page or by questionnaire, as the latter discourages the operator from taking time to solve any problems he/she encounters, and could increase the number of errors in the final database. However, after the first or second day of data entry, it is a good idea to set daily quotas for questionnaires entered in the database by each data entry operator.

4.3.5 Paper questionnaire management during data entry

Tracking the filled-in questionnaires is extremely important during the entire survey process, ensuring that none is lost or entered more than once.

a. The filled-in questionnaires should be organized by cluster, in boxes or in envelopes. Clearly write the name of the cluster on the envelope or box. Identify an area where questionnaires will be stored prior to data entry, and a second area, clearly separate from the first, where questionnaires will be stored post–data entry. Identify a third area, near the data entry computers used, where the envelopes/boxes of questionnaires currently being entered can be stored. Finally, identify a fourth area where questionnaires with problems can be stored until their problems can be resolved.
b. Prepare a register in which all of the clusters and the numbers for the filled-in questionnaires in each of the clusters are recorded. This can be done by recording the name of the cluster and the questionnaire numbers contained in that cluster (e.g. 130052 through 130075). When the data entry operator selects a cluster for entry, she/he should sign and date the register to assume responsibility for that cluster's questionnaires. This will also help to track the questionnaires (which are also all clearly numbered).
c. Each data entry operator should work on one envelope of questionnaires at a time, and should be responsible for all of the questionnaires in that envelope. As data entry for each questionnaire is completed, the questionnaire should be returned to its envelope/box. This will prevent the loss or misfiling of questionnaires.
d. When the data entry operator starts to enter information from a questionnaire, he/she should record his/her name and/or code on the cover page of the paper questionnaire. If desired, this code can also be entered into the database during data entry, allowing easy identification of data entry operator during analysis.
e. When all information from a questionnaire is entered, a clear mark (e.g. a large checkmark or a slash) should be made across the entire front page using a highlighter. Once all questionnaires in an envelope/box are entered, the same mark should be made on the envelope/box.
f. When the data entry operator puts the envelope/box containing the already entered questionnaires in its designated area (see point a.), she/he should sign and record the date in the register.

4.4 HOUSEHOLD DATA ANALYSIS AND PROCESSING

4.4.1 Objective

This section is designed to help analysts (who should already have an in-depth knowledge of statistics) to analyse the household survey data generated by CFSVAs. These guidelines do not elaborate on common statistical or data management techniques, as this information is beyond the scope of these guidelines and must be acquired through academic course work and/or on-the-job training and supervised experience.

Why analyse primary data?
The CFSVA Food and Nutrition Security Conceptual Framework describes how various factors influence the food security situation and vulnerability of households. Using information obtained from various sources, the analyst describes and evaluates household food security status, the factors that influence household food security, the livelihood strategies employed, and the health and nutritional status and other livelihood outcomes at the household level.

Information generated by the CFSVA is used to explain how different households are exposed to risk and how they manage to cope. This information is combined with data obtained from secondary sources to describe the geographic, economic, and social context and explain the risk factors that influence the extent of vulnerability and the capacity to cope with shocks.

4.4.1.1 A note on statistical software

Because WFP uses SPSS for most of its quantitative data analysis, the guidance presented here focuses on that programme. However, experienced statisticians may choose to use other software packages.

For most cluster analysis, and often for principle component analysis, WFP-VAM uses ADDATI,[62] but SPSS can process this analysis, too.

For anthropometric z-score calculations of under-5s (stunting, wasting, underweight), WHO ANTHRO 2005[63] is used. Epi Info is essentially obsolete unless the ENA add-on for EPI Info is used.

4.4.2 Preparation for the analysis

4.4.2.1 Hierarchical data structure

CFSVAs consist primarily of household data. However, information gathered at the household level often includes data on individual household members, such as age, sex, children's education, nutritional status of mothers and children (under 5), women's childcare practices, and women's knowledge of HIV/AIDS. This may result in multiple "units," or cases, from each household. Additionally, data may be gathered at the village or community level that is pertinent to each household in the community (such as presence of schools and health clinics).

These data need to be organized into several data files, one for each unit of analysis, corresponding to the level at which the data were collected. For example, the member-level information (e.g. age, sex, children's education) should be saved into one file, while household-level information (e.g. assets, expenditure, current food consumption) should be saved into a separate file. There should be a separate file for anthropometric

[62] This software can be downloaded for free at: http://cidoc.iuav.it/~silvio/addawin_en.html
[63] This software can be downloaded for free at: http://www.who.int/childgrowth/software/en/

information for children (sex, age in months, height, and weight). Meanwhile, a different file should be used for child-care data. Similarly, if village-/cluster-level information is collected, a separate data file would be needed.
For CFSVAs, data can generally be organized in up to five data sets:

1. Village
2. Household
3. Individual
4. Mother
5. Child

It is essential to develop a data management plan before data entry.[64] Data entry application may automatically produce the five data sets (if designed to do so) or one large data set that needs to be reorganized into several data sets.

To obtain information about each of these levels, each of these data sets needs to be analysed. However, the analyst may desire to combine information from different data sets into one combined data set. Using SPSS, queries cannot be made between individual data sets; therefore, data sets must be merged in SPSS using a "many-to-one" relationship.

The merging of different data sets needs to be done so that member-level data sets can add to the information gathered from the household. Data analysis should in general be done only at the lowest level contained in that data set (i.e. member level), as described in Box 4.8.

Box 4.8: Relating child nutrition with other indicators

An analyst wants to look at child malnutrition as it relates to the household water source. There can be multiple children in each household, so the relationship between these two data sets (child and household) when merging is many children to one household. Using SPSS, this means merging the household data set into the child data set. This preserves all child information and keeps the number of children in the data set the same. However, in this merging process, some household information is lost (e.g., those with no children under 5) and some is duplicated (e.g. where there is more than one child in a household).

The resulting data set is used only for child-level queries (e.g. to answer the question "What percentage of wasted children live in households with unsafe drinking water supply?")

This merged data set cannot answer the question "What percentage of households have an unsafe water supply?" because some households were duplicated (i.e. those with more than one child) and others were deleted (i.e. those without any children). This question should be analysed within the household data set.

This merged data set cannot answer the question "What percentage of households have a wasted child as one of the members?" Analysis in this direction (from higher to lower aggregation level) is uncommon, and generally not recommended. To answer such questions, merging in an alternate direction would be required.

[64] See Section 4.2.5

4.4.2.2 Organization of the database

Organizing the database is an important step to getting a clear idea of the variables the analyst is going to consider. It is also helpful to manage and analyse the data by different individuals. **It is a good practice to make a copy of the database and keep it in a separate folder.** The following are key aspects of database organization:

a. Verify that all **variable names** clearly identify the question in the questionnaire. This can be easily done by using the question's code. Do not change the variable names unless it is absolutely necessary. Changing a name may complicate the identification of the particular variable when comparing it to the original raw data (from MS Access or another data entry tool), especially for other analysts who might access the data later. Additionally, if additional cases are to be appended to the database, differing variable names will impede the process.

b. Often variable names are cryptic; therefore it is necessary to enter **variable labels** to clarify what each variable is. A well-designed data entry programme, properly exported to SPSS, will already have appropriate labels for all variables, but this should be carefully checked. If the labels are clearly written and correctly spelled, it will be easier and quicker to create tables for reporting.

Box 4.9: Example of variable names and labels

In the questionnaire:

HQ5.1b What is the main source of drinking water for your family? 1…, 2…, 3…

The variable name could be HQ5.1b.
The variable label could be "drinking water source."

c. The variable type should also be correctly identified (usually "string variable" for letter/word values, and "numeric variable" for numbers). For categorical variables, it is necessary to enter the value labels for each variable, following the coding from the questionnaire. This information is essential for analysing the data and also for cleaning the categorical variables in the data set (see section 4.4.2.3 on data cleaning).

d. Identifying the measure (scale, ordinal, or nominal) of a variable is a key part of database organization. This information enables the software to conduct appropriate analyses with specific variables.

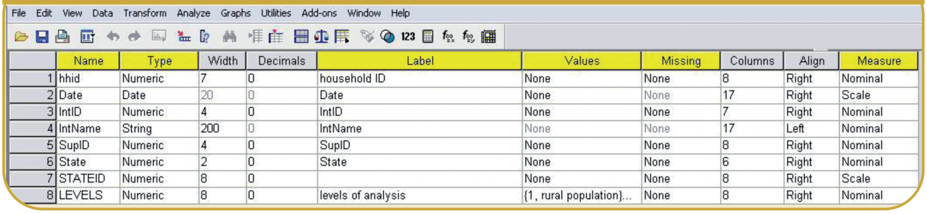
e. During data cleaning it is also important to specify whether a variable has one or more missing values. It is not uncommon for an analyst to be the first person to discover that a variable is missing values. Coding missing data can be done in several ways. However, each variable, and possibly even different analyses of the same variable, will have their own specific needs, and so there is no cut-and-dried rule for dealing with missing data.

f. Another step involves data recoding. This process is particularly useful for categorical variables. For example, yes/no questions are best coded as 1/0 (not 1/2). The 1/0 option is preferred because a simple mean illustrates the frequency, and in case of regression analysis, the yes/no questions are already recoded as binary variables, with the sign of the regression coefficient pointing in the intuitively correct direction. If the data entry programme is well designed, this should not be necessary. Boolean variables from MS Access translate automatically in SPSS into 1 (yes/present/true) and 0 (no/absent/false). A good and essential rule is not to lose data during recoding. For example, do not recode and simultaneously replace a continuous variable with a categorical variable. Instead, keep the original variable and create a new categorical variable.

g. A good practice is to keep only those variables meaningful to the analysis in the final dataset. If the analyst creates too many working variables before arriving at the final working data set, these variables will need to be deleted after computation, otherwise the size of the database will increase exponentially and become difficult to manage.

Figure 4.7 gives an example of all the fields that must be organized before data analysis.

Figure 4.7: Example of a database


File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
123

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	hhid	Numeric	7	0	household ID	None	None	8	Right	Nominal
2	Date	Date	20	0	Date	None	None	17	Right	Scale
3	IntID	Numeric	4	0	IntID	None	None	7	Right	Nominal
4	IntName	String	200	0	IntName	None	None	17	Left	Nominal
5	SupID	Numeric	4	0	SupID	None	None	8	Right	Nominal
6	State	Numeric	2	0	State	None	None	6	Right	Nominal
7	STATEID	Numeric	8	0		None	None	8	Right	Scale
8	LEVELS	Numeric	8	0	levels of analysis	{1, rural population}...	None	8	Right	Nominal

4.4.2.3 Data cleaning

Data cleaning is an essential step in data analysis. Every dataset contains some errors, and a significant amount of time in data analysis is spent "cleaning" the data.

Data cleaning can commence once the data are organized into different files. Data cleaning does not mean simply confirming that the data recorded on the paper questionnaires is the same as that in the dataset. It also entails several iterative steps of checking the dataset(s) to ensure that the data are credible.

Usually the cleaning of CFSVA data sets is done in several stages. The initial part of the data cleaning can be done with the software used for data entry[65] (most often MS Access). Cleaning should not be done as an automatic process but, rather, as a critical, well-thought-out series of recorded decisions.

UNIQUE ID

The first step in the data cleaning process is to ensure that the total number of households in the data set equals the total number of filled-in paper questionnaires. It is important to carefully review the data sets to confirm that all questionnaires have been entered only once and that all unique IDs are truly unique. This step ensures that data sets can later be merged and other household specific variables added, if necessary.

[65] See section 4.3.3.

In the more recent versions of SPSS, there is now an option, in the “Data” menu, to “Identify Duplicate Cases,” which makes this process very simple. If this option is not available, running a “frequency of the household ID” can be useful for detecting the presence of duplicate households. Household IDs resulting in a frequency of 2 or more have duplicates.

Additionally, it is a good idea at this point to take a random selection of questionnaires and compare them to the database (this should also happen as part of the quality control step during data entry). This verifies that all questions are being entered correctly and that variables are not being mislabelled.

Check the variables

The next step in data cleaning become more subjective, thus it is important not to make any permanent changes to the data unless you are absolutely confident in the decision. Regularly save backups of the database (without replacing earlier backups) so that any changes made can be undone at any time. These steps include checking for outliers and checking for errors/inconsistencies.

Check for outliers

An outlier is an observation that is numerically distant from the rest of the data. Statistics derived from data sets that include outliers will often be misleading. In most samplings of data, some data points will be further away from their expected values than what is deemed reasonable. In the presence of outliers, any statistical test based on sample means and variances can be distorted. Estimated regression coefficients that minimize the sum of squares for error (SSE) are also sensitive to outliers. Outliers can be caused by data collection/data entry errors or by extreme observations that for some legitimate reason do not fit within the typical range of other data values (High 2000).

Check the distribution of data values by levels of a categorical variable, if available. This procedure should always be one of the first steps in data analysis, as it will quickly reveal the most obvious outliers. For continuous or interval data, visual aids such as a dot plot or scatter plot are good methods for examining the severity of any outlying observations. A box plot is another helpful tool, since it makes no distributional assumptions, nor does it require any prior estimate of a mean or standard deviation. Values that are extreme in relation to the rest of the data are easily identified.

Running a frequency table or simple descriptive statistics could also be useful for detecting outliers. Working with outliers in numerical data effectively can be a rather difficult experience. Neither ignoring nor deleting them is a good solution. If nothing is done with the outliers, the results will describe essentially none of the data – neither the bulk of the data nor the outliers. Even though the numbers may be perfectly legitimate, if they lie outside the range of most of the data, they can cause potential computational and inference problems (High 2000). Outliers are not “missing,” just too high or low given our expectations; hence they should not be recoded as missing data. There are a couple of ways to deal with outliers.

- Since means are sensitive to extreme values, median values can be used instead of means.
- Maintain the "raw" version of the data, which retains the outliers, but create a "processed version" in which new variables are created that, for example, replace outliers with medians. Create a variable that denotes whether an outlier has been replaced by a median. That way, no data are discarded.

In both cases, it is crucial to report how outliers were managed.

Box 4.10: Possible effects of outliers

- Bias or distortion of estimates (especially of the arithmetic mean)
- Inflated sums of squares (which make it unlikely to partition sources of variation in the data into meaningful components)
- Distortion of p-values (statistical significance, or lack thereof, can be due to the presence of a few, or even one, unusual data value)
- Faulty conclusions (it is quite possible to draw false conclusions if irregularities in the data have not been investigated)

Check for errors/inconsistencies

Impossible values are often found in data sets, in spite of the filters used in the data entry programmes. Sometimes the values are absolutely impossible or contradictory to the information given in prior questions.

Box 4.11: Example of inconsistent values

- The number of the household members is 50.
- People with little or no land had a considerable harvest.

Once an inconsistent value has been identified, the data should be checked on the paper questionnaire to exclude the possibility of data entry error. If the data was entered wrongly in the paper questionnaire, the analyst should be able to decide whether to exclude the value, based on his/her experience and the type of variable. The rule is to change "absolutely impossible" values into "missing values," that is, if there is no way to determine if it is too time-consuming to investigate for the correct value. However, this is a subjective choice and should be approached with absolute caution.

Sometimes, even if the answer appears clean when you compare it with another variable, there is still an evident contradiction. In this case, the rules for an efficient data cleaning will be first to check the original questionnaire; if the answer is not there, look at other variables that can have a connection to those contradictory variables. If even this solution does not yield results, record the value of both the variables as missing.

Box 4.12: Example from Laos database

In the child database there is information about child demographics and household demographics.

Section 1, question 1.7: the enumerator should complete the information about the household demographics (number of people in the different sex and age groups).
Section 10, questions 10.3 and 10.4: the enumerator entered the information about the child's age and sex. In the database, the analyst found many inconsistencies that were difficult to solve, including households where the number of people in a specific age group does not match the number of children measured.
In this situation, the analyst, after cross-checking with the paper questionnaire, should try to find the truth in other variables (e.g. by looking at the variables related to education to see if the child was included in the wrong age group) or exclude the case from the analysis.
In other cases, the sex of children is different in the two sections. In this case, the information in Section 10 should be more accurate because the children were present during the measurements. So the analyst changed the variable 1.7 based on the information collected in 10.4.

Usually a questionnaire is developed with the flow of the questions kept in mind. In many cases there are skips in the questionnaire that allow the interviewer to bypass questions not applicable to a particular respondent. The data should be entered accordingly. For example, if a household did not cultivate any land, questions regarding harvest and crop types are not applicable to them. However, a well-designed data entry programme should automatically skip the fields that are not applicable to a household based on the information entered for the filter question.

4.4.2.4 General rules for data cleaning

- Do not start guessing, predicting, or assigning values. Even if a value seems obvious, do not make a change unless it is supported by clear evidence and the change is recorded.
- Bear in mind that most often the database will be managed/analysed by different people after the cleaning. Prepare an easy-to-read, clean database in which all the variables have their basic information; this will reduce mistakes and minimize time spent trying to comprehend the data set.
- Save a copy of the unchanged database before making any changes to it.
- Recode the history of changes in a syntax file. A record of the changes will be invaluable and can help replicate the same cleaning in different backups.
- In case of contradiction/inconsistency/doubts in the database between variables:
  1. Check the original questionnaire.
  2. Check the validity of the data by comparing them to other variables in the database.
  3. Change the value only if you are 100 percent sure.
  4. Consider the case as missing data.
  5. Keep the syntax for recoding automatic changes (if any).