USAID
FROM THE AMERICAN PEOPLE

DHS WORKING PAPERS

Optimal Sample Sizes for Two-stage Cluster Sampling in Demographic and Health Surveys

Alfredo Aliaga

Ruilin Ren

2006 No. 30

DEMOGRAPHIC AND HEALTH RESEARCH

July 2006

The *DHS Working Papers* series is an unreviewed and unedited prepublication series of papers reporting on research in progress based on Demographic and Health Surveys (DHS) data. This research was supported by the East-West Center. Additional support was provided by the United States Agency for International Development (USAID) through the MEASURE DHS project (#GPO-C-00-03-00002-00). The views expressed are those of the authors and do not necessarily reflect the views of USAID, the United States Government, or the organizations with which the authors are affiliated.

MEASURE DHS assists countries worldwide in the collection and use of data to monitor and evaluate population, health, and nutrition programs. Additional information about the MEASURE DHS project can be obtained by contacting ORC Macro, Demographic and Health Research Division, 11785 Beltsville Drive, Suite 300, Calverton, MD 20705 (telephone: 301-572-0200; fax: 301-572-0999; e-mail: reports@orcmacro.com; internet: www.measuredhs.com).

The Optimal Sample Sizes for Two-Stage Cluster Sampling in Demographic and Health Surveys

Alfredo Aliaga and Ruilin Ren
ORC Macro

July 2006

Corresponding author: Alfredo Aliaga, Demographic and Health Research Division, ORC Macro, 11785 Beltsville Drive, Suite 300, Calverton, MD 20705. Phone: 301-572-0940, Fax: 301-572-0999, Email: alfredo.aliaga@orcmacro.com.

Abstract

This paper examines the optimal sample sizes in a two-stage cluster sampling, a sampling procedure used in most Demographic and Health Surveys (DHS), which are interview surveys of household members in a certain age group. Determining optimal sample size is a critical step in a DHS survey because it requires a trade-off between the budget available and the desired survey precision

The households in a survey area are stratified according to type of residence (urban-rural) crossed by administrative/geographical regions. In the first stage, a number of primary sampling units (PSUs), or clusters, are selected from a sampling frame independently in each stratum. The sampling frame is usually a complete list of enumeration areas (EAs) created in a recent population census. After the selection of EAs and before the second-stage selection, a household listing and mapping operation is conducted in each of the selected EAs. This operation updates the outdated population information in the sampling frame and provides a list of all of the households residing in each EA with a location map. In the second stage, a fixed number of households are selected from the newly constructed household list in each of the selected EAs, and all household members in a certain age group (e.g., all women age 15-49 and all men age 15-59) in the selected household are selected for the survey.

This two-stage sampling procedure has several advantages: it provides good coverage, is simple to implement, and allows for control of field-work quality. In order to achieve both economy and good precision, sample sizes at both stages of the survey must be determined in such way that they minimize the sampling error under a given sampling cost.

This paper investigates the optimal sample sizes in different situations in DHS surveys, based on experiences of actual surveys. The results show that for an average cluster size of 100-300 households, for moderate intracluster correlation and cost ratio, the optimal second-stage sample size is about 20 women per cluster. The results also show that for most of the DHS surveys the sample sizes met the optimal standard or were within tolerable limits of relative precision loss.

Key Words: Cluster; Cost Ratio; Demographic and Health Survey (DHS); Enumeration Area (EA); Intracluster Correlation; Primary Sampling Unit (PSU); Sampling Design; Two Stage Cluster Sampling.

1. Introduction

Over the past 20 years, the Demographic and Health Surveys (DHS) program of ORC Macro (Macro) has implemented and/or provided technical assistance for about 200 surveys in about 80 countries in Africa, South America, Southeast Asia, and West Asia. Macro has accumulated a great deal of experience in sampling design, data collection, and data analysis. Our experience tells us, for example, that in a two-stage sample, a second-stage sample of 20-30 women of a certain age group per cluster is adequate for gathering data on most of the survey indicators covering contraception prevalence, fertility preferences, infant and child mortality, and knowledge and behavior regarding sexually transmitted infections.

Because determining optimal sample size is a trade-off between the budget available and the desired survey precision, it is important to establish theoretical support for the practice of two-stage cluster sampling. In this paper, we present research results concerning optimal sample sizes in DHS surveys of different populations. Our results prove that the empirical second-stage sample size in DHS surveys meets the need for both cost control and survey precision, or are within the tolerable limits of relative precision loss.

All DHS surveys are in developing countries where statistics are often incomplete or not up to date. Due to the fact that sampling frames, which are statistical categories derived from a national census, are usually outdated, Macro's policy is to use simple sampling design to facilitate exact implementation and control of the fieldwork. Thus most DHS surveys are based on stratified two-stage cluster samplings.

In the first stage, *primary sampling units* (PSUs) are selected from a frame list with probability proportional to a size measure; in the second stage, a fixed number of households (or residential dwellings) are selected from a list of households obtained in an updating operation in the selected PSUs. A PSU is usually a geographically constructed area, or a part of an area, called an *enumeration area* (EA), containing a number of households, created from the most recent population census. In most cases, a complete list of the EAs is available with such basic information as their geographical location, rural-urban characteristics, total population, and total number of households. Cartographic materials delimitating the boundaries of the EAs are also available. However, in most cases, because a population census is only conducted every ten years, the important information concerning an EA (e.g., number of households residing within it) is often outdated and needs to be updated. The updating operation consists of listing all of the households residing in the selected EAs and recording for each household the basic information, such as name of household head, street address, and type of residence. The listing procedure provides a complete list of the households residing in the selected EAs, which then serves as the sampling frame for the second-stage sampling for household selection.

Due to the cost of conducting a listing operation, it is not possible to do one for the entire sampling frame. A standard practice of Macro is to conduct a listing operation only on the EAs selected in the first stage. But the cost of listing the selected EAs still represents a major expense. It is important therefore to determine, at the sampling design stage, the number of clusters to be selected and the number of individuals (the *sample take*) to be interviewed in each cluster, in order to achieve the desired precision within the survey budget. The optimal sample take is a function of the *cost ratio* and the *intracluster correlation*.

The *cost ratio* of a DHS survey represents the cost of interviewing a cluster compared to the cost of interviewing an individual. (The costs of interviewing a cluster mainly include the cost of household listing and of traveling between clusters for household listing and for individual interviews; the costs of individual interviews are mainly the interview cost and the travel cost within a 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 travel between clusters is expensive, and it is desirable to select fewer clusters and interview more individuals per cluster. To the contrary, better precision is achieved by selecting more clusters and interviewing fewer individuals per cluster.

The *intracluster correlation* on survey characteristics plays an important role in determining the sample size in the second stage. The intracluster correlation measures the similarity of the individuals on the survey characteristic within a cluster. A high intracluster correlation means that there are strong similarities between the individuals within the same cluster; therefore a large sample take per cluster will decrease the survey's precision. A low intracluster correlation means weak similarities between the individuals within the same cluster; therefore a large sample take will decrease the survey cost.

We suggest that it become a standard practice in a DHS survey to implement, in between the two sampling stages and within each selected PSU, a household listing and mapping operation to detect any distributional changes that might have occurred since the sampling frame was created. If the sampling frame is outdated, this operation is extremely important for calculating correct sampling weight. It also provides a sampling frame for the second-stage sampling and guarantees an exact implementation.

2. Optimal sample size in different stages

Determining optimal sample size is a trade-off between the budget available and the desired survey precision. Since almost all of the indicators in DHS surveys are proportions, it is easy to determine the total sample size—for example, the total number of women age 15-49 to be interviewed—needed for a specified precision for several main indicators, at either the national level or at the specific domain level. However, for a given total sample size, the survey cost varies a great deal depending on the number

of PSUs to be selected and how the sample individuals are distributed within the 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 surveys the numbers of households to be selected are constants for both urban and rural areas, 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 will assume that the PSUs are all of equal 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

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 to implement the survey; while c_2 represents the cost per individual interview (the *interviewing cost*) and the cost of traveling within the PSU.

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

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

where $f_1 = n/N$ and $f_2 = m/M$ are the first and second stages' sampling fractions, respectively. The variance among the PSU means and the variance among subunits within the PSU, respectively, are:

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

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

The minimization of the variance in expression (2) under given total cost gives the solution (Cochran 1977):

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

We know from practice the value of c_1 / c_2, but we do not know the value of S_2 / S_u. To calculate 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:

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

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

$$\frac{S_2^2}{S_u^2} \cong \frac{1-\rho}{\rho} \tag{5}$$

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

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

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 (the number of second-stage sampling units in the cluster). In fact, the cluster size has little effect on the sampling error if the second-stage 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 high, we draw fewer PSUs and more subsampling units within each PSU. If $\rho > 0$, there is a strong intracluster homogeneity, and we draw fewer secondary sampling units and more PSUs. If $\rho \le 0$, there is a strong intracluster heterogeneity, and we take all of the secondary sampling units in the selected PSU and use fewer PSUs to decrease the sampling cost.

3. Calculation of the optimal sample size

The calculation of the optimal sample size has now been turned to the calculation of the intracluster correlation. Intracluster correlation is not a measurement of sampling error, and it is rarely calculated in survey data analysis. But it is closely related to *design effect* (*deft*), which is another parameter for measuring survey design efficiency. Design effect is sometimes 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. The design effect of complex surveys was first considered by Kish (1965) and then studied by Kish and Frankel (1974), and it is now widely used as a measure of efficiency of complex survey designs. (See Särndal, Swensson and Wretman 1992; more detailed studies are found in Park and Lee 2001, 2002, and 2004). Let *deft* denote the design effect of the survey, which is defined by Kish (1995) as:

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

where $Var(\hat{\bar{\bar{Y}}})$ is the actual variance of a mean estimator for the two-stage survey; S^2/nm is the approximate variance of the mean estimator if the sample had been 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 sampling fraction is negligible $f_1 \cong 0$, the above expression of *deft* can be simplified as:

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

This compares to a single-stage cluster sampling, where the *deft* is given by:

$$deft^* = \sqrt{1+(M-1)\rho}$$

It is interesting to note that the *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 is, the more precise the survey will be. Therefore, a two-stage sampling is better than a cluster sampling if the intracluster correlation is positive, since 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 a simple random sampling.

The value of *deft* may depend on other survey parameters, such as sampling weights. The variation of sampling weights contributes to the sampling variance and therefore to the *deft* (see Kish 1987; Park and Lee 2004). The sampling weight's influence on the *deft* is very small for DHS surveys because the surveys are usually 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. The difference is usually small if the population census is not too old. Therefore, for simplicity, in this study we ignore the influence of sampling weight on the *deft*.

The value of the *deft* is calculated for most of the important indicators in DHS surveys. This information can be used to estimate the value of ρ for a survey. Supposing that the *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:

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

Using c_1 / c_2 and $\hat{\rho}$ obtained from previous surveys, Table 3.1 describes the calculation of the optimal sample take for eight different 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 3.2 gives the optimal sample take in function of c_1 / c_2 and the intracluster correlation.

Table 3.1 Optimal sample take calculated for currently married women age 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}
Cambodia	10	1.34	33	0.025	20
Uganda	10	1.37	25	0.037	16
Jordan	12	1.32	12	0.067	13
Ethiopia	12	1.65	34	0.052	15
Haiti	15	1.92	33	0.084	13
Turkey	27	1.26	20	0.031	29
Burkina Faso	48	1.67	32	0.058	28
Togo	52	1.30	31	0.023	47
Average	**20**[1]	**1.48**	**28**	**0.047**	**23**

[1] The average value of the cost ratio is a weighted average reached by using the number of clusters in the survey as weights.

Table 3.2 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 from 48 surveys shows that the overall average value of the intracluster correlation is 0.06 (see Table 5.1). In Table 3.2, for the cost ratio c_1/c_2 between 20-25, the optimal sample take is 18-20 women age 15-49. But in all DHS surveys, the second-stage sampling unit is the household, so we need to convert this number to the optimal number of households to be selected in each PSU according to the average number of women age 15-49 per household in the country being surveyed. The DHS surveys show that the number of women age 15-49 per household varies from 0.9 to 1.4. In order to get the expected total number of successful interviews of women age 15-49 in the survey,

we must also take the nonresponse into account. Our experience shows that the total response rate (household response rate multiplied by woman response rate) is approximately 90 percent. This means the optimal sample take by adding the nonresponse is 22-25 households if the average number of women age 15-49 per household is 0.9, and 14-16 households if the average number is 1.4 women.

4. Evaluation of precision loss when using a non-optimal sample size

In Section 3 we showed that the optimal sample take can only be calculated approximately from previous surveys, and therefore the sample size actually used is usually different from the optimal one. We thus must consider the loss of precision due to the use of a 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 in Section 3 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 on 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 to be the ratio of the standard error minus 1:

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

The RPL is thus a measure of the 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 percent, compared with the case where an optimal sample take is used. Table 4.1 describes the precision loss for the eight countries' data used in Table 3.1. The maximum RPLs occur for Ethiopia and Haiti, with losses of 9 percent and 11 percent, respectively. For these two countries, the actual sample take is more than twice as large as the optimal sample take. For other countries, the RPL is less than 5 percent. Table 4.2 lists the RPL values for various combinations of m_0, m_{opt}, ρ and c_1/c_2. From Table 4.2, for the average intracluster correlation 0.05 and the cost ratio

c_1 / c_2 between 10 and 15, the average optimal sample take is between 14 and 17. The precision loss is not important if the actual sample take is between 20 and 30, which are the most frequent sample takes in DHS surveys, where the RPL varies from 2 to 7 percent.

Table 4.1 Precision loss due to not using the optimal sample take for the eight countries studied in Table 3.1

Country	c_1 / c_2	m_0	$\hat{\rho}$	m_{opt}	*RPL*
Cambodia	10	33	0.025	20	0.03
Uganda	10	25	0.037	16	0.02
Jordan	12	12	0.067	13	0.00
Ethiopia	12	34	0.052	15	0.09
Haiti	15	33	0.084	13	0.11
Turkey	27	20	0.031	29	0.02
Burkina Faso	48	32	0.058	28	0.00
Togo	52	31	0.023	47	0.02

Table 4.2 Precision loss for various combinations of m_0, m_{opt}, ρ and c_1/c_2

$c_1/c_2=5$			$c_1/c_2=10$			$c_1/c_2=15$			$c_1/c_2=20$		
$\rho=0.01$											
m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL
25	22	0.00	35	31	0.00	40	39	0.00	45	44	0.00
30	22	0.01	40	31	0.01	45	39	0.00	50	44	0.00
35	22	0.02	45	31	0.01	50	39	0.01	55	44	0.00
40	22	0.03	50	31	0.02	55	39	0.01	60	44	0.01
$\rho=0.02$											
m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL
20	16	0.01	25	22	0.00	30	27	0.00	35	31	0.00
25	16	0.02	30	22	0.01	35	27	0.01	40	31	0.01
30	16	0.04	35	22	0.02	40	27	0.02	45	31	0.02
35	16	0.06	40	22	0.04	45	27	0.03	50	31	0.03
$\rho=0.03$											
m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL
15	13	0.00	20	18	0.00	25	22	0.00	30	25	0.00
20	13	0.02	25	18	0.01	30	22	0.01	35	25	0.01
25	13	0.05	30	18	0.03	35	22	0.03	40	25	0.03
30	13	0.08	35	18	0.05	40	22	0.04	45	25	0.04
$\rho=0.04$											
m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL
15	11	0.01	20	15	0.01	20	19	0.00	25	22	0.00
20	11	0.04	25	15	0.03	25	19	0.01	30	22	0.01
25	11	0.07	30	15	0.05	30	19	0.03	35	22	0.03
30	11	0.11	35	15	0.08	35	19	0.05	40	22	0.05
$\rho=0.05$											
m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL
15	10	0.02	15	14	0.00	20	17	0.00	20	19	0.00
20	10	0.06	20	14	0.02	25	17	0.02	25	19	0.01
25	10	0.10	25	14	0.04	30	17	0.04	30	19	0.02
30	10	0.15	30	14	0.07	35	17	0.07	35	19	0.04
$\rho=0.10$											
m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL
10	7	0.02	10	9	0.00	15	12	0.01	15	13	0.00
15	7	0.08	15	9	0.03	20	12	0.04	20	13	0.02
20	7	0.15	20	9	0.07	25	12	0.07	25	13	0.05
25	7	0.22	25	9	0.12	30	12	0.11	30	13	0.08
$\rho=0.15$											
m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL	m_0	m_{opt}	RPL
10	5	0.05	10	8	0.01	10	9	0.00	15	11	0.01
15	5	0.14	15	8	0.06	15	9	0.03	20	11	0.05
20	5	0.23	20	8	0.12	20	9	0.07	25	11	0.08
25	5	0.31	25	8	0.18	25	9	0.12	30	11	0.13

5. Urban and rural differences in sample take in DHS surveys

All DHS surveys use a region crossed by urban-rural stratification. The consideration of demographic differences in urban and rural areas supports the relatively independent designs in urban and rural areas and therefore the region crossed by urban-rural stratification. To save costs, the sample take in the second-stage sampling in urban clusters is always smaller than, or at maximum equal to, the sample take in rural clusters. From equation (6), the optimal sample take is an increasing function of the cost ratio c_1 / c_2 and a decreasing function of the intracluster correlation. The combined effect of the two factors on the optimal sample take is difficult to determine. Table 5.1 illustrates that for most of the selected indicators the intracluster correlation is stronger in rural areas than in urban areas. Therefore, on assuming the same cost ratio c_1 / c_2, the sample take in rural clusters should be smaller than that in urban clusters (see Table 5.2). However, it is obvious that the cost ratio c_1 / c_2 in urban areas is usually smaller than that in rural areas because of lower travel expenses. Although we are not able to calculate the cost ratio by urban and rural areas separately because the survey budget was considered for a country as a whole, Table 5.3 illustrates the optimal sample take on the assumption that the cost ratio is $c_1 / c_2 = 10$ in urban areas and $c_1 / c_2 = 15$ and $c_1 / c_2 = 20$ in rural areas, respectively. The calculated optimal sample takes show that, in general, the sample takes in urban areas are smaller than those in rural areas. A small sample take in an urban area will result in a relatively larger number of urban clusters being selected if a fixed number of individuals is to be selected in the urban area. Because the urban area of most developing countries represents only a small proportion of the whole country, a smaller sample size is usually allocated to the urban area. The strategy of selecting a relatively larger number of PSUs with a smaller sample take in an urban area will benefit the estimation precision and compensate for the relatively smaller sample size in the urban area.

Table 5.1 Intracluster correlations and their group averages for selected indicators averaged over 48 different surveys for total, urban, and rural samples

Variable	Intracluster correlation ρ		
	Total	**Urban**	**Rural**
Medical care	**0.16**	**0.13**	**0.16**
Medically delivered	0.22	0.22	0.19
Mother received tetanus	0.12	0.09	0.12
Have health card	0.15	0.08	0.16
Immunized	0.21	0.12	0.24
Given ORS	0.12	0.15	0.10
Knowledge of contraception	**0.14**	**0.11**	**0.11**
Know modern method	0.15	0.12	0.12
Know any method	0.14	0.11	0.11
Know source for method	0.12	0.11	0.10
Background or lifetime variables	**0.07**	**0.08**	**0.06**
Illiterate	0.08	0.07	0.08
Ever used contraception	0.08	0.08	0.08
Ideal family size	0.06	0.06	0.06
Age at first marriage	0.05	0.06	0.04
Children ever born to women age 40-49	0.08	0.11	0.06
Current use of contraception	**0.04**	**0.03**	**0.05**
Using any method	0.05	0.04	0.06
Using modern method	0.04	0.03	0.06
Using IUD	0.04	0.04	0.04
Using pill	0.04	0.02	0.05
Using condom	0.03	0.02	0.09
Using public source	0.03	0.03	0.04
Sterilized	0.03	0.02	0.04
Child health	**0.04**	**0.04**	**0.03**
Had diarrhea in past 2 weeks	0.03	0.03	0.03
Height for age	0.05	0.07	0.04
Weight for age	0.04	0.04	0.04
Weight for height	0.02	0.02	0.02
Fertility	**0.02**	**0.03**	**0.02**
Births in past 5 years	0.03	0.04	0.02
Currently married	0.03	0.03	0.02
Children 0-4 years	0.02	0.03	0.02
Births 1-4 years	0.02	0.03	0.02
Children ever born	0.02	0.02	0.02
Children weighted	0.02	0.02	0.02
Births 5-9 years	0.01	0.02	0.01
Children 1-2 years	0.01	0.01	0.00
Current fertility intentions	**0.02**	**0.02**	**0.02**
Want no more children	0.02	0.02	0.02
Want to delay next child	0.01	0.01	0.01
Infant mortality	**0.02**	**0.02**	**0.01**
Infant mortality past 1-4 years	0.02	0.01	0.01
Infant mortality past 5-9 years	0.01	0.02	0.01
Dead	0.02	0.02	0.02
Total average	**0.06**	**0.055**	**0.06**

Note: The bold figures are the group average or total average values.

Table 5.2 Optimal sample take for selected indicators given in Table 4.1 according to different levels of the cost ratio

Variable	$c_1/c_2 = 10$		$c_1/c_2 = 15$		$c_1/c_2 = 20$	
	Urban	**Rural**	**Urban**	**Rural**	**Urban**	**Rural**
Medical care	**8**	**7**	**10**	**9**	**11**	**10**
Medically delivered	6	7	7	8	8	9
Mother received tetanus	10	9	12	10	14	12
Have health card	11	7	13	9	15	10
Immunized	9	6	10	7	12	8
Given ORS	8	9	9	12	11	13
Knowledge of contraception	**9**	**9**	**11**	**11**	**13**	**13**
Know modern method	9	9	10	10	12	12
Know any method	9	9	11	11	13	13
Know source for method	9	9	11	12	13	13
Background or lifetime variables	**11**	**12**	**14**	**15**	**16**	**17**
Illiterate	12	11	14	13	16	15
Ever used contraception	11	11	13	13	15	15
Ideal family size	13	13	15	15	18	18
Age at first marriage	13	15	15	19	18	22
Children ever born to women 40-49	9	13	11	15	13	18
Current use of contraception	**18**	**13**	**23**	**16**	**26**	**19**
Using any method	15	13	19	15	22	18
Using modern method	18	13	22	15	25	18
Using IUD	15	15	19	19	22	22
Using pill	22	14	27	17	31	19
Using condom	22	10	27	12	31	14
Using public source	18	15	22	19	25	22
Sterilized	22	15	27	19	31	22
Child health	**15**	**17**	**19**	**21**	**22**	**24**
Had diarrhea in past 2 weeks	18	18	22	22	25	25
Height for age	12	15	14	19	16	22
Weight for age	15	15	19	19	22	22
Weight for height	22	22	27	27	31	31
Fertility	**20**	**25**	**24**	**30**	**28**	**35**
Births in past 5 years	15	22	19	27	22	31
Currently married	18	22	22	27	25	31
Children 0-4 years	18	22	22	27	25	31
Births 1-4 years	18	22	22	27	25	31
Children ever born	22	22	27	27	31	31
Children weighted	22	22	27	27	31	31
Births 5-9 years	22	31	27	39	31	44
Children 1-2 years	31		39		44	
Current fertility intentions	**26**	**26**	**31**	**31**	**36**	**36**
Want no more children	22	22	27	27	31	31
Want to delay next child	31	31	39	39	44	44
Infant mortality	**24**	**27**	**30**	**33**	**34**	**38**
Infant mortality past 1-4 years	31	31	39	39	44	44
Infant mortality past 5-9 years	22	31	27	39	31	44
Dead	22	22	27	27	31	31
Total average	**11**	**10**	**14**	**12**	**16**	**14**

Note: The bold figures are the group average or total average sample take computed based on the respective average intracluster correlation given in Table 5.1.

Table 5.3 Urban and rural differences in optimal sample take for selected indicators given in Table 5.1

Variable	Urban $c_1 / c_2 = 10$		Rural $c_1 / c_2 = 15$		20
	ρ	m_{opt}	ρ	m_{opt}	m_{opt}
Medical care	**0.13**	**8**	**0.16**	**9**	**10**
Medically delivered	0.22	6	0.19	8	9
Mother received tetanus	0.09	10	0.12	10	12
Have health card	0.08	11	0.16	9	10
Immunized	0.12	9	0.24	7	8
Given ORS	0.15	8	0.10	12	13
Knowledge of contraception	**0.11**	**9**	**0.11**	**11**	**13**
Know modern method	0.12	9	0.12	10	12
Know any method	0.11	9	0.11	11	13
Know source for method	0.11	9	0.10	12	13
Background or lifetime variables	**0.08**	**11**	**0.06**	**15**	**17**
Illiterate	0.07	12	0.08	13	15
Ever used contraception	0.08	11	0.08	13	15
Ideal family size	0.06	13	0.06	15	18
Age at first marriage	0.06	13	0.04	19	22
Children ever born to women 40-49	0.11	9	0.06	15	18
Current use of contraception	**0.03**	**18**	**0.05**	**16**	**19**
Using any method	0.04	15	0.06	15	18
Using modern method	0.03	18	0.06	15	18
Using IUD	0.04	15	0.04	19	22
Using pill	0.02	22	0.05	17	19
Using condom	0.02	22	0.09	12	14
Using public source	0.03	18	0.04	19	22
Sterilized	0.02	22	0.04	19	22
Child health	**0.04**	**15**	**0.03**	**21**	**24**
Had diarrhea in past 2 weeks	0.03	18	0.03	22	25
Height for age	0.07	12	0.04	19	22
Weight for age	0.04	15	0.04	19	22
Weight for height	0.02	22	0.02	27	31
Fertility	**0.03**	**20**	**0.02**	**30**	**35**
Births in past 5 years	0.04	15	0.02	27	31
Currently married	0.03	18	0.02	27	31
Children 0-4 years	0.03	18	0.02	27	31
Births 1-4 years	0.03	18	0.02	27	31
Children ever born	0.02	22	0.02	27	31
Children weighted	0.02	22	0.02	27	31
Birth 5-9 years	0.02	22	0.01	39	44
Children 1-2 years	0.01	31	0.00		
Current fertility intentions	**0.02**	**26**	**0.02**	**31**	**36**
Want no more children	0.02	22	0.02	27	31
Want to delay next child	0.01	31	0.01	39	44
Infant mortality	**0.02**	**24**	**0.01**	**33**	**38**
Infant mortality past 1-4 years	0.01	31	0.01	39	44
Infant mortality past 5-9 years	0.02	22	0.01	39	44
Dead	0.02	22	0.02	27	31
Total average	**0.055**	**11**	**0.06**	**12**	**14**

6. A more general cost function model

Although the application of the previous approach is under a simple cost function model which does not treat separately the travel cost between clusters, its simplicity makes it easier to understand several aspects of the problem. A more general cost function model was discussed by Hansen, Hurtwiz and Madow (1953). For *n* selected clusters and a sample take of *m* units, the total survey cost is:

$$C = c_0\sqrt{n} + c_1 n + c_2 nm$$

providing optimal *m* as

$$m_{opt} = \sqrt{\frac{c_1 + c_0 a}{c_2}\frac{1-\rho}{\rho}}$$

where

$$a = \left\{\sqrt{1+4\frac{C}{c_0}\frac{c_1 + c_2 m_{opt}}{c_0}} - 1\right\} / \left\{\frac{c_1 + c_2 m_{opt}}{c_0}\right\}$$

and

$$n_{opt} = \frac{a^2}{4}$$

The optimal values m_{opt} and n_{opt} can be calculated in a recursive way.

The first term in the above cost function model reflects the travel cost between clusters and is proportional to the square root of the number of PSUs selected. The c_0, coefficient of $\sqrt{n}$, is a value that is proportional to the square root of the survey area, i.e., $c_0 = k\sqrt{A}$, where *A* is the covered area, and *k* is a constant value reflecting the total cost per mile traveled. The other coefficients, c_1 and c_2, are defined as in the first cost model. Examples of the c_0 values are given in Table 6.1. We can apply the following four steps to compute optimal values for *m* and *n* :

i. Obtain the values of *C,* c_0, c_1 and c_2 from the last survey. We will assume homogeneity of transport cost per mile and of interviewer cost per mile traveled among all the countries. The value of c_0 can be calculated as in the following table.

Table 6.1 Average travel cost per mile traveled and the coefficient c_0

Country	Area (square miles)	Average transport cost per mile	Average interviewer cost per mile	Average total cost per mile	Estimated c_0 value
Cambodia	69,898	0.45	0.30	0.75	198
Uganda	91,134	0.45	0.30	0.75	226
Jordan	35,467	0.45	0.30	0.75	141
Ethiopia	471,778	0.45	0.30	0.75	515
Haiti	10,714	0.45	0.30	0.75	78
Turkey	300,948	0.45	0.30	0.75	411
Burkina Faso	105,869	0.45	0.30	0.75	244
Togo	21,925	0.45	0.30	0.75	111

ii. Find the estimate value of ρ for the indicator needed in the country under consideration. A useful source for such values is "An analysis of sample designs and sampling errors of the Demographic and Health Surveys" (Lê and Verma 1997). In Table 8.1 of that study, the estimated ρ for using modern methods characteristics is about 0.04. With not much variation in the estimated ρ value, as an illustration we assume homogeneity throughout the countries.

iii. Substitute the values of c_0, c_1 and c_2 and use an initial value of m in the expression for the initial value of a :

$$a = \left\{ \sqrt{1 + 4\frac{C}{c_0}\frac{c_1 + c_2 m}{c_0}} - 1 \right\} / \left\{ \frac{c_1 + c_2 m}{c_0} \right\}$$

iv. Calculate the optimal value m_{opt} for given a by:

$$m_{opt} = \sqrt{\frac{c_1 + c_0 a}{c_2}\frac{1-\rho}{\rho}}$$

and then calculate

$$n_{opt} = \frac{a^2}{4}$$

Steps (iii) and (iv) must be performed in a recursive way by using m_{opt} obtained in the previous step as initial value of *m* in the next step, until m_{opt} and n_{opt} converge.

So far we have assumed that a fixed budget was available for a survey, and the problem was to find the values for *m* and *n* that provide the smallest standard error (or variance). However, in some

situations a consideration may be to minimize the cost budget for a specific standard error. Cochran (1977) shows that these two apparently different problems are essentially the same and have the same solution apart from the value of *a*. Assuming that the given precision ε is specified as the relative standard error (or the coefficient of variation) of the mean estimator, then m_{opt} and n_{opt} are given by the above formula except that now *a* is calculated by

$$a = \sqrt{\frac{4\rho C_v^2}{\varepsilon^2}\left[1+\frac{1-\rho}{m\rho}\right]}$$

where C_v^2 represents the coefficient of variation of the population.

References

Cochran, W.G. 1977. *Sampling techniques*. New York: John Wiley & Sons Inc.

Hansen, M., N. Hurwitz, and W. Madow. 1953. *Sample survey methods and theory*, Volume 1, Chapter 6, Sections 12-21. New York: John Wiley & Sons.

Kish, L. 1965. *Survey sampling*. New York: John Wiley & Sons Inc.

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

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

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

Lê, T.N., and V.K. Verma. 1997. *An analysis of sample designs and sampling errors of the Demographic and Health Surveys*. Demographic and Health Surveys Analytical Reports, No. 3. Calverton, Maryland, USA: Macro International Inc.

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

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

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

Särndal, C.-E., R. Swensson, and J. Wretman. 1992. *Model assisted survey sampling*. New York: Springer-Verlag.