Food and Agriculture Organization
of the United Nations

GLOBAL STRATEGY TO IMPROVE AGRICULTURAL AND RURAL STATISTICS

TRAINING IN AGRICULTURAL STATISTICS

Module 0: Statistical review

2017

Statistical review

0.1 Definitions (1/14)

- Population
- Target population
- Statistical unit
- Sampled population
- Sampling unit
- Data
- Statistics
- Collection period
- Reference period

0.1 Definitions (2/14)

- **The population** consists of all the individuals having identification or definition criteria in common.

- Example:
  - Population of farms
  - Population of farmers
  - Male farm labour
  - Agricultural assets, equipment and machinery
  - All agricultural assets, equipment and machinery in good condition for the current agricultural year

✓ *This population could coincide with the target population* or the sampled population* in certain situations*

* These concepts are defined in the following slides

0.1 Definitions (3/14)

- **The target population** or scope of a survey is the population to which the desired information relates. It should be carefully defined for each study.

- **The statistical unit** (or individual) is any component, all of which together form the population or universe. In other words, it is any component of the population (target). It can be of any nature (village, household, hamlet, cultivated plot, business, etc.), for example:

Population	Statistical unit
Farm labour	Any farm employee or any farm operation
Farm assets, equipment and machinery	A tractor

0.1 Definition of concepts (4/14)

- **The sampled population**: population actually included in the scope of the study. It does not always coincide with the target population. This situation could occur in the following cases:
  - Methodological problem: difficulty directly identifying the desired statistical units and selecting them.
  - Problem of coverage: exclusion of some units. Difficulty accessing reporting units.

If the two populations are different, the sampled population should be reasonably consistent in terms of coverage and matching the target population so that the survey results are relevant.

0.1 Definition of concepts (5/14): Sampled population

- Methodological problem: difficulty directly identifying the desired statistical units and selecting them.

  - In practice, it is sometimes difficult to identify and select statistical units directly and have access to them.
  - Example: Inventory of available agricultural equipment and its condition (working) in a given agricultural year.
  - In this case, the statistical unit is the agricultural equipment. However, it is virtually impossible to gain direct access to agricultural equipment to list it and assess its working order.

0.1 Definition of concepts (6/14): Sampled population

- Methodological problem: difficulty directly identifying the desired statistical units and selecting them.
  - On the other hand, access seems easy to agricultural households (agricultural holders/holdings). So the agricultural household (agricultural holder/holding) will be the sampling unit.
  - The equipment thus available in households (agricultural holders/holdings) will be identified and the required information collected.

0.1 Definition of concepts (7/14): Sampled population

- Problem of coverage: exclusion of some units. Difficulty accessing reporting units.

- A few examples are:
  - Exclusion of isolated areas (owing to relatively high travel costs)
  - Exclusion of agricultural activities carried out by some institutions (prisons, etc.)
  - Non-exhaustive sampling frame list, due to missing information in some units

0.1 Definition of concepts (8/14)

- **Sampling unit**
  - consists of each “member” of the sampling frame. This frame is a comprehensive list of the statistical units of a given population
  - is the unit directly subject to a selection operation.
  - can be:
    - an agricultural holding
    - a farm plot
    - a household
    - a child
    - a housing
    - a school
    - a health training

- **Analytical unit:** level to which the analysis relates. The agricultural holding can be, for example, the sampling unit (agricultural holdings are then selected), but the analysis can relate to plots which are then the analytical units.

0.1 Definition of concepts (9/14)

- **The reporting unit** is an intermediary which supplies information on each statistical unit (e.g. a farm manager interviewed about the holding's plots, a mother asked about her children, or a head teacher asked about the school).

- **The observation unit or unit of interest**
  - is the unit on which information is requested. For example:
    - agricultural holdings, plots for which the farm manager has provided information;
    - children for whom the mother has provided information;
    - a state primary school for which the head teacher has provided information.
  - It is therefore the object that has been measured
  - It is the basic unit observed
  - For human populations, it is an individual

0.1 Definition of concepts (10/14)

- **An item of data** is the basic component of a wider information system. When statisticians produce data, they try to measure or count phenomena (individuals or activities) which are part of the real world.

  Examples of data:

  - Number of cows on a farm
  - Surface area of a field
  - Number of people in a household
  - Number of children in a family

Data is not very useful by itself. It must be organized into statistics to make it understandable and usable.

0.1 Definition of concepts (11/14)

- **Statistics** are compilations of numerical facts and figures. These facts and figures are created from data and are organized so they can be used. They appear in tables, diagrams, graphs or maps.

  They can come from:

  - Surveys (censuses or by sampling);
  - Opinion polls;
  - Administrative data (for example, imports and exports).

A distinction should be made between official statistics (produced by government bodies recognized by the national statistical system) and unofficial statistics (privately produced).

0.1 Definition of concepts (12/14)

- **Statistics** is a mathematical science which focuses on the collection, analysis, interpretation or explanation and presentation of data.

- **Information** is data processed and communicated (i.e. made available to the public in some form):
  - So data that is not disseminated is not information
  - It cannot be used generally, but instead relates to a specific issue
  - In this respect information can be used to support decisions in a variety of situations.

0.1 Definition of concepts (13/14)

- **The collection period** is the period during which data is collected in the field. It should guarantee good control of sample identification processes (neutralization of seasonal effects in particular).

Example:

Agricultural surveys are generally conducted during the period covering the crop cycle. When organizing the schedule for interviewers, the crop calendar should be taken into consideration, along with the crop growth cycle and differences between cereals, tubers and roots, vegetables, cash crops, fruit production and other crops.

0.1 Definition of concepts (14/14)

- **Reference period:** This is the period to which the data relates. It depends on the survey objectives. Depending on the case, it is an interval of time (week, month, year, agricultural season, etc.) or a specific date. It should be noted that variables can have different reference periods in the same survey.

Example:

The reference period of a crop production survey is the agricultural season. The reference period for births, purchases and natural deaths of livestock depends on the species. It is generally one year for cattle, six months for small ruminants and pigs, and one month for poultry.

0.2 Steps of a statistical survey (1/2)

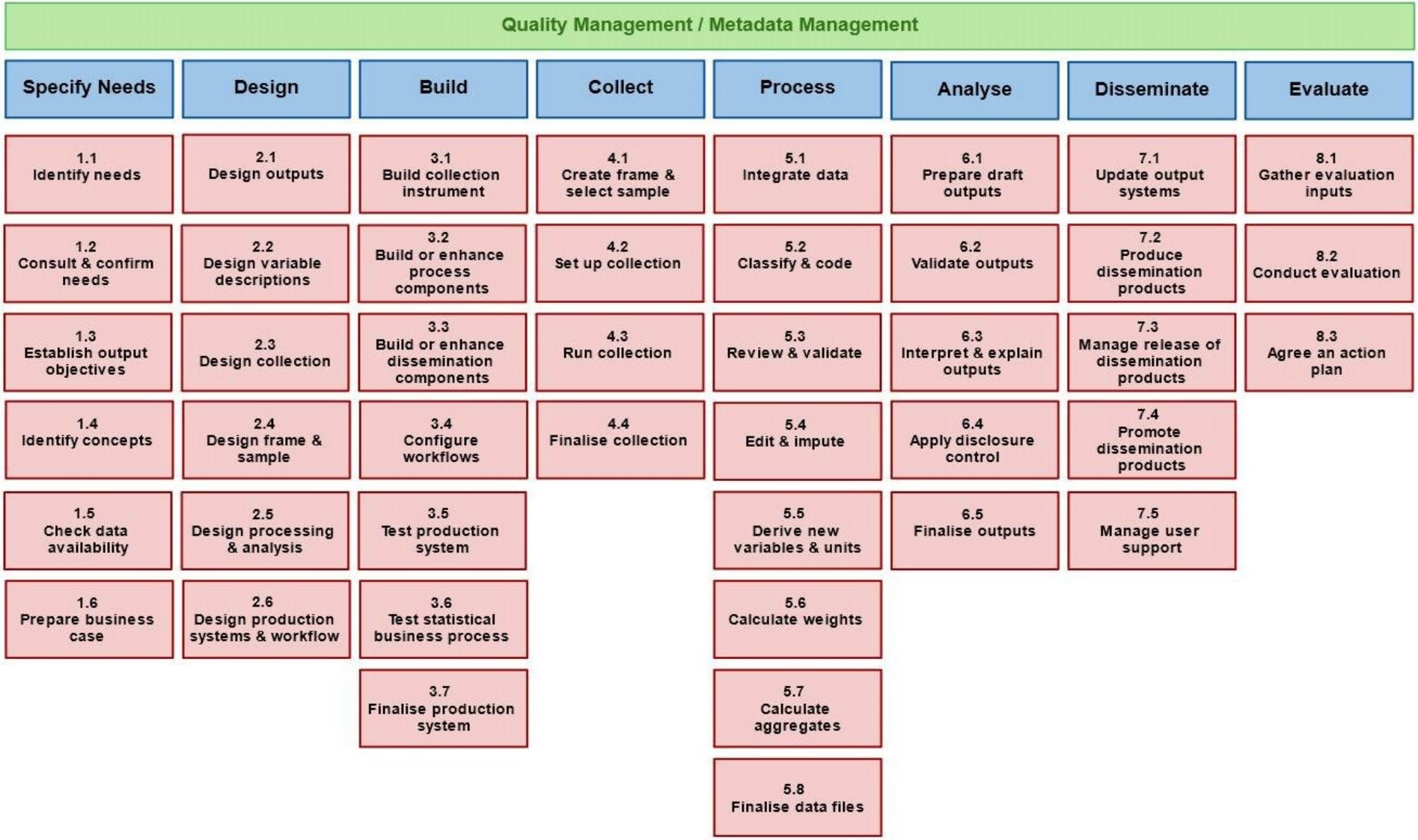
The main steps in conducting a survey are:

- Identifying information needs / Defining the objectives of the survey and resources
- Determining the collection period and reference period
- Work plan, budget
- Choosing the sampling frame and units
- Defining the sample design
- Data collection method

0.2 Steps of a statistical survey (2/2)

- Designing technical documents (questionnaires, instruction manuals, etc.)
- Recruiting and training staff
- Testing and/or pilot surveys
- Organizing and monitoring field activities
- Data processing (input, tabulation, processing and analysis)
- Preparing the report and disseminating the results

Figure 1: Different steps in a survey
Source: GSBPM v5



0.3. Sampling method (1/10)

- The choice of an optimal sample design is crucial. It helps to determine:
  - sample size
  - selection procedures
  - derived estimators, and their theoretical accuracy
- There are three main types of sample design:
  - List sample design
  - Area sample design
  - Multiple-frame sample design
- *When choosing a sample design, the desired accuracy of the data and the available resources should always be taken into account.*

0.3. Sampling method (2/10)

- The sampling method is selected from a sampling frame. However, if the frame is not reliable, a design with at least two stages is generally adopted:

  - At the first stage, a sample of the primary sampling units (PSU) (village, data production section, counting zone, etc.) is set, preferably according to strata related to the variables to be measured (e.g. agro-ecological area, high-, medium- or low-agricultural density area, etc.) and containing the survey units. A listing of all the survey units is then undertaken within each PSU;

  - At the second stage, a sample of the secondary sampling units (SSU) is set from the survey units previously listed within the PSU.

0.3. Sampling method (3/10)

- The principle of the sample design adopted is to end up with a sample where each survey unit has a non-zero chance (probability) of being included in the sample (probabilistic sample).

- However, to avoid biased estimates by favouring some survey units over others as regards the probabilities of belonging to the sample,
  - it is sometimes preferable to also consider **equality of probabilities**: all **operations of a similar** type (same stratum) **have the same** (or almost the same) **chance of belonging to the sample** will be the principle to be adopted in the two-stage sampling design.

0.3. Sampling method (4/10)

Once a sample of holdings has been selected, should it be kept indefinitely or renewed periodically for agricultural surveys?

- Ideally the same holdings should be surveyed regularly. The accuracy of estimating the variability of variables is far greater in this case.

- But in practice, this has the following drawbacks, among others:
  - the response burden of the units surveyed which eventually results in a refusal to cooperate
  - data produced for a season from data from previous seasons, by the enumerator
  - changing the structure of the units surveyed (e.g. households) over time

0.3. Sampling method (5/10)

The 2 stage sampling technique

- PSU are most often selected with unequal probabilities. This probability for each PSU is:

$$pi = \frac{Pi}{P} = \text{taille relative de l'unité i}$$

Avec:

$P = \text{taille totale des UP}$

$Pi = \text{taille de l'UP i } (i = 1,2,....,N)$

$N = \text{nombre total d'unités primaires (UP)}$

- The probability of each PSU is therefore related to a ”size” variable which can be the number of inhabitants, the number of households, the number of holdings, etc.

0.3. Sampling method (6/10)

The 2 stage sampling technique

- PSU sampling is carried out with replacement or without replacement;
- Sampling with replacement: this allows a sampling unit to be selected more than once. The probability of selection of a unit in any selection is constant;
- Sampling without replacement: Contrast to the previous selection. The probability of selection of a unit in each selection varies and depends on the previous selection. This procedure is complex and difficult to apply in surveys;
- Sampling with replacement can be similar to sampling without replacement. These two types of selection differ very little when the sampling fraction is small.

0.3. Sampling method (7/10)

The 2 stage sampling technique

- Holdings are selected from a complete listing of SSU undertaken within the sampled PSU; this listing is decisive in the extrapolation of data and the correct sampling procedure of holdings to be surveyed;

- SSU can be selected by stratification (large, medium and small holdings, for example) to avoid bias. If there is insufficient information available to carry out such stratification, empirical knowledge can be applied.

0.3. Sampling method (8/10)

Sample size

- The size of the sample in a two-stage survey consists of the size of the sample of PSU (e.g. village, counting area, etc.) and the number of SSU (e.g. holdings) to be selected per sampled PSU;

- The number of SSU sampled per PSU depends on:
  - The degree of dispersion of SSU in the PSU with regard to the variable of interest
  - The contribution of the second stage to the precision of the estimates

0.3. Sampling method (9/10)

Sample size

- The number of PSU sampled is determined after calculating the total size n of units to be surveyed with:

$$n = \frac{CV(y)^2}{CV(\bar{y})^2} = \frac{CV(y)^2}{d^2}$$

$CV(y)$ = *Coefficient of variation of y*

$\bar{y}$ = *Estimator of the mean*

d = *Relative precision desired for the mean*

- This size *n* should be adjusted according to the design effect (*Deff*) and the non-response rate (*r*) such that:

$$n' = n \times Deff \times (1 + r)$$

0.3. Sampling method (10/10)

Sample size

- If sampling takes stratification into account, the size of the strata is determined according to the aim set;

  - Local precision aim:

  The necessary size of each stratum is determined separately according to the aim set. The total size will be the sum of the various sizes.

  - Overall precision aim:

  The size n of the sample is determined first. This size is then allocated between the strata according to the following methods:

    - Equal allocation
    - Proportional allocation
    - Neyman allocation

0.4. Data collection (1/2)

- Data collection is the operational phase, also called the field phase;
- It is important to carry out awareness raising before this phase. Awareness raising should target:
  - The local and customary administrative authorities, opinion leaders and organized social groups to seek their participation in conveying the information related to the study objectives to their populations.
  - The target population to seek their cooperation and willingness to provide the information sought and to inform them about the schedule of the field work.

0.4. Data collection (2/2)

Key points for successful field work

- The chronology of the various phases which should necessarily be carried out in succession
- The identification of sampling units
- The identification of survey units
- The training of field work staff (interviewers, controllers)
- The assurance of statistical quality

0.5. Data processing (1/7)

- The main steps:

Recording questionnaires

Organization of questionnaires for coding, entry and editing

Data entry and editing

Data entry after or during data collection and editing by double entry and/or consistency checking (data type, field length, etc.)

Verification process (data editing)

Verification of the validity of response codes, consistency of a response with previous items, etc.

Imputation of missing data and processing of partial and total non-responses

Look for reasons for these non-responses and use the appropriate methods to correct them

Extrapolation

Calculating weighting coefficients to obtain representative results in data analysis

0.5. Data processing (2/7)

The different types of missing data

- Missing Completely at Random: MCAR: no link between the missing information and the whole data set. It is entirely random and cannot be explained.

- Missing at Random: MAR: link between the missing information and any variable in the data set. This is also ”missing conditionally at random”;

  e.g. Women are less likely to give their age and weight than men.

0.5. Data processing (3/7)

The different types of missing data

- Missing Not at Random: MNAR: link between the absence of the information and its value. This is also ”non-ignorable non-response”.

e.g. People with a high income are less likely to declare it.

0.5. Data processing (4/7)

Methods of processing missing data

- Analysis of all complete cases (this applies to MCAR, MAR): It involves excluding all units with missing data or results.
- Analysis of available cases (this applies to MCAR, MAR): It involves excluding a variable or a series of variables because of their high rate of missing data.
- Weighting according to the rate of missing data (this applies to MCAR, MAR): It involves finding a means of reweighting the sample to re-establish its representativeness.

0.5. Data processing (5/7)

Methods of processing missing data

- Imputation of missing values: This methods involves replacing missing data with plausible values. Here are a few methods of imputation:

  - **Mean imputation**. In this case, you should make sure that the distribution does not contain extreme values (likely to affect the means)

  - **Last observation carried forward** (historical imputation)

  - **Regression imputation**

  - **The use of information inferred from connected observations** (deterministic or deductive imputation). This applies to MNAR;

  - **Nearest neighbour imputation**. This is imputation from a donor who has similar characteristics. This applies to MAR and MNAR;

0.5. Data processing (6/7)

Methods of processing missing data

- Hot-deck imputation: This involves replacing the missing value by the value of the previous unit in the file, or the value of the last unit found, preferably in the same geographical entity. This replacement unit must be sufficiently similar to the unit for which information is missing. This method can be readily automated.

- Cold-deck imputation: This involves using information external to the survey relating to the unit for which information is missing. For example, this approach is particularly applicable in panel surveys.

0.5. Data processing (7/7)

Extrapolation: Example of a two-stage sample

- n: Number of units in the sample
- p_i: probability of inclusion of unit i
- W_i: survey weight of unit i
- W: total survey weight of the sample design
- We have:

$$W_i = \frac{1}{p_i} \quad \text{and} \quad W = \prod_{i=1}^{n} W_i = \prod_{i=1}^{n} \frac{1}{p_i}$$

0.6. Data analysis (1/5)

- This can be defined as the process of converting raw data into relevant information by analytical and logical reasoning;
- This phase begins with the definition of indicators and their calculation formulae. But differences are observed at this level in the agricultural statistics produced;
- These differences cause problems with the comparability of data between countries and sometimes within regions of the same country.

0.6. Data analysis (2/5)

Reasons for these differences in the agricultural statistics produced

- Difference in design and definition according to the country, its traditions, its culture, its statistical practices;
- Difference in data sources for calculating the same indicator;
- Differences in demographic estimates and denominators: use of different projection methods;
- Inadequate institutional coordination between national, regional and international stakeholders on the one hand and between national stakeholders on the other;
- Incomplete metadata (information on data content).

0.6. Data analysis (3/5)

Tabulation and statistical software

- Tabulation is a process that involves determining numbers of individuals or cases corresponding to specified combinations of characteristics from records in a dataset.
- The specifications of tabulations must be understandable both for specialists and staff responsible for data processing, and be sufficiently detailed so that the staff responsible for data processing do not take decisions concerning tabulation contents.

0.6. Data analysis (4/5)

Tabulation and statistical software

- There are many statistical software packages available for data analysis, some of which are listed below:
  - Open-source software packages: R (a free application of language S) and DAP (the free version of the program SAS);
  - Public domain programs: CSPro (Census and Survey Processing System, mainly used to capture, tabulate, map and disseminate survey and census data), Survey Solutions (SuSo) and Epi Info (specialized in epidemiology);
  - Free software: GeoDa (free software for spatial data analysis, geovisualization, spatial autocorrelation and spatial modelling), QGis (an open-source geographical information system) and WinBUGS [Bayesian analysis software using Markov chain Monte-Carlo (MCMC) methods];

0.6. Data analysis (5/5) Tabulation and statistical software

- Commercial software:
  - **EViews** (econometric analysis software),
  - **Stata** (general statistical software),
  - **SAS** (general statistical software),
  - **S-PLUS** (general statistical software) and
  - **SPSS** (general statistical software).

0.7. Data dissemination

- This phase refers to all means used to make data public, including:

  - Publication of documents, in particular press releases, periodicals and special issues

  - Electronic dissemination of statistics – for example on CD-ROM, USB stick or via the internet

  - Sending statistics in a printed or electronic version in response to direct requests

  - Setting up automated systems to provide access to statistics on request by telephone or internet

- To ensure that data are used to their full potential, it is important to consult users to determine the dissemination method most suited to their needs.

0.8. Data quality management (1/6)

- Quality should be aimed at in the production system (institutional aspects, human, material and financial resources) and in the outputs (tools, methodology, operations);
- It should be an integral part of statistical activities through internal and external checks at all steps in the programming and statistical compilation process;
- This search for quality is a requirement for international comparability. But it must be well planned as part of on-going official statistics compilation and management activities in order to reduce related costs.

0.8. Data quality management (2/6)

The prerequisites of data quality are:

- Favourable legal and institutional framework
- Appropriate human, financial and material resources for agricultural statistics compilation programmes
- Acknowledging that:
  - Statistics contain information relevant to the area of specialization;
  - Quality is a condition that governs all statistical compilation work.

0.8. Data quality management (3/6)

- There are several international references for data quality management. They have been set up by organizations such as:
  - International Monetary Fund (IMF) with the General Data Dissemination System (GDDS), the Special Data Dissemination Standard (SDDS) and the Data Quality Assessment Framework (DQAF);
  - The United Nations Statistics Division with the National Quality Assurance Framework (NQAF)
  - Paris21 with the questionnaire on Statistical Capacity Building Indicators (SCBI)

0.8. Data quality management (4/6)

- …… They have been set up by organizations such as:
  - Statistical institutes: Statistics Canada, Eurostat, INSEE, etc.
  - The FAO by setting up an integrated surveys system, an integrated database and the emergence of new technologies (PDA, GPS, remote sensing).

- The majority of international data quality frameworks are based on the framework developed by the IMF: the Data Quality Assessment Framework (DQAF).

0.8. Data quality management (5/6)

The dimensions adopted by the internmational organizations to define the quality of official statistics:

- **The relevance** of the information expressing how the information meets the real needs of users
- **The accuracy and reliability** of the statistical information expressing the extent to which the information correctly describes the phenomenon it should assess
- The statistical information is **timely and punctual** taking into account its publication date in relation to its reference date

0.8. Data quality management (6/6)

The dimensions adopted by the internmational organizations to define the quality of official statistics:

- **Accessibility and clarity** which refer to the ease with which the information can be obtained from the data producer
- **Interpretability or metadata** characterized by the availability of additional information necessary for its interpretation (metadata)
- **The consistency and comparability** of the statistical information which is guaranteed when this information can be reconciled with other statistical information in a general analytical framework

Figure 2 : Data quality assessment framework

FAO	UNSD	UNECE	OECD	EUROSTAT	IMF
Relevance	Relevance	Relevance	Relevance	Relevance	Prerequisites of quality
					Methodological soundness
Accuracy and reliability	Accuracy and reliability	Accuracy	Accuracy	Accuracy	Accuracy and reliability
Timeliness and punctuality	Timeliness and punctuality	Timeliness and punctuality	Timeliness	Timeliness and punctuality	Serviceable: Timely availability for decision making
Accessibility and clarity	Accessibility and clarity	Accessibility and clarity	Accessibility and interpretability	Accessibility and clarity	Accessibility
	Metadata				Assurance of integrity
Comparability and coherence	Comparability and coherence	Comparability	Coherence	Comparability and coherence	Serviceable: Timely availability for decision making
		Considered more relevant at the level of the organization	Credibility		Prerequisites of quality Prerequisites of integrity

Exercises

- **Exercise 1: Definition of the population**
- **Exercise 2: Sampled population vs target population**

THANK YOU