1.1 Sampling methodology for crop production practices survey at landscape level in India

A. Samaddar[1], A. Ajay[2], A. Keil[2], Ajit Gupta[3], Ranjit Kumar Paul, Alka Arora, Sudeep Marwaha, Soumen Pal and S.N. Islam

[1]International Rice Research Institute (IRRI)
[2]International Maize and Wheat Improvement Center (CIMMYT)
[3]Indian Agricultural Statistics Research Institute (IASRI), New Delhi

E-mail: a.samaddar@irri.org

Sampling method is one of the important factors which determines the correctness of survey results. There are many ways of drawing survey samples depending upon the need and situation. If anything goes wrong in selecting samples, survey results get distorted. Cereal System Initiative for South Asia (CSISA) in collaboration with Indian Council of Agriculture Research (ICAR) in India planned to gather information about current crop production practices at large scale. This electronically enabled survey had to be implemented through *Krishi Vigyan Kendra* (KVK) of each district. In 2018, the survey was implemented in 50 districts across five eastern states (Bihar, Uttar Pradesh, Odisha, West Bengal and Chhattisgarh). By 2019, approximately 100 districts have to be covered. The objectives of this electronically enabled survey were to fill existing data gaps, generate recent data-based evidences, derive better insights and facilitate informed decisions by policy makers. Sampling methods used for the survey have been detailed in this article.

Data Collection

There are several methods of data collection that can be applied in field surveys. These methods fall into two broad categories:

Quantitative method – Data are collected in a random sample of 'observation units', typically farm households. Random sampling is required to ensure that the sample is representative of a larger underlying population, e.g. farm households in the district. From each observation unit, the same set of information is elicited using a structured questionnaire. If the observation unit is the farm household, this means that each respondent farmer is asked exactly the same questions. Appropriate statistical methods are used for data analysis. For example, we collect data on wheat yields from a random sample of farm households in district X. If our sample is sufficiently large, the average yield we find in our sample will be an adequate estimate of the average yield that farmers in district X attain overall.

Qualitative method – This approach is commonly used when the research topic is complex and requires deeper understanding. Focus group discussion (FGD) and key informant interviews (KII) are some of the very popular methods to collect qualitative data. Other than in the quantitative approach, we will prepare an interview guideline for data collection, rather than a questionnaire with fully formulated questions and potential response options. The guideline helps us ensure that we cover all relevant aspects during the interviews/discussions, but each such event will differ from the other. For instance, when we discuss a given topic with two groups of farmers separately, the two groups will almost certainly give different kinds of reactions and inputs, leading the discussion in somewhat different directions. It is the task of the researcher to react flexibly and follow up on such diverse inputs, rather than sticking to a list of pre-defined questions, as is done in a quantitative survey.

Both quantitative and qualitative methods have their own merits and limitations. None is 'better' than the other; rather, the two approaches complement each other, and which of the two is more appropriate depends on the research question to be addressed. If we want to get a 'representative' picture of what practices farmers are using, how these technologies are performing, and what farmers' perceptions are regarding the benefits of these practices and the constraints to their adoption, we need to use the quantitative approach. If we want to delve into great depth or get farmers' views on sensitive or highly complex issues, or we want to investigate particularly contrasting cases/settings, we should pursue the qualitative approach. When, we follow a qualitative approach, we often select the villages where we conduct FGDs or KIIs according to certain criteria, e.g. villages with good market access versus very remote villages;

this means that, in contrast to the quantitative approach, we often use purposive sampling rather than random sampling to select our research villages. Findings from qualitative research cannot be generalized to the population, but they can be used to highlight (contrasting) cases or conditions that require further investigation. Consequently, the sample size (e.g. number of selected villages) in qualitative research is usually very small.

Often, a 'mixed-methods' approach is recommended, combining the strengths of both quantitative and qualitative approaches. Following is an overview of the advantages and disadvantages of the two approaches.

	Quantitative	Qualitative
Advantages	• Results can be extrapolated to a larger, underlying population • Efficient and easy digital data collection using structured questionnaire • Relatively quick basic statistical data analysis	• Information can be obtained relatively quickly and inexpensively • More suitable for sensitive or complex issues than quantitative approach • Flexibility to follow up on unexpected aspects as they arise during data collection
Disadvantages	• Relatively costly and time-consuming, depending on sample size • Less suitable for sensitive or highly complex issues (e.g. power relations etc.)	• Results cannot be extrapolated to a larger population (e.g., each FGD represents a case study) • Data collection and analysis require greater skill than applying a structured questionnaire

Sampling

In the context of field surveys, sampling is a process in which a predetermined number of respondents are selected from a larger population. The methodology used for selecting respondents from a larger population depends on the type of analysis being performed. All sampling methods can broadly be categorized into two:

- Probability sampling
- Non-probability sampling

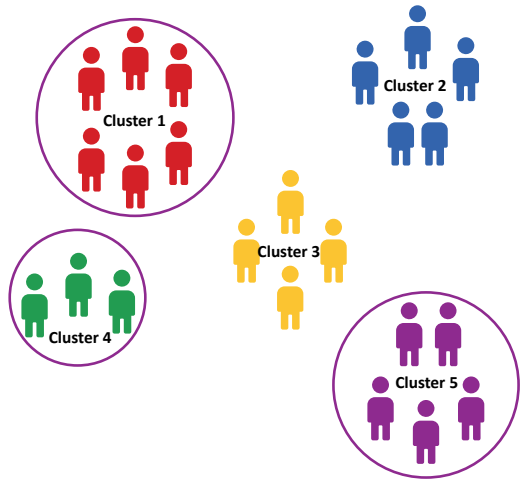
The difference lies between the above two is whether the sample selection is based on randomization or not. In case of randomization, every element gets equal chance/probability to be selected and to be part

of survey. Before start, it is important to understand basic terminologies used in sampling.

- **Element (or observation unit):** This is the unit about which information is sought.
- **Sampling unit:** This is the element or elements available for selection at a given stage in the sampling process.
- **Sampling frame:** This is the list of sampling units available for selection.
- **Population (or universe):** This is the aggregate of all the elements defined prior to selection of the sample.

The ongoing Landscape Diagnostic Survey (LDS) of cereal crops used single stage cluster sampling, a type of probability/random sampling method.

Probability sampling → Cluster sampling Single stage cluster sampling

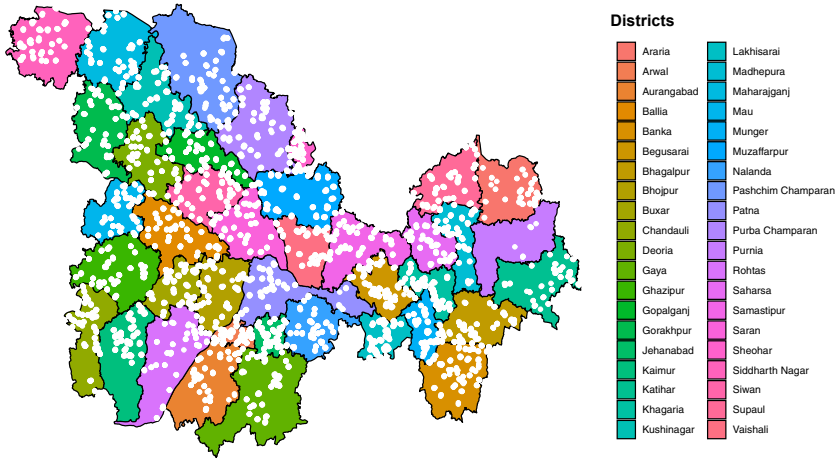


In cluster sampling, entire population is divided into clusters and then the clusters are randomly selected. To apply single stage cluster sampling, samples are drawn randomly from the selected clusters. All the elements of the cluster are used for sampling (Singh, 2018).

Accordingly, LDS selected villages within a district considering village as one cluster then selected farm households within each cluster/village. So, in our perspective, the above terminologies refer to:

Population: All villages of a district then all households of the village

Landscape diagnostic survey (LDS) in Eastern UP and Bihar



Sampling frame: Rural villages with >30 and <5,000 households then all households of the village

Sampling units: Villages first then farm households

Elements: Farm households

In broad terms, the sampling process comprises following five steps:

- Step 1: Define the population
- Step 2: Select a sampling procedure
- Step 3: Construct the sampling frame
- Step 4: Determine the sample size
- Step 5: Select the sample

Sample Size

The larger the sample size, the more precise the estimates will be, such as average yields or the percentage of farmers using a given technology. In other words, with a larger sample, we can be more confident that our results will be relatively close to what we would find in the population as a whole. Related to this, the larger the sample size, the more likely we are to detect statistically significant differences between groups (e.g., differences in wheat yields between farmers who sowed before November 15 and those who sowed thereafter). However,

the gains in precision decrease quickly at the margin with increasing sample size.

We suggest that KVKs aim at a sample size of 210 randomly selected farm households in their district to assess farmers' current practices; for most purposes, this sample size achieves a good balance between data precision on the one hand and cost of data collection on the other. We further suggest that sample households be spread across 30 randomly selected villages to capture an adequate degree of across-village variation, e.g. in terms soil conditions, infrastructure, and market access (factors which may influence the outcomes that we are interested in).

Village Selection

Probability proportionate to size method of random sampling was used to select villages. It refers to a sampling technique where the probability that a particular sampling unit will be chosen in the sample is proportional to a known variable such as number of households. It can also be called unequal probability sampling, because one is actually increasing the odds that a subject will be chosen in the sample based on its size. It is used when the populations of sampling units vary in size. If the sampling units are selected with equal probability, the likelihood of a sampling unit with a large population being selected for the survey is actually lesser than the likelihood of elements from a sampling unit with a small population. This reduces standard error and bias by increasing the likelihood that a sampling unit from a larger population will be chosen over a sampling unit from a smaller population. To illustrate this method, consider the example of four villages of varying sizes given in the table below:

Village name	Number of households (HHs)	Cumulative number of HHs	HH ID range	Probability of selection
A	200	200	1 -200	200/1000 = 20%
B	300	500	201 – 500	300/1000 = 30%
C	100	600	501 – 600	100/1000 = 10%
D	400	1000	601 – 1000	400/1000 = 40%

To select villages using probability Proportionate to size method, we generate random numbers within the range 1 – max. HH ID. In the

example above, we would type the formula = randbetween (1, 1000) into Excel; a village is selected if the random number falls within its HH ID range, thus making the probability of its selection proportionate to its size. For example, the random number 461 would fall into village B (HH ID range 201 – 500); hence, village B would be selected. We would continue generating random numbers (pressing the F9 key) until the desired number of villages is selected. If a random number falls within an already selected village, we simply continue pressing F9 until we get a random number that falls within a new village. This method for selection of villages was done based on the 2011 census data which contain the number of resident households in each village of a given district.

Household Selection

Once the 30 villages are selected using probability proportional to size method, 7 households in each village need to be selected through simple random sampling. In the simple random sample there is only one type of sampling unit, for instance all households residing in one village. Simple random sampling is a sampling technique where every item in the population has an equal chance of being selected in the sample.

This means that we need a complete list of households in that one village. This is our sampling frame for household selection. LDS used voter list of the respective village to construct sampling frame. These voter lists of villages were downloaded from election commission websites of the respective states. These lists are generally available in PDF version. Unique house numbers were treated as single household. Using 'R' software, these PDF type voter lists were processed in batch to generate random house numbers. The output was available as single excel file with 30 worksheets (one sheet per village) having desired random numbers for survey. This is an efficient way of doing household level randomization.

The process can also be done alternatively using MS Excel. But, one need to enlist all unique house numbers of the village. Once, it is compiled, number the households consecutively from 1 to max, where max stands for the total number of households in the village. For example, if there are 150 households in the village, the numbers would run from 1 through 150. Open an MS Excel spreadsheet and select cell A1. Use the function 'randbetween' to create a random number that lies between a

specified minimum and maximum. The minimum is usually '1', i.e. the first element in our sampling frame. The maximum depends on the number of elements in our list. In our example it is 150; we therefore type:

= randbetween (1,150) and press Enter

Assume you want to select 7 households randomly: select cell A1, click on the lower right corner of cell A1 and drag it down until you reach cell A7. You now have a list of 7 random numbers available, which all lie between 1 and 150. Now simply copy the random numbers and paste them in column B as values. Now, tick off all the households that have been selected according to the list of random numbers.

Plot Selection For Crop-Cut

Follow these 10 steps:

- Refer to the selected 7 households in this village
- Select the farmer whom you meet first out of these selected 7
- Ask him for his largest wheat/rice plot – consider this largest plot for crop-cut
- Take farmer's consent for crop-cut
- Crop-cut has to be taken from 2 spots in the selected largest plot
- Size of each of these 2 spots (quadrats) are 2 m × 2 m
- Get on the corner of the plot, move diagonally for almost 5 m and select it as your first spot for taking samples
- Similarly, repeat the procedure from the another corner of this plot and mark second spot
- Finish crop-cut from these two spots and record – total above ground biomass grain weight, and moisture percent
- Use Open Data Kit (ODK) Form – 'Crop Cut Form' to enter these readings along with other basic information asked in this form.

Reference

Singh, S. (2018, July 26). Retrieved from Towards Data Science: https://towardsdatascience.com/sampling-techniques-a4e34111d808