RESEARCH PAPER | Social Science | Volume : 5 | Issue : 4 | April 2015 | ISSN - 2249-555X

Use of Sample Weight in Household Survey: Principles and Practice

KEYWORDS: disproportionate design, design weight, sample design

Bhim Raj Suwal

Associate Professor, Central Department of Population Studies (CDPS), Tribhuvan University, Kirtipur, Kathmandu

ABSTRACT *Household survey generally aims to provide sample estimates for different groups of population. Therefore, prior to sample selection, population is stratified especially based on geographical characteristic of the population and allocated adequate sample size to each stratum. In allocating sample to different strata, survey may follow equal or unequal probability selection method. Sample estimates produced with unequal selection method are said to be biased meaning that they are not valid estimate of population unless adjusted with "weight". This article aims to review basic principles of sample weight, provide evidences of use of sample weight in official surveys of Nepal, demonstrates procedure to derive sample weights for different types of sample design, assign them to sample data, and derive weighted sample values as unbiased estimate of the population.*

INTRODUCTION

In many instances, stratification is an essential element in household survey design. Stratification is generally defined as a process of dividing survey population into different groups prior to sample selection. Although stratification can be done using any type of population characteristics, it is general practice to do it based on spatial characteristics such as rural-urban residence, states, regions, sometimes district, and so on. Official national surveys conducted in Nepal at various points of time like Nepal Demographic and Health Survey (NDHS), Nepal Living Standard Survey (NLSS), and Nepal Labour Force Survey (NLFS) generally stratifies survey population in amalgamation of rural-urban residence, development region and ecological belts. It is believed that stratification greatly helps enhance representativeness of the sample and sample efficiency on certain conditions. Sample design with stratification is generally called stratified sampling in which each stratum constitutes domain of study. Domain of study is also called as "estimation domain" mainly because survey wants to provide separate estimate for each domain with acceptable level of statistical precision.

Kish and Kalton has indicated that a sample design that assigns uniform selection probability to the study domains is known as proportionate design or proportionate stratification. Disproportionate design is described as an opposite of proportionate design. When population follows highly skewed distribution, proportionate design may not yield adequate sample size for small-sized domains. In this context, small-sized domains have to be either oversampled or combined with other domains. However, if there is a necessity of studying small-sized domains as a separate sub-group of population, it is necessary to oversample them generally by reducing size of sample in other study domains. Oversampling of certain study domain leads to differential selection probability across the study domain hence sample design tends to become a disproportionate design.

Disproportionate sample design was commonly adopted in official national surveys like NLFS and NDHS conducted in different points of time in Nepal. Verma, Kalton and ILO/SPA-FI have indicated that disproportionate design is also considered to be an efficient method of studying rare elements in which whole population is stratified based on degree of concentration of rare elements and domain with higher concentration of rare elements is oversampled.

Kish, Kalton and Cochran have argued that combined estimate of the study domain (or estimate for total) obtained from the disproportionate design is considered to be biased. Only weighted average of the domainwise estimate can solve the problem. This requires weighting of sample cases during data analysis. Therefore, prior to data analysis, weight factors are calculated and assigned to each sample cases in the database. These weight factors are generally known as "design weight" or "base weight". This article demonstrates the ways in which weights are determined for different types of sample design, assigned them to the sample data, and derived weighted estimates as an unbiased estimate of the population.

DEFINING SAMPLE WEIGHTS

In probability sampling, size of the survey population (or population or totality of the elements) must be known to the samplers. We cannot draw probability sample from a population whose size is not known. Sampling frame provides information on the size of the population.

If we know the size of population from which we draw sample, then we can immediately calculate two design statistics: i) inclusion probability of elements, and ii) sampling rate. Let us say we select 20 elements (n) from a set of 200 elements (N) (Table 1). Probability of an element being selected at first draw (p_1) is calculated as 1/200=0.005. According to Frerichs, the overall probability of sample selection can be calculated as 0.0.005*200=0.01 or (1/200)*1200 or 20/200). It is just addition of probability from first to the last draw (nth draw) such as 1/200+1/16500+……+nth draw. The overall probability of sample selection is an indicator of inclusion probability of all elements (p). Inverse of inclusion probability, i.e. 1/p, according to Yansaneh, works as weight which is also generally terms as sampling rate (sr). Sampling rate tells us about how many population units are represented by a sample unit. In this case, one sample unit represents 1/0.10 or 10 units in the population. Alternatively, sampling rate is given by N/n (or 200/10=10). If we multiply number of sample units (n) by sampling rate (sr), we obtain total number of population units, or 10*20=200. This shows that we can obtain population values multiplying the sample value by sampling rate. Therefore, sampling rate is generally called as multiplier or expansion factor. Hahs-Vaughn has indicated this type of weight as raw weight (denoted here by 'w'). CBS 2009, KC, Subedi and Suwal, and KC & Suwal in Nepal have used raw weight in analyzing survey data.

Non-Stratified Element Sampling

Single-stage sample design is generally known as element sampling in which elements are directly selected. It does not involve any other complexities associated with stratification and selection of clusters. Therefore, in this design, all the population elements are arranged in a single list and selected required number of units as sample with lottery or random number table method. Selection probability (p) for this design is given by overall selection probability, i.e. n/N and corresponding weight by 1/p as shown above. The left part of the database from the wi column in Table 1 corresponds to the SRS design in which data on monthly income of the household (y_i), possession of refrigerator (z_i, 1=possess, 2=do not possess) are given for 20 sample households. Each value given under y_i and z_i are called value of the characteristics. Prior to data analysis, we derive weight factors as 1/p or 1/0.10=10 which is same as derived in the previous section. The next step is to assign the weight to each sample case in a separate column. In SPSS, weights can be assigned to sample cases by using compute command (COMPUTE w=10). Value of 10 under the w_i column of the database indicates that each household in the sample represents 10 households in the population. Therefore, addition of weights (w_i) over all sample households is equal to the size of the population, i.e. $\sum w_i$=N or 200. Alternatively, size of the population can be obtained by w*n or 10*20=200. This indicates that weight factor can be used to project (extrapolate) sample values of y_i and z_i as well and calculate mean for the population ($\bar{Y}$). For example, based on the total income of the household in the sample, i.e. 624400 (Table 1), we can project total household income in the population by (w_i*y_i) which is 6244000 (10*624400). Mean income for the population is given by the weighted mean such as $\sum(w_i*y_i)/(w_i*n)$ which is equal to 6244000/200 =31220. The variable z_i is a binomial variable for which proportion (p) is an appropriate measure of mean. The database shows that, of the 20 sample households, some 8 have refrigerator (y) and the rest 12 don't have. Then, proportion of the households who have refrigerator is given by (y/n) or 8/20=0.400 = 40.0%. Proportion of the households who have refrigerator for the survey population as a whole is calculated as weighted average of sample cases such as (w*y)/(w*n) =(10*8)/(10*20) =80/200 =0.4 or 40.0%.

Since sample and population mean for both variables is the same, we can conclude here that sample mean is unbiased estimator of the population mean. In SPSS, we execute weight before data analysis session by the WEIGHT command (WEIGHT BY w).

Stratified Element Sampling

Stratified element sampling is a special case of element sampling in which whole population is divided into two or more groups. The right part of the database in Table 1, when disregarded two bold lines between 5th and 6th, and 15th and 16th records, corresponds to the stratified element sampling. Instead of drawing sample from a single list of the population, in stratified element sampling, we divide population into two or more groups of population, prepare separate list of population elements for each study domain, and determine and select samples accordingly.

In this example, two study domains have been assumed, denoted respectively by D1 and D2 where we again assume, out of 200 population elements, 50 are in D1 and the rest 150 in D2. Of the 20 sample households, we allocate equal number to each domain, i.e 10 households. If samples were allocated with proportionate method, sample size for D1 would be 5 (50/200)=0.25*20) and the rest 15 for D2 (150/200=0.75*20). The corresponding selection probability and weight for D1 would be 5/50=0.10 and 1/0.10=10 respectively. For D2, it would be 15/150=0.10 and 1/0.10=10. The size of weight for both study domains is the same and corresponds with that we derived for non-stratified element sampling.

A comparative view of domainwise sample size obtained from the above two methods indicates that, with equal allocation, 5 more households have been assigned to D1 than that would have been assigned by proportionate allocation method. This indicates a situation of over representation of D1 in the sample (upward bias of sample). This has resulted in under representation of D2 by 5 households (downward bias). This is generally known as imbalance represent of the study domain in the sample.

Let us suppose we implement the sample size obtained from the equal allocation method. Since our sample with equal allocation method is upward bias to D1 and downward bias to D2, aggregate estimate of mean without use of weight cannot be unbiased. Therefore, weight factors have to be derived at domain level and use them during data analysis. Weight for each study domain (wh) is calculated as $1/p_h$, where $p_h=(n_h/N_h)$ being n_h and N_h sample and population size in the respective study domain. In the non-stratified element sampling (Table 1), size of weight does not vary across the sample cases, but in stratified sample it varies when disproportionate allocation of sample is made. For example, using the same data given in the last paragraph, selection probability for D1 is 0.200 (10/50) with corresponding weight of 5 (1/0.20). For D2, it is 0.067 (10/150) and 15 (1/0.067) respectively. This shows that different size of weight has to be used for D1 and D2. Therefore, we assign a weight values of 5 to all 10 sample cases of D1 and 15 to all cases of D2. We can assign these values to each sample cases with COMPUTE command of SPSS. Before we assign weight, we have to have one additional column in the database to indicate study domain such as 'strata' coded with 1 and 2 for D1 and D2 respectively. The SPSS command looks like this,

IF (strata=1) w=5.

IF (strata=2) w=15.

This type of weight is also called as raw weight which is generally used to extrapolate sample value to population values. But most of the large-scale household surveys present survey findings equivalent to sample by making correction of the imbalance representation of the study domain on the aggregate sample estimates. This is in fact managed with "relative weight" (w'_h), as noted by Hahs-Vaughn. In its simplest form, relative weight is given by a ratio of population proportion in a domain h (Pr_h) to the proportion of sample (pr_h). Alternatively, relative weight can also be derived as a ratio of raw weight for a particular domain (w_h) to the mean of the raw weight ($\bar{w}$). It is denoted by

$w'_h = w_h / \overline{w}$, where, $\overline{w} = (\sum_h^H w_h)/H$ (H=total number of study domain).

Therefore, domainwise raw weight provides base data for relative weight. It is mainly because raw weight also reflects domainwise actual relative distribution of the population. For example, size of weight for D1 and D2 as calculated above is 5 and 15 respectively, and when divided them by their total which is 20 (5+15), the relative size of raw weight is 0.25 or 25% and 0.75 or 75% which is equal to the the domainwise actual population distribution.

Based on the raw weight given in Table 1, mean of the raw weight is (5+15)/2=10. The relative weight for D1 is 5/10=0.5. For D2, it is 15/10=1.5. Relative weight less than 1 suggests over representation of the domain in the sample, hence sample values should be adjusted downward during data analysis. Weight factor greater than 1 indicates an opposite situation requiring upward adjustment of the sample values. Weight factor equal to 1 indicates none of the above problems. The relative weights can be assigned to each sample cases in the database again using compute command of SPSS like,

IF (strata=1) w'=0.5.

IF (strata=2) w'=1.5.

Estimate of mean for total using relative weight is given by $\sum(w'_h*y_h)/\sum(w'_h*n_h)$, or [(0.5*312600)+ (1.5*311800)]/(0.5*10)+(1.5*10)=31200 which is equal to estimate of population mean derived above. From this, we conclude that sample mean weighted by relative weights is unbiased estimator of the population mean. NDHS conducted in different points of time in Nepal uses relative weight.

Stratified Cluster Sampling

Stratified cluster design is generally known as complex sample design in which, like stratified element sampling, there are two or more study domains and within each domain several clusters are formed or existing civil or geographical units are treated as cluster. Within each cluster, there are elements (here households). In Nepal, ward of the Village Development Committees (VDCs) and municipalities, which is the smallest political-administrative unit, are treated as cluster while conducting nationally representative surveys like NDHS, NLFS, NLS. According to the 2011 population census of Nepal, a rural ward on an average consists of 124 households with minimum of 1 to maximum of 6246 households. Urban clusters are much larger than the rural wards with average number of 1297 households. The number of households in urban wards ranges from minimum of 102 to maximum of 22715 households. Generally, large-sized clusters are segmented with equal size rule and required number of sub-ward is selected randomly.

In cluster sampling, both clusters and households constitute selection units in which clusters as Primary Sampling Units (PSUs) are selected at the first stage and, households, as Ultimate Sampling Units (USUs) at the second stage. If clusters are segmented and person (s) is/are to be selected from the households, then sample selection in this design involves four stages. This exercise, however, assumes two-stage selection design – selection of cluster and household.

In two-stage cluster design, samplers have to determine sample size of PSUs as well as USUs. Let n_h be the sample size for domain h, a_h be the sample size of clusters for domain h, and 'b' be the fixed number of households to be selected from each sampled cluster (it is same for all domains). Value of 'b' as noted by Turner is determined in consideration with degree of homogeneity of the elements in the cluster. Once value of 'b' is determined then sample size for clusters for each study domain (a_h) is given by n_h/b.

Once sample size for each study domain is determined, the next step is to determine appropriate method of sample selection. Broadly, there are three types of sample selection methods: SRS (simple random sample), systematic random sampling and PPS (probability proportionate to size). According to Frerichs, in a cluster sample like this, PPS must be used for the selection of cluster. The advantage of this method is that it can ensure epsem and provide unbiased estimate by the use of weight even when a disproportionate design is used. Selection of clusters with SRS on the other hand cannot ensure epsem even if we use proportionate design. Therefore, use of SRS method for the selection of clusters should be avoided. Household selection can be made either with SRS or systematic sampling procedure.

In PPS, selection probabilities are determined by the size of the cluster (number of households in each cluster called as measure of size (MOS). The relation is such that cluster with larger size has higher probability of being selected than that with smaller size, and a large cluster likely to be selected a_h times. Therefore, probability of a cluster being selected is given by relative size of each cluster weighted by the number of cluster to be selected from each study domain. It is denoted by $(a_h*B_{hj})/N_h$, where a_h=the number of cluster to be selected from domain h, B_{hj}=total number of elements in the j^{th} ,cluster of the domain h, and N_h=total number of elements in domain h. Element (here household) are generally selected with SRS or systematic random procedure and selection probability is given by b/B_{hj}. Since cluster design involves two-stage selection of sample, it is necessary to combine two-stage selection probabilities. Combined probabilities are generally called as overall probabilities which is calculated by multiplying first and second stage selection probabilities, or $[((a_h*B_{hj})/N_h)*(b/B_{hj})]$ when clusters are selected with PPS.

It is to note here that selection probabilities and weight factors (raw and relative weights) for the cluster sampling should be derived at cluster level. For this, listing of selected clusters with household counts generally called as measures of size (MOS) (B_{hj}) including size of sub-sample within each study domain is required.

Figure 1 depicts cluster, household and combined selection probabilities for D1 and D2. Probabilities here are calculated assuming a_h=2 (same for D1 and D2, thinner bold line in Table 1 divides domain into clusters), b=5 and N_h=50 for D1 and 150 for D2. The Figure shows that combined probabilities in both study domains follow a straight line despite great variation in cluster and household selection probabilities. This pattern of probability is always true mainly because, as noted by Babbie (1990), these two procedures have equalizing effect on ultimate (or here noted as "combined") probabilities of selection of all elements as elements in larger clusters stand a poorer chance of selection within their cluster than elements in small clusters.

Volume : 5 | Issue : 4 | April 2015 | ISSN - 2249-555X

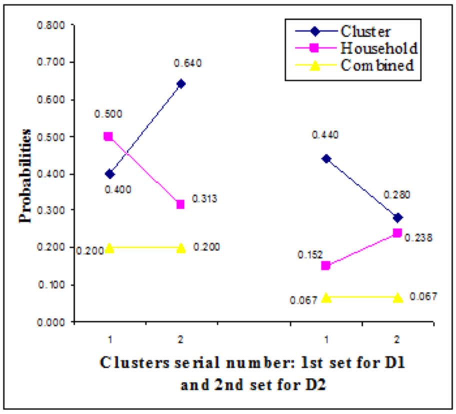

Figure 1: Cluster, household and combined selection probabilities

Now, as usual, we calculate raw weight for each domain as an inverse of combined selection probability and relative weight as a ratio of raw weight to its mean. Raw weight for each cluster of D1 and D2 is equal to 1/0.200=5 and 1/0.067=15 respectively. This is exactly same as we found for stratified element sampling. Likewise, mean of the raw weight is 10 [(5+15/2)] and relative weight is 0.5 (5/10) and 1.5 (15/10) respectively for D1 and D2. Now we can assign weights to the sample data and use in the data analysis. However, assigning weights to the data require one more column in the database (Table 1) like PSU to indicate cluster number. In this example, size of weight remains the same over the clusters of a domain. But it may vary if non-response is to be adjusted with weight. If so, then it is necessary to assign weights based on the PSU code.

TABLE - 1
HYPOTHETICAL DATABASE

hh (i)	y_i	z_i	w_i	strata	w_h	w'_h	psu
1	63000	1	10	1	5	0.5	1
2	21500	0	10	1	5	0.5	1
3	45000	0	10	1	5	0.5	1
4	15000	0	10	1	5	0.5	1
5	18000	1	10	1	5	0.5	1
6	19500	0	10	1	5	0.5	2
7	23200	1	10	1	5	0.5	2
8	27500	0	10	1	5	0.5	2
9	28900	0	10	1	5	0.5	2
10	51000	1	10	1	5	0.5	2
11	10000	0	10	2	10	1.5	1
12	33200	1	10	2	10	1.5	1
13	26800	0	10	2	10	1.5	1
14	25000	0	10	2	10	1.5	1
15	27000	0	10	2	10	1.5	1
16	31200	1	10	2	10	1.5	2
17	25300	1	10	2	10	1.5	2
18	60000	0	10	2	10	1.5	2
19	51500	1	10	2	10	1.5	2
20	21800	0	10	2	10	1.5	2

CONCLUSIONS

The basic logic behind the use of sample weight is to obtain unbiased estimate by making adjustment of imbalance representation of study domains in the sample. Imbalance representation occurs when disproportionate sample design is used. Procedures to weights are discussed basically in relation to four types of sample design: non-stratified element sampling, proportionate stratified element sampling, disproportionate stratified element sampling, and disproportionate cluster design.

The outcome of the whole exercise (Table 1) in this article indicates that the former two types of sample design correspond in terms of the size of sample weight, being weight size of 10 for each element. On the other hand, later two designs apply same size of weight, i.e. 5 and 15 for D1 and D2, despite the fact that they are two contrasting designs, one involving cluster selection and another not involving it. Evidences suggest that similarity in the size of weight in these two contrasting design is mainly due to the role of PPS method that must be applied to select cluster. In cluster sampling, use of any other method than PPS to select cluster provides unusable outcome.

Here, the varying size of weight according to the sample design (Table 1) requires clarification because different size of weight produces different size of projected population values. In non-stratified element and stratified proportionate design, we project all the value in the database by the weight factor of 10 while in disproportionate element and cluster designs we project values in D1 by 5 and in D2 by 15. The total projected values obtained for any groups consisting equal number of sample cases will be different. Obviously, estimate of population mean between the groups also tends to vary. In this context, it is important to note here that the former two types of sample design, as noted by Kish and Kalton, are the basic sample design that can produce most précised estimate. Any deviation from these two designs may lead to higher variance making the estimates less precise.

REFERENCE [1]Babbie, E. (1990), Survey research methods, Wadsworth Publishing Company. | [2]Central Bureau of Statistics (CBS) (2009), Report on the nepal labour force survey 2008, CBS. | [3]Cochran, W.G. (1977), Sampling techniques (3rd Edition, Wiley. | [4]Frerichs, R.R., (2004), Rapid Surveys | unpublished), http://www.ph.ucla.edu/epi/ rapidsurveys/RScourse/chap4rapid_2004.pdf | [5]Hahs-Vaughn , D. L. (2005), A primer for using and understanding weights with national datasets. The Journal of Experimental Education, 73 (3), 221–248. | [6]International Labour Office (ILO)/International Programme on the Elimination of Child Labour (IPEC), Special Action to Combat Forced Labour (SPA-FL) (2011), Hard to see, harder to count: survey guidelines to estimate forced labour of adults and children, ILO/SPS-FL. | [7]Kalton, G. (1983), Introduction to survey sampling, SAGE Publications. | [8]Kalton, G. (n.d.), Sampling and oversampling Rare subpopulations, IMPS Sessions. | [9]KC, B. K. and Suwal, B. R. (2013), Forced labour of adults and children among internal migrant workers in Nepal (focusing on internal migrant workers in kathmandu valley) (unpublished), ILO/ SPA-FL. | [10]KC, B. K., Subedi, G. and Suwal, B. R. (2013), Forced labour of adults and children in the agriculture sector of Nepal, ILO Country Office Kathmandu -Series No. 11. | [11]Kish, L. (1995), Survey sampling, John Wiley & Sons, Inc. | [12]Ministry of Health [Nepal], New ERA, and ORC Macro (2002), Nepal demographic and health survey 2001, Calverton-Maryland, USA, Family Health Division/Ministry of Health, New ERA, and ORC Macro. | [13]Ministry of Health and Population (MOHP) [Nepal], New ERA, and ICF International Inc. (2012), Nepal demographic and health survey 2011, Ministry of Health and Population/Nepal, New ERA, and ICF International, Calverton, Maryland. | [14]Pradhan, A., Aryal, R. H., Regmi, G., Ban, B. and Govindasamy, P. (1997), Nepal family health survey 1996, Calverton-Meryland, Ministry of Health [Nepal]; New ERA, and Macro International Inc. | [15]Turner, A. G, 1994, Master sample for multi-purpose household surveys in Nepal: detailed sample design, National Planning Commission (NPC), Kathmandu. | [16]Verma, V. (2008), Sampling for household-based surveys of child labour, International Labour Organization. | [17]Yansaneh, 2003, I. S. (2003), Construction and use of sample weights (draft), United Nations Secretariat, Statistics Division (ESA/STAT/AC. 93/5).