CLUSTER OPTIMAL SAMPLE SIZE FOR DEMOGRAPHIC AND HEALTH SURVEYS

Alfredo Aliaga and Ruilin Ren
ORC Macro, Inc., United States
alfredo.aliaga@orcmacro.com

For practical purposes and simplicity, the sample design used in the Demographic and Health surveys is a two-stages clustered sample. In general, the sampling frame is a complete list of enumeration areas (EAs) created in a recent population census (around 100 households per EA). In a second stage, a prefixed number of households is selected from each EA. All household members (all women ages15-49 in particular) are selected for interviewing. This presentation looks at nearly optimum sample sizes and compares them with different situations. The results show that for most of the Demographic and Health surveys conducted, the sample size per cluster met the optimum size with a tolerable precision loss.

INTRODUCTION

The *Demographic and Health Surveys* (DHS) program of *Macro International, Inc.* (Macro), is a worldwide project initiated by the *U.S. Agency for International Development* (USAID). Throughout the past 20 years, the program has implemented and/or provided technical assistance in about 80 countries for about 200 surveys in Africa, South America, Southeast Asia, and West Asia. We have accumulated many experiences from the practice in every major step of a survey, from sampling design, to data collection, to data analysis. Experience tells us that with a two-stage sample, the second-stage sample size (i.e., the number of women to be selected in each cluster) within the range of 20–30 women per cluster is good for most of the survey indicators covering contraception prevalence, fertility preferences, infant and child mortality, and knowledge and behavior regarding sexually transmitted infections. However, theoretical evidence of this survey practice is important and must be provided. In this paper, we present some of our research results concerning optimal sample sizes in DHS among different population situations. The results prove that the empirical second-stage sample size in DHS met the optimal request in terms of cost and precision, or was within the tolerable limits in terms of relative precision loss.

All of the DHS conducted and/or assisted by Macro are in underdeveloped countries where statistics are underdeveloped, too. Regarding the usually outdated sampling frame (in most of the cases) and various difficulties in implementing the survey, Macro's sampling policy is to use simple sampling design that facilitates exact implementation and easy control of the fieldwork. Based on this policy, most of the DHS are stratified, two-stage cluster samplings. In the first stage, a number of *Primary Sampling Units* (PSU) are selected from a frame list with probability proportional to a size measure; in the second stage, a prefixed number of households (or residence dwellings) are selected from a list of households obtained in an updating operation in the selected PSU. A PSU is usually a geographically constructed area, or a part of the area, called *Enumeration Area* (EA) that contains a number of households created in the last population census. In the majority of cases, a complete list of the EAs is available with basic information on their geographical location, rural-urban area, total population, total number of households, etc. Also included are cartographic materials delimitating the boundaries of the EAs. However, in most of the cases, regarding the length between two population censuses (usually 10 years), the important information concerning the size measures of the EA (e.g., number of households residing in the EA) needs to be updated. The updating operation consists of listing all the households residing in the selected EAs and recording for each household basic information such as name of household head, street address, and type of residence, etc. The procedure provides a complete list of the households residing in the selected EAs, which serves as the sampling frame for the second stage's sampling for household selection.

Regarding the cost of the listing operation, it is impossible to do the listing for the whole sampling frame. A routine exercise of Macro is to conduct the listing operation only on the EAs selected in the first stage. Even the cost of listing the selected EAs represents a major survey cost. Therefore, it becomes important to decide the number of clusters to be selected and the number of individuals (hereafter abbreviated as "sample take") to be interviewed in each cluster in the

sampling design stage in such a way that the desired survey precision and the available total survey budget are met. The solution depends on the cost ratio and the intracluster correlation. The cost ratio represents the relative cost of interviewing a cluster (mainly including the cost of household listing and the cost of traveling between clusters for household listing and for individual interviews) to the cost of interviewing an individual (mainly including the interview cost and the travel cost within the cluster). The cost ratio varies from country to country, depending on the population density, the level of urbanization, and the infrastructure of the country. When the cost ratio is high, it means that the between-clusters travels are expensive and it is desirable to select fewer clusters and interview more individuals per cluster. On the contrary, it is desirable to select more clusters and interview fewer individuals per cluster in order to achieve better precision. Apart from the cost ratio, the intracluster correlation on survey characteristics plays an important role in determining the second stage's sample size. The intracluster correlation measures the similarity of the individuals on the survey characteristic within cluster. A high intracluster correlation means strong similarities between the individuals within the same cluster; therefore, a large sample take per cluster will decrease the survey precision. While a low intracluster correlation means weak similarities between the individuals within the same cluster, a large sample take will decrease the survey cost. The optimal sample take is a function of the cost ratio and the intracluster correlation.

OPTIMAL SAMPLE SIZE IN DIFFERENT STAGES

As for all surveys, sample size determination is a trade-off between the budget available and the desired survey precision. Because almost all the indicators in DHS are proportions, it is easy to determine the total sample size (i.e., the total number of women aged 15–49) needed for a specified precision for several main indicators at the national level and/or at the specific domain level. However, for a given total sample size, the survey cost varies a lot, depending on the number of PSUs to be selected and how the sample individuals are distributed in selected PSUs. The number of PSUs needed for obtaining the specified number of individuals varies according to the number of households to be selected in each selected PSU. For simplicity, in DHS, the number of households to be selected are constants for urban and rural areas, respectively, except for some special cases where self-weighting is requested for disclosure concerns. An equal second-stage sample size simplifies the determination of the total sample size. For simplicity, we assume the PSUs are all equal in size (in practice, the variation of the PSU size is rarely important). Suppose a simple cost function:

$$C = c_1 n + c_2 nm \tag{1}$$

where C is the total cost of the survey, not including the fixed cost;
c_1 is the unit cost per PSU for household listing and interview;
c_2 is the unit cost per individual interview;
n is the total number of PSUs to be selected; and
m is the number of individuals to be selected in each PSU.

Apart from the fixed cost of the survey, which is subtracted from the total cost, c_1 represents the cost per PSU, including mainly the cost associated with activities for updating the household list (the listing cost) and the cost associated with traveling between the PSUs for survey implementation; while c_2 represents the cost per individual interview (the interviewing cost) and the cost associated with traveling within the PSU.

The objective is to determine the optimal sample sizes in different sampling stages in order to minimize the sampling error under a given sampling budget. The DHS surveys are two-stage surveys, the first stage's sampling is a systematic sampling with probability proportional to the EA size; the second stage's sampling is a systematic sampling of equal probability and fixed size across the EAs. In terms of precision, this sampling procedure is usually better than the procedure with simple random sampling at both stages. So a conservative solution to the above problem is to suppose that the both stage samplings are simple random sampling without replacement. Furthermore, for simplicity, assume the PSUs are all equal size *M*. The variance of the sample mean is given by (Cochran 1977):

$$Var(\hat{\bar{\bar{Y}}}) = \frac{1-f_1}{n}S_1^2 + \frac{1-f_2}{nm}S_2^2 = \frac{1}{n}S_u^2 + \frac{1}{nm}S_2^2 - \frac{1}{N}S_1^2 \quad (2)$$

where $f_1 = n/N$ and $f_2 = m/M$ are the first and second stage's sampling fraction, respectively.

$$S_1^2 = \frac{1}{N-1}\sum_1^N \left(\bar{Y}_i - \bar{\bar{Y}}\right)^2 ,$$

$$S_2^2 = \frac{1}{N(M-1)}\sum_1^N \sum_1^M \left(Y_{ij} - \bar{Y}_i\right)^2$$

are the variance among the PSU means and the variance among subunits within the PSU, respectively. The minimization of the above variance (2) under given total cost gives the solution (Cochran, 1977):

$$\begin{cases} m_{opt} = \dfrac{S_2}{S_u}\sqrt{c_1/c_2} \\ n_{opt} = \dfrac{C}{c_1 + c_2 m_{opt}} \end{cases} \quad \text{with } S_u^2 = S_1^2 - \frac{1}{M}S_2^2 \quad (3)$$

We know from practice the value of c_1/c_2, but we do not know the value of S_2/S_u. For calculating the optimal value of m_{opt}, we must find a way to estimate this variance ratio. Let ρ be the *intracluster correlation coefficient,* defined as:

$$\rho = \frac{2\sum_i \sum_{j<k} \left(Y_{ij} - \bar{\bar{Y}}\right)\left(Y_{ik} - \bar{\bar{Y}}\right)}{(M-1)(NM-1)S^2}$$

After some basic algebraic calculations, it is easy to find that:

$$\begin{cases} S_1^2 \cong \dfrac{1}{M}S^2\left[1+(M-1)\rho\right] \\ S_2^2 \cong S^2(1-\rho) \\ S_u^2 \cong S^2\rho \end{cases} \quad (4)$$

Therefore, the variance ratio S_2^2/S_u^2 is given by:

$$\frac{S_2^2}{S_u^2} \cong \frac{1-\rho}{\rho} \quad (5)$$

Using this approximation in expression (3), we have the approximate optimal sample sizes:

$$\begin{cases} m_{opt} = \sqrt{\dfrac{1-\rho}{\rho}c_1/c_2} \text{ , if } \rho > 0; \quad m_{opt} = M, \text{ if } \rho \le 0 \\ n_{opt} = \dfrac{C}{c_1 + c_2 m_{opt}} \end{cases} \quad (6)$$

It is interesting to note that the optimal sample take depends explicitly on the cost ratio $c_1/c_2$and the intracluster correlation ρ, but not on the cluster size, that is, the number of the second stage's sampling units in the cluster. In fact, the cluster size has little effect on the sampling error if the second stage's sample size is fixed. The optimal sample take is an increasing function of c_1/c_2 and a decreasing function of ρ. This means that if the sampling cost of drawing a PSU is important, we draw fewer PSUs and more subsampling units in each PSU. If ρ>0, a larger value of ρ means a strong intracluster homogeneity, so we draw fewer secondary sampling units and more PSUs. If $\rho \le 0$, this means a strong intracluster heterology, so we take all of the secondary sampling units in the selected PSU and fewer PSUs to decrease the sampling cost.

CALCULATION OF THE OPTIMUM SAMPLE SIZE

The calculation of the optimal sample size has been turned to the calculation of the intracluster correlation. Intracluster correlation is not a sampling error measurement, and is rarely calculated in survey data analysis, but it is closely related to another survey design efficiency measurement parameter called *design effect*, which sometimes is calculated along with sampling

error calculation. For example, design effects are calculated for most of the key indicators in DHS surveys. Therefore, the calculation of the intracluster correlation can be achieved through the calculation of the design effect. Design effect of complex surveys was first considered by Kish (1965) and then studied by Kish and Frankel (1974). It is now widely used as a measure of efficiency of complex survey designs (see Särndal, Swensson and Wretman, 1992). More detailed studies are seen in Park and Lee (2001, 2002, 2004). Let *deft* denote the design effect of the survey, which is defined as in Kish (1995):

$$deft = \sqrt{\frac{Var(\hat{\bar{\bar{Y}}})}{S^2 / nm}} \tag{7}$$

where $Var(\hat{\bar{\bar{Y}}})$ is the actual variance of a mean estimator for the two-stage survey and S^2 / nm is the approximate variance of the mean estimator, if the sample was drawn by simple random sampling without replacement with the same total sample size nm:

$$Var_{SRSWOR}\left(\hat{\bar{\bar{Y}}}\right) = \frac{1 - f_1 f_2}{nm} S^2 \tag{8}$$

with S^2 denoting the total population variance. Using the results given in expression (4), the variance of the sample mean for a two-stage sampling given in expression (2) can be written as:

$$Var\left(\hat{\bar{\bar{Y}}}\right) \cong \frac{1 - f_1}{n} \frac{1}{M} S^2 \left[1 + (M-1)\rho\right] + \frac{1 - f_2}{nm} S^2 (1 - \rho) \tag{9}$$

According to the definition of *deft* in (7), it can be calculated that the value of *deft* for a two-stage sampling is given by:

$$deft = \sqrt{(1 - f_1) f_2 \left[1 + (M-1)\rho\right] + (1 - f_2)(1 - \rho)} \tag{10}$$

If the first stage's sampling fraction is negligible $f_1 \cong 0$, the above expression of *deft* can be simplified as:

$$deft \cong \sqrt{1 + (m-1)\rho} \tag{11}$$

compared to a single-stage cluster sampling where *deft* is given by: $deft^* = \sqrt{1 + (M-1)\rho}$. It is interesting that *deft* for a two-stage sampling depends on the intracluster correlation and the sample take, but not on the cluster size. For a given intracluster correlation $\rho > 0$, the smaller the second stage's sample size, the better the precision of the survey. Therefore, a two-stage sampling is better than a cluster sampling if the intracluster correlation is positive, as we have $deft < deft^*$ if $m \neq M$. When $m = M$, the two-stage sampling is degenerated to a cluster sampling; when $m = 1$, the two-stage sampling is approximately equivalent to simple random sampling.

The value of *deft* may depend on other survey parameters, such as sampling weight. The variation of sampling weight values contributes to the sampling variance, therefore to the *deft* (Kish, 1987; Park and Lee, 2004). But the sampling weight's influence on *deft* is small for DHS surveys, as the surveys usually are designed for self-weighting. But the self-weighting property is broken down by the differences between the number of households listed and the census number of households in each cluster. Usually, the difference is small if the population census is not too old. Therefore, for simplicity, we ignore the sampling weight influence on *deft* in this study.

Since the value of *deft* is calculated for most of the important indicators in all DHS surveys, this information can be used to estimate the value of ρ for the current survey. Suppose that *deft* and the sample take per cluster for a specific indicator in a country's previous DHS were $deft_0$ and m_0, respectively. According to expression (11), the value of ρ can be estimated by:

$$\hat{\rho} = \frac{deft_0^2 - 1}{m_0 - 1} \tag{12}$$

Therefore, an approximate solution of the optimal sample take is given by:

$$\begin{cases} m_{opt} = \sqrt{\dfrac{1-\hat{\rho}}{\hat{\rho}} c_1 / c_2} \ , \ if \quad \hat{\rho} > 0 \\ m_{opt} = M, \quad if \quad \hat{\rho} < 0 \end{cases} \tag{13}$$

The optimal sample size for the first stage's sampling is then:

$$n_{opt} = \frac{C}{c_1 + c_2 m_{opt}} \tag{14}$$

When using c_1 / c_2 and $\hat{\rho}$ obtained from previous surveys, Table 1 calculates the optimal sample take for eight countries based on the indicator *currently married women, 15–49, currently using any contraceptive method,* which has a moderate *deft* among all other indicators. Table 2 gives the optimal sample take in function of c_1 / c_2 and the intracluster correlation.

Table 1: Optimal sample take calculated for currently married women, aged 15–49, currently using any contraceptive method. Based on $deft_0$, m_0, and c_1 / c_2 from past surveys

Country	c_1 / c_2	$deft_0$	m_0	$\hat{\rho}$	m_{opt}
Country 1	10	1.34	33	0.025	20
Country 2	10	1.37	25	0.037	16
Country 3	12	1.32	12	0.067	13
Country 4	12	1.65	34	0.052	15
Country 5	15	1.92	33	0.084	13
Country 6	27	1.26	20	0.031	29
Country 7	48	1.67	32	0.058	28
Country 8	52	1.30	31	0.023	47
Average	20*	1.48	28	0.047	23

*The average value of the cost ratio is a weighted average by using the number of clusters in the survey as weights.

Table 2: Optimal sample take based on different values of c_1 / c_2 and ρ

c_1 / c_2	Intracluster correlation ρ													
	0.01	0.02	0.03	0.04	0.05	0.06	0.08	0.10	0.12	0.14	0.16	0.20	0.25	0.30
2	14	10	8	7	6	6	5	4	4	4	3	3	2	2
3	17	12	10	8	8	7	6	5	5	4	4	3	3	3
5	22	16	13	11	10	9	8	7	6	6	5	4	4	3
7	26	19	15	13	12	10	9	8	7	7	6	5	5	4
10	31	22	18	15	14	13	11	9	9	8	7	6	5	5
12	34	24	20	17	15	14	12	10	9	9	8	7	6	5
15	39	27	22	19	17	15	13	12	10	10	9	8	7	6
17	41	29	23	20	18	16	14	12	11	10	9	8	7	6
20	44	31	25	22	19	18	15	13	12	11	10	9	8	7
25	50	35	28	24	22	20	17	15	14	12	11	10	9	8
30	54	38	31	27	24	22	19	16	15	14	13	11	9	8
35	59	41	34	29	26	23	20	18	16	15	14	12	10	9
40	63	44	36	31	28	25	21	19	17	16	14	13	11	10
45	67	47	38	33	29	27	23	20	18	17	15	13	12	10
50	70	49	40	35	31	28	24	21	19	18	16	14	12	11

A study of selected indicators throughout 48 surveys (see Table 1) shows that the overall average value of the intracluster correlation is around 0.06. From Table 2, for the cost ratio c_1 / c_2 between 20–25, the optimal sample take is between 18–20 women aged 15–49. But in all of the DHS surveys, the second stage's sampling unit is household, so we need to convert this number to optimal number of households to be selected in each PSU according to the average number of women aged 15–49 per household in a specific country. The DHS surveys show the number of women aged 15–49 per household varies from 0.9 to 1.4. In order to get the expected

total number of women aged 15–49 with successful interview in the survey, it needs to take the nonresponse into account, too. Our experiences show that the total response rate (household response rate multiplied by woman response rate) is around 90%. This means the optimal sample, taken by adding the nonresponse, is around 22–25 households, if the average number of women aged 15–49 per household is 0.9, and around 14–16 households, if the average number is 1.4 women.

EVALUATION OF PRECISION LOSS WHEN USING A NONOPTIMAL SAMPLE TAKE

Earlier we saw that the optimal sample take can be calculated only approximately from previous surveys, and therefore the actually used sample take is usually different from the optimal one. We thus must consider the precision loss due to the use of the nonoptimal sample take. Assuming that the actually used sample takes are m_0, n_0, with design effect noted as $deft_0$, from the results obtained above, it is easy to see that the variance of the mean estimate is approximately equal to:

$$Var\left(\hat{\bar{\bar{Y}}}'\right) = deft_0^2 \frac{1-f_{01}f_{02}}{n_0 m_0} S^2 = \frac{1-f_{01}f_{02}}{n_0 m_0} S^2 \left[1+(m_0-1)\rho\right]$$

with $n_0 = \dfrac{C}{c_1 + c_2 m_0}$. While the variance of the mean estimate using the optimal sample sizes is:

$$Var\left(\hat{\bar{\bar{Y}}}\right) = deft^2 \frac{1-f_1 f_2}{n_{opt} m_{opt}} S^2 = \frac{1-f_1 f_2}{n_{opt} m_{opt}} S^2 \left[1+(m_{opt}-1)\rho\right]$$

with $n_{opt} = \dfrac{C}{c_1 + c_2 m_{opt}}$. Assuming that $f_1 f_2 \cong 0$, $f_{01} f_{02} \cong 0$, the variance ratio is:

$$\frac{Var\left(\hat{\bar{\bar{Y}}}'\right)}{Var\left(\hat{\bar{\bar{Y}}}\right)} = \frac{1+(c_1/c_2)/m_0}{1+(c_1/c_2)/m_{opt}} \frac{1+(m_0-1)\rho}{1+(m_{opt}-1)\rho}$$

The *relative precision loss* (*RPL*) is defined as the ratio of the standard error minus 1:

$$RPL = \sqrt{\frac{1+(c_1/c_2)/m_0}{1+(c_1/c_2)/m_{opt}}} \sqrt{\frac{1+(m_0-1)\rho}{1+(m_{opt}-1)\rho}} - 1 \qquad (15)$$

RPL is thus a measure of increase of the half-length of the confidence interval due to not using the optimal sample take. For example, a value of 0.25 for *RPL* means that the half-length of the confidence interval will be increased by 25%, compared to the case where the optimal sample take is used.

REFERENCES

Cochran, W. G. (1977). *Sampling Techniques*. New York: John Wiley and Sons, Inc.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons, Inc.

Kish, L. and Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B*. 36, 1-22.

Kish, L. (1987). Weighting in Deft². *The Survey Statistician*, June 1987.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.

Park, I. and Lee, H. (2001). The design effect: Do we know all about it? *2001 Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Park, I. and Lee, H. (2002). A revisit of design effects under unequal probability sampling. *The Survey Statistician*, 46, 23-26.

Park, I. and Lee, H. (2004). Design effects for the weighted mean and total estimators under complex survey sampling. *Survey Methodology*, 30, 183-193.

Thanh, N. L. and Vijay, K. V. (1997) An analysis of sample designs and sampling errors of the Demographic and Health Surveys. *Demographic and Health Surveys Analytical Reports*, No. 3, Macro International, Inc.