RMS Basic Sampling Guidance

Sampling describes the selection of a subset of individuals or households from a population of interest in a controlled way that allows us to make generalisations about the characteristics of that population without interviewing everyone.

Probabilistic sampling as a controlled procedure with known selection probabilities of study units is needed to achieve statistically representative results that are reliable and likely to reflect the true indicator value in the whole population. While **non-probabilistic sampling** techniques like snowball sampling are sometimes proposed for quantitative studies, **their use can normally not be recommended** if the goal is to produce summary statistics or indicators about a pre-defined population.

There is a multitude of ways to select a probabilistic sample. All of these follow the same principles, namely that the selection probabilities of study units are known, and that these selection probabilities are positive for all units in the population. While tools like the RMS sampling decision tree and sample size calculators can assist in defining a sampling scheme, sampling is a highly technical matter and will normally require support from technical experts for example from the regional DIMAs.

The way a sample is drawn has major implications for the usability and usefulness of survey data, and careful sample size calculations can lead to considerably more efficient designs and cost savings than ad-hoc and poorly thought through designs.

The **sample size** for a household survey depends on several factors, including the study design, the level of disaggregation required, the pre-defined precision goals such as the margin of error, and the sampling technique used. However, it is important to note that the sample size is not always the main driver of cost, and a **larger population does not necessarily require a higher sample size**. Study goals, precision goals and disaggregation requirements need to be well established and agreed on before statisticians can calculate the required sample size for a survey.

After defining study population and study goals, **identifying a viable sample frame** is one of the most important steps in building a sampling scheme (note that for the RMS, a template is available to help you with exploring sample frames).

A **sample frame** is the information that is being used to select units to be surveyed from the population of interest. Ideally, we would have a complete and up-to-date registration list of the population of interest as sampling frame. From this list, we can randomly select households or individuals to be included in their sample, ensuring that each person or household has an equal probability of being chosen in a so-called **simple random sample** (SRS). While in UNHCR we do sometimes have such a list-based sampling frame in proGres or government-maintained administrative records, in practice this is not always possible. For example, the list may be incomplete or outdated, or we might be interviewing populations not registered in proGres. In these cases, we need other methods to select our sample.

One common approach if a simple random sample is not an option is **multi-stage cluster sampling**: we divide the population into geographic clusters such as villages or camp blocks, randomly select a sample of these clusters, and then sample individuals or households in each selected cluster (or, depending on the sampling design, interview everyone in the selected clusters). This can be more cost-effective than trying to sample individuals directly because we have to visit only the selected clusters. In order to ensure that the sample is still representative of the population as a whole, we will need to calculate the probability of each person being selected and use this information to weight the results. The sampling frame in this case consists of a list of clusters and (estimated) population sizes per cluster, while sampling of respondents within selected clusters can be done with random route or listing procedures. Other probabilistic sampling techniques include **respondent-driven** and **time-location sampling**.

In **stratified sampling**, the population is divided into subgroups (strata) based on relevant characteristics, and a random sample is drawn from each stratum. It can be combined with other techniques such as cluster sampling and SRS. This method is useful when the population is heterogeneous and there is a need to guarantee precision goals for specific sub-groups for example by age and gender. It is important to note that representativeness as a statistical quality does not guarantee inclusiveness.

There are **several sources of bias that can affect the representativeness of survey results**, even when probabilistic sampling is used. For example, matching the sample to the population in terms of disaggregation criteria such as gender or household composition does not guarantee that the sample will be representative. In addition, the size of the sample is not directly related to its representativeness, and data even from very large samples can yield biased results if selection and response probabilities have been calculated incorrectly. Another challenge we may face is access limitations to certain regions. For instance, it may be difficult or impossible to collect data from certain areas due to security concerns or other factors. In these cases, the sample may not be fully representative of the population, and the results of the study may be less generalisable.

To improve the representativeness of results, we use a variety of **techniques to adjust for known differences between the sample and the population**. These techniques include survey weights, which are used to account for the sample design, and post-stratification, which is used to adjust for differences between the sample and the population that are not accounted for by the sample design. These adjustments can help to reduce bias and improve the reliability of the results. Overall, careful planning and data analysis are essential for obtaining representative results from survey data.

Please note: An RMS sampling decision tree is available to serve as a methodological framework for RMS quality assurance as well as being a tool for colleagues with expert statistical / survey expertise to decide on most appropriate methodological approaches.