The Economic Journal of Nepal, Vol. 36, No. 1 & 2, January-June 2013 (Issue NO. 141) © CEDECON-TU

A Note on Sample Design

Bhim Raj Suwal[1]

Abstract

As a most scientific approach to sampling, probability sampling has been a common tool for conducting general household survey. Sample design generally refers to the way in which population elements are included in the sample. In this regard, two contrasting approaches to sample design - element and cluster - have been evolved. At the same time, sample design also has to take into account the various population sub-groups through stratification process. When the concept of stratification involves in sampling, four basic forms of sample designs applicable to general household survey can be envisaged. With this, sampling and estimation process tends to be more complex and a single method to sample selection, evaluation of sample design and estimation do not work. In this context, based on the available literatures and author's own work experience on sampling, this article describes various types of sample design for large-scale general household survey, methods of evaluating their efficiency, and the process of sample selection and estimation.

Introduction

Sample survey has been an essential tool for social research in contemporary world. It is a technique in which part of the elements are selected from the entire population (the term population is used to refer to the totality of elements) in order to study the whole. The selected part constitutes sample. On the basis of the methods of sample selection techniques, sample can be random or non-random. Random sample follows principle of probability in which selection probability of each unit included in the sample has known or non-zero probability (Kish, 1995). This is considered to be the most scientific method of sample selection in which human biasness in sample selection is avoided through the process of randomization. The selection of non-random samples on the other hand does not follow the principle of probability because samples are drawn with subjective judgment, convenience or in consideration with objective of the study (Kalton, 1998). Non-random samples, though not a scientific sample, can be used in a variety of situations where random sample cannot be applied.

Large-scale general household survey generally utilizes probabilistic sample when quantitative data are to be produced and findings are to be generalized to the survey population. For any such survey, a clear and reliable sample plan is always required because a reliable sample plan can minimize sampling errors and enhances precision of the estimate. A sample plan has four essential components: sample design, determination of sample size, methods of sample selection, and estimation. Sample design is basically related with the ways in which population units are included in the sample. In this regard, two contrasting

1 Dr. Suwal is an Associate Professor at Central Department of Population Studies (CDPS), Tribhuvan University, Kirtipur.

designs – *element* and *cluster* design - have been evolved and are in common use (Cochran, 1997; Kish, 1995; Kalton, 1983). The task of developing a sample design tends to become more complex when it involves the concept of stratification. Then, the probabilistic sample designs generally take four different forms such as element sampling, stratified element sampling, cluster sampling and stratified cluster sampling. A single method to sample size, design statistics, sample selection and estimation does not work in relation to different sample designs. Any discrepancy in the choice of methods of sample selection and estimation leads to a biased sample resulting in a loss of precision. This article, based on available literatures and author's own work experience on sampling, introduces four different types of probability sample design for large-scale general household survey, describes methods of evaluating their efficiency, and the process the sample selection and estimation.

Element Sampling

The term "element" in sampling literature is used to refer to the units for which information is sought (Kish, 1995). Elements are also sometimes called as ultimate unit or object. Elements may be households, persons, group of person, events, institutions, and any other physical units. In element sampling, elements are selected directly and selection of sample is completed in a single stage. This design does not involve any other concepts such as stratification, clustering, multiple stages of selection, etc. (Figure 1). Therefore, element sampling is also synonymously called as Simple Random Sampling (SRS), single-stage design or simple design (Kish, 1995; Kalton, 1983). Element sampling is considered to be the natural starting point of discussion on sampling and provides a basic design against which all other designs are compared (Kalton, 1983). This is also the most preferred design as this design yields the lowest variance, hence the most efficient design. Therefore, if a sample design is to be developed for a survey, then applicability of this design should be considered at the first step. If possible, it is strongly recommended to use this design.

Assuming binomial distribution of key study variables, sample size for a cross-sectional element sampling is determined in consideration with three statistical parameters: degree of population variability measured by proportion (p), desired level of precision measured as standard errors ($se_{(p)}$) and confidence level (usually 95%). The formula is given as

$$n' = [1.96^2 * (p * q)] / (se_{(p)}) \quad \text{.....(i)}$$

Where

n' = initial estimate of sample for element sampling,

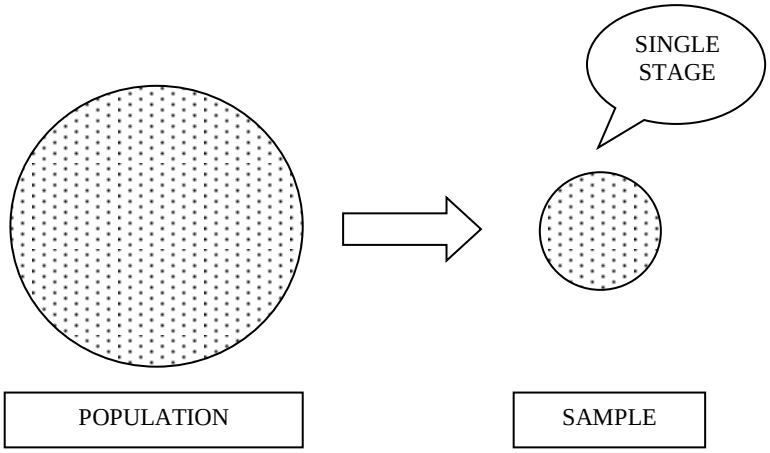
q = (1-p),

1.96 = standard normal variate at 95 percent confidence level.

Then, desired sample size (n_{srs}) is estimated by adjusting the effect of finite population correction (fpc) such as n_{srs}= n'/[(1+n'/N)] where N=population size (Kish, 1995; Krotki, 1997). When size of N is large, effect of fpc tends to be negligible. In this situation, fpc can be ignored. Final sample size is calculated by inflating desired sample size (n_{srs}) with response rate such as n/r where r=response rate. In addition, an adjustment in the final sample size can be made generally at higher side in consideration with analytical plan. Overall selection probability of sample (pr) for this design is calculated as n_{srs}/N. Sampling

rate is given as inverse of pr, i.e. 1/pr where it represents number of population elements represented by a sample unit.

Figure 1: Sketch Map of Element Sampling



Note: Each dot in the map represents element.

It is to note here that probability sampling cannot be drawn without sampling frame. Usually list of all population elements constitute sampling frame. Population here refers to totality of the elements. Since element sampling has a single domain of study, a single list of all population elements with proper identifiers constitute sampling frame.[2] Table 1 presents hypothetical example of sampling frame for 20 households (household as element) with "region" as identifier.

Table 1: Hypothetical example of sampling frame for SRS design

SN	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Region (R=Region Name)	R1	R1	R1	R1	R1	R1	R1	R1	R2	R2	R2	R2	R2	R3	R3	R3	R3	R3	R3	R3
Name of household head (N=Name)	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	N15	N16	N17	N18	N19	N20

Most often single identifier may not be sufficient to identify household of given serial number. Therefore, it is essential to use identifiers like district, VDC/municipality, ward number, locality, house number, flat number, room number, name of household head, etc. Once sampling frame of this kind is obtained, sample can be selected with the help of lottery or random number table method. Computer assisted selection method also can be applied for this. Despite being most efficient design, use of this design is largely constrained by two factors (Cochran, 1977). Firstly, if survey population is large, then it may not be possible to obtain sampling frame. Secondly, if survey has to cover large geographical areas, field survey costs especially travel cost significantly increases. It is mainly because samples are

2 Generally it is recommended that elements in the list should be arranged/listed in a serpentine fashion.

scattered all over the survey areas and field staffs have to keep on travelling much more frequently from one place to another for locating and interviewing respondents.

Imagine that a nationally representative sample survey of households is to be carried out in Nepal. Sample size determination shows that a total of 5,000 households can fairly represent the whole nation. According to the most recent population and housing census of Nepal 2011, there are 5,423,297 households in Nepal scattered all over the country. For this design, a current list of all households currently residing in the country with proper identifiers constitutes sampling frame as shown in Table 1. But none of the secondary sources provides list of households for the country as a whole that can be used for the sampling purpose. In the absence of such list, we can construction of current list of households for the country but this requires huge cost. Secondly, let us imagine that we have been able to obtain/construct a complete list of households and selected 5,000 households from the list with the help of lottery or random number table methods. In this case, sample will be scattered all over the country, most probably to 5,000 locations. Surveying 5,000 households from nearly 5,000 locations requires great amount of money. Completion of the field survey also takes long time.

Although element sampling appears to be less feasible to a survey covering large geographical areas, this design is quite feasible to a small area. It is mainly because travel involves less time and financial cost in a small area.

Imagine that a representative sample survey of households in a ward of a VDC (Village Development Committee) is to be carried out in Nepal. Sample size determination shows that a total of 300 households out of 600 can fairly represent the whole ward. In element sampling, we select 300 households directly from the list of 600 households. Even if reliable list of household for the ward is not available from the secondary sources, a complete list of household can be constructed through door to door visit. It does not involve high costs. Once list of households is obtained, lottery or random number table method can be applied to the list in order to select 300 households. Obviously, samples in this case will be scattered all over the ward, but it is quite easier to access all the sample households with less cost because ward is a small geographical area.

Mean for the element sampling is calculated as simple mean (Kalton, 1983). Let $y_i$be the values of the characteristics of element, i, n be the total sample size, then sample mean ($\bar{y}$) for element sampling is calculated as

$$\bar{y} = \sum y_i / n \quad \text{.....(ii)}$$

The above formula applies to the quantitative characteristics of the population. Therefore, y is a continuous variable. Let us assume that we selected 12 households (n=12) with lottery method from the list given in Table 1. The household size for 12 sampled households is 5, 2, 3, 4, 5, 6, 2, 3, 4, 4, 3, 5. The average household size according to the formula (ii) is calculated as (5+2+3+4+5+6+2+3+4+4+3+5)/12=3.8.

Proportion (p) or percentage is another area of interest where p is the mean for the binomial distribution. Let y_i be the count of element, i, having particular qualitative characteristics, and n be the total number of elements, p is calculated as

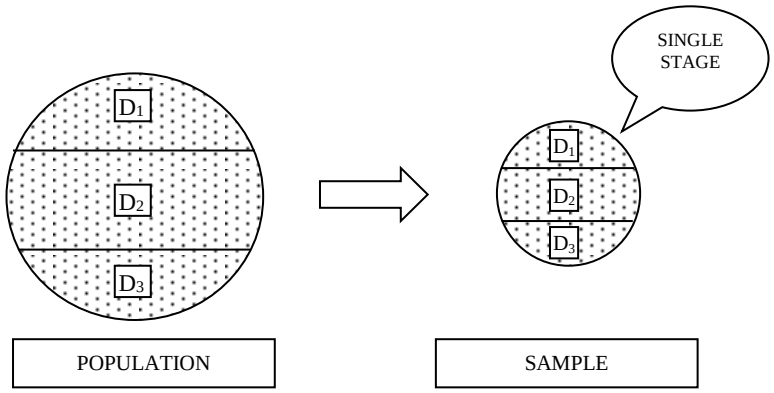
$$p = y / n \quad \text{.....(iii)}$$

Imagine that, of the 12 sampled households, 7 have toilet facility and the rest 5 do not have. Proportion of the households who have toilet facility is according to the formula (iii) is calculated as 7/12=58.3 percent.

3. Stratified Element Sampling

The basic idea about stratified element sampling is that population be stratified before selection of sample but we apply principle of element sampling in selecting sample as described in Section 1 (direct selection of elements) . In its simplest term, stratification is a way of creating different sub-groups of population based on selected characteristic (s) of the population which are called stratifying variable. In Table 1, region can be used as stratifying variable and all the 20 households can be classified into three groups – households residing in region 1, region 2 and region 3. The three categories of region here constitute domains of the study denoted by D1, D2, D3(Figure 2). Now, the design tends to become a multiple domain design. Stratification, in fact, is not an essential element for all the survey. But it is general practice to adopt this when there are sufficient evidences that population sub-groups varies significantly in terms of the values of the survey variables and there is a need to examine such variation with better representation of each sub-group in the sample. In this design, sub-group of population is said to be better represented in the sample, which generally helps gain precision (Cochran, 1977; Kalton, 1983).

Figure 2: Sketch Map of Stratified (multiple domains) Element Sampling



The creation of multiple study domains makes this design more complex in terms of the way in which we estimate sample size, select sample, and derive design and sample statistics. Generally, sample size for one domain is estimated at the first step and assigned same size for each of the remaining domain (Turner, 2003). Thus, total sample size can be obtained as a cumulative of the sample size over all the study domains. By this method, we assign equal size of sample to all domain of the study even if size of the population varies by the study domain. This method of sample size determination is valid because effect of population size on the size of sample tends to be negligible when population size is large (Kish, 1995; Bartlett, Kotrlik& Higgins, 2001; Krejecie& Morgan, 1970).

In this design, we will be interested in examining design statistics (probability and sampling rate) for total as well as study domains. Table 2 shows that when equal size of sample is implemented in the unequal-sized domain, sample design tends to be a

disproportionate design.[3] Disproportionate design implies that domain-wise relative share of sample does not correspond to the relative share of the population. In such a situation, probability of sample selection (pr_h^e) and sampling rate (sr_h^e) defined as inverse of sample selection probability ($1/pr_h^e$) does not correspond across the study domain and the total (Table 2). Then, such sample design violates the principle of *epsem* (equal probability selection method) and encounters a problem of under and over representation of the domain. Table 2 presents that if equal allocation method is adopted, mountain belt will be highly over represented in the sample with 1,667 sample size (as against 336 in proportionate design), while all other domain will be under-represented. In this situation, correction of such imbalance representation should be made. It can be done in two ways: i) by using proportionate allocation method (col. 7, Table 2), or ii) by using weight during estimation where weight is population proportion indicated with W_h in table 2 (Kalton, 1983; Kish, 1995). Calculation of sample selection probabilities and sampling rates for this design follows same technique described in Section 2 but it is essential to examine design statistics for all domain of the study (col. 5, 6, 8, and 9 in Table 2).

In stratified sampling, each domain constitutes an independent unit of the study. Therefore, we select independent sample from each domain meaning that sample selection process of one study domain should not affect another domain. Independency of the domain can be maintained by separating out the sample selection process from one domain to another. Therefore, it is necessary to construct separate sampling frame for each domain and select required samples accordingly. List of population elements disaggregated by study domain serve as sampling frame for this design. When sampling frame is obtained for each domain, then, any random method can be applied to the sampling frame for the selection of sample.

Any stratified design like this enables us to examine sample estimates at aggregate (total) as well as domain level and compare them. In fact, this is the basic intent of the stratification. Let $\bar{y}_h$ be domainwise mean, y_{hi}=value of the characteristics of the elements i in domain h, and n_h=number of sample elements in domain h, then domainwise mean ($\bar{y}_h$) is given by

$$\bar{y}_h = \sum y_{hi} / n_h \cdot \qquad \text{(iv)}$$

Overall mean for the design is computed as weighted mean,

$$\bar{y} = \sum W_h \bar{y}_h \qquad \text{(v)}$$

where

$\bar{y}$ = overall mean, W_h=population proportion for the domain h (Kalton, 1983).

Imagine that average household size for mountain, hills and Tarai is 5.7, 5.3 and 5.8 respectively. The weighted mean for total (overall mean) is computed as (0.0671*5.7)+ (0.4670*5.3)+(0.4659*5.8)=5.6. Alternatively, weighting also can be done by assigning relative weight to each sample case and using it during data analysis (Hahs-Vaughn, 2005).

3 Here, an attempt is made to present example with actual data, instead of using hypothetical data from the Table 1.

Table 2: Equal and Proportionate Allocation of Sample and Indicators of Sample Design

Ecological Belt (domains, h)	Total Households (M_h)	Proportion (W_h)	Equal sample size			Proportionate allocation		
			Sample size * (n_h^e)	Probability (pr_h^e)	Sampling rate (sr_h^e)	Sample size (n_h^p)	Probability (pr_h^p)	Sampling rate (sr_h^p)
(1)	(2)	(3)=(2)/5427302	(4)=5001/H	(5)=(4)/(2)	(6)=1/(5)	(7)=(3)*5001	(8)=(7)/(2)	(9)=1/(8)
Mountain	364,120	0.0671	1667	0.004578	218	336	0.000921	1085
Hill	2,534,430	0.4670	1667	0.000658	1520	2335	0.000921	1085
Tarai	2,528,752	0.4659	1667	0.000659	1517	2330	0.000921	1085
Total	**5,427,302**	**1.0000**	**5001**	**0.000921**	**1085**	**5001**	**0.000921**	**1085**

* Hypothetical sample size, H refers to total number of strata which is 3.
Source: Information on the number of household is adopted from CBS, 2012.

Table 3: Proportionate Allocation of Cluster and Household Sample Size

Ecological Belt (domain, h)	Total households (M_h)	Proportion (W_h)	Sample size (n_h)*	Number of sample cluster assuming (b=20) (a_h)	Adjusted Sample size (n^a_h)**	Probability (a*b)/M_h (pr_h)***	Sampling rate (sr_h)
(1)	(2)	(3)=(2)/5427302	(4)=(3)*10000	(5)=(4)/20	(6)=(5)*20	(7)=(6)/(2)	(8)=1/(7)
Mountain	364,120	0.0671	671	34	680	0.001868	535
Hill	2,534,430	0.4670	4670	234	4680	0.001847	542
Tarai	2,528,752	0.4659	4659	233	4660	0.001843	543
Total	**5,427,302**	**1.0000**	**10000**	**501**	10020	0.001846	542

* Hypothetical sample used in Section 2. Instead of 500 sample cluster, we use 501 as determined by b=20. The increase in the sample clusters is due to rounding effect.

** Refers to adjusted sample size with rounding effect in estimating the number of sampled clusters.

***In proportionate design like this, there should not be difference the probability of selection (p_h) and sampling rate (sr_h). Small difference seen here in the value of probability and sampling rate is due to rounding effect.

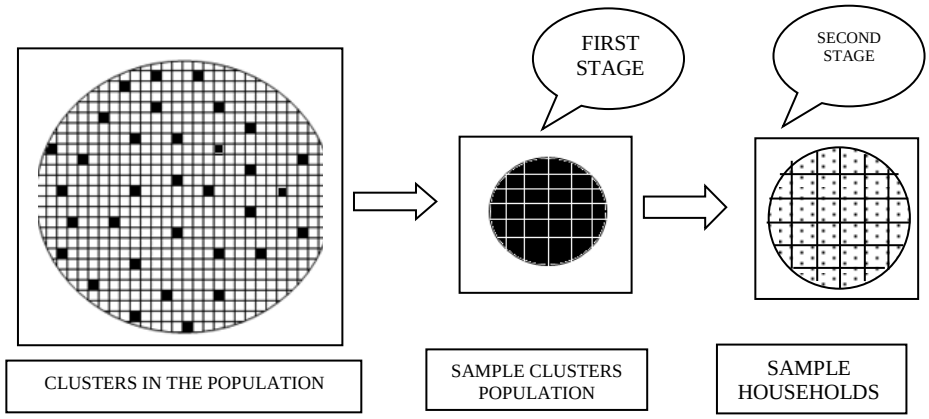
Source: Table 1 for total households.

Except stratification, the other feature of this design is similar to that of element sampling described in Section 2. Therefore, applicability of this design is again largely constrained by availability of reliable sampling frame for large population, and huge travel cost involved when survey population is scattered all over the large geographical areas. This design, however, is another most preferred design as we can have significant gain in precision due to stratification. Therefore, it is argued that variance of this design is generally expected to be not greater than that of element sampling (Kalton, 1983). Therefore, design effect (deff.) measured as a ratio of variance for proportion (p) from stratified element sampling [$var_{st(p)}$] to the element sampling [$var_{el(p)}$] is expected to be less than that of element sampling, i.e. $var_{st(p)}/var_{el(p)}<1$.

Cluster Sampling

Cluster sampling provides a more feasible alternative to element sampling to conduct survey in a large geographical area with much less cost (Cochran, 1997). Cluster is defined in terms of small spatial unit from which more than one element is included in the sample. By doing so, it adds more complexity in the sampling and estimation process. Therefore, cluster design is often called as *complex* design (Kish, 1995; Kalton, 1983). The term "cluster sample" or "cluster design" is generally used to refer to a design without stratification. Disregarding the regional boundary, we can group 20 households given in table 1, into 7 clusters consisting of 3 households in the first 6 clusters and 2 households in the last one cluster. Cluster may or may not be of equal size. Figure 3 shows a sketch map of population and sample for the cluster design where population is divided into various small clusters and clusters are selected as sample. Note that each square in the map constitutes a cluster and the shaded clusters with black are the clusters included in the sample.

Figure 3: Sketch Map of Cluster Sampling



Note: Elements are shown only in second stage, although clusters consist of elements. The sketch map is presented just to represent the concept, so number of sample clusters may not be same.

Cluster sampling involves more than one stage of sample selection. If clusters and households are primary and ultimate sampling units[1] respectively, then sample selection is

1 Primary Sampling Unit (PSU) is defined as the sampling unit selected as the first stage of selection. Ultimate sampling unit refers to the final unit, or elements selected at the final stage of sample selection.

completed in two stages. At the first stage, we select clusters and, at the second stage, elements. Clusters can be selected with the help of any random method such as probability proportion to size (PPS), lottery, or random number table methods. There are two ways of selecting elements from the clusters: i) include all elements in the sample, or ii) select sub-sample of the elements. Sub-samples can be selected using any probability method such as lottery, random number table, or systematic random sampling method. We can select same or different size of sub-sample from each cluster. But it is to note here that if the feature of *epsem* is to be maintained in the design where clusters are of unequal-size, it is necessary to allocate samples proportionately to each study domain, select clusters with PPS and take uniform size of sub-sample from each cluster (Frerichs, 2004; Suwal, 2012).

In cluster sampling, determination of sample size involves more complex process than that of element sampling. It is because sample size should be estimated at two levels, total and cluster, where total sample is distributed over the clusters. In other words, total sample is an aggregate of the sub-sample taken from the clusters. In general, taking of uniform size of sub-sample is preferred. However, taking uniform size of sub-sample may not be possible in a design when clusters are of unequal size and intended to include all elements in the sample.

Sample size for a cluster design is determined in consideration with possible effect of homogeneity (Turner, 2003; Turner, 1994). Degree of homogeneity is measured in terms of intra-class correlation. The problem of homogeneity arises in cluster sample mainly because we take more than one element from each cluster and elements included in the sample, as compared to the element sampling, are more likely to be alike in terms of the survey variables. Due to this, variance from the cluster design is often said to be higher than that of element sampling. Design effect measured as a ratio of variance for proportion (p) obtained from the cluster design ($var_{ct(p)}$) is expected to be greater than that of obtained from the element sampling ($var_{el(p)}$), i. e. $var_{ct(p)}/var_{el(p)} > 1$ (Cochran, 1977; Kish, 1995).

Design effect greater than unity implies that cluster design is less efficient than the element sampling. Therefore, cluster design is generally said to be less efficient than the element sampling. However, it is argued that sampling efficiency of the cluster design can be improved by using larger size of the sample. The question of how much larger sample size should be used depends on possible effect of homogeneity on the sample estimates. Generally, two-fold increase in the sample size of the element sampling is required (Turner, 1994). Therefore, total sample size for the cluster design (n_{ct}) can be obtained multiplying n_{srs} by 2 where 2 is design effect. Applying this rule to the hypothetical sample size of 5,000 for n_{srs}design indicates that a total of 10,000 sample size (5,000x2) is required for a cluster design. The value of design effect used here assume that our cluster design is 2 times less efficient than the element sampling but try to make it as efficient as element sampling by using two times larger sample size.

Once total sample size is determined, the next step is to determine sub-sample for the clusters. Here, sub-sample is denoted by "b". Size of sub-sample to be taken from each cluster again should be determined in consideration with degree of homogeneity of the elements. The relation is such that, higher the degree of homogeneity of the elements included in the sample, for a given size of the sub-sample, higher the variance of the design. (Turner,1994). Likewise, for a given level of homogeneity of the elements, variance of the design tends to increase with larger size of sub-sample taken from the clusters. Therefore, it is generally suggested to take size of sub-sample as small as possible. But we cannot take very small size of sub-sample because taking of too small sub-sample will not ease the survey much if large geographical areas have to be covered. Depending upon the nature of

the survey variables, taking sub-sample of 15-25 elements from each cluster is common. Assuming n_{srs}=5,000, design effect=2 and sub-sample of 20 elements per cluster, the number of sample clusters to be selected denoted by 'a' is 500 [or (5,000*2)/20)]. Total sample size for cluster design, n_{ct}, is equivalent to (a x b)=(500 x 20) = 10,000. Surveying 10,000 households from the 500 clusters scattered in 500 locations requires much less cost than the 5,000 households in element sampling scattered in 5,000 locations. The overall probability of sample selection (pr) is calculated as

$$pr = [(a * b) / M] = n / M \quad \text{(vi)}$$

where,

(a*b) = total sample size equivalent to n

a = total number of sample clusters

b = size of sub-sample per cluster

M = total number of elements in the survey population.

Assuming selection of clusters with PPS method, probability of an element being selected from a particular cluster (p_α) is calculated as combined selection probability of cluster and household as given below.

$$pr_\alpha = [(a * m_\alpha) / M] * (b / m_\alpha) \quad \text{(vii)}$$

where,

a = total number of sample clusters

m_α = total number of elements in the α^{th}cluster (called as measures of size, MOS)

M = total number of elements in the survey population ($\sum m_\alpha$).

b = size of sub-sample

In a cluster design which includes all the elements in the sample, the value of the term n/m_α tends to be always unity. Therefore, this design even being cluster sampling is equivalent to a single-stage design (Cochran, 1977). When sub-sample of certain size is taken from each cluster, then the value of b/m_α varies across the clusters. Hence, a cluster sample design should be treated as two-stage design.

As usual, overall sampling rate is calculated as an inverse of pr, i.e. 1/pr. Sampling rate for a cluster α (sr_α) is given as $1/pr_\alpha$. Depending upon the size of cluster and sub-sample taken from each cluster, value of pr_α either remains the same or varies all over the clusters. Any variation in the value of pr_α implies that samples are not selected with equal probabilities. As a result, sample design tends to become a *non-epsem* design. This situation, even in the unequal-sized clusters, however, does not encounter in a condition when clusters are selected with PPS, and taken uniform size of sub-sample from the clusters. In a condition of equal-sized clusters, taking of equal sub-sample from each cluster is the best strategy to maintain equal selection probabilities. In case, the design is implemented with unequal selection probabilities of samples, then overall estimate of mean and variance requires use of weight where weight is proportion of population by clusters.

At each stage of sample selection, sampling frame is necessary. The nature of sampling frame varies according to the nature of sampling unit – cluster or element. For the selection of clusters, we need a list of clusters with count of the elements. For example, if we take ward of VDC (Village Development Committee) as cluster and households have to be

sampled in a survey of a district, then, a list of all wards of the district with the count of households will serve as sampling frame for the selection of cluster. Such list can be obtained either from national census or other reliable sources such as voter's list. However, before use of such list, it is necessary to ensure reliability of the information in the list. Construction of such list, however, may not be possible for the survey through door to door visit because it involves large financial costs. Once sampling frame is obtained, then, PPS method can be applied to the list for the selection of clusters.

At the second stage, we select households from each cluster included in the sample. For household selection, we need a list of households residing in each sampled cluster at the time of survey. This list serves as sampling frame for the selection of households. It is desirable to construct current list of households during field survey through door-to-door visit and do sampling. If this method is followed, field staffs should be provided rigorous training on construction of list and sample selection. During listing operation, utmost care should be given for the accuracy of the information. The objective of the household listing operation should be "none of the households be listed more than once and none of them be missed out" from the list. Household list can also be obtained from the secondary sources, but its reliability should be checked. If not reliable, then such list should be used only after correction of the errors because use of erroneous sampling frame is a source of sampling bias (Kish, 1995).

Overall mean for this design is calculated as weighted average of cluster mean when the design does not follow *epsem*. Let $y_{\alpha\beta}$ be the value of the characteristics y of the elements β in cluster α, and 'b' be the size of sub-sample per cluster, then mean for a cluster ($\bar{y}_\alpha$) is given by

$$\bar{y}_\alpha = \sum y_{\alpha\beta} / b \quad \text{..... (viii)}$$

Overall mean is given by

$$\bar{y}_c = \sum (w_\alpha * \bar{y}_\alpha) \quad \text{..... (ix)}$$

Where,

$\bar{y}_c$ =mean for the cluster sampling

w_α = clusterwise proportion of population.

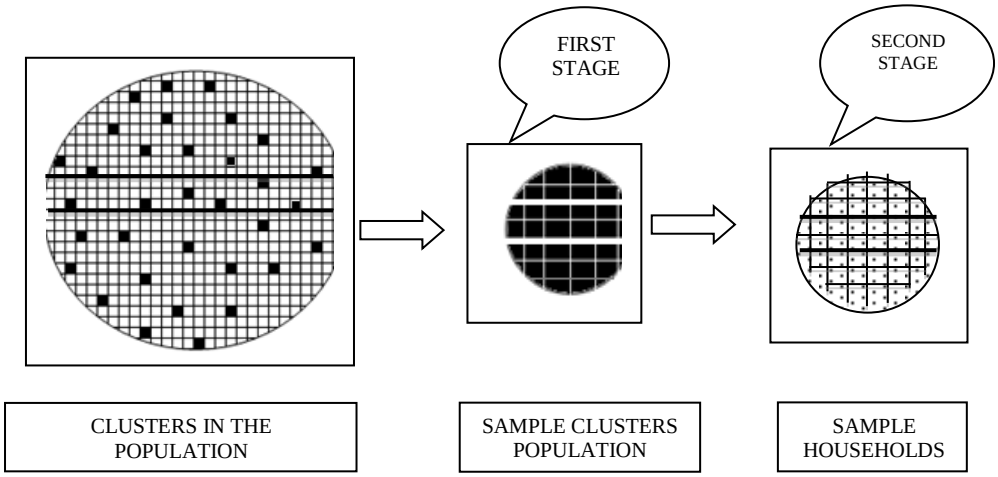
Stratified Cluster Design

Stratified cluster design is special case of stratified design described in Section 3 in which, instead of elements, we use clusters as sampling units within each strata or domain of study. Therefore, this design draws basic features of stratification from stratified element sampling and the cluster approach from the cluster design. In stratified cluster design, each domain consists of clusters, and clusters consist of elements. From the data given in table 1, we can create two clusters in each region consisting of 2-4 households. Figure 4 presents sketch map of stratified cluster design where thick lines cuts off the whole population and samples into 3 domains of the study.

Sample size determination for this design (for total and cluster) follows the same procedure as in the cluster design. But, since this design would have more than one domains of the study, it is necessary to estimate/allocate sample size for each domain with the procedure. Table 3 presents domain-wise proportionate allocation of sample using

population information presented in Table 2; total sample size of 10,000, and sub-sample of 20 elements per cluster. Once sample size for clusters and elements is determined, design-related indicators such as probability of sample selection for each domain (pr_h) including sampling rates can be derived with the procedure (equation vi) (for data, table 3). At the same time, combined selection probabilities of cluster and households also can be derived for each domain of the study using equation (vii).

Figure 4: Sketch Map of Stratified Cluster Sampling



Note: Elements are shown only in second stage, although clusters consist of elements. The sketch map is presented just to represent the concept, so number of sample clusters may not be same.

Design-related indicators presented in table 3 (col. 7 & 8) reveal that a proportionate design is an *epsem*design at domain level.[2]This is because elements from each domain have equal chance of being included in the sample resulting in uniform sampling rate across the study domain. In a design like this, however, one can select clusters through any other techniques than the PPS by taking samples in four different ways without proportionate allocation such as i) same number of clusters, same number of elements, ii) same number of clusters, different number of elements, iii) different number of clusters, same number of elements, and iv) different number of clusters, different number of elements. Use of any such combinations of sample size without proportionate allocation and PPS method to select sample cluster cannot maintain basic feature of the *epsem*if uniform size of sub-sample is not taken when clusters are of unequal size. (Frerichs, 2004; Suwal, 2012). However, such a choice could be necessary in a situation when proportionate distribution of sample cannot fulfill the objectives of the study. If so, then weighting of sample cases is necessary during data analysis so that effect of imbalance representation of the population could be eliminated.

This design draws basic features of sampling process from the stratified element sampling (Section 3) where we select independent sample from each domain. As mentioned earlier, if we take sub-sample from each cluster, sample selection will be completed in two stages – selection of cluster at the first stage and households at the second.

2 Small difference in probability and sampling rate is due to rounding effect of sample size.

In this design, we may be interested in examining mean for each domain and total (overall). Let $y_{h\alpha\beta}$ be the value of the characteristics of elementsβ of clusterα in domain h, and n_h be the sample size for domain h, then domain mean ($\bar{y}_h$)is given by

$$\bar{y}_h = \sum_{\alpha}^{a} \sum_{\beta}^{b} y_{h\alpha\beta} / n_h \quad \text{..... (x)}$$

Overall mean is calculated as weighted average of domain mean where weight is domainwise proportion of the population (W_h). Therefore, overall mean for the stratified cluster sampling ($\bar{y}_{sc}$)is given by

$$\bar{y}_{sc} = \sum W_h * \bar{y}_h \quad \text{..... (xi)}$$

Conclusions

Probability sampling is well-recognized as scientific method of sampling, although non-probability methods are also widely used to deal various situations. The general household survey commonly uses probability sampling. The basic idea of probability sampling is that human biasness in sample selection is avoided by randomization. Element and cluster sampling are the two basic form of probability sampling. But they contrast each other because the former is based on the idea of direct selection of element, which is considered to be the most efficient design. On the other hand, the cluster design is based on selection of sample in group (cluster), which is considered to less efficient design. At the same time, the general household surveys generally demand stratified sampling. When the concept of stratification is taken into consideration, the probability design takes four different forms – element sampling, stratified element sampling, cluster sampling and stratified cluster sampling.

Even being the most efficient design, element sampling cannot be applied to a large population and geographical areas. It is mainly because reliable sampling frame may not be available. Even if available, element sampling involves huge field survey costs when survey has to cover large geographical areas. Cluster design is considered to be the most feasible alternative to this, but it is a more complex design. As being less efficient design, cluster design also can be used with nearly equal efficiency of element sampling. It requires increase in the size of sample. The element and cluster sampling are two contrasting sample design not only in relation to the sample size, but also the way in which design related statistics are derived, and estimation is done. Especially in cluster design, samplers can choose different combination of sample size of cluster and elements as well as method of sample selection. But the basic stance of probability sampling is that any sample design should not be deviated from the principle of *epsem*. Deviation from the *epsem* causes sampling bias and loss of precision. Therefore, choice of appropriate combination of sample size and sample selection methods is utmost desirable.

References

Bartlett, J. E., Kotrlik, J. W.,& Higgins, C. C. (2001). Organizational research: Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, 19 (1), 43-50.

CBS (Central Bureau of Statistics) (2012).*National population and housing census 2011 (National Report).* Kathmandu: Central Bureau of Statistics.

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

Frerichs, R. R., (2004). *Rapidsurveys* (unpublished), retrieved from http://www.ph.ucla.edu/epi/rapidsurveys/RScourse/chap4rapid_2004.pdf

Hahs-Vaughn , D. L. (2005). A primer for using and understanding weights with national datasets. *The Journal of Experimental Education,* 73 (3), 221–248.

Kalton, G. (1983). *Introduction to survey sampling.* Newbury Park: SAGE Publications.

Kalton, G. (1983). Survey sampling. *Encyclopedia of Statistical Sciences,* 9, 111-119.

Kish, L. (1995). *Survey sampling*, New York: John Wiley & Sons, Inc.

Krejecie, R.V., & Morgan, D.W. (1970). Determining sample size for research activities.*Educational and Psychological Measurement*, 30 (3), 608.

Krotki, K. (1997). *Methods of survey sampling* (unpublished Teaching Material). Michigan: Institute of Social Research, University of Michigan.

Suwal, B. R. (2012). A Practical Consideration on proportionate sample design.*The Economic Journal of Nepal,* 35 (2), 127-138.

Turner, A. G., (2003). *Sampling Strategies,*Retrieved from http://unstats.un.org/unsd/demographic/meetings/egm/Sampling_1203/docs/no_2.pdf

Turner, A.G. (1994). *Master sample for multi-purpose household surveys in Nepal: Detailed sample design*, Kathmandu: National Planning Commission.