PLAN FOR DATA PROCESSING AND ANALYSIS

Plan for data analysis

- The procedures for the analysis of data collected through **qualitative** and **quantitative** techniques are quite **different**.
  - For **quantitative data** the starting point in analysis is usually a ***description of the data for each variable*** for all the study units included in the sample.
  - Processing of data may take place during data collection or when all data has been collected; description and analysis are usually carried out *after* the fieldwork has been completed.

Plan ...

- For **qualitative data** it is more a matter of ***describing, summarizing and interpreting the data*** obtained *for each study unit* (or for each group of study units).

- Preparation of a plan for data processing and analysis will provide you with better insight into the feasibility of the analysis to be performed as well as the resources that are required.

Plan for data analysis

- Data presentation will cover how the findings will be presented in order to answer the study objectives.
- **Data** should be **presented** in a ***simple, clear, and easily understandable way.***
- Only the most important and relevant data must be presented

Data Analysis and Interpretation

- **Think about analysis EARLY**
- Start with a plan
- Code, enter, clean
- Analyze
- Interpret
- **Reflect**
  - → What did we learn?
  - → What conclusions can we draw?
  - → What are our recommendations?
  - → What are the limitations of our analysis?

Why do I need an analysis plan?

- It helps the researcher assure that at the end of the study:
  - all the information (s)he needs has indeed been collected, and in a standardized way;
  - (s)he has not collected unnecessary data which will never be analyzed.
- To make sure the questions and your data collection instrument will get the information you want.
- Think about your “report” when you are designing your data collection instruments

What should the plan include?

- When making a plan for data processing and analysis the following issues should be considered:
  - Sorting data
  - Performing quality-control checks
  - Data processing, and
  - Data analysis.

I. SORTING DATA

- If you have **different study populations** you obviously would number the questionnaires *separately*.

II. PERFORMING QUALITY CONTROL CHECKS

- Most important data checks,
  1. Accuracy of data
  2. Checking outliers
  3. Handling missed values

Accuracy

- Verify the accuracy of sample of completed data.
- Entered twice & compared /use verifier to .../

III. DATA PROCESSING - quantitative data

- Manually
- Computer

Data processing in both cases involves:

- **Categorizing** the data,
- **Coding**, and
- **Summarizing** the data

DATA PROCESSING…..

1. **Categorizing**
   - categorical & numerical variables

- For categorical variables the categories have been decided upon beforehand.
- However, there should always be a category called ‘Others, specify . . .’ which can only be categorized afterwards. (not more than 5%)

- ✓ The **numerical variables** are often better collected without any pre-categorization.
- ✓ If your categories had been wrongly chosen you cannot reclassify the data anymore.

2. Coding

- is assigning a separate (non-overlapping) numerical, letter code for separate answers and missing values.

Grouped as PRE, POST and RECODING

- ***Pre-coding*** ;-when questionnaire being written so that ***coding categories*** appear in the questionnaire
- ***Post-coding***:- After respondents have answered questions.
- ***Recoding***:-changing earlier coding to facilitate meaningful analysis.
  - ✓ Done after data have been collected.

Cont'd …

- **Missing values** should **not** be entered as a '***blank***', because some statistical packages interpret blanks as zeroes.

- Ideally, a code should be chosen to denote a missing value (e.g. a code '9' or '99' or '999' is often used missing values).

Data processing…..

- Common responses should have the same code in each question

For example:

- Yes (or positive response) code - Y or 1
- No (or negative response) code - N or 2
- Don’t know code -D or 8
- No response/unknown code - U or 9

3. Summarizing the data

1. Data master sheets

- On a data master sheet all the answers of individual respondents are entered by hand.

For example, the straightforward answers of the smoking questionnaire for male smokers could be processed as follows

Table 1: Master sheet for smokers (males)

No.	Q1 Sex	Q2 Age		Q6 No of cig.		Q7 Age onset		Q9 Tried to				Q14 Cough > 2wks		Q14 Cough/ chest ever	
								reduce		stop					
		Yrs	Cat	No	Cat	Yrs	Cat	Yes	No	Yes	No	Yes	No	Yes	No
1	M	18	(1)	10	(2)	12	(2)	1x			√		√		√
2	M	35	(3)	30	(4)	20	(4)		NR	1x			√	√	
3	M	54	(4)	15	(2)	14	(2)	10x		3x		√		√	
Etc.															
Total	31	Av	35	Av	20	Av	18	26	4 + 1NR	19	12	5	26	11	20

2. Compilation by hand (without using master sheets)

- If only **one person** is doing the compilation use **manual sorting**.
- If a team of **2 persons** work together use either **manual sorting** or **tally counting**.

- To do **manual sorting** the basic procedure is to:
  - ✓ Take one question at a time, for example, 'use of health facility',
  - ✓ Sort the questionnaires into different piles representing the various responses to the question, e.g., hospital/ health centre/ traditional practitioners) and
  - ✓ Count the number in each pile.

Compilation by hand….

- When you need to sort out subjects who have a certain combination of variables (e.g. females who used each type of health facility) sort the questionnaires into piles according to the first question (gender), then subdivide the piles according to the response to the other question (use of health facility)

Compilation by hand…

- To do **tally counting** the basic procedure is:
  - ✓ One member of the compiling team reads out the information while the other records it in the form of a tally.
  - ✓ After tally counting, add the tallies and record the number of subjects in each group.

3. Computer compilation

- The **computer should not be used** if your **sample is small** and the data is mainly generated by open questions (**qualitative data**),

- The larger the sample, the more beneficial in general the use of a computer will be.

Computer compilation……

- **Computer compilation consists of the following steps:**

  I. Choosing an appropriate computer program
  II. Data entry
  III. Verification or validation of the data
  IV. Programming (if necessary)
  V. Computer outputs/prints

I. Choosing an appropriate computer program

The most widely used programs are:

- **Excel**, easily accessed by most of us and ***simple to use***
- **Epi Info**, a very consumer friendly program for ***data entry***
- **SPSS**, menu driven computer program which is good for ***data analysis***

ii. Data entry

- Is the transfer of data from questionnaire to **appropriate computer program** of a computer.
- During data entry, the information relating to each subject in the study is keyed into the computer in the form of the relevant code.
- When working on computers, note to:-
  - Saving your works frequently
  - Keep back-up (more than one copy)
  - Enter twice, if resource permits.

iii. Verification

- This involves giving the appropriate commands to identify errors.

Example:

- ✓ The computer can be instructed to identify and print out all subjects where the ‘sex’ column has a code different from 1 (F) or 2 (M).

iv. Programming

- **Communicate effectively** with computer personnel .

- You as a researcher should tell the computer personnel:-
  - ✓ the names of all the variables,
  - ✓ their locations.
  - ✓ ....to be analyzed and compared .

v. Computer outputs

- The computer can do all kinds of analysis and the results can be printed.
- It is important to decide whether each of the tables, graphs, and statistical tests that can be produced makes sense and should be used in your report.

Data Presentation

- Analysis of quantitative data involves the production and interpretation of frequencies, tables, graphs, etc., that describe the data.
- Collected data can be **presented** as follows:
  - Raw numbers, texts
  - Measures of central tendency
  - Percentage
  - Tables
  - Graphical Presentations

Tables

- A table is designed to stand alone from the text.
- Since a table is intended to communicate information, it should be easy to read and understand.
- It should have a **clear title** and the rows and columns must be clearly ***labeled***.
- The **title of the table** should be before the table (**at the top**)

Presentation

- The **title of the table** should **answer** the question ***what, when, where, how classified***?
- If data is not original, the **source** should be stated
- Simple frequency table will be used to describe the level of some variables with the occurrence of the categories of the variables
  - If numbers are large it is better to use **(relative frequencies)**

Presentation

- Cross tabulation will be also used to reveal relation between different variables, such as:
  - the level of education and utilization of FP
  - Place of birth and infant mortality
  - Type of training and performance

- Testing for statistical significance for differences between categories (if needed) and testing for significant relationships (statistical inference).

Cont'd …

- Cross-tabulations-Further analysis of the data usually requires the combination of information on two or more variables in order to describe the problem or to arrive at possible explanations for it.

- Descriptive cross-tabulations.
- Analytic cross-tabulations

- In order to visualize how the data can be organized and summarized it is useful to construct DUMMY cross-tabulations.

- A **DUMMY TABLE** contains all elements of a real table, except that the ***cells are still empty***.

Cont'd …

When you construct a dummy table

- If a dependent and an independent variable are cross-tabulated
- All tables should have a clear title and clear headings for all rows and columns.
- All tables should have a separate row and a separate column for totals
- All tables related to a certain objective should be numbered and kept together

Graphical presentation

- Appropriately drawn graphs will help readers to grasp the point quickly and easy to memorize
- The type of graph that we need to use depends on the type of variable (categorical or numerical)
- The choice between graph and table depends on the number of categories the variable has and the complexity.
- **The title of the graph should be written <u>at the bottom</u>**

Writing results

- **Computer programs are garbage-in-garbage out(GIGO)**
- **Forbidden to copy** computer output as it is.
- Not allowed to copy and paste table output of computer.
- You need to form your own table and only take the numbers and interpret.
- Unexpected results should be checked and should be corrected
- **Don’t copy results if you didn’t understand well**