Far East Journal of Theoretical Statistics
Volume 29, Number 2, 2009, Pages 129-136
Published Online: December 16, 2009
This paper is available online at http://www.pphmj.com

ESTIMATING PROPORTIONS WITH UNEQUAL SAMPLING PROBABILITIES: THE BASU'S ELEPHANT PROBLEM REVISITED

JOSE M. PAVÍA

Departamento de Economía Aplicada
Facultad de Economía
Area de Métodos Cuantitativos
Universidad de Valencia
Campus Els Tarongers
46022 - Valencia, Spain
e-mail: pavia@uv.es

Abstract

When sampling a finite universe with unequal selection probabilities, the unbiased estimators recommended can yield inadmissible outcomes. Although Basu early warned about this weakness of the Horvitz-Thompson estimation method, the very unrealistic Basu's example rapidly relaxed statisticians and this crucial detail was almost forgotten and systematically omitted in textbooks. In predicting proportions, however, this shortfall is more visible and can appear even with reasonable sampling plans. This paper shows it first using an unrealistic laboratory design and later through a real data application. After mentioning a feasible solution, this note proposes to open a debate and suggests return to survey-sampling and inference foundations on the search for a solution that can widely be accepted.

2000 Mathematics Subject Classification: 62D05, 62A01.

Keywords and phrases: Horvitz-Thompson estimator, finite populations, multi-stage sampling, sampling with replacement, sampling without replacement, unequal probability sampling.

This research was supported by the Spanish MICINN project CSO2009-11246.

Received July 22, 2009

1. Introduction

It is well known that we statisticians have a special fondness for unbiased and linear estimators. Thus, it is not surprising that after Horvitz and Thompson [9] discovered an unbiased linear estimator to calculate population means and totals when sampling finite universes without replacement with unequal selection probabilities and, later, Godambe [5] showed that no unbiased estimate of the population mean exists with minimum variance for all populations within Horvitz-Thompson's main classes of linear competitors, the Horvitz-Thompson estimator (HT) was broadly adopted (e.g., Kish [11] and Ogus [12]). And despite Basu [2, pp. 212-213) rapidly showing its weaknesses, from that point onwards, researchers turned to proposing different rules for calculating admissible unbiased estimates of the sampling error and to developing more efficient sampling systems with an optimal choice of selection probabilities.

The straightforward logic behind HT, which weighs each drawn value inversely to its probability of being selected, triggered the use of statistics based on HT being adopted in fields as diverse as forestry, medicine or economics and almost all sampling books recommend statistics derived from HT in multi-stage designs without replacement (where the chief use of unequal-probability sampling emerges). Unfortunately, this strategy leads in a more than negligible number of cases to illogical estimates when approximating proportions in both single and multi-stage sampling designs. Apparently, however, many authors' textbooks have forgotten to warn practitioners about this fact (see, e.g., Ardilly and Tillé [1], Cochran [3], Pérez López [14], Särndal et al. [17], Thompson [20] and Tryfos [21]). This paper illustrates this by means of (i) a very simple single-stage simulated example and (ii) with real data from a two-stage cluster sampling design and provides an alternative tentative solution in the aim of opening a debate in search of a broadly accepted answer to this issue.

Suppose a sample, S_n, of n units must be chosen without replacement from a population of N zeros and ones $(y_1, y_2, \cdots, y_N)$, such that the probability of the jth unit being drawn is π_j $(j = 1, 2, \cdots, N)$. It is well known that the HT estimator for the proportions of ones in the population is given by equation (1),

$$\hat{p} = \frac{1}{N} \sum_{j \in S_n} \frac{y_j}{\pi_j}. \tag{1}$$

As example of multi-stage sampling design, consider a simple two-stage sampling design in which the *N* units are divided into *K* clusters of N_i units. In the first-stage *k* clusters are chosen without replacement with probabilities π_i $(i = 1, 2, \cdots, K)$ of being selected and in the second-stage n_i units are randomly selected in each one of the *k* clusters selected in the first-stage. Then, in these circumstances, the HT estimator for the proportions of ones in the population is given by equation (2), where y_{ij} is the value observed for the *j*th unit drawn in *i*th cluster and $\hat{p}_i$ is the proportion of ones estimated in *i*th cluster,

$$\hat{\hat{p}} = \frac{1}{N}\sum_{i=1}^{k}\frac{N_i}{\pi_i}\sum_{j=1}^{n_i}\frac{y_{ij}}{n_i} = \sum_{i=1}^{k}\frac{N_i}{N\pi_i}\hat{p}_i. \tag{2}$$

It is easy to prove that both, $\hat{p}$ and $\hat{\hat{p}}$, are unbiased estimators of the proportion of ones in the sample. The reason is simple: the expected value of the sum of weights is one in both equations, $E\left(\sum_{j\in S_n}\frac{1}{N\pi_j}\right) = E\left(\sum_{i=1}^{k}\frac{N_i}{N\pi_i}\right) = 1$. For instance, defining α_j the sample inclusion indicator function (i.e., $\alpha_j = 1$, if *j*th unit being drawn, and 0, otherwise) and writing the estimator as a sum over the population, it follows straightforward that:

$$E(\hat{p}) = E\left(\frac{1}{N}\sum_{j\in S_n}\frac{y_j}{\pi_j}\right) = \frac{1}{N}E\left(\sum_{j=1}^{N}\alpha_j\frac{y_j}{\pi_j}\right)$$

$$= \frac{1}{N}\sum_{j=1}^{N}E(\alpha_j)\frac{y_j}{\pi_j} = \frac{1}{N}\sum_{j=1}^{N}\pi_j\frac{y_j}{\pi_j} = p.$$

The fact that the expected value of the sum of weights is one, however, does not imply that weights not adding up to one can be obtained for a particular sample and that, therefore, single proportion estimates higher than one or groups of proportion estimates whose aggregation exceeds unity can occur.

2. Examples

In the first example, an experiment is simulated in which two units from a universe of zeros and ones and size five are drawn without replacement and unequal

selection probabilities in order to estimate the proportion of ones in the population. Although it is a very unrealistic example, it is useful to show in a simple way that illogical estimates can be obtained with the HT estimator. The second example takes real data from an electoral survey performed in one of the last regional elections held in Spain to show that if the theoretically recommended HT estimator had been used, it would have resulted in party shares adding up to more than 100 percent.

2.1. A laboratory design

Let $U = \{u_1, u_2, u_3, u_4, u_5\} = \{1, 1, 0, 0, 1\}$ be a universe of size five of zeros and ones. Suppose a sample of two units is chosen without replacement from U, where the probabilities of being selected in the first extraction are, respectively, $\{.5, .2, .1, .1, .1\}$ and in the second extraction the probabilities of selection are re-weighed according to the first extraction. Thus, the probability (adjusted to three decimals) of each unit being drawn is, respectively, $\{.792, .467, .247, .247, .247\}$; and, therefore, the HT estimate that would be obtained for each one of the ten possible samples would be those presented in Table 1.

Table 1. Estimates for the proportion of ones in the population using the HT estimator

Sample	HT estimate	Sample probability	Sample	HT estimate	Sample probability
$\{u_1, u_2\} = \{1, 1\}$	0.6812	0.3250	$\{u_2, u_4\} = \{1, 0\}$	0.4286	0.0472
$\{u_1, u_3\} = \{1, 0\}$	0.2526	0.1556	$\{u_2, u_5\} = \{1, 1\}$	**1.2376**	0.0472
$\{u_1, u_4\} = \{1, 0\}$	0.2526	0.1556	$\{u_3, u_4\} = \{0, 0\}$	0.0000	0.0222
$\{u_1, u_5\} = \{1, 1\}$	**1.0616**	0.1556	$\{u_3, u_5\} = \{0, 1\}$	0.8090	0.0222
$\{u_2, u_3\} = \{1, 0\}$	0.4286	0.0472	$\{u_4, u_5\} = \{0, 1\}$	0.8090	0.0222

As can be deduced from Table 1, although (according to theory) the expected value of HT estimates fits the actual population mean (0.6), the use of the HT estimator would produce illogical estimates for two of the samples ($\{u_1, u_5\}$ and $\{u_2, u_5\}$), which together have more than a 20% probability of occurring.

2.2. A real data example

On 27th May 2007 citizens from several Spanish regions were called to renew their regional parliaments. Divided into three constituencies (Alacant, Valencia and Castello), the voters from the Valencia region had to choose ninety-nine

representatives. In each constituency, seats are distributed using the d'Hondt rule (see, e.g., Pavía-Miralles [13]) among the parties receiving the greatest support. According to survey polls, for this election only three parties had real chances of obtaining seats in the parliament: the conservative party (PP), the socialist party (PS) and a coalition between the communist party and a leftwing nationalist-regionalist party (CpP).

On election day, an exit-poll was conducted to advance final outcomes. In each constituency, a two-stage cluster sampling design was performed to collect data. In the first stage, a number of precincts (with all the precincts having the same probability of being selected) were randomly selected without replacement among all the precincts of the constituency. In the second stage, a pollster was sent to interview a sample of voters from each of the selected precincts. In practice, however, the interviews take place outside the voting building and the voters interviewed come from the several precincts located in the corresponding polling place. Thus, actually the primary units selected in the first stage were polling places and, due to the number of precincts varying among stations, they had unequal probabilities (albeit approximately proportional to their size) of being selected.

In the constituency of Castello, the 287, 427 voters were distributed to cast their ballots in 729 precincts, divided into 343 locations. With the sampling plan devised above, twenty-five of these locations were selected in Castello to collect data. According to the theory, the two-stage statistic given by equation (2) is the best linear unbiased estimator to predict the proportions of votes that each party would reach. However, if that statistic had been used, the estimates would have been: PP 52.86%, PS 41.90%, and CpP 12.17%, which sum up to more than 100% (and this without adding up the estimates corresponding to the rest of the parties and null and blank votes!). Similarly, illogical results would have been obtained in the constituency of Valencia (1,862, 566 voters), but not in Alacant (1, 168, 358 voters) because in that case the sum of the weights was virtually one: 0.995.

3. Discussion

In almost all the methods of sampling based on multi-stage survey designs, the population units are drawn with unequal probabilities of selection. It is usual, therefore, to construct population parameter estimators as weighted functions of the observed values, where the probabilities of selection are incorporated both implicitly

and explicitly into the estimator functions in order to balance observations. Usually the weights are proportional to strata or cluster size and therefore, their sum adds up to one and no problems arise. However, when the sum of weights differs from unity, incongruent estimates can become visible when predicting proportions.

One obvious answer would be to take the HT estimator as a reference and to recalculate its weights to sum one. This strategy, which in the case of equation (2) would yield equation (3), involves, to a certain extent, recovering Godambe's old idea (Godambe [5]) of considering that the three elements (observations, labels and sample) contain information and should be taken into account when constructing population estimators. In particular, he proposes to derive the estimators from the likelihood function associated to the complete sampling design after imposing some appropriate criteria,

$$\hat{\bar{p}}^{*} = \frac{\sum_{i=1}^{k} \frac{N_i}{\pi_i} \hat{p}_i}{\sum_{h=1}^{k} \frac{N_h}{\pi_h}}. \tag{3}$$

The solution given in (3), however, has two main drawbacks: (i) it needs a theoretical justification and; (ii) it would require developing a mathematical expression for the variance of this estimator (it must be noted that the weights are random as they depend on the particular units drawn). The ratio adjustment approach may perhaps offer a solution for the first drawback – in fact, the use of ratio estimators was the first approach proposed (Hajek [7]) in order to fix this problem – and, on the other hand, the second shortfall could be overcome, the level of complexity depending on the design of the sample, by using subsample replication techniques (see, e.g., Woodroof [22]).

Alternatively, although finite population sampling is possibly the only area of statistics where the primary mode of analysis is based on randomization distribution rather than on statistical models (Smith [18] and Smith [19]), it might be worth taking into account the specific characteristics of the kind of universes dealt with when estimating proportions and to suggest models that make it possible to incorporate both the sample design and some parametric assumptions to generate reliable and efficient estimators (Särndal [16]), reopening, for this particular problem, the old debate between design based inference and model based inference (e.g., Royall [15]).

A different approach could be to adopt Fisher's position, according to which randomization is relevant before the data are collected, but not in the analysis of the data (Fisher [4]). This would imply abandoning Neyman's paradigm of building estimators based on all possible random samples that could be drawn (Neyman [10]). After all, what the user of the survey requires is an efficient estimate together with a measure of its accuracy for the particular survey in question.

The examples and posterior discussion herein will hopefully arouse interest and lead others to investigate sampling systems of this type in order to reach a solution that can generally be accepted. It must be noted, likewise, that while the examples displayed deal with proportions and sampling without replacement, this problem is not exclusive to them. First, the estimator proposed by Hansen and Hurwitz (Hansen and Hurwitz [8]) to sample with replacement has similar limitations. Secondly, the use of the HT estimator to estimate totals and means could yield a clear overestimation in whatever population.

Acknowledgements

I wish to thank Carl Särndal and Santiago Murgui for their answers to my queries about the issue and Generalitat Valenciana for their help in providing the detailed exit-poll raw data. The usual disclaims apply.

References

[1] P. Ardilly and Y. Tillé, Sampling Methods: Exercises and Solutions, Springer, New York, 2006.

[2] D. Basu, An essay on the logical foundations of survey sampling, Foundations of Statistical Inference, Godambe and Sprott, eds., Holt, Rinehart and Winston, Toronto, 1971, pp. 203-233.

[3] W. G. Cochran, Sampling Techniques, John Wiley & Sons, New York, 1977.

[4] R. A. Fisher, Statistical Methods and Scientific Inference, Oliver and Boyd, Edinburgh, 1956.

[5] V. P. Godambe, A unified theory of sampling from finite populations, J. Roy. Statist. Soc. Ser. B 17 (1955), 269-278.

[6] V. P. Godambe, Foundations of survey-sampling, Amer. Statist. 24(1) (1970), 33-38.

[7] J. Hajek, Comments to "An essay on the logical foundations of survey sampling", Foundations of Statistical Inference, p. 236, Godambe and Sprott, eds., Holt, Rinehart and Winston, Toronto, 1971.

[8] M. H. Hansen and W. N. Hurwitz, On the theory of sampling from finite populations, Ann. Math. Statistics 14 (1943), 333-362.

[9] D. G. Horvitz and D. J. Thompson, A generalization of sampling without replacement from a finite universe, J. Amer. Statist. Assoc. 47 (1952), 663-685.

[10] J. Neyman, On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, J. R. Stat. Soc. 97 (1934), 558-625.

[11] L. Kish, Survey Sampling, John Wiley & Sons, New York, 1965.

[12] J. L. Ogus, A note on the "Necessary Best Estimator", J. Amer. Statist. Assoc. 64 (1969), 1350-1352.

[13] J. M. Pavía-Miralles, Forecasts from nonrandom samples: the election night case, J. Amer. Statist. Assoc. 100 (2005), 1113-1122.

[14] C. Pérez López, Muestreo Estadístico, Conceptos *y* Problemas Resueltos, Pearson, Madrid, 2005.

[15] J. Royall, Linear regression models in finite population sampling theory, Foundations of Statistical Inference, Godambe and Sprott, eds., Holt, Rinehart and Winston, Toronto, 1971, pp. 259-279.

[16] C. E. Särndal, The calibration approach in survey theory and practice, Survey Methodology 33 (2007), 99-119.

[17] C. E. Särndal, B. Swensson and J. Wretman, Model Assisted Survey Sampling, Springer-Verlag, New York, 1992.

[18] T. M. F. Smith, The foundations of survey sampling: a review, J. Roy. Statist. Soc. Ser. A 139 (1976), 183-204.

[19] T. M. F. Smith, Sample surveys 1975-1990: an age of reconciliation?, Int. Stat. Rev. 62 (1994), 5-19.

[20] S. K. Thompson, Sampling, Wiley, New York, 2002.

[21] P. Tryfos, Sampling Methods for Applied Research, John Wiley & Sons, New York, 1996.

[22] J. Woodroof, Bootstrapping: as easy as 1-2-3, J. Appl. Statist. 27 (2000), 509-517.