SAMPLING METHODOLOGY AND SIZE

Step 1: Select the adequate sampling methodology

Beforehand, the survey population as well as the sampling unit must be known: the survey population will usually be the refugee population of a camp/site (but can in some cases include host population), and for WASH KAP Surveys, the sampling unit is always the household.

It is very important to make sure that the households surveyed during the process are selected randomly. The more randomly the households are selected, the more representative the results will be of the whole camp/site. There are different ways of sampling, depending on the information you already have on individual households. The details on how to select the right option are depicted in the following figure:

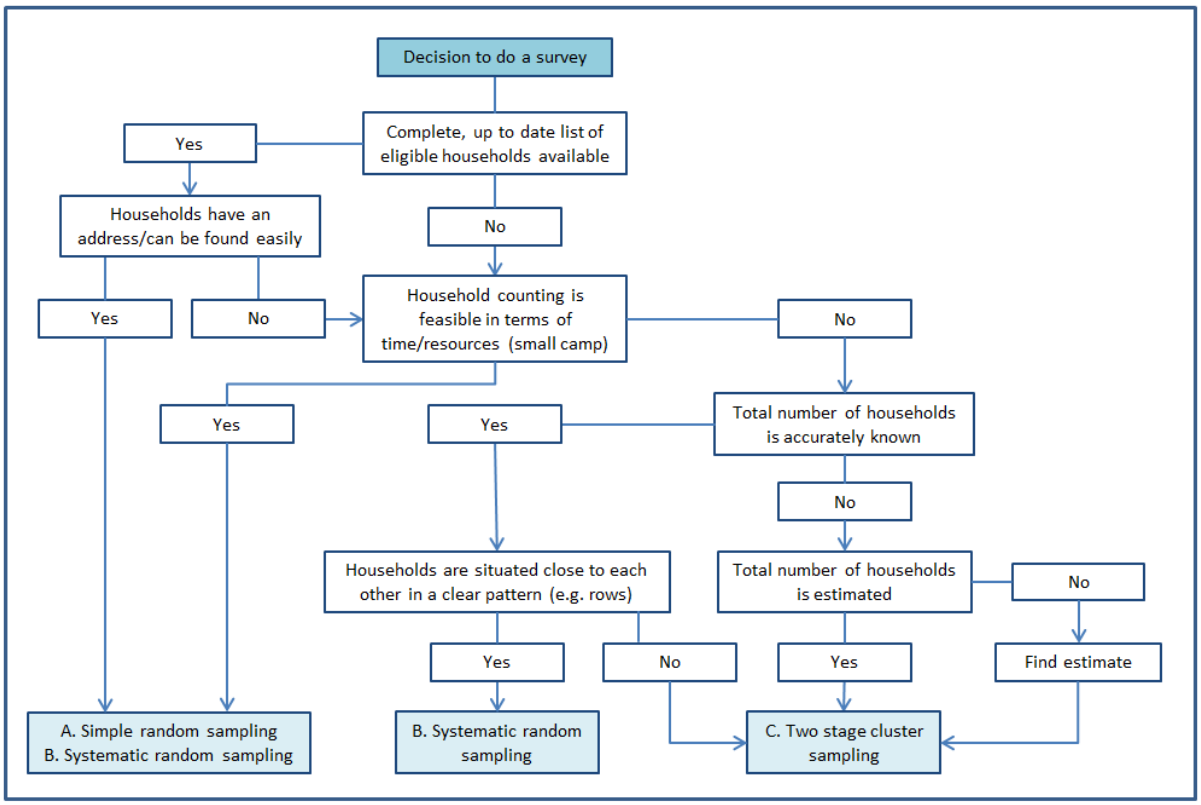


Figure 1: Random sampling methodologies

In the case of random sampling (simple and systematic), an up to date list of all the households in the camp must be available or derivable from a map, or the total number of households must be known accurately. If this information is not available, an exhaustive counting and chalk-marking of the households can be organized (pending available resources). These two methods are the best options for camps, and should be targeted as they will allow saving time and precious resources, while being easily applicable on the ground.

In case of settlements or very large camps where population information is unavailable, or where households are not organized in a structured way, cluster sampling will likely be used.

Step 2: Calculate the sample size of a WASH KAP Survey

1) Sample size calculation

The first step for obtaining a sample for the Standardized WASH KAP Survey is to calculate its size. The sample size (number of households to be surveyed) is calculated using the recommended (and widely used) formula below:

$$n = \frac{t^2 \times p \times q}{d^2} \times d_{eff}$$

With:

- n being the calculated sample size
- t being the error risk parameter (use 1.96, for a confidence interval of 95%)
- p being the expected prevalence (use 0.5 - 50% prevalence - in normal situations)
- q = 1-p is the expected non-prevalence (which is 50% in normal situations)
- d being the relative desired precision (for simple/systematic random sampling, use 5% precision in normal situations, 10% in some cases)
- d_{eff} being the design effect in case of cluster sampling (use 1 for random sampling, 2 for cluster sampling)

The calculated sample size then needs to be adjusted based on the total number of households and the anticipated non-response. The different parameters of the formula and the adjustments are further explained below.

Under normal conditions, the most common sample sizes are the following:

- 360 households for random sampling with 5% precision
- 100 households for random sampling with 10% precision (should be used only in case of important resources limitations – doesn't allow intra-camp comparisons)
- 210 households for cluster sampling

The calculated sample size for cluster sampling will usually be around 200. However, since we use 30 clusters, the total sample size must be a multiple of 30: this is the reason why 210 households is the commonly used sample size for cluster sampling.

In case of very large camps that are split in well-known and clear sub-areas (not more than 5) it can make more sense to do a separate survey (10% precision, 100 households) per sub-area then a general survey (5% precision, 400 households) in the whole camp. That will allow for better comparison of different levels of WASH services between the sub-areas, while keeping sample sizes of a similar magnitude.

- **Expected prevalence of the indicators in the population of interest (p)**

This is based on the expected answers to the survey's questions. If answers to one question is expected to be in general evenly spread (e.g. 50% no and 50% yes, for that specific question), then

the expected prevalence should be 50%. If you expect all the answers to be less balanced (40%-60% or 30%-70%) then the expected prevalence will be 40% or 30%.
The prevalence used to calculate the sample size should be based on the expected prevalence of the major indicators while favoring those that are closer to 50%. Indeed, using a prevalence of 50% means that all your indicators will be estimated adequately (those that have a prevalence of 50% as well as all the other ones). On the other hand, if you use a prevalence of 30%, indicators that are usually 50%-50% or 40%-60% will not be calculated with enough accuracy, as the sample size will not be big enough.
Though the closer to 50% the expected prevalence is the bigger the sample size will be, it is recommended to use a prevalence of 50% in a WASH KAP Survey.

- **Precision of measuring the survey estimate (d)**

In that cell must be entered the precision with which we want to have the survey results. A high precision (under 5%) will considerably increase the sample size, whereas a low precision (10%) will decrease it. The precision selected for the survey should be a balance between ensuring that key indicators are estimated with adequate precision while being feasible in light of available programmatic resources. A precision of 5% is recommended when possible for WASH KAP surveys, while if available resources are insufficient a precision of 10% can be used for this type of household survey.

Precision backward adjustment (absolute precision):

The sample size obtained with the chosen precision is usually based on a conservative 50% prevalence. However, for indicators for which the prevalence ends up being lower (e.g. 25% of respondents have access to soap) while the sample size was calculated for 50% prevalence, the precision will increase for those particular indicators.

It is possible to adjust the precision of specific indicators after the survey for those indicators where the prevalence was lower than 50%. Based on the same formula, the precision for these indicators can be adjusted using:

$$d = t \times \sqrt{\frac{p \times q \times d_{eff}}{n_{sur}}}$$

With n_{sur} being the number of surveys actually collected during the survey.

- **Design effect in case of cluster sampling (d_{eff})**

The design effect's purpose is to adjust upward the sample size in case of cluster sampling - usually to account for differences within a camp/site population (these differences can be differences between an old caseload and new arrivals or between communities with different ethnic origins and thus different practices etc…).
In case of simple or systematic random sampling, the design effect will always be 1. When using cluster sampling, the design effect will be 2. This is a conservative value that will work for all the indicators.

2) Sample size adjusted to the size of the camp/site (number of households)

The sample size calculated must then be adjusted to the camp population (total number of households in that camp). This doesn't change much the sample size in very large camps, but can be beneficial in smaller camps (less than 5'000 households for example) as it will reduce the sample size and can save time, energy and resources on the field. The adjustment formula is the following:

$$n_b = \frac{n \times N}{n + N - 1}$$

With:

- n_b being the sample size adjusted to the size of the site
- N being the site total number of households

3) Sample size adjustment to anticipated non-response rate

Once you have calculated your sample size, it needs to be adjusted again upwards to account for the expected non-response rate. This is to make sure that at the end of the survey we will have the required number of filled forms. The formula used for that is detailed below:

$$n_{fin} = \frac{n_b}{1 - r}$$

With:

- n_{fin} being the adjusted calculated sample size taking into account expected non-response rate
- r being the expected non-response rate

The expected non-response rate is the proportion of the households we expect to be unavailable, or refuse to participate. If we expect that 5% of the households (1 out of 20) will not be available or refuse to participate, the expected non-response rate is 5%. If we expect that 1 out of 10 households will not participate, the non-response rate would be 10%. The anticipated non-response rate can be based on previous year's experiences, but additional factors need to be weighed in such as seasonal migrations. If you have no such information, you can safely use 5%.

Please note that the final calculated sample size will not automatically give you representativeness for all the questions, but only for those questions that can be answered by each and every household.
For example, questions on feminine hygiene will be answered only by households where women are present, which will be less than the required sample size for representativeness; specific questions on type of water treatment used will be answered only by those households who treat water, which is not the whole initial sample, etc.
If you absolutely need representativeness for these questions, the initial sample size would need to be increased accordingly.
It is however recommended not to do this, but to instead handle the results from non-representative sub-samples with additional care.

4) Excel ready to use sample size calculator

For easy calculation of the sample size (number of households to be surveyed) using the methodology described above, the Excel tool 'Sample Size Calculator' can be used and will automatically generate your sample size depending on the parameters you have entered. Once all the parameters have been entered in the cells highlighted in light blue, you will get your final sample size in the cell J4 which has bold characters:

	A	B	C	D	E	F	G	H	I	J
1										
2		Sample Size Calculator								
3		Confidence interval	Expected prevalence of the indicators in the population of interest	Precision (+/-) of measuring the survey estimate	Design effect in case of cluster sampling (otherwise leave 1.0)	Sample size (number of households) needed	Total number of households on the site	Sample size adjusted to the total number of households	Anticipated non-response rate	**Sample size adjusted for anticipated non-response**
4		95%	50%	5%	1.0	385	5,000	358	5%	**377**
5										

Figure 2: Excel Sample Size Calculator

The calculated sample size can then be rounded up upwards if need be, but not downwards.

Step 3: Random sampling of the households to be surveyed

N.B: This methodology is applicable to refugee camps, sites and in some cases settlements, but needs to be adapted for urban areas.

Now that you know the sample size of the number of households for the survey, it is important to make sure that the households surveyed during the process are selected randomly. Depending on the sampling methodology selected in Step 1, follow the instructions below to randomly select which individual households will be surveyed.

A. Simple random sampling

For simple random sampling, you need to have a complete and up to date list of all the households in the camp/site. This list should at least have a precise address linked to each household, so that it can be found easily during the survey (or if not an address, surveyors should have another way of finding it easily). With this method, the list doesn't necessarily need to be in order (households that are next to each other geographically don't need to be next to each other in the list). If no list is available but household counting for all the camp is feasible resources-wise, than simple random sampling can be used as well. The surveyors will count and number all the households, marking them with chalk or marker.
If a list is used, each household in that list is given a number, and the households needed to get to the desired sample size are selected randomly from the list. This allows for an equal probability for each household to be selected for the survey, in a completely independent way.

Simple random sampling can be done easily using Excel as described in **Annex A**.

B. Systematic random sampling

In order to do systematic random sampling, you need to either have a complete and up to date list of all the households in the camp/site that needs to be surveyed, or know accurately the total

number of households. A sample frame can also be obtained by using an up to date map of the camp, with sufficiently high definition to be able to single out households.

If you don't have a list and just know the total number of households, those households need to be well organized, for example in blocks or rows. .

First, the sampling step must be calculated using the following formula:

$$k = \frac{N}{n_{fin}}$$

With:

- k being the sampling step (or sampling interval)
- N being the site total number of households
- n_{fin} being the adjusted sample size

The first household to be surveyed must be randomly selected between 1 and k. Then, to select the remaining houses to be surveyed, every k^{th} household starting from the initial randomly selected household is selected (as k will likely not be an integer, each non-integer selected must be rounded up to the nearest integer).

It is important to beware of existing patterns concealed behind the skipping pattern as this could affect the randomness of the selection: for example, if every 10^{th} household is selected and it turns out that in that specific camp every 10^{th} household is at the corner of a neighbourhood, all households selected could be corner households and this would affect the randomness of the survey results as these households might have distinct specificities in the levels of WASH services they receive.

Like simple random sampling, systematic random sampling gives every household an equal chance of being selected. It is recommended to use this method whenever possible. In order to achieve that, one should try to get access to one of the following prior to a WASH KAP Survey:

- Up to date list of households in the site
- Up to date map of the site
- Counting of the households, marked with a number using chalk for field recognition

If your camp is organized by known groups or strata (e.g. geographically defined units such as camp areas or sectors where refugees are assigned to live) then you may wish to stratify your sample accordingly. This can be done both for simple and systematic random sampling, simply by apportioning the sample size to those sub-areas based on their respective numbers of households. For example, if you need a sample of 360 households and the camp is divided in two parts with 1200 and 2400 households, you would select respectively 120 and 240 households in those two parts. This is especially relevant when doing systematic random sampling without using a list, and just using the configuration of the camp. Surveyors will work in different areas, with sampling steps calculated for these areas based on their respective sample size and total household size.

Systematic random sampling with list can be done easily using Excel as described in **Annex B**.

C. Two stage cluster sampling (Most likely option for a settlement)

In the case where it is impossible to gather a list of all the households, get an accurate total number of these households, or if the camp is very large and the population is dispersed, cluster sampling is the method that needs to be applied, as it only requires an estimate of the total number of households. With this method, rather than going to so many different locations to get one household each time, surveyors will go to fewer locations and conduct multiple surveys at each stop (cluster).

It is the most frequently used method of random sampling for surveys in the field. This method is done in 2 stages, described below. Those steps can appear to be cumbersome, but they are required to achieve a good randomness of the sample and therefore obtain valid results for the survey.

First of all, it is necessary to split the camp/site in known geographic areas or sectors that have known population figures. This is especially the case for large camps. If you don't have data on the population of the different areas the camp is split in, good estimates can be used. These areas should be large enough to have at least 7 households (this is further explained below). Areas should have clearly defined boundaries such that the survey team will be able to select households from an area and not the adjacent one.

The first stage is to equally allocate clusters within those areas (clusters are sub-areas where the interviews will be held). This is done by using probability proportional to population size method, so that every household in every area of the camp has an equal chance of being selected, whether it is a small sized or large sized area.

This technique needs the estimated cumulative number of households for all the areas to be calculated in a table. To each area a range will be allocated, which will be between the cumulative population sizes from the area just before on the list +1, up to the cumulative population size including that area. A cluster sampling interval is then calculated using the formula below:

$$i = \frac{N}{c}$$

With:

- i being the cluster sampling interval
- N being the total number of households (cumulated number for all the areas)
- c being the number of clusters needed

A random starting point is selected between 1 and i. This cluster will be allocated to the area for which the number of this starting point is situated in the range of cumulative household numbers of that area. The sampling interval will then be added to that starting point, allocating a second cluster, and that process is continued until all the clusters are allocated to areas. Some large areas might have multiple clusters allocated to them, while some smaller areas might have no clusters allocated at all.

Cluster allocation is further explained and can be done easily using Excel, as described in **Annex C**.

Once all the clusters are allocated, the second stage is to randomly select the households within each cluster. The method of household selection described below is one of many. Other methods can be used as long as they guarantee equal chance of being selected to households in the clusters.

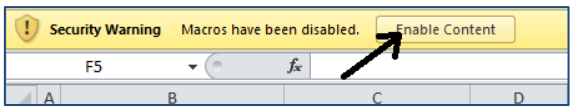
If one area was allocated one cluster, the number of households surveyed for this cluster will be selected randomly from all the households in that area. If multiple clusters were allocated to a single area, then that area will need to be segmented in the correspondent number of sub-areas, by using existing administrative sub-divisions or by using geographical features (roads, paths, rivers or riverbeds, hills, etc.). Any available map will help with that purpose. The number of households must be as equal as possible in each of these sub-areas. The surveys needed from each cluster will then be selected randomly from the households in the respective sub-area.
In case the number of sub-areas in a given area is larger than the number of clusters allocated to that area, then the sub-areas that will represent a cluster must be selected randomly from all the sub-areas in that area.
In order to select the households for a cluster, the surveyors must go to the centre of the sub-area represented by that cluster. They will spin a pen or a bottle on a flat surface on the ground, and then follow the direction indicated until the border of the sub-area, counting and numbering the households as they go. They will then sort randomly one of these households, and do the first interview there. They will then go to the second closest household for the second interview, then from that household select the second closest household for the third and so on (being careful not to go out of the limits of the sub-area, and not to interview twice the same household) until they have surveyed the 7 households required.

Please note that using any of the 3 methods described in this manual, absent households must not be replaced by the closest available household. It must be marked, and given a second or third visit at different times/days, after what it can be recorded as non-respondent (this is taken into account in the adjustment of the sample size to anticipated non-response rate).

Annex A – Using Excel to do simple random sampling

For simple random sampling, you need to have an up to date list of all the existing households. The first step is to allocate a number to every household, from 1 to N, N being the total number of households.

Open the Excel file '1c – Simple Random Sampling' in the sampling toolbox, and enable the macros by clicking on the "enable content" bottom as shown below:



Once you have done that, fill in the cell in light blue with the number of households in the camp. That is your total number of households N. The list of households will automatically appear in one column (B), and each household will be allocated a random number.

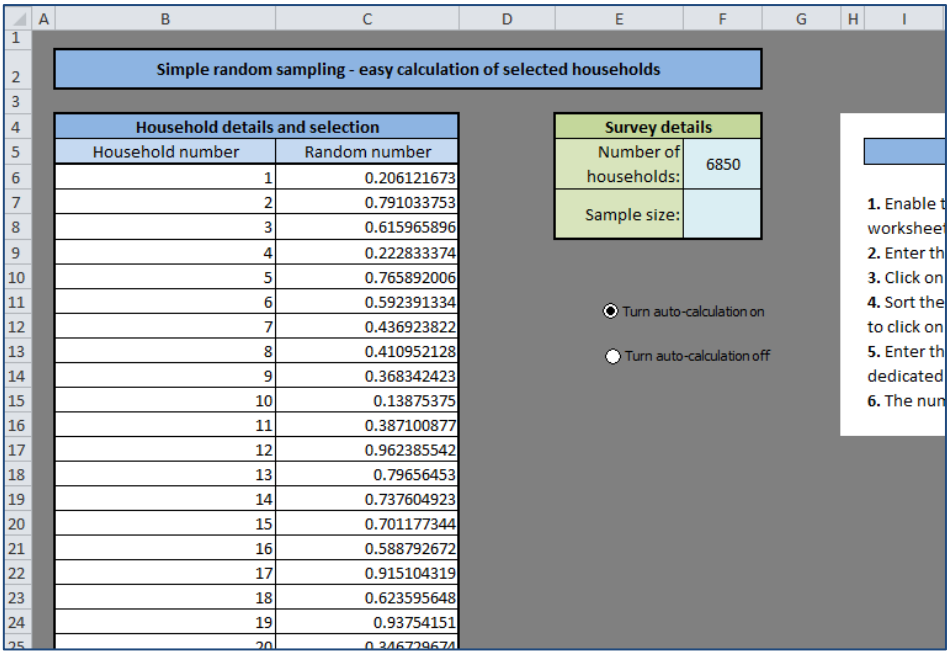
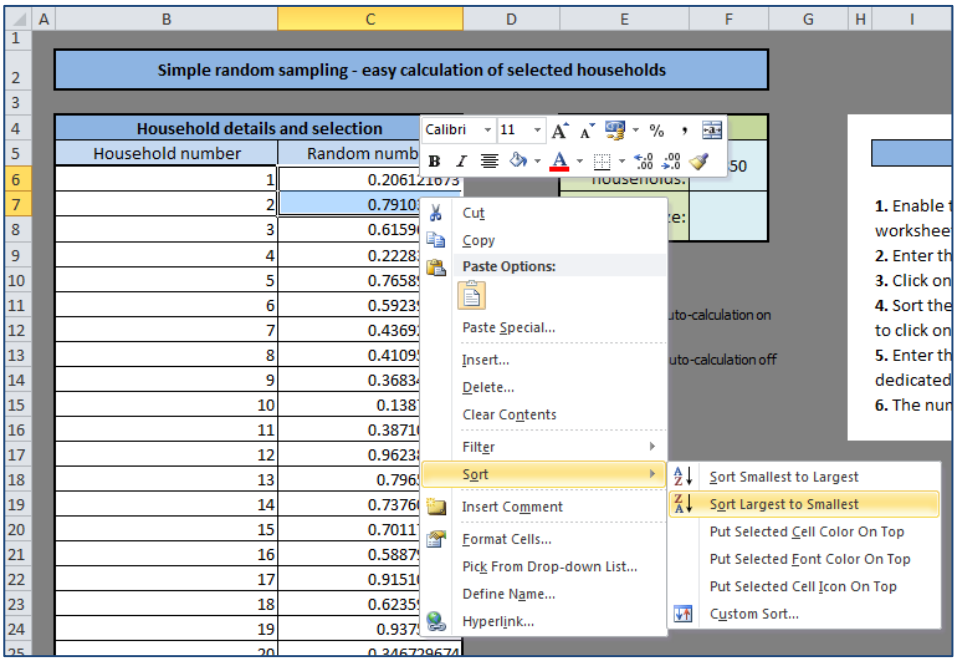


Figure 3: Production of random numbers

As random numbers are recalculated every time the worksheet changes (basically every time you use a command on the worksheet) it is important to freeze them by clicking on the "Turn auto-calculation off" button in the worksheet.

To reorganize randomly your list of households, it must be sorted by random number from the largest to the smallest. In order to do so, select the first two random numbers from the list, right-click and click on 'Sort' and 'Sort Largest to Smallest'. A box will appear, click on 'Expand the selection' (see figure below).



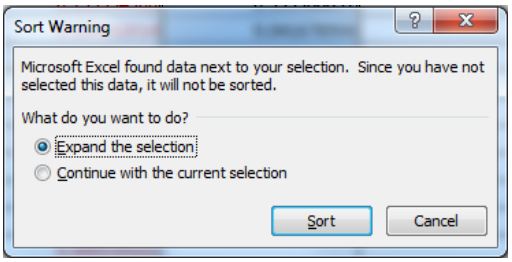


Figure 4: Sorting the random numbers

Now your households are randomly reorganized. In order to highlight the numbers of the households selected for the survey, you can enter the sample size in the second light blue cell. This will highlight in purple the households selected for the survey (that will be the first n_{fin} households of the sorted list (n_{fin} being the adjusted sample size).

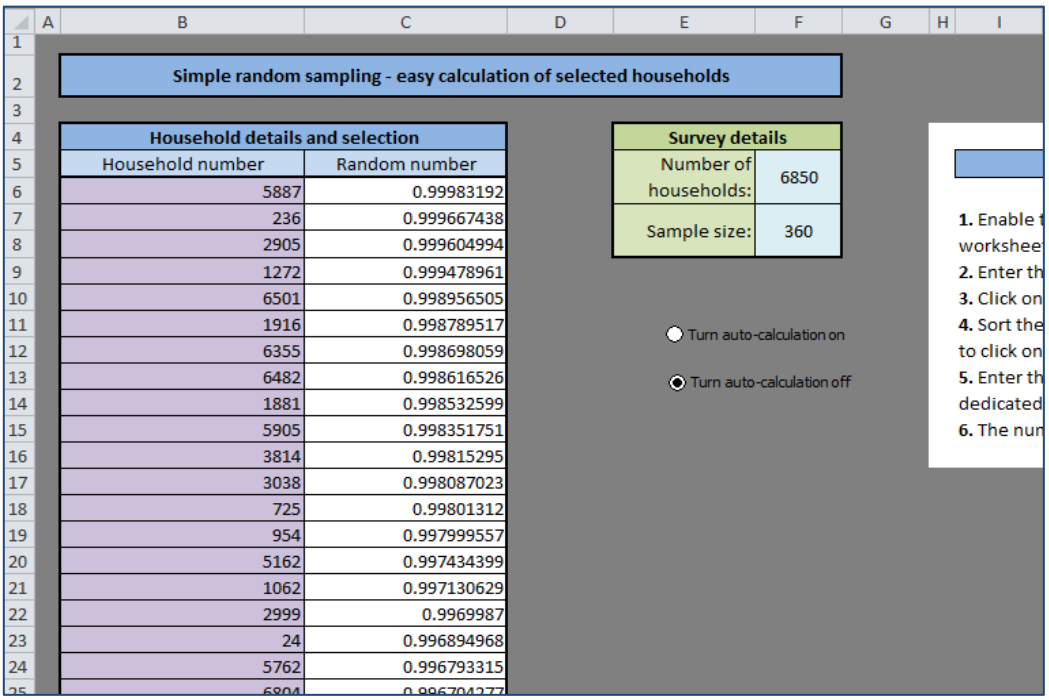
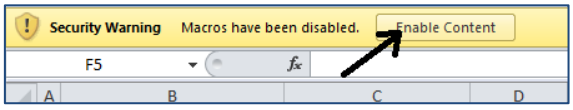


Figure 5: Highlighting the selected households

Annex B – Using Excel to do systematic random sampling <u>with list</u>

You need to have an orderly list of all the households (or numbers of all the households) so that the numbers produced by Excel can be linked easily to households. This means that like for simple random sampling, each of the household is allocated a number between 1 and N (N being the total number of households) in an orderly manner.

Open the Excel file '1d – Systematic Random Sampling' in the sampling toolbox, and enable the macros by clicking on the "enable content" bottom as shown below:



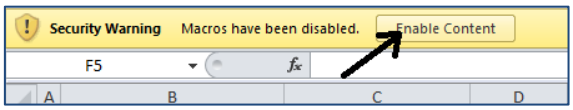
The two cells highlighted in light blue must then be filled: total number of households (N) and the sample size (n) that you have calculated for this survey. The sampling step and a random starting

point will be calculated automatically, and all the selected households will appear in the table in the table.
You can then freeze that selection by clicking on the "Turn auto-calculation off" button in the worksheet. This will prevent the file from recalculating the random starting point (thus changing the entire selection) every time you type a command on the worksheet.

Annex C – Using Excel to do the first step of two stage cluster sampling

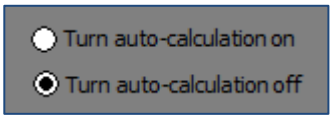
With this method, clusters must be allocated to the areas you have defined within the camp. The sizes of these areas are taken into account, while still giving every single house in the camp the same chance of being selected. Some areas might be allocated multiple clusters, while other areas will be allocated no clusters at all.

Open the Excel file '1e - Two stage random sampling' in the sampling toolbox, and enable the macros by clicking on the "enable content" bottom as shown below:



Then fill in the two left columns (B and C) in light blue the information you have on all the different sectors/areas names and their respective population figures (in terms of household numbers).

Once you have done that, the cumulative and total number of households are automatically calculated. You then just need to enter in the cell in light blue "Number of clusters" the number of clusters you have chosen for this survey. It is usually 30.
Once this number is entered, the clusters are automatically (and randomly) allocated to the different areas you have entered in the worksheet. The number of clusters per area is given in the column AS, per camp area you have entered. As random functions are involved in this worksheet, you may want to click on the "Turn auto-calculation off" button to freeze the result.



Now you have a camp/site geographically separated in distinct areas, and for each area a number of clusters (or no cluster at all) assigned based on the number of households of each area. You can now move on to the second stage of this method as described in part C of step 3 in this document.

Additional information on random sampling

If you want additional information on sampling size calculation or random sampling methods, please visit the SMART Methodology website on the following link:
http://smartmethodology.org/survey-planning-tools/
It will give you access to a detailed manual and to the ENA software used for nutrition surveys.