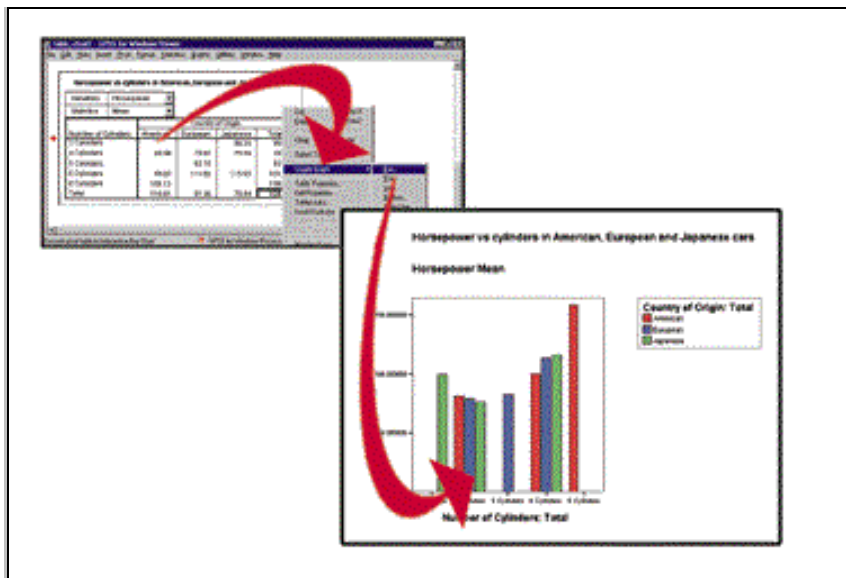


Poverty analysis with SPSS



Course Manual

Compiled by Klaus Röder
– Beira, Mozambique November 2004
– Ver:041112

TABLE OF CONTENTS

| | | |
|----------|--|-----------|
| 1 | INTRODUCTION | 4 |
| 1.1 | CONVENTIONS OF TEXT STYLES | 5 |
| 1.2 | HOW TO CHOOSE OPTIONS FROM MENUS | 5 |
| 2 | A GENERAL VIEW OF SPSS | 6 |
| 2.1 | HOW TO START AND STOP SPSS | 6 |
| 2.2 | SPSS WINDOWS | 6 |
| 2.3 | HOW TO EDIT DATA | 8 |
| 2.4 | EXERCISES | 9 |
| 2.5 | THE FIRST ANALYSIS WITH SPSS: THE ANALYSIS OF FREQUENCY | 10 |
| 2.6 | GENERAL CONSIDERATIONS ABOUT MEASUREMENT LEVELS: | 12 |
| 2.7 | THE VARIABLE DEFINITION IN SPSS | 12 |
| 2.8 | EXERCISES | 13 |
| 2.9 | THE TRANSFORMATION OF VARIABLES | 14 |
| 2.10 | DEFINING VARIABLES | 14 |
| 2.11 | VARIABLE NAMES IN SPSS | 14 |
| 2.12 | VARIABLE TYPES | 14 |
| 2.13 | MISSING VALUES | 15 |
| 2.14 | COMPUTE VARIABLES | 15 |
| 2.15 | RECODE VARIABLES | 16 |
| 2.16 | TRANSPOSE VARIABLES | 17 |
| 2.17 | SORT CASES | 17 |
| 2.18 | SELECT CASES | 17 |
| 2.19 | EXERCISES | 19 |
| 3 | THE TRANSFORMATION OF DATA FILES | 20 |
| 3.1 | MERGING FILES: GENERAL CONSIDERATIONS | 20 |
| 3.2 | MERGE FILES: ADD CASES | 20 |
| 3.3 | MERGE FILES: ADD VARIABLES | 21 |
| 3.4 | THE AGGREGATION OF THE DATA | 22 |
| 3.5 | EXERCISES | 24 |
| 4 | FUNCTIONS FOR THE INITIAL ANALYSIS OF DATA | 25 |
| 4.1 | FREQUENCIES | 25 |
| 4.2 | EXPLORATIVE DATA ANALYSIS | 26 |
| 4.3 | EXERCISES | 28 |
| 5 | THE PRESENTATION OF DATA | 29 |
| 5.1 | LIST CASES | 29 |
| 5.2 | THE TABLES MODULE | 29 |
| 5.2.1 | <i>Basic Tables</i> | 29 |
| 5.2.2 | <i>General Tables</i> | 31 |
| 5.3 | EXERCISES: | 33 |
| 5.4 | GRAPHICS AND CHARTS | 34 |
| 5.4.1 | <i>Graphs from the Graphs Menu</i> | 34 |
| 5.4.2 | <i>Graphs resulting from procedure calls</i> | 35 |
| 5.4.3 | <i>Graphs from the Interactive Graphs Menu</i> | 35 |
| 5.4.4 | <i>Using the Chart Editor</i> | 36 |
| 5.4.5 | <i>Interactive Graphs and the Interactive Chart Editor</i> | 37 |
| 5.5 | EXERCISES: | 39 |

| | | |
|----------|---|-----------|
| 6 | THE POVERTY LINE | 40 |
| 6.1 | THE CONCEPT OF POVERTY AS DEFINED BY THE IAF | 40 |
| 6.1.1 | <i>The poverty headcount index</i> | 41 |
| 6.1.2 | <i>The poverty gap index</i> | 41 |
| 6.1.3 | <i>The squared poverty gap index</i> | 41 |
| 6.2 | EXERCISES:..... | 43 |
| 7 | REGRESSION ANALYSIS | 44 |
| 7.1 | BACKGROUND: | 44 |
| 7.2 | SPSS - LINEAR REGRESSION | 44 |
| 7.3 | AN EXAMPLE WITH SPSS AND THE IAF DATA OF SOFALA | 45 |
| 7.4 | AS CARACTERÍSTICAS DO MODELO DA REGRESSÃO | 46 |
| | <i>The chicken example</i> | 46 |
| | <i>A Graph of the Regression Equation</i> | 47 |
| 7.5 | MODELS WITH TWO OR MORE PREDICTORS | 49 |
| 7.6 | EXERCISES..... | 51 |
| 8 | BIBLIOGRAPHY | 52 |
| 8.1 | SPSS..... | 52 |
| 8.2 | STATISTICS, POVERTY ANALYSIS | 52 |
| 8.3 | WWW | 52 |

1 Introduction



The aim of this course is to introduce the participants briefly to the statistics program SPSS. This program is among the best known and most widely distributed statistics programs. For many years, since the days of the mainframe, where this program was originally developed, it stands for a very efficient symbiosis between computers and statistical analysis.

Since this course deals with poverty statistics, the special emphasis is on the use of SPSS for poverty statistics.

Since the course focuses on some statistical methods for analysis, it might be the side effect of this course for some participants to refresh old theory and gain new insight into some elementary statistical methods. The course addresses the beginner to SPSS as well as the experienced SPSS user, it assumes that the participant is equipped with a basic statistical background. The result should be that the participants are familiar with some basic concepts of SPSS and understands the use of some more sophisticated methods to analyze data for poverty statistics.

The development of computers has been dramatic over the recent years, so the software for analyzing and communication of statistical results has been developed as well. Although the statistical theory behind the software has not changed, if the focus remains on the less exotic and sophisticated methodology, the impact on user friendliness and "look and feel" of the software was tremendous.

The actual version 10 of SPSS is fully integrated into WINDOWS as well as network computing and eases the use and access notably for beginners. The Common User Access(CUA) of WINDOWS and the integrated help features facilitates the learning process and helps even experienced users handling the program. The data and information exchange between SPSS and other WINDOWS applications eases and speeds up the work of the analyst.

This introduction will not replace a SPSS reference manual, but the use of this manual will accompany the participants through the set of guided exercises and together with the on-line tutorial and the on-line help should be sufficient for the aims of the course

1.1 Conventions of text styles

Throughout this manual we will follow certain conventions. We hope, that this will make the text easier to understand.

| | |
|-------------------|---|
| Alt] | You are instructed to type the indicated key |
| [Alt] + [Shift] | You are instructed to type the indicated keys together |
| [End],[_] | You are instructed to type the indicated keys one after the other |
| [Files/Open/Data] | Menu choices, you can use mouse or keyboard to activate the menu choice |

1.2 How to choose options from menus.

There are two methods of choosing items from a menu:

With a mouse: Click on the name of the menu that you want. The name of the menu is highlighted and the list of menu items drops down. You can then view the items and click on the particular item that you want. (Or click on the menu name, holding your finger on the mouse button. Without releasing the button, drag through the menu items in order. They become highlighted in succession as you drag through the list. When the item that you want is highlighted, release the mouse button.) To cancel: click outside the menu and menu bar.

With the keyboard: Press the [Alt] key and the underlined letter (often the first) in the menu name. When the menu drops down, press the underlined letter in the command name. To cancel: press [Esc].

Neste manual serão seguidas determinadas convenções, de modo a facilitar a compreensão do texto, como seja:

2 A general view of SPSS



The objective of this chapter is to give you a general view of the program. You should learn which are the main elements of SPSS, the types of windows and files SPSS uses

2.1 How to start and stop SPSS

- Click two times on the SPSS icon to start
- Choose File/Exit to leave the program

2.2 SPSS windows

SPSS uses the following most important types of window:

Data editor: A rectangular, spreadsheet-like display of the working data file. You can edit the data in the data window, add or delete variables, or change their attributes, except when the SPSS processor is modifying the data. Use File menu commands **New** and **Open** to clear data from the data editor, or to open an existing data file in the data editor. Use File menu commands **Save** or **Save As** when the data editor is active to save the data file. You cannot close the data editor, although you can minimize it.

The file extension used by SPSS for this type of file is .sav.

Viewer. All statistical results, tables, and charts are displayed in the Viewer. You can edit the output and save it for later use. A Viewer window opens automatically the first time you run a procedure that generates output. Results are displayed in the Viewer. You can use the Viewer to:

Browse results.

- Show or hide selected tables and charts.
- Change the display order of results by moving selected items.
- Move items between the Viewer and other applications.

The Viewer is divided into two panes:

- The left pane of the Viewer contains an outline view of the contents.
- The right pane contains statistical tables, charts, and text output.

You can use the scroll bars to browse the results, or you can click an item in the outline to go directly to the corresponding table or chart. You can click and drag the right border of the outline pane to change the width of the outline pane.

Through the Viewer you can call various other editors

- **Pivot Table Editor.** Output displayed in pivot tables can be modified in many ways with the Pivot Table Editor. You can edit text, swap data in rows and columns, add color, create multidimensional tables, and selectively hide and show results.
- **Chart Editor.** You can modify high-resolution charts and plots in chart windows. You can change the colors, select different type fonts or sizes, switch the horizontal and vertical axes, rotate 3-D scatterplots, and even change the chart type.
- **Text Output Editor.** Text output not displayed in pivot tables can be modified with the Text Output Editor. You can edit the output and change font characteristics (type, style, color, size).

The file extension used by SPSS for this type of file is .spo.

Syntax window with Syntax Editor: You can paste your dialog box choices into a syntax window, where your selections appear in the form of command syntax. You can then edit the command syntax to utilize special features of SPSS not available through dialog boxes. You can save these commands in a file for use in subsequent SPSS sessions.

If you have more than one open Viewer window, output is routed to the designated Viewer window. If you have more than one open Syntax Editor window, command syntax is pasted into the designated Syntax Editor window. The **designated** windows are indicated by an exclamation point (!) in the status bar. You can change the designated windows at any time.

The **designated** window should not be confused with the **active** window, which is the currently selected window. If you have overlapping windows, the active window appears in the foreground. If you open a new Syntax Editor or Viewer window, that window automatically becomes the active window and the designated window.

The file extension used by SPSS for this type of file is .sps

2.3 How to edit data

The data editor is the basic data file, where each line represents one case (observation). For example each interviewed person is a case. Each variable contains information about the cases: for example a variable contains information the sex of the individual and its relation with the head of the household. In the data editor you can add and to edit the data but if you cannot execute calculations or include formulas.

You can enter data case-by-case and value-by-value, but this is rarely done. Usually data entry tools are use like specialized software or generic software like ACCESS tables and entry forms

The screenshot shows the SPSS Data Editor window for the file 'DataIndividuaisSofalaSecçãoBsemEtiquetas.sav'. The data is displayed in 'Data View' mode. The table contains 19 rows of data. The columns are: 'regio', 'a1', 'a2', 'memnumbe', 'b1', 'b2', 'b3', 'b4', 'b5', 'b6', 'var', 'var', and 'w'. The first 10 rows have 'regio' values of 07 and 'a1' values of 469. The next 3 rows have 'regio' values of 07 and 'a1' values of 469. The last 6 rows have 'regio' values of 07 and 'a1' values of 469. The 'memnumbe' column contains values ranging from 1 to 10. The 'b1' through 'b6' columns contain numerical values. The 'var' and 'w' columns are empty.

| | regio | a1 | a2 | memnumbe | b1 | b2 | b3 | b4 | b5 | b6 | var | var | w |
|----|-------|-----|-----|----------|----|----|----|----|----|----|-----|-----|---|
| 1 | 07 | 469 | 001 | 1 | 1 | 1 | 1 | 51 | 4 | . | | | |
| 2 | 07 | 469 | 001 | 2 | 2 | 1 | 2 | 47 | 4 | . | | | |
| 3 | 07 | 469 | 001 | 3 | 2 | 1 | 3 | 30 | 6 | . | | | |
| 4 | 07 | 469 | 001 | 4 | 1 | 1 | 3 | 20 | 1 | . | | | |
| 5 | 07 | 469 | 001 | 5 | 2 | 1 | 3 | 18 | 1 | . | | | |
| 6 | 07 | 469 | 001 | 6 | 2 | 1 | 3 | 15 | 1 | . | | | |
| 7 | 07 | 469 | 001 | 7 | 1 | 1 | 3 | 10 | . | 1 | | | |
| 8 | 07 | 469 | 001 | 8 | 1 | 1 | 3 | 08 | . | 1 | | | |
| 9 | 07 | 469 | 001 | 9 | 2 | 1 | 3 | 11 | . | 1 | | | |
| 10 | 07 | 469 | 001 | 10 | 2 | 1 | 3 | 07 | . | 1 | | | |
| 11 | 07 | 469 | 003 | 1 | 1 | 1 | 1 | 58 | 4 | . | | | |
| 12 | 07 | 469 | 003 | 2 | 2 | 1 | 2 | 46 | 4 | . | | | |
| 13 | 07 | 469 | 003 | 3 | 1 | 1 | 3 | 16 | 1 | . | | | |
| 14 | 07 | 469 | 006 | 1 | 1 | 1 | 1 | 31 | 4 | . | | | |
| 15 | 07 | 469 | 006 | 2 | 2 | 1 | 2 | 19 | 4 | . | | | |
| 16 | 07 | 469 | 006 | 3 | 2 | 1 | 3 | 04 | . | 1 | | | |
| 17 | 07 | 469 | 008 | 1 | 1 | 1 | 1 | 31 | 6 | . | | | |
| 18 | 07 | 469 | 010 | 1 | 1 | 1 | 1 | 37 | 4 | . | | | |
| 19 | 07 | 469 | 010 | 2 | 2 | 1 | 2 | 22 | 4 | . | | | |

Variable and value labels.

Besides defining the data type of variables, labels can be attributed to variables and values be used in statistical reports and graphs. For example, to attribute the label " Sex " to the variable B1 and labels 'Male ' and ' Female ' to numerical values 1 and 2 of this variable.

2.4 Exercises

General observation. The data of the IAF 2002/03 are the property of the INE. These data are confidential and anonymous (one does not know the personal location, as address and locality and the individual details on the inquired families and individuals). The improper use of these data is of the INE consists of an infraction of the Law of Statistics (7/96) and would be punished as such.

- 1.) Use the data file **DadosIndividuaisSofalaSecçãoB.xls**. Compare the file with the first two pages of the questionnaire of IAF 2003
 - a. How many questions can be counted
 - b. How many individuals had been interviewed
- 2.) Use the menu [File/Open/Data] to open a new empty data file. Save the file as **DadosIndividuaisSofalaSecçãoBsemEtiquetas.sav**. Recover the data file immediately after this.
- 3.) Use "Variable View" of the **Data Editor** to add the following variable and value labels:

| Variable | Variable Label | | Value | Value Label |
|----------|-------------------------|--|-------|-----------------|
| REGION | * No label * | | | |
| A1 | Area Enumeração | | | |
| A2 | Agregado Familiar | | | |
| MEMNUMB | Número de Membros | | | |
| B1 | Sexo | | 1 | Masculino |
| | | | 2 | Feminino |
| B2 | Nacionalidade | | 1 | Moçambicana |
| | | | 2 | Outra |
| B3 | Relação com chefe do AF | | 1 | Chefe |
| | | | 2 | Cônjuge |
| | | | 3 | Filho/a |
| | | | 4 | Pai/mãe |
| | | | 5 | Outros parentes |
| | | | 6 | Sem parentesco |
| B4 | Idade em Anos | | | |
| B5 | Estado Civil | | 1 | Solteiro/a |

| Variable | Variable Label | | Value | Value Label |
|----------|----------------|--|-------|-------------------------------|
| | | | 2 | Casado/a |
| | | | 3 | Casado/a em reg.poligamia |
| | | | 4 | Unido maritalmente |
| | | | 5 | Divorciado/a ou Separado/a |
| | | | 6 | Viúvo/a |
| B6 | Mãe viva? | | 1 | Sim |
| | | | 2 | Não |

2.5 The first analysis with SPSS: The analysis of frequency

The principal steps of the SPSS analysis

There are few simple basic steps for the analysis of data with the SPSS:

Enter and edit the data in the **Data Editor**

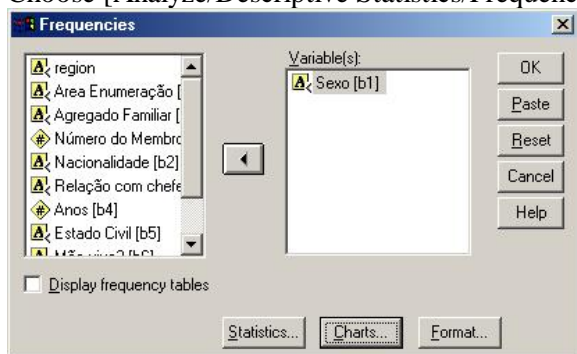
Select a menu option to create tables, to calculate statistics or to create charts.

Select the variables to be used in the analyses.

Select the options of the chosen procedure and verify the results.

After this, start with the first simple statistical analysis: How many women and men are in the group of respondents. Analyze the frequency of sex:

Choose [Analyze/Descriptive Statistics/Frequencies..].



Then choose the variables using the dialog box, deselect *Display frequency tables* and choose the option [**Charts**]:

With the choice of *Bar Chart(s)* and [**Continue**] in the second and the choice of [**Paste**] in the first window you get the following results: a syntax file, which you can save (syn01.sps) and another result: tables and graphics. The command to

Choosing variables for Frequencies

pressing the button in the syntax window,

if the cursor is positioned somewhere on the SPSS command line and all of the syntax file has been selected with your mouse.

You can also press [**OK**] to display the results immediately

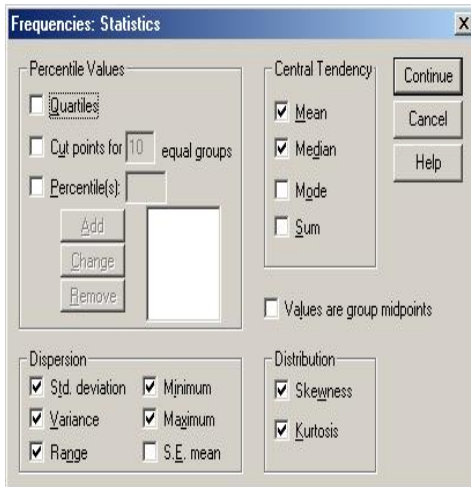
The content of the syntax file:



analyze the frequency will be executed by choosing [**Run/All**] or

FREQUENCIES

```
VARIABLES=b1 /FORMAT=NOTABLE  
/BARCHART FREQ  
/ORDER= ANALYSIS .
```



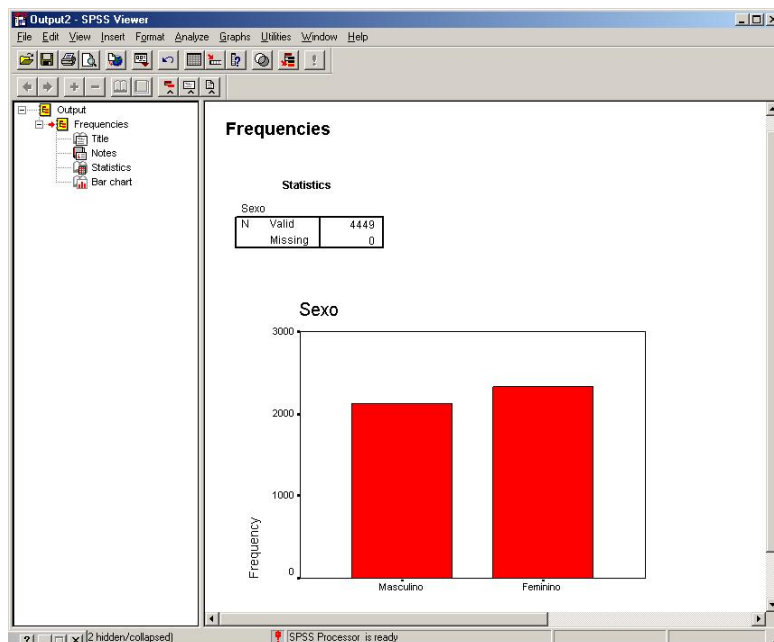
By choosing the option **[Statistics]** in the first window, you could choose to display statistics for another variable B4(age). After choosing *Histogram with Normal Curve* as the **[Graphs]** Option, the syntax file will be looking like this

```
FREQUENCIES  
VARIABLES=b4 /FORMAT=NOTABLE  
/STATISTICS=STDDEV VARIANCE RANGE  
MINIMUM MAXIMUM MEAN MEDIAN  
SKEWNESS  
SESKEW KURTOSIS SEKURT  
/HISTOGRAM NORMAL  
/ORDER= ANALYSIS .
```

Choosing Statistics for Frequencies



Although not being necessary to execute an syntax file instead of using the commands of the menu, its use will be of great utility in the future, for that if it recommended to be used



The results will be displayed in the **Viewer**

2.6 General considerations about measurement levels:

You can specify the level of measurement as scale (numeric data on an interval or ratio scale), ordinal, or nominal. Nominal and ordinal data can be either string (alphanumeric) or numeric. Measurement specification is relevant only for:

Chart procedures that identify variables as scale or categorical. Nominal and ordinal are both treated as categorical.

You can select one of three measurement levels:

Scale. Data values are numeric values on an interval or ratio scale (e.g., age, income). Scale variables must be numeric.

Ordinal. Data values represent categories with some intrinsic order (e.g., low, medium, high; strongly agree, agree, disagree, strongly disagree). Ordinal variables can be either string (alphanumeric) or numeric values that represent distinct categories (e.g., 1=low, 2=medium, 3=high). Note: for ordinal string variables, the alphabetic order of string values is assumed to reflect the true order of the categories. For example, for a string variable with the values of low, medium, high, the order of the categories is interpreted as high, low, medium -- which is not the correct order. In general, it is more reliable to use numeric codes to represent ordinal data.

Nominal. Data values represent categories with no intrinsic order (e.g., job category or company division). Nominal variables can be either string (alphanumeric) or numeric values that represent distinct categories (e.g., 1=Male, 2=Female).

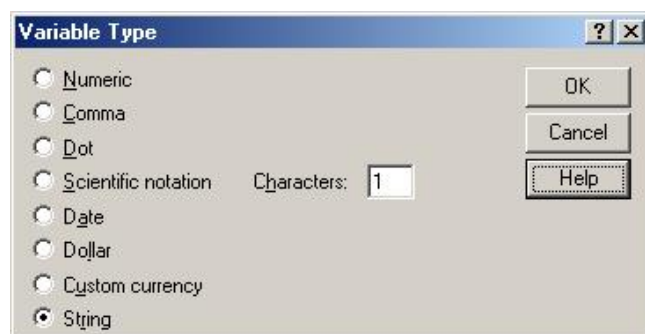
With the choice of an alphanumeric and nominal variable you will not be able to calculate statistics of distribution or of central tendencies

2.7 The variable definition in SPSS

To define the type of variable in the SPSS, click two times in the name of the variable on the top of the column or in the " Variabel View " in the column " Type ".
Select the type

By default:

- String (alphanumeric) variables are set to nominal.
- String and numeric variables with defined value labels are set to ordinal.
- Numeric variables without defined value labels but less than a specified number of unique values are set to ordinal.




2.8 Exercises.

Use SPSS to analyze the data of the file

DadosIndividuaisSofalaSecçãoBcomEtiquetas.sav:

1. Prints the frequency table for men and women in the sample
2. Modify the types of variables, if necessary or appropriate (Modify B4 = numeric)
3. Analyze the variable B4, display the statistics of central tendency, dispersion and distribution. Avoid to print the frequency table (Click in the main window of this analysis and deselect “*Display Frequency Tables*”)
4. Print a histogram of this variable and comment the differences of the distribution of this variable in relation to a normal distribution

2.9 The transformation of variables

 Before being able to analyze the data, usually some preparatory work is necessary, the transformation of variables and their respective values. You should know the possibilities and the reasons why these transformations are necessary and how they are done.

2.10 Defining variables

To define new variables or change the format of existing variables, double-click a column in the Data Editor, or select a cell in the column and choose [Define Variable] from the [Data] menu. You can also avoid to define variables and type data directly into the specified cells. Then the variable names will be provided by SPSS by default: the first variable named will be VAR00001, the second VAR00002 etc.

2.11 Variable names in SPSS

SPSS variable names can contain up to 8 characters. The first character of a variable name that you define must be a letter or @. Each subsequent character must be a letter, a numeral, a period, an underscore, or one of \$#@, except that the final character cannot be a period. Uppercase and lowercase letters are equivalent in variable names. You cannot use as a variable name: ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO, WITH.

2.12 Variable Types

The following are valid variable types:

- Numeric
- Comma
- Dot
- Scientific Notation
- Date
- Dollar
- Custom Currency
- String

The two types most frequently used are described briefly below:

Numeric Variable

Defines a numeric variable whose values are displayed in standard numeric format, using the decimal delimiter specified in the International control panel. The data editor accepts numeric values in standard format; or in scientific notation, provided that both the E and the sign of the exponent are entered (1E+1 but not 1E1).

String Variable

Defines a string variable, whose values can contain any characters up to the defined length. Upper- and lower-case letters are considered distinct. Some SPSS commands cannot use string variables, since calculations with strings are not defined. Strings defined as longer than 8 characters in length ("long strings") are restricted in use.

There are functions to convert numeric into string variables and vice versa. Even though you will learn about SPSS functions later, these two important ones are explained below:

| | |
|---------------------------------------|---|
| <i>STRING(numexpr,format) String</i> | Returns the string that results when numexpr is converted to a string according to format. The second argument format must be a format for writing a numeric value. |
| <i>NUMBER(strexpr,format) Numeric</i> | Returns the value of the string expression strexpr as a number. The second argument, format, is the numeric format used to read strexpr. Thus if name is an 8-character string containing the character representation of a number, <i>NUMBER(name, f8)</i> is the numeric representation of that number. |

2.13 Missing values

There are two types of missing values in SPSS, user-missing values and system-missing values.

Any blank numeric cells in the data rectangle are assigned the system-missing value, which is displayed as a period. You can also assign values that identify information missing for specific reasons, and then instruct SPSS to flag these values as missing. System-missing and user-missing values are handled specially by SPSS statistical procedures and by data-transformation commands.

2.14 Compute variables

Menu: [Transform/Compute]
Command Language: *COMPUTE, IF*

To compute values for all cases, type the name of a single Target Variable. It can be an existing variable, or a new variable to be added to your working data file. Build an Expression for the values that should be assigned to that variable.

To compute values selectively, for cases satisfying some logical condition, select the **[If]** pushbutton as well. If the Target Variable is or will be a string variable you must select **[Type] & [Label]** to indicate its length. If you are computing a new variable, you can select **[Type] & [Label]** to specify a variable label for it.

To build an expression, either paste components into the *Expression* field, or place the cursor there and type.

Paste variable names by copying them from the source list at the left.

Paste numbers and operators from the calculator pad.

Paste functions by copying them from the function list at the right.

String constants must be enclosed in quotation marks or apostrophes.

Numeric constants must be typed in American format, with the dot as a decimal delimiter.

If the Transformation Options setting (in Edit/Preferences) is *Calculate values immediately* (the initial setting), SPSS reads the data file and carries out the data transformation when you select **[OK]**. If the setting is *Calculate values before used*, SPSS carries out the transformation as the next procedure reads the data, so the Data Editor will not reflect the transformation until then. Additional features are available in the SPSS command language.

2.15 Recode variables

Recode into Same Variables

Menu: [Transform/ Recode/into Same Variables...]

Command language: *RECODE*

Move the variables that you want recoded according to the same specifications into the Variables list box. Select Old and New Values and specify how to recode the variable(s). If you want to recode only the cases satisfying a logical condition, select **[If]** and specify the condition.

Recode into Different Variables

Menu: [Transform/ Recode/into Different Variables...]

Command Language: *RECODE...INTO*

Move the variables of which you want copies recoded according to the same specifications into the *Input Variable -> Output Variable* list box. Select each variable individually in that box and, in the Output Variable group, type the Name (and optional Label) of a new Output Variable. Select **[Change]** to apply the name and label. If you modify the name or label, you must select **[Change]** again.

Select **[Old and New Values]** and specify how to recode the variable(s). If you want to recode only the cases satisfying a logical condition, select **[If]** and specify the condition.

See end of paragraph 3.6 for Transformation Options.

2.16 Transpose variables

Menu:[Data/Transpose]

Command Language: *FLIP*

Transpose transposes the rows and columns in the data file so that cases become variables and variables become cases. Select at least one variable. These variables are columns in the working data file, and will be transposed into cases (rows) in the new data file. Variables that you do not select will be dropped. SPSS reports the variable names in the transposed file, and creates a variable *CASE_LBL* containing the old variable names.

If the working data file contains an ID or name variable with unique values, you can move it into the *Name Variable* box, and its values will be converted into variable names for the transposed data file

2.17 Sort cases

Menu:[Data/Sort Cases]

Command Language: *SORT CASES*

Select a variable and move it into the *Sort by* box. To sort in descending order, select the variable in the *Sort by* box and press [**D**escending]. You can select additional sort variables. Each additional sort variable specifies the order of cases within the group that have identical values for all preceding sort variables. Ascending order is represented by (*A*) in the list of sort variables. String variables are sorted alphabetically. Descending order is represented by (*D*) in the list of sort variables. String variables are sorted in reverse alphabetical order.

The dialog boxes for the Data/Split File menu and the Statistics/Summarize/ Split File menu can automatically sort your data file, so it is not necessary to do so yourself before those commands.

2.18 Select cases

Menu: [Data/Select Cases]

Command Language: *SELECT IF; SAMPLE; FILTER; USE*

Choose one of the alternatives in the Select group:

All cases indicates that you wish to process all of the cases in your working data file.

If condition is satisfied indicates that you wish to process only the cases satisfying a logical condition. If you have specified a condition, it appears beside the **[If]** Pushbutton. If not, or if you want to change the condition, click on **[If]**. This alternative sets the variable *filter_\$*, discussed below.

Random sample of cases indicates that you wish to select cases randomly for processing. If you have entered sampling specifications they appear beside the **[Sample]** Pushbutton. If not, or if you want to change them, click on **[Sample]**. This alternative sets the variable *filter_\$*, discussed below.

Based on time or case range indicates that you wish to process only those cases falling within a range of specified dates. Click on **[Range]** to define the dates to be included in processing.

Use filter variable indicates that you want to use the values of an existing numeric variable to control case filtering. Select the variable from the list at the left. Cases for which the filter variable has the value 0 are excluded from analysis.

Choose one of the alternatives in the *Unselected cases are* group:

Filtered excludes unselected cases from processing, but retains them in the data file. You can re-enter this dialog box and change the Select alternative to restore the cases.

Deleted drops the cases from the data file. To recover them, you must Open the original data file again (assuming you have not replaced it with the reduced version).

If you execute this command with the **[If]** or **[Sample]** alternative and request that unselected cases be filtered, a variable named *filter_\$* is calculated as 1 (selected) or 0 (not selected). You can edit the values of *filter_\$* in the Data Editor; cases with a value of 0 for a filter variable are excluded from analysis, while cases with any other valid numeric value are included. However, be aware that this variable holds the current filtering status. If you do choose to edit its value, this affects which cases are filtered.


2.19 Exercises

Use SPSS to analyze the data of the file
DadosIndividuaisSofalaSecçãoBcomEtiquetas.sav:

It uses SPSS to transform the data

1. Define at missing values for at least one variable (candidates: *Idade em Anos* e *Estado Civil*).
2. Define a new variable `gr_age`
3. Compute values for this variable `gr_age` for each group with an interval of 10 years (`gr_age = 1` for an age between 0 and 10 years – include the upper value, `gr_age = 2` for an age between 10 and 20 years etc.). Use the command [Transform/Compute]
4. After the first operation use the command key [Paste] and use the syntax file for the rest of the operations
5. Which age group is the biggest and which is the number of persons in this group?
6. Sort the file according to the age
7. Sort the file according to the age and the civil status. Controle if there are married children, widows/widowers etc.
8. Select only the heads of households with more than 50 years of age. How many are there?
9. Select only the female heads of households. How many are there?

3 The transformation of data files

 Before data analysis can start, very often data files have to be prepared and transformed. It may be that files have to be imported from different sources, e.g. data bases like ORACLE, EXCEL or xBASE, sometimes several files have to be added, sometimes they have to be merged. In this chapter you should learn how this transformation is done and why this might be necessary.

3.1 Merging Files: General considerations

The dialog-box interface to the Merge Files facility combines two data files: the working data file, and a data file saved in SPSS format (called the "external file"). Two types of merge are supported:

1. Add Cases adds all of the cases from the external file to the end of the working data file. The new file includes whatever variables were shared by the two files, by default. You can choose to include variables that were in only one file; cases coming from the other file will have system-missing or blank values for such variables. You can also exclude variables.
2. Add Variables adds some or all of the variables in the external file to corresponding cases in the working data file. You can define "corresponding" cases in any of three ways:

- a. Case order. If you do not specify a key variable, cases are simply aligned according to their order in both files. This is risky unless you are certain that case orderings in the two files correspond, so usually type b. or c. of merging is selected.
- b. Key variable(s). If you specify a key variable, cases are aligned according to the values of that variable, which typically has unique identifying values. If necessary, you can use more than one key variable to specify the correspondence of cases. Both files must be sorted by the key variable(s). Unmatched cases in either file enter the new working data file, with system-missing or blank values for the variables that come from the other file.
- c. Table lookup. If you specify a key variable and indicate that one of the files is a "table," cases are aligned according to the key variable, but cases appearing only in the table do not contribute to the new working data file. The table simply provides data for new variables added to the other file. Both files must be sorted by the key variable(s).

If you have selected [*Calculate values before used*] in [Edit/Options.../Data], the files are not merged until you execute another command, such as a statistical procedure.

3.2 Merge Files: Add Cases

Menu:

[Data/Merge Files/Add Cases]

Command Language: *ADD FILES*

An initial dialog box asks you to identify the file from which you want to add cases.

The list box at the right, *Variables in New Working Data File*, contains all the variables shared by both files. If this list is satisfactory, press [OK]. The list box at the left, *Unpaired Variables*, contains variables that could not be matched across the files, because they had different names or different types. Variables marked with (*) are in the working data file. Variables marked with (+) are in the external file. Select one of the following for instructions:

Including a variable that is in one file only.

Including a variable with two different names.

Splitting a variable into two variables (giving different names for variables with identical names from two sources).

Excluding a variable.

Renaming a variable in the left list box.

Renaming a variable in the right list box.

If you select *Indicate case source as variable* a variable is added to the new working data file which has the value 0 for cases from the original working data file, and the value 1 for cases from the external file.

3.3 Merge Files: Add Variables

Menu: [Data/Merge Files/Add Variables]

Command Language: *MATCH FILES*

An initial dialog box asks you to identify the file from which you want to add variables.

Specify the type of match:

If you want to match the cases in the two files by sequential order alone, leave *Match cases* on key variables in sorted files unchecked.

If both files are sorted by one or more "key variables" and you want to match the cases according to their values: select *Match cases* on key variables in sorted files, choose one of the three alternatives for the type of keyed match, and move the key variable(s) from the list of Excluded Variables into the bottom list of Key Variables.

Specify the variables for the new working data file:

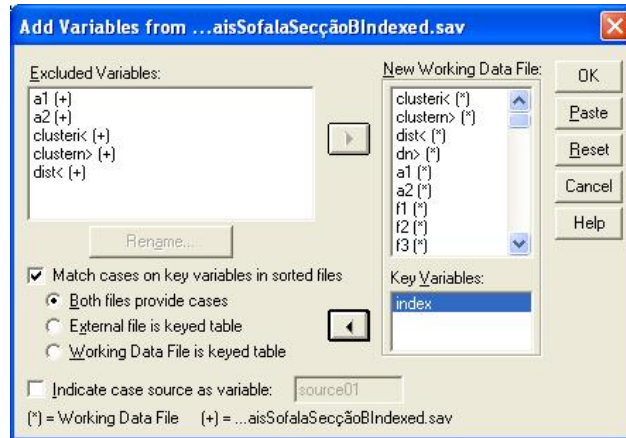
The list box at the right, *Variables in New Working Data File*, initially contains all the variables in the working data file, marked (*), followed by all differently-named variables in the external file, marked (+). You can add variables to this list, or remove them from it.

The list box at the left, *Excluded Variables*, contains variables that will not normally be included in the new working data file, because they have the same name as variables in the working data file. Initially all of them are from the external file, and are marked (+).

Variables marked with (*) are from the (old) working data file. Select one of the following for instructions:

Replacing the values of an existing variable.
Adding a variable with a conflicting name
Excluding a variable.
Using different-named key variables.
Replacing working-file data from
external file.

If you select *Indicate case source as variable* a variable is added to the new working data file which has the value 0 for cases that were in the original working data file, and the value 1 for cases that were added from the external file, and may therefore be questionable. (This option is disabled when you select one of the keyed table alternatives, since cases are never added from a keyed table.). To change the name of a variable in the *Excluded Variables list*, select it and press [**R**ename].



**Dialog box for MATCH FILES
according to keyed variable**

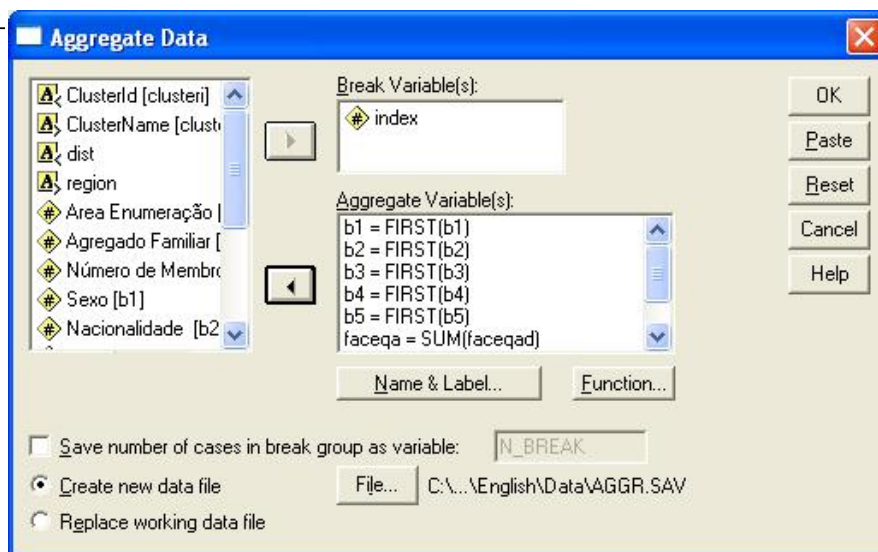
3.4 The aggregation of the data

Menu: [Transform/Aggregate] Command: AGGREGATE

Aggregate Data combines groups of cases into single summary cases and creates a new aggregated data file. Cases are aggregated based on the value of one or more grouping variables. The new data file contains one case for each group. For example, you could aggregate county data by state and create a new data file in which state is the unit of analysis.

[Break Variable(s)]. Cases are grouped together based on the values of the break variables. Each unique combination of break variable values defines a group and generates one case in the new aggregated file. All break variables are saved in the new file with their existing names and dictionary information. The break variable can be either numeric or string format.

[Aggregate Variable(s).]
Variables are used with aggregate functions to create the new variables for the aggregated file. By default, Aggregate Data creates new aggregate variable names using the first several characters of the source variable name followed by an underscore and a sequential two-



Dialog box for AGGREGATE on break variable “index”

digit number. The aggregate variable name is followed by an optional variable label in quotes, the name of the aggregate function, and the source variable name in parentheses. Source variables for aggregate functions must be numeric.

You can override the default aggregate variable names with new variable names, provide descriptive variable labels, and change the functions used to compute the aggregated data values. You can also create a variable that contains the number of cases in each break group.


3.5 Exercises

It uses the SPSS to analyze the data in the file
DadosIndividuaisSofalaSecçãoBcomEtiquetas.sav:

Use o SPSS para transformar os dados

1. Transform the variables A1, A2, B1, B3, B4 e B5 into numeric variables, if this has not been done before..
2. Use the commando [Transform/Compute] to create a new variable `index` with the calculation: `COMPUTE index = a1*1000+a2.`
3. Use the commando [Transform/Compute] to create a new variable `FacEqAd` ("Fator do Equivalente Adulto" with the calculation:
`IF (b4 >17) FacEqAd = 1.`
`IF (b4 <= 17 & b4 > 10) FacEqAd = 0.7.`
`IF (b4 <= 10 & b4 > 5) FacEqAd = 0.5.`
`IF (b4 <= 5) FacEqAd = 0.2.`
4. Save the file as
DadosIndividuaisSofalaSecçãoBcomEtiquetasIndexed_EqAd.sav
5. Use the commando [Transform/Aggregate] to aggregate the data of the file on the variable `index` as *break variable*. Include all variables a1-b5 as aggregate variables: Use the function `First()` to be applied for all variables except for the new variable `FacEqAd` and `memnumb`. There you must use `Sum()` and `N()`
6. Open the new file AGGR.SAV and save it as
DadosIndividuaisSofalaSecçãoBcomEtiquetasAggregated.sav. You now have the Individual archive aggregated an to be combined with the data of the Household. You should sort the file on the variable `index` and save it.
7. Open the file **DadosAggegadosFSofalaSecçãoF_G.sav** Use the command [Transform/Compute] to create a new variable `index` as in the Exercise 4.).
8. Use the command [Transform/Add Files/Match Variables] e adds the file **DadosIndividuaisSofalaSecçãoBcomEtiquetasAggregated.sav** . Choose *Match cases on key variables in sorted files* (the files have both to be sorted on the variable `index`) e *Both file provide cases*. Choose the common variable `index` as key variable and combine the two files. Save the file as
DadosAgregadosFamiliaresComIndividuais.sav . This archive contains characteristics of the households and the heads of the households to be used in the next chapters.

4 Functions for the Initial Analysis of Data

 In this chapter you will learn to use SPSS functions. There will be very little explanation about the statistical theory behind the functions, existing knowledge is assumed. If you are not sure, what the method does, please consult a statistical text book. The main objective of this chapter is to explain the handling of the functions and the interpretation of the results. You should learn both this in this chapter.

4.1 Frequencies

Menu: [Analyse, Descriptive Statistics /Frequencies...]
Command Language:FREQUENCIES

Move one or more variables into the *Variable(s) list*. A frequency table will be produced for each variable in this list, showing the number of cases for each value.

Select [**Statistics**] to request univariate statistics.
Select [**Charts**] to request bar charts or histograms.
Select [**Format**] to control the order of values in the table, its arrangement on the page, the display of an index, and the display of value labels.

If you deselect the [*Display frequency tables*] box, no frequency table is displayed, but any statistics or charts that you request will appear.

[Statistics]

| | |
|------------------------------------|--|
| <i>Quartiles</i> | Values that divide the cases into four equal-sized groups. |
| <i>Mean</i> | The arithmetic average; the sum divided by the number of cases. |
| <i>Cut points for equal groups</i> | Values that divide the cases into some number of equal-sized groups |
| <i>Median</i> | A value that is greater than half of the actual data and less than the other half. The 50th percentile. |
| <i>Percentiles</i> | Values above and below which certain percentages of cases fall. |
| <i>Mode</i> | The most commonly occurring value. If several values share the greatest frequency of occurrence, each of them is a mode. The Frequencies procedure reports the smallest of such multiple modes. |
| <i>Sum</i> | The sum or total of the values, across all cases with nonmissing values. |
| <i>Std. deviation</i> | A measure of dispersion around the mean, equal to the square root of the variance. The standard deviation is measured in the same units as the variable itself. |
| <i>Minimum</i> | The smallest value of a numeric variable. |
| <i>Skewness</i> | A measure of the asymmetry of a distribution. Positive skewness indicates that the more extreme values are greater than the mean; negative skewness indicates that the more extreme values are less than the mean. |

| | |
|------------------|--|
| <i>Variance</i> | A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself. |
| <i>Maximum</i> | The largest value of a numeric variable. |
| <i>Kurtosis</i> | A measure of the extent to which a distribution is "tail-heavy," compared to a normal distribution. Positive kurtosis indicates more cases in the extreme tails than in a normal distribution with the same variance. |
| <i>Range</i> | The difference between the largest and smallest values of a numeric variable; the maximum minus the minimum. |
| <i>S.E. mean</i> | The standard error of the mean of the summarized variable for all the cases in the cell. Available for summarized variables and their totals. |

If you select Values are group midpoints, the statistics in Percentile Values and the median statistic are calculated under the assumption that your data have been grouped, and that the actual values in the data are the midpoints of the original groups.

[Charts..]

A **bar chart** displays the frequency count for each value as a separate bar. Bar charts are appropriate for variables measured at the nominal level, or for variables with few distinct values. Values for which the count is zero do not appear. On bar charts, the scale axis can be labeled by *actual frequency counts* or by *percentages*.

A **histogram** displays the frequency count for ranges of values as separate bars. Histograms are appropriate for variables for which it is meaningful to group adjacent values. Space is left for ranges into which no cases fall. A normal curve can be superimposed on a histogram as an aid in judging whether the variable is normally distributed.

(You can obtain a group of histograms that all use the same scale by selecting from the [Statistics/Summarize/Explore..] submenu, or by using the SPSS command language.)

[Format..]

The frequency table can be arranged according to the actual values in the data, or according to the count (frequency of occurrence) of those values, and in either ascending or descending order. However, if you request a histogram (in [Charts]) or you request percentiles or equal cut points (in [Statistics]), the table is always arranged in ascending order of values.

[Suppress tables with more than] a specified number of values displays the remaining tables in standard format.

4.2 Explorative data analysis

Like the *Frequencies* function delivers some general statistical information about the data set, the function *Explore* allows to observe statistics and listings about all observations as well as groups of observations. It incorporates some graphical techniques which are especially useful for a first analysis of the data set, comparison among groups of responses and test of normality as well as explicative graphics (like boxplots and stem and leaf

graphics) or a comparison of variances. Like the other functions in this chapter, this is a first approach to examine or explore the data, especially suited for a first glance at a bulky data set.

Menu: [Analyse, Descriptive Statistics, Explore]

Command language: EXAMINE

Move at least one variable into the Dependent List. Statistics and plots from this command will describe the distribution(s) of the variables in this list. You may move one or more grouping variables into the Factor List. The values of factor variables define groups of cases, which will be described separately.

You may move a variable into the Label Cases by box. This variable is used to identify unusual cases in some of the output. Select one of the alternatives in the Display group to choose plots, statistics or both for the output. These alternatives activate the **[Plots]** and **[Statistics]** pushbuttons:

- | | |
|---------------------|---|
| [Plots] | allows you to control the output of plots. |
| [Statistics] | allows you to request descriptive statistics and a listing of cases with the 5 smallest and the 5 largest values. |
| [Options] | allows you to determine the treatment of missing values. |

Before starting the analysis, let us explain briefly the used techniques:

The Boxplots shows the average number, quartiles and Outliers (isolated values). The "boxes" of Boxplot contain 50% of the values that fall between the 25% and 75% percentiles, and the "whiskers", the lines that extend from the box to the highest and lowest values, covering the extreme percentiles, excluding outliers. (outliers = values that are located at a distance of more 1.5 times the distance of the difference between the 25% and 75% percentiles from the respective percentile limit). A line across the box indicates the median.

Boxplots. These alternatives control the display of boxplots when you have more than one dependent variable. Factor levels together generates a separate display for each dependent variable. Within a display, boxplots are shown for each of the groups defined by a factor variable and are shown side by side for each dependent variable. This display is particularly useful when the different variables represent a single characteristic measured at different times.

Descriptive. The Descriptive group allows you to choose stem-and-leaf plots and histograms.

Normality plots with tests. Displays normal probability and detrended normal probability plots. The Kolmogorov-Smirnov statistic, with a Lilliefors significance level for testing normality, is displayed.

4.3 Exercises

Utilize o SPSS para analisar os dados no arquivo

DadosAgregadosFamiliaresComIndividuaisCompleto.sav

1. Analyze of the frequency of the highest degree of education of the head of household (C3)
2. Analyze of the frequency of the highest degree of education of the head of household (C3) between male and female heads of households. Which will be a problem of this analysis?
3. Make a exploratory analysis of the household size(memnumb) and the gender of the head of household (B1)
4. Count the frequency of head of households employed of the public sector, the private sector and working on own account (E8).
5. Create a new variable for a new age group (4 values only). Make a exploratory analysis of the relation between these age groups and gender of the head of household (B1) and the variables: *grau do ensino mais alto obtido* (C3) and *alguma vez frequentou a escola* (C2)
6. Analyze the normality of the distribution of the above mentioned variables with a exploratory analysis and frequency analysis, look at the extremes in each group (age groups and sexo(B1)).
7. Analyze the variables *Esteve doente ou ferido nas ultimas duas semanas* (D2) and *Quantos dias ficou sem trabalhar* (D3) for the above mentioned age groups.

5 The Presentation of Data



The goal of statistical analysis is to produce results. These results have to be communicated to the users of statistics, sometime these are statisticians, sometimes they are not. So the presentation has to be comprehensive, informative and attractive. This means that the statistician has to produce results and present them according to his or her audience. The explications necessary to understand results will always be better. Graphics and charts should always be preferred as means of presentation even as a supplement to more traditional presentation. This chapter deals with the possibilities SPSS offers to publish and display statistical data, listing, tabulation and graphical representation. This is in brief the techniques you will use for most statistical publication. You should learn these methods and functions, you should know when and how to use them.

5.1 List Cases

Menu: [Analyze/Reports/Case Summaries]
Command Language: LIST

displays the values of variables for cases in the data file
Move one or more variables into the Variable(s) list. The values of the selected variables will be displayed in the listing file.

To list a subgroup of the cases that start with the first case, selects *Limit cases to .. e* indicate the number of the last case to be listed.
You can group variables by moving a variable to the box *Grouping Variable(s)*

5.2 The Tables Module

The supplementary module Tables is a combination of list functions, report functions, crosstabs and descriptive statistics. All these SPSS functions merge into a set of commands which allow to produce tables almost ready for publication. Without any problems these tables produced can be incorporated in any report using the ordinary data exchange opportunities WINDOWS offers.

5.2.1 Basic Tables

produces publication-quality tables displaying crosstabulations and subgroup statistics. It allows nested layouts. The same statistics are reported for all variables summarized in the table.

Menu: [Analyze/ Custom Tables/Basic Tables]
Command Language: TABLES

Move one or more variables into the Summaries list, or into any of the three Subgroups lists, or both. At minimum, you need only one variable in any one of these boxes. If you specify a variable for Summaries, you should select **[Statistics]** to indicate which statistic(s) you want to see.

To obtain simple frequency tables or crosstabulations containing counts or percentages, leave the Summaries list empty and move grouping variable(s) into the Subgroups lists.

To obtain subgroup statistics, move the variable(s) for which you want statistics into the Summaries list and the grouping variable(s) into the Subgroups lists. Select **[Statistics]** to specify the desired statistics. Be sure to select appropriate statistics for the summarized variables. To obtain overall statistics, move the variable(s) for which you want statistics into the Summaries list and select **[Statistics]** to specify the desired statistics. Be sure to select statistics that are appropriate for the summarized variables.

The three Subgroups lists let you define cells to display subgroups of cases in any combination of:

Down the page (as separate rows).

Across the page (as separate columns). The combination of Down and Across generates a crosstabular display.

(Separate Tables) Spread across Separate Pages. Only one layer of the table is visible at a time. You can view other layers after the table is displayed in the Viewer by double-clicking the table and clicking the arrows on the layer pivot icon. *

When you specify more than one grouping variable in any one dimension, you can choose between the alternatives *All combinations (nested)* and *Each separately (stacked)*. (The General Tables command offers more detailed control over the relation between multiple grouping variables in a dimension.)

Select **[Statistics]** to specify the statistics in the table. For statistics on summary variables, select the variable in the Summaries list first. By default the means of summary variables are displayed if there are any summary variables, or the case counts if there are none.

Select **[Layout]** to specify the summary dimension, the statistics dimension, the relative position of summary variables and subgroups in the summary dimension, and to suppress variable labels for grouping variables.

If you have entered a variable into any of the Subgroups lists, you can select **[Totals]** to request total statistics (after the subgroup statistics) for grouping variable(s).

Select **[Format]** to specify margins, column sizes, borders, the appearance of empty cells and undefined statistics, and the characters at which labels can split between lines.

Select **[Titles]** to specify lines of text for the table header, footer, and top-left corner.

5.2.2 General Tables

provides the greatest amount of control over the content and format of tables, at the cost of somewhat more complex specifications. Different statistics can be displayed for different variables. The General Tables command allows you to nest some variables and stack others. It also supports tables involving multiple-response and multiple-dichotomy sets.

Menu: [Analyze/ Custom Tables/General Tables]

Command Language: *TABLES*

Specify where *Statistics Labels Appear* to determine where the labels for cell statistics appear. (These alternatives are sometimes unavailable, when only a single location is possible.)

Move one or more variables into any or all of the Row list, the Column list, or the Layer list. You can use either the regular variables in the source list at the upper left, or you can use Multi Response sets. (To add multiple-response or multiple-dichotomy sets to the Multi Response list for use in the table, select [Multi Response Sets].)

In order for the [OK] and [Paste] buttons to be available:

1. Something must be down the side of the table. You must put a variable in the Row list unless you specify that statistics labels appear down the side.
2. For multi-page tables, something must be across the top, too. If you put a variable into the Layer list, you must also put one into the Column list, unless you specify that statistics labels appear across the top.

When a variable is highlighted in the Rows, Columns, or Layers box, you can use the Selected Variable controls at the right to modify its function in the table:

Defines cells identifies a grouping variable, each of whose categories forms a row or column in the table, or a separate subtable. You can display counts or various percentages for the categories of a grouping variable.

Is summarized identifies a summary variable, one for which summary statistics appear in the table. The first summary variable you specify defines the summary dimension.

The default summary statistic is the mean and is not labeled; you can label it by selecting the summary variable and visiting the Edit Statistics subdialog.

While a summary dimension is specified, the words Summary Dimension appear above the variable list for that dimension, and you cannot select *Is summarized* for a variable in another dimension.

Select [Omit Label] to suppress the variable label of the selected variable.

Multiple variables in any of these lists are stacked in their dimension within the table unless you nest them. The [Nest] button allows you to nest a selected variable within categories of the variable that precedes it in the list for its dimension. The [Unnest] button allows you to move a nested variable out of its nested position.

Nesting A nested variable is displayed "within" each category of a grouping variable or a multiple-response set. Up to 4 levels of nesting are allowed. You cannot nest a variable "within" a summarized variable or a total, and you cannot nest a multiple response set within another multiple-response set.

To nest a variable within categories of a variable below it in the list, remove it from the list and add it again at the bottom. For subgroup summary statistics, nest the summarized variable within categories of the grouping variable.

[Edit Statistics] lets you specify which statistics will be displayed for the selected variable. You cannot assign statistics to a variable within which another variable is nested. Once you have assigned statistics to a variable, they become the default statistics for other variables of the same type in the same dimension. This makes it easy to assign the same statistics to several variables in the statistics dimension.

[Insert Total] allows you to add total statistics after the selected variable's group statistics. You can modify the label that will be used for the total statistics in the Total Label: field. The same statistics are normally displayed for a total as for the variable totaled; but by selecting the total and clicking **[Edit Statistics]**, you can specify different statistics for the line or column containing the total statistic(s).

Select **[Format]** to set margins, page dimensions, borders, and the treatment of missing cells and data.

Select **[Titles]** to specify lines of text for the table header, footer, and top-left corner.

5.3 Exercises:

Use SPSS to present the following results:

Tables of:

1. Main occupation of the head of household (frequency)
2. Main occupation of the head of household according to sex and group of age
3. Main occupation of the head of the family according to highest school degree obtained.

5.4 Graphics and Charts

Graphics are certainly the most attractive, comprehensive and persuasive means to communicate between the producer and consumer of statistical analysis. They are certainly recommended if little time or space is available to explain the thesis of the analyst. The publication of statistics via the media like television or newspapers is inconceivable without the presence of graphics. Graphs and charts however are usually not self-explained and the analyst must be careful to explain and comment the charts in an appropriate way, either by the possibilities of the graphs itself (legends etc.) or by simple text files. You should learn in this section the use of graphics, their modification and their importance in the context of the presentation of your data..

Creating Charts

There are mainly three ways of creating charts and plots in SPSS:

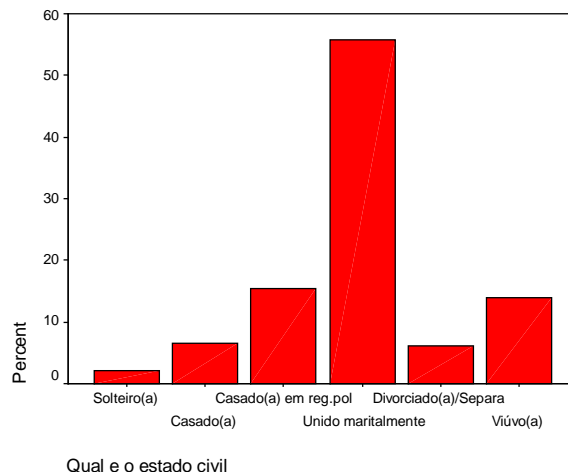
1. Commands from the Graphs menu on the main menu bar produce charts or plots. The main menu bar is available when the Data Editor, an output window, or a syntax window is active.
2. Certain statistical procedures produce plots. Examples include boxplots in the Explore procedure or scatterplots generated from Linear regression.
3. Interactive Charts

To create an interactive chart, select Interactive from the Graphs menu. For example, to create a bar chart, select Bar from the Interactive menu. Drag and drop variables from the source list to the target lists.

Modify Charts

There are mainly two ways of modifying charts and plots in SPSS:

1. Using the Chart Editor
2. Using the Interactive Charts E



5.4.1 Graphs from the Graphs Menu

To create a chart...

- Select [Graphs] from the menu bar.
- Select the type of chart you want from the Graphs menu.
- Choose the icon for the specific type of chart you want.
- You also need to indicate how your data are organized.

Graphs from the Graphs Menu

Options of Menu [Graphs]

- Bar allows you to generate a simple, clustered, or stacked bar chart from your data.

-
- Line allows you to generate a simple or multiple line chart from your data.
 - Area allows you to generate a simple or stacked area chart from your data.
 - Pie allows you to generate a simple pie chart or a composite bar chart from your data.
 - High-Low allows you to plot pairs or triples of values, for example high, low, and closing prices.

 - Pareto creates Pareto charts, bar charts with a line superimposed showing the cumulative sum.
 - Control produces the most commonly-used process-control charts.
 - Boxplot allows you to generate boxplots showing the median, interquartile range, outliers, and extreme cases of individual variables.
 - Error Bar allows you to generate boxplots showing the median, interquartile range, outliers, and extreme cases of individual variables.
 - Scatter allows you to generate a simple or overlay scatterplot, a scatterplot matrix, or a 3-D scatterplot from your data.
 - Histogram allows you to generate a histogram showing the distribution of an individual variable.
 - Normal P-P plots the cumulative proportions of a variable's distribution against the cumulative proportions of the normal distribution.
 - Normal Q-Q plots the quantiles of a variable's distribution against the quantiles of the normal distribution.
 - Sequence produces a plot of one or more variables by order in the file, suitable for examining time-series data.
 - Time Series: Autocorrelations calculates and plots the autocorrelation function (ACF) and partial autocorrelation function of one or more series to any specified number of lags, displaying the Box-Ljung statistic at each lag to test the overall hypothesis that the ACF is zero at all lags.
 - Time Series: Cross-correlations calculates and plots the cross-correlation function of two or more series for positive, negative, and zero lags.

5.4.2 Graphs resulting from procedure calls

As in many previous examples, graphs can be produced as a result of SPSS procedures. Examples are the bar charts and histograms of the *FREQUENCIES* procedure.

5.4.3 Graphs from the Interactive Graphs Menu

A limited set of chart types can be selected through the Menu option: [Graphs/Interactive ..].

The following chart types are also available as interactive charts.

- Bar
- Line
- Pie
- Histogram
- Boxplot

- Error Bar
- Scatterplot

5.4.4 Using the Chart Editor

The Chart Editor is where you modify, print, and save charts that you have created elsewhere. You can change almost any attribute of a chart here. You can also change a chart from one type to another (for example, change a bar chart to a line chart). To edit the overall chart, rather than a specific chart object, select a chart editor procedure from the Gallery, Chart, or Series menu.

To edit a chart object (such as an axis, title, legend item, bar or line), first select that object by clicking on it once. "Handles" appear on the selected object. (Instead of handles, a rectangle appears around selected text.) Selecting an object, such as a bar, selects all objects of that class. For example, selecting a bar in a simple bar chart selects all bars in that chart. Selecting one bar in a clustered bar chart selects all bars in that series.

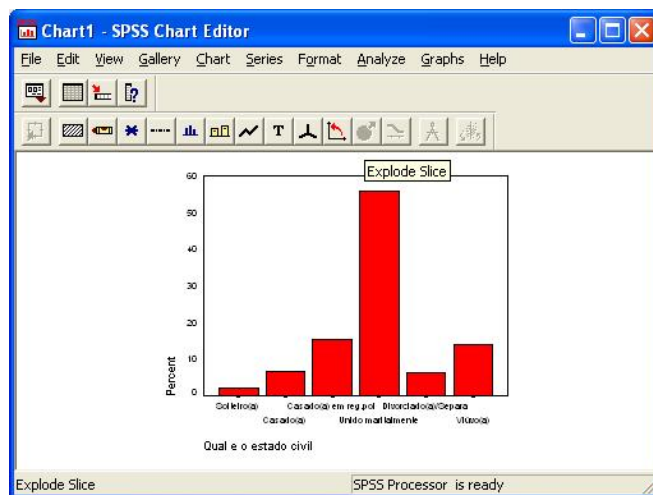
After selecting the object, choose the appropriate palette for editing the attribute that you want to change. Select the palette in either of two ways:

1. Make a selection from the Attributes menu.
2. Select one of the attribute buttons on the chart window icon bar.

You can also select chart objects by double-clicking on them. When you double-click an object, an appropriate palette is automatically opened.

Attributes in Charts

All of the commands on the Attributes menu are available on the Icon bar when you are in the Chart Editor.



The Chart Editor

- Point selection. Click on the point selection tool to turn point selection mode on and off in the Chart window when a scatterplot or a boxplot is current.
- Fill Pattern allows you to change the fill of objects in an existing chart.
- Color allows you to change the color of a selected object.
- Markers allows you to change the markers on a chart or plot.
- Line Style allows you to change line styles.
- Bar Style allows you to change bar styles for charts that contain bars.
- Bar Label Style allows you to label bars with the numeric values that they represent.

- Interpolation allows you to select an interpolation method for a line or scatterplot series.
- Text allows you to change the font and point size of text within a chart or plot.
- Rotation rotates 3-D scatterplots, using a dialog box that shows only the axes.
- Swap Axes interchanges the vertical and horizontal axes of a chart.
- Explode Slice visually removes or "explodes" the selected slice from a pie chart.
- Break Line at Missing breaks a line chart where missing values should be.
- Chart options. Allows you to modify an existing chart.
- Spin Mode rotates 3-D scatterplots directly, hiding only the labels. It is slower than 3-D Rotation when many data points are present.

5.4.5 Interactive Graphs and the Interactive Chart Editor

To Select and Modify a Chart Object

Activate the chart (double-click it).

Right-click the object to be modified.

The object is selected and a context menu is displayed. You can choose an action or a dialog box from the context menu.

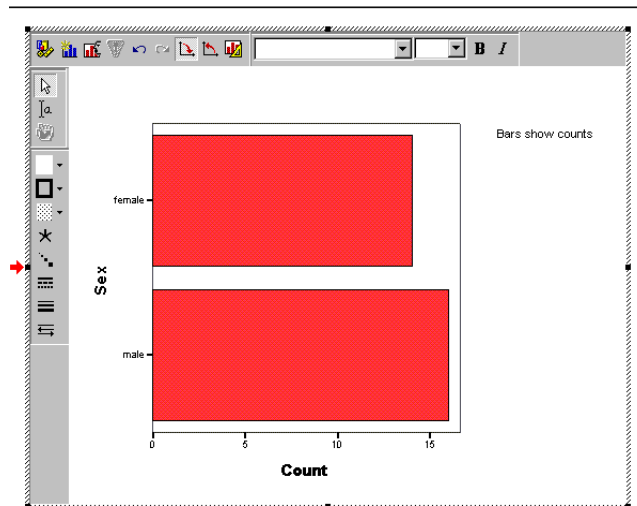
or

Click the object to be modified.

Select a command from the menus at the top of the Viewer.

or

Double-click the object to be modified.



The Interactive Chart Editor

This displays a dialog box for editing the object you clicked. Double-clicking a graphical element displays the element dialog box. For example, double-clicking a bar displays the Bars dialog box.

An interactive chart has access to the variables used to create it. Using the Assign Graph Variables dialog box, you can change variable assignments whenever the chart is activated.

If the data file that was used to create the chart is still open and nothing has changed in the Data Editor since the chart was created, the chart is labeled Interactive in the status bar of the Viewer, and you can change any of the variable assignments, using all of the variables available in the active data file. These charts use the Interactive Chart Editor.

If Save all data with the chart is selected on the Interactive tab of Options and the data file has been changed or a new data file has been opened, the chart is detached from the data file and is labeled Interactive (detached) in the status bar of the Viewer. For a detached chart, you can still change variable assignments; however, only variables used while the chart was actively connected to the open data file are now available. The Assign Graph Variables dialog box contains a message about available variables.

All interactive charts become static charts if Save only summarized data is selected on the Interactive tab of Options (Edit menu of the Viewer). In a Static chart, only the summarized data is saved, and you cannot change variable assignments, although you can change attributes such as color, fill, and symbol size. You can turn an Interactive chart into a Static chart by right-clicking and selecting Save only summarized data from the context menu. A static chart cannot be changed to an interactive chart.

Non-interactive charts and charts created in previous versions of the software are not interactive. The chart objects can be edited for attributes such as color, but the variable assignments cannot be changed. These charts use the Chart Editor window available in previous releases of the program.


5.5 Exercises:

Use SPSS to present the following results:

Graphs for:

1. Main reasons not having worked in last the 7 days (variable E5)?
2. *Highest school degree of head of household (C3) by Main professional activity (E9) and Gender of head of household (B1) (graph 3-D)*
3. *Main professional activity (E9) and Tipo de habitação, Construção de paredes de..(F7)*

6 The poverty Line

 The word „poverty“ is constantly used in political, economical and social reports. It is permanently used to describe, accuse and compare. But how is poverty defined? What is meant, if reports state, that 40% of the population is poor or 20% are „extremely“ poor? The concept of poverty lines will be briefly explained in this chapter. It must be clear, that only economic poverty is meant, if the word „poverty“ is used henceforth.

6.1 The concept of poverty as defined by the IAF

There are several different approaches to determine poverty and to measure poverty. In the IAF evaluation¹ the MPF follows the methodology of the cost of basic necessities (CBN) to construct specific regional lines of poverty (Ravallion 1994,1998). In the CBN approach, the line of total poverty is constructed as the addition of a food poverty and a non-food poverty line. With a poverty line constructed, the households, which spend less than the line of poverty at a per capita base is considered poor. Although this approach is well accepted, also very transparent and easy to adapt to different levels of disaggregating the data source, this paper will try also to add some analytical thoughts on some other approach to define poverty on a grass root level because the IAF approach has two severe shortcomings: certainly missing a substantial part of the ultra poor and vulnerable population and missing part of the rural population in remote and inaccessible areas. These shortcomings have to be accepted to reduce costs and overheads of a national survey as such. However on a more disaggregated level it would be extremely important to spot characteristic poverty profiles in space and time.

As the food poverty line is relatively well explained and especially the flexible food basket approach seems to be very reasonable. The line of poverty of the essential non-food products for 2002-03 uses the fractions of expenditures (non-food shares) derived from the IAF derivatives in 1996-97. So some thought were dedicated in this article on the non-food items as well as on the poverty related indicators, especially Health, Education, Employment and Access to other Services as well as to Income.

Similar to the 1996-97 IAF three measures had been used to measure the poverty. All they are members of classes P of indices of poverty of Foster-Greer-Thorbecke (1984)² which are used of a routine form to measure the poverty. Mathematically, all the indices of this class have the form:

$$P_{\alpha} = \frac{1}{n} \sum_{y \leq z} \left(1 - \frac{y}{z}\right)^{\alpha}, \alpha \geq 0$$

Where: n is the population, y is the per capita consumption, z is the poverty line, and α is a not negative parameter. We use the measures with $\alpha = 0, 1,$ and $2,$ that they correspond

¹ **Pobreza E Bem-Estar Em Moçambique: Segunda Avaliação Nacional** - Março de 2004 Direção Nacional do Plano e Orçamento, Ministério de Plano e Finanças Gabinete de Estudos, Ministério de Plano e Finanças Instituto Internacional de Pesquisa em Políticas Alimentares (IFPRI) Universidade de Purdue

² **Greer, J., J. Greer, and E. Thorbecke.** 1984. A class of decomposable poverty measures. *Econometrica* 52 (3): 761-765

to the incidence of the poverty, the index of depth of the poverty, and to the squared index of depth of the poverty, respectively.

6.1.1 The poverty headcount index

is the ratio of the population whose per capita consumption is below of the line of the poverty. This index can mathematically also be express as $P_0 = q/n$, where q is the number of poor people in one given region and n is the total population of the region.

6.1.2 The poverty gap index

is the average distance, in percentage, where the measured consumption is below of the poverty line using all the aggregates in the sample where the aggregates who live above of the line of the poverty receive value zero. Mathematically, that is the same that the average of difference between the levels of consumption of the poor persons and the line of poverty (express as a ratio of the poverty line), multiplied by the incidence of the poverty. Thus the index of depth of the poverty catches changes in the poverty structure that the index of poverty does not detect, because the index of depth of the poverty measures " How poor are the poor persons ". For example, if all the poor persons remained below of the line of the poverty and all the not poor ones remained above of the poverty line, but the incomes of the poor persons went up, many would say that the poverty decreased. The incidence of the poverty would not move to reflect this improvement, but the index of depth of the poverty will go to decrease, to show that the poor persons are not so poor as they were before.

6.1.3 The squared poverty gap index

is the squared average of the depth of the poverty. It measures the severity of the poverty, and takes in account the inequalities between the poor persons. For example, if transference is made from a person only slightly below of the line of the poverty to a person very much below of the poverty line, the squared poverty gap index will be reduced because the difference of standards of living of the poor between the poor persons will have improved. In contrast, such transference would not affect nor the poverty headcount index, nor the poverty gap index.

These indices can be fully disaggregated, that is indices calculated for subgroups add up to the full index.

| | Poverty Line Average expenses in Meticaís per Person per Day |
|--------------------------|---|
| Sofala e Zambezia rural | 5473 |
| Sofala e Zambezia urbana | 8775 |

It will become more clearly through the explicit formulas:

To know the ratio of the poor households/persons in relation to the total population ratio, it is calculated by:

$$P_0 = q / n$$

wher n is the number of the observations in the population

q is the number of households/persons to be considered poor (number of observations below the poverty line)
and z is always the value of the poverty line

Embora este valor meça a percentagem dos pobres nada diz, no entanto, sobre as características dos dados da amostra relativamente aos pobres. A fórmula seguinte mede a profundidade da pobreza, significando que um valor maior representa um parte substancial da população se encontra mais afastada da LdP do que um valor menor (tem mais ‘muito` pobres se este valor fosse elevado, tem menos ‘muito` pobres se este valor fosse mais baixo). De modo simplificado, pode concluir-se que uma profundidade elevada da pobreza reflecte a existência de uma parte significativa da população extremamente pobre, enquanto uma profundidade baixa reflecte uma parte reduzida da população extremamente pobre, relativamente aos pobres “médios”.

Although this value measures the percentage of poor nothing is said, however, about the characteristics of the data of the sample concerning the poor households/persons. The following formula measures the depth of poverty, meaning that a bigger value represents one substantial part of the population is further away from the LdP than for a lesser value (many “very poor” has a higher value, few “very poor” has a lower value). In simplified way, it can be conclude that a high poverty gap index shows a bigger portion of the population being “extremely poor”, while a lower one shows a portion of the population being “extremely poor”.

$$P_1 = q / n * (z - y^p) / z$$

Where y^p is the mean of the poverty indicators (in our case household expenses) y_i among the poor.

The thirs formula for the squared poverty index:

$$P_2 = q / n * \left[(z - y^p) / z + (S_p / z)^2 \right]$$

Where S_p is the standard deviation of the y_i among the poor.

Perhaps the most important characteristic of these indices are the possibility of decomposition, or either, that the indices can be applied to the sub-groups (mutually exclusive sub-sets in the mathematical sense), for example, urban and rural households, and the indices if each group can be calculated independently. The total index can be calculated using a (weighted) mean of the sub-group indices.

6.2 Exercises:

Use SPSS to present the following results:

Use the following file **DadosAgregadosFamiliaresComIndividuaisCompleto.sav**

1. Calculate the expenditures per capita and the expenditures for Adult Equivalent.
Create two new variable and use [Transform/Compute]
2. Use SPSS and/or EXCEL to calculate the indices of poverty ($\alpha = 1, 2, 3$) in Sofala and for the following groups
 - a) Heads of households by gender (male and female)
 - b) “Big households” > 6 persons and “small households” ≤ 6
 - c) Urban and rural households

7 Regression Analysis



Linear Regression estimates the coefficients of the linear equation, involving one or more independent variables, that best predict the value of the dependent variable. For example, you can try to predict a salesperson's total yearly sales (the dependent variable) from independent variables such as age, education, and years of experience. Regression analysis is one of the most frequently used procedures to analyze relationships among statistical variables. This chapter offers a brief introduction into the basics of Regression Analysis and the introduction into the SPSS tools to apply these to some analysis of Poverty.

7.1 BACKGROUND:

What is (Linear) Regression:

Calculates the statistics for a line by using the "least squares" method to calculate a straight line that fits best your data, and returns an array that describes the line. Because this function returns an array of values, it must be entered as an array formula.

The equation for the line is:

$$y = m*x + b \text{ or } y = m^1*x^1 + m^2*x^2 + \dots + b \text{ (if there are multiple ranges of x-values)}$$

where

x^n are the independent or exogenous variables

y is the dependent or endogenous variable

b is the constant or intercept

m^n are the regression coefficients or predictors

There are several statistical indicators illustrating the 'quality' of the relation between the variables

Remember EXCEL:

Menu: [Tools/Data Analysis/Regression]

This is the summary output table, which includes an

- Anova (Analysis of Variance) table,
- coefficients,
- standard error of y estimate,
- R2 values,
- number of observations
- standard error of coefficients.

The main indicator for the quality, how well one (or many) variable(s) explain the dependent variable are the R and R squared values

Close to zero means little or no relation

Close to +1 means strong relation

7.2 SPSS - Linear Regression

Menu: [Analyze/ Regression]

Command Language: Regression

Select one numeric Dependent variable, and one or more numeric Independent variables.

You can control the entry of Independent variables into the analysis in two ways: by grouping them into blocks, and by choosing the method by which the variables in each block are processed. The available methods are *Enter*, *Remove*, *Stepwise*, *Backward*, and *Forward*. Starting with the first block, SPSS applies the selected method to all of the variables in the block, and then proceeds to the next block if there is one.

Select **Statistics** for additional statistics.

Select **Plots** for residual scatterplots, histograms, outlier plots, or normal probability plots.

Select **Save** to create new variables containing predicted values, residuals, and related statistics.

Select **Options** to change the criteria used in the stepwise methods, to request regression through the origin, or to control the treatment of missing value

- Estimates are the coefficients themselves.
- Confidence intervals are 95% confidence intervals for the coefficients
- Covariance matrix gives the variances and covariances among the coefficient estimates.
- *Descriptives* provides the means and standard deviations of each variable in the analysis, plus a correlation matrix (with one-tailed significance level and number of cases for each correlation).
- Model fit statistics include multiple R, R squared and adjusted R squared, standard error of the estimate, and an analysis-of-variance table.

additionally

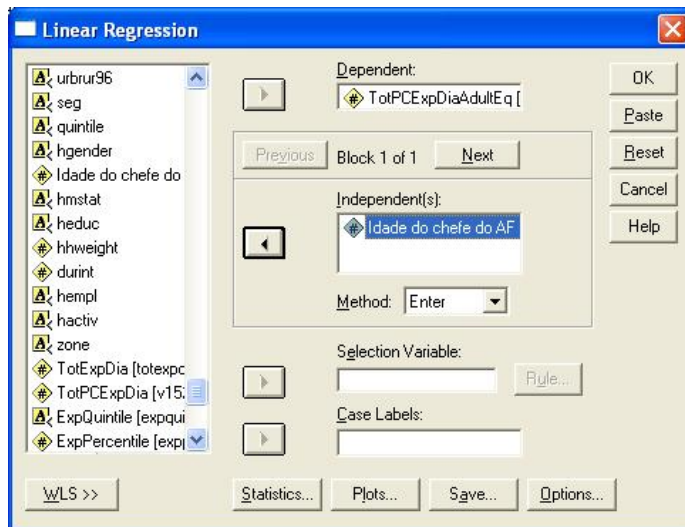
- Durbin Watson displays the Durbin-Watson test for serial correlation of the residuals.

7.3 An Example with SPSS and the IAF Data of Sofala

The Question: Is there a relation between the poverty and the age of the head of the household (AF)

Relate Expenses (the dependent variable and indicator of poverty - Name of the variable:

totpcexp) and the age of the head of household (Name of the variable: *b4*)



7.4 As características do modelo da regressão

The results of the output are shown below

Coefficients^a

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95% Confidence Interval for B | |
|-------|----------------------|-----------------------------|------------|---------------------------|--------|------|-------------------------------|-------------|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | -14839.1 | 85772.551 | | -.173 | .865 | -200139.382 | 170461.260 |
| | Idade do chefe do AF | 4297.214 | 2209.000 | .493 | 1.945 | .074 | -475.040 | 9069.469 |
| | HHSIZE | -17372.5 | 10108.764 | -.436 | -1.719 | .109 | -39211.127 | 4466.184 |

a. Dependent Variable: TotPCExpDiaAdultEq

The *Coefficients* are the regression coefficients

The *Confidence Intervals* are the 95% confidence intervals for the coefficients

The *Covariance matrix* gives the variances and Covariances between the estimated coefficients .

The *Descriptives* supplies the measures of central tendencies and standard deviations of each variable in the analysis and a matrix of the correlations.

Model Summary^b

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
|-------|-------------------|----------|-------------------|----------------------------|---------------|
| 1 | .334 ^a | .111 | .048 | 94451 | 1.722 |

a. Predictors: (Constant), Idade do chefe do AF
b. Dependent Variable: TotPCExpDiaAdultEq

Modelo com uma variável independente

The *Model summary* contains model statistics and it includes the R, the R squared and the estimate of the adjusted R, the standard error of the estimate and a table of the analysis of the variance.

In addition: the *Durbin Watson* shows the results of the Durbin-Watson test., a analysis of the residuals to test the autocorrelation between variables. One can assume a autocorrelation between variables if the DW is below 1.08 (positive correlation) or above 2.92(negative correlation)

Model Summary^b

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
|-------|-------------------|----------|-------------------|----------------------------|---------------|
| 1 | .525 ^a | .276 | .165 | 88479 | 1.341 |

a. Predictors: (Constant), HHSIZE, Idade do chefe do AF
b. Dependent Variable: TotPCExpDiaAdultEq

Modelo com duas variáveis independentes

Nota-se que o segundo modelo explica melhor a “Pobreza” de uma maneira mais satisfatória.

variable (Household size) explains " Poverty " in a more satisfactory way.

One notices that the second model, including another

The chicken example

To explain better the concepts of regression we use a very simple example. 16 hens in a “capoeira” lay eggs with the following weekly frequency. The proprietor of the hens

assumes that the age of the hens has to do with the number of eggs laid. He makes a table of the egg production of one week determined with eggs ranks the hens according to their egg production and includes the age of each hen in months.

We then calculate the regression with eggs as the dependent variable and age as the independent variable.

| | Ovos | Idade (meses) |
|-------|------|---------------|
| Gal1 | 0 | 2 |
| Gal2 | 0 | 1 |
| Gal3 | 1 | 4 |
| Gal4 | 1 | 7 |
| Gal5 | 1 | 10 |
| Gal6 | 2 | 14 |
| Gal7 | 2 | 12 |
| Gal8 | 2 | 9 |
| Gal9 | 2 | 6 |
| Gal10 | 3 | 8 |
| Gal11 | 3 | 10 |
| Gal12 | 4 | 23 |
| Gal13 | 4 | 8 |
| Gal14 | 4 | 9 |
| Gal15 | 5 | 12 |
| Gal16 | 5 | 18 |

We arrive at an equation of simple regression in the following form:

$$\text{Ovos} = 0.523 + 0.200 * \text{Idade}$$

The quality of the Regression ($R^2=0.4714$) is good but not exceptional. The residuals indicate a positive auto-correlation, showing a Durban-Watson coefficient < 1 . This however is not a problem because it is not about a

chronological series.

The P-P plots are used generally will determine if the distribution of a variable is in accordance with one determined type of distribution (normal in this case). If the variable selected will be in accordance with the distribution of the test, the points gather around a

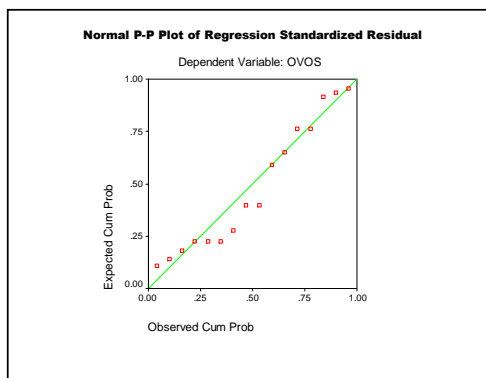
straight line, as this is the case of our

| Model | | | |
|--------------|----------------|-------------------------------|-------------|
| | Unstandardized | 95% Confidence Interval for E | |
| Coefficients | Coefficients | Lower Bound | Upper Bound |
| | B | | |
| (Constant) | 0.523 | -0.812 | 1.859 |
| IDADE | 0.200 | 0.079 | 0.322 |

Dependent Variable: OVOS

The inclination is the relation between the vertical change and the horizontal change of the line (plain for more variables) of regression. The intercept or the constant indicates where the line (plain) of the regression intercepts the Y axes; $x = 0$ (that is, when the expected value of the independent variable is 0, the intercept is to the distance of the origin until the a regression line).

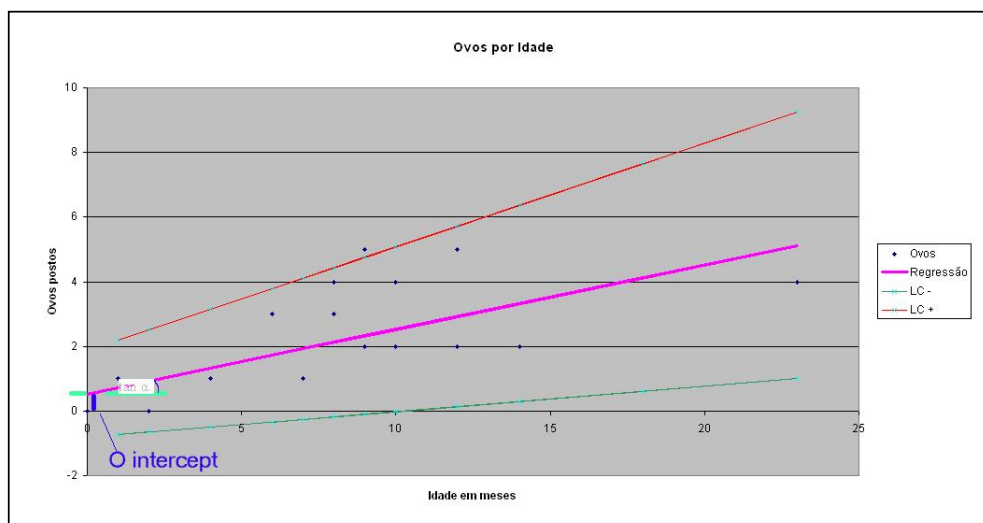
| Model Summary | | | |
|---------------|-------------------------------|----------|---------------|
| Model | R | R Square | Durbin-Watson |
| a | 0.6866 | 0.4714 | 0.7383 |
| b | Predictors: (Constant), IDADE | | |
| | Dependent Variable: OVOS | | |



regression.

A Graph of the Regression Equation

To explain this we include the data of the regression in a graph created with EXCEL



The eggs appear as diamond points in the graph. The fat central line in the graph indicates the equation of the regression. The Intercept is the distance between point 0 and the point where the line of the regression crosses the Y-axes. The coefficient of the regression ($x=0.2$) is the tangent of the angle between the line of the regression and X. axes, often interpreted as: While the dependent variable varies in one unit the independent variable in accordance to the coefficient of the variable.

In the example: with one month more of age, it is expected that the hens lay 0.2 more eggs. Getting 5 months older, it can be expected that a hen lays an extra egg.

To represent the errors in the model, extracts a vertical line from each point to the regression line. The length of these segments between the line and the points of the graph is called the residual and is an estimator for the errors in the model. SPSS uses the method of Least Square Estimates to calculate the inclination and intercept. This method minimizes the squared sums of residuals squares (that is, the sum of the squared vertical line segments).

In the equation, y is the dependent or the explained variable, what you are trying to predict; x is the independent variable or the predictor. The intercept and the inclination are coefficients of the model or the equation of the regression.

If the model will be a good explanation of the relation between the variables, you can use the estimates of the coefficients to forecast the value of the dependent variable for new cases.

In the example: It can be predicted how many eggs a hen puts with the age of 36 months. Use the formula: $Ovos = 0.523 + 0.200 * Idade$ and the result is 8 eggs. Already the weakness of the forecasts of the regression of this simple form is understandable. It does not take the life cycle of a hen into account. The old hen does not lay more eggs, even the opposite! Moreover, any statistical forecast is subjects the uncertainty of the static induction. They are the confidence intervals, that allow to calculate the lower confidence line (LC -) and the

upper (LC+) with the respective values for intercept and the coefficient; then the equation for lower confidence line is

$$\text{OvosLC-} = -0.812 + 0.079 * \text{Idade}$$

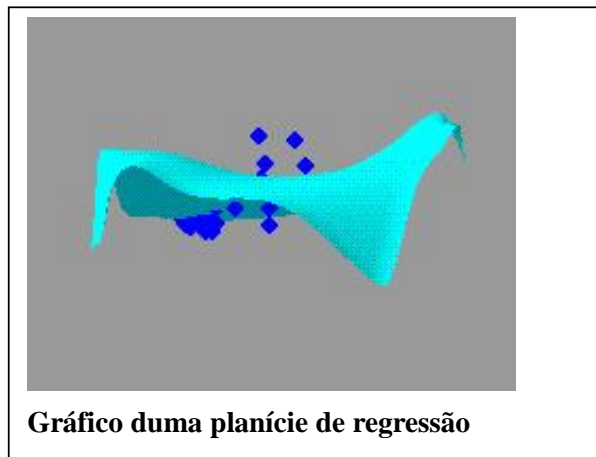
and the equation for upper confidence line is

$$\text{OvosLC+} = 1.859 + 0.322 * \text{Idade}$$

In the graph these lines appear to be sketched with a finer line. Depending on this sample, it can be expected that, for another hen and in 19 of 20 cases, the relation between eggs and age will fall into this confidence band. However in one case among between 20, this value will be outside this band. Another problem for a regression based forecast.

7.5 Models with two or more predictors

Adicionando uma segunda variável independente: A equação com uma variável independente é o modelo para a regressão linear simples; a equação com duas variáveis é



um modelo para a regressão múltipla. O SPSS permite que você inclua mais de duas variáveis independentes em correlações de uma regressão múltipla. Se a correlação de Pearson entre as duas variáveis for significativa com um valor menos de 0.1 que indica a hipótese que a correlação é 0 (nenhuma relação linear entre as variáveis) está rejeitado. Quando O seu modelo inclui mais do que uma variável independente e você escolhe *Colinearity diagnostics* na opção [Statistics..] SPSS indica correlações

para todos os pares das variáveis

Adding one second independent variable: The equation with one independent variable is the model for simple the linear regression; the equation with two or more variable is a model for the multiple regression. SPSS allows that you it more than includes as many independent variable as you like for a multiple regression. This does not necessarily improve the quality (R^2) of the model.

If the correlation of Pearson between the two independent variables will be significant with a value less than 0,1, the hypothesis that the correlation is 0 (no linear relation between the variables) is rejected. When its model includes more than one independent variable, you choose *Colinearity diagnostics* in the option [Statistics..] SPSS indicates correlations for all pairs of the variables

The graph of this model with two variables already is not a line any more but a plain. For more dimensions (more variables) it is even more difficult to visualize plains of the regression.

Normality is not required for the estimates of the coefficients. To make tests and intervals of the confidence of the estimate, however, these subsequent assumptions are required:

- the errors are distributed with normal distribution and an average of 0
- the errors have a constant variance.
- the errors are independent, not correlated among themselves.

These assumptions are verified studying the residual ones of the model. The statistics of Durbin-Watson (available in the linear regression: in the sub-dialogue of [Statistics..]) can be used to test the correlation of series of residual of adjacent variables.

Model Summary. The value of R (also called multiple R). When there is only one independent variable, R is the simple correlation between the independent and dependent variable (see the Same correlation in the Correlations table above). R^2 is the square of this value and often is interpreted as the proportion of the total variation of the dependent variable accounted for by the independent variable (fertilizer consumption “explains“ ???% of the variability of life expectancy).

R and R^2 ranges from 0 to 1. If there is no *linear* relation between the dependent and independent variable, R^2 is 0 or very small. If all the observations fall on the regression line, R^2 is 1. This measure of the goodness of fit of a linear model is also **called the coefficient of determination**.

A caution. Be careful about concluding, “If fertilizer consumption is decreased, the population will live longer.“ There may be an *association* between fertilizer consumption and life expectancy. However, these data come from an observational study, not a controlled experiment; so any statements about cause-and-effect relationships can be misleading. Association is not the same as causation..

If an investigator observes the values of the independent and dependent variables for a set of subjects (cases), association does not establish causation. If an investigator does an experiment where he or she sets the values of the independent variable (for example, six specific doses of a drug) and watches the effect on the dependent variable, there may be little question that the results were *caused* by the independent variable.

7.6 Exercises

Use the following file of the IAF of Sofala

DadosAgregadosFamiliaresComIndividuaisCompleto.sav

1. Use SPSS calculating the equation of a linear regression for the Sofala data series to relate: Poverty / expenses (the dependent variable) and gender of head of household (B2) and highest educational level (C3),
2. Try to improve the model, by using employment variables (E..)
3. Use variables about housing characteristics (F..). Are these exogenous or endogenous variables? Discuss the results and the validity of these results
4. Calculate a line of the poverty on the base of the IAF, independent of the known values. Discuss the criteria to establish this measure. Which expenditures must be considered and in which dimension? Calculate three groups of Wealth/Poverty Rich, Medium, Poor
5. Use forecasts for the groups calculated in the exercise 4.) of calculating the line of poverty and wealth groups with the use of "the best " predictors (variables that explain better than others).Discuss and present the results.
6. Calculates the forecasts of these indices for some sub-samples:
 - a. The AF with female heads,
 - b. Urban and Rural
 - c. Grouped by Districts

8 Bibliography

8.1 SPSS

SPSS 10.0 Manuals - SPSS Inc. Headquarters, 233 S. Wacker Drive, 11th floor Chicago, Illinois 60606

SPSS 10.0 Regression Models ISBN 0130179043

SPSS 10.0 for Windows Student Version ISBN 0130280402

Discovering Statistics Using SPSS for Windows : Advanced Techniques for Beginners (Introducing Statistical Methods series) by Andy Field (Paperback)

SPSS/PC+ Basics and Graphics - Gerhard Brosius McGrawHill ISBN 3-89028-132-X

SPSS/PC+ Advanced Statistics and Tables Gerhard Brosius McGrawHill ISBN 3-89028-157-5

SPSS Base System und Professional Statistics - Gerhard Brosius, Felix Brosius International Thomson Publishing ISBN 3-929821-62-1

8.2 Statistics, Poverty Analysis

How to Lie With Statistics by Darrell Huff, Irving Geis

Deaton, A., and S. Zaidi. 1999. *Guidelines for constructing consumption aggregates for welfare analysis*. Research Program in Development Studies Working Paper No. 192. Princeton, N.J.: Princeton University.

J.-L-Dubois, D.Blaizeau - Connaître les conditions de vie de ménages dans les pays en développement –Tome 3: Analyser les résultats- - Ministère de la Coopération et du développement- ISBN 2-11-884855-3

Foster, J., J. Greer, and E. Thorbecke. 1984. A class of decomposable poverty measures. *Econometrica* 52 (3): 761-765.

Ravallion, M. 1994. *Poverty comparisons*. Chur, Switzerland: HarwoodAcademic Publishers.

-' 1998. *Poverty lines in theory and practice*. Living Standards Measurement Study Working Paper No. 133. Washington, D.C.: World Bank.

-' 2001. *Growth, inequality and poverty: Looking beyond averages*. Policy Research Working Paper No. 2558. Washington, D.C.: World Bank.

Tarp, E, C. Arndt, H. T. Jensen, S. Robinson, and R. Heltberg. 2002a. *Facing the development challenge in Mozambique: An economywide perspective*. Research Report 126. Washington, D.C.: International Food Policy Research Institute.

Tarp, F., K. R. Simler, C. Matusse, R. Heltberg, and G. Dava. 2002b.. *Economic Development and Cultural Change* 51 (1): 77-108.

8.3 WWW

<http://www.spss.com/>

http://www.ats.ucla.edu/stat/spss/library/sp_hist.htm