Journal of Official Statistics
Vol. 7. No. 3, 1991, pp. 367–373
© Statistics Sweden

Reconciling Fixed Interview Workloads With Self-Weighting Sampling When Size Measures are Defective

Chris Scott[1] *and Ben Amenuvegbe*[2]

Abstract: When probability proportional to size sampling is used at the area sampling stage, a self-weighting sample can be achieved with a fixed take of households in each area. This convenient result breaks down if the size measures are imperfect. The paper shows how it is still possible to retain the advantages of self-weighting and fixed interviewer workloads by allowing the number of interviewers per area unit to become a random variable.

Key words: Sampling; self-weighting; developing country surveys.

1. Introduction

A classic, and often used, sample design consists of selecting area units with probability proportional to size, then selecting a fixed number of households in every unit. This provides a household sample which combines two advantages: the interview workload in each area is constant and the household sample is self-weighting.

Fixed interview workloads are popular with survey organizers. They simplify the planning and supervision of field work, particularly where repeat visits to households are required after a fixed time interval, as in many third world living standards surveys (Ainsworth and Muñoz 1986). Moreover an exactly constant workload reduces the temptation for the interviewer to over-report non-response or to under-enumerate households in the listing operation.

Self-weighting is desirable, particularly in developing countries, because of its simplicity. Anyone who has worked in a developing country's statistical office is likely to have noticed how often confusion is caused by weights. They have to be individually computed, checked, standardized, input into the data file, used or not used according to the nature of the analysis, and explained in the file documentation and the survey report. Users may need weighted frequencies to enable them to re-group categories, but at the same time many will want to know the actual, unweighted n's when shown data from a sample. All of these are small matters in a well staffed and efficient statistical office. But in many offices in developing countries they are not negligible.

These considerations account for the popularity of the above simple design.

[1] Consultant, 60 Highgate Hill, London N19 5NQ, UK.
[2] Australian National University.

Acknowledgement: The authors thank the Associate Editor for helpful suggestions and the Ghana Statistical Service for permission to quote data on the Ghana Living Standards Survey.

Unfortunately the system works, in the sense of reconciling a fixed cluster take with self-weighting, only if accurate size measures are available in advance of sample selection for every area unit. In developing countries, in particular, this condition is not easily satisfied (Scott 1987; World Fertility Survey 1975). The only available size measures are likely to come from the latest census; even if accurate at the time of their collection, they go quickly out of date.

In this paper we describe a modification of the above standard design which maintains both constant workloads and self-weighting even when the initial measures of size are defective. The method seems to be new and is being applied for the first time at the authors' suggestion in the Living Standards Survey in Ghana.

The method is proposed not on grounds of greater sampling efficiency but simply as a procedure which assists the survey organizer in a developing country to achieve two convenient organizational features simultaneously without serious loss of sampling efficiency.

2. The Standard Model

We assume a two-stage sample with area units (AUs) in the first stage and households in the second stage.

We shall use Σ_u to denote summation over the universe and Σ_s summation over the sample.

If $\{p_{1i}\}$ is the set of first-stage selection probabilities and the ith AU has size-measure M_i, then

$$p_{1i} = kM_i \tag{1}$$

where k is a constant. If p_{1i} is interpreted as the probability of unit i being included in the sample, k is equal to a/M where a is the number of AUs selected and $M = \Sigma_u M_i$. Thus

$$p_{1i} = aM_i/M. \tag{2}$$

In the standard model M_i is exactly the number of households in the ith AU.

If we now select m_i households with equal probability from among the M_i in the ith AU, the second-stage (conditional) probability in that AU will be

$$p_{2i} = m_i/M_i. \tag{3}$$

It follows from (2) and (3) that the overall inclusion probability for any household in the ith AU will be

$$f = p_{1i}p_{2i} = am_i/M. \tag{4}$$

For self-weighting, we require f constant. Equation (4) shows that this is attained if m_i is constant ($= b$, say). If we place one interviewer in each sample AU this arrangement will provide fixed interviewer workloads equal to b. Thus the conditions of fixed workload and self-weighting are satisfied simultaneously, with $f = ab/M$. However, we require a knowledge of M_i for every AU in the sampling frame prior to sampling.

3. Modification of the Standard Model

Let M_i' be the current number of households in the ith AU. This may not be equal to the census figure M_i.

If we continue to use the same method in these circumstances equation (4) becomes

$$p_{1i}p_{2i} = \frac{am_i}{M}\frac{M_i}{M_i'} \tag{5}$$

and with fixed $m_i = b$ we obtain as the overall probability for households

$$f_i = p_{1i}p_{2i} = \frac{ab}{M}\frac{M_i}{M_i'}. \tag{6}$$

If self-weighting were still assumed, we would be under-weighting AUs that had grown abnormally since the measures of size were determined, and over-weighting those that had grown less than the average or contracted. Such growth in the number of households could well be associated with economic variables and the resulting bias might be appreciable, especially in an economic survey.

It has usually been assumed that this leaves only two choices: to maintain self-weighting but abandon the fixed workload, or to drop the self-weighting requirement. If self-weighting is to be retained we need $p_{1i}p_{2i} = f$, constant. With $p_{2i} = m_i/M_i'$ this will require a sample take of

$$m_i = M_i'p_{2i} = M_i'f/p_{1i}$$
$$= \frac{Mf}{a}\frac{M_i'}{M_i} \tag{7}$$

If we write q_i for m_i/b, where b is now the *target* workload size, equation (7) becomes

$$q_i = \frac{Mf}{ab}\frac{M_i'}{M_i} \tag{8}$$

The quantity q_i is then the number of sample workloads that would be required to yield a self-weighting sample. Our problem arises because q_i will not, in general, be equal to 1.

However, a closer look at the practical constraints suggests a solution. In reality we do not need to insist on one workload *per AU* as long as we have one workload *per interviewer*. By allowing more than one interviewer, or no interviewer, in a given AU we are able to weaken the constraint on q_i. We now require only that q_i be an integer. This suggests the solution of replacing q_i by an integer q_i' selected in such a way that its expectation is equal to q_i as given by (8). This will maintain self-weighting with a fixed workload per interviewer, although the number of interviewers allocated to an AU will be variable (and possibly zero). Incidentally, instead of placing two interviewers in one AU we may prefer to have the same interviewer stay for two survey periods in the same AU.

One common feature of developing-country sampling makes this solution particularly attractive. Although the only size measures for AUs available before sampling are likely to be quite imprecise, an accurate up-date is routinely available for the selected AUs after sampling. This arises because, in the absence of any address lists or household lists, it is standard practice in most developing countries to conduct a household listing operation in the selected AUs to provide a sampling frame for household selection. Thus the M_i' will be made available routinely. This makes it possible to compute the values q_i after area selection and before household selection.

Given q_i, the desired determination of the integer q_i' can be made by the following procedure. If q_i lies between integers e and $e + 1$ we make a choice between e and $e + 1$ with probability $e + 1 - q_i$ of choosing e, and probability $q_i - e$ of choosing $e + 1$. It is easily verified that this gives an expected value of q_i. However, such a procedure, if carried out independently for each i, would be burdensome and yield an unnecessarily variable sample size. A more convenient method is to use a form of systematic sampling with probability proportional to q_i. We list the selected AUs, each with its q_i, cumulate the q_i values, and select at a fixed interval among the cum q_i. Any AU against which a sample point falls is considered "re-selected." For each re-selection we count one sample workload of just b households. The expected number of workloads allocated to the ith AU will then be proportional to the calculated q_i. The example in Section 3 may clarify the procedure.

The sample design may be summarized as

follows:

First stage	Probability of inclusion of *i*th AU in sample: $p_{1i} = aM_i/M$.
Re-selection	Expected number of selections (hence of workloads allocated) for *i*th AU conditional on 1st stage: q_i/I, where I = systematic sampling interval used.
Households	Conditional selection probability: $p_{2i} = b/M_i'$.

Since these three operations are independent, the overall selection probability for households is the product of the three expectations, namely

$$\frac{aM_i}{M}\frac{q_i}{I}\frac{b}{M_i'} \qquad (9)$$

or, substituting for q_i from (8),

$$f/I. \qquad (10)$$

Since this is constant, the sample is self-weighting.

If we select with interval $I = 1$ every household is ultimately selected with probability f. In the example below, selection is made with this value of I.

EXAMPLE

(1)	(2)	(3)	(4)	(5)	(6)	(7)
AUs selected	Census households M_i	Households listed M_i'	$q_i = \frac{M}{M_0}\frac{M_i'}{M_i}$	Cumulative q_i	Sampling sequence	Number of households for interview
001	120	140	1.06	1.06	0.20	16
002	200	286	1.30	2.36	1.20; 2.20	32
003	170	157	0.84	3.20	3.20	16
004	141	152	0.98	4.18		–
005	198	196	0.90	5.08	4.20	16
.	.	.	.	.	.	.
.	.	.	.	.	.	.

NOTES

Col. (4) Total census households: M = 2,600,000
Total households, current estimate: M_0 = 2,860,000
Formula for q_i is (12), with $a' = a$.

Col. (6) Sampling interval: I = 1.00
Random start: A = 0.20
Sampling sequence: $A + rI$, $r = 0, 1, 2\ldots$
Each term of the sampling sequence is entered against the AU whose cum q_i is the first to equal or exceed it.

Col. (7) The number of terms of the sequence entered against an AU indicates the number of workloads to be allocated to it.
Workload size: b = 16.

4. Sample Size

Let a' be the number of sample workloads desired (which need not necessarily be chosen equal to a, the number of AUs selected).

We have already assumed that the universe total $M = \Sigma_u M_i$ is available. However we will not normally know the universe total of M_i'. Suppose that our best estimate for this, presumably from some source outside the survey, is M_0. Then the sampling rate may be estimated as the ratio of the target sample size to M_0:

$$f = a'b/M_0. \tag{11}$$

Substituting this value for f in (8), we obtain

$$q_i = \frac{a'}{a}\frac{M}{M_0}\frac{M_i'}{M_i}. \tag{12}$$

Since all these quantities are available we have a fixed selection probability calculable in advance and all estimates are self-weighting, whether of population totals or means. By cumulating the q_i and selecting at interval 1 in the cum q_i column we obtain an expected number of workloads equal to the target figure a' and the overall probability of a household's inclusion in the sample will be that given in (11). (In practice it will be simpler to cumulate M_i'/M_i instead of q_i and select at interval $(a/a')(M_0/M)$ instead of 1.) However, in estimating totals there will be a bias in so far as M_0 differs from $\Sigma_u M_i'$.

An objection that might be raised to the method proposed is that we are obtaining self-weighting only by deliberately ignoring our knowledge of the actual values q_i': we could make better inferences by conditioning on the q_i'. This is true but the method proposed is nevertheless legitimate: our aim is not to achieve the best possible estimate but to present an unbiased procedure that does not involve weighting and that yields a constant interviewer workload.

5. Sampling Error and Costs

Survey efficiency and sampling efficiency are not the same thing. The reasons cited in Section 1 for preferring fixed workloads and self-weighting do not include sampling efficiency, though they may be none the less pressing for that. Even where sampling efficiency is not the main motive, however, the survey organizer will wish to know the sampling error implications of adopting the "re-selection" method.

We need first to specify a standard design against which to compare sampling errors for a given cost. We propose to compare the re-selection (RS) method, with $a' = a$, to the following standard design: the same number a of AUs is selected, with probabilities proportional to M_i, with a fixed take of b households in each AU and with weights of M_i'/M_i applied in the estimation. We may term this standard of comparison "fixed-take weighted" (FW). Here the main elements of the cost are kept constant in the comparison: the household listing operation covers the same number of AUs and the expected household sample sizes are the same in both cases. Thus it only remains to compare the sampling errors of the two designs.

We make two simplifying assumptions about the RS design:

a. We assume that all AUs receive either 0, 1 or 2 workloads, ignoring the rare case of 3 or more. This assumption is made for simplicity of presentation; it is obviously not crucial.

b. We assume that the population variance S^2 and the intra-class correlation ρ for the variable under study do not vary systematically with the number of workloads allocated to the AU.

We may now regard the sample as made up of two post-strata:

Stratum 1: (1-2h)a AUs selected, with b households selected in each.

Stratum 2: ha AUs selected with 2b households selected in each.

For the estimate of a mean $\bar{Y}$ we have:

$$\hat{\bar{Y}} = (1\text{-}2h)\bar{Y}_1 + 2h\bar{Y}_2 \qquad (13)$$

and

$$\sigma^2(\bar{Y}) = (1\text{-}2h)^2\sigma^2(\bar{Y}_1) + (2h)^2\,\sigma^2(\bar{Y}_2) \qquad (14)$$

ignoring the contribution to the variance due to sampling variability in h.

The within-stratum variance can be obtained in terms of S^2 and ρ by the well known formula (Kish 1965)

$$\sigma^2 = \frac{S^2}{ab}[1 + (b - 1)\rho] \qquad (15)$$

where b secondary units are selected in each of a primary units. Substituting this in (14) for each stratum, using the stratum values given above for the a's and b's, we obtain after simplification:

$$\sigma^2(\bar{Y}) = \frac{S^2}{ab}[1 + (b - 1)\rho] + \frac{S^2}{a}2h. \qquad (16)$$

The second term here is the increase in variance over that of a design using a fixed take b with no weights applied (FNW).

Turning to the standard-of-comparison design FW, the need to weight the estimate will increase the variance (relative to FNW) by a loss factor $L = C^2(w)$, where $w_i = M_i'/M_i$ and $C(w)$ is the ratio of the standard deviation of the w_i to their mean. Again we are assuming that the weights are not related systematically to the population variance or the intraclass correlation. (Kish 1965.) Thus, the variance increases relative to FNW and will be:

For the RS design: $2b\rho h/[1 + \rho(b\text{-}1)]$

For the FW design: $C^2(w)$.

Some typical values might be $b = 16$, $h = 0.1$, and $C^2(w) = 0.08$. This makes the RS design more efficient than FW for any variable for which $\rho < 0.04$. With $b = 8$ instead of 16, this changes to $\rho < 0.08$. In practice ρ values vary widely by variable and by country (see for example Verma, Scott and O'Muircheartaigh 1980, p. 449), but the great majority fall between 0.01 and 0.1. It seems fair to conclude that sampling efficiency, in the traditional restricted sense of sampling error per unit cost, does not argue unequivocally for or against the re-selection method in most practical applications.

For given values of a and b the relative variance for the RS design increases with h while that for the FW design increases with $C^2(w)$. Note that $C^2(w)$ is approximately proportional to h, because the q distribution is roughly poissonian. Thus outdatedness or inaccuracy in the size measures will have approximately the same proportional effect on both methods as regards sampling precision.

For estimation of the error variance after the survey we can regard the initial primary stage selection and the reselection as a single process of sampling with replacement with probability $a'M_i'/M_0$. This leads to the standard variance estimate (Cochran 1977, p. 307)

$$\sigma^2(\bar{Y}) = \frac{1}{n'(n' - 1)}\sum_{i=1}^{n'}(\bar{y}_i - \bar{y})^2 \qquad (17)$$

where $\bar{y}_i$ is the element mean in the ith AU, $\bar{y}$ is the whole-sample mean, and n' is the number of distinct AUs selected, that is, a(1-h). This method can be applied within the first stage strata used in sampling.

6. Application in Ghana

In the first application of the method in Ghana, with $a = a' = 200$, $b = 16$, $M_0 =$

$M = 2{,}444{,}836$, the expected sample size was 3,200 and the sample size obtained by summing the q_i across the AU sample was 3,458. The sampling interval was adjusted to yield exactly a' workloads.

Using a census that was three years old at the time of the household listing, the number of workloads allocated per AU among the initial sample of 200 AUs was:

Workloads per AU	No. of AUs
0	22
1	157
2	20
3	1
	200

As the years pass after the census, this distribution would presumably spread gradually.

Where more than one workload was allocated the team stayed in the AU for more than one cycle, rather than two or more teams being sent to the same AU at the same time. No special problems appeared to arise in the field from these arrangements.

7. References

Ainsworth, M. and Muñoz, J. (1986). The Côte d'Ivoire Living Standards Survey: Design and Implementation. Living Standards Measurement Survey Working Paper No. 26, Washington, D.C.: World Bank.

Cochran, W.G. (1977). Sampling Techniques. 3rd ed. New York: John Wiley.

Kish, L. (1965). Survey Sampling. New York: John Wiley.

Scott, C. (1987). Sampling Manual, Demographic and Health Surveys. Basic Documentation-8. Columbia, Maryland: Institute for Resource Development.

Verma, V., Scott, C., and O'Muircheartaigh, C. (1980). Sample Designs and Sampling Errors for the World Fertility Survey. Journal of the Royal Statistical Society, Ser. A, 143, 431–473.

World Fertility Survey (1975). Manual on Sample Design. The Hague: International Statistical Institute.

Received March 1988
Revised February 1991