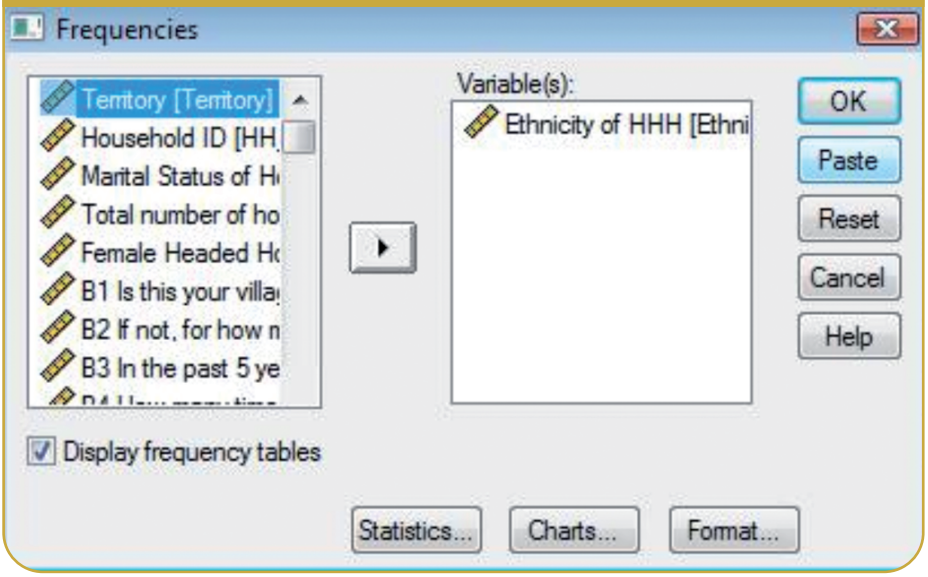
4.4.3 Data analysis

4.4.3.1 Standard practices

Figure 4.8: Pasting syntax



This section discusses several standard practices that are useful in CFSVA data analysis. By following these general guidelines, CFSVA analysts will have more compatible methods of organizing their analyses.

Use of syntax

In SPSS, the syntax is a log of all transformations and procedures used. Syntax can easily be generated in SPSS. Using syntax is a matter of personal experience and preference, ranging from only minimal use for data cleaning to use for conducting all analyses.

In most of the interactive menu for a transformation or procedure in SPSS, there is a "Paste" option. This will save the syntax of the command or transformation in the **most recently** opened syntax file (see Figure 4.8).

It is recommended to keep several syntax files: one for data cleaning, one for key complex transformations, and others for the main analytical steps. Keeping the syntax of transformations and key analysis is good practice. That way, if you are ever uncertain about how a variable was created or corrected, you can go back and check.

Syntax can be copied, pasted, and edited in the SPSS Syntax Editor. It is a good practice to write a brief description of the procedure before each syntax command. Separating different parts of the analysis is also considered good practice.

Advantages of using syntax

There are several reasons for using syntax. It can improve efficiency and transparency, and also save time.

A lot of transformations for different variables are similar. A clever use of the Copy/Paste and eventually the Find/Replace function in the Syntax Editor can avoid repetitive chores, save time, and even reduce the chance for error in transformations. Some analytical procedures might take more manipulation than others. This can easily be done by using the interactive menus in SPSS. However, if an analyst wants to change the procedure after the analysis, he/she will have to rebuild the entire procedure, reset all options, and reselect all the variables. Clearly it takes a lot of unnecessary extra time before the same procedure with the same results can be reproduced. If the syntax is saved, however, no time is wasted at all.

At a later date, the analyst or a colleague might want to review the transformations performed or the exact way a procedure was conducted. Use of syntax will enable a review of the draft analysis and make necessary corrections before finalizing the report. It is not uncommon to identify inconsistencies in the data or problems in previous procedures. Moreover, others should be able to see exactly how new variables were defined. Using syntax will give instant access to the formula used to create the new variables. It is recommended to keep a complete syntax of key transformations and procedures.

Data backup

Frequently save backups of the data set. An erred transformation of a variable or a manipulation of the data file can result in lost data. Each successive version of a database should be independently saved. This allows the retrieval of original data if such a mistake occurs.

Labelling variables and values

The cleaned database should have complete variables and values recorded. When creating new variables, be sure to enter suitable variable labels and values of categorical variables. This will enable another analyst to easily and quickly interpret the variables in the data set. As described in section 4.4.2.3, it is recommended that the variable names reflect the questionnaire number, and the variable label provide a more detailed and widely understood name. Calculated variables should have an appropriate name, and a label that clearly identifies the variable to future analysts.

4.4.3.2 Types of variables

The different types of data collected in a household survey, including age, sex, income, assets, and names of districts/provinces, can be categorized according to their measurement scale. Four measurement scales are generally used in statistics: nominal, ordinal, interval, and ratio. Nominal and ordinal variables are considered to be **categorical variables**, while interval data and ratio variables are considered **continuous variables.**

Categorical variables

A categorical variable is one for which each response can be put into a specific category. The categories are usually labelled and coded. These categories must be

both mutually exclusive and exhaustive. Mutually exclusive means that each possible survey response should belong to only one category, whereas, exhaustive requires that the categories cover the entire set of possibilities. If the age categories are 0–6, 7–12, 13–18, they are mutually exclusive, as a person can never be in two of these categories at the same time. To be exhaustive, we have to add a category ("19 or more") so that all possible cases are covered. Categorical variables can be either nominal or ordinal.

Nominal: A nominal variable describes a name or category. Contrary to ordinal variables, there is no "natural ordering" of the set of possible names or categories. Sex, household status, and type of dwelling are examples of nominal variables. Another example is type of crop, which could be categorized as 1 = wheat, 2 = rice, 3 = maize, 4 = sorghum, 5 = millet, 6 = other. Nominal variables cannot be analysed using means; the mode can be used.

Ordinal: Ordinal variables order (or rank) data in terms of degree. Ordinal variables do not establish a numeric difference between data points. They indicate only that one data point is ranked higher or lower than another (Shawna, J. *et al.* 2005). They are not customarily analysed using means. The variable "food consumption group," for example, is ordinal because the category "acceptable food consumption" could be considered better than the category "poor food consumption." There is some natural ordering, but it is limited since we do not know by how much "acceptable food consumption" is better than "poor food consumption."

Example: Variable: Food consumption group, where:

1 = poor food consumption
2 = borderline food consumption
3 = acceptable food consumption

Binary: A special type of categorical variable is a dichotomous/binary variable, which is a nominal variable consisting of only two categories (or levels). Observations can be classed into two groups: male/female, group 1/group 2, true/false, yes/no. All cases having a certain characteristic. For example, "household has female head" could be coded with a value 1, for "yes", while all other cases without that characteristic could be coded with a value 0, for "no". Coding 1/0 allows calculations, which are normally not possible with a nominal variable. For example, the percentage of cases having the given characteristic (e.g. percentage of female household heads) corresponds with the average of the variable.

Continuous variables

A variable is said to be continuous if it can take an infinite number of real values. Continuous variables are interval or ratio variables.

Interval variables

The numbers assigned to objects have all the features of ordinal measurements, and in addition, equal differences between measurements represent equivalent intervals. That is, differences between arbitrary pairs of measurements can be meaningfully

compared. Operations such as addition and subtraction are therefore meaningful. The zero point on the scale is arbitrary; negative values can be used. However, ratios between numbers on the scale are not meaningful, so operations such as multiplication and division cannot be carried out directly. For instance, the phrase "today it is 1.2 times hotter in degrees Celsius than it was yesterday" is not very useful or meaningful; in degrees Fahrenheit it might be 1.4 times hotter. Stating that the birth year of person A is 5 percent higher than the birth year of person B is also not useful or meaningful.

The central tendency of a variable measured at the interval level can be represented by its mean, median, or mode, with the mean giving the most information. Variables measured at the interval level are called interval variables, or sometimes scaled variables, though the latter usage is not obvious and is not recommended. Examples of interval measures include temperature in Celsius scale or Fahrenheit scale.

Ratio variables

A ratio variable, has all the properties of an interval variable, but also a clear definition of 0.0. When the variable equals 0.0, there exists none of that variable. Variables such as height and weight are ratio variables. Operations such as multiplication and division are therefore meaningful. The zero value on a ratio scale is non-arbitrary. The central tendency of a variable measured at the ratio level can be represented by its mode, its median, its arithmetic mean, or its geometric mean, as with an interval scale.

Table 4.22: Summary of statistical measures by type of variable

Okay to compute	Nominal	Ordinal	Interval	Ratio
Frequency distribution	Yes	Yes	Yes	Yes
Median and percentiles	No	Yes	Yes	Yes
Add or subtract	No	No	Yes	Yes
Mean, standard deviation, standard error of the mean	No	No	Yes	Yes
Ratio, or coefficient of the variation	No	No	No	Yes

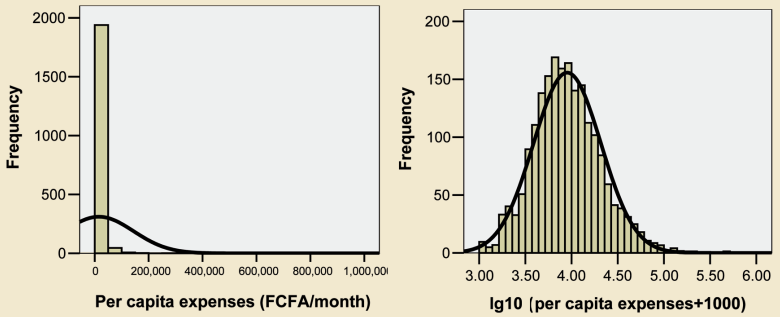
A continuous variable can be categorized into a categorical variable to facilitate months: 24–35 months, 36–47 months, and 48–59 months. The actual age of the children collected in a survey can be categorized into these groups by giving a code such as 0–5 months = 1, 6–11 months = 2, and so on.

Distributions of continuous variables

The distribution of continuous variables should be considered during analysis. Most common procedures and statistics assume that variables are normally distributed. When the distribution of a continuous variable is highly non-normal, alternative data summaries should be used (e.g. median and not mean), and tests of significance that do not assume normality should be used (non-parametric tests). Sometimes, variables

can also be transformed to have a more normal distribution. This transformed variable can then be used, for instance in a regression or a principal components analysis (PCA).

Figure 4.9: Distribution of per capita expenditure



Expenditures, revenue, and production typically have a skewed distribution, as in this example from the Cameroon CFSVA. The mean is 15,500 FCFA/month per capita, however, a few households have monthly per capita expenditures reaching 500,000 to 1,000,000 FCFA/month.

Using the per capita expenditures in a regression model or correlation, or running a t-test or ANOVA to compare means, would violate the assumption of normal distribution. Hence a logarithmic transformation was applied, and the resulting distribution resembles the normal. The transformed variable can now be used in correlations, regressions, PCA, etc.

Constructing indicators: Transformations

Simple or complex transformations are needed to create many of the key CFSVA indicators. Mathematical operations in the compute procedure, categorizing of values in the recode procedure, and other transformations possible in SPSS allow for the combining of different variables or the reconfiguration of variables into desired indicators.

Changing values of a variable

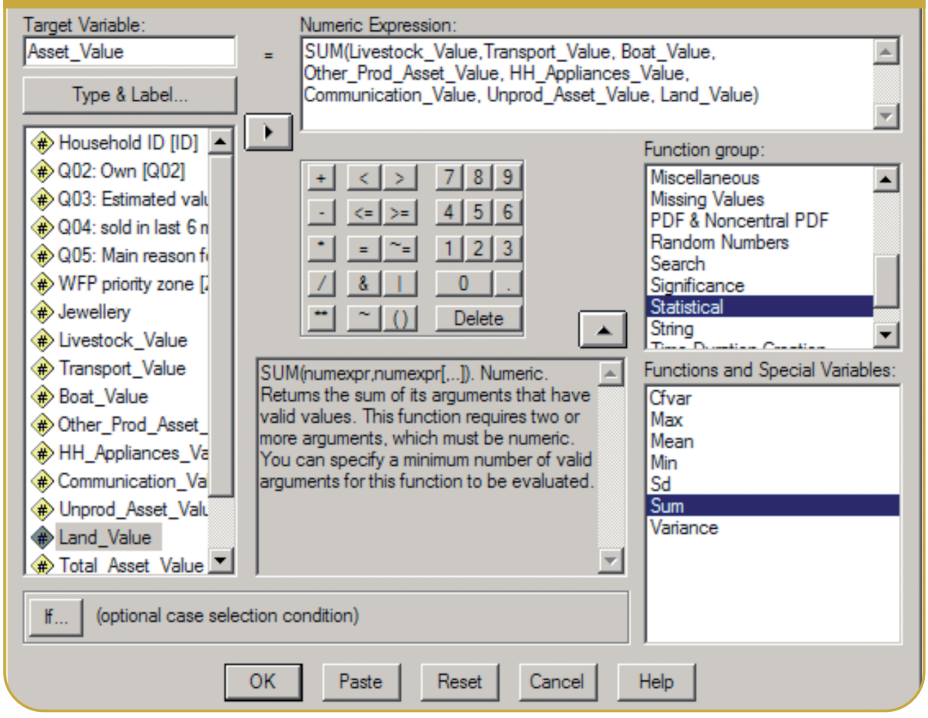
The "Recode" command in SPSS is generally used to change the values in a variable. Variables can be recoded into the same variables or into different variables. Generally, it is recommended that recoding be done into a different variable so that the original data is not lost, particularly in the case of an error in recoding. This command can be used to recode one categorical variable into a new categorical variable, or one continuous variable into a new categorical variable.

It is worth aggregating categories or transforming a continuous variable into a categorical when the list of possible answers is very long (e.g. the relationship of the HH members with the HH head), and/or when some answers have been chosen by few households.

Calculating new variables

New variables can be calculated using the "Compute" command in SPSS. To help compute new variables, SPSS has a number of mathematical operations, including addition, subtraction, multiplication, and division. However, the "Addition" command does not work if the variables added contain missing values. Using the "Sum" command (Figure 4.10) addresses the problem.

Figure 4.10: Computing a new variable using "sum" command



If the analyst adds variables that contain missing values, it is recommended that he/she not use the "Addition" command (+). If it is used, the sum will have a missing value every time there is a missing value in one of the added variables. The command "Sum" would treat all the missing values as if they were "zero," thus not increasing the number of missing values in the variable "Sum."

"Logical operators" can be used to set up conditions ("If" command in SPSS) to create a new variable. Box 4.13 includes a list of the logical operators available in SPSS.

With careful use of these operators, new variables can be constructed for CFSVA analysis. For example, CFSVAs often collect age of children, which is a continuous variable. However, analysts typically create age categories to generate cross tables with other variables like enrolment or drop out.

Box 4.13: Commonly used logical operators in SPSS

<	less than
>	greater than
<=	less than or equal to
>=	greater than or equal to
=	equal to
~=	not equal to
|	or
and	logical "and"

Box 4.14 demonstrates the use of logical operators to create age groups from a variable called "childage."

Box 4.14: Use of logical operators in SPSS

```
If (childage <= 7) agegroup = 1
If (childage >7 and childage <= 12) agegroup = 2
If (childage >12) agegroup = 3
```

However, be careful of missing value codes. For example, if "no answer" is coded as 99, this could be miscategorized as 3, meaning a child more than 12 years old.

The compute command can also use more advanced mathematical functions. For instance, the square root or the logarithmic transformation could be used to normalize a skewed distribution, and TRUNC or RND can be used to categorize continuous variables.

Box 4.15: Some useful mathematical functions in SPSS

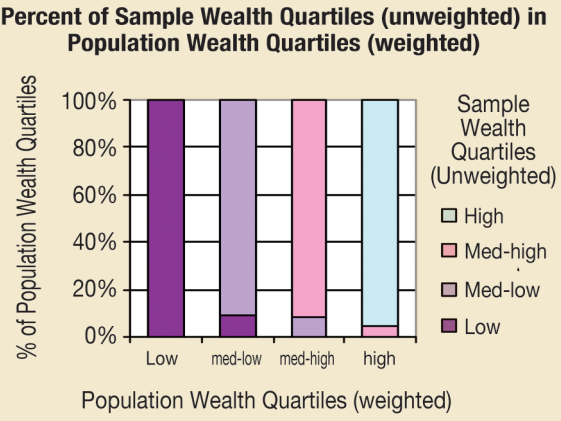
ABS(var)	the absolute value of a variable ABS 13.8=13.8; abs(-13.8)=13.8
RND(var)	the rounded value of a variable: RND(13.8)=14
TRUNC(var).	the truncated value of the variable: TRUNC(13.8)=13
SQRT(var)	the square root of the variable
lg10(var)	the base 10 logarithm of a variable
ln(var)	the natural logarithm of a variable
exp(var)	e raised to the power of the variable

Calculating n-tiles

N-tiles (usually quintiles) can be calculated automatically using SPSS. Under "Transform," use "Rank cases." Under "Rank types," select "n-tiles," and indicate the number of tiles desired, then continue. To deal with tied ranks, usually the mean is used (under "Ties," select "Mean"), although there may be circumstances where other methods should be used.

Most CFSVA data sets use probability weights, hence n-tiles should be calculated with weights on, so that (in the case of quintiles) 20 percent of the weighted sample lies in each quintile, allowing the quintiles to be applied to the population they are representing.

Figure 4.11: Example of incorrect classification of HHs into wealth quartiles



Based on various assets, a wealth index was calculated in the Cameroon CFSVA. When creating quartiles of this index, using the sampling weights, the results can be extrapolated to 25 percent of the population being classified in each quartile. If, however, the weights are turned off when creating the quartiles, 25 percent of the sample household are classified in each "sample wealth quartile," which does not coincide with population quartiles.

The example from the Cameroon CFSVA shows that if the weights are not turned on, 6 percent of rural households are classified in the wrong category.

COMPUTING RATIOS

At household level

Ratios are simply calculated using the "compute" function. Particular care must be taken with 0 values (division by 0 generates a missing value), and that missing values (such as 99,888) are coded and recorded as "missing" in the variable view.

At aggregate level

Although most ratios can be calculated at the household level, it makes more sense for some to be computed at the aggregate level. The enrolment rate at the household level, for example, is often 1 or 0 or missing (when there are no school-age children), whereas the enrolment rate of an entire subgroup is more meaningful. Remember, a statement for a ratio calculated at the household level is different from the same statement for a ratio calculated at the aggregate level, and should be reported as such.

Box 4.16: Example of aggregate ratio vs. household ratio

Calculating dependency ratio in both ways:
Household A has 4 children and 2 productive adults. The dependency ratio is 4 to 2, or 2.
Household B has 1 child, 1 elderly person, and 8 productive adults. The dependency ratio is 2 to 8, or 0.25.
The AVERAGE household dependency ratio = ((2 + 0.25)/2) = 1.125
The AGGREGATE dependency ratio = (sum of all children and elderly/sum of all adults) = (6/10) = 0.600

In order to calculate a ratio at the aggregate level, the values of the variables for the denominator and the variable for the numerator are first aggregated. For example, if calculating school attendance, the sum of all the children attending school in the sample needs to be calculated, and then the sum of all the school-age children needs to be calculated. Then the ratio for that level of aggregation is calculated, using the aggregates in the denominator and numerator. Alternatively, the average for the sample of the numerator and the denominator can be calculated independently, and then these averages are divided to achieve the aggregate ratio. The confidence interval of certain rates can be wide, especially for subgroups. It is therefore important to estimate the error correctly.

Calculation of anthropometric indicators (z-scores)

Epi Info, from the CDC, and Anthro, from WHO, are the two most commonly used applications for anthropometric data analysis. The current standard recommendation for CFSVAs is to calculate and report under-5 anthropometry (stunting, wasting, underweight) using both the NCHS and the new WHO references. Analysis of nutrition should then use the WHO reference data. Only the software WHO Anthro 2005 can currently calculate z-scores using both reference scores.

4.4.3.3 Descriptive statistics

Descriptive statistics are those used to describe characteristics of a sample or population, and they involve exploring the distribution of one variable (frequency) or the distributions between two or more variables (cross tabs). In SPSS, descriptive statistics can be most easily produced by the Frequencies command. Together with simple graphics, they form the basis for virtually every quantitative analysis of data.

Means

a) Simple mean of continuous variables

The (arithmetic) mean is the sum of all the values divided by the number of cases (considering only valid cases and excluding missing cases). It can be used for continuous data (that is, data measured on an interval or ratio scale). The mean is a measure of the variable's central tendency. Statistics such as mean (and standard deviation, defined further on) assume a normal distribution and are appropriate for quantitative variables with symmetric, or normal, distributions. Typical means calculated in CFSVAs include number of household members and monthly income.

b) Value codes of binary variables

While defining the coding of variables, it is recommended to use 1 for "yes" or "present," and 0 for "no" or "not present." In a household survey, the population mean of a variable coded in this way corresponds with the proportion of households that answered "yes" or where the reply was "present," and when multiplied by 100, gives the percentage prevalence of "yes."

Similarly, one could agree to specify 0 for "male" and 1 for "female." However, for the sex of children, to be used in anthropometric data transformations,[66] "male" should always be male = 1 and female = 2. In this case, means cannot be used easily to calculate proportions and percentages.

Medians

The median is the middle of a distribution when the values are ranked from highest to lowest, meaning half of the values are above the median and half are below the median, the fiftieth percentile, i.e. the middle value of a set of observations ranked in order.

66. This is the convention for Epi Info from the CDC and Anthro from WHO.

If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not sensitive to outlying values, unlike the mean, which can be affected by a few extremely high or low values.

The median is a robust statistic, appropriate for quantitative variables that may or may not meet the assumption of normality. It is preferable to use the mean when the data are not normally distributed. For example, some expenditure and income data have very skewed distributions, and the median may be a better summary than the mean. The median can be used with measurement scales that are at least ordinal (that is ordinal, interval, or ratio).

Modes

Modes are rarely used in describing CFSVA indicators. However, they may occasionally be of use. The mode is defined as the most frequent variable value. In the set of values 1, 2, 3, 3, 3, 4, 5, 5, 6, 6, 7, 8, 9, the mode is 3, as it appears more frequently than any other value. There can be more than one mode. Using the mode is more appropriate when there are only a few possible values of the variable (for instance, to describe the household size, the mode could be used). The same is true for categorical variables (for example, "farming" is the most common livelihood strategy of the households.

PERCENTAGES/PROPORTIONS

Frequencies

Frequencies is one of the more common descriptive functions used in the analysis of CFSVAs. They are most commonly used to produce global prevalence (prevalence for the whole data set). SPSS gives two prevalence results: percentage and valid percentage. Percentage includes the missing cases (system missing and those coded as missing) in the denominator, whereas valid percentage excludes the missing cases and includes only the cases with data in the denominator. The analyst needs to interpret which of these to report. If the missing values are assumed to be not different from the values with valid responses, the "valid percentage" statistic should be given; if the missing values are different (nonexistent options, not applicable), the analyst should consider reporting the "percentage" as compared to the total population. Valid percentage is the most common; however, in certain cases percentage may be more relevant.

Box 4.17: Example highlighting the difference in interpretation

In this example from Laos, households were first asked if they cultivated land in the last year. If they responded "no," then the enumerators skipped the rest of the agriculture section, leaving the questions blank. If the response was "yes," then the following question was asked: "What was your main crop cultivated in the past year?" In the data entry, this question on the main crop was left blank (system missing) if there was no response. A frequency of the main crop cultivated results in the following table:

(cont...)

(...cont)

Main crop cultivated

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Glutinous rice	2827	72.0	82.1	82.1
	White rice	365	9.3	10.6	92.7
	Maize	65	1.7	1.9	94.6
	Beans	4	0.1	0.1	94.7
	Cassava	3	0.1	0.1	94.8
	Vegetables	11	0.3	0.3	95.1
	Fruits	19	0.5	0.5	95.6
	Tobacco	1	0.0	0.0	95.7
	Groundnuts and other nuts/seeds	6	0.2	0.2	95.8
	Other	143	3.7	4.2	100.0
	Total	3444	87.7	100.0	
Missing	System	482	12.3		
Total		3926	100.0		

When looking at glutinous rice, two statements can be made. Using the percentage, it can be stated "72 percent of all households cultivated glutinous rice as their main crop in the past year." This statement is of ALL households. Using valid percentage, it can be stated "of the households practicing agriculture in the last year, 82 percent cultivated glutinous rice as their main crop."

Cross tabulations

Cross tabulations are another way of exploring frequencies, and are one of the most common descriptive tools used in CFSVA analysis.

Unlike frequencies, cross tabs include only the valid cases. When calculating prevalence in SPSS, three options are commonly used: percentage rows, percentage columns, and percentage total. This will determine how SPSS calculates the prevalence in each cell. For each cell, the numerator remains constant: the number of valid cases belonging to the two groups. The denominator in percentage rows is the total number of cases in the row cells. The denominator in percentage columns is the total number of cases in the column cells. The denominator in total percentage is the total number of cases in ALL cells, which is equal to the valid number of cases in the data set.

Understanding this difference between percentage rows, columns, and choosing the correct one to report is critical. As a general rule, the "independent variable" should be put in the columns and the "dependent" in the rows; column percentages are more important than row percentages. For instance, we can look at the influence of wealth on food consumption. If we put wealth in the column, we should focus on the column percentages and compare them across the rows (food groups).

In Table 4.23, three food consumption groups were cross tabulated with the quintiles of wealth score (the first being the poorest, the fifth being the richest).

Table 4.23: Example of cross tabulation

			Quintiles of wealth score					
			1	2	3	4	5	Total
Food consumption groups	Poor Food Consumption	% within Food consumption groups	67.1%	13.4%	8.5%	3.7%	7.3%	100.0%
		% within Quintiles of wealth score	7.1%	1.4%	0.9%	0.4%	0.8%	2.1%
		% of Total	1.4%	0.3%	0.2%	0.1%	0.2%	2.1%
	Borderline Food Consumption	% within Food consumption groups	38.4%	26.8%	17.8%	11.9%	5.1%	100.0%
		% within Quintiles of wealth score	20.3%	14.2%	9.3%	6.3%	2.7%	10.6%
		% of Total	4.1%	2.8%	1.9%	1.3%	0.5%	10.6%
	Acceptable Food Consumption	% within Food consumption groups	16.6%	19.2%	20.6%	21.4%	22.2%	100.0%
		% within Quintiles of wealth score	72.6%	84.4%	89.8%	93.3%	96.5%	87.3%
		% of Total	14.5%	16.8%	18.0%	18.7%	19.4%	87.3%
Total		% within Food consumption groups	20.0%	19.9%	20.1%	20.0%	20.0%	100.0%
		% within Quintiles of wealth score	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	20.0%	19.9%	20.1%	20.0%	20.0%	100.0%

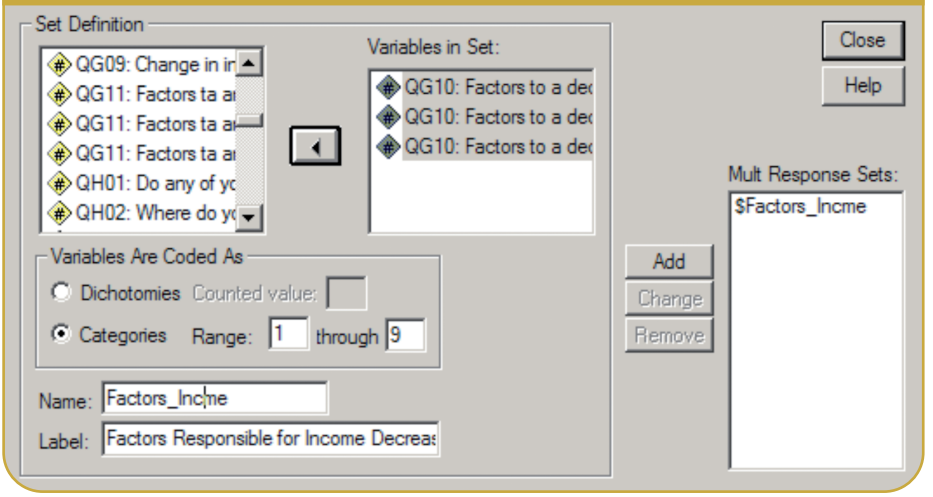
Table 4.23 shows the relationship between the food consumption of households and wealth status. It shows that food consumption increases stepwise by wealth quintile. More than 67 percent of households in the first wealth quintile reported poor food consumption, while 22.2 percent of households in the fifth wealth quintile and 21.4 percent of households in the fourth wealth quintile reported good food consumption.

4.4.3.4 Analysing multiple responses

A number of questions in CFSVAs call for multiple responses from the respondents. For example, a typical CFSVA questionnaire includes questions about major crops cultivated, main income sources, and food sources. All these questions generate multiple answers. A household might have obtained its food from different sources, including food bought from the market, produced, received as food for work, and borrowed from neighbours. It is a common practice to create separate variables for each of these answers, and these answers are mutually exclusive

The multiple response feature in SPSS allows one to analyse variables taking multiple responses into account. The first step is to define variable sets. All the variables containing (multiple) responses should be inserted into the “Variable in Set.” The next step is to select the type of variables included in the variable set and their range of values. After giving the name of the “Multiple Response Set,” add the variable set to the Multiple Response Set. Now the multi-response set is ready for analysis. To analyse, go to “Multiple Response,” then “Analyse,” and select the desired analysis.

Figure 4.12: Define multiple response sets



It is a good practice to paste the command to a syntax file for future reference. This will also help the analyst if she/he wants to regenerate the table.

Box 4.18 presents syntax to compute percentages. It essentially computes percentages of responses and percentages of cases.

Box 4.18: Syntax to compute percentages

```
mult response
groups=$Income_decline_factors 'Factors responsible for income decrease' (QG10_V1
QG10_V2 QG10_V3 (1,9))
/frequencies=$Income_decline_factors.
```

The output of the analysis is presented in Table 4.24. The column that presents percentage of responses calculates percentage of total responses. For example, in this case 623 households responded to the question "What are the factors that led to a decrease in your household income?" A household could answer 3 different responses from a list of 9. Sixty-five households identified 3 factors, 194 households identified 2 factors, and 364 households identified only 1 factor responsible for income decrease.

Altogether, 623 households responded to this question with 947 responses. The percentage of responses column calculates percentage of responses and the percentage of cases column calculates the percentage of households that responded to this question (623 in this example).

Table 4.24: Output of the analysis

		Responses		Total
		N	Percent	
Factors responsible for income decrease	Loss of employment	124	13.1%	19.9%
	Loss of crop/animal	65	6.9%	10.4%
	Prolonged illness of income earner	152	16.1%	24.4%
	Death of income earner	48	5.1%	7.7%
	Decrease in remittance income	8	0.8%	1.3%
	Loss of asset	147	15.5%	23.6%
	Exposure to natural disaster	124	13.1%	19.9%
	Market failure	157	16.6%	25.2%
	Other	122	12.9%	19.6%
Total		**947**	**100.0%**	**152.0%**

4.4.3.5 Measures of variation

Variance and standard deviation

Variance is a measure of dispersion around the mean of a continuous variable, equal to the sum of squared deviations from the mean divided by one less than the number of cases (degrees of freedom). The variance is, therefore, the average squared distance between the mean and the observations made (and so is a measure of how well the model fits the actual data). However, the variance is measured in units that are the square of those of the variable itself. The **standard deviation** is obtained by calculating the square root of the variance.

The **standard deviation** of a distribution, a measure of dispersion based on a deviation from the mean (which are squared, summed, and averaged and then the square root is taken), has the same unit as the original observations and can be used for data measured on an interval or ratio scale. A large standard deviation (relative to the mean; also called coefficient of variation) indicates that the data points are distant from the mean. In this case, the mean may not be an accurate representation of the data. A standard deviation of 0 would mean that all the scores were the same.

In a normal distribution, 68.27 percent of cases fall within one standard deviation on either side of the mean, 95.45 percent of cases fall within two standard deviations, and 99.73 percent fall within three standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95 percent of the cases would be between 25 and 65 in a normal distribution.

Confidence intervals

Confidence intervals enable analysts to make statements on the precision of their estimates. For example, if 30 percent of households were observed in the representative sample to be female headed, confidence intervals could be added to state that "the analysts are 95 percent sure that between 25 and 37 percent of households (in their sampling universe) are female headed." Confidence intervals should always be used in CFSVAs when reporting highly standardized indicators such as stunting, wasting,

underweight, low BMI, and low MUAC. Confidence intervals can be used in CFSVAs when reporting key indicators such as percentage of food insecure, percentage of poor food consumption. Confidence intervals are not typically reported for descriptive indicators in the text of a CFSVA; however, one should strive to include them in annex tables.[67]

4.4.3.6 Tests of significance

Significance is a statistical term that indicates how sure the researcher is that a difference or relationship exists. Tests of significance help the researcher and the audience know if differences between groups are real or by chance. When a statistic is significant, it simply means that one can be very sure that it is reliable and can be referred to the entire population. It does not mean the finding is important or that it has any decision-making utility. This significance, produced by the statistical tests discussed here, is referred to as the p-value (probability value).

The p-value can be interpreted as the probability of a difference occurring by chance alone. If all other biases are eliminated or accounted for, then one can assume that when this p-value is small, the differences are due to a factor other than chance. The cut-off for significance most often used is 0.05. If a p-value is less than 0.05, then assume that the relationship observed is real not by chance. Usually p-values are reported by their actual value to three decimal places, or as >0.05, <0.05, <0.01, or <0.001. Significance levels, when appropriate, are usually reported in the body of the report and in the annex tables.

In this section, some of the more commonly used statistical tests are presented. However, there is a wealth of further statistical tests, many available in SPSS. For a more complete guide to tests of significance, see Discovering Statistics Using SPSS,[68] or any other statistical manual or textbook.

It is very important to note that CFSVAs employ cluster sampling methods, which require special analytical approaches to calculating significance levels and confidence intervals. Standard packages such as the basic SPSS package, do not compute accurate p-values for surveys that are sampled using a cluster design. The appropriate statistical analyses can be obtained using the SPSS Complex Samples module or other special software.

Table 4.25 provides some guidance on what tests of significance to use when comparing different types of data. Keep in mind that this is a generic list and should be used only as a guide. It is the analyst who should decide which test is appropriate for what analysis based on a number of factors.

67. Automatic applications exist for computing confidence intervals for percentages. It is good practice to report them, at least for key indicators.
68. A. Field, 2005, *Discovering Statistics Using SPSS*, 2nd ed., London: SAGE Publications Ltd.

Table 4.25: Example of test of significance for different types of variables

	Dependent variable	Independent variable	When to use	Example	Procedure
Independent T-test	Continuous	Categorical binomial	To compare differences in the means of two groups (identified by the categories of the binomial variable) To see if the difference is statistically significant ($p<0.05$)	Compare the mean z-scores of male and female children	Run the independent samples T-test; Report the two means; Check if the T value is statistically significant ($p<0.05$)
One-way ANOVA: Post-hoc Multiple Comparisons	Continuous	Categorical	To compare differences in the means of three or more groups (identified by the categories of the categorical variable)	Compare the mean z-score by residence status (IDP, refugee, or resident HHs)	Run the One-Way ANOVA post-hoc procedure Check if the categorical variable explains in a significant way some of the observed variation through the F-test. Check which differences are statistically significant ($p<0.05$) through the post-hoc tests (e.g., REGWQ, Tukey HSD, Games-Howell, etc.)
Chi-square	Categorical	Categorical	To detect whether there is a statistically significant association between two categorical variables	Explore the association between food consumption groups and ethnic groups	Compute the Chi-square and report the value Check if the value is statistically significant ($p<0.05$) (The Chi-square helps determine whether the association is statistically significant)
Bivariate Correlation	Continuous	Continuous	To assess the general association between two variables (i.e.,one variable increases/decreases when another increases/decreases)	Correlation between children's height and weight	Compute the Pearson Correlation Coefficient and report the value Check if the correlation is statistically significant (two tailed tests) ($p<0.05$)
Simple Linear Regression	Continuous	Continuous/ Categorical binomial (0/1 values)	To measure how the dependent variable changes with a one-unit increase in the independent variable	Regressing food consumption score by wealth index	Run Simple Linear Regression Model Report R^2 adjusted, B value Check and report if B is statistically significant ($p<0.05$)
Multiple Linear Regression	Continuous	Two or more continuous/ categorical binomial (0/1 values)	To measure how the dependent variable changes with a one-unit increase in the independent variable (controlling by the other variables in the model)	Regressing food consumption score by wealth index and gender of the HH head	Run Multiple Linear Regression Model Report R^2 adjusted, B values Check and report if B values are statistically significant ($p<0.05$)
Multivariate General Linear Model (GLM)	Continuous	2 or more continuous variable and/or 2 or more categorical variables	GLM combines ANOVA and Regression to analyse the effects of more than one independent variable on the dependent variable (and to see how these independent variables interact)	Analyse the effects of ethnic group, province, and wealth index on the food consumption score	Run a Multivariate GLM Interpret the output from main ANOVA table Report R^2 adjusted, B values
Logistic Regression	Categorical	Two or more continuous variables and/or two or more categorical variables	To predict the probability of an event occurring for a given household	Predict which households are more likely to be food insecure according to the province of residence and WI	Run a Logistic Regression Check the overall fit of the model Check which variables significantly predict the outcome (in SPSS, check table "Variables in the equation")

In a typical CFSVA, the most commonly used tests of significance include the Chi-square test, z-test, t-test, and the ANOVA. For further information on tests of significance, consult a statistics manual.

4.4.3.7 Multivariate analysis

Multivariate analysis in statistics describes a collection of procedures involving analysis of more than one statistical variable at a time. In design and analysis, these techniques are used to perform studies across multiple dimensions while taking into account the effects of all variables.

Regression

Regression analysis is a technique used for the modelling and analysis of numerical data consisting of values of a dependent variable (response variable) and of one or more independent variables (explanatory variables). The dependent variable in the regression equation is modelled as a function of the independent variables, corresponding parameters ("constants"), and an error term. The error term is treated as a random variable. It represents unexplained variation in the dependent variable. The parameters are estimated so as to give a "best fit" of the data. Most commonly the best fit is evaluated by using the "least squares method," but other criteria have also been used.

Data modelling can be used without there being any knowledge about the underlying processes that have generated the data; see http://en.wikipedia.org/wiki/Regression_analysis—cite_note-Berk-0#citenote-Berk-0; in this case the model is an empirical model. Moreover, in modelling, knowledge of the probability distribution of the errors is not required. Regression analysis requires assumptions to be made regarding probability distribution of the errors. Statistical tests are made on the basis of these assumptions. In regression analysis, the term model embraces both the function used to model the data and the assumptions concerning probability distributions.

Regression can be used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modelling of causal relationships. These uses of regression rely heavily on the underlying assumptions being satisfied.

Regression analysis is complex, and therefore cannot be adequately covered in these guidelines. However, a few key concepts and guiding principles are presented here. CFSVAs often, but not always, use regression analysis. Nutritional analysis makes frequent use of regression techniques.

a) When is regression used for CFSVAs, and why?

In CFSVA, regression is primarily used to understand causal relationships between the variables that are important for decision-making purposes. For example, to explore the relationships between stunting (height-for-age, an indicator of chronic malnutrition) and dietary quality while controlling for sanitation facilities, access to potable water, mother education, and household income, the analyst may want to set up an OLS (Ordinary Least Square) model, where height-for-age is a dependent variable and all of the other variables mentioned could be explanatory variables.

However, before estimation of the model, the analyst has to test for:

- **multicollinearity:** one or a combination of explanatory variables is strongly correlated to another explanatory variable;
- **heteroskedasticity:** the variance of the error terms changes across observations;
- **specification error:** the model is wrong, by missing important explanatory variables or by having other incorrect assumptions; and
- **endogeneity:** when an explanatory variable is itself a function of the dependent variable.

Necessary correctional measures need to be taken if the tests identify multicollinear variables, heteroskedasticity or omitted variables. If any of the explanatory variables is found to be endogenous, the variable has to be replaced by suitable instrumental variables. It is important to understand that a simple regression involving only the dependant variable and one independent variable is similar to a Pearson correlation.

b) Why control for other factors – confounders?

Confounders refer to factors that relate both to the dependent (outcome) and independent variable of interest. For example, we can hypothesize that children under 5 tend to be more underweight (low weight for age z-score) in female-headed households. We can run a simple compare means, or a t-test, to explore the differences in mean z-score between male- and female-headed households. However, the critical question is whether the sex of the household is the only factor responsible for the nutritional status of the children in the household? In a regression analysis, one could enter both education level and sex of the head of the household. It may be found in this analysis that the head of household no longer has a significant effect on underweight z-score, but that education level does. This could likely arise because female heads of household, in this example, often have lower education levels than male heads of households. The regression analysis controls for the difference in sex when it estimates the effect of education, and vice versa. Hence we can say "controlling for education, sex of household head is not significantly related to underweight z-score."

Female-headed households is still important as an identifier of vulnerable households for targeting, since gender may be more easily identified than education level, even if we know that the real reason is that most female household heads have a low education level.

c) Why explore interactions?

Interactions illuminate how a cause of food insecurity might be modified by another variable. For example, if we look at sanitation, water source, and underweight status of children under 5, we might find that improving sanitation has no effect unless in the presence of a safe source of water. In the regression, the two variables (sanitation and water) are simply multiplied together to give the interaction term. Environmental variables, economic factors, education, and age are common effect modifiers (variables that result in statistical interactions).

d) Coefficient of determination

In linear regression models, R^2 is a statistic that gives some information about the merit or fit of a model. R^2 is the square of the correlation coefficient between the dependent variable and the estimate it produced by the independent variables, or equivalently

defined as the ratio of regression variance to total variance. It is a measure of determination of how well the regression line approximates the real data points. An R-squared of 1.0 indicates that the regression line perfectly fits the data, and 0.0 indicates that one term does not help to know the other term at all. For regression applied to household surveys, it is normal to find an R^2 between 0.15 and 0.25, or (exceptionally) a little higher.

Principal component analysis and cluster analysis[69]

This section discusses the following two multivariate analysis techniques:

- Principal component analysis (PCA) – which belongs to the factor analysis family; and
- Cluster analysis – which belongs to the classification family.

Both techniques can typically be used to reduce the complexity of the data set for exploratory purposes: factor analysis, to reduce the selected variables into a lesser number of factors; and cluster analysis, to group all cases into a number of groups. Detailed explanations of how PCA and cluster analysis techniques work are beyond the scope of these guidelines; interested readers should refer to specialized textbooks for more information. This section presents a simplified summary of the statistical ideas on which PCA and cluster analyses are currently used in CFSVA and FSMS, as well as of the terminology used throughout those analyses.

In addition to SPSS, several commercial statistical software packages perform both PCA and cluster analysis. With the support of WFP and FAO, VAM typically also uses a software developed explicitly for socio-economic and food security analysis (ADDATI, or the brand new Windows version ADDAWIN). This software was designed for the use of food security specialists. It includes preselected algorithms proven to be suitable when analysing socio-economic and nutrition data for food security and vulnerability analyses. It uses the output of the PCA as the input for the cluster analysis and facilitates the final interpretation of the outputs providing the cluster results in terms of the original input variables.

ADDATI/ADDAWIN cannot perform factor analysis with rotation. For this type of multivariate analysis, the software normally used in VAM is SPSS.

Principal component analysis

Factor analysis is used to study the patterns of relationships among many dependent variables, with the goal of discovering the underlying variations that affect them. The inferred underlying variables are called factors. Principal component analysis (PCA) uses a factor extraction method to form uncorrelated linear combinations of the observed variables. The first component explains maximum variance. Successive components explain progressively smaller portions of the variance and are all uncorrelated with each other.

PCA is one technique of multivariate analysis that applies to continuous variables. The objective of PCA is twofold:

69. This chapter strongly benefits from inputs and quotations from: WFP/VAM, *Household Food Security Profiles (Thematic Guidelines)*, April 2005; S. Griguolo, *ADDATI Users' Manual*, July 2003, IUAV; S. Landau and B.S. Everitt, *A Handbook of Statistical Analyses Using SPSS*, 2004; Chapman & Hall/CRC, Andy Field, Discovering Statistics using SPSS, 2005; SAGE ADDATI help; and SPSS help.

- to identify and describe the underlying relationships among the variables by creating new indicators (called "factors" or "principal components") that capture the essence of the associations between variables; and
- to reduce the complexity of the data, saving a limited number of these new variables that is sufficient to keep the most relevant aspects of the description with a minimal loss of detail.

PCA yields as many principal components as there are initial variables. However, the contribution of each principal component to explaining the total variance found among all variables will progressively decrease from the first principal component to the last. As a result, a limited set of principal components explains the majority of the matrix variability, and principal components with little explanatory power can be removed from the analysis. The result is data reduction with relatively little loss of information.

It is recommended to use the rotated solution in most circumstances. Rotation of the result will give a new solution: the new factors explain the same variance and can be much better interpreted as the underlying dimensions. The analyst can understand the "meaning" of each dimension and then decide which of the uncovered dimensions he wants to use in subsequent analyses.

Uses of PCA

Two of the most common uses of PCA in CFSVA analysis are briefly described here:

Scoring on the first principal component (single factor solution)

PCA might be used to build a synthetic composite index moving from more than one variable. In this case, the first principal component is taken as the new variable on which statistical units might be measured and ranked. For example, variables from a household survey that are associated with wealth (quality of housing, assets owned by the household, etc.) can be used to perform a PCA. Since what those variables have in common is related to wealth, the first component can be interpreted as a **wealth index**.

While the single factor solution is very straightforward and easy to interpret, its use should be limited to specific cases. If the variables are uni-dimensional, then an exploratory analysis showing a single factor grouping, where all items "hang together," is supportive. If the variables used in the analysis have multiple facets or dimensions, a single factor solution would not be able to maintain a minimum description for all of the analysed statistical units. In this case, trying to capture the underlying variations of all of the input variables through a unique index would undermine the instrument's construct validity. That is because the use of a small selection of measuring items (i.e. the information maintained by one factor only) could lead to false and confusing results that would not reflect the complexity of the original data.

Using principal component(s) as input variable(s) for follow-up analysis (specifically cluster analysis)

PCA creates a set of new variables or components that, being perfectly uncorrelated, explain different portions of the original total variance.

As one of the main purposes of PCA is to reduce the dimensionality of the data set, components are ranked by their decreasing contribution to explaining the total

variance. It is hence possible to remove components with little explanatory power. If the main purpose of PCA is to describe statistical units on the basis of the relationships among selected variables, data reduction is a secondary objective. Furthermore, if the final aim is to cluster units based on those relationships, it is recommended that analysts keep as many principal components needed to capture up to 80 percent of the total variance. Such a high level of consistency with the original complexity of the data would ensure a good reflection of the relationships among variables. It would also guarantee that particular combinations of variables' values were maintained and not smoothed too much through a high data reduction approach.

When data reduction is the primary objective, the analyst may want to remove more components. One rule of thumb is that the Eigen value of each extracted component should be higher than 1; an alternative is to keep as many factors (after rotation) that still have a readily understandable meaning. For a subsequent clustering, the analyst can even exclude factors irrelevant for the clustering activity to be undertaken.
For example, average decadal rainfall, Normalized Difference Vegetation Index (NDVI), and other climatic data were used to conduct a factor analysis in Sudan. Four underlying, meaningful factors were retained. The third factor, related to rainfall during the dry season, was not used for clustering, since rain during that season has little economic importance.

Cluster analysis

Clustering data is a common technique for statistical data analysis. Clustering is the classification of similar objects into groups or, more precisely, the partition of a data set into subsets (clusters) so that the data in each subset share some common features, often proximity according to some defined distance measure.

Note that cluster analysis is an exploratory data analysis tool that aims at sorting different objects into groups such that the degree of association between two objects is maximal if they belong to the same group, and minimal otherwise.

Clustering is one of the key parts of many large data set analyses. In fact, it is not possible to analyse and describe the situation of each statistical unit (be it a household, a village, a district, a region, etc.) separately, since there might be too many. There is a clear need to identify main patterns of similar characteristics. Clustering involves some kind of subjectivity based on analysts' choices of specific methods for clustering.

In conducting cluster analysis, analysts are often faced with two questions: How many clusters are there in the data set? and What is the compactness (inertia) of each cluster?

How many clusters in the data set?

A given data set does not contain a definitive number of clusters. First, because cluster analysis involves a series of iterations performed by statistical software, there will be some variance in the number of clusters and the assignment of particular households to clusters each time the analysis is performed, depending on the initial "cluster seeds." Second, several different methods and algorithms can be used to produce clusters, and the number of clusters produced will vary depending on the type of clustering method used. For very large data sets, the partition method, with a random selection of the initial centres, seems to be most appropriate. Specific algorithms to

improve the quality of a partition are implemented, being different in different software packages.

What is the compactness of each cluster (inertia)?

The measurement of the dispersion or compactness of each cluster is called **inertia or internal variance of the cluster.** The degree of inertia within and among clusters provides a useful means of determining the final number of clusters that best fits the data.

There are no standard thresholds indicating what level of inertia is good, acceptable, or poor, and the final decision remains with the analyst. However, the ratio between the inertia of the overall cloud (the dispersion found among all units in the dataset) and the inertia within each cluster should be maximized. Doing so ensures that the similarity among units belonging to the same cluster (e.g. within clusters) is high, while the similarity between clusters is very low (e.g. maximizing intra-cluster homogeneity and inter-cluster heterogeneity).

One of the strengths of ADDATI (or ADDAWIN) is that it incorporates a specific formula (objective function) to calculate the intra- and inter-cluster inertia as a measurement of partition optimality with a given number of clusters. In other words, it measures how compact a set of clusters is.

In addition, the clustering in ADDATI displays a graph that plots how the value of the objective function decreases when the number of clusters retained is increased. By inspecting this graph, the user can focus on one or more promising partitions, with a number of clusters within the range he/she would like to obtain and a value of the objective function sufficiently high. This tool indeed helps the analyst find a number of final clusters, which is a fair **trade-off** between the level of synthesis achievable (few clusters are always more convenient) and a significant level of homogeneity of characteristics within the clusters (provided by the value of the objective function that represents the rate of information maintained).

One of the common uses of cluster analysis in CFSVA is to create groups of households with similar food consumption patterns, for profiling purposes or for further analysis. Typically the principal components are used to create the clusters.

Cluster analysis is also used to categorize households that share similar livelihood strategies into **livelihood groups**. The aim of CFSVA livelihood grouping is not to replace a comprehensive livelihood analysis but to utilize livelihood strategies as a basis for classifying populations.

4.4.4 Key references: Household data analysis

- Chapman and Hall/CRC Andy Field. 2005. *Discovery Statistics Using SPSS*. Sage ADDATI help; and SPSS help.
- Griguolo, S. 2003. ADDATI User's Manual. IUAV, July.
- High, R. 2000. *Dealing with Outliers: How to Maintain Your Data's Integrity.* Computing News. UO Computing Center. University of Oregon.
- Landau, S., and B. S. Everitt. 2004. *A Handbook of Statistical Analyses Using SPSS*, http://cc.uoregon.edu/cnews/spring2000/outliers.html.
- Shawna, J., K. Marcus, C. McDonald, T. Wehner, and M. Palmquist. 2005. *Introduction to Statistics.* Writing@CSU. Colorado State University Department of English. Retrieved 12/31/2007 from http://writing.colostate.edu/guides/research/stats/.
- United Nations, Department of Economic and Social Affairs, Statistics Division. 2005. *Household Sample Surveys in Developing and Transition Countries.* Studies in Methods, Series F No. 96.