Sample design strategies

Regional Training Course on Agricultural Cost of Production Statistics

21– 25 November 2016, Daejeon , Republic of Korea

1 – Definitions and general remarks

• **Surveys, censuses and administrative records** are the 3 main data sources for agricultural data:

=> Sampling design is one of interrelated elements of survey design and must be developed in relation to these elements

•**A sample is a subset of the units defining the universe of the population** of interest

• **Well-designed samples have the capability of making inferences to the population** with known probabilities of selection and measures of sampling variability

• **Well-designed samples for national estimates will require a surprisingly small number of farms** (cf. GS Handbook on Master Sampling Frame)

• To obtain these benefits, some basic principles on sampling design have to be respected

2 – What is a good sample ?

- A good sample should:
  - o **Adequately cover the population of interest**
  - o **Be small enough** to limit survey costs and complexity but
  - o **Large enough** to provide results with an acceptable error level
- **The quality of the sample is dependent on:**
  - o **The quality and exhaustivity of the frame** from which the sample is drawn
  - o if the frame is incomplete and/or biased towards certain activities, the sample will also suffer from the same shortcomings.

3 - A few words on sample frames (1/2)

- **A sample frame is the set of all the units defining the universe of the population of interest**. Examples:
  - o All farms with an area larger than 3 ha
  - o All commercial farms
  - o All small-scale cacao plantations, etc.
- **Units can be areas or segments**: area frame
- **Units can be codes** defining an individual, household, farm, etc.: list frames
- **Area frames are generally more exhaustive than list frames**
- **List frames tend to be more precise than area frames** (the actual unit is directly identified, as opposed to groups or fractions of units for area frames)

3 – A few words on sample frames (2/2)

- Samples can be drawn from list or area frames:
  - o **individually** (for the later, usually in several stages), or
  - o **in combination**, the most frequent case in agricultural surveys
- A classical sampling procedure for an agricultural survey :
  - o The national territory is first partitioned in large zones (ex: regions);
  - o Within each of these regions, a sample of sub-zones is randomly selected (ex: municipalities);
  - o Within each of the sampled sub-zones, the list of all household involved in farming activities is established;
  - o Finally, a sample of these households is randomly selected for the purpose of the survey.

4 – Elements of sampling design (1/2)

Sample design has two aspects:

• **A selection component**: rules and operations of including members of the population into the sample

=> The accuracy of the sample design rests on two elements :

- o The sampling frame used (or developed) should be as complete, correct and current as possible, and
- o Appropriate sample selection techniques

4 – Elements of sampling design (2/2)

• **An estimation component**: computing sample statistics, which are sample estimates of population values

=> Four points are essential in the estimation process :

- oThe specification of the (function of) parameter(s) to be estimated ;
- o An estimator of the (function of) parameter(s) ;
- o The variance of the estimator of the (function of) parameter(s);
- o An estimate of the variance of the estimator of the (function of) parameter(s)

5 – Sample designs: Random stratified sampling (1/2)

In its simplest version, this method consists in:

• **Partitioning the population of interest**, on the basis of one or more auxiliary variables: farm size, commodity produced, region, etc

• **Within each group (stratum), a sample of units is randomly selected**: the sample of the survey is therefore the sum of the strata-samples

• **The advantages** of the stratification is that:

- o for a given sample error, the sample size and budget can be reduced compared to simple random sampling
- o Conversely, with a given budget and sample size, stratification permits a reduction in sample variance.

5 – Sample designs: Random stratified sampling (2/2)

• Generally, **the size of the strata-samples is chosen to be proportional to the size of the strata** (PPS – Probability Proportional to Size):

o If the a stratum comprises 20% of the universe, the corresponding strata-sample will represent 20% of the total sample

o This is what is generally referred to as "representative" sampling

o The sampling rate for each strata is equal to the sampling rate in the population:

$$n_h/N_h = n/N \Leftrightarrow n_h/n=N_h/N$$

• **Systematic sampling is often used to select the units of the sample**:

o The sampling step is *k=N/n*, where n is the sample size and N the population size

o A random number *r* comprised between 1 and *k* is chosen

o The sequence {*r*,*r+k*,*r+2k*,….,*r+nk*} is the selected sample

6 – Sample designs: Multistage stratified sampling (1/2)

• **Multi-stage sampling is when samples are drawn iteratively** (samples of samples)

• This sampling procedure is also often referred to as **cluster sampling**

• **An example of three-step stratified sampling** procedure combining area and list frames frequently used in agricultural surveys:

o Step 1: within each administrative region, a number of enumeration zones (primary sampling units) is randomly selected;

o Step 2: within each selected enumeration zone, a number of villages (secondary sampling units) is randomly selected (the full list of villages is obtained from the previous population census);

o Step 3: within each selected village, all the households involved in agricultural activities (final sampling units, list obtained from the population census) are selected for the survey.

6 – Sample designs: Multistage stratified sampling (2/2)

- Advantages :
  - o **Less costly**: it is not necessary to have the auxiliary information for all the units of the population
  - o **Ensures statistical representativeness** at different levels: provinces and country, for example
  - o **More precise** than simple stratified sampling
  - o **More secure**: all provinces/regions are certain to be represented in the sample as long as they are used as stratification variables

- Main disadvantage: **the sample size is unknown** because each sampling unit is randomly selected and has a variable size. Implications:
  - o **The budget cannot be determined beforehand**; and
  - o **The computation of sampling errors for the mean of the variable of interest (*m*=*y/n*) is complex** because it is a ratio of two random variables.

7 – Determination of the sample size (1/2)

- **The sample size *n* is directly related to the desired precision** of the results:
  - o Assuming the objective is to estimate *m=y/n*, the Central Limit Theorem says that there is a 95% chance that *m* lies between *m-2*sig/n* and *m+2*sig/n,* where *sig* the estimated standard-deviation of the variable of interest in the population.
  - o If sig is known and the width of the confidence interval is set (to *r* units) n is determined by: *4*sig/n=r*
  - o sig (or var) is known for simple random designs and simple estimators
  - o Example: for simple random sampling, the variance of mean of the objective variable is: *var(m)=(1-f)sig(y)/n*

7 – Determination of the sample size (2/2)

- **Sample size is a pivotal feature** in overall sample design:
  - o It depends on survey objectives, resources, desired precision, anticipated non-response…
  - o It determines the number of data collectors to hire and their work load, etc.

- **The sample size has direct implications on the budget** of the survey:
  - o **Survey costs have a fixed component**: cost of installations, cost of statisticians at headquarters, etc.
  - o **and variable component**, measuring the cost of surveying one additional unit: travel costs of interviewers, paper, etc.

- The sample size is in practice the result of the **compromise between budget and precision/accuracy**.

8 – Putting the pieces together (1/2)

To obtain the desired results, the sampling design must:

- **Be stratified** to have a good representativeness and control of the sample size over population subgroups;

- **Be based on a complete, accurate and up-to-date sample frame,** to obtain maximum accuracy ;

- **Use of selection and estimation techniques that minimize bias and maximize accuracy**;

- **Be measurable (known probabilities of selection)** so that sampling errors can be estimated to provide users with a reliable error measure of the results

8 – Putting the pieces together (2/2)

The sampling must:

- **Be done in stages** to efficiently choose elements of the population of interest

- **Be multistage** if necessary to reduce cost

- **Make a wise use of clusters** of elements (balance between cost reduction and precision)

- **Have a sample with an appropriate size** so that costs and precision are optimally controlled

9 – References

- **Handbook on agricultural cost of production statistics** (Draft), Global Strategy Publications, 2012

- **Sampling Methods for Agricultural Surveys** (1989), FAO Statistical Development Series 3, FAO, Rome. Accessible online at: http://www.fao.org/fileadmin/templates/ess/ess_test_folder/Publications/SDS/3_sampling_method_for_agricultural_survey.pdf

10 – Annex: an optimal sample size

An "optimal" sample size may be determined using the following approach:

• **Determine a desired sample size** associated with a simple random sampling design:

- o $SamplingError(y)_{SRS} = (1-n/N).Variance(y)$, which implies:
- o $n = n'/ [1+n'/N]$ where $n' = Variance(y) / SamplingError(y)_{SRS}$

• **Adjust this sample size** using the relative precision of the design at hand compared to the simple random design (Deff coefficient):

- o $Deff = SamplingError(y) / SamplingError(y)_{SRS}$
- o $n'_{final} = Deff. n'$
- o $n_{final} = n'_{final}/ [1+n'_{final}/N]$