



FAO Standard Seed Security Assessment

SAMPLING AND SAMPLE SIZE

Sampling is the selection of a representative part of a population in order to determine parameters or characteristics of the whole population. A **sample** therefore is the proportion of the population selected for the study or investigation. A **Sampling unit** is that unit about which information is collected and that provides the basis of analysis. In survey research, elements are people or certain types of people. **Sampling methodology** or design refers to a set of rules or procedures that specify how a sample is to be selected. This can either be probability or non-probability. In probability sampling it is possible to assign levels of confidence about the reliability of the sample from a statistical point of view. In non probability sampling it is not. In SSA, households are selected using probability sampling, whereas all other sampling units (individuals for community FGD, traders, officials, NGO representatives, seed producers) are not. In the case of units, representativeness relies upon purposive sampling.

In probability sampling, sampling is often done so as to ensure a 95% level of confidence in the estimates derived from the sample. The number of units required to achieve this will increase as the heterogeneity within the overall population increases. Thus the sample required for 95% confidence in a population with consists of residents, IDPs and refugees will probably be higher than a population consisting of just residents. In situations where there is heterogeneity, separate samples would need to be calculated for each population group, increasing the size of the overall sample.

As a rule of thumb, in relatively homogenous population groups of 10,000 units or more which are normally distributed, the sample sizes necessary to achieve different kinds of accuracy are given in table 4.1.

Table 4.1 Level of reliability and Sample size for a population above 10,000 households

Level of reliability	75%	80%	85%	90%	95%
Sample size	133	165	210	270	384

When the population size is smaller than 10,000, there will be some reduction in the sample size, as shown in Table 4.2.

Table 4.2. Reliability level and population adjusted sample size

Population size	Reliability level				
	75%	80%	85%	90%	95%
500	105	124	148	176	217
1,000	117	142	174	213	278
2,000	125	152	190	238	322
4,000	129	159	200	253	350
6,000	130	161	203	258	361
8,000	131	162	205	261	366
10,000	131	162	206	263	370
20,000	132	164	208	266	377
50,000	133	164	209	269	381
100,000	133	165	210	270	383
200,000	133	165	210	270	383
500,000	133	165	210	270	384

At the end the final sample size will be a trade-off between the theoretical best practices, applying the above mentioned formulas and the available resources – human, financial, logistics and time. This trade off will influence the final level of reliability at the selected site level. The final decision on the sample size during surveys preparatory phase often requires the responsible person to make a decision as he/she has to make a tradeoff between those two components (statistically optimal sample size and available resources).

Once the overall sample size has been determined at one administrative level (eg regional/provincial level), this sample size can be divided proportionally to the populations of entities at the next administrative level down (e.g. Counties or district). If geographical entities at the lower administrative level have similar population sizes, then the overall sample size can be divided equally among the selected entities. The different sizes of sample will have implications for the time required to cover each district, as illustrated in Box 4.1:

Box 4.1 Example on determining sample size at county or district level

- a) The overall sample size at regional level is 384, and within the region, three counties have been selected to represent the various agro-ecologies. The size of each sub-sample is determined by the relative size of the population in each of the districts. If the populations of the districts are roughly equal then a sample of 128 households needs to be interviewed from each of the three counties.
- b) On average, a team of five is able to interview 20 households, 3-5 LM traders and 1-3 agro vet. Therefore, each team will need a minimum of $(126/20)$ 7 working days to cover each district
- c) If the three counties on the other hand have proportionally very different population size, Say $X=9000$; $Y= 7000$ and $Z= 4000$ (Total 20000), then the sample size for each district would be:
 - $X = 9/20 \times 384 = 173$
 - $Y = 7/20 \times 384 = 134$
 - $Z = 4/20 \times 384 = 77$

Therefore, a team of five going to X will require 9 days $(173/20)$, Y will require 7 days $(134/20)$ and Z will require 4 days $(77/20)$.

2.5.5 Sampling methods at community level for Household Questionnaire

Sampling of households within a village after determination of sample size is done by use of probability sampling. Simple **random sampling** would be the most appropriate process, as each of the potential sampling unit (households) within the population has the same opportunity to be selected under this approach. This however is time consuming, especially where the population is dispersed over a wide geographic area. **In most cases systematic random sampling** may be preferred for practical reasons.

Box 4.2. Example on systematic sampling

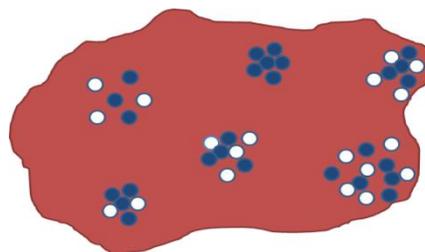
In a systematic sampling, a prefixed pattern is normally used, to logically skip certain number of household at an interval. If for instance it is estimated that there are 20 households along a given transect, and the enumerator moving along that transect expect to interview 5 households along that transect, then he/she can determine the sample interval (SI) as follows

$$SI = \frac{\text{Number of households (e.g. 20) along the transect}}{\text{Number of households (5) to be interviewed}} = 4$$

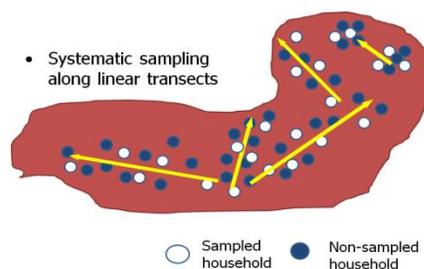
The sampling interval therefore is 4, and the enumerator will therefore skip 4 households after every sample taken. The distance to be covered by a numerator walking along the transect needs to be put into consideration. Therefore, transect needs not to be more than 3 km.

The pattern of settlements and the heterogeneity of the village will determine the type of systematic sampling. The pattern of settlements can be noticed as the team drives into the village and or through some quick questioning of local residents. Generally speaking, there are basically three common types of settlements:

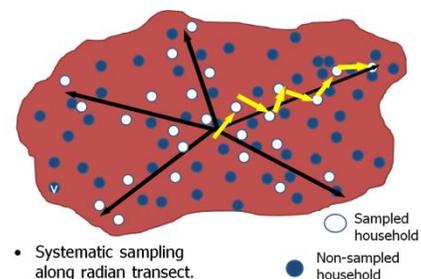
- a) *Clustered households* – this consists of groups of 4-10 households clustered within a very close proximity (radius of 20-10m) to each other and separated with a reasonable distance (200-800m) from another cluster (Fig. 4.1). This is very common in agro-pastoral communities. Here, the most appropriate method is to randomly sample 4-8 clusters and within each cluster, randomly sample 3-5 households.



- b) *Linear settlement* – Here the households normally settle along natural or man-made features such as at the foot of the mountains, along the river beds, along the trunk roads. However, crop (cultivation) fields might be scattered within a village. Here systematic sampling (Box 4.2) is normally deployed along the linear pattern of settlement. As the team drives into the village, they need to take note of how far the households are from one another as this will also influence the number of households that could be interviewed in that village/location. The wider the distance between neighboring households, the more time needed by the enumerators to move from one household to another.



c) *Scattered households* – Here there is no well recognized pattern of settlements. The households are considered randomly scattered with the village. Here, like in linear settlement, the team needs to take note of how far the households are from one another as also. The most appropriate sampling method is the systematic sampling along 3-6 radian transects depending on the number of enumerators. Each of the enumerators could move long one transect, sampling households at regular intervals.



NOTE: These sampling schemes assume homogeneity. If the village is split into different groups (e.g. IDPs, returnees and residents) then this would ideally require sub-sampling.