The treatment of uncertainty in the results

Regional Training Course on Agricultural Cost of Production Statistics

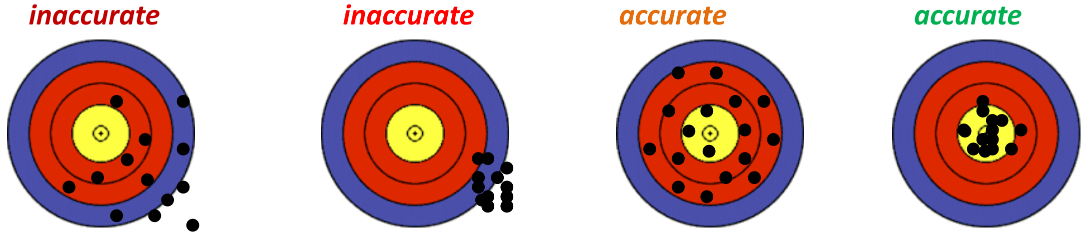
21– 25 November 2016, Daejeon , Republic of Korea

1 – Uncertainty: definition and characteristics (1/2)

- **Uncertainty is generated by the randomness** of the data collection and estimation process.

- **Uncertainty impacts the precision** of the results but not the accuracy: results may be uncertain (highly dispersed around their mean) but accurate (the estimated mean is close to the “true” mean) and vice-versa.

Uncertain and inaccurate | *Certain and inaccurate* | *Uncertain and accurate* | *Certain and accurate*



1 – Uncertainty: definition and characteristics (2/2)

• Errors generating uncertainty (loss in precision) include:

  o **Sampling errors**: the loss in precision due to the selection of a random sample of the population

  o **Errors due to estimation procedures** such as regression or any technique which involves statistical inference.

• **Non-random (or non-sampling) errors** are systematic errors made at different stages of the data collection and compilation process:

  o Asking the wrong questions, inappropriate sample coverage, systematic editing mistakes, etc.

  o These errors can be minimized through an appropriate planning, management and follow-up. They are difficult/costly to correct ex-post.

  o These errors affect accuracy, not uncertainty/precision

2 – Why is it important to take uncertainty into account ?

• **Important policy decisions are based on surveys**: uncertainty may lead to adverse policy impacts based on badly designed policies... or no decision at all !

• **Surveys always lead to uncertain results**, even when designed in a statistically sound way: users have to elaborate strategies that take uncertainty into account

• **Varying degrees of uncertainty may be accepted**, depending on the nature of the survey: lower uncertainty may be required for surveys touching vital topics such as public health

=> It is essential to provide users with information on uncertainty, be it of a quantitative or qualitative nature

3 – Sampling errors: definition and characteristics

• **They originate from the extrapolation of the results** obtained for a randomly selected sample to the population of interest as a whole

• **They provide a quantitative information on the quality of the sample** selection process:

=> **high sampling errors lead to uncertain results**: they will be difficult to interpret and of little use

• The magnitude of the sampling error depends on:

- **The type of sample selection method**: simple random, stratified, multistage, etc.
- **The size of the sample**

4 – Sampling errors: measurement

• **Uncertainty can be appropriately measured by the variance** or standard-deviation of the variable of interest:

=> Sampling errors are measured by the variance of the variable of interest due to the sampling procedure (sampling variance)

• **Sampling variance formulas are known for simple sample designs**, i.e. when the inclusion probability of the final sampling unit is known and when the variable of interest to estimate has a "simple" expression (total or mean, for example).

=> See corresponding Annex of the manual

• When sampling variance can be measured, **point estimates of the variable of interest can be completed with confidence bands**

• **For complex sample designs** (multistage, etc.) a**nd/or complex functions to estimate** (ratio of two variables for example) **sampling variance can only be approximated using complex methods**

5 – Other random errors

• The process of collecting and compiling data may include stages where estimations are made using statistical inferences, such as:

o The **correction of non-response** using regression techniques, hot or cold-deck imputation, etc.

o An example was given in this seminar with the use of **model-based techniques for cost allocation**

• **These statistical operations generate uncertainty** in the results: in most cases, it can be measured by simulation or other techniques.

• **These estimations generally affect part of the sample**, rendering its impact on the whole sample difficult to measure.

6 – Global or combined uncertainty

• As seen, **different sources of uncertainty exist**, some sampling-related and others not

• When the estimated objective variable is affected by several sources of uncertainty , the combined uncertainty can be measured:

o $\sigma(y) = \sqrt{\sigma_1^2(y) + \sigma_2^2(y)}$ if the two sources of uncertainty (1 and 2) **are uncorrelated**

• **In more complicated cases, uncertainty cannot be summed.** For example when:

o The objective variable is a function of other variables, etc.

o The sources of uncertainty are correlated

7 – How to present information on uncertainty (1/2)

- When quantifiable, **uncertainty measurements should be included in the results along with the point estimates**. Most commonly, these include:
  - o **Sampling standard-deviation** and associated confidence bands, using the usual risk probabilities of 5% or 10%
  - o **Standard-deviation due to estimation methods** and associated confidence bands
  - o **Combined standard-deviation**, when there are several uncertainty sources
  - o Any other information on the probability distribution of the estimate

8 – How to present information on uncertainty (2/2)

Even when uncertainty cannot be measured:

- **it can be taken into account in the presentation of the results**. Common options are:
  - o To **present the results for sub-groups of the population of interest**, such as deciles, quintiles, etc.
  - o To provide **inter-decile and inter-quartile ranges**
  - oThe variables used for the grouping should be:
    - ▪ **The variable of interest itself**: for example, if average farm income is the objective variable, average income by income deciles/quintiles can be provided
    - ▪ **Variables correlated to the variable of interest**: farm income by deciles/quintiles of farm size, etc.

- **The statistician should provide qualitative information** on its possible magnitude

9 – Country example: Zambia

•

Share of total maize production (%)	1 31.4%	2 27.1%	3 20.1%	4 12.8%	5 8.7%	farmer mean	per 50 kg bag mean
	Total Cost Quintile (ZMK/50 maize kg)						
Costs of production (ZMK/50kg)	---------------------------------Mean-------------------------------						
Total cash expenditures	6,411	12,239	14,969	19,449	27,482	16,111	18,630
Expenditures plus household labor and assets (excl. land)	15,567	29,078	42,776	64,341	118,953	54,152	40,739
Land annual rental	3,364	4,835	6,633	9,152	15,102	7,818	4,720
Total Cost (incl. land cost)	18,931	33,914	49,409	73,493	134,055	61,970	45,459

Source: MACO/CSO Crop Forecast Survey 2010.
Note: a) Fertilizer and seed costs include both subsidized and commercially acquired inputs.

• **Countries** (developed and developing) **do not generally disseminate error/uncertainty estimates** even though such estimations are done and used internally

10 – References

• ***Handbook on agricultural cost of production statistics*** (Draft), pp. 53 – 55 and Annex 1, Global Strategy Publications, 2012

• Bell (1999), ***A Beginners Guide to Uncertainty Measurement***, Measurement Good Practice Guide n°11 (Issue 2)

• ***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