Rapid Guide to Survey Sampling

Sampling is one of the most important parts of your surveys. In most cases, our surveys do not have the resources required for collecting information from all members of our target population. A correct sampling strategy enables us to survey only a part of the population while ensuring that the results are valid for the entire target population. However, sampling is prone to many mistakes which can compromise the validity of your data. IndiKit therefore prepared a brief guide to help you ensure maximum quality of your survey's sampling.

CHOOSING WHO TO SURVEY

First, define your **target population** – the people we are interested in collecting data from. The most common target populations are:

- people whom your intervention aims to help (its direct beneficiaries)
- other people who might have benefited indirectly (can be used for assessing the intervention's spillover effect)
- other stakeholders of your interventions whose opinions and other data you need to assess
- members of your comparison group (see below)
- various population groups living in a given area (commonly used for needs assessments)

Keep in mind that one survey can have several different target populations. For example, a baseline survey of a nutrition-sensitive agricultural intervention might target children under 5 years (for measuring the prevalence of undernutrition), their caregivers, trained farmers as well as agriculture extension workers. Such surveys then require separate samples for each target group.

Control and Comparison Groups

Most relief and development surveys are interested in assessing the changes they brought to the populations of their concern. A frequent assumption is that by comparing the baseline and endline data we can see changes our intervention delivered. However, how do we know that these changes were caused by our intervention and not some external factors, such as changing weather (affecting harvest), market prices fluctuations (influencing access to food) or certain trends the entire country is experiencing, such as decreasing poverty? The most common approach to assessing **which changes can be attributed to our intervention** is comparing the situation of our intervention's beneficiaries with the situation of a similar group of people who were not exposed to the intervention. The result should tell us what would have happened if our intervention was not implemented. There are two most common approaches:

- Using a Control Group: This approach randomly divides the intervention's potential beneficiaries in two groups: the treatment group (which receive the planned assistance) and the control group (which does not receive any assistance). It then compares the changes experienced by the treatment and control group members. Due to obvious ethical and programming-related issues, this approach is used primarily when the intervention's primarily **objective is to gain evidence** (e.g. of a certain product or approach's effectiveness). It is suitable for purely research-based or pilot interventions which are expected to provide evidence required for an effective scale-up of the tested approach (product, etc.). The control group members do not necessarily have to be left without any assistance – a so called **phase-in design** lets the treatment group participate first while the control group receives assistance once the initial testing is over. In doing so, all participants can receive the same assistance, only at a different time.
- Using a Comparison Group: In many contexts, using control groups is not an appropriate option. In such cases, the next best approach is to use a **comparison group** comprising of people with very similar characteristics as your intervention's beneficiaries. For example, if your project targets farmers living in 50 communities, your comparison group can be farmers living in communities which are not exposed to your intervention but have a very similar environment, agronomic practices, access to infrastructure, and other essential characteristics.

Sample Selection

Once you know who your target population is, the next question is how to select the individual survey participants. You can take advantage of the following, most common approaches:

- Simple Random Sampling gives every member of the target population an equal chance of being chosen to participate in the survey. The most common selection method is preparing a list of all the target population members and then selecting them based on randomly generated numbers (provided by, for example, Excel's RAND function). If the list is not available (typically for larger populations), proceed by using other sampling techniques, such as Cluster Sampling or Multi-Stage Cluster Sampling.
- Cluster Sampling is a good alternative if you do not have a list of the entire target population or if there is a large geographic area to be covered. It involves randomly selecting groups, not individuals. For example, if we are interested in surveying the population of an entire town, we can divide it first into geographic units (clusters), such as neighbourhoods or streets, which have roughly the same population and characteristics (this is very important). Then we randomly select a number of those clusters and survey all the members of our target population living in the selected clusters.
- Multi-Stage Cluster Sampling involves randomizing at two or more levels. For example, as in the case above, divide the area into geographic units (clusters) of roughly the same population and characteristics. Subsequently, randomly select several clusters. The number will depend on your capacity and resources. As the next step, you do not survey every member of the target population living in the cluster, but select them randomly from each cluster, typically via selecting the households and surveying a member of the household who is a member of your target population. Random selection of households is based on a rough estimate of the number of target population members living in the given cluster (sometimes you can also use Google Maps to make the estimate). Randomly select a starting point (for example, randomly selected intersection). When you get to this place, spin a pen to choose the direction you will walk in first. To decide which household you will visit first, choose randomly a number from 1 to 10 and visit the given household (e.g. if you randomly choose 5, visit the 5th household). Then continue surveying every Xth household (e.g. every 7th household) based on the total number of households and your sample size (so called sampling interval). When you get to the end of the area in the given direction, you can spin the pen again and continue in a new direction until you have surveyed the required number of respondents.
- Stratified Sampling allows the identification of sub-groups (strata) within a population and creation of a sample which reflects their actual representation in the population. The individual members from each sub-group are chosen by simple random sampling. The size of the sub-groups in the overall sample should be proportional to the entire target population. For example, if women represent 60% of an emergency project's beneficiaries, the proportion of female respondents should equal to 60%.
- Convenience Sampling (Purposive Sampling) is used when randomization in sample selection is too complicated or not possible. The interviewer selects as respondents any members of the target population which are available nearby, such as people on the streets, at a market, etc. While this approach saves time, the data it provides are not likely to be representative for the given population. Convenience sampling can lead to misleading conclusions – for example, surveying clients of a rural health clinic would tell you very little about health-care needs of the general population as you will not have collected data from people who do not frequent health clinics (for example, because they cannot afford it or are healthy).
- Snowball Sampling involves finding several members of the target population and then using them to find more respondents (asking them for referrals to others). Each "wave" of respondents is used to contact the next, so the sample slowly grows like a snowball rolling down a hill. While this technique does not provide a representative sample of the target population, it is useful for accessing hard-to-reach respondents.

Keep in mind that sampling is **prone to many selection biases**, such as the interviewers incorrectly replacing an unavailable survey participant with another respondent or conducting a needs assessment on a market day when economically more active and younger people are away, resulting in gaining data of limited representativeness. Therefore, **always give clear instructions** of when to collect the data, how to select the survey participants and what to do when the intended participants are not available. Furthermore, use IndiKit's QIV Checklists for monitoring and improving the quality of interviewers' work.

SAMPLE SIZE CALCULATION

Baseline, endline and other surveys requiring precise quantitative data need to use a representative sample of respondents. Its required size can be easily calculated by using IndiKit's on-line Sample Size Calculator. The only information which you need to enter are:

- **Margin of Error** (also called confidence interval) is the amount of error that your survey's findings can tolerate. For example, if your sample uses a 5% margin of error and you survey shows that 63% of women are exclusively breastfeeding, you can be sure that if you had surveyed all breastfeeding women in the given population, the result would between 58% (63-5) and 68% (63+5). The lower the margin of error you choose, the larger your sample size will be. While the most common margin of error is 5%, surveys focusing on data which are likely to record only small changes (such as prevalence of undernutrition) often only use 2-3%.
- **Population Size** is the total number of people you are choosing your random sample from. This can, for example, be the total number of your beneficiaries or number of women living in a given district.
- **Confidence Level**, tells you how sure you can be of your results. Since it is very rare that relief and development surveys would use anything other than a 95% confidence level, IndiKit's Sample Size Calculator enters this value automatically.

A common mistake occurs when surveys include **questions which can only be answered by a portion of the respondents**. For example, while a survey's target population could be mothers with children aged 0-23 months, some of the questions may focus only on mothers with children aged 0-6 months or 12-15 months. Since only a smaller portion of the survey's respondents have children of such an age, the number of responses will be relatively low, resulting in the survey findings having limited precision (often raising the margin of error up to 10%). In such a case, in order for the survey to provide acceptably precise data, you should significantly increase its sample size.

Sample Size for Follow-Up Surveys

Follow-up surveys are usually mid-term and/or endline surveys. While most of them allow us to use a new sample of respondents (following the same methodology as the baseline survey did), some surveys require interviewing the same respondents which participated in the baseline survey. However, these respondents may not be available of may refuse to participate in the follow-up survey. This result in so called **sample attrition**, sometimes also called a **loss to follow-up** – the percentage of respondents who participated in the baseline survey but did not participate in the mid-term or endline survey. The loss to follow-up is calculated as:

$$\text{loss to follow-up} = \left(\frac{\text{number of respondents at endline}}{\text{number of respondents at baseline}} \times 100 \right)$$

A large loss to follow-up can bias the results, because the baseline and end-line data are not directly comparable. The "**5-to-20 rule**" can be used to interpret the validity of outcomes. This rule states that:

- if less than 5 percent of the baseline respondents are lost to follow-up, the loss probably results in minimal impact on the validity of outcomes; and
- if more than 20 percent of the baseline population is lost to follow-up, the loss threatens the validity of results (in this case, caution is advised in making conclusions based on the outcomes obtained)

> Do you have a suggestion for improving this Rapid Guide's content? Send it to us please!
> Would you like this Rapid Guide to be available in a different language? Get in touch with us!

RESOURCES USED FOR PREPARING THIS GUIDE:

- PIN (2014) Data Collection: A Practical Guide to Collecting Data
- PIN (2014) Impact Evaluation: A Practice Guide to Designing and Administering Impact Evaluations
- FAO, Analysing the Data and Reporting the Results