**Table 10. Probabilistic sampling**

| | |
|---|---|
| Simple random sampling | The sampling frame includes all the population, and each member of the population has the same chance of being selected. |
| Systematic sampling | Like simple random sampling, respondents are selected but at regular intervals. For instance, families in a household survey, instead of being selected by randomly extracting the number of their house, might be selected at an interval of 10 houses after randomly selecting the first. |
| Stratified sampling | If a population presents mixed characteristics, it might be better to first divide it by these characteristics and then apply a simple random sampling to each group. For instance, instead of doing a simple random sampling on an entire IDP population in a district, a researcher might want to differentiate IDP communities in formal camps, IDP communities in informal settlements, etc. |
| Cluster sampling | Cluster sampling consists in the creation of subgroups, with each subgroup having features similar to those of the population. Instead of sampling individuals from each subgroup, the researcher randomly uses entire subgroups. For instance, instead of sampling an entire province's population, a researcher will first identify villages, then randomly select villages. |

**Table 11. Non-probabilistic sampling**[41]

| | |
|---|---|
| Convenience sampling | A sample includes individuals that the researcher can access easily – for instance, migrants and refugees crossing at an accessible flow-monitoring point. |
| Voluntary response sampling | It is still based on ease of access. The researcher does not identify the respondents. Rather, the respondents can volunteer. For instance, a researcher advertises in community centres that he/she is looking for candidates for a research. Or he/she publishes an online survey on social media to which anybody can reply. |
| Snowball sampling and response-driven sampling | It is based on chain referral and particularly useful with hard-to-reach populations. One respondent identifies another respondent, that respondent identifies another one, and so on. |
| Purposive sampling | The sampling is based on the judgement of the researcher, who wants to select a specific group functional to the research. For instance, a researcher might actively seek and include people with disabilities to focus on their specific experience. |



[41] The previous chapters stressed that for information management activities for research purposes, the involvement of VoTs or former VoTs is not a prerequisite and often not the most advisable solution, to mitigate protection risks. It is important to emphasize again that VoTs are a typical example of a hard-to-reach population (see Section 4.1), for which probabilistic sampling strategies are usually not applicable.

**Key simplified concepts to understand probabilistic sampling**

### What does *randomized* mean?

**Randomized does not mean by chance, and randomization always requires knowing the total number of the reference population**. In a simple randomized sample, all participants (out of the reference population) have the same probability of being selected, and that probability is known.

*If there is a population of 100, each person has 1 out of 100 probabilities to be selected. If there is a population of unknown dimensions, the researcher cannot know the probability each person has of being selected.*

**The sample is supposed to be, in a way, a micro-version of the population of interest**. Randomization is the systematic way of eliminating selection bias. A selection bias occurs when the sampled individuals differ from the population of interest in a systematic way. Probabilistic sampling strategies give the population parameter of interest (the population parameter is the true "answer" – the estimate that would be found if the researcher had access to the whole population, instead of a sample) a better chance of being accurately represented by the sample.

### What is *confidence level*?

**Confidence level tells how sure the researcher is**. A 95 per cent confidence level means that if the very same exercise were repeated 100 times (each time randomly extracting another sample from the same reference population), in 95 out of those 100 times, the researcher would receive the same answer.

The more confident the researcher wants to be (99%), the larger the sample size must be (see below). To be 100 per cent sure, the researcher should ask everybody in the population, and not just a sample.

### What is *confidence interval* and *margin of error*?

**Confidence interval is a range of plausible values**, based on the confidence level. For instance, the researcher is 95 per cent sure (confidence level) that the answer is likely to fall somewhere between value A (lower value) and value B (upper value). **The margin of error tells how precise the answer is. It is a way to express sampling error**. The difference between the average answer from A (average – margin of error = A) and B (average + margin of error = B) is the margin of error.

Sample mean ± margin of error = interval between A and B = confidence interval.

Keeping the same confidence level, the more accurate a researcher wants to be, the smaller the margin of error has to be and the larger the sample size (see below).

### How big should a sample be?

In sum, the sample size depends on how sure and how precise a researcher needs the answer to be. **A sample is not proportional to the total reference population**.

Here are sample sizes for a population of 100, 1,000, 10,000 and 100,000 people.

| Population size | 95% confidence level, 10% confidence interval (95/10) | 95% confidence level, 5% confidence interval (more precise) (95/5) | 99% confidence level, 10% confidence interval (more confident) (99/10) | 99% confidence level, 5% confidence interval (more confident and more precise) (99/5) |
|---|---|---|---|---|
| 100 | 50 | 80 | 63 | 88 |
| 1,000 | 88 | 278 | 143 | 400 |
| 10,000 | 96 | 370 | 164 | 625 |
| 100,000 | 96 | 383 | 167 | 662 |

### What does this mean?

If the researcher randomly extracted 100 times a sample of 383 people from a population of 100,000, in 95 times out of 100, the researcher would receive an answer that falls 5 percentage points below or above the average that the researcher would have obtained if he had interviewed the whole population. If in the sample, 48 per cent of respondents picked a choice, the researcher is 95 percent sure that 48% ± 5 (from 43% to 52%) of the total population would have picked the same choice.