Training Course on "Basic Data Analysis for Price and Foreign Trade Data"

Ms. Münevver Nilgün Egemen
Statistical, Economic and Social Research and Training Centre
for Islamic Countries - (SESRIC)

Central Department of Statistics and Information
Saudi Arabia, Riyadh
(24-26 December 2011)

Data Quality

and

Data Validation

Quality;

"Fitness for its intended use"

Data Quality

Basic aspects of data quality;

- Characteristics of data collection process

- Characteristics of the statistical product

- Perception of the statistical product by user

Data Quality

Basic aspects of data quality;

Process Quality

Identification and measurements of key process variables effecting quality and cost

Product Quality

User demand product quality; timeliness, accuracy and accessibility

User Needs

Determines quality indicator

Data Quality

Components of data quality;

1. Relevance

2. Accuracy

3. Timeliness and punctuality

4. Accesibility and clarity

5. Comparability

6. Coherence

Data Quality

Relevance

Requires identification of user groups and their needs

- Coverage of the required topics
- Use of appropriate concepts

Characterised by the degree to which the data serves to address the purposes for which they are requested by users

Data quality

Accuracy

Closeness between the values provided and the unknown true values

- Sampling errors
- Non-response rate

Characterised by the degree to which the data Correctly estimate or describe the quantities or charecteristics they are designed to measure

Data quality

Timeliness and punctuality

The length of the time between availability of data and the phonemonen they describe

It implies the existence of a publication schedule and reflects the degree to which data are released in accordance with it.

Data quality

Accesibility and clarity

How readily product can be located and Accessed and user can easily be understand and properly use data

It requires different modes of dissemination availability of metadata and user support services to ensure proper use and analysis of the data by the user

Data quality

Comparability

Adherence to the international standards

It requires compliance with international standards in terms of methodology and Classifications used

Data quality

Coherence

It is the degree to which they are logically connected and mutually consistent

Within data set
Across data set
Over time
Across countries

Data quality

Table 1:   List of Standard Quality Indicators (Eurostat 2005d)

Quality com-ponent	Indicator	1=Key 2=Supportive 3=Advanced
Relevance	R1. User satisfaction index	3
	R2. Rate of available statistics	1
Accuracy	A1. Coefficient of variation	1
	A2. Unit response rate (un-weighted/weighted)	2
	A3. Item response rate (un-weighted/weighted)	2
	A4. Imputation rate and ratio	2
	A5. Over-coverage and misclassification rates	2
	A6. Geographical under-coverage ratio	1
	A7. Average size of revisions	1
Timeliness and Punctuality	T1. Punctuality of time schedule of effective publication	1
	T2. Time lag between the end of reference period and the date of first results	1
	T3. Time lag between the end of reference period and the date of the final results	1
Accessibility and Clarity	AC1. Number of publications disseminated and/ or sold	1
	AC2. Number of accesses to databases	1
	AC3. Rate of completeness of metadata information for released statistics.	3
Comparability	C1. Length of comparable time-series	1
	C2. Number of comparable time-series	1
	C3. Rate of differences in concepts and measurement from European norms	3
	C4. Asymmetries for statistics mirror flows	1
Coherence	CH1. Rate of statistics that satisfies the requirements for the main secondary use	3

Data Validation

Basic steps in a statistical activity

Definition of data requirements

Planning of data collection methodology

Preparing questionnaire and metadata files

Data Validation

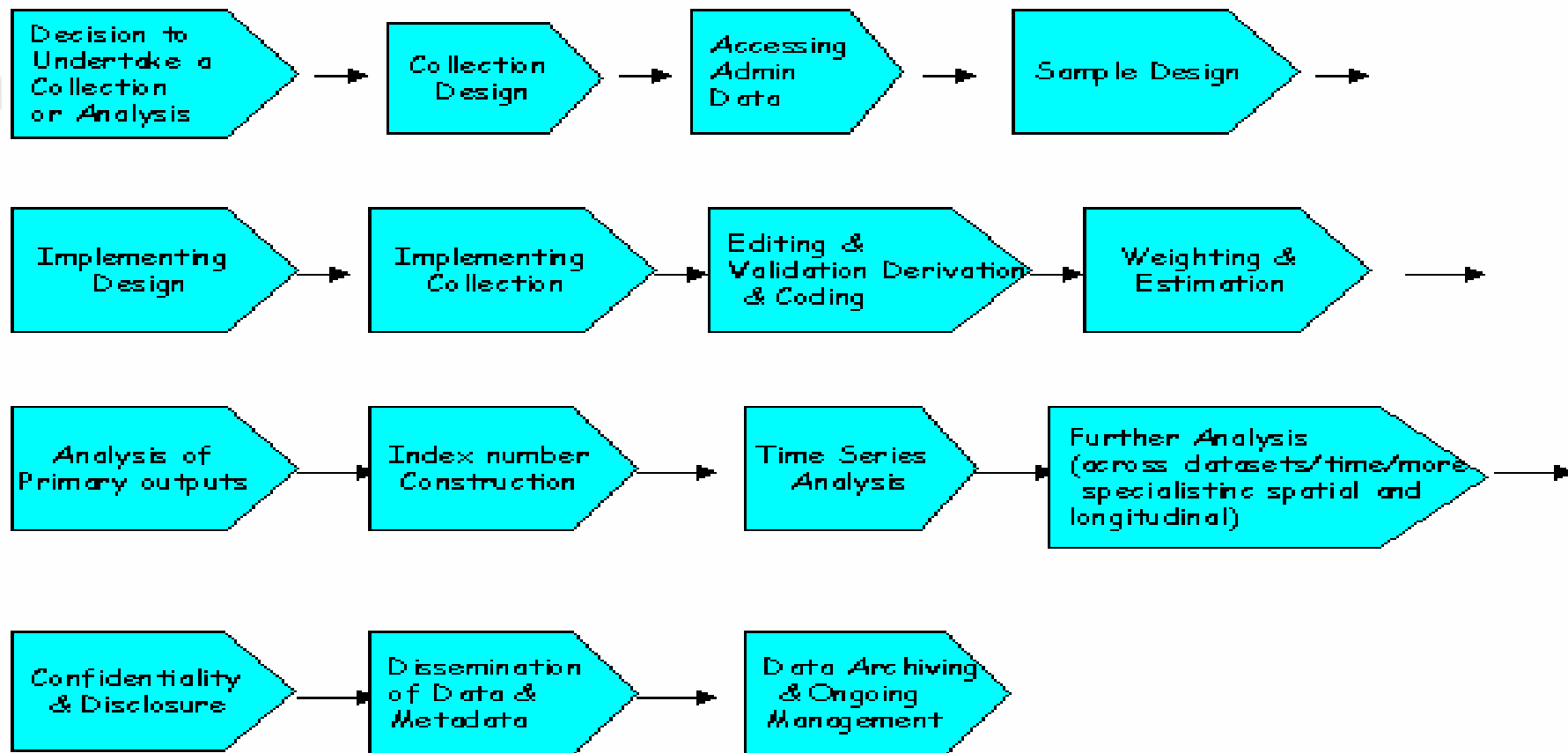## Basic steps in a statistical activity

Collection of data

Data capture, editing and analysis of data

Dissemination of data

Data Validation

STATISTICAL VALUE CHAIN

Decision to Undertake a Collection or Analysis → Collection Design → Accessing Admin Data → Sample Design →

Implementing Design → Implementing Collection → Editing & Validation Derivation & Coding → Weighting & Estimation →

Analysis of Primary outputs → Index number Construction → Time Series Analysis → Further Analysis (across datasets/time/more specialistic spatial and longitudinal) →

Confidentiality & Disclosure → Dissemination of Data & Metadata → Data Archiving & Ongoing Management

METADATA

Data Editing

Data editing; is the application of

checks that identify missing, invalid or

inconsistent entries or that point to data records

that are potentially in error.

40% of resources used

Most difficult and time consuming process in data collection

Data Editing Process

1. **Survey managament**
   - *Completeness checking (survey and questionnaire level)*
   - *Quality control*
2. **Data Capture**

   Conversion of data to electronic media
   - *Heads down - no error detection*
   - *Heads up - data entry with a review at time of entry* requires subject matter knowledge by the individuals entering the data

OIC ANKARA
SESRIC

Data Editing Process

3. Data review ; *Error detection and Data analysis*

Manual data review may occur prior to data entry. The data may be reviewed and prepared/corrected prior to key-entry.

Automated data review may occur in a batch or interactive fashion. It is important to note that data entered in a heads-down fashion may later be corrected in either a batch or an interactive data review process. CATI and CAPI Interactive data

3. Data review ; *Interactive fashion*

Computer-Assisted Interviewing (CAI) combines interactive data review with interactive data editing while the respondent is an available source for data adjustment.

Computer-Assisted Telephone Interviews (CATI)
Computer Assisted Personal Interviewing (CAPI)
Computer Assisted Web Interniewing (CAWI)

can perform validation at the time of data collection but requires more qualified staff

3. Data review ;

Item level - Validations at this level are generally named "range checking".
Example: age must be > 0 and < 120.

Questionnaire level - This level involves across item checking within a questionnaire.
Example: If married = 'yes' then age must be greater than 14.

Hierarchical - This level involves checking items in related sub-questionnaires.
Example household questionnaires the sum of the individual's data for an item does not exceed the total reported for the household.

Data Editing Process

4. Data adjustment; *Manual or automated imputations*

Deterministic - where only one correct value exists, as in the missing sum at the bottom of a column of numbers. A value is thus determined from other values on the same questionnaire.

Model based - use of averages, medians, regression equations, etc. to impute a value. It requires use of specialised softwares (expert systems)

Macro Editing
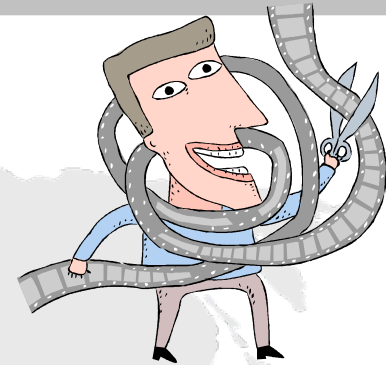
- A set of checks based on statistical analysis of respondent data

  e.g., the ratio of two fields lies between limits determined by a statistical analysis of that ratio for presumed valid reporters.

- A statistical edit may incorporate cross-record checks

  e.g., the comparison of the value of an item in one record against a frequency distribution for that item for all records.

- It may also use historical data on a firm-by-firm basis in a time series modelling procedure.

Statistical Classifications

Classification

Grouping and organising information meaningfully and systematically into a standard format that is useful for determining the similarity of ideas, events, objects or persons
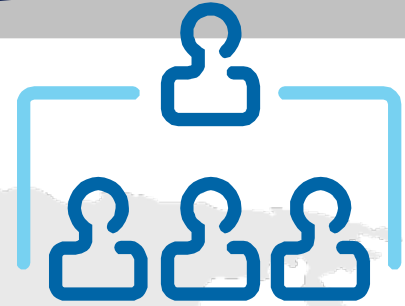
Statistical Classifications

Statistical classification

A classification having a set of discrete categories which may be assigned to a specific variable registered in a statistical survey or in a administrative file, used in the production and presentation of statistics

Statistical Classifications

International standard classifications (ISC)

developed and adopted by international organisations to ensure correct implementation of agreements and standardise national and international communication. They are the products international agreements among NSOs.

Statistical Classifications

International standard classifications (ISC)

Classifications of economic activities cover all economic activities - from agriculture to services

   ISIC Rev.3 International Standard Industrial Classification of All Economic Activities

Classification of products used for statistics on many aspects of the production and use of products

   CPC Ver.2 Central Product Classification
   SITC Rev.4 Standard International Trade Classification

Statistical Classifications

International standard classifications (ISC)

ISCED2011- UNESCO -International Standard Classification of Education – Education Statistics

ISCO-08 – ILO - International Standard Classification on Occupation – Labour Force

ICF 2001 – WHO- International Classification on Functioning, Disability and Health – Health Statistics

Statistical Classifications

http://unstats.un.org/unsd/cr/registry/regct.asp?Lg=1

References

1. Statistical Data Editing, Volume No.1, Methods and Tehniques, UN, 1994

2. Handbook on Improving Quality by Analysis of Process Variables, Eurostat,

3. Standard Statistical Classifications: Basic Principles, UNSD,

Thank You !

http://www.sesric.org
© 2011