Sampling strategy

The sample size proposed for the selected country is designed to get sufficiently precise estimates of each tier at the national level as well as the zonal (urban and rural) level. This section, at first, presents a discussion on the factors that should be taken into consideration in the determination of sample size calculation (1.1) and provides a justification for the proposed sample size for the selected countries (1.2). Then, it explains the stratification process (1.3), and finally, the sample weighting calculation (1.4).

1.1. Issues in the determination of the sample size for a survey

The major issues considered in determining the appropriate sample size for a survey are:

1. The precision of the survey estimates (Sampling error);
2. The quality of the data collected by the survey (Non-sampling error); and
3. The cost in time and money of data collection, processing, and dissemination.

The following sub-sections discuss each of these issues in turn.

1. The precision of the survey estimates

The concept of the precision of a sample survey estimate is crucial in determining the sample size. By definition, a sample from a population is not a complete picture of it. However, an appropriately drawn random sample of reasonable size can provide a clear picture of the characteristics of that population, certainly sufficient for policy implication or decision-making purpose. From a sample of households, one can collect data and generate a sample (or survey) estimate of a population parameter. The population parameter value of characteristics of interest is generally unknown.

The formula to calculate the sample size is:

$$n = \frac{z^2 r(1-r) f k}{e^2} = \frac{z^2 r(1-r)[1+\rho(m-1)]k}{e^2} \qquad (1)$$

where:

n = Sample size in terms of number of households to be selected.

z = z-statistics corresponding to the level of confidence desired. The commonly used level of confidence is 95% for which z is 1.96.

r = Estimate of the indicator of interest to be measured by the survey.

f = Sample design effect. It represents how much larger the squared standard error of a two-stage sample is when compared with the squared standard error of a simple random sample of the same size. Its default value for infrastructure interventions is 2.0 or higher, which should be used unless there is supporting empirical data from similar surveys that suggest a different value. The sample design effect has been included in the sample size calculation formula (1) and is defined as: f = 1 + ρ (m – 1).

ρ = Intra-cluster correlation coefficient. It is a number that measures the tendency of households within the same Primary Sampling Unit (PSU) to behave alike in regards to the variable of interest. ρ is almost always positive, normally ranging from 0 (no intra-cluster correlation) to 1 (when all households in the same PSU are exactly alike). For many variables of interest in LSMS surveys, ρ ranges from 0.01 to 0.10, but it can be 0.5 or larger for infrastructure related variables.

m = Average number of households selected per PSU.

k = Factor accounting for non-response. Households are not selected using replacement.[1] Thus, the final number of household interviewed will be slightly less that the original sample size eligible for interviewing. For most developing countries, the non-response rate is typically 10% or less. So, a value of 1.1 (= 1 + 10%) for k would be conservative.

e = Margin of error, sampling errors or level of precision. It depends very much on the size of the sample, and very little on the size of the population.

2. The quality of the data (Non-sampling error)

Besides sampling errors, data from a household survey are vulnerable to other inaccuracies from causes as diverse as refusals, respondent fatigue, measurement errors, interviewer errors, or the lack of an adequate sample frame. These are collectively known as non-sampling errors. Non-sampling errors are harder to predict and quantify than sampling errors, but it is well accepted that good planning, management, and supervision of field operations are the most effective ways to keep them under control. Moreover, it is likely that management and supervision will be more difficult for larger samples than for smaller ones (Grosh, M. E., & Muñoz, J., 1996, p56). Thus, one would expect non-sampling error to increase with sample size.

3. The cost of data collection, processing, and dissemination

The sample size can affect the cost of the survey implementation dramatically. It will also affect the time in which the data can be collected, processed and made available for analysis. The availability of survey firm and cost for each country would affect the total cost of survey implementation, too. Thus, the cost of data collection, processing, and dissemination should be considered in determining the sample size for each country.

1.2. Sample size calculation

Sample surveys are appropriate for the collection of national and relatively large geographic domain level data on topics that need to be extensively explored. The main purpose of this survey is to identify and analyze the energy access tiers (tier 0 to tier 5) both at the national level and at the zonal (urban and rural) level. Equation (1) in the previous section indicates the formula to calculate the sample size. Given that the concept of the MTF has been recently introduced and the aim of this global survey is to establish the baseline of monitoring energy access globally, the indicator of interest (r) is unknown. Thus, the sample size for each selected country is calculated using the prevalence rate of 50% as the most conservative choice and to achieve the minimum margin of error (standard errors are inversely proportional to the square root of the sample size: $e=z^*\sigma/\sqrt{n}$). Since the non-response rate is typically under 10% in developing countries (United Nations, 2011), a value of 1.1 for k (non-response rate), therefore, would be considered a conservative choice (United Nations, 2011, p42). The number of households selected per PSU (m) is 12 (DHS normally visit 20-35 households per PSU, while socioeconomic surveys 6-16 households per PSU); however, it can be modified depending on the level of homogeneity in a given PSU and community. Due to the characteristics of infrastructure variables/indicator, we select 0.45 for intra-cluster relation coefficient (ρ), consequently, the design effect (f) will be equal to 6 ($f = 1 + \rho\ (m - 1)$) (Grosh, M. E., & Muñoz, J., 1996, p59).

[1] The sample size should be calculated to reflect the experience from the country in question. Hence, we will introduce the possibility of replacement of certain households in particular countries if needed. In this case, a different weight will be considered when preparing the estimates.

The number of analytic domains has a large impact on the sample size and strategy, too. An analytic domain can be defined as the analytic sub-groups for which equally reliable data is required for the analysis. The sample size is increased by a factor equal to the number of domains desired, because it does not depend on the size of the population itself.

In the process of defining a strategy to calculate the sample size for the selected countries, we have considered two approaches: one calculating, at first, the national sample size as one analytic domain and, then, allocating the sample size proportional to urban and rural population; the other is calculating, at first, the sample size using the distribution between urban and rural as two analytic domains and, then, adding these two values to obtain the national sample size. Besides, these two approaches have taken into consideration data on sample size by margin of error, ranging from approximately 4% to 5.5% at national level and from nearly 5% to 11% at zonal level. Considering the results obtained, it has been chosen to keep those of the second approach, which for a margin of error of 6% at urban and rural levels gives a national sample size of roughly 3,500 households with an error of 4.2% (Table 1). However, final sample size for each country can be modified depending on the stratification strategy.

Table 1. Calculation based on two analytic domains

Sample size			Margin of Error		
National	**Urban**	**Rural**	**National**	**Urban**	**Rural**
3,500	***1,750***	***1,750***	***4.2%***	***6.0%***	***6.0%***
3,000	1,500	1,500	4.6%	6.5%	6.5%
2,600	1,300	1,300	5.0%	7.0%	7.0%

Within each cluster/state, PSUs are selected with probability proportional to its measure of size (PPS) and households are selected with equal probability within each PSU (the definition of this approach is reported in United Nations, 2011).

1.3. Oversampling

In some cases, it is required to oversample specific geographic sub-areas or population sub-groups for specific project or research purpose. Sample size for oversampling areas should be calculated considering the estimated impact of the project (if the purpose of oversampling is to measure the impact of the intervention), or the precision of the estimates. At the same time, we also need to consider the non-sampling error and cost into the sample size calculation for oversampling area. After the sample size for oversampling area is estimated, we need to think how we will allocate the sample for oversampling.

a. If the unit of sub-groups/areas (e.g. Region / Department / Province) is in line with the highest geographical/administrative units used in the stratification strategy: We need to allocate sample proportionally to population (estimates) of each stratum. After comparing the sample size for oversampling areas and the sample size for the area after initial allocation (PPS), we need to add more sample size to the oversampling areas to obtain the precision to meet the purpose of the oversampling.

b. If the unit of sub-groups/areas (e.g. district / village tract / commune) is lower/smaller: We need to separate these sub-groups from our core sampling frame. Since the unit of oversampling areas is smaller than our unit of geographical/regional stratification, we won't be able to predict how many of households will be sampled in the initial distribution. To be cost-efficient, we would like to treat these oversampling areas as a separate stratum in the stratification strategy and allocate the number of households to this stratum.

1.4. Stratification

Once we determine the sample size, we need to develop a stratification strategy, which is the process of dividing households into homogeneous smaller groups called strata and then sampling separately for each stratum following certain rules. Stratification often improves the representativeness of the sample by reducing sampling error. Each stratum is treated as an independent population. As explained in the final session (1.4) sampling weights need to be used to analyze the data. This section provides guidelines on stratification for the MTF Global Survey.

The guidelines provided in this section are general, and ideally, this is what we aim to achieve in the stratification of the sample for the selected countries. However, these guidelines may not apply identically to all 16 selected countries where MTF surveys are going to be implemented as these countries may well vary in their geographical structure and population distribution within and across geographical units. That is, country-specific modification of the guidelines is likely, and such modification will be covered in the country-specific data collection reports.

Before discussing the stratification strategy it is useful to go over the criteria that will guide the overall stratification strategy. Such criteria are:

1. Equal allocation between urban and rural areas. This is established during sample size calculation. This will help conduct disaggregate and in-depth analysis for urban and rural areas, which are statistically sound.
2. While the parameters of interest for the MTF study are access to grid electricity and access to non-solid fuel, the first one will be used in the stratification. However, to make the analysis representative of the underlying population, sampling weights will be used that take into account the actual distribution of both grid users and non-solid fuel users in the population.
3. Sample will have 50-50 distribution of grid users and non-users. This will help us carry out in-depth analysis of both groups. Again, sample weights will be used in the analysis to compensate oversampling of either group.
4. Twelve households will be sampled from each village or urban block (PSU).

With these criteria in place we can proceed to develop stratification strategy as follows.

Allocation of strata:

An important exercise in stratified sampling design is the allocation of sample units to different strata. This is done by defining allocation rules for each stratum, which rules are not the same for all strata. In this subsection, we address the sample allocation with rural areas in mind, while the same allocation rules may very well apply to urban areas. For MTF survey, we use two strata and describe next the allocation rules for each of them.

a. ***State (or Province or Region):*** This is the highest-level geographical unit. It may also be called region, province or other name depending on the country. The rules of state selection are:
   i) Unless there are specific reasons to do otherwise, we select all states from a country.
   ii) The distribution of sample households will be proportional to actual population distribution of the states in urban and rural, respectively. Let us assume that there are three states A, B and C in a country with population N_A, N_B and N_C, respectively, so that the total population of the country is, $N = N_A + N_B + N_C$. Let us also assume that the sample size is n. So the number of sample households in state A, B and C are given by n_A=$(N_A/N)n$, n_B=$(N_B/N)n$, and n_C=$(N_C/N)n$, respectively.

b. ***Village (PSU)***: This is the primary sampling unit (PSU) for the survey. It is also called community, block, Enumeration Areas (EA), etc. Villages are allocated as follows:

i) From each State, we need to divide all villages into urban and rural since State is a combination of urban and rural PSUs.

ii) Since twelve households will be selected from each village, so total number of villages allocated to the State will be 1/12 of total number of households allocated to the State (Number of villages to be selected in the State = Number of households allocated to the State / 12).

iii) From each of the urban or rural State, all the villages will be divided into two groups: electrified and non-electrified ones.

iv) If a village has lower than 3% of electrification rate it will be defined as non-electrified for practical purpose. Any village above 3% of electrification rate will be grouped into electrified villages.

Let us say, we calculate the sample size for a country as 3,500. Using 50-50 urban-rural allocation we have to sample 1,750 households from urban areas and 1,750 from rural areas. In the case of Rwanda, for example, we calculate a slightly smaller sample size of 3,300. Using 50-50 urban-rural allocation criteria, we roughly have to sample 1,620 households from urban areas and 1,680 from rural areas. Rwanda has 5 States with following population (or household) distribution in rural and urban areas. Table 2 shows the allocation of strata for rural and urban areas.

Table 2 Allocation of states, villages and households for urban and rural areas

Rural			Urban		
States	**Villages**	**Households**	**States**	**Villages**	**Households**
Kigali (3.1% HHs)	5 villages (60/12)	60 HHs (1,680*3.1%)	Kigali (4.95% HHs)	70 wards (840/12)	840 HHs (1,620*49.5%)
Northern (17.8% HHs)	25 villages	300 HHs	Northern (9.3% HHs)	10 wards	120 HHs
Southern (26.9% HHs)	35 villages	420 HHs	Southern (13.2% HHs)	15 wards	180 HHs
Eastern (27.4% HHs)	40 villages	480 HHs	Eastern (10.7% HHs)	15 wards	180 HHs
Western (24.7% HHs)	35 villages	420 HHs	Western (17.3% HHs)	25 wards	300 HHs
Total	140 PSUs	1,680 HHs	Total	135 PSUs	1,620 HHs

Selection of sample units in strata:

After various sample units are allocated we select sample units for each stratum. What follows are the sample selection steps for different strata.

a. ***State***: Since all states are included in the sample, sample selection for states is straightforward.

b. ***Village***: The sample frame for village (or PSU) selection is the most recent census. Villages are selected as follows:

i) Villages are selected in a way so that the aforementioned criteria 3) and 4) are satisfied, that is, there will be a 50-50 distribution between grid users and non-users in the sample, and each village will have 12 households. However, in some cases, it will not be easy to keep 50-50 distribution between on-grid user and off-grid users due a country specific condition and situation. If that is the case, a field coordinator should be able to keep the ratio (electrified households/non-electrified households) less than 1.1.

ii) A list of villages is prepared for each sampled State based on the availability of grid electricity. That is, there are two lists of villages within each State: one with villages with access to electricity and the other one with villages without grid electricity. For all practical purposes, villages with a very low access rate (<3% households have grid connection) can be considered as those without electricity access. The lists of the two types of villages may be available from the National Statistical Office (NSO), National Grid authorities (utility companies), or other sources. This is something to be sorted with the survey firm. Once the list is prepared, villages can be randomly selected using any statistical package or MS Excel. There can be three scenarios:

A. *The State has both villages with electricity and villages without electricity*. In the case of Rwanda, 3 villages with electricity and 2 villages without electricity are randomly selected from the lists using a statistical package.
   i. Special case I (Number of electrified PSUs in the State is less than the number of electrified PSUs allocated to the State): We will select all the electrified PSUs in the State and oversample non-electrified PSUs. To keep the ratio between on-grid and off-grid users to less than 1.1, we will oversample electrified PSUs in other States.
   ii. Special case II (Number of non-electrified PSUs in the States is less than the number of non-electrified PSUs allocated to the State): We will select all the non-electrified PSUs in the State and oversample electrified PSUs. If the ratio between on-grid and off-grid users is less than 1.1, we do not need to oversample non-electrified PSUs in other States.

B. *All the villages in the State have access to electricity (or only few villages do not have access to electricity - e.g. if less than 2% of villages do not have access to the grid in the State then the threshold should be adjusted by the World Bank team)*. This is a special case. In this case, all the villages are randomly selected from the list of the villages.

C. *No villages in the State have access to electricity*: In this case, all the villages are randomly selected from the list of the villages. This State will be paired with another State where all sampled villages have electricity.

c. Households: Selection of household is the last step in stratification. While we will adopt a general procedure in selecting households, it is likely to change based on the context. Household selection will depend very much on the type of villages selected. That is, the different village selection scenarios discussed before need to be considered. Household selection process is described below.

A. *The State has both villages with electricity and villages without electricity*. As mentioned, in this eg. 3 villages with electricity and 2 villages without electricity are selected from these States. Households from these villages will be selected in a way so that the 50-50 ratio of grid and non-grid users can be maintained. To do so, from villages with electricity, we will select 10 households with electricity and 2 households without electricity. And, from villages without electricity, we will select 12 households. Table 3 shows this selection. Since there are 3 villages with electricity and 2 without electricity, this selection ensures the 50-50 distribution of the users and non-users of grid.[2]

Table 3: Household selection from States having both villages with electricity and villages without electricity

Villages	HHs with electricity per village	HHs without electricity per village	All HHs per village
Villages with electricity (3)	10	2	12
Villages without electricity (2)	-	12	12
Total for 5 villages	30 HHs from 5 villages (10*3=30)	30 HHs from 5 villages [(2*3)+(12*2)=30]	60 HHs from 5 villages (30+30=60)

[2] There are a number of ways villages and households can be selected to maintain the 50-50 ratio. A field coordinator can adjust the household selection considering the context and limitation that she/he is facing in the activity.

This selection also ensures that some non-grid households will be picked from the villages with grid access. Thus, this selection allows us to make a distinction between two types of non-grid households: those from villages with grid access and those from villages without grid access.

B. *All the villages in the State have access to electricity*. This is a special case when no village without electricity can be found. In this case, from each of the 5 villages we will select 6 households with electricity and 6 without electricity. Table 5 shows the distribution.

Table 4: Household selection from States in which all sampled villages have electricity

Villages	**HHs with electricity per village**	**HHs without electricity per village**	**All HHs per village**
Villages with electricity (5)	6	6	12
Total for 5 villages	30 HHs from 5 villages	30 HHs from 5 villages	30 HHs with electricity and 30 HHs without electricity from 5 villages

This selection ensures that we will have 50:50 ratio of grid and non-grid households in each State.

C. *No villages in the State have access to electricity*. In this case, we select 12 households from each of the 5 villages of the State (say, State A). Thus, we get 60 households without electricity. To compensate for this oversampling of households without electricity, we will select from another, let's say State B, 5 villages with electricity. This selection ensures that we will have 50:50 ratio of grid and non-grid households over two States. Table 5 shows the distribution.

Table 5: Household selection from States where no villages have electricity

States	**HHs with electricity**	**HHs without electricity**	**Total HHs from villages**
State A 5 villages without electricity	-	12 HHs per village	60 HHs without electricity
State B 5 village with electricity	12 HHs per village	-	60 HHs with electricity
Total for 2 States	60 HHs	60 HHs	60 HHs with electricity and 60 HHs without electricity from 10 villages

Selection of households (implementation):

Household selection is a more involved process than village selection, and so, it is discussed separately here. Once we determine the number of households to be selected, including grid and non-grid users, from different villages in the sample, we need to select those households randomly from the villages. This will involve the following steps.

a. Ideally, we would like to have a list of households, with their grid access status, for each of the villages that are sampled. Such lists may be available from the NSO, or other sources. Then using a statistical package we can select the households randomly.
b. In case the list of households is not available or outdated, the survey firm has to build such a list. It is better if such a list can be built before the scheduled survey starting date. Depending on the capability of the survey firm, availability of necessary information, planning and logistics, such a list may be developed during pre-survey activities (such as questionnaire finalization and translation, enumerator hiring and training, entering questionnaires into CAPI, and pretesting of the questionnaires). During this period, few staff members of the survey firm will go to villages and make a list of the households, including their electrification status. Once such a list is prepared, sample households can be selected using a statistical package.
c. In a worst-case scenario, the list of households may not be prepared before the survey. In that case, survey team will have the responsibility of selecting sample households during the survey. This will be done as follows.
   i. On the first day of arriving in a survey village, the survey team will make a list of households in the village, including their electrification status. Such a list will include a serial number, some identification information (the household head's name, for example) and electrification status. Once the list is complete, sample households can be randomly selected. This is demonstrated in Table 6. Let us assume that a village has 150 households, 100 of which having access to grid electricity, and 50 without such access. The first and third columns show household listing for grid and non-grid users, respectively, which will be produced at the end of listing operation. With this list, we generate 6 random numbers from 1 to 100 for grid users and another 6 random numbers from 1 to 50. The second and fourth columns show the numbers generated. These households are to be sampled and surveyed. Considering the possibility of non-response we can generate 8 numbers instead of 6.

Table 6: Sampling of households from household list

Listing serial of HHs with electricity	6 random numbers for households with electricity	Listing serial of HHs without electricity	6 random numbers for households without electricity
1	3	1	5
2	12	2	9
3	26	3	13
.	47	.	29
.	78	.	34
.		.	48
50	93	48	
51		49	
.		50	
.			
98			
99			
100			

This is just one way to select households randomly. There are other ways to do so, namely, random walk, Kish grid and so on. The method to be adopted will be decided after consultation with the survey firm.

Stratification for urban area:

From urban areas, same number of households will be selected as from rural areas, and the same stratification guidelines will be followed. Instead of villages, PSUs for urban areas will be urban blocks (may also be called wards or by other names).

1.5. Sample weighting calculation

In our survey we deal with representative samples randomly selected from the target populations. This representativeness of the sample must be considered to ensure that any statistical inferences drawn from the survey data is valid. For this purpose, we use sampling weights calculated for each interviewed household to make the sample more like the target population (ICF International, 2012; United Nations, 2011).

What exactly is a sampling weight? As explained in detail in this section, a sampling weight is an inflation factor. Weighting for household surveys involves three processes: calculation of the design (or base) weights, adjustments for non-response and adjustments for post-stratification. It is crucial that sampling weights (or probabilities) at each stage of selection are cautiously calculated, applied, and recorded in any data analysis. If the exact weights that compensate for differences between census and survey measures of size are used, the resulting survey estimates will be unbiased. Failure to adjust the weights accordingly produces biased estimates, leading to incorrect conclusions.

There are a few reasons why using sampling weights is required, even though the effect of sampling weights on survey indicators may be small: 1. to obtain valid statistical inference; 2. to correct or at least reduce bias; and 3. to keep the weighted sample distribution close to the target population distribution, particularly if oversampling is applied in certain strata or domains.

As explained in the next section, at first we have to determine the probabilities of selection of sampled units, and then construct sampling weights. The probability of selection of a sampled unit depends on the sample design used to select the unit. The calculation of sampling weights begins with the calculation of the design weight for each sampled unit, in order to reflect their unequal probabilities of selection.

Basically, the design weight of a sampling unit (household in our case, but it can be individual in other cases) is the inverse of the overall probability with which the unit was selected in the sample. That means, if a unit has probability p_i to be included in the sample, then its design weight is: $w_i = 1/p_i$.

When multi-stage designs are considered, is essential that the design weights reflect the probabilities of selection at each stage. Assuming that, for example, $w_{ij,b}$ is the design weight for the j^{th} household, $w_{ij,nr}$ is the weight attributable to compensation for non-response, and $w_{ij,nc}$ is the weight attributable to the compensation for non-coverage; consequently, then the overall weight of the household is:
$w_{ij} = w_{ij,b} * w_{ij,nr} * w_{ij,nc}$.

Design weights calculation

Let's now assume that the survey sample is drawn with two-stage, stratified PSU (or cluster) sampling, hence, design weights is calculated based on the separate sampling probabilities for each sampling stage and for each PSU. We have:

P_{1hi}: probability of selecting the i^{th} PSU/cluster in stratum h in stage 1

P_{2hi}: probability of selecting the household within the i^{th} PSU/cluster in stage 2

Assuming that n_h is the number of PSUs selected in stratum h; M_{hi} is the measure of size of the PSU used in the first stage's selection, which means it is the number of households residing in the PSU according to the sampling frame (or census); $\sum M_{hi}$ is the total measure of size in the stratum h. The probability P_{1hi} of selecting the i^{th} PSU in the sample is thus:[3]

[3] Since PSU and EA are considered identical and the EA is not segmented in our analysis, we do not need to multiply P_{1hi} by the factor b_{hi}, which is the proportion of households in the selected PSU compared to the total number of households

$$P_{1hi} = \frac{n_h \; M_{hi}}{\text{å} \; M_{hi}}$$

$$P_{1hi} = \frac{\text{\# PSUs selected in stratum h * \# HHs in the PSU}_i \text{ in stratum h (from census)}}{\text{total \# HHs in stratum h}}$$

Assuming that t_{hi} is the number of households selected in the PSU *i* in stratum *h*, and L_{hi} is the number of households listed in the household listing operation in PSU *i* in stratum *h*. The second stage selection probability P_{2hi} for each household in the PSU is thus:

$$P_{2hi} = \frac{t_{hi}}{L_{hi}}$$

$$P_{2hi} = \frac{\text{\# HHs selected in the PSU}_i \text{ in stratum h}}{\text{\# HHs listed in the PSU}_i \text{ in stratum h}}$$

Consequently, the overall selection probability of each household in PSU *i* of stratum *h* is the product of the selection probabilities of the two stages:

$$P_{hi} = P_{1hi} \times P_{2hi}$$

Finally, we can calculate the design weight for each household in PSU *i* of stratum *h* as the reverse of its overall selection probability:

$$d_{hi} = 1/P_{hi}$$

As we will explain in an example at the end of this section, the calculation of the design weight is not very difficult; nonetheless, complications usually result from the fact that the design parameters involved in the above calculation are not available because they are not well documented.

Correction for non-response

Usually, non-response is common in surveys. For this reason, the design weight calculated above that is based on sample design parameters is not enough for all analyses. For example, DHS program (ICF International, 2012) confirms that rich urban households in developed regions are less likely to respond to the survey than their counterparts in poor rural and less-developed areas respectively; furthermore, individuals with higher levels of education are less likely to respond to the survey than those with lower levels of education, and men are less likely to respond to the survey than women.

in EA *i* in stratum *h*; in other words, in this case $b_{hi} = 1$. Otherwise, the probability of selecting PSU *i* in the sample would be: $P_{1hi} = \frac{n_h M_{hi}}{\sum M_{hi}} \times b_{hi}$

In general, correcting for unit non-response is required to calculate a response rate for each homogeneous response group; subsequently, the design weight has to be divided by the response rate for each response group.

Assuming that the response groups coincide with the sampling strata, we need to calculate the sampling weight by first calculating the various response rates for unit non-response. Here we consider only PSU and household levels response rate and not the individual levels response rate, given that the survey is at household level.

- *PSU/Cluster level response rate:*

Assuming that n_h is the number of PSUs selected in stratum *h* and n_h^* is the number of PSUs interviewed. The PSU level response rate in stratum *h* is:

$$R_{ch} = n_h^* / n_h$$

- *Household level response rate:*

Assuming that m_{hi} is the number of households found in PSU *i* of stratum *h* and m_{hi}^* is the number of households interviewed in the PSU. The household response rate in stratum *h* is:

$$R_{hh} = \sum d_{hi} m_{hi}^* / \sum d_{hi} m_{hi}$$

where d_{hi} is the design weight of PSU *i* in stratum *h*. The summation is over all PSUs in the stratum *h*.

The household sampling weight of PSU *i* in stratum *h* is obtained by dividing the household design weight (previously calculated) by the product of the response rate at PSU and at household levels, for each of the sampling stratum:

$$D_{hi} = d_{hi} / (R_{ch} \times R_{hh})$$

The household sampling weight above can then be used to calculate any indicators at the household level. Given that, as previously mentioned, a sampling weight is an inflation factor, the weighted sum of households interviewed is calculated as:

$$T = \sum \sum D_{hi} m_{hi}^*$$

This is an unbiased estimate of the whole number of residential households of the country. The summation is over all PSUs and strata in the full sample.

Caveat: the increase in sampling variance caused by weighting. Weights in the analysis of survey data are introduced with the aim of reducing the bias in the estimates; however, weights could also increase the variances of such estimates.

Example on Weights Calculation

In order to help the reader in understanding the weights calculation explained in the previous paragraphs, let's consider the following example. State A (Stratum 1) has two clusters: rural and urban zones, clusters 1 and 2 respectively. After having randomly selected 20 PSUs from cluster 2 (urban zone), we need to calculate the sampling weights. Table 7 reports details and weighs calculation for three of the 20 randomly selected PSUs.

Table 7. Weights calculation based on two analytic domains

Stratum	Unique PSU ID	PSU #	Census Pop.	Census HHs	Average Census HH Size	Cluster Type 1=Rural, 2=Urban	# HHs in PSU from Listing
1	101001	1	482	96	5.0	2	94
1	101011	2	429	98	4.4	2	90
…	…	…	…	…	…	…	…
1	102007	20	43	8	5.4	2	17

# HHs Selected for Interview	# HHs Interviewed	Probability of Selection in 1st Stage P_1	Probability of Selection in 2nd Stage P_2	Non-Response Adjustment	Weight
8	8	0.154	0.085	1	76.16
8	7	0.158	0.089	1.14	81.63
…	…	…	…	…	…
8	7	0.01	0.47	1.14	188.88

In stratum 1:

Total PSUs: 184. Total HHs: 12,444. Total Population: 61,468. # PSUs selected in stratum 1 cluster 2: 20

$$P_1 = \frac{\#\text{PSUs selected in stratum } 1 * \#\text{HHs in PSU}_2 \text{ in stratum } 1 \text{ (from census)}}{\text{total } \#\text{HHs in stratum } 1} = \frac{20*98}{12{,}444} = 0.158$$

$$P_2 = \frac{\#\text{HHs selected in PSU}_2 \text{ in stratum } 1}{\#\text{HHs listed in PSU}_2 \text{ in stratum } 1} = \frac{8}{90} = 0.089$$

$$\mathit{Non\text{-}resp\ adjustment} = \frac{\#\text{HHs selected}}{\#\text{HHs interviewed}} = \frac{8}{7} = 1.14$$

$$\mathit{Weight} = \frac{\mathit{Non\text{-}Resp\ Adjustment}}{P_1 * P_2} = \frac{1.14}{0.158 * 0.089} = 81.63$$

Trimming of Weights

As soon as the calculation and the adjustment of the weights are done in order to compensate for imperfections, it is prudent to examine the distribution of the adjusted weights. Weights that are exceptionally large, even if they affect only a small portion of sampled cases, can cause a considerable increase in the variance of survey estimates. A usual practice is to trim extreme weights to a maximum value, with the intention of limiting the associated variation in the weights (thus reducing the variance of survey estimates), and avoiding that a small number of sampled units could determine the overall estimate. In other words, trimming weights replaces outlier weights to reduce the variance of the resulting estimations. However, this introduces some bias in the estimates and needs to be carefully considered against increases in precision.

For a stratified design, the process of weight trimming should be done within each stratum. At first, it is necessary to identify an upper bound for the original weights, and, subsequently, adjust the entire set of weights so that the sum of the trimmed weights is the same as the sum of the original weights.

Assume that w_{hi} is the final weight for the i^{th} unit in stratum h, and w_{hB} is the upper bound for the weights specified for stratum h. The trimmed weight for the i^{th} sampled unit in stratum h is, thus:

$$w_{hi(T)} \begin{cases} w_{hi} & if\ w_{hi} < w_{hB} \\ w_{hB} & if\ \ w_{hi} \geq w_{hB} \end{cases}$$

Subsequently, the trimmed weights for the whole sample has to be adjusted so that their sum is equal to the sum of the original weights.[4] Assuming that F_T denotes the ratio of the sum of the original weights to the sum of the trimmed weights, as:

$$F_T = \frac{\sum_h n_h w_h}{\sum_h n_h w_{h(T)}}$$

$$F_T = \frac{\text{Sum of weights}}{\text{Sum of triggered weigh}}$$

The final (adjusted) trimmed weight for the h^{th} stratum is defined as:

$w^*_{h(T)} = F_T\, w_{h(T)}$

and, thus, we obtain:

$$\sum_h n_h w^*_{h(T)} = \sum_h n_h w_h$$

Let's recall the example shown above in Table 2, and extend it with two extra columns in order to illustrate the trimming procedure (Table 7bis).

[4] To simplify the discussion, let's assume constant weights within strata in order to drop the subscript *i*.

Table 7bis. Weights calculation based on two analytic domains

Stratum	Unique PSU ID	PSU #	Census Pop.	Census HHs	Average Census HH Size	Cluster Type 1=Rural, 2=Urban	# HHs in PSU from Listing
1	101001	1	482	96	5.0	2	94
1	101011	2	429	98	4.4	2	90
…	…	…	…	…	…	…	…
1	102007	20	43	8	5.4	2	17

# HHs Selected for Interview	#HHs Interviewed	Prob. of Selection in 1st Stage P_1	Prob. of Selection in 2nd Stage P_2	Non-Response Adjustment	Weight	Weight Trimmed	Final HH Weight
8	8	0.154	0.085	1	76.16	76.15	80.10
8	7	0.158	0.089	1.14	81.63	81.63	85.86
…	…	…	…	…	…	…	…
8	7	0.01	0.47	1.14	188.88	103.70	109.08

In stratum 1:

Total PSUs: 184. Total HHs: 12,444. Total Population: 61,468. # PSUs selected in stratum 1 cluster 2: 20. Upper bound for weights: 172.70. Replace Value: 103.70. Sum of weights: 1728.09. Sum of triggered weights: 1642.91.

In this example, we identify outlier weights. More specifically, in case the weight is above the maximum weight of 172.70 (the value selected considering the weights in the 99th percentile) the original weight is truncated and replaced with 103.70 (the highest value of the weights in the 98th percentile) as presented in the table above. In this case, the ratio of the sum of the original weights to the sum of the trimmed weights is:

$$F_T = \frac{\sum_h n_h w_h}{\sum_h n_h w_{h(T)}} = \frac{\text{Sum of weights}}{\text{Sum of triggered weights}} = \frac{1728.09}{1642.91} = 1.05$$

Finally, trimmed weights are re-scaled so that they sum up to the original value $\sum_h n_h w_h = 1728.0$, by multiplying each weight $w_{h(T)}$ by F_T =1.05. The final household weight is thus calculated as follow:

$w^*_{h(T)} = F_T * w_{h(T)}$

Adjustments for post-stratification

Weighting for household surveys involves a last process, which is represented by adjustments for post-stratification.

To sum up, sample weights is an essential part of the analysis of household survey data worldwide, in particular in less developed countries. As argued, the introduction of weights reduces biases due to imperfections in the sample related to various kinds of error due to the impossibility of a designed survey to achieve information from some units in the target population. The usage of sampling weights complicates the survey process in many ways. For instance, weights have to be calculated for each stage of sample selection; subsequently, they have to be adjusted to account for different kinds of imperfections in the sample; and, finally, the weights have to be recorded and used appropriately in all subsequent data analyses.

Reference

Grosh, M. E., & Muñoz, J. (1996), A manual for planning and implementing the Living standards measurement study survey. Washington, D.C.: World Bank.

United Nations (2011), Designing household survey samples: Practical guidelines. New York: United Nations Publications. United Nations. (2003). Sampling strategies. New York: United Nations Publications.

ICF International (2012), *Demographic and Health Survey Sampling and Household Listing Manual*. MEASURE DHS, Calverton, Maryland, U.S.A.: ICF International