CHAPTER VII

PROCESSING THE DATA

This chapter is for survey coordinators and technical resource persons. It gives guidance to help you:

✔ Carry out data entry.

✔ Check and edit the data, and create a "clean" data file for analysis.

✔ Evaluate the quality of the survey data.

✔ Calculate estimates of the indicators.

INTRODUCTION

Data processing should take place as quickly as possible. First results of the survey, the preliminary report, should appear quickly, within a few weeks after the field work ends. A full debriefing of the survey team, with a review of the initial results, contributes to a good final report. The enthusiasm and interest generated by the field work will evaporate quickly when the field work ends, so be sure to keep the leverage the survey results can give you by reporting them quickly.

As part of field procedures to reduce error, interviewers should check through the questionnaires when they complete an interview. Errors can often be corrected on the spot, without the need to re-interview. Field supervisors should also check through all questionnaires immediately after collecting them from interviewers. Are they complete? Are they clear? Are the "skip" instructions followed correctly? The survey manager should also check all questionnaires in the earliest stage of the field work, and continue to check samples of the questionnaires throughout the survey.

Data entry is best begun while interviewers are still in the field. This allows you to spot and correct mistakes that certain interviewers or teams may be making. Serious problems that may escape the field supervisor’s notice can then be picked up quickly, in time to re-train field staff and correct serious errors. Field supervisors are responsible for checking the questionnaires for completeness and consistency, and for classifying any responses the interviewer could not code. Except for checking on the numbers of questionnaires returned and for completing identifying numbers on the forms, it should not be necessary for office staff to do any editing or coding of the questionnaires that are returned to the office.

The data-entry system must be carefully designed and data-entry staff must be well trained and motivated. The programs for data entry, checking and analysis should be set up and verified using data from the pilot survey before the main field work phase begins. Any problems should be ironed out, and the programs modified if necessary, before the survey data start to come in from the field.

There are four main phases for the computerized data-processing of the questionnaires:

1. data entry
2. data checking and editing
3. frequency distributions
4. variable construction and tabulations

If any modifications have been made to the questionnaire, the data entry and analysis programs will have to be modified by a programmer with excellent knowledge of the EPI INFO program package.

A variety of software is available for help in performing these functions. With this handbook, a complete program for data entry, checking and analysis to calculate all the indicators of the seven Mid-Decade Goals (that is, those that can be measured with a multiple-indicator survey) is supplied for use with the EPI INFO (Version 6) package of programs. The programs are also listed in Appendix 3.

EPI INFO was developed by the Centers for Disease Control and Prevention in collaboration with WHO (available from USD, Inc., 2075 A West Park Place, Stone Mountain, GA 30087, USA) and can perform all four phases of the data processing. A copy of EPI INFO (Version 6) is included with this handbook.

The data processing supervisor should set up the data entry system in advance. Advice for doing so is found in Appendix 3.

ENTERING THE DATA INTO THE COMPUTERS AND PRODUCING "CLEAN" DATA FILES

Data Entry

As soon as data from one cluster arrive back at headquarters, data entry should begin. Data entry should be carried out in small batches (e.g., a cluster at a time).

Data from the questionnaire should be entered for each household in the order they are collected: the household data from the listing page, the data on all children on the child roster, the water and sanitation and salt iodization modules, education data for each child over school-entry age, tetanus toxoid module for each mother, care of acute respiratory illness (CARI) module for each mother and, last, child health modules for each child. Once all the data for one household are entered, the clerk moves on to enter the data for the next household.

Note that missing information during the field work is recorded as 9's—for example, 9, 99, 999, and so on. Blank fields at data-entry time and inapplicable responses (e.g., for answers to optional questions or skipped questions) are stored as blank fields in the data base. As data for each household are entered, the program performs *range checks*. These ensure that no number is entered for an item that is out of the given range of responses for that item.

✎ **EXAMPLE:**
The date of interview should be recorded during the period of field work. A date of interview recorded as 01 11 93 is outside the range of dates allowed for this survey. If the data entry program has established the range for this variable, the computer will recognise that this date entry is invalid.

There are two approaches to entering data. Data can be entered into the computer twice, by two different data-entry clerks. The two files are then compared for inconsistencies, and other data checks performed. Another approach is to enter the data only once, but to perform a number of checks to prevent errors at the time the data are entered. Interactive data-checking can slow the data-entry process, but can avert many common data-entry errors made by clerks. Consistency checks are then run and the file corrected. For these multiple-indicator surveys, we advise the latter option, which is the approach that is used in the data-entry programs provided with this handbook.

Box 7.1
PROS AND CONS OF DOUBLE ENTRY

Pros

Data entry errors are virtually eliminated if done correctly. This option, followed by batch-error checking, may be cheaper and faster when professional data entry clerks are used.

Cons

Double entry doubles the cost of data entry. It complicates questionnaire flow in the office, which can also lead to errors.

A possible compromise is to enter some questionnaires twice as a way to spot-check the accuracy of data-entry clerks and identify systematic errors that are being made.

Data Checking

Experience has shown that range errors are almost always data-entry errors and can be usefully checked for and corrected immediately at data-entry time. Consistency errors, however, must be resolved by returning to the household or by carefully examining the questionnaire. Consistency checking is therefore better carried out as a separate step, with errors reported on hard copy that can be used for marking in corrections.

The EPI INFO data-entry program supplied with this handbook automatically applies a large number of checks to these data, either alerting the data-entry clerk whenever an out-of-range response is entered, or reporting inconsistencies on the final pass through the data-checking program.

☞ Once the data are in headquarters, it is usually not possible to return to the households for correcting inconsistent answers. The study coordinator or a senior supervisor will have to make "best guesses" based on the existing information to correct the questionnaire. If that is not possible, the variable should be treated as "missing."

After data for an entire cluster are entered, the batch should be checked for consistency and edited. The program checks to ensure that all variables appropriate to the case are in the file (that is, the structure of the file is correct) and are within the correct range. The program performs internal consistency checks to ensure that the data have been recorded and entered properly. The data-entry instructions found in Appendix 3 discuss procedures to follow when errors occur. Reference to the original questionnaire at this stage can often help to sort out problems in the data.

✎ **EXAMPLE:**

Consistency checks verify that the following requirements are met:

— The date of birth for each child must precede the date of interview.

— Dates of birth for all children must be within the past 15 years.

— If the water source is "piped-in dwelling," then the distance should be "on premises."

— If a breastfed child received "ONLY breast milk," then no other items should be answered affirmatively.

— The date of first DPT should precede second DPT, which should precede third DPT.

— Anthropometric indices are not flagged as invalid due to errors in age- or weight-data entry.

It may be useful to look at the number of errors for each team or interviewer, so that further training may be carried out in the early stages of field work.

Once corrections have been noted for reported inconsistencies, the corresponding records are updated on-line. The checking program should then be run again and the procedure repeated until all errors that can be corrected in the office have been rectified. After these checks are performed, the data are "clean" and the files are ready to be analysed.

EVALUATING THE QUALITY OF THE SURVEY DATA

The next step is to execute *frequency distributions* of all the variables in your data file and the possible values they can take (see Table 7.1). This will allow you to check the data again for inappropriate entries and to note the number of missing values in the data.

Table 7.1. Example of a frequency distribution

Item number	**Household water supply source**	**Distribution**
1	Piped-in dwelling	146
2	Public tap	440
3	Tube well or borehole	102
4	Protected dug well	95
5	Unprotected dug well	660
6	Pond, river or stream	35
7	Tanker-truck, vendor	55
9	Other	15
.	Missing	6
Total		**1,554**

You may need to look back at the original questionnaire to check data entries. You may then need to correct the raw data file and perform this step again.

☞ Referring to the example in Table 7.1: Suppose you want to examine the 15 questionnaires with "other" water source. You will need to program the computer to select only those records that have a value of "9" (other) for the variable "water source." You must then tell the computer to list the identification number of the households with these values. You can use the identification number to go back to examine those questionnaires.

You should keep a record of how many cases in the file were incomplete, how many households were listed but not contacted and what errors were noted in the data. The per cent of eligible households which were not contacted should be examined and reported. A response rate below 90–95 per cent means that the survey does not necessarily represent the entire population.

One way to check the quality of the data is to examine the internal consistency of the data collected. Consistency checks which were run at the time of data entry can be run again for the entire data set to determine the number of records for which inconsistencies remain. (See the example above on requirements for consistency.)

A high number of unresolved inconsistencies is an indicator of sloppy data collection (and sometimes falsification of data) and could damage the validity of results.

A second indicator of data quality is the per cent of responses which are "don't know" or "other." These percentages can be obtained by examining a frequency distribution for the items in the data set. A high per cent of "don't know" indicates that respondents did not understand the questions or that the information requested is too difficult to report. Results based on such variables are questionable. A high per cent of "other" responses indicates that the categories on the questionnaire do not capture the responses most commonly given.

Another indicator of data quality is how well they compare with other existing data. For example, the distribution of mothers and children by age and the ratio of boys to girls should be similar to the age distributions and sex ratios found in a recent census or other recent survey of the same population. Even in the absence of a recent census or survey, the number of boys and girls enumerated should be roughly the same. The number of children in each one-year age group should also be similar, with a slight decline in numbers with increasing age.

CALCULATING THE INDICATORS

The basic tables necessary to calculate each indicator are listed in Appendix 4, and the programs to produce the estimates are given in Appendix 3. All the estimates can be produced with the programs found on the diskette supplied with this handbook. Otherwise, a standard package such as SPSS PC can be used.

You can obtain some of the indicators—for example, the per cent of households with iodized salt—directly from responses to variables on the questionnaire. For most indicators, however, you will have to examine the responses to a number of questions to determine the information needed in the indicator. For example, access to safe water is based on both source of water and distance to source. In these cases, a variable must be created indicating whether or not the household, mother *or* child fits the definition of the indicator.

☞ The EPI INFO program supplied with this handbook will have to be adapted to report on indicators that vary from country to country—for example, water and sanitation.

Most analysis packages, including SPSS and Epi Info, will report a table of information on the variables created for the indicators. To report on the indicator, the appropriate number must be read off the table.

✎ **Example:**

A table on low weight-for-age might be generated, as follows:

Value label	Value	Frequency	Per cent	Valid per cent	Cumulative per cent
≥ −2 SD	0	1,183	77.9	79.0	79.0
< −2 SD	1	314	20.7	20.9	100.0
	.	21	1.4	missing	
Total =		**1,518**	**100.0**	**100.0**	
Valid cases: 1,497					*Missing cases: 21*

SD = standard deviation

The number to report for Indicator 11.1 in this case would be 20.9 per cent, the per cent of children whose weight-for-age was low, among those who had a valid weight-for-age in the data set. In the Epi Info package included with this handbook, the indicators are printed directly using the REPORT function, so that only the necessary information is shown.

Anthropometric indices need to be calculated for each child, based on the measured weights and heights. Epi Info uses a separate routine to compare a child’s age and weight (or height) to an international standard, and then assigns a score for each child based on his or her deviation from the median value of the standard (a z-score). The calculations can be done at the time of data entry, as is done in the Epi Info package supplied here, or can be done in a subsequent processing of the data using a special anthropometry package such as ANTHRO (formerly CASP) or EPINUT. The advantage of Epi Info is that the indices can be viewed on the screen as soon as the height and weight data are input so that the data-entry staff can immediately detect indices which are implausibly high or low (usually due to keying errors when these data are entered). Indices below −6.0 z-scores or above 6.0 are generally considered to be implausible.

✎ **EXAMPLE:**

The output generated by the REPORT program supplied with this handbook looks like this:

Proportion of Children with Low Weight-for-Age Z-Scores
(Indicator 11.1,2)

Total Children	**WAZ < −2**	**WAZ < −3**
1,497	21%	4%

☞ In some cases, the analysis package will not generate the indicator itself. For example, a survey will not give the absolute numbers in the population. To convert the per cent of the population with access to safe and convenient water (the number reported by the REPORT program) into absolute numbers (Indicator 13.1), it is necessary to multiply this percentage by the total population size.

Estimating the Margin of Error for Each Indicator

Most analysis packages will estimate the standard error of percentages and means, assuming that the sample taken was a "simple random sample."[1] However, these standard errors are generally underestimated because they do not take into account the fact that the data come from cluster sample surveys. Analysis packages specifically designed to work with cluster sample surveys are needed to correctly estimate standard errors. EPI INFO (Version 6) has a module called CSAMPLE to compute correct standard errors for percentages and means, although this module is not integrated within the ANALYSIS module. Appendix 3 contains instructions for generating the standard errors and confidence intervals for each indicator using this module.

The EPI INFO CSAMPLE program will automatically report the estimate along with its margin of error. See Appendix 3 for instructions on using this program.

[1] The formula for calculating the *standard error* (s.e.) for a proportion, assuming a simple random sample, is:

$$s.e. = \surd[p \times (1 - p) \div n]$$

where *p* is the proportion of interest and *n* is the sample size in the group. The *margin of error* for this proportion is (2 × s.e.). The *confidence interval* for the proportion is calculated as *p* ± the margin of error.

✎ **EXAMPLE:**

The margin of error for the indicator and its 95 per cent confidence interval are shown. For instance, in analysing weight-for-age, the output was:

Per cent with low weight-for-age is 21.0% with a confidence interval of (19.3, 22.7).

Calculating Weighting Factors for Non-Self–Weighting Samples

If separate sampling frames were used for different regions at the first stage of sampling, the national sample was not chosen with probability proportional to size (PPS). This may also happen if you stratified according to some other factor (e.g., urban/rural or slum/non-slum) and took different sampling fractions (proportions) in each stratum. These samples are not "self-weighting," and you must weight your sample when you report national estimates. That is, you must ensure that each separate subsample—for example, each separate region—contributes only what it would contribute if the survey sample at the national level had been chosen with PPS. The procedure for calculating weights is illustrated in Table 7.2 using a hypothetical example.

☞ Create a new variable to represent the weighting factor. This factor must be used to calculate *national* estimates from your separate subnational surveys. Each case will then be weighted by this variable.

For example, suppose the country has a population comprising 1,333,415 households and consists of seven regions, each of which you surveyed independently. The total sample of households was equal to 11,312.

To calculate the appropriate weighting factor for each survey sample, create a table such as Table 7.2.

☞ To calculate the actual coverage for the entire sample, construct a variable representing the weighting factor:

```
IF (REGION = 1) THEN WTFACTOR = 2.73
IF (REGION = 2) THEN WTFACTOR = 0.48
IF (REGION = 3) THEN WTFACTOR = 0.99
IF (REGION = 4) THEN WTFACTOR = 0.26
IF (REGION = 5) THEN WTFACTOR = 1.77
IF (REGION = 6) THEN WTFACTOR = 0.41
IF (REGION = 7) THEN WTFACTOR = 0.23
```

The analysis should then be weighted by the variable, WTFACTOR. The output will give national-level indicators properly weighted to allow for the non-self–weighting sample design.

☞ This calculation can also be done with a hand-calculator. Multiply each regional estimate by its weighting factor, add up the resulting products, and divide that sum by the sum of the weights.

✎ **EXAMPLE:**

(1) Region	(2) Estimate	(3) Weighting factor	(4) Estimate × WTFACTOR
1	0.85	1.7	1.445
2	0.75	0.2	0.15
3	0.70	1.0	0.70
Total		∑ = 2.9	2.295

National estimate weighted average: (sum of column 4 ÷ 2.9) = 0.79

Table 7.2. Illustrative table for calculating weighting factors

Stratum	(1) Population size (households)	(2) Stratum population as a proportion of national population	(3) Final sample size	(4) Stratum sample as proportion of total sample	(5) Weighting factor
		(col. 1 ÷ 1,333,415)	*n (households)*	(col. 3 ÷ 11,312)	(col. 2 ÷ col. 4)
1	550,088	.412	1,707	.1509	2.730
2	93,088	.070	1,664	.1471	.476
3	192,580	.144	1,648	.1457	.988
4	47,873	.036	1,560	.1379	.261
5	330,312	.248	1,587	.1403	1.768
6	77,124	.058	1,592	.1407	.412
7	42,350	.032	1,544	.1374	.233
Total	1,333,415	1.000	11,312	1.000	

Sex- and Age-Specific Estimates

☞ To calculate each indicator by the sex of the child, you must select from the data file children of only one sex, and run the analysis for this group separately. Do the same for children of the other sex.

Follow the same procedure if you want to report indicators for narrow age groups, or for subnational groups such as regional, district or rural/urban groups. Select each group separately, and run the analysis program to get a report for one group. Repeat the process for each group for which you want separate indicators calculated. If you are not using the EPI INFO analysis program, you may find it easier to cross-tabulate the variable for sex of child, or other similar variables that define subnational groups (for example, regional, urban/rural).

☞ Make sure that you label your printouts clearly (e.g., "boys only") to avoid later confusion.

Now you are ready to write your preliminary report, and to display these indicators graphically.