widehat{R} + \widehat{R}\widehat{R}_c - \widehat{R}_c\widehat{R}P_B + \widehat{\beta} - \widehat{\beta}\widehat{R}_c + \widehat{\beta}\widehat{R}_c P_B}{\widehat{R}_c - \widehat{R}} =$$

$$= \frac{\widehat{\beta}P_B(\widehat{R}_c - \widehat{R}) + (\widehat{R}_c - 1)(\widehat{R} - \widehat{\beta})}{\widehat{R}_c - \widehat{R}} = \widehat{\beta}P_B + \frac{(\frac{\widehat{q}_A}{\widehat{q}_B} - 1)(\frac{\widehat{p}_A}{\widehat{p}_B} - \widehat{\beta})}{\frac{\widehat{q}_A}{\widehat{q}_B} - \frac{\widehat{p}_A}{\widehat{p}_B}} =$$

$$= \widehat{\beta}P_B + \frac{\frac{1-\widehat{p}_A - 1 + \widehat{p}_B}{1-\widehat{p}_B}\frac{\widehat{p}_A - \widehat{\beta}\widehat{p}_B}{\widehat{p}_B}}{\frac{(1-\widehat{p}_A)\widehat{p}_B - \widehat{p}_A(1-\widehat{p}_B)}{(1-\widehat{p}_B)\widehat{p}_B}} =$$

$$= \widehat{\beta}P_B + \frac{(\widehat{p}_B - \widehat{p}_A)(\widehat{p}_A - \widehat{\beta}\widehat{p}_B)}{\widehat{p}_B - \widehat{p}_A\widehat{p}_B - \widehat{p}_A + \widehat{p}_A\widehat{p}_B} =$$

$$= \widehat{p}_A + \widehat{\beta}(P_B - \widehat{p}_B).$$

We observe that the optimum regression estimator given by (4) coincides with the regression type estimator given in Rueda et al. (2011b). This relationship between the ratio and regression estimators show that the regression type estimator is more efficient than the ratio estimators $\widehat{p}_r$ and $\widehat{p}_{r.q}$.

A variance estimator of the optimum ratio estimator is given by

$$\widehat{V}(\widehat{p}_{r.opt}) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{A_i - \widehat{A}_i}{\pi_i}\frac{A_j - \widehat{A}_j}{\pi_j},$$

where $\widehat{A}_i = \widehat{p}_A + \widehat{\beta}(B_i - \widehat{p}_B)$.

Acknowledgment

This work is supported by the project PYR-2010-20 (PYR-GENIL) of the "Campus de Excelencia Internacional CEI BioTIC GENIL (CEB09-0010)", which belongs to the CEI programme of the "Ministerio de Ciencia e Innovación" in Spain.

REFERENCES

Blyth, C. R., Still, H. A. (1983). Binomial confidence intervals. J. Am. Stat. Assoc. 78: 108-116.

Cohen, G. R., Yang, S. Y. (1994). Mid-p confidence intervals for the Poisson expectation. Stat. Med. 13: 2189-2203.

Fleiss, J. L., Levin, B., Paik, M. C. (2003). Statistical methods for rates and proportions 3rd edn. Wiley, New Yersey.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat. Med. 17: 857-872.

Rueda, M.M., Muñoz, J.F., Arcos, A., Álvarez, E. (2011a). Estimators and confidence intervals for the proportion using binary auxiliary information with applications to the estimation of prevalences. J. Bioph. Stat. 21, 526-554.

Rueda, M.M., Muñoz, J.F., Arcos, A., Álvarez, E. (2011b). Indirect estimation of proportions in natural resource surveys. Mathematics and Computers in Simulation. In press.

Särndal, C.E., Swensson, B. Wretman, J. (1992) *Model Assisted Survey sampling.* Springer Verlag.