

# Introduction to SPSS 19.0

## Authors:

Nicholas Fritsche

Casper Voigt Rasmussen

Morten Christoffersen

Niels Yding Sørensen

Jesper Pedersen

Rasmus Porsgaard

Martin Klint Hansen

Ulrick Tøttrup

Morten Mondrup Andreassen

Rasmus Maarbjerg

## Last updated:

June 2012

# Table of contents

1. INTRODUCTION .....	1
2. SPSS IN GENERAL .....	2
2.1 Data Editor .....	2
2.2 FILE-menu .....	2
2.3 EDIT-menu .....	2
2.4 VIEW-menu .....	2
2.5 DATA-menu .....	2
2.6 TRANSFORM-menu .....	2
2.7 ANALYZE-menu .....	2
2.8 GRAPHS-menu .....	3
2.9 UTILITIES-menu .....	3
2.10 HELP-menu .....	3
2.11 Output .....	3
2.11 Syntax editor .....	4
2.12 Chart editor .....	5
3. DATA ENTRY .....	6
3.1 Manual data entry .....	6
3.1.1 Making a new dataset .....	6
3.1.2 Open an existing dataset .....	7
3.2 Import data .....	8
3.2.1 Import data from Excel, SAS, STATA etc. ....	8
3.2.2 Import of text files .....	8
3.3 Export data .....	8
3.4 Dataset construction .....	8
4. DATA PROCESSING .....	10
4.1 Data menu .....	10
4.1.1 Defining dates (time series analysis) .....	10
4.1.2 Sorting observations .....	10
4.1.3 Transposing of data .....	10
4.1.4 Aggregation of data (in relation to a variable) .....	10
4.1.5 Splitting files .....	11
4.1.6 Select cases .....	12
4.1.7 Weight Cases .....	13
4.2 Transform .....	14
4.2.1 Construction of new variables .....	14
4.2.2 Count numbers of similar observations .....	15
4.2.3 Recode variables .....	17
4.2.4 Ranking Cases .....	18
4.2.5 Automatic Recode .....	18
4.2.7 Construction of time series .....	20
4.3 Recode (join) .....	21

4.3.1 Join using the dialog box .....	22
4.3.1.1 Recode into Same Variables .....	22
4.3.1.2 Recode into Different Variables .....	23
4.3.2 Recoding using the syntax .....	23
4.4 Missing values .....	24
5. CUSTOM TABLES .....	26
5.1 Custom Tables output .....	27
6. TABLES OF FREQUENCIES AND CROSSTABS .....	28
6.1 Custom Tables .....	28
6.1.1 Table of frequencies output .....	29
6.2 Crosstabs .....	30
7. DESCRIPTIVES .....	32
7.1 Output for Descriptive Statistics .....	32
8. FREQUENCIES .....	33
8.1 Frequencies output .....	34
9. PLOTS .....	36
9.1 Histograms .....	36
9.2 Chart Editor .....	36
9.3 Reference line .....	37
9.4 Trend Line .....	38
9.5 Editing Scales .....	39
10. TEST OF NORMALITY, EXTREME VALUES AND PROBIT-PLOT .....	40
10.1 Explore output .....	41
11. CORRELATION MATRICES .....	42
11.1 Correlation matrix .....	42
11.2 Bivariate Correlation output .....	43
12. COMPARISONS AND TEST OF MEANS .....	44
12.1 Compare means .....	44
12.2 One sample T-test .....	44
12.2.1 Output .....	45
12.3 Independent samples T-Test .....	45
12.3.1 Output .....	46
12.4 Paired Samples T-Test .....	47
12.4.1 Output .....	47
13. ONE-WAY ANOVA .....	49
13.1 Output .....	50
14. GENERAL ANALYSIS OF VARIANCE .....	53
14.1 GLM output .....	57
14.2 Test of assumptions .....	59
14.2.1 Homogeneity of variance .....	59
14.2.2 Normally distributed errors .....	60
14.2.3 Independent errors .....	62

15. REGRESSION ANALYSIS .....	64
15.1 Test of design criteria .....	69
15.1.1 Zero mean: $E(\varepsilon_i) = 0$ for all $i$ .....	69
15.1.2 Homoscedasticity: $\text{var}(\varepsilon_i) = \sigma^2$ for all $i$ .....	69
15.1.3 Mutually uncorrelated: and $\varepsilon_j$ uncorrelated for all $i \neq j$ .....	73
15.1.4 Uncorrelated with $x_1, \dots, x_k$ : $\varepsilon_i$ and $x_j$ are uncorrelated for all $i$ and $j$ .....	73
15.1.5 Normality: $\varepsilon_i \sim \text{i.i.d.} - N(0, \sigma^2)$ for all $i$ .....	74
15.2 Further Topics .....	76
15.2.1 LM test for Heteroscedasticity .....	76
15.2.1.1 Output .....	77
15.2.2 WLS .....	78
16. LOGISTIC REGRESSION .....	79
16.1 The procedure .....	79
16.2 The output .....	80
17. TEST FOR HOMOGENEITY AND INDEPENDENCE .....	81
17.1 Difference between the tests .....	81
17.2 Construction of the dataset .....	81
17.3 Running the tests .....	82
17.4 Output .....	84
17.5 Assumptions .....	85
18. FACTOR .....	86
18.1 Introduction .....	86
18.2 Example .....	86
18.3 Implementation of the analysis .....	87
18.3.1 Descriptives .....	88
18.3.2 Extraction .....	89
18.3.3 Rotation .....	90
18.3.4 Scores .....	90
18.3.5 Options .....	91
18.4 Output .....	91
19. CLUSTER ANALYSIS .....	95
19.1 Introduction .....	95
19.2 Hierarchical analysis of clusters .....	95
19.2.1 Example .....	95
19.2.2 Implementation of the analysis .....	96
19.2.2.1 Statistics .....	97
19.2.2.2 Plots .....	98
19.2.2.3 Method .....	98
19.2.2.4 Save .....	99
19.2.3 Output .....	100
19.3 K-means cluster analysis (Non-hierarchical cluster analysis) .....	102
19.3.1 Example .....	102
19.3.2 Implementation of the analysis .....	102

19.3.2.1 Iterate .....	104
19.3.2.2 Save.....	104
19.3.2.3 Options.....	105
19.3.3 Output.....	105
20. NON-PARAMETRIC TESTS .....	108
20.1 Cochran's Test.....	108
20.2 Friedman's Test.....	109
20.3 Kruskal Wallis Test.....	111

# 1. Introduction

The purpose of this manual is to give insight into the general use of SPSS. Different basic analyses will be described, which covers the statistical techniques taught at the bachelor level. Further more difficult techniques, which are taught and used at the master level, can be found in the manual: *Cand.Merc. Manual for SPSS 16*.

This manual only gives examples on how to do statistical analysis. This means that it does not give any theoretical justification for using the analysis described. It will only be of a descriptive nature where you can read how concrete problems are solved in SPSS. Where found necessary there are made references to the literature used on the statistics course taught on the bachelor level. Here you can find elaboration of the statistical theory, used in the examples. The following references are used in the manual:

**Keller (2009)** : Currently used textbook Keller. Managerial Statistics. 8e. 2009.

Intern undervisningsmateriale E310 "Notesamling til Statistik" 2011.

The examples in this manual are based on various datasets. The folder containing these can be downloaded through the following link:

[http://www.studerende.au.dk/fileadmin/www.asb.dk/servicekatalog/IT/Analysevaerktoejer/SPSS/SPSS\\_Manual\\_Files.zip](http://www.studerende.au.dk/fileadmin/www.asb.dk/servicekatalog/IT/Analysevaerktoejer/SPSS/SPSS_Manual_Files.zip)

Most of the examples are based on the *Rus98eng.sav* dataset. If another dataset is used, it will be mentioned. The different datasets can be found in the same folder as the dataset mentioned above.

The following changes and modifications have been made in this new version

- Some screenshots and descriptions have been updated to SPSS version 19.0
- All known errors have been corrected

Any reports on errors in the manual can be addressed to [analytics@asb.dk](mailto:analytics@asb.dk).

## 2. SPSS in General

SPSS consists of four windows: A Data Editor, an Output window, a Syntax window and a Chart Editor. The Data Editor is further divided into a *Data view* and a *Variable view*. In the Data Editor you can manipulate data and make commands. In the *Output window* you can read the results of the analysis and see graphs and then it also works as a log-window. In the Chart Editor you can manipulate your graphs while the *syntax window* is used for coding your analysis manually.

### 2.1 Data Editor

At the top of the Data Editor you can see a menu line, which is described below:



### 2.2 FILE-menu

The menu is used for data administration, this means opening, saving and printing data and output. All in all you have the same options as for all other Windows programs.

### 2.3 EDIT-menu

Edit is also a general menu, which is used for editing the current window's content. Here you find the CUT, COPY, and PASTE functions. Furthermore it is possible to change the font for the output view and signs for decimal (place) when selecting OPTIONS.

### 2.4 VIEW-menu

In the VIEW menu it is possible to select or deselect the Status Bar, Gridlines etc. Here you also change the font and font size for the Data Editor view.

### 2.5 DATA-menu

All data manipulation is done in the *Data menu*. It is possible to manipulate the actual data in different ways. For instance you can define new variables by selecting *Define Variables...* sort them by selecting *Sort Cases...* etc. A further description of these functions can be found in chapter 4 (Data processing).

### 2.6 TRANSFORM-menu

Selecting the Transform menu makes it possible to recode variables, generate randomized numbers, rank cases, define missing values etc.

### 2.7 ANALYZE-menu

This is the "important" menu, where all the statistical analyses are carried out. The table below gives a short description of the most common methods of analysis.

Method	Description
<b>Reports</b>	Case- and report summaries
<b>Descriptive statistics</b>	Descriptive statistics, frequencies, plots etc.
<b>Tables</b>	Construction of various tables
<b>Compare Means</b>	Comparison of means. E.g. by using t-test and ANOVA
<b>General Linear Model</b>	Estimation using GLM and MANOVA
<b>Generalized Linear Model</b>	Offers an extension of the possibilities in Regression and General Linear Model. I.e. estimation of data that is not normally distributed and regressions with interaction between explanatory variables.
<b>Mixed Models</b>	Flexible modeling which includes the possibility of introducing correlated and non-constant variability in the model.
<b>Correlate</b>	Different associative measures for the variables in the dataset.
<b>Regression</b>	Linear, logistics and curved regression
<b>Loglinear</b>	General log-linear analysis and Logit.
<b>Classify</b>	Cluster analysis.
<b>Data Reduction</b>	Factor.
<b>Scale</b>	Item analysis and multidimensional scaling.
<b>Nonparametric Tests</b>	$\chi^2$ binominal, hypothesis and independent tests.
<b>Time Series</b>	Auto regression and ARIMA.
<b>Survival</b>	Survival analysis.
<b>Multiple response</b>	Table of frequencies and cross tabs for multiple responses.
<b>Missing Value Analysis</b>	Describes patterns of missing data.

## 2.8 GRAPHS-menu

If a graphical overview is desired the menu *Graphs* is to be used. Here it is possible to construct histograms, line, pie, and bar charts etc.

## 2.9 UTILITIES-menu

In this menu it is possible to get information about type and level for the different variables. If for some reason it is not desired to directly use the data for the variables given in the editor, then it is possible to construct a new dataset using the existing variables. This is done under *Utilities -> Define variable sets*. It will then in the future be possible to use the new dataset constructed. This is done through *Utilities -> Use Variable sets*.

## 2.10 HELP-menu

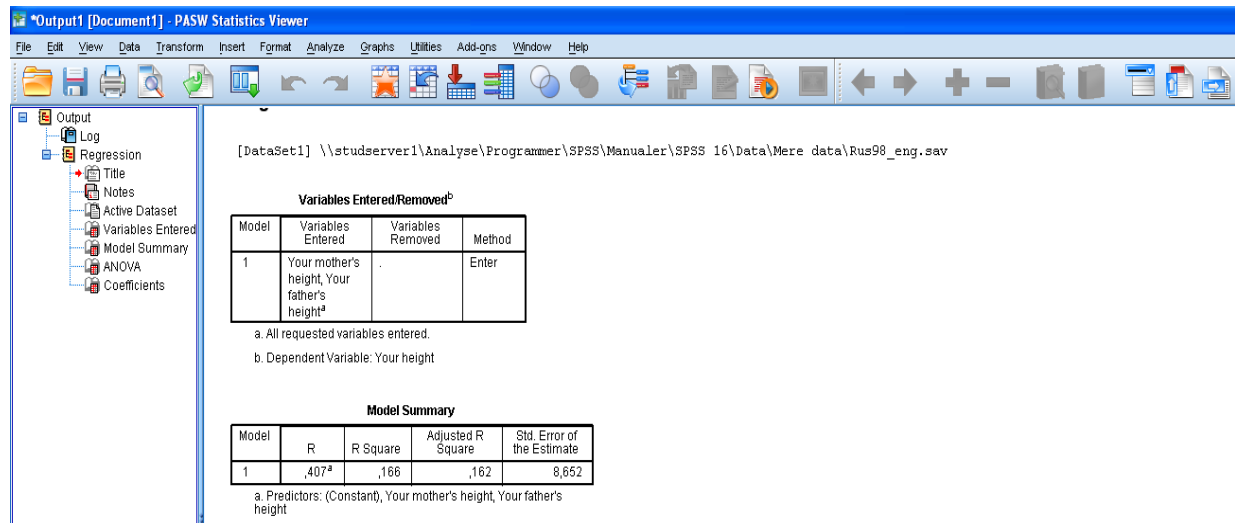
In the help menu it is possible to search for help about how different analysis, data manipulations etc. are done in SPSS. The important menu is *Topics* where you can enter keywords to search for.

## 2.11 Output

The output window works the same way as just described in section 2.1 and 2.2. Though in the *Edit* menu there is a slight difference; the *Copy Objects* option. This function is recommendable when tables and like are to be copied from SPSS into another document. By using this function the copied object keeps its original format!



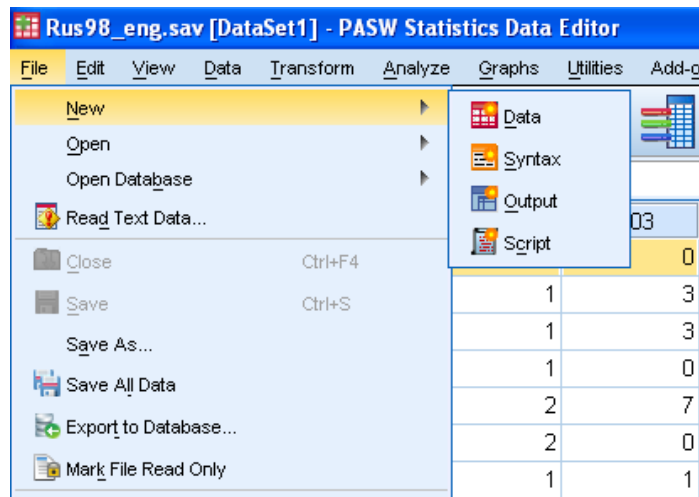
As mentioned earlier the Output window prints the results and graphs generated through the analysis and it also functions as a log-menu. You can switch between the *Editor* and *Output window* under the menu *Window*. The *output window* is constructed to give a very good over-view letting the user (de)select the different menus to be seen.



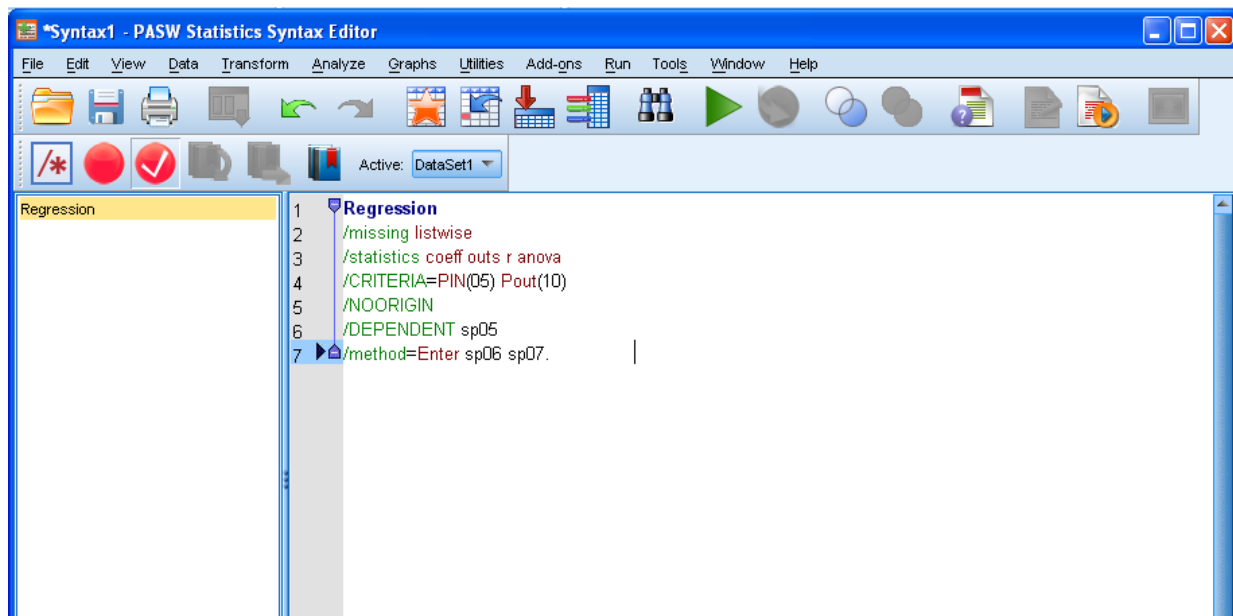
To see the output from an analysis simply double click on the analysis of interest in the menu on the left side of the screen, and the results will appear in the right hand side of the screen. If errors occur, a log menu will appear. As can be seen from the above window there is a sub-menu called *Notes*. In this submenu you find information about the time the analysis was performed, and under what conditions. By default this menu is not visible, but by double clicking it, you can open and review it. Moreover SPSS prints the syntax code for the selected tests in the output window. The syntax code can in this way be reused and altered for additional analysis. Furthermore the syntax code can be used to document the way in which the analysis has been done.

## 2.11 Syntax editor

The syntax editor is the part of SPSS where the user can code more advanced analyses, which might not be available in the standard menu. This function works pretty much like the statistical program SAS. To open the syntax window you select *File => New => Syntax*



When the syntax option is selected an empty window will show on the screen.



In the window you can enter the program code you want SPSS to perform. Here the code from the regression seen above is typed in. When the code is ready to be run you highlight it (with your mouse) and select *Run => Selection* or press the *Play* button in the menu bar.

When carrying out a mean based analysis you can always see the related syntax by clicking on the paste key in the analysis window.

## 2.12 Chart editor

The chart editor is used when editing a graph in SPSS. For further detail, on how to do this, see chapter 9.

## 3. Data entry

There are two ways to enter data into SPSS. One is to manually enter these directly into the *Data Editor* the other option is to import data from a different program or a text file. Both ways will be described in the following.

### 3.1 Manual data entry

Entering data manually into SPSS can be done in two ways. Either you can make a new dataset or you can enter data into an already existing dataset. The latter option is often useful when solving statistical problems based on dataset already available.

#### 3.1.1 Making a new dataset


Entering data into SPSS is very simple since the way to do it is similar to the way you enter data into e.g. Excel. Rows equal observations and columns equal variables. This means that the left column in the dataset (which is always grey) is the observation numbers and the top row (also grey) is the variable names. This is illustrated in the below figure where there are two variables; *VAR00001* and *VAR00002*. These two variables have 9 observations, which e.g. could be the year 1990-1998 or 9 respondents.

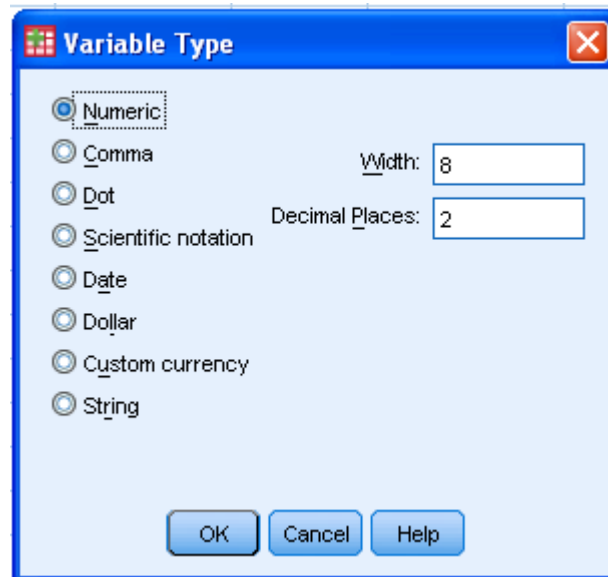
	VAR00001	VAR00002	var
1	3,00	2,00	
2	2,00	3,00	
3	2,00	2,00	
4	2,00	4,00	
5	2,00	5,00	
6	4,00	4,00	
7	3,00	3,00	
8	5,00	3,00	
9	6,00	3,00	
10	.	.	
11			

When SPSS is opened, the *Data Editor* is automatically opened and this is where you enter your data. Alternatively you could choose *File => New => Data*.

Before you start entering your data it would be a very good idea first to enter a name and de-fine your variables. This is done by selecting *Variable view* in the bottom left corner. Alternatively you can double click the variable and the result will look almost like you can see below:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	VAR00001	Numeric	8	2		None	None	8	Right	Scale
2	VAR00002	Numeric	8	2		None	None	8	Right	Scale

As you can see it is now possible to name the variables. Under *Type* you define which type your variable is (numeric, string etc.) By placing the marker in the *Type* cell, a button like this:  appears. This button indicates that you can click it and a window like below will show:



*Numeric* is selected if your variable consists of numbers. *String* is selected if your variable is a text (man/woman). The same way you can specify *Values* and *Missing*.

By selecting *Label* you get the option to further explain the respective variable in a sentence or so. This is often a very good idea since the variable name can only consist of 8 characters. *Missing* is selected when defining if missing values occur among the observations of a variable.

In *Values* you can enter a label for each possible response value of a discrete variable (e.g. 1 = man and 2 = woman).

When entering a variable name the following rules must be obeyed in SPSS for it to work:

- The name has to start with a letter and not end with a full stop (.).
- Do not enter space or other characters like e.g. !, ?, ', and \*.
- No two variable names must be the same.
- The following names is reserved for SPSS use and cannot be used:

ALL	NE	EQ	TO	LE	LT	BY
OR	GT	AND	NOT	GE	WITH	

When all variable names are entered and defined you can start entering your data. This is done in the "Data view" where you put your cursor in the cell you want to enter your data. When all data is entered you select *File => Save As...* in the menu to save your new dataset.

### 3.1.2 Open an existing dataset

If the dataset already exists in a SPSS file you can easily open it. Select *File => Open...* and the dataset will automatically open in the *Data Editor*.

## 3.2 Import data

Sometimes the data is available in a different format than a SPSS data file. E.g. the data might be available as an Excel, SAS, or text file.

### 3.2.1 Import data from Excel, SAS, STATA etc.

If you want to use data from an Excel file in SPSS there are two ways to import the data. (1) One is to simply mark all the data in the Excel window (excluding the variable names) you want to enter into SPSS. Then copy and paste them into the SPSS data window. The disadvantage by using this method is that the variable names cannot be included meaning you will have to enter these manually after pasting the data. (2) The other option (where the variable names are automatically entered) is to do the following:

- 1) Open SPSS, select *Files => Open => Data*.
- 2) Under *Files of type* you select Excel, press 'Open', and the data now appear in the *Data Editor* in SPSS.

### 3.2.2 Import of text files

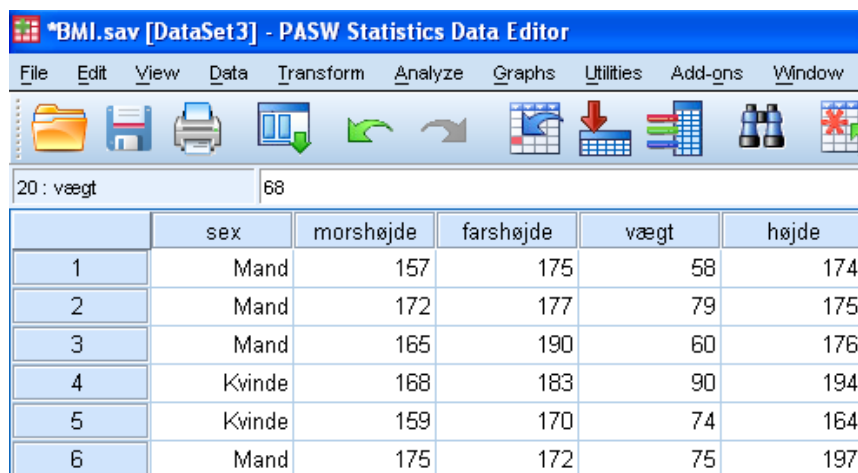
Importing text files requires that the data are separated either by columns or a different separator like tab, space, full stop etc. Importing is done by selecting *Read Text Data* in the *File* menu. You will then be guided through how to specify how the data are separated etc.

## 3.3 Export data

Exporting data from SPSS to a different program is done by selecting *File => Save As...* Under *Save as type* you select the format you want the data to be available in e.g. Excel.

## 3.4 Dataset construction

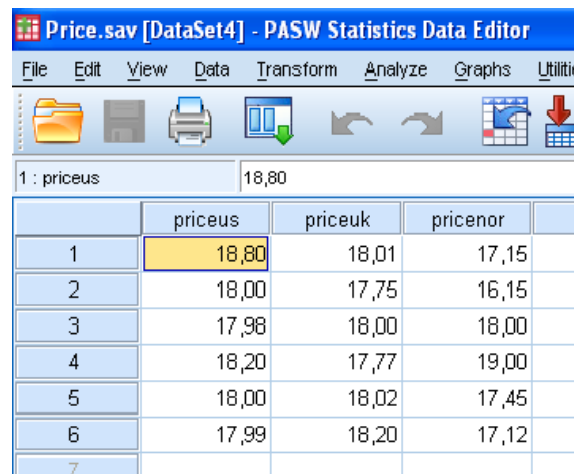
When you want to use your dataset in different statistical analyses it's important to construct the actual dataset in the right way in order to be able to carry out the analysis. You have to keep in mind that the construction of the dataset depends on which analyses you want to perform. In most analyses you have both a dependent and one or more independent variables. When you want to make an analysis each of these different variables must be separated as shown below.



	sex	morshøjde	farshøjde	vægt	højde
1	Mand	157	175	58	174
2	Mand	172	177	79	175
3	Mand	165	190	60	176
4	Kvinde	168	183	90	194
5	Kvinde	159	170	74	164
6	Mand	175	172	75	197

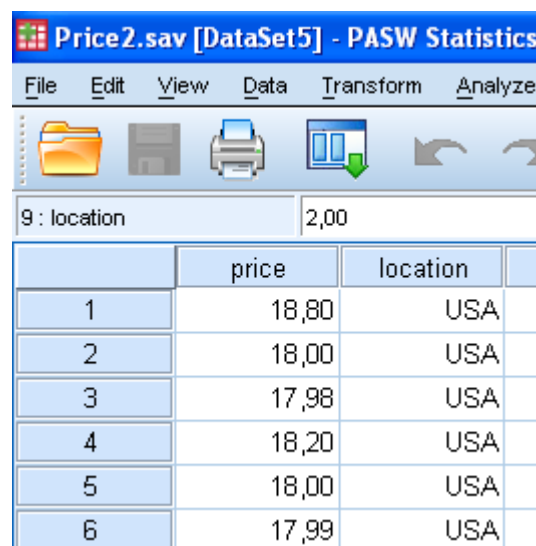
In this example a possible analysis could be a regression where you would predict a persons weight by his height. In these kind of analyses it's necessary that each variable is separated so they can be defined as dependent and independent variables respectively.

If you want to do other kinds of analyses it's often necessary to construct your dataset in another way, by using a grouping variable. This is often the case in experimental analysis, where you are measuring a variable under different treatments. An example could be that you have measured some price index's in different countries, and want to test whether there is any statistical differences between them. To do this you have to construct your dataset, so you'll have a grouping variable containing information about which country the price index is from. Below the earlier mentioned construction method is shown. This construction is mainly used in the regression analysis.



	priceus	priceuk	pricenor
1	18,80	18,01	17,15
2	18,00	17,75	16,15
3	17,98	18,00	18,00
4	18,20	17,77	19,00
5	18,00	18,02	17,45
6	17,99	18,20	17,12
7			

The dataset as it should be constructed is shown below. Here you have the grouping variable containing information about which country the price index is from. This construction is used in most other analyses such as T-test and analysis of variance.



	price	location
1	18,80	USA
2	18,00	USA
3	17,98	USA
4	18,20	USA
5	18,00	USA
6	17,99	USA

As you can see, the construction of dataset depends on which analysis you want to carry out.

## 4. Data processing

When processing data, two menus are of high importance; *Data* and *Transform*. In the following the functions that are used most frequently under these menus will be described.

### 4.1 Data menu

Global transformations to the SPSS dataset are done in the data menu. This might be transformations like transposing variables and observations, and dividing the dataset into smaller groups.

#### 4.1.1 Defining dates (time series analysis)

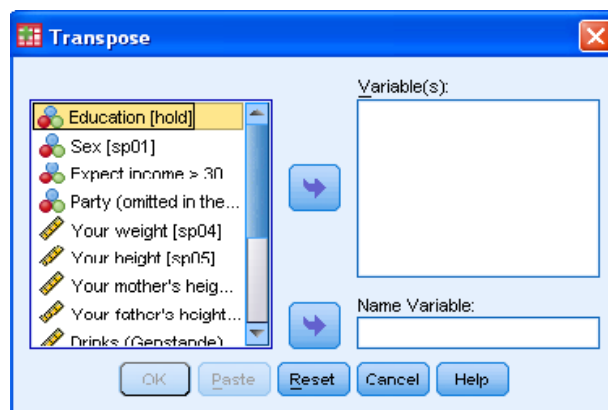
Under the menu *Define Dates...* it is possible to create new variables, which define a new continuous time series that can be used for a time series analysis. After having defined which time series the observations follow, you click 'OK' and a new variable will automatically be constructed.

#### 4.1.2 Sorting observations

Sorting observations based on one or more variable is done using the menu *Sort Cases...* It should be noted that when sorting the dataset, you could easily run into trouble if a later analysis of time series is to be done. This problem can be solved by making observation numbers as shown above, before sorting the cases.

#### 4.1.3 Transposing of data

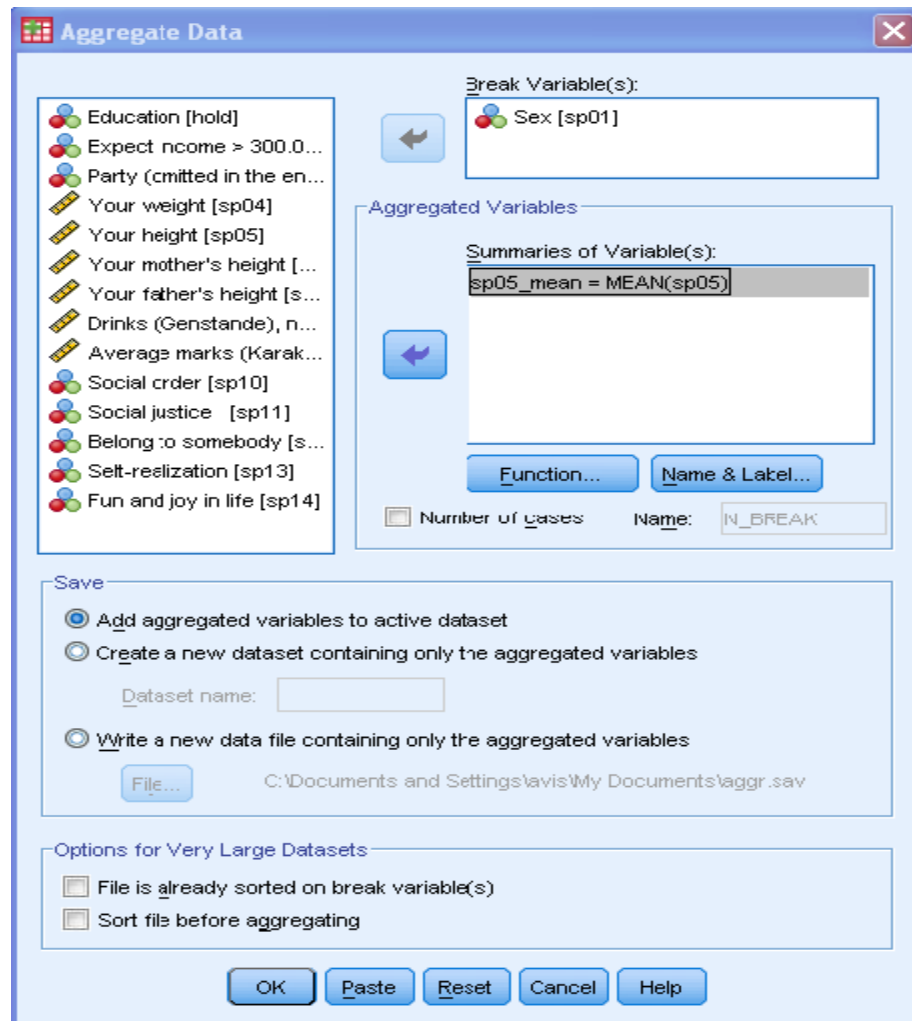
Transposing data, so that the columns turn into rows and the other way around, is done using the menu *Transpose...*



Those variables you want to include in the new dataset should be marked in the left window. By clicking the top arrow they are moved to the top right window where you can see all the variables included. In the field *Name Variable* you can enter a variable containing a unique value if you want the output to be saved as a new variable.

#### 4.1.4 Aggregation of data (in relation to a variable)

In the menu *Aggregate* it is possible to aggregate observations based on the outcome of a different variable. For instance, if you have a dataset obtaining the height and sex of several respondents, an aggregation of the variable sex, would result in a new dataset. In this new dataset each observation states the average height of each sex – meaning one observation for each sex. When selecting *Aggregate...* the following window appears:



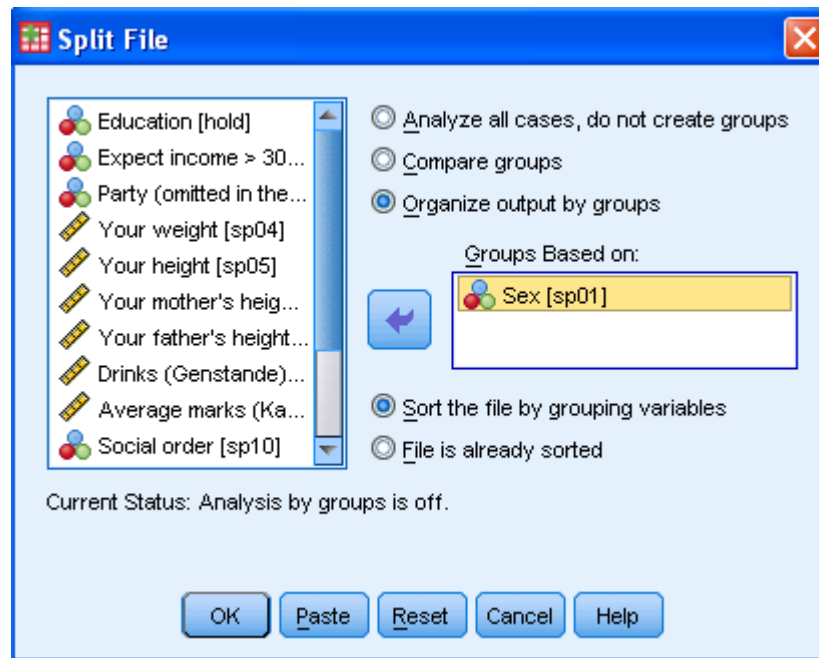
The variables you want aggregation for, is to be moved to the *Break Variables(s)* (In the ex-ample shown above it would be the variable *Sex*). Those variables that are to be aggregated should then be moved to the *Aggregate Variable(s)* (the variables *age* and *Height*). In the 'Function...' you must define which statistical function to be used for aggregating the variables. Names of new variables can be defined by clicking 'Name & Label...'.

If you mark *Number of cases ...* a new variable will appear which includes the number of observations that are aggregated for each variable. Finally you need to decide where the new file should be saved. (This is done using the bottom menu.)

#### 4.1.5 Splitting files

The menu *Split Files* splits data files into two or more. This means that each time a new test is performed, not one output will be shown but instead the number of outputs will correspond to the number of possible outcome for each split group.

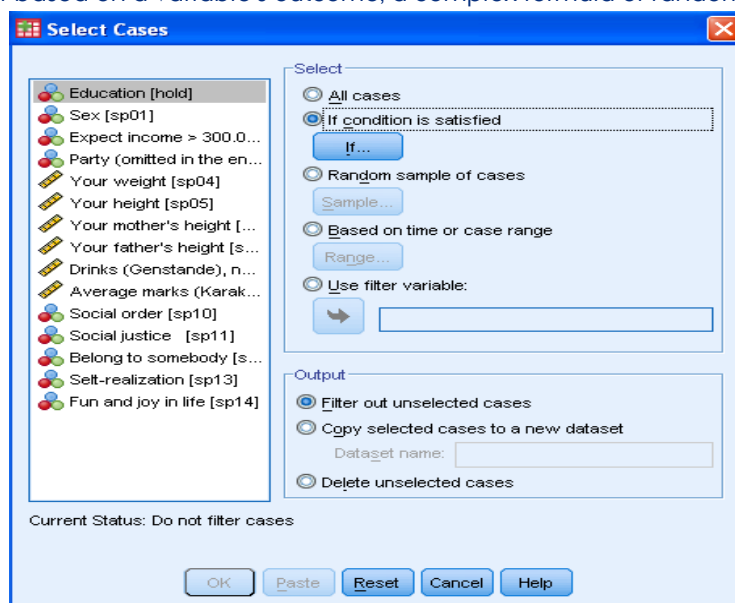




If you choose to group using more than one variable the output will first be grouped by the variable shown in the top of the list, then further grouped by the next into subgroups and so forth. Note that you at most can group by 8 variables. Also note that if the observations are not sorted in the same way you want to group them you need to *mark Sort the file by grouping variables*. By marking the *Compare Groups* button the split files will be presented together so it is easier to compare these. By clicking the *Organize output by groups* button the split files will not be presented together.

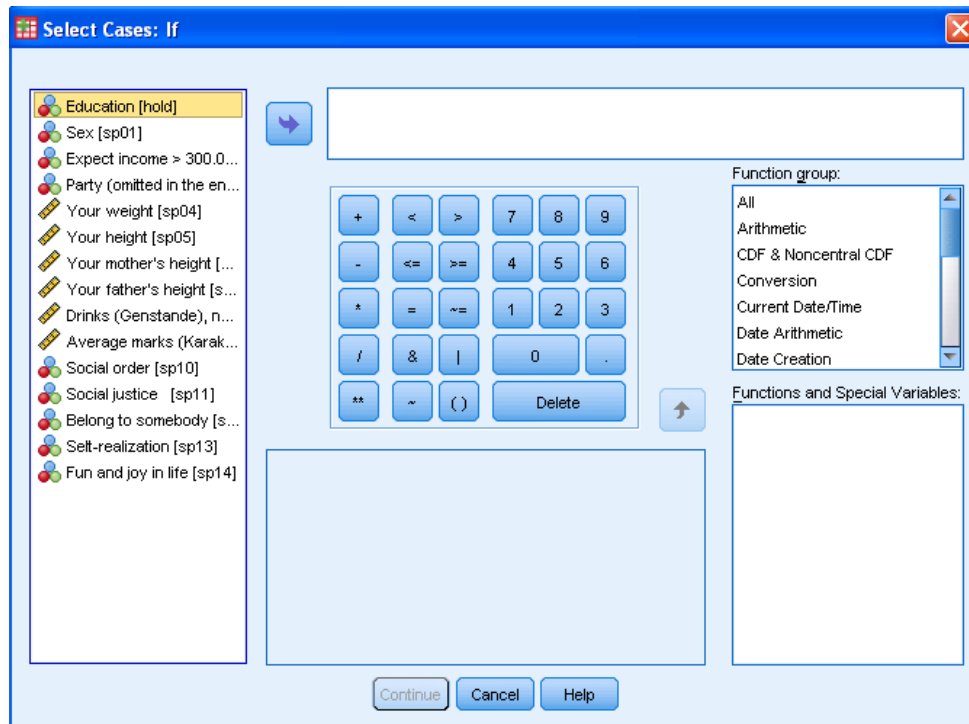
#### 4.1.6 Select cases

In the *Select Cases* different methods are presented to include only observations that fulfill a certain criteria. These criteria are either based on a variable's outcome, a complex formula or random selection.



In the window shown above you can see the different functions for selecting data.

Choosing the 'if...' button it is possible to make a complex selection of the observations to be included in the analysis. By clicking the button you get the following window:



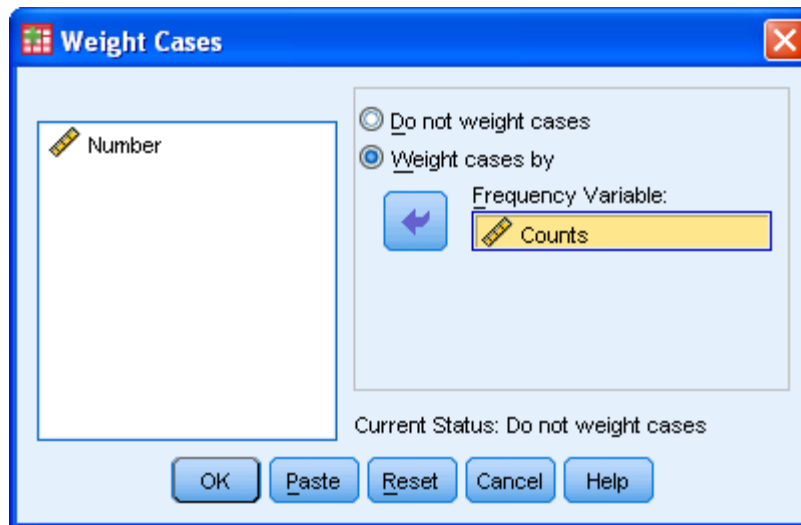
It is possible now to specify which observations you would like to select. This is simply done by writing a mathematical function, where the observations you want to be included fulfill the function's criteria.

The other buttons are very similar to the above described 'if...' button and therefore they will not be described.

The last thing you must do is to decide whether the data you have excluded from the selection should be deleted or just filtered – we suggest that you filter your data because this way you can always correct your selection. By filtering, SPSS adds a new variable named *filter\_\$*. The value of this variable is either 0 or 1 for deselected and selected cases respectively. If you no longer want the data to be filtered you simply select *All Cases...* and all observations in your dataset will be included in your analysis. You should note that if you have chosen to delete the non-selected data and have saved the dataset AFTER deleting them the data are lost for good and cannot be restored!

#### 4.1.7 Weight Cases

In the menu *Weight Cases* it is possible to give each observation different weights for analyzing purposes. For instance you have a large dataset of frequency counts, then instead of entering the raw scores of each individual case, each combination of scores is along with the total frequency count for that group. When using the weight command, the following window will appear.



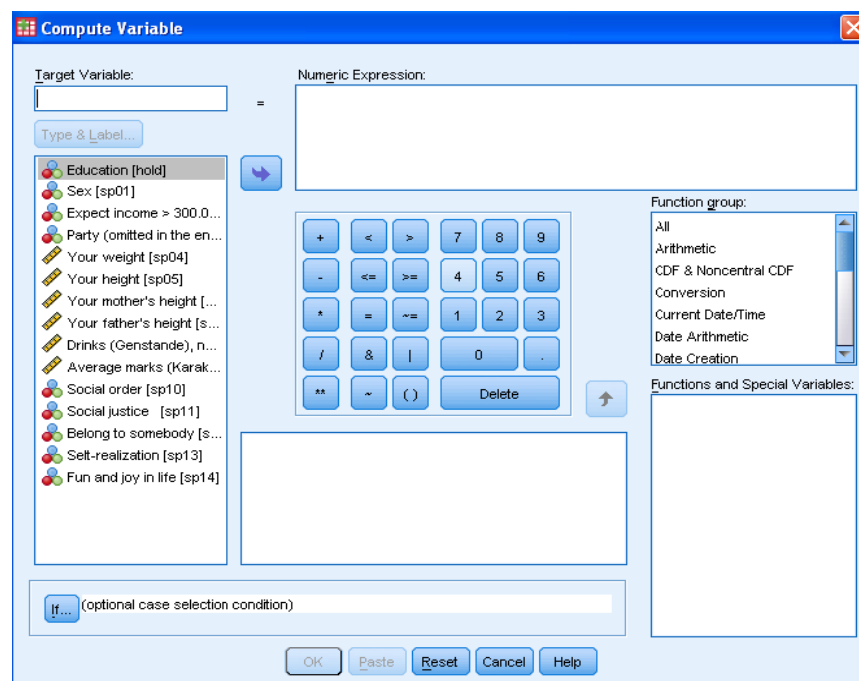
The variable you want weighted is to be moved to the *Frequency Variable* (in the example shown above it would be the variable *Counts*). Then click ok. In the right bottom corner there should now be a text with *Weight On*. This means that each combination of scores is along with the total frequency count for that group.

## 4.2 Transform

If you want variables to be changed or construct new ones this can be done using the menu *Transform*.

### 4.2.1 Construction of new variables

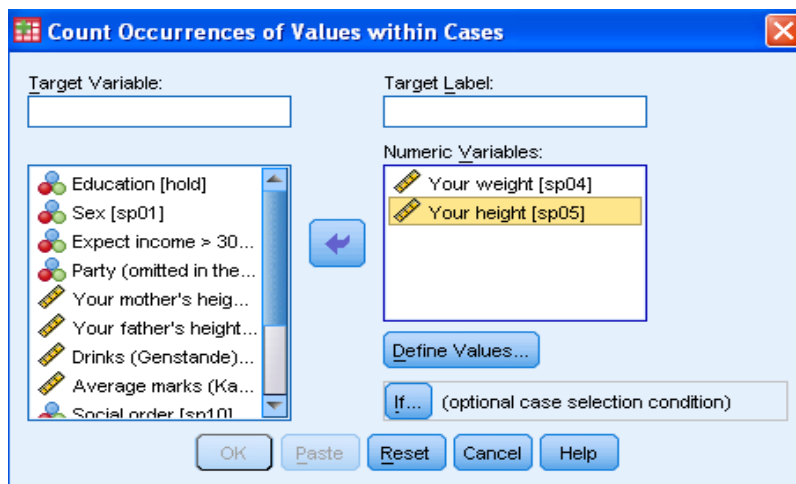
The menu *Compute...* constructs new variables using mathematical transformation of other variables. If you choose this menu the following window will appear:



If you want to construct a new variable you must first define a name for it in the Target Variable. The value of the new variable is to be defined in the *N*umeric *E*xpression by using a mathematical function. This is much simpler than it sounds. Just choose the existing variables you want to include and use these designing the formula/Numeric Expression. Then you click 'OK' and the new variable is being constructed automatically.

#### 4.2.2 Count numbers of similar observations

By choosing the menu *Count...* it is possible to construct a new variable that counts the number of observations for specified variables. E.g. a respondent (case) has been asked whether (s)he has tried several products. The new variable shows how many products the respondent has tried – how many selected variables (s)he has said yes to. The window looks as can be seen below:



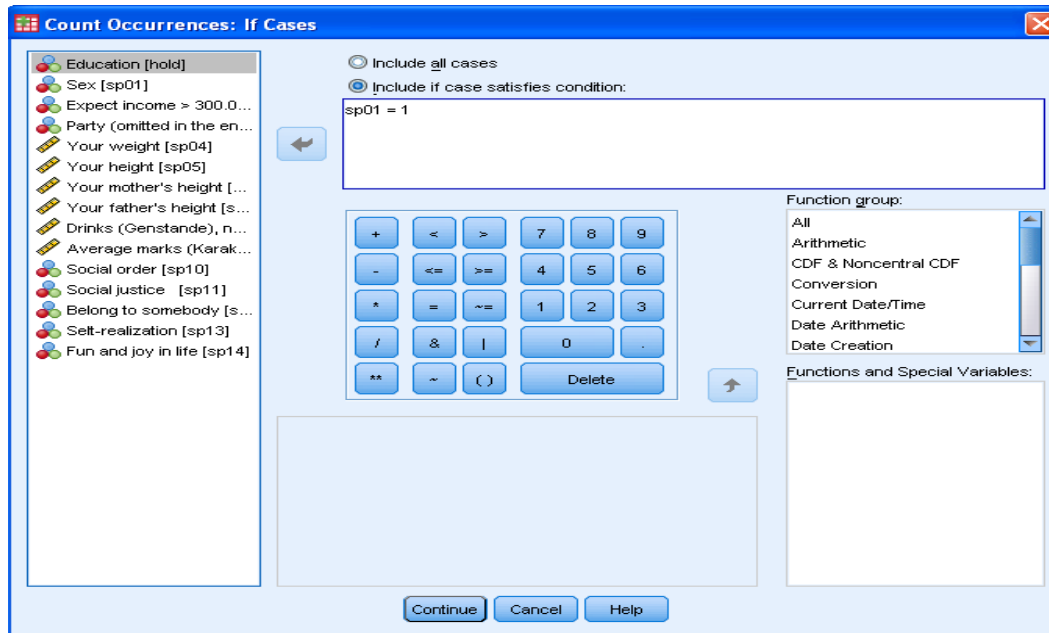
In the *T*arget Variable the name of the new variable is to be written. Then the variables you want to be included in the count are moved to the *N*umeric *V*ariables by selecting them from the left hand window and using the arrow to move them.

The rest of the window will be explained by an example. A count is to be done on how many women fulfill the following criteria:

- Height between 170 and 175 cm.
- Weight  $\leq$  65 kg.

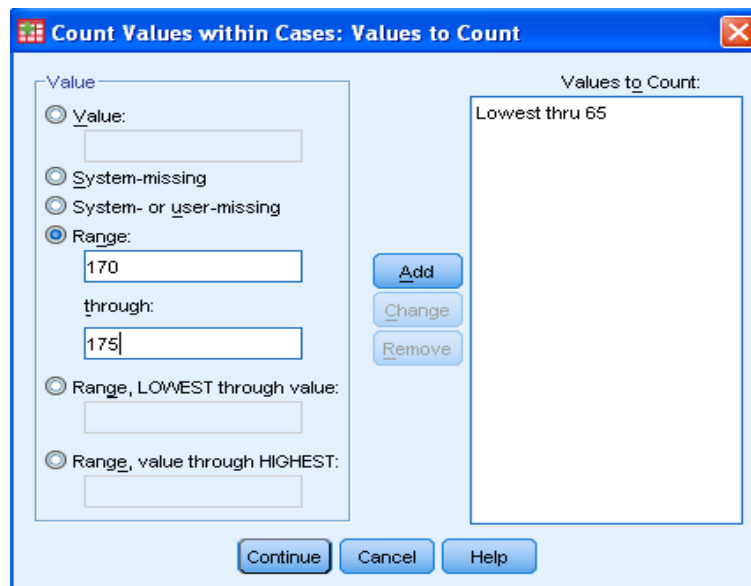
First choose the name of the new variable and move the variable *Height* and *Weight* as specified above.

Now you need to specify which variables the count is to be limited to include (Women=1). This is done by clicking 'if' and the following window will appear:



Since you only want women to be included in your analysis you specify that the variable sex = 1 (meaning only women is to be included). It should be noted that only numeric variables can be used. If your data do not have this format you can easily change it by using *Automatic Recode...* (See section 4.2.5). When the selection is done you click: 'Continue'.

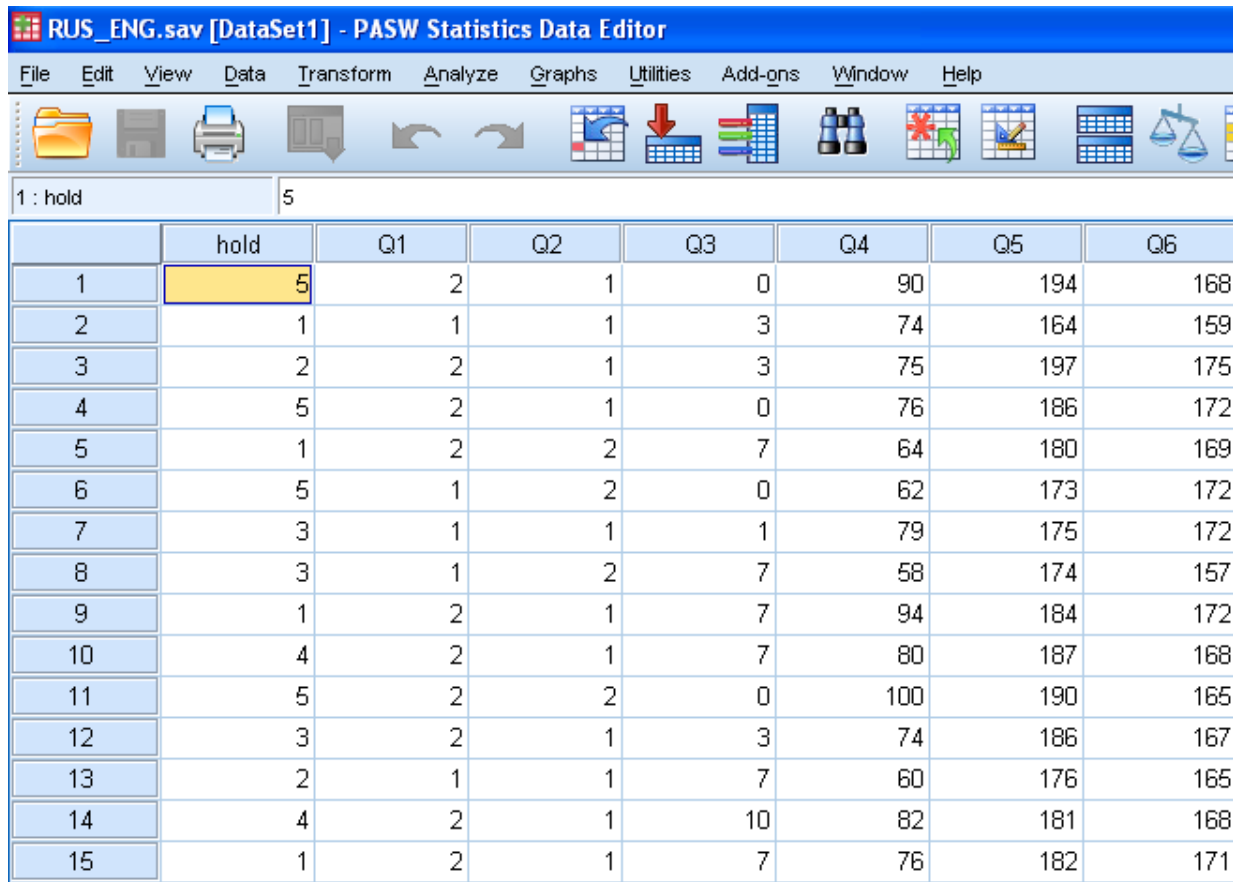
Next you must define the value each variable can take in order to be included in the count. This is done in the *Count Occurrences of Value within Cases* menu and here you click 'Define Values...' and the following window will appear:



You now have different options. You can decide to use a specified value, an interval, a mini-mum or maximum value. In the example shown above you first specify the wanted value (65) for the variable *Weight*. This is done under *Range, LOWEST through value*: In *Value* you simply write 65 and click 'Add...'. For the variable *Height* you want to use an interval,

which is done by clicking *Range*, and specify the minimum (170) and maximum (175) values and click 'Add' again. When all the criteria are specified and added you click 'Continue'.

By running the above example you should get the following output:



	hold	Q1	Q2	Q3	Q4	Q5	Q6
1	5	2	1	0	90	194	168
2	1	1	1	3	74	164	159
3	2	2	1	3	75	197	175
4	5	2	1	0	76	186	172
5	1	2	2	7	64	180	169
6	5	1	2	0	62	173	172
7	3	1	1	1	79	175	172
8	3	1	2	7	58	174	157
9	1	2	1	7	94	184	172
10	4	2	1	7	80	187	168
11	5	2	2	0	100	190	165
12	3	2	1	3	74	186	167
13	2	1	1	7	60	176	165
14	4	2	1	10	82	181	168
15	1	2	1	7	76	182	171

From the output above you can see that e.g. respondent number 2 fulfill 0 of the criteria's (both height and weight), respondent number 6 fulfill 2. Respondent number 1 is not a part of the variable because it is a man.

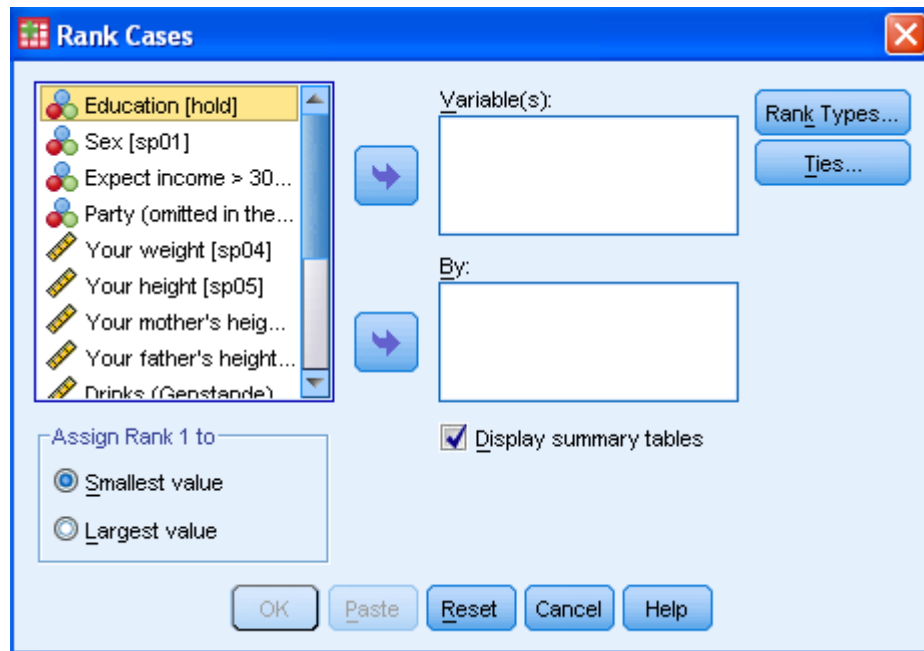
#### 4.2.3 Recode variables

Recoding variables is done when a new variable is to be created based on values from an existing variable or an existing variable needs to be recoded (e.g. the value of men, which now is assuming the value 2 in the dataset needs to be recoded into the value 0. Then we get a so-called dummy variable, which is equal to 1 if the respondent is a woman, and otherwise is equal to 0 if the respondent is a male.

Recoding of variables is a broadly used technique in e.g. regression analysis, logit models and log-linear models (see later chapters).

#### 4.2.4 Ranking Cases

If the dataset needs to be ranked this is done using *Rank Cases*. The following window will appear:

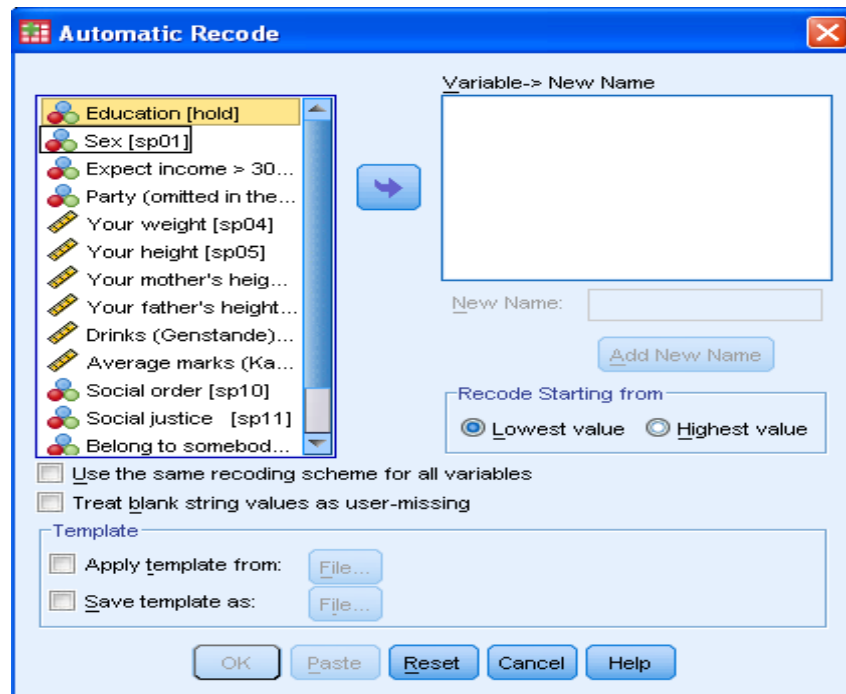


In the field: Variable(s) the variables that are to be ranked are typed (or moved using the arrow). In the field By you enter the variable you want to rank by. By clicking 'Rank Types...' it is possible to choose different ways of ranking the data. By clicking 'Ties...' it is possible to choose the method you want to use if there are more than one similar outcome. The table shows the results of the different methods when using 'Ties...'

Value	Mean	Low	High
10	1	1	1
15	3	2	4
15	3	2	4
15	3	2	4
16	5	5	5
20	6	6	6

#### 4.2.5 Automatic Recode

If a string variable is to be recoded into a numeric variable this is most easily done using *Automatic Recode...*. The following window will appear for specification:



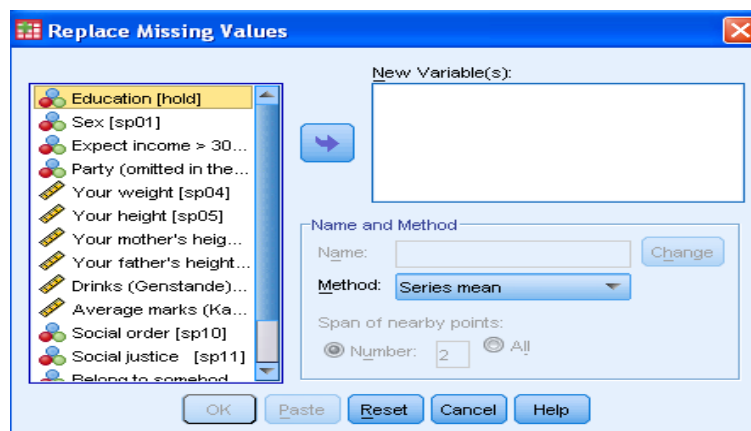
If you e.g. desire to recode the variable *sex*, which is a string variable (Male, Female), into a numeric variable with the values 1 and 2 you do the following: First you select the variable you want to recode (*sex*). Then you have to rename the new variable by using the 'Add New Name' button. Now SPSS automatically constructs the new variable and gives it values starting at 1 ending at the number equal to the number of different outcomes for the string variable.

#### 4.2.6 Replacing missing values

If the dataset includes missing values it can result in problems for further analysis. Because of that it is often necessary to specify a value. For an elaboration of the problems with missing values see section 4.4

The replacement can be done using the: *Replace Missing Values...*

If you select the menu the following window will appear:

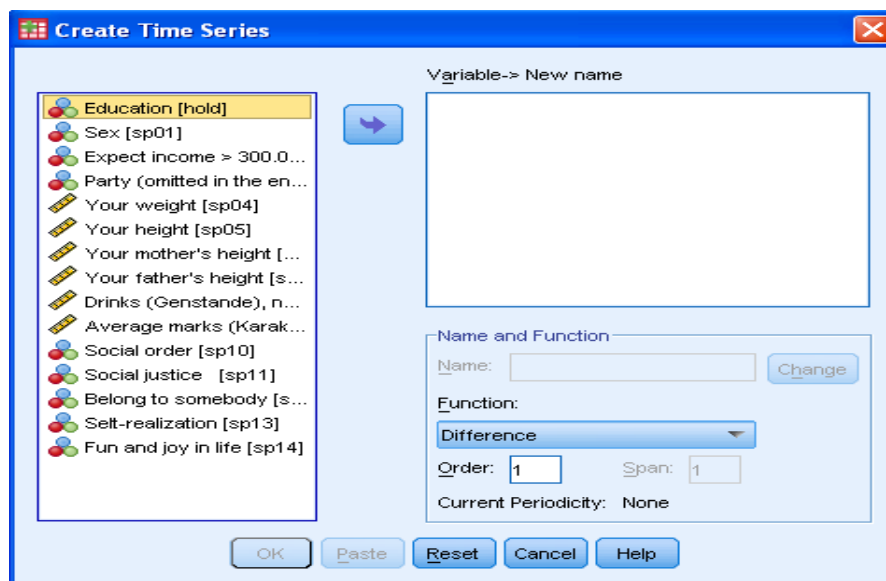




First you select the variables for which you want to specify the missing values and then select the method you want to use. You can choose to use the average of the existing values (Series mean), use an average of the closest observations (Mean of nearby points), use linear interpolation etc. If you want to choose the *Mean of nearby points* you need to specify the nearest observations. This is done by selecting; *Span of nearby points*, where the value specified determines how many of the earlier observation should be included in the calculation. By clicking 'OK' SPSS creates a new variable where the missing values are replaced. SPSS names the new variable automatically, but you can also specify it yourself by selecting 'N<sub>gme</sub>'.

#### 4.2.7 Construction of time series

Using the menu *Create Time Series...* it is possible to create new variables, as a function of al-ready existing numeric time series variables.



First you specify the variable to be used for the time series. This is simply done by selecting the desired variable and clicking the arrow. In the *Order* box you then specify the number of times you want to lag the variable. Finally you specify which method to use for the calculation (Difference, lag etc.) When done you click 'OK' and SPSS automatically creates the new variable.

### 4.3 Recode (join)

Both logit- and log-linear analysis use table of frequencies, which can be described as a count of how many times a given combination of factors appears (see the table below).

Obs (cells)	Factor1 (i)	Factor2 (j)	Frequencies <sub>ij</sub>
1	Male	1	9
2	Male	2	5
3	Male	3	3
4	Male	4	8
5	Female	1	5
6	Female	2	2
7	Female	3	10
8	Female	4	7

From the table above you can see that there are 10 respondents, which was a female (*factor1*) and scored 3 on *factor2* (second last row). It is often necessary or just interesting to join and recode observations – E.g. if the assumption of a model about a minimum expected count is not fulfilled.

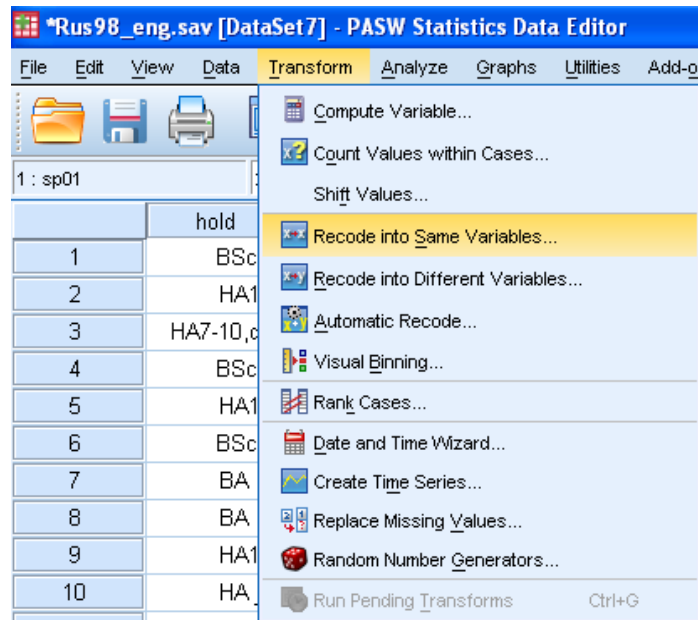
When recoding you join several levels. By doing this you increase the number of observations in each cell. E.g. in the example shown above it would often be recommended that level 1 & 2, and level 3 & 4 in the variable *factor2* are joined respectively. This will reduce the number of cells in our table of frequencies to consist of only 4 cells but each now including more respondents – see the table below.

Obs (cells)	Factor1 (i)	Factor2 (j)	Frequencies <sub>ij</sub>
1	Man	1 (1+2)	14
2	Man	2 (3+4)	11
3	Woman	1 (1+2)	7
4	Woman	2 (3+4)	17

It must be noted that joining levels rely on a subjective evaluation of whether it makes sense to join these levels.

#### 4.3.1 Join using the dialog box

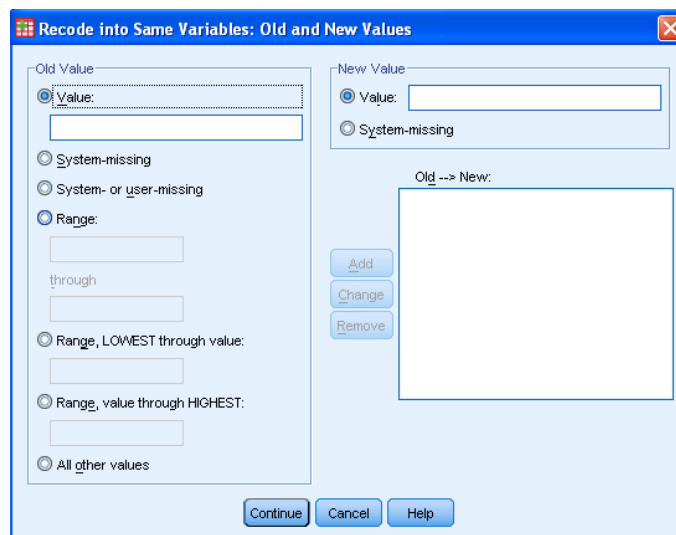
As can be seen in the window below recoding can be done either into the same variables or into a new (different) variable.



##### 4.3.1.1 Recode into Same Variables

By selecting *Recode => Into Same Variables...* it is possible to recode already existing variables. This can be done for both numeric and string variables.

In the first window you select the variable you want to recode. If more variables are selected they must be of the same type. To select the variables to be recoded click 'if...' and they can be selected using logic relations. It is also possible to select all variables. Next you click the menu 'Old and New values...' and the following window appears:



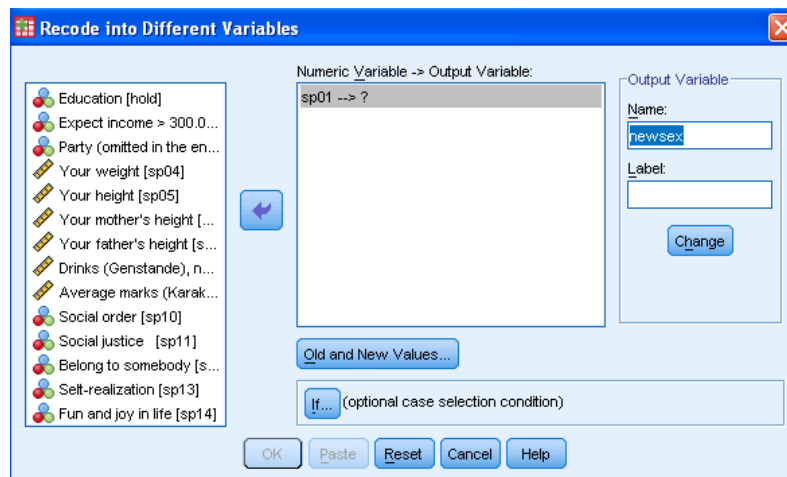
In *Old Value* you specify which values are to be recoded. If it is only a single value you want to specify you choose the first field *Value* and enter the value. If you are to recode non-defined missing values you choose the field *System-missing*. If the variables are defined as missing values or unknown, you choose *System- or user-missing*. Please noted that this is a very important feature to use, when recoding variables including missing values cf. section 4.4 below.

Last if it is a range or an interval you choose and specify the range in one of the next three options.

In the right hand side of the window you define the new value you want the old values to be replaces by. After the recoding is defined you click 'Add'. When all the recoding has been specified you click 'Continue' and 'OK' and SPSS does the recoding automatically.

#### 4.3.1.2 Recode into Different Variables

Instead of recoding into the same variable you can choose to recode *Into Different Variables*. Now it is possible to create a new variable from existing ones. Also here you can both recode numeric and string variables. The window looks as seen below:

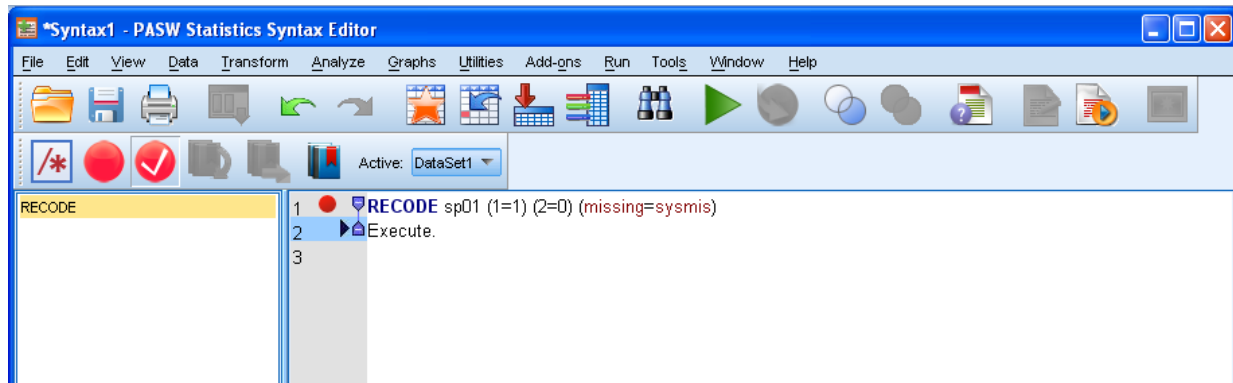


In the left hand side you choose the variable you want to recode. In the right hand side you specify the name for the new variable you are to compute. When specified click; 'Change', and the combination is added to the list. If it is not desired to recode all cases you can use the 'if...' menu to define in which situations you want the recode.

The values the recoded observations are to take can be specified in the menu; 'Old and New Values...'. A new window will appear where you choose which values are to be replaced. The procedure is equivalent to what is described in section 4.3.1.1.

#### 4.3.2 Recoding using the syntax

A different method is manually recoding in the syntax. To do that you choose *File => New => Syntax* and a window similar to the one below will automatically open.



### Explanations:

RECODE: Start procedure!

- The “recode” procedure takes all values = 2 and gives them the new value “0”. The values = 1 will be given the value 1. The missing values will be sysmis. This means that they are missing, and will not be included in further analysis. Please note that these changes will be recorded into the already existing variable.

EXECUTE: The procedure will be executed.

*Always remember that every statement must end with a full stop – a dot (.).*

## 4.4 Missing values

The term, missing values, is defined as non-respondents / empty cells within a variable.

The problem with missing values is mostly pronounced when working with data collected by a questionnaire. The problem arises as some of the respondents have chosen not to answer one or more of the questions posed.

Before you carry out any statistical analyses, it is important to consider how to deal with these missing values. The most commonly used method is to define which value of the variable that represents a missing value cf. section 3.1.1. When the variable takes on this particular value the observation is excluded from any analyses performed, thus only leaving in the respondents who actually answered the question.

Another but not so frequently used method is the one described in section 4.2.6 where a missing value is replaced by a specific value, for instance the mean of the other observations and then included in any subsequent analyses. This method is not applicable when dealing with data from a questionnaire, however it is most often used with time series data when you want to remove any holes in the series

Another situation where it is important to focus on missing values is in conjunction with data manipulation. For instance if you want to recode a variable, there is a risk that you may unintentionally change a missing value so it will be included in subsequent analyses. An example of this is given below, where the variable education with the following levels:

0 = missing value 1 = HA, 2 = HA(dat), 3 = HA(int), 4 = HA(jur)

is recoded into a new variable with the following levels 1 = HA and 2 = other educations. If this is done as described in section 4.3.1 (*Transform => Recode => Into different...*) you put 1 = 1 and else = 2 as shown below.

**Recode into Different Variables: Old and New Values**

**Old Value**

- ☐ Value:
- ☐ System-missing
- ☐ System- or user-missing
- ☐ Range:
- ☐ Range, LOWEST through value:
- ☐ Range, value through HIGHEST:
- ☒ All other values

**New Value**

- ☒ Value:
- ☐ System-missing
- ☐ Copy old value(s)

Old --> New:

1 --> 1  
ELSE --> 2

Buttons: Add, Change, Remove

☐ Output variables are strings Width: 8

☐ Convert numeric strings to numbers ('5'-->5)

Buttons: Continue, Cancel, Help

As a result of this recoding, all the missing values are now assigned with the value 2, which mean they will be included in any subsequent analyses. This may result in false conclusions not supported by the real data.

To prevent this from happening it is important to make sure that the missing values are preserved after the recode. In the example above, you can do this, by using the option "system- or user-missing" as shown in the dialog box below.

**Recode into Different Variables: Old and New Values**

**Old Value**

- ☐ Value:
- ☐ System-missing
- ☐ System- or user-missing
- ☒ Range:
- ☐ Range, LOWEST through value:
- ☐ Range, value through HIGHEST:
- ☐ All other values

**New Value**

- ☐ Value:
- ☒ System-missing
- ☐ Copy old value(s)

Old --> New:

1 --> 1  
SYSMIS --> SYSMIS  
ELSE --> 2

Buttons: Add, Change, Remove

☐ Output variables are strings Width: 8

☐ Convert numeric strings to numbers ('5'-->5)

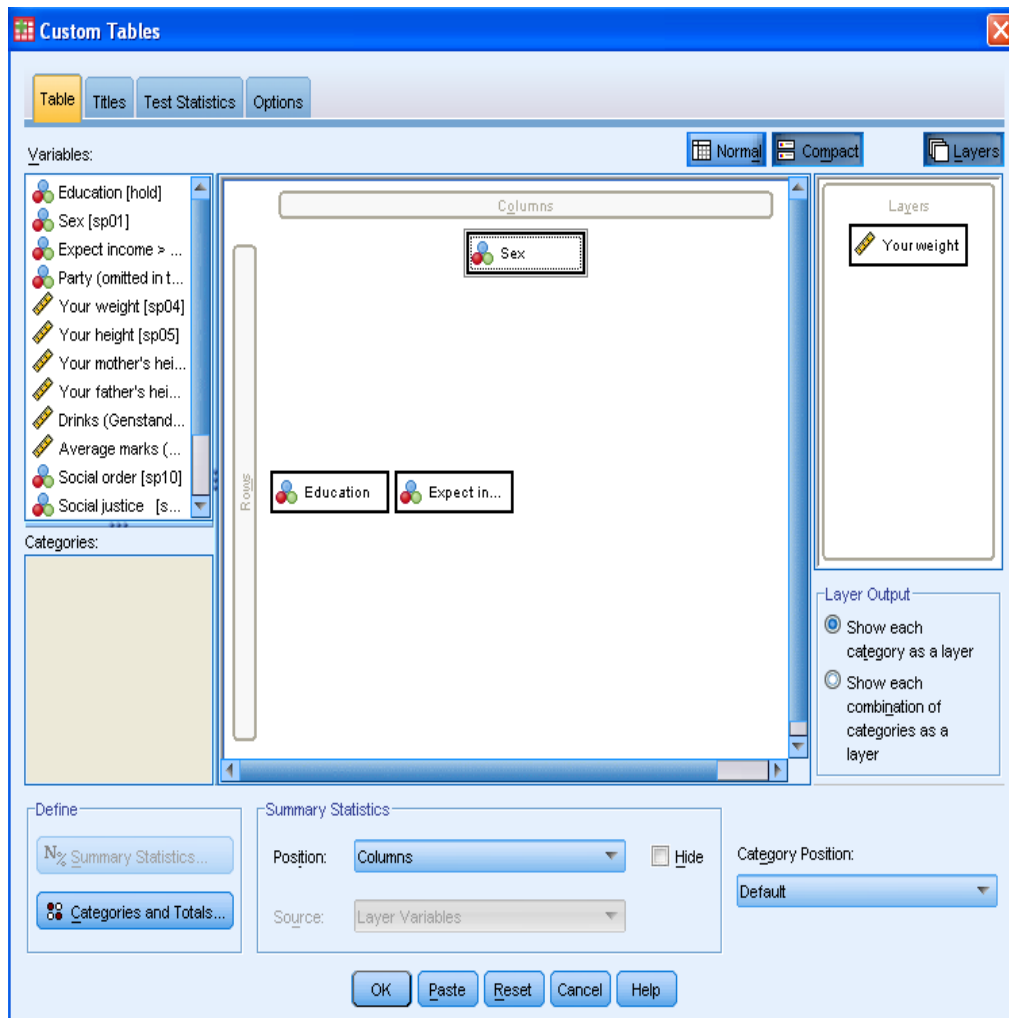
Buttons: Continue, Cancel, Help

If you recode your missing values in this way, you will be certain that they are preserved in the new variable created.

## 5. Custom Tables

In SPSS you can also do custom tables, which describe the relationship between variables in a table of frequencies. These tables can either be simple two-dimensional tables or multiple dimensional tables. To make simple tables you do the following:

*Analyze => Tables => Custom Tables*



If you want to make a table with multiple dimensions you need to press the *Layers* button. Otherwise the table will only consist of two dimensions.

- The variables you want displayed, means and other descriptive measures are dragged into the *Rows* section.
- The *Normal* button makes it possible to preview the table you are about to produce. Whether you chose to use the *Compact* or the *Normal* button is a matter of taste.
- In *Summary Statistics...* it is possible to include other measures than mean, which is set as default. You can e.g. select the minimum or maximum value. You need to click on the variables you want statistics calculated for in order to activate the *Summary Statistics* button.

- In *Titles...* it is possible to give the table a title or insert a time stamp.

## 5.1 Custom Tables output

Below is an example for the output of a Custom Table. The output shows the average weight split into groups based on sex, education and expected income.

**Table 1**

Your weight				Sex	
				Female	Male
				Mean	Mean
Education	HA1-6	Expect income > 300.000	Yes	61	78
			No	61	74
	HA7-10,dat	Expect income > 300.000	Yes	63	78
			No	62	78
	BA int	Expect income > 300.000	Yes	61	77
			No	61	75
	HA jur	Expect income > 300.000	Yes	62	78
			No	55	74
	BSc B	Expect income > 300.000	Yes	58	73
			No	61	83

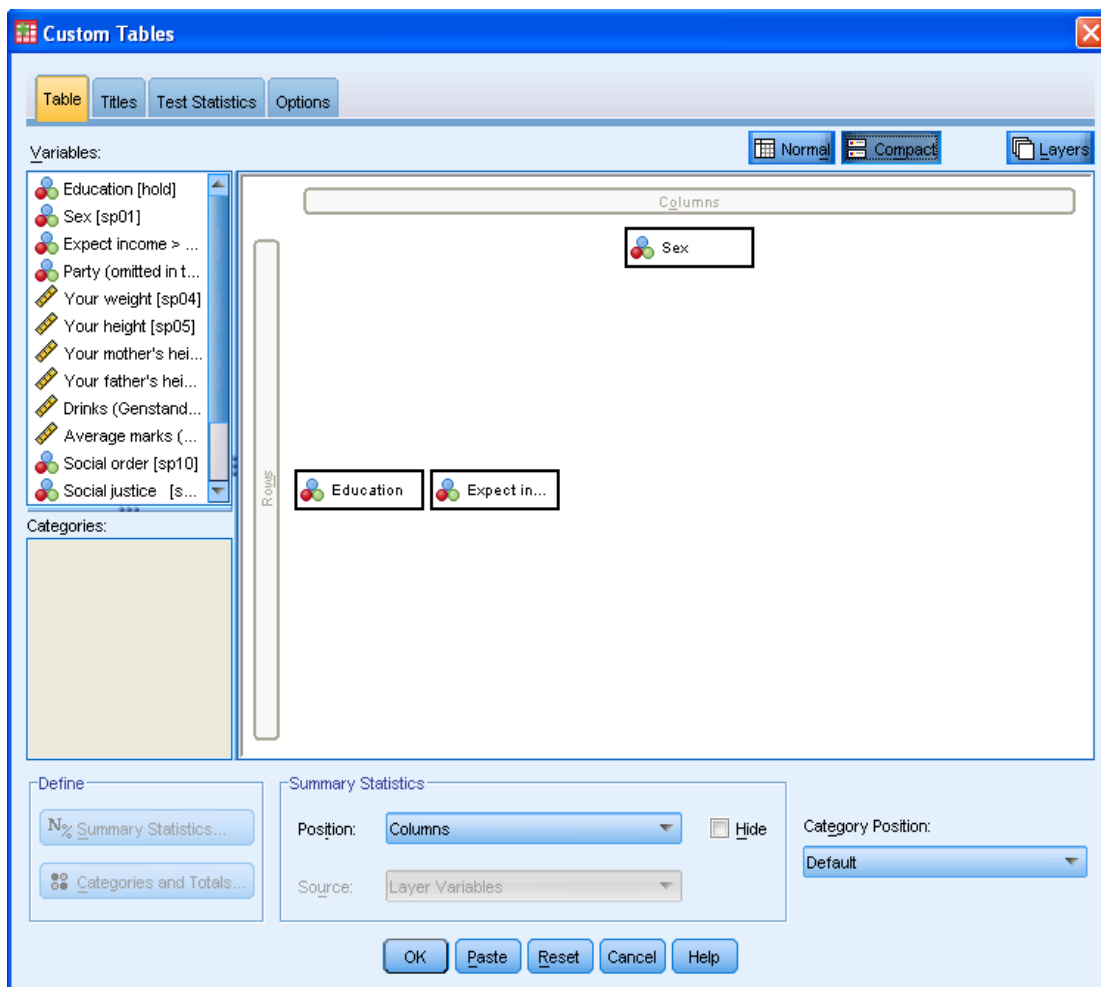


## 6. Tables of Frequencies and crosstabs

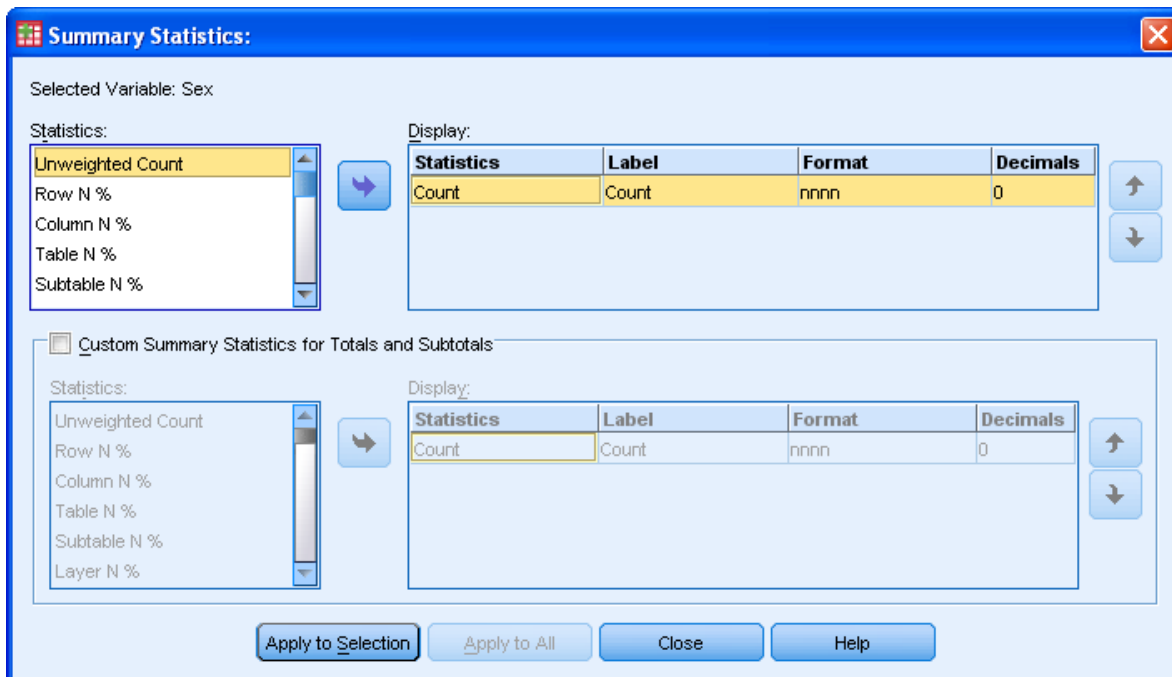
### 6.1 Custom Tables

Custom Tables can also be used to produce tables of frequencies, crosstabs and more. To make a table of frequencies you select *Custom Tables* as described above.

A window similar to the one below will be shown.



- In *Rows* you enter the variables, which are to be counted.
- In *Columns* you enter subgroups - if any.
  - If you include a variable in *Layers* you will get a table of multiple dimensions.
- In *Summary Statistics...* you have the option to get the percentage of each group. E.g. by clicking on the *Sex* variable and pressing *Summary Statistics*, you get the window shown below. Here you can add a percentage for the row.



- In 'Titles...' the title of the tables can be changed.

### 6.1.1 Table of frequencies output

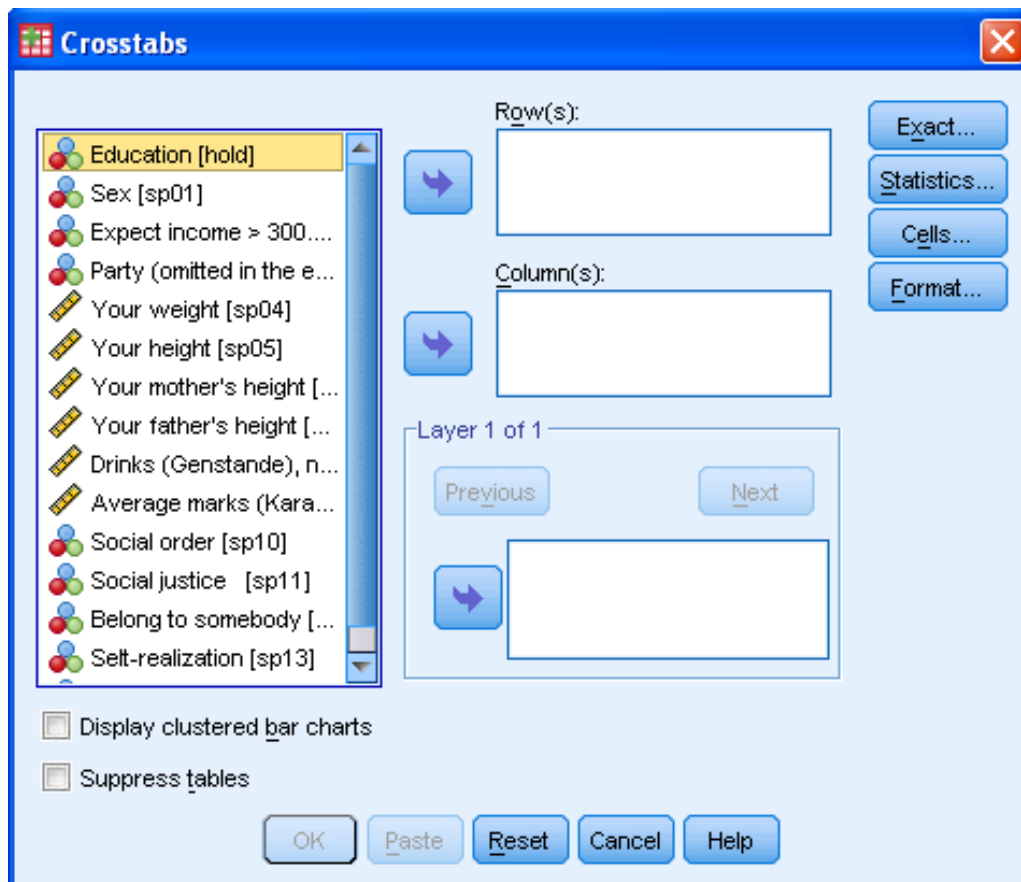
Below you see an example of a table of frequency output, corresponding to the options set in the example above.

				Sex			
				Female		Male	
				Count	Column N %	Count	Column N %
Education	HA1-6	Expect income > 300.000	Yes	37	69,8%	75	74,3%
			No	16	30,2%	26	25,7%
	HA7-10,dat	Expect income > 300.000	Yes	36	78,3%	76	80,0%
			No	10	21,7%	19	20,0%
	BA int	Expect income > 300.000	Yes	21	60,0%	28	84,8%
			No	14	40,0%	5	15,2%
	HA jur	Expect income > 300.000	Yes	26	96,3%	26	86,7%
			No	1	3,7%	4	13,3%
	BSc B	Expect income > 300.000	Yes	6	60,0%	15	75,0%
			No	4	40,0%	5	25,0%

The table shows what percentage of the students, that expects to earn above 300.000 in the future, based on their sex and education. As can be seen only 60 % of the female BA(Int.) students expect to earn more than 300.000 while 96,3 % of the female HA(Jur.) students do.

## 6.2 Crosstabs

Crosstabs shows the relationship between two nominal or ordinal scaled variables. To make a crosstab chose: *Analyze=> Descriptive Statistics=> Crosstabs* and the following box appear. The variable you want in rows is moved to Row(s) and the variable you want for column is moved to Column(s). In the following example the relationship between sex and education will be investigated.



It is also possible to test for homogeneity and independence in your output How to do this is described in section 17.

When you press *ok*, a crosstab with the frequencies within the different combinations will appear in the output. It is also possible to get percentages in the table. This can be done by pressing *Cells*. Then the following window appears, here it is possible to get percentages both for each row, each column and in percent of the total. Marking respectively Row, Column and Total, does this.

**Crosstabs: Cell Display**

**Counts**

☒ Observed

☐ Expected

**Percentages**

☒ Row

☒ Column

☒ Total

**Residuals**

☐ Unstandardized

☐ Standardized

☐ Adjusted standardized

**Noninteger Weights**

☒ Round cell counts    ☐ Round case weights

☐ Truncate cell counts    ☐ Truncate case weights

☐ No adjustments

Continue Cancel Help

This leads to the following output, where you can see that there on HA jur. are 27 females which is 15.6% of all women, 47,4% of those how study HA jur. and 5.9% of all students.

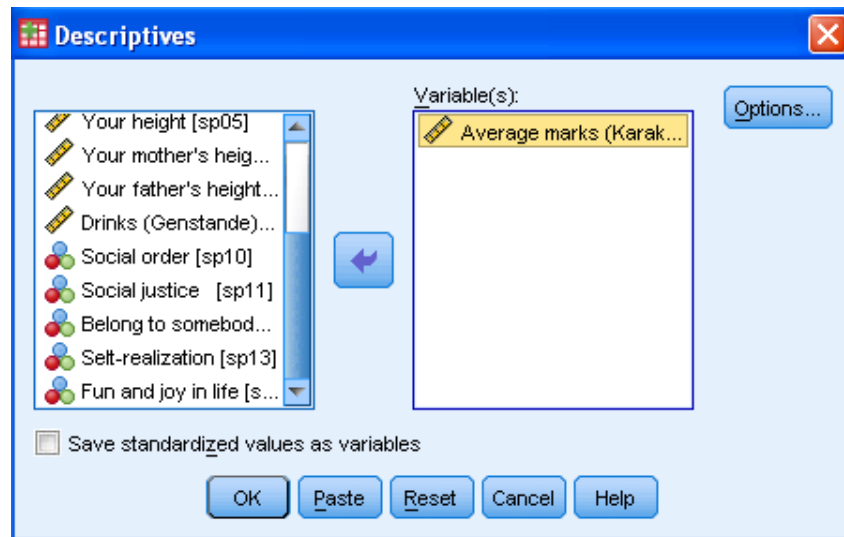
**Sex \* Education Crosstabulation**

			Education					
			HA1-6	HA7-10,dat	BA int	HA jur	BSc B	Total
Sex	Female	Count	53	47	36	27	10	173
		% within Sex	30,6%	27,2%	20,8%	15,6%	5,8%	100,0%
		% within Education	34,2%	32,6%	52,2%	47,4%	33,3%	38,0%
		% of Total	11,6%	10,3%	7,9%	5,9%	2,2%	38,0%
	Male	Count	102	97	33	30	20	282
		% within Sex	36,2%	34,4%	11,7%	10,6%	7,1%	100,0%
		% within Education	65,8%	67,4%	47,8%	52,6%	66,7%	62,0%
		% of Total	22,4%	21,3%	7,3%	6,6%	4,4%	62,0%
	Total	Count	155	144	69	57	30	455
		% within Sex	34,1%	31,6%	15,2%	12,5%	6,6%	100,0%
		% within Education	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% of Total	34,1%	31,6%	15,2%	12,5%	6,6%	100,0%

## 7. Descriptives

Often it is desirable to get some descriptive measures for a selected variable. Descriptives include measures like mean, standard deviation etc. To get descriptive measures you select: *Analyze => Descriptive Statistics => Descriptives*

A window like the one below will be shown.



- In *Variable(s)* you include those variables you want to have descriptive measures for.
- If you tick *Save standardized values as variables*, the standardized variables will be saved in a new variable in the current dataset.
- In '*Options...*' you select the descriptive statistics you want to be included in the output.

### 7.1 Output for Descriptive Statistics

Below is shown what the output for descriptive statistics could look like, depending on the different selections you have made. In this case descriptive statistics for the average marks are shown.

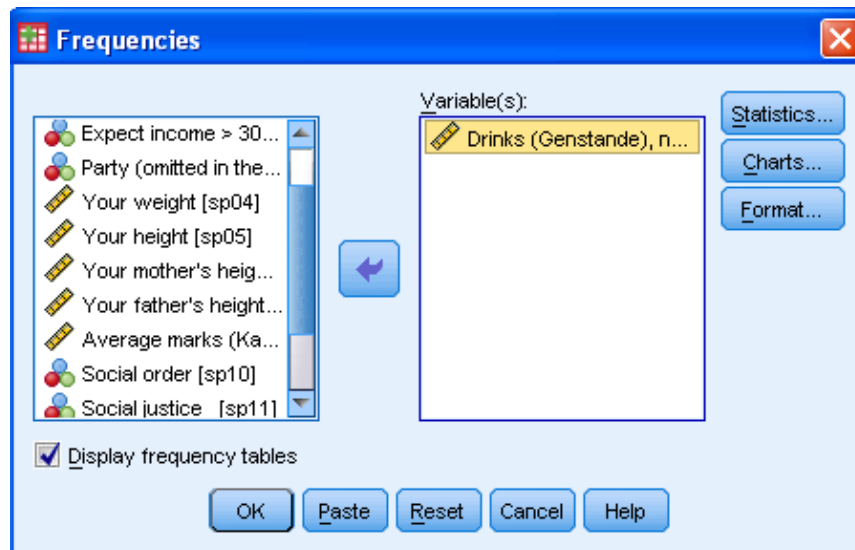
**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
Average marks (Karakter) at qualifying exam	445	6,3	10,4	8,476	,7382
Valid N (listwise)	445				

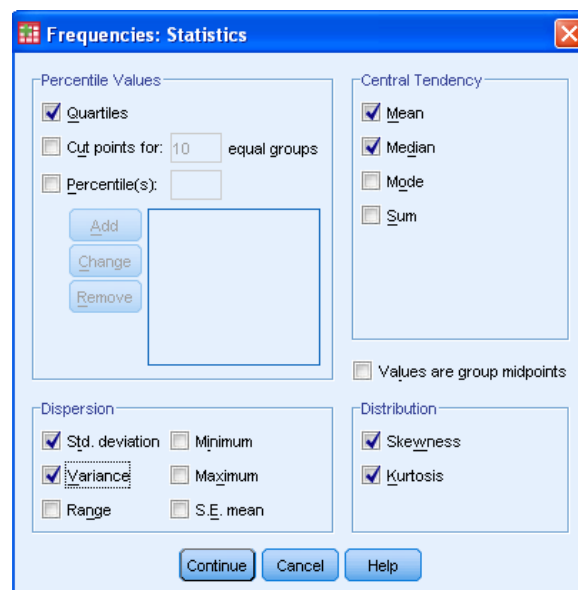
## 8. Frequencies

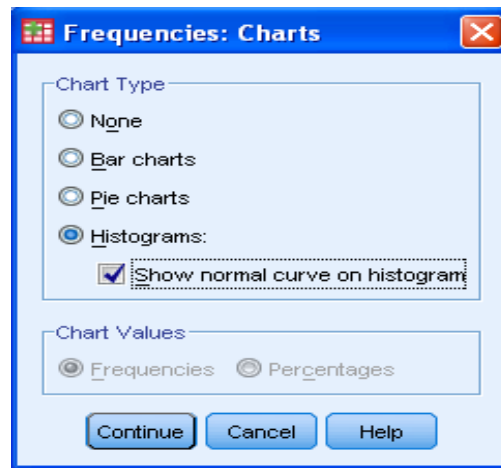
As could be seen in the former sections, it is possible to select both descriptive statistics and frequencies at the same time. *Frequencies* are used if you want to see quartiles and plots of the frequencies. To do this you select the following in the menu bar: *Analyze => Descriptive Statistics => Frequencies*

The following will appear on your screen:



- In *Variable(s)* you include the variables you wish to have measures for.
- If the menu '*Statistics...*' is selected it will be possible to include descriptive statistics and different percentages. E.g. standard deviation, variance, median etc.





- In 'Charts...' it is possible to make plots of the table of frequencies.
- In 'Format...' you can format the tables to make it look like you want it to – almost!

## 8.1 Frequencies output

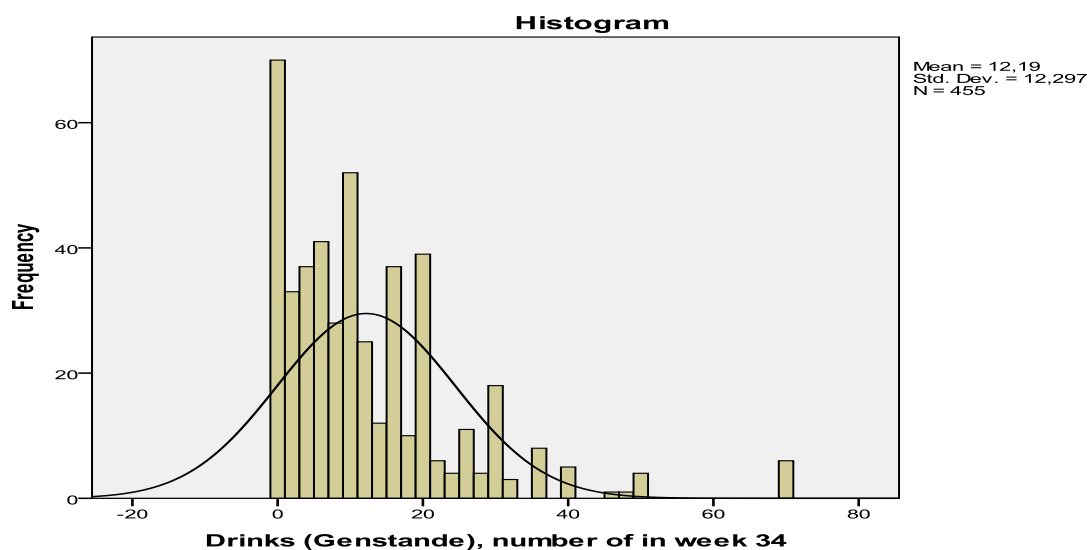
The frequencies' output will look somewhat similar to the one shown below:

Statistics		
Drinks (Genstande), number of in week 34		
N	Valid	455,000
	Missing	,000
Mean		12,193
Median		10,000
Std. Deviation		12,297
Variance		151,214
Skewness		1,888
Std. Error of Skewness		,114
Kurtosis		5,155
Std. Error of Kurtosis		,228
Percentiles	25	3,000
	50	10,000
	75	18,000

**Drinks (Genstande), number of in week 34**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	70	15,4	15,4	15,4
	1	6	1,3	1,3	16,7
	2	27	5,9	5,9	22,6
	3	11	2,4	2,4	25,1
	4	26	5,7	5,7	30,8
	5	32	7,0	7,0	37,8
	6	9	2,0	2,0	39,8
	7	11	2,4	2,4	42,2
	8	17	3,7	3,7	45,9
	9	6	1,3	1,3	47,3
	10	46	10,1	10,1	57,4
	11	2	,4	,4	57,8
	12	23	5,1	5,1	62,9
	13	4	,9	,9	63,7
	14	8	1,8	1,8	65,5
	15	30	6,6	6,6	72,1
	16	7	1,5	1,5	73,6
	17	5	1,1	1,1	74,7

Like any other output in SPSS the output layout varies depending on the options selected. The above shown output looks exactly what it would look like with the options mentioned in this section.



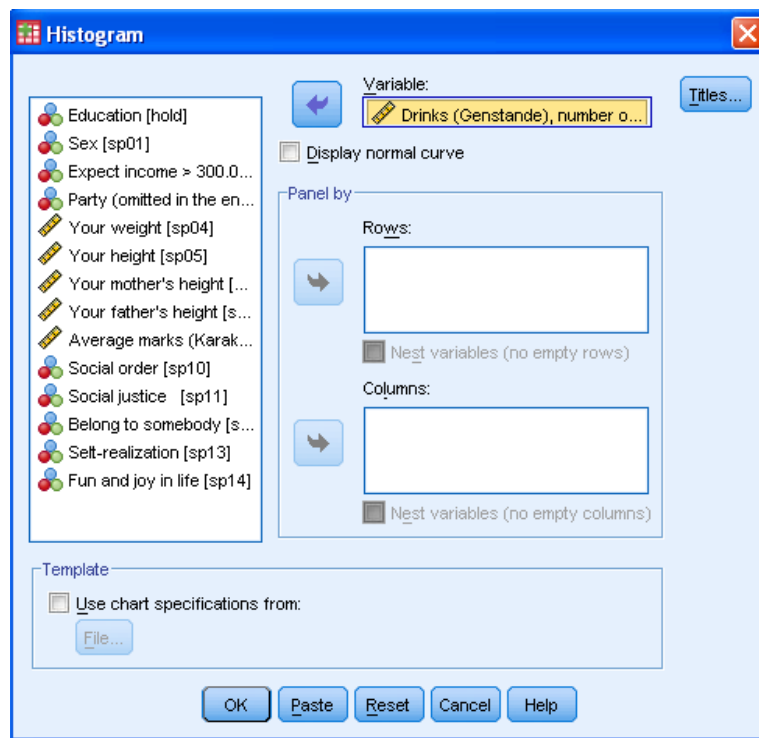


## 9. Plots

### 9.1 Histograms

In many cases it is relevant to make a histogram of a variable where you can see the distribution of the respondent's answers. This can easily be done in SPSS by choosing *Graphs => Legacy dialogs => Histogram*.

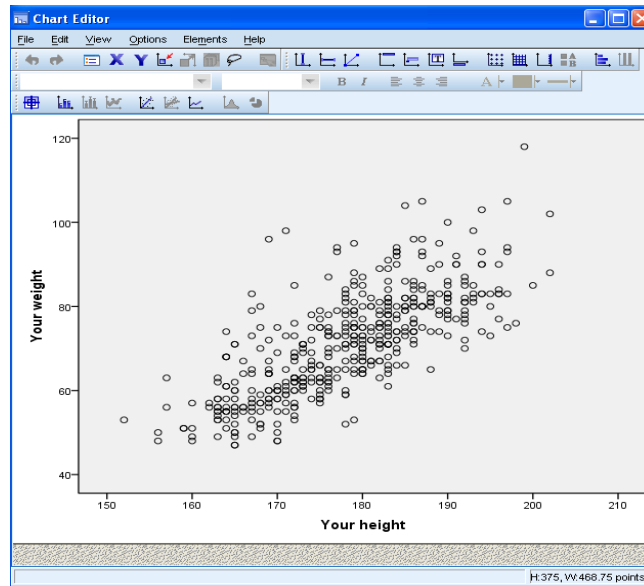
Then the below shown window will appear. Under variable you choose the variable that should be used in the histogram. If you want to have a normal curve on the histogram this can be done by marking *Display normal curve*. By pressing "Titles" you can make titles and insert footnotes on the graph. If you wish to have more histograms of the same variable sex grouped by another variable for example sex, this can be done by moving the variable over in the box *Rows* (the histograms will appear under each other) or the box *Columns* (the histograms will appear beside each other).



Under section 8.1 there is an example of a histogram showing the units of drinks people drank in the "rusuge" with a normal curve on.

### 9.2 Chart Editor

In the *Chart Editor* it is possible to edit plots and charts. To activate this editor you must double click the graph you want to edit. The *Chart Editor* is a separate window like the *Data Editor* and the *Output viewer*. The graph will be grey, when you have double clicked on it for editing (as can be seen below) until you have closed down the window. The graph below is produced via *Graphs => Legacy Dialogs => Scatter/Dot => Simple Scatter* and choosing *Your height* as *X-Axis* and *Your weight* as *Y-Axis*.

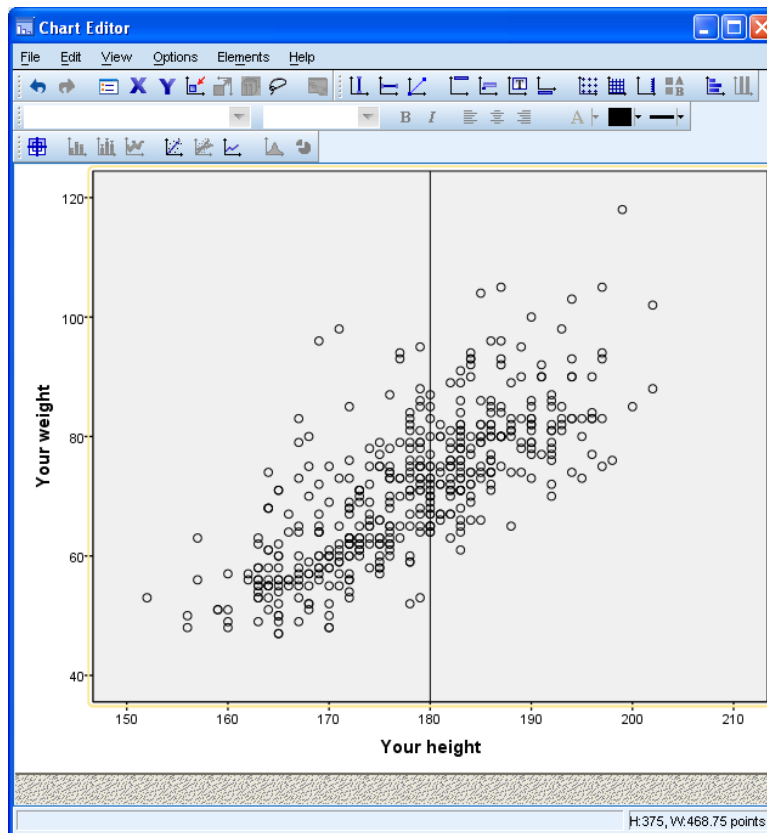


In the *Chart Editor* you can edit the graph in many ways, insert reference lines etc. The way it works is similar to Excel's Chart Editor and will be described below.

### 9.3 Reference line

First select *Options => X-(or Y) Axis Reference Line* on the menu bar, depending of which type of reference line is needed. Then you need to specify where the line should be positioned. This is done in the menu window:

The reference line will now look like the one below:

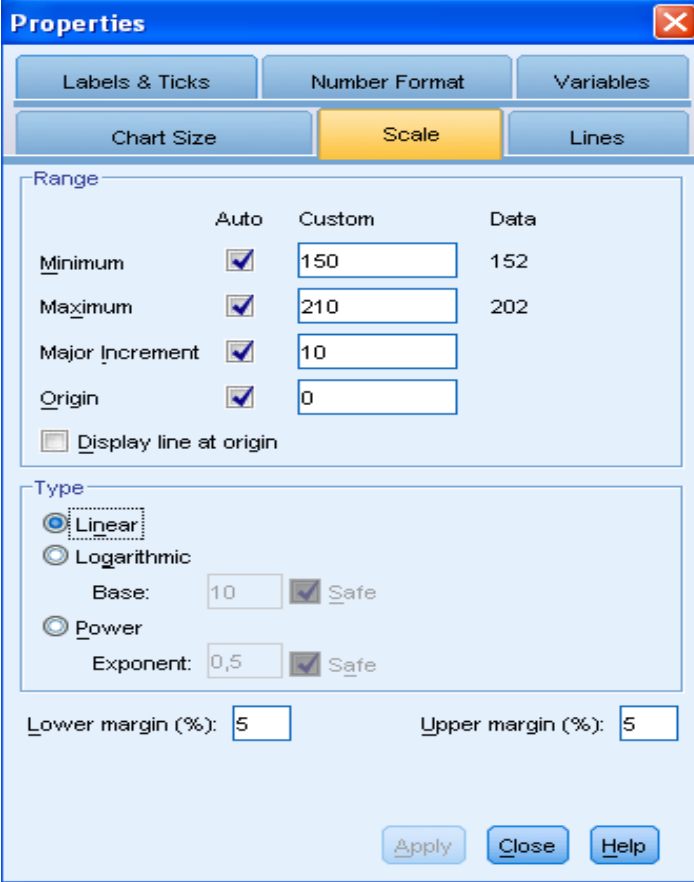


## 9.4 Trend Line

To enter a Trend Line you select Elements => *Fit line at total*. In the submenu *Fit line* it is possible to make a curved line and confidence interval for the regression line if desired.

## 9.5 Editing Scales

Often it is desirable to edit the scales. This is done in the *Edit => X/Y Select X/Y Axis*. Then you will be given the option to specify the scale for the axis, change their titles etc.



The image shows the 'Properties' dialog box in SPSS, specifically the 'Scale' tab. The dialog box has a title bar with a close button. Below the title bar are four tabs: 'Labels & Ticks', 'Number Format', 'Variables', and 'Scale'. The 'Scale' tab is selected and highlighted in yellow. Below the tabs are two main sections: 'Range' and 'Type'. The 'Range' section has a table with columns 'Auto', 'Custom', and 'Data'. The 'Type' section has radio buttons for 'Linear', 'Logarithmic', and 'Power'. At the bottom, there are input fields for 'Lower margin (%)' and 'Upper margin (%)', and three buttons: 'Apply', 'Close', and 'Help'.

	Auto	Custom	Data
Minimum	<input checked="" type="checkbox"/>	150	152
Maximum	<input checked="" type="checkbox"/>	210	202
Major Increment	<input checked="" type="checkbox"/>	10	
Origin	<input checked="" type="checkbox"/>	0	
<input type="checkbox"/> Display line at origin			

Type

☒ Linear

☐ Logarithmic

Base: 10 ☒ Safe

☐ Power

Exponent: 0,5 ☒ Safe

Lower margin (%): 5 Upper margin (%): 5

Apply Close Help

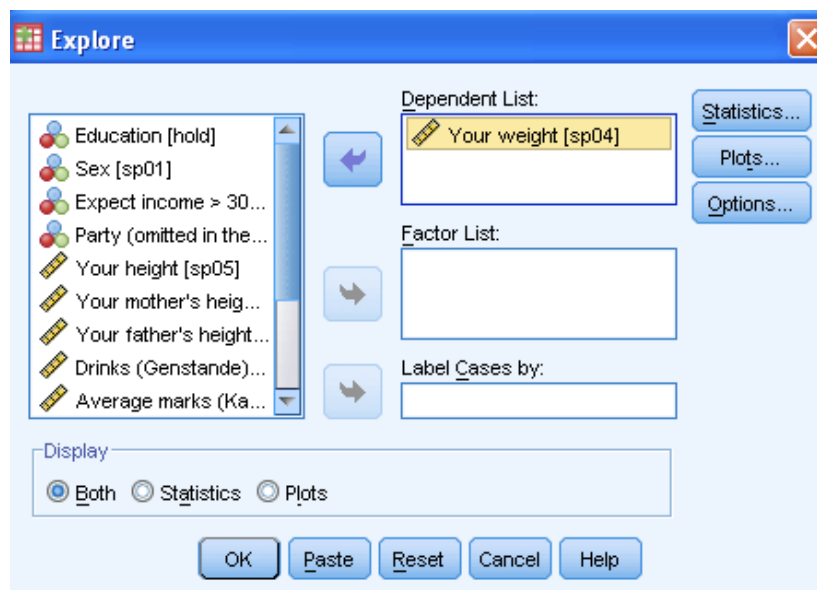
## 10. Test of normality, extreme values and probit-plot

This chapter will show you how to test for normality and make probit plots. By doing this you can check if the assumptions, for the test you want to perform, are satisfied. Also you can use this as an explorative test to identify observations, which have an extreme value (also called outliers). Sometimes you actually want to exclude these extreme values to get a better test result. Test for normality has the following hypothesis.

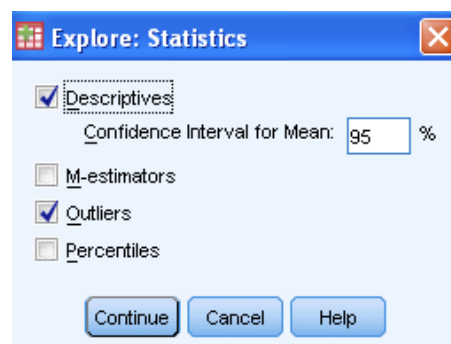
$$H_0 : \text{Norm distributed}$$

$$H_1 : \text{Not- Norm distributed}$$

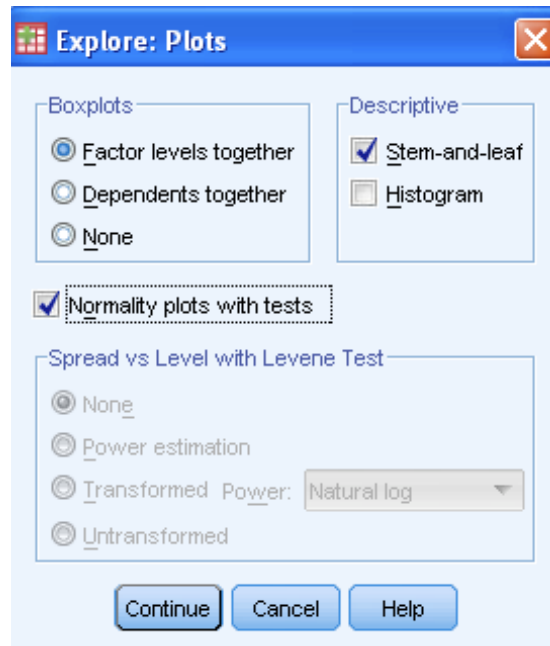
Test for normality and probit plot can be done by selecting: *Analyze => Descriptive Statistics => Explore*. The following window will appear on the screen:



- In *Dependent List* you insert the variables to be tested.
- In *Factor List* it is possible to divide the dependent variable based on a nominal scaled variable.
- In *Display* you must tick *Both* if you want to include both a plot and test statistics in your output.
- In '*Statistics...*' it is possible to select the level of significance and extreme values. As shown.



- In 'Plots...' select *Normality plots with tests*, as shown below. The interesting part here is the two tests that are performed; "Kolmogorov-Smirnov-test" and "Shapiro-Wilk-test" (the latter are only used if the sample size does not exceed 50).



- In 'Options...' you get the possibility to exclude variables in a specified order or just re-report status.

## 10.1 Explore output

The following is just a sample of the output, which appears with the above selected settings. The first table shows the test of normality, while the second table shows statistics about possible outliers.

Tests of Normality						
	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Your weight	,053	451	,004	,986	451	,000

a. Lilliefors Significance Correction

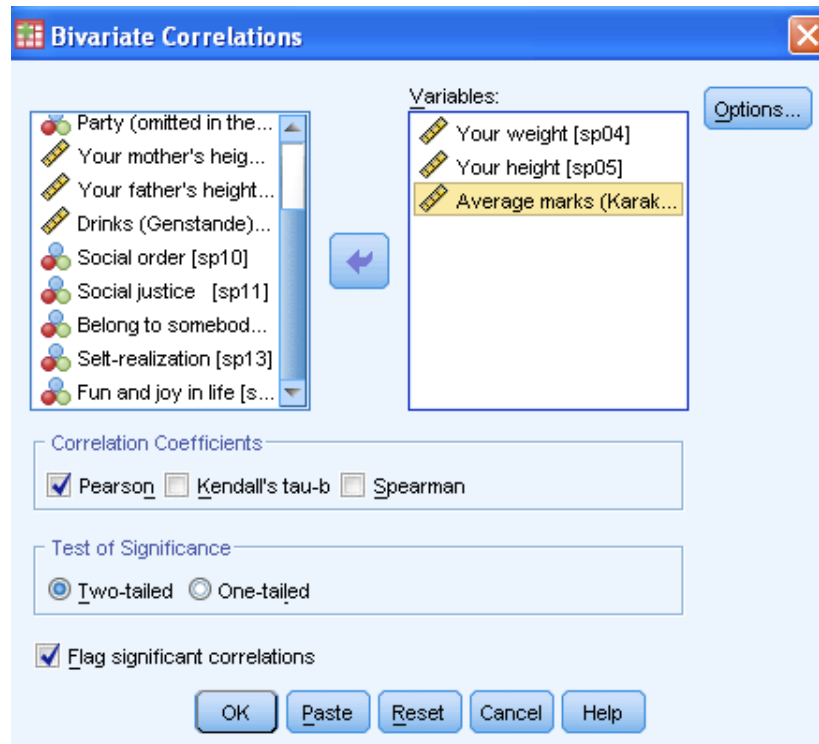
As it can be seen from the output we have a p-value on 0,4%. This means we reject  $H_0$  and therefore we cannot say it is normal distributed. The Shapiro-Wilk test gives us a p-value on 0% and therefore the data is not from a normal distributed population.

## 11. Correlation matrices

In SPSS there are *three methods* to make a correlation matrix. *One of them (Pearson's bivariate correlations)* is the most frequently used and will be described in the following.

### 11.1 Correlation matrix

The most used correlation matrixes is the following: *Analyze => Correlate => Bivariate...*



- In *Variables* you insert the variables you want to correlate.
- In *Correlation Coefficients* you mark the correlations you want to be calculated. The most used choice is Pearson!
- In *Test of Significance* you select the test form – one or two tailed. Note that the significant correlations as default will be shown with a \*/\*\* because of *Flag significant correlations*. It should also be noted that significant correlations does not indicate that the variables are significant in a regression analysis.
- In 'Options...' you can calculate means and standard deviations

## 11.2 Bivariate Correlation output

Correlations				
		Your weight	Your height	Average marks (Karakter) at qualifying exam
Your weight	Pearson Correlation	1	,749**	-,124**
	Sig. (2-tailed)		,000	,009
	N	451	450	442
Your height	Pearson Correlation	,749**	1	-,029
	Sig. (2-tailed)	,000		,546
	N	450	454	444
Average marks (Karakter) at qualifying exam	Pearson Correlation	-,124**	-,029	1
	Sig. (2-tailed)	,009	,546	
	N	442	444	445

\*\* . Correlation is significant at the 0.01 level (2-tailed).

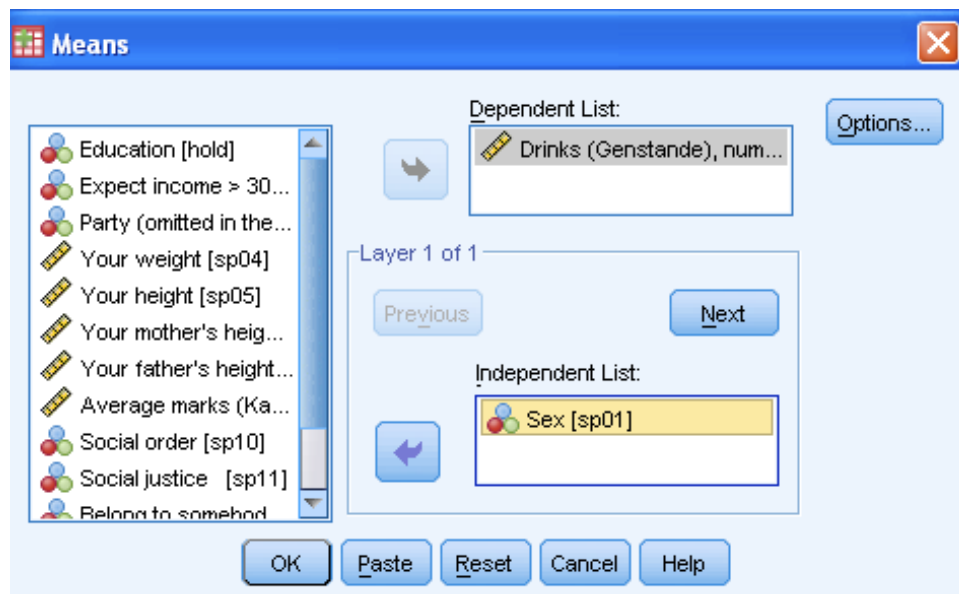
As can be seen from the output, there are significant correlations between the variables *your height* and *your weight*. On the contrary the correlation between *average marks* and the other variables is very small. Further the output-table shows two-tailed levels of significance for correlations between each variable and the total number of observations included in the correlation test.



## 12. Comparisons and test of means

### 12.1 Compare means

When you want to compare means grouped by another variable, this is possible by choosing *Analyze => Compare means => Means*. The variable you want the mean of should be put in dependent list, in the following example this will be the number of drinks. The variable that you want to group by should be put in *Independent List*, in this example sex. By pressing *Options* it is possible to choose different statistical measures that should appear in the output as standard the means, number of observations, and the standard deviations are shown.



This gives the following output where the means for males and females easily can be compared.

#### Report

Drinks (Genstande), number of in week 34			
Sex	Mean	N	Std. Deviation
Female	7,20	173	9,011
Male	15,26	282	13,033
Total	12,19	455	12,297

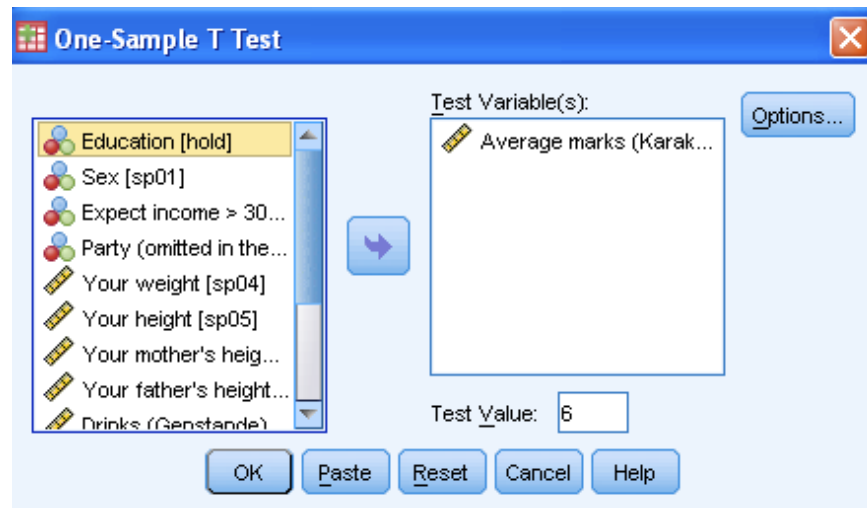
### 12.2 One sample T-test

A simple T-test is used, when you want to test whether the average of a variable is equal to a given mean; i.e. *one sample T-test*. E.g. you might want to test if the average mark for students at BSS is equal to the value 6. The hypothesis for this two-sided test would look like this:

$$H_0 : \mu_{Ave.mark} = 6$$

$$H_1 : \mu_{Ave.mark} \neq 6$$

The test procedure is the following: *Analyze => Compare means => One-sample T-Test*



Select the variables and enter the test value in the *Test Value* field. The value must be the same for each variable! Under 'Options...' you select the confidence level you want to use. As default this is set to 95%.

### 12.2.1 Output

In the following output it is tested whether the average mark for students at BSS is equal to the expected value 6.

One-Sample Test						
	Test Value = 6					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Average marks (Karakter) at qualifying exam	70,764	444	,000	2,4762	2,407	2,545

In the output both the t-value and the confidence interval are given. The most interesting thing to look at is the *Sig.* column, which gives the p-value of the test. As can be seen the p-value is almost zero, which indicates that the  $H_0$  hypothesis must be rejected; meaning that it cannot be said, with 95% confidence, that the mean of the tested variable is equal to 6.

## 12.3 Independent samples T-Test

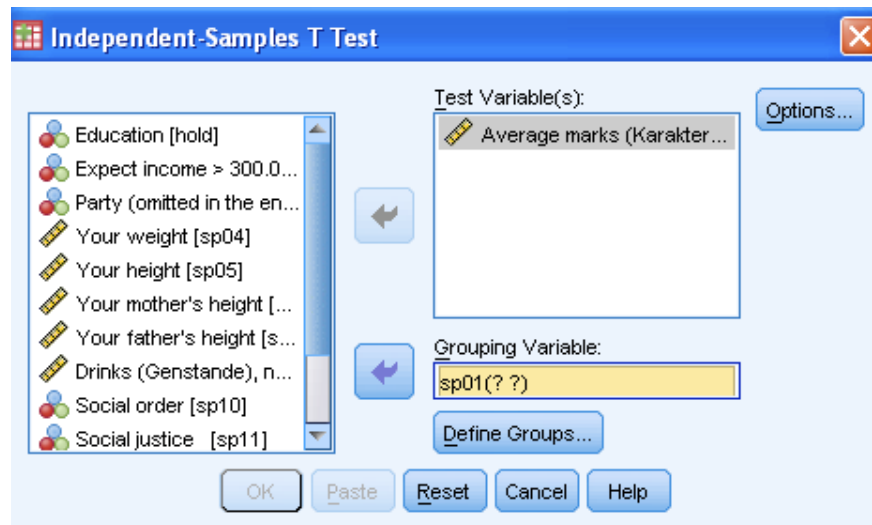
If you want to compare two means based on two independent samples you have to make an independent sample t-test. E.g. you want to compare the average mark for students at ASB for women versus men. The hypothesis looks as follows:

$$H_0 : \mu_{ma rkmen} = \mu_{ma rkwo men} \Leftrightarrow \mu_{ma rkmen} - \mu_{ma rkwo men} = 0$$

$$H_1 : \mu_{ma rkmen} \neq \mu_{ma rkwo men} \Leftrightarrow \mu_{ma rkmen} - \mu_{ma rkwo men} \neq 0$$

The test can only be performed for two groups. If you need to test more than two groups you need to use another test (ANOVA or GLM – see section 13 and 14). The test is performed by choosing the following:

Analyze => Compare Means => Independent-Samples T-test



- The variable *Average marks* is selected as the test variable.
- The variable *sex* is selected as grouping variable and 'Define Groups...' is used to specify the groups. In our example the two groups are: 1 (women) and 2 (men).
- Under 'Options...' you select the confidence interval to be used.

### 12.3.1 Output

The output will look like this (just a sample):

**Group Statistics**

Sex		N	Mean	Std. Deviation	Std. Error Mean
Average marks (Karakter) at qualifying exam	Female	170	8,534	,7123	,0546
	Male	275	8,440	,7528	,0454

**Independent Samples Test**

		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
Average marks (Karakter) at qualifying exam	Equal variances assumed	,195	,659	1,303	443	,193	,0938	,0720	Lower: -,0477 Upper: ,2352
	Equal variances not assumed			1,320	373,200	,188	,0938	,0710	Lower: -,0459 Upper: ,2334

The first table shows descriptive statistics, for the selected variable, after the split up. The last table shows the independent-samples T-test. To the left is Levene's test for the equality of variance. With a test value of 0,195 and a p-value of 0,659 we accept that there is variance equality. On the basis of this acceptance, we should use the first line to test the equality of the means. This gives a  $t_{obs}=1,303$  and a p-value of 0,193. Thereby we accept the null-hypothesis and we cannot, on the basis of the test say that there is a difference between the average mark for men and women.

## 12.4 Paired Samples T-Test <sup>1</sup>

A farmer has in the summer compared two combine harvesters. The farmer has used two farmhands to test them. They were tested on the same mark, right next to each other. This means they were exposed to same weather and same top-soil. The farmer has been testing which combine harvest that could produce the most.

In this example, a paired sample t-test is to prefer. The production is measured for production\_a for combine harvester a, and production\_b for combine harvester b. The dataset *Paired Sample t-test.sav* for the following test can be found in the downloaded zip-folder (see top of document)

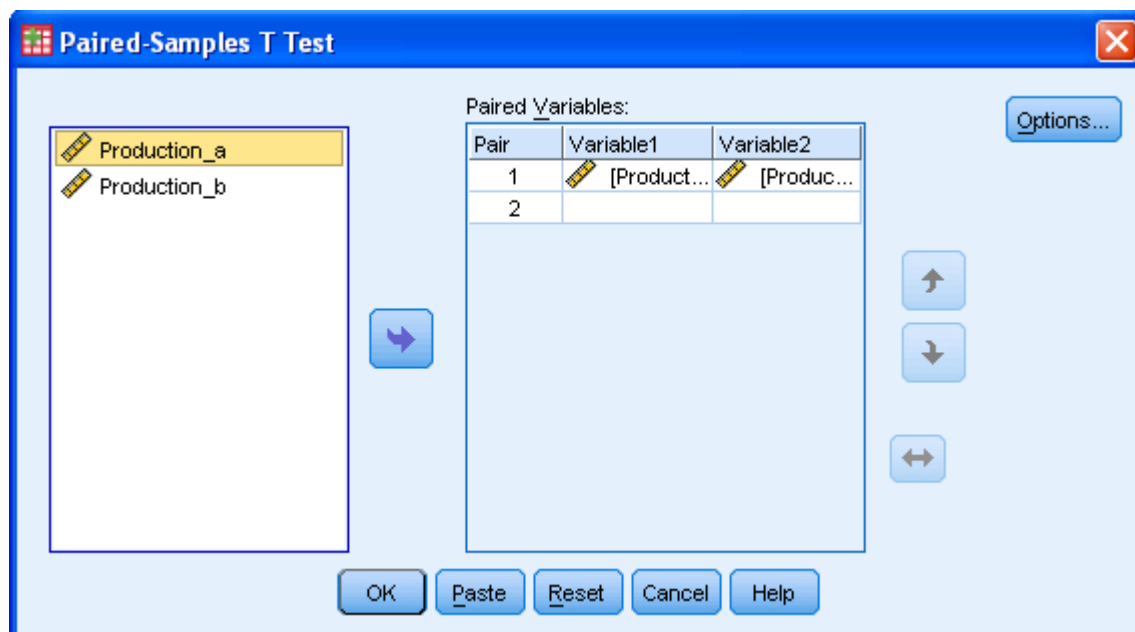
The hypothesis looks as follows:

$$H_0 : \mu_{\text{produktion\_a}} = \mu_{\text{produktion\_b}} \Leftrightarrow \mu_{\text{produktion\_a}} - \mu_{\text{produktion\_b}} = 0$$

$$H_1 : \mu_{\text{produktion\_a}} \neq \mu_{\text{produktion\_b}} \Leftrightarrow \mu_{\text{produktion\_a}} - \mu_{\text{produktion\_b}} \neq 0$$

The analysis is performed by selecting: *Analyze => Compare means => Paired-Samples T-test.*

Then the two variables are moved into the *Paired Variables* field:



The output will look almost like the one for the Independent samples T-Test. Note that both variables have to be selected before moved into *Paired Variables*.

### 12.4.1 Output

The output will look like this:

<sup>1</sup> Keller (2009) ch. 13.3 and E310 p. 25-26

**Paired Samples Test**

		Paired Differences				t	df	Sig. (2-tailed)	
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower				Upper
Pair 1	Production_a - Production_b	-2.00000	15.05747	4.34672	-11.56706	7.56706	-.460	11	.654

In the output both the t-value and the confidence interval are given. The most interesting thing to look at is the *Sig.* column, which gives the p-value of the test. As can be seen the p-value is 65,4%, which indicates that the H0 hypothesis cannot be rejected; meaning that it cannot be concluded that there is a difference between the two combine harvesters.

## 13. One-Way Anova<sup>2</sup>

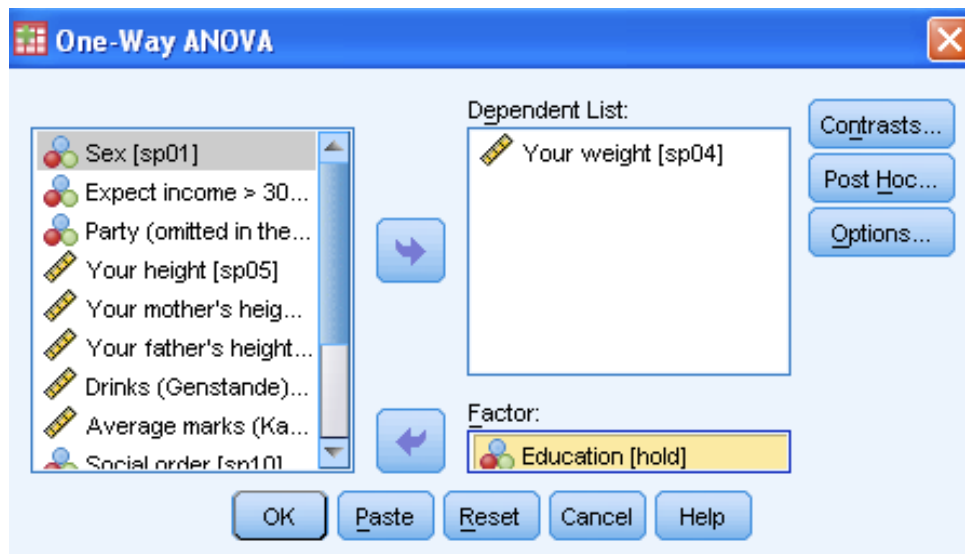
To test the hypothesis of equal means between more than two groups, an ANOVA test is to be applied. In the following a one-way ANOVA test will be shown through an example, where we want to test whether the weight of the students at BSS is the same between the different educations.

The following hypotheses will be tested:

$H_0$ : There is no difference in the population means ( $H_0 : \mu_{education\_1} = \mu_{education\_2} = \mu_{education\_3}$ )

$H_1$ : There is a difference in the population means ( $H_1$ : at least two means differ)

The test is done by selecting: *Analyze => Compare means => One-Way ANOVA*. Then a window looking like the one below appears and the dependent variable is selected and moved to the *Dependent list* box. The classification variable is moved to the *Factor* box.

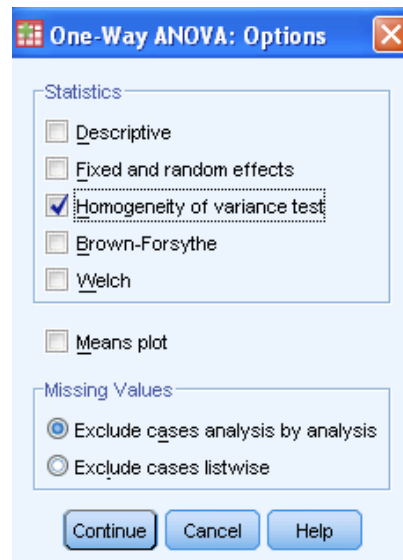


In this example the dependent variable is *your weight*. The classification variable is *education*, which describes the different educations at BSS. It should be noted that this variable must not be a string. If the variable is a string variable it must first be recoded into a numeric variable.

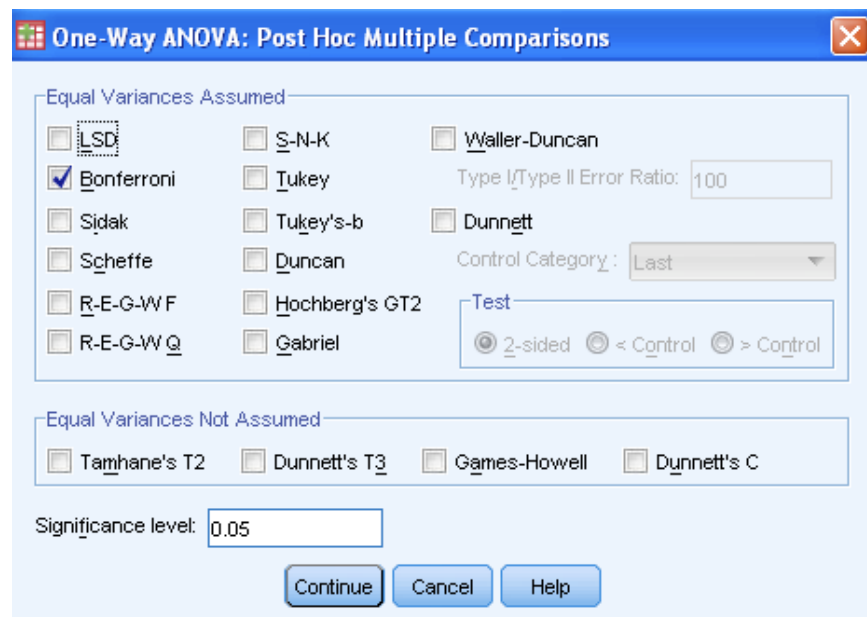
Further details must be specified before the test is to be completed:

- By selecting 'Options...' it is possible to include descriptive measures and tests for homogeneity of variances between groups (Levene's test), which is one of the tests of assumptions being performed before an analysis of variance.

<sup>2</sup> Keller (2009) ch. 15.1.



- By selecting 'Post Hoc...' it is possible to do several tests of differences between the groups.



This is done based on the outcome from the test of equal variance. It is usually recommended to use Bonferroni's test, which is selected in this test as well.

## 13.1 Output

The output from an ANOVA test is shown below

**Test of Homogeneity of Variances**

Your weight

Levene Statistic	df1	df2	Sig.
,736	4	446	,568

**ANOVA**

Your weight

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1064,936	4	266,234	1,788	,130
Within Groups	66408,332	446	148,898		
Total	67473,268	450			

**Post Hoc Tests****Multiple Comparisons**

Dependent Variable: Your weight

Bonferroni

(I) Education	(J) Education	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
HA1-6	HA7-10,dat	-,928	1,422	1,000	-4,94	3,08
	BA int	3,398	1,769	,554	-1,59	8,39
	HA jur	1,962	1,894	1,000	-3,38	7,30
	BSc B	1,637	2,436	1,000	-5,24	8,51
HA7-10,dat	HA1-6	,928	1,422	1,000	-3,08	4,94
	BA int	4,327	1,791	,161	-,73	9,38
	HA jur	2,890	1,913	1,000	-2,51	8,29
	BSc B	2,566	2,452	1,000	-4,35	9,48
BA int	HA1-6	-3,398	1,769	,554	-8,39	1,59
	HA7-10,dat	-4,327	1,791	,161	-9,38	,73
	HA jur	-1,436	2,184	1,000	-7,60	4,73
	BSc B	-1,761	2,669	1,000	-9,29	5,77
HA jur	HA1-6	-1,962	1,894	1,000	-7,30	3,38
	HA7-10,dat	-2,890	1,913	1,000	-8,29	2,51
	BA int	1,436	2,184	1,000	-4,73	7,60
	BSc B	-,325	2,752	1,000	-8,09	7,44
BSc B	HA1-6	-1,637	2,436	1,000	-8,51	5,24
	HA7-10,dat	-2,566	2,452	1,000	-9,48	4,35
	BA int	1,761	2,669	1,000	-5,77	9,29
	HA jur	,325	2,752	1,000	-7,44	8,09



From the first table it can be seen that the assumption of equal variances can be accepted (homogeneity of variance) since the p-value is above 0,05. Further it can be concluded from the middle table that the H0 hypothesis is not rejected since the p-value is 0,13. This indicates that there are no differences between the mean weights based on the different educations. The bottom table, which shows the differences between the groups, is not relevant in this case, but it should be mentioned that if H0 is rejected the differences are specified in this table indicated by a (\*).

## 14. General Analysis of Variance<sup>3</sup>

An analysis of variance is another statistical method to determine the existence of differences in group means. The one-way ANOVA described above only allows for one classification factor (one-way), whereas the following analysis allows multifactor analysis (i.e. randomized block design or two-factor ANOVA). In SPSS these are called GLM (General Linear Mean) procedures.

The following table will show which experimental design to apply in different cases.

Datatype? (groupingvar./testvar.)	Objective	Experimental design?	Identify where there is significant difference.
Nominal / Interval	Compare means	Randomized: One Way ANOVA	Bonferroni's simultaneous confidence intervals (or LSD or Tukey)
Nominal & Nominal / Interval		Block design: Two Way ANOVA (Sample blocked by known variances in the test variable (reduces the SSE))	
Nominal & Nominal / Interval		More factors: Two Factor ANOVA (Testing for mean differences across two factors)	Interaktion significant: Interpret on a Profile Plot.  If interaction is <i>not</i> included in the final model: Bonferroni's

In the following example it will be tested if the average mark of the exam qualifying for enrollment at the business school can be said to be influenced by sex and education as well as by an interaction between the two factors (i.e. a two-factor ANOVA).

*The full model* looks like this:

Average Marks =  $\mu$  + sex + education + sex\*education

$$Y = \mu + \alpha + \beta + \Gamma$$

<sup>3</sup> E281 ch. 7

The hypothesis for the test looks as follows

$\alpha$   $H_0$  : Sex has no effect on the average mark ( $\alpha = 0$ )

$H_1$  : Sex has a effect on the average mark ( $\alpha \neq 0$ )

$\beta$   $H_0$  : Education has no effect on the average mark ( $\beta = 0$ )

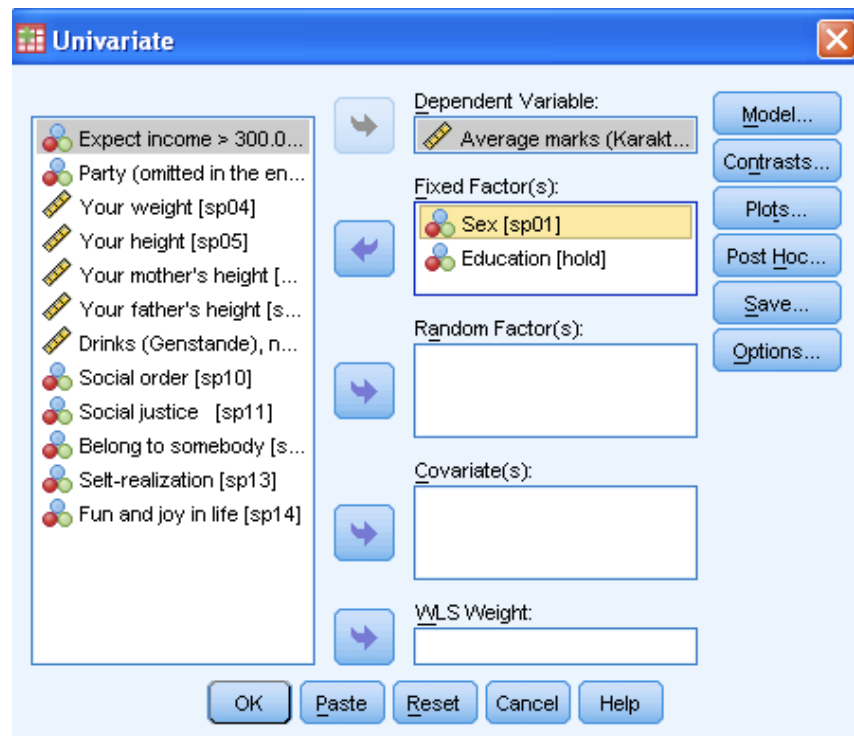
$H_1$  : Education has a effect on the average mark ( $\beta \neq 0$ )

$\Gamma$   $H_0$  : The combination of sex and education have no effect on the average mark ( $\Gamma = 0$ )

$H_1$  : The combination of sex and education have a effect on the average mark ( $\Gamma \neq 0$ )

Please note that in a two factor anova, there are 3 hypotheses one of each effect and one the interaction effect.

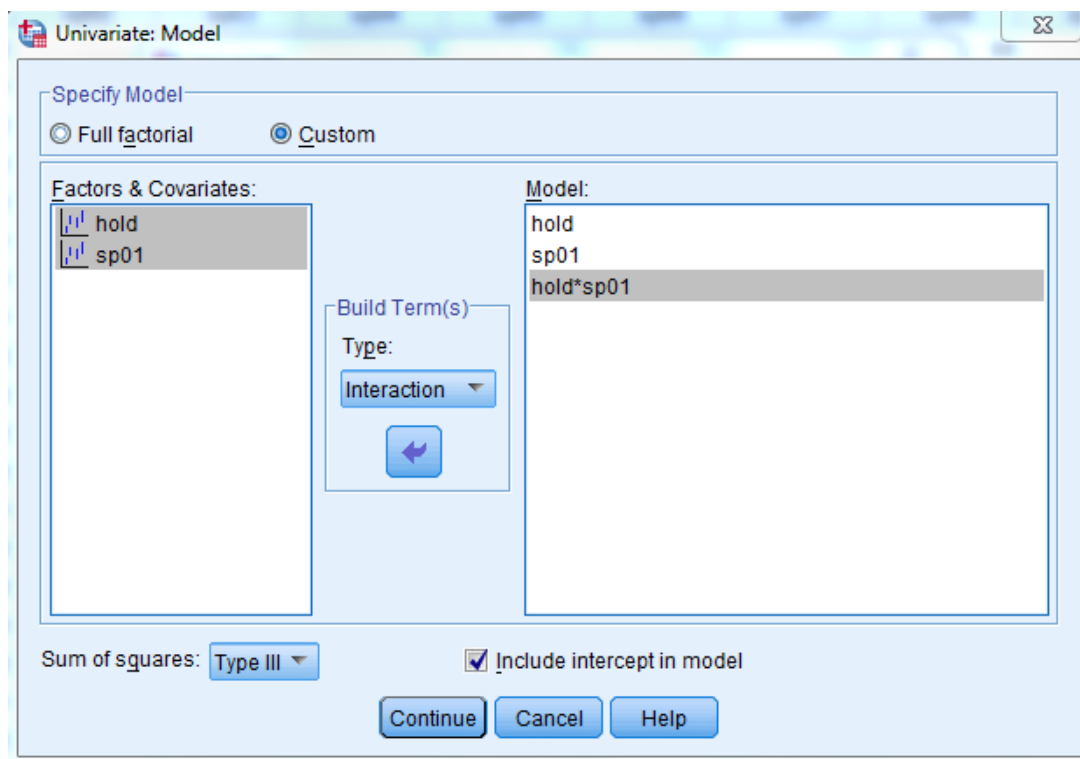
To test the model you select: *Analyze => General Linear Model => Univariate* and the following window will appear:



The dependent interval scaled variable is moved to the *Dependent variable* box. In the *Fixed factor(s)* box the classification variables are inserted. In our example this would be the nominal scaled variables *sex* and *education*.

Next the relevant model must be specified:

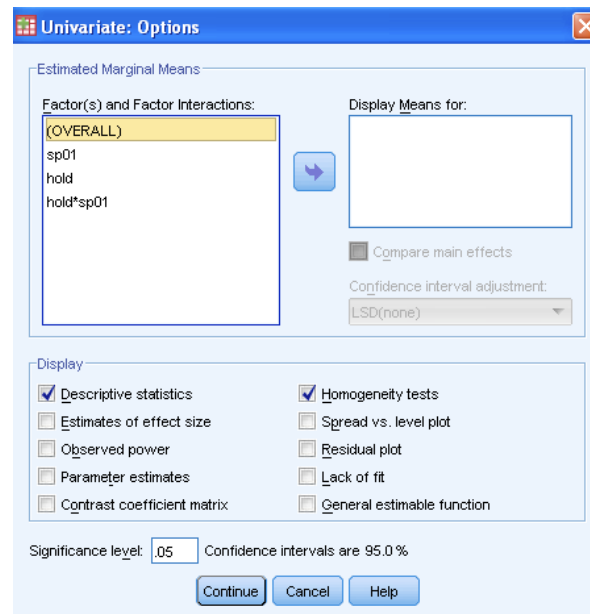
- When selecting 'Model...' you can either select *Full factorial* model or *Custom* model. In the former all the interaction levels are estimated and in the latter you are to specify the model yourself. We recommend you to use the latter model: *Custom* since it makes the eventually later reduction of the model easier. *Sum of squares* should always be set to *Type 3*.



You specify the model by clicking the effects you wish to include in the model and put them in the *Model*. In *Build Term(s)* you choose if you are making either main effects or interaction effects from the selected variables. If you want to include the interaction between *sex* and *education* in your model, you should select both variables and choose *Interaction* in the *Build Term(s)* and click the arrow button. It is important that you enter the effects in the right order so that the main effects are at the top and next the 1st order interaction effects, then 2nd order interaction effects etc.

- By selecting 'Options...' it is possible in the output to include the mean for each factor by placing them in the box *Display Means for* as shown below. If you want the grand mean  $\bar{X}$  included in the output, (*Overall*) should also be chosen.

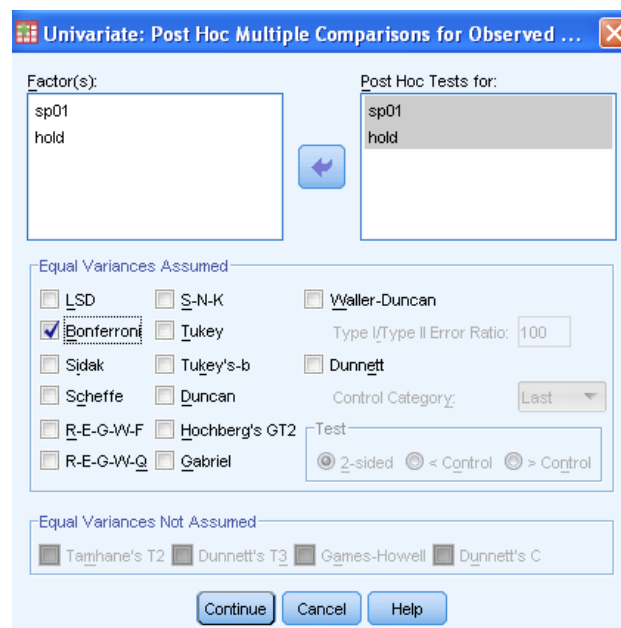
Further in SPSS it is possible to compare main effects (not interaction effects) by selecting *Compare main effects*.



*Homogeneity tests* should also be selected since it has a direct influence on the estimation.

Furthermore it is possible to change the level of significance for tests such as *Parameter estimates*, which by default is 0,05.

- By selecting 'Post hoc...', tests can be done for differences between the groups. If this is done based on the assumption of equal variance. The tests can be carried out using confidence intervals based on the Bonferroni principle e.g.



In the above window you move the factor(s) you wish to *Post hoc tests for*. Then you select the test you want completed. In our example the Bonferroni test is selected. Please note that simultaneous confidence levels for interaction effects cannot be computed this way.

## 14.1 GLM output

The output from the analysis is very similar to the output from ANOVA – though there are some improvements. It includes more information than the former method. The output will vary de-pending on the different options selected.

The ANOVA table is shown below:

Tests of Between-Subjects Effects					
Dependent Variable: Average marks (Karakter) at qualifying exam					
Source	Type I Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	9.545 <sup>a</sup>	9	1.061	1.985	.039
Intercept	31971.302	1	31971.302	59847.555	.000
hold	7.733	4	1.933	3.619	.006
sp01	.729	1	.729	1.364	.243
hold * sp01	1.083	4	.271	.507	.731
Error	232.382	435	.534		
Total	32213.230	445			
Corrected Total	241.928	444			

a. R Squared = .039 (Adjusted R Squared = .020)

As can be seen from the table both the main effect sex and the interaction are insignificant. Based on the hierarchical principle the interaction effect is always excluded first from the model, which then is re-tested before excluding the main effect sex. This is done by choosing 'Model...' and deleting the interaction from the model and moving it to the Factors & covariates. Then the test is run again. After this is done, it turns out that the main effect sex is still insignificant and it is then removed in the same way as the interaction effect. The table below shows the final model including only education as a significant explanatory variable. This means that we cannot reject the hypothesis for  $\alpha$  and  $\gamma$ . The only hypothesis we can reject is  $\beta$ .

As can be seen at the table below, the determination coefficient  $R^2$  is shown. It should be noted that you normally don't make conclusions based on this values in conjunction with analysis of variance. If you want to evaluate the significance of the model, the F-test should be used instead.

Tests of Between-Subjects Effects					
Dependent Variable: Average marks (Karakter) at qualifying exam					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7.733 <sup>a</sup>	4	1.933	3.632	.006
Intercept	19880.024	1	19880.024	37350.270	.000
hold	7.733	4	1.933	3.632	.006
Error	234.194	440	.532		
Total	32213.230	445			
Corrected Total	241.928	444			

a. R Squared = .032 (Adjusted R Squared = .023)

In the table below the models descriptive parameters are printed with both mean standard deviation and count.

### Descriptive Statistics

Dependent Variable: Average marks (Karakter) at qualifying exam

Education	Sex	Mean	Std. Deviation	N
HA1-6	Female	8.496	.7784	52
	Male	8.410	.7924	102
	Total	8.410	.7924	102
HA7-10,dat	Female	8.523	.6548	47
	Male	8.441	.7574	97
	Total	8.441	.7574	97
BA int	Female	8.820	.5754	35
	Male	8.555	.6399	33
	Total	8.555	.6399	33
HA jur	Female	8.193	.7364	27
	Male	8.293	.6512	30
	Total	8.293	.6512	30
BSc B	Female	8.722	.6476	9
	Male	8.723	.8738	13
	Total	8.723	.8738	13
Total	Female	8.534	.7123	170
	Male	8.440	.7528	275
	Total	8.440	.7528	275

As can be seen from the output the grand mean for sample is 8.476. While for instance BScB has an average of 8.723 and women from BA int has an average of 8.820.

As indicated by the table above there is some difference between the average marks across the different educations. The question is whether these differences are significant in a statistical sense. This is examined in the table below where the Post Hoc test is performed. Significant differences between groups, at significance level 0,05, are marked with the symbol (\*).

The conclusion for the test is that BA(int.) students have a significant higher average mark than HA(jur). At the same time it can be concluded that there are no significant differences between the remaining educations.

It should be noted that if the final model contains interaction effects, it doesn't make sense to use the described method to compare the different levels of the main effects. Instead the method on bonferroni intervals for interaction effects described in section 21.5 should be used. This allows for the necessary computations to be made to draw any conclusions about the interaction effect in the model.

**Multiple Comparisons**

Average marks (Karakter) at qualifying exam  
Bonferroni

(I) Education	(J) Education	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
HA1-6	HA7-10,dat	-.029	.0847	1.000	-.268	.210
	BA int	-.252	.1064	.182	-.552	.048
	HA jur	.193	.1133	.887	-.126	.513
	BSc B	-.284	.1666	.892	-.754	.186
HA7-10,dat	HA1-6	.029	.0847	1.000	-.210	.268
	BA int	-.223	.1075	.386	-.527	.080
	HA jur	.222	.1144	.524	-.100	.545
	BSc B	-.255	.1673	1.000	-.727	.217
BA int	HA1-6	.252	.1064	.182	-.048	.552
	HA7-10,dat	.223	.1075	.386	-.080	.527
	HA jur	.446*	.1313	.008	.075	.816
	BSc B	-.032	.1793	1.000	-.537	.474
HA jur	HA1-6	-.193	.1133	.887	-.513	.126
	HA7-10,dat	-.222	.1144	.524	-.545	.100
	BA int	-.446*	.1313	.008	-.816	-.075
	BSc B	-.477	.1835	.096	-.995	.040
BSc B	HA1-6	.284	.1666	.892	-.186	.754
	HA7-10,dat	.255	.1673	1.000	-.217	.727
	BA int	.032	.1793	1.000	-.474	.537
	HA jur	.477	.1835	.096	-.040	.995

Based on observed means.  
The error term is Mean Square(Error) = .534.

\*. The mean difference is significant at the .05 level.

## 14.2 Test of assumptions

A number of assumptions must be met to ensure the validity of the above analysis of variance. The following three assumptions will be checked in this section

- 1) Homogeneity of variance
- 2) Normally distributed errors
- 3) Independent error terms

### 14.2.1 Homogeneity of variance

To test for homogeneity of variance between the different groups in the analysis, use Levenes' test. In the following the assumption will be tested using Levene's test. The hypothesis for the test is

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_r^2$$

$$H_1 : \text{At least two are different}$$

To make SPSS run Levene's test, the function must be activated during the actual test for analysis of variance. This is done by choosing Analyze => General Linear Model => Univariate and under the 'Options...' button activate Homogeneity tests. This results in the following table being added to the original analysis of variance output.



### Levene's Test of Equality of Error Variances<sup>a</sup>

Dependent Variable: Average marks (Karakter) at qualifying exam

F	df1	df2	Sig.
2,359	4	440	,053

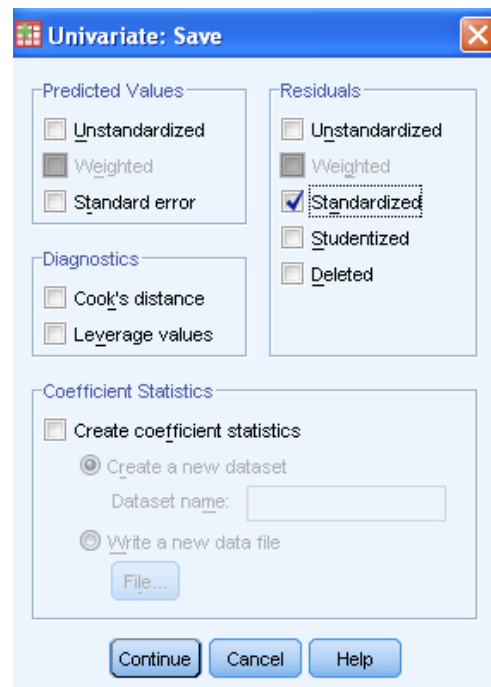
Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+hold

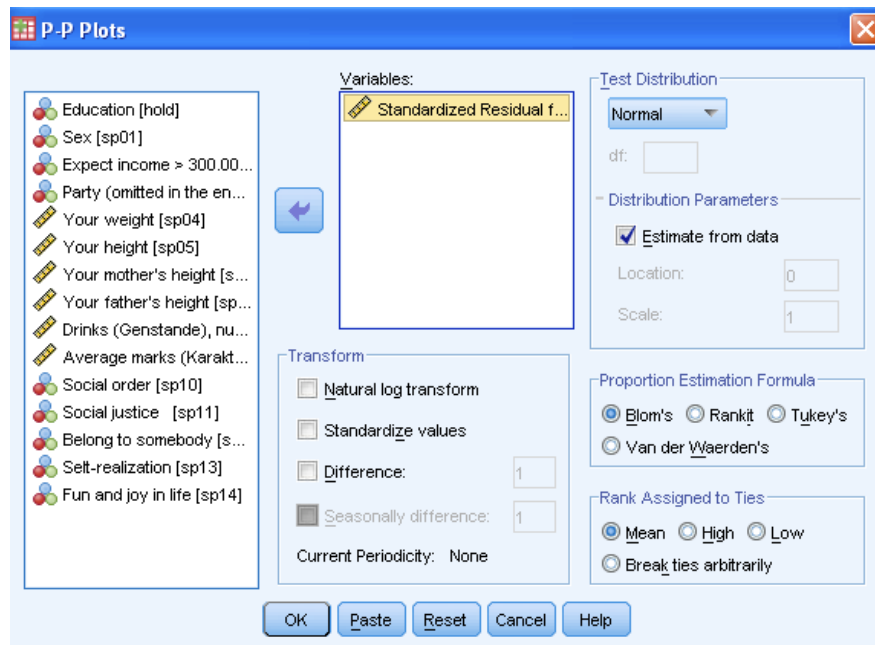
As it can be seen from the table above, the p-value for the test is 0,053, which means that we cannot reject the H0 hypothesis. In this case there is homogeneity of variance and the assumption is satisfied, even though the conclusion is quite sensitive to changing the level of significance.

#### 14.2.2 Normally distributed errors

The easiest way to test the assumption of normally distributed errors is by making a probit plot based on the standardized residuals. To generate the standardized residuals choose the sub-menu "Save" when performing the Univariate (Analysis of Variance) test, and select the "Standardized Residuals" option as illustrated below.

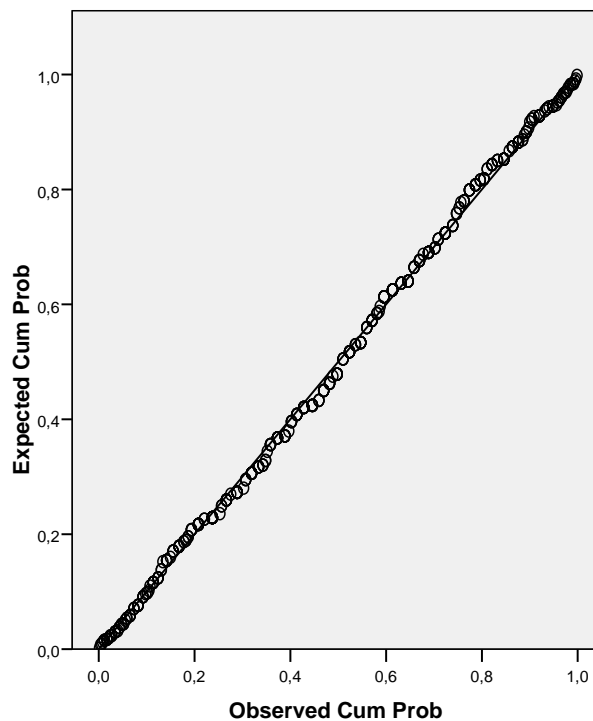


Next choose "Analyze=> Descriptive Statistics => P.P Plots ..." from the main menu and test the constructed residuals against a normal distribution.



As the observed standardized residuals are closely located around the 45 degree line, the assumption concerning normally distributed errors is assumed fulfilled.

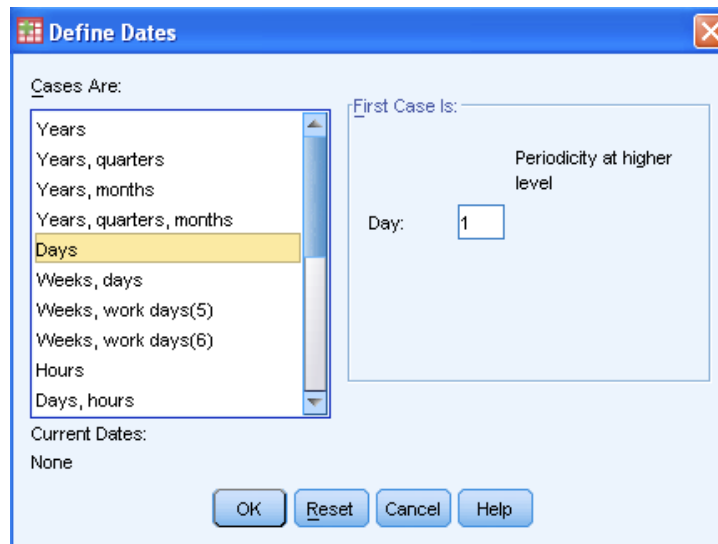
**Normal P-P Plot of Standardized Residual for sp09**



### 14.2.3 Independent errors

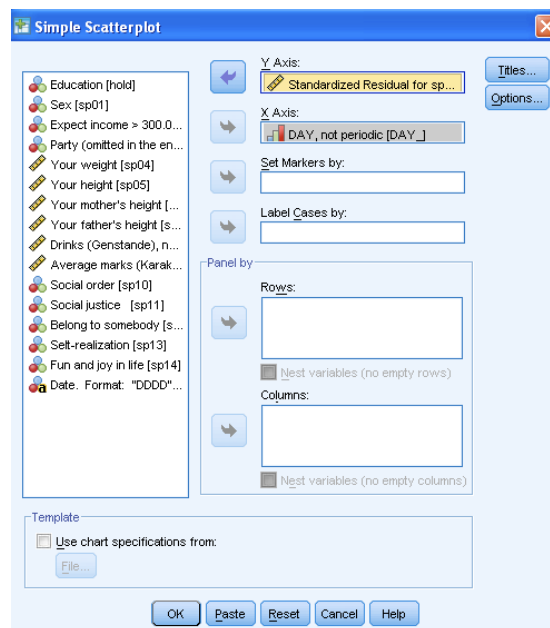
To test the assumption of independent errors is by making a scatter/dot. Here it is possible to plot the standardized residual against an observation number. If the assumption is to be satisfied, there should be no systematic variation/pattern in the plot.

If the dataset doesn't contain the observation numbers, they can be added under the menu Data => Define Dates...

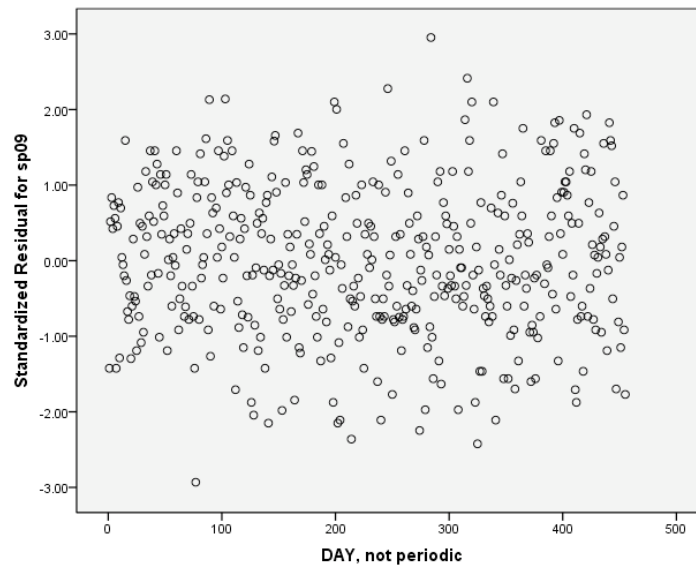


In the dialog box choose Days and press 'OK'. SPSS will now generate a new variable named Day.

Next choose "Graphs => legacy Dialogs => Scatter/Dot..." from the main menu and test the standardized residuals against observations.



Choose Standardized Residuals to be on the Y axis and the Day variable to be on the X axis, press ok.



## 15. Regression Analysis <sup>4</sup>

This chapter has been adapted by Jensen, Juhl & Mikkelsen.

A regression model is based on the assumption that a dependent variable can be explained by a linear relationship with one or more explanatory variables.

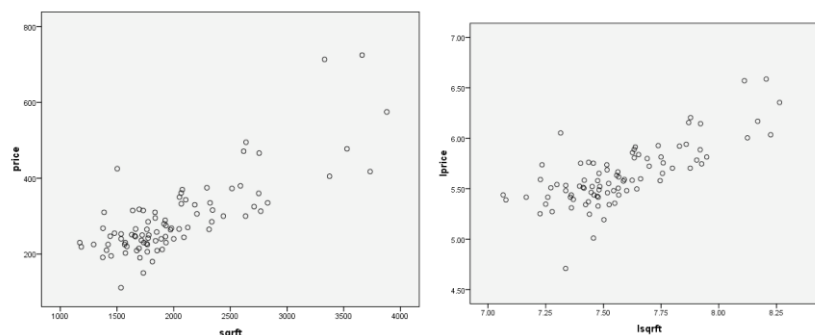
As an example we consider an equation that describes the determination of house prices. Let's say we have access to some observations on the following variables:

Price	=	House price
Lotsize	=	Size of the lot, in feet
Sqft	=	Square footage
Bdrms	=	Number of bedrooms
Colonial	=	1 if home is colonial style

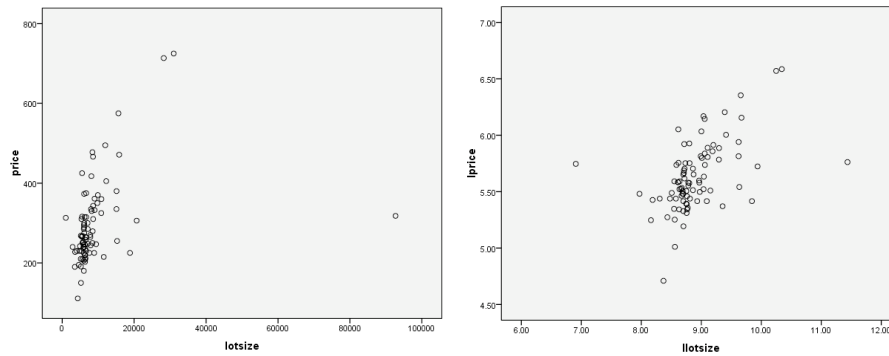
The dataset Regression.sav for the following test can be found in the downloaded zip folder (see top of document). It is somewhat backward to start from some variables that you have, so now we pretend that we start from the top of our step-by-step list:

**Step 1:** Problem statement: What determine the traded price on a residential house?

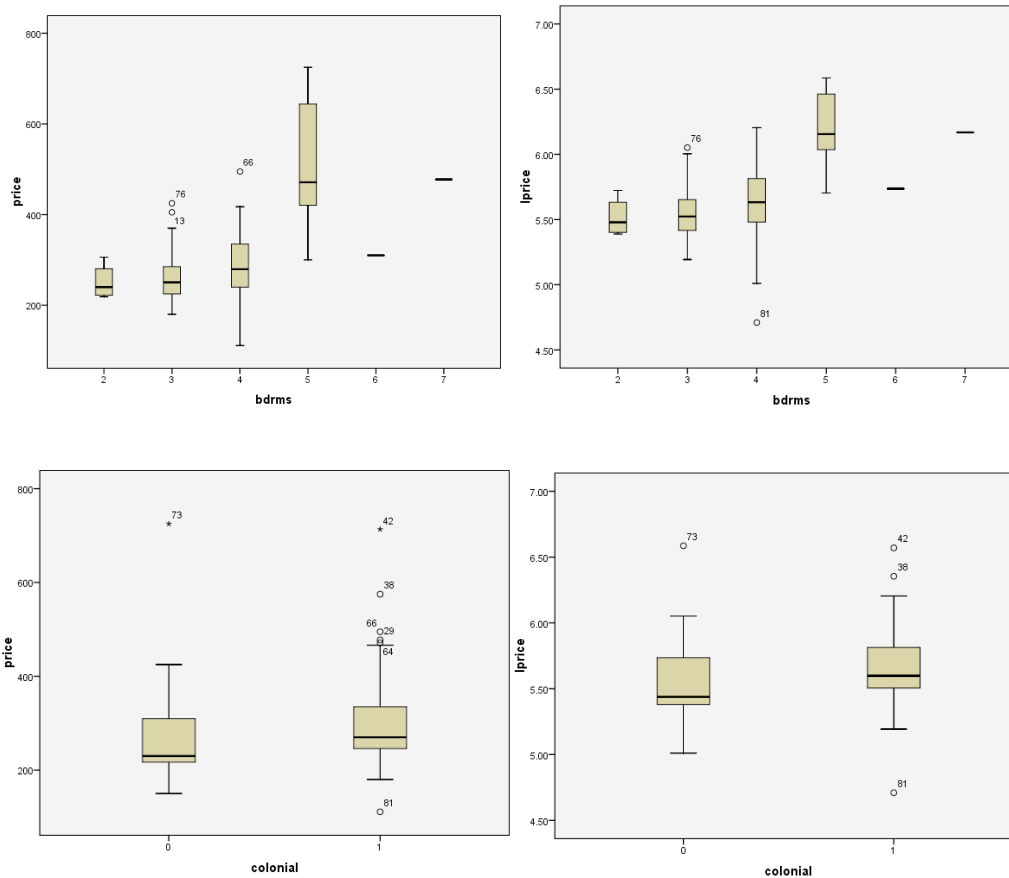
**Step 2:** There are of course many more relevant variables than the five we have access to, but we will have to do with these. The dependent variable is price, and the remaining four variables are the regressors. The variables price, lotsize, sqft and bdrms are all ratio scaled (even though bdrms is discrete). Colonial is a dummy variable, so it is a nominal. In this example all the regressors are actually variables of interest. Below are plots of the dependent variable against the regressors. We both do plots in levels and for the cases where it makes sense we also try log-arithmetic versions. The scatter/dots can be performed under Graphs => Legacy Dialogs => Scatter/Dot:



<sup>4</sup> Simple linear regression: Keller (2009) ch. 17. Multiple regression: Keller (2009) ch. 18



For the variables `bdrms` and `colonial` we do a boxplot since they are not interval scaled. The box plot can be performed under Graphs => Legacy Dialogs => Boxplot.



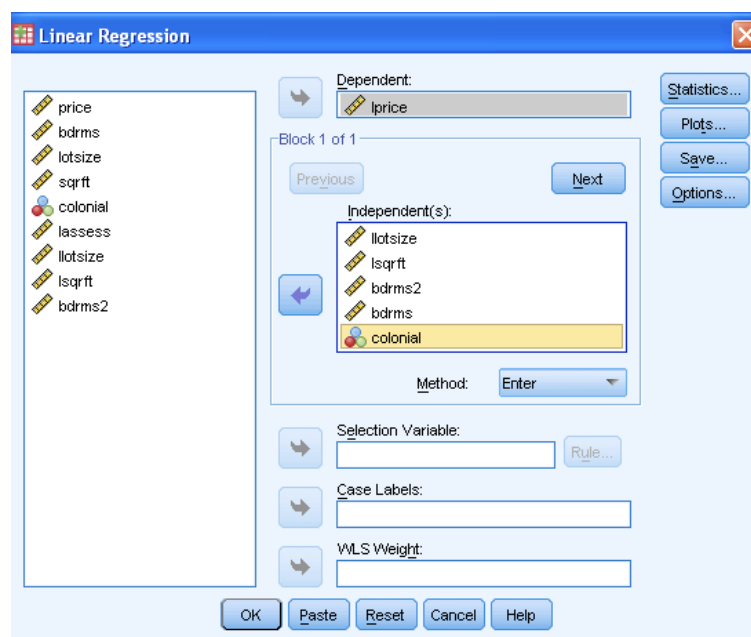
**Step 3:** Based on the plots and a wish to work with elasticities whenever possible we could start with the following regression equation:

$$\text{price}_i = \beta_0 + \beta_1 \text{llotsize}_i + \beta_2 \text{lsqrft}_i + \beta_3 \text{bdrms}_i + \beta_4 \text{colonial}_i + \varepsilon_i$$

It would actually make sense to expand this model with quadratic terms of at least `bdrms`. In theory and from the preliminary plot we should have the feeling that there should be decreasing returns to the number of bathrooms one puts into a house,

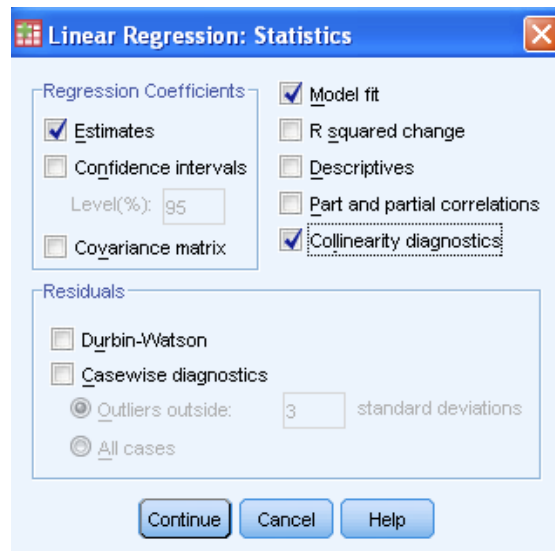
$$\text{price}_i = \beta_0 + \beta_1 \text{llotsize}_i + \beta_2 \text{lsqrft}_i + \beta_3 \text{bdrms}_i + \beta_4 \text{bdrms}_i^2 + \beta_5 \text{colonial}_i + \varepsilon_i$$

**Step 4:** to estimate the model using OLS, and get the output and save the variables we have to go to Linear Regression, which can be found under Analyze => Regression => Linear. The following window will appear.



- In Dependent you insert the dependent variable (left hand side), which in our example is the variable: `lprice`.
- The independent explanatory variables (left hand side) are to be inserted in the Independent(s) box. (`llotsize`, `lsqrft`, `bdrms2`, `bdrms` and `colonial`).
- Method is used when eliminating the independent explanatory variables automatically (either stepwise, remove or backward method), based on a given significance level in a multiple regression. This is not done by default.
- Selection variable is used to limit the analysis to include only the observations, which have a given value for a given variable.

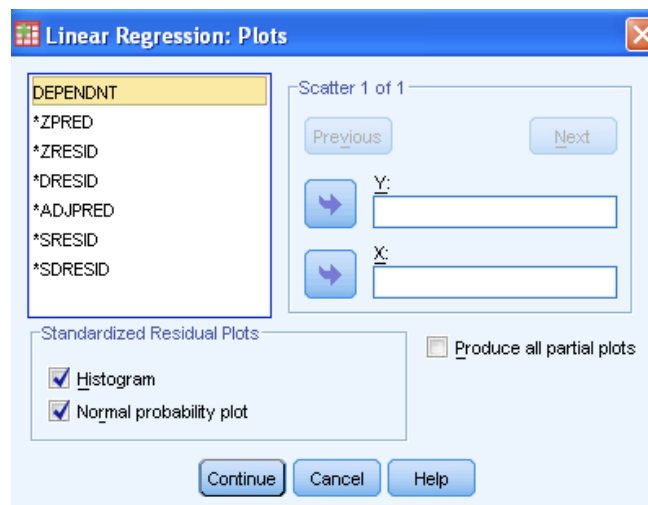
By selecting statistics we get the following window.



The 'Linear Regression: Statistics' dialog box is shown. It has two main sections: 'Regression Coefficients' and 'Residuals'. In the 'Regression Coefficients' section, the following options are checked: 'Estimates', 'Model fit', and 'Collinearity diagnostics'. The 'Confidence intervals' section has 'Level(%)' set to 95. The 'Residuals' section has 'Durbin-Watson' and 'Casewise diagnostics' checked. Under 'Casewise diagnostics', 'Outliers outside: 3 standard deviations' is selected. At the bottom are 'Continue', 'Cancel', and 'Help' buttons.

Here we have make sure to tick the collinearity diagnostics, since we are going to use it in assumption discussion, then click continue.

By selecting Plots we get the following window.

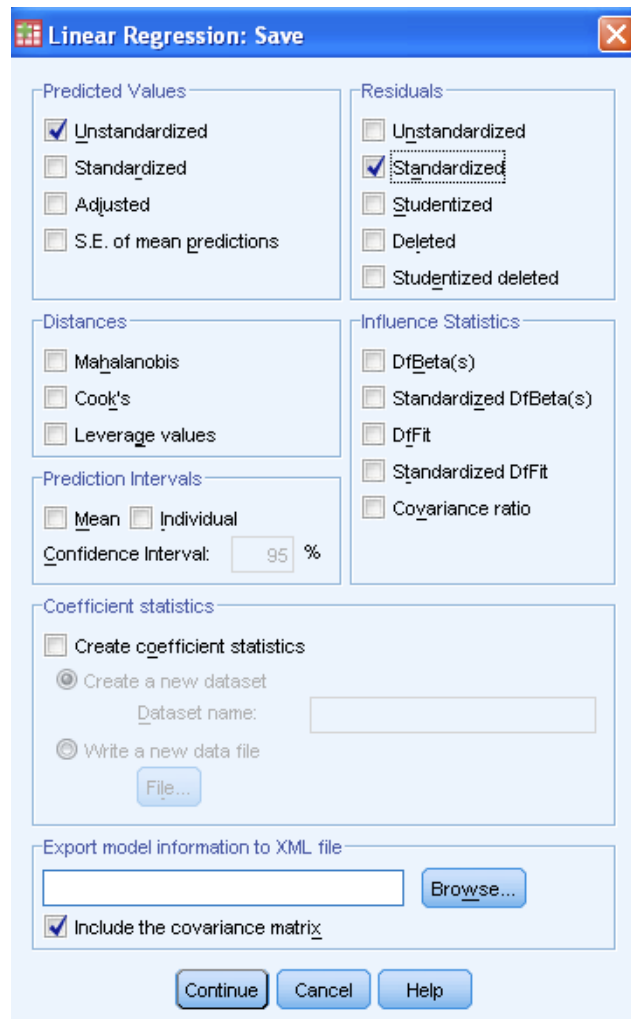


The 'Linear Regression: Plots' dialog box is shown. It has a list of dependent variables on the left: 'DEPENDNT', '\*ZPRED', '\*ZRESID', '\*DRESID', '\*ADJPRED', '\*SRESID', and '\*SDRESID'. On the right, there are 'Previous' and 'Next' buttons, and fields for 'Y:' and 'X:'. Below this, the 'Standardized Residual Plots' section has 'Histogram' and 'Normal probability plot' checked. There is also a checkbox for 'Produce all partial plots'. At the bottom are 'Continue', 'Cancel', and 'Help' buttons.

Here it is important to tick Histogram and Normal probability plot. This is as well going to be used in the assumption discussion. Continue has to be pressed when this is done.



By selecting Save, we get the following window.



The 'Linear Regression: Save' dialog box is shown. It contains several sections with checkboxes and options:

- Predicted Values:** ☒ Unstandardized, ☐ Standardized, ☐ Adjusted, ☐ S.E. of mean predictions.
- Residuals:** ☐ Unstandardized, ☒ Standardized, ☐ Studentized, ☐ Deleted, ☐ Studentized deleted.
- Distances:** ☐ Mahalanobis, ☐ Cook's, ☐ Leverage values.
- Prediction Intervals:** ☐ Mean, ☐ Individual, Confidence Interval: 95 %.
- Influence Statistics:** ☐ DfBeta(s), ☐ Standardized DfBeta(s), ☐ DfFit, ☐ Standardized DfFit, ☐ Covariance ratio.
- Coefficient statistics:** ☐ Create coefficient statistics, ☒ Create a new dataset (Dataset name: ), ☒ Write a new data file (File...).
- Export model information to XML file:** (Browse...), ☒ Include the covariance matrix.

Buttons at the bottom: Continue, Cancel, Help.

In Predicted Values the Unstandardized has to be ticked and in the Residuals the Standardized has to be ticked. Then click continue.

Now we can perform the test. Click ok and the following output appear.

Coefficients <sup>a</sup>								
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics		
	B	Std. Error	Beta			Tolerance	VIF	
1 (Constant)	-.947	.679		-1.395	.167			
lotsize	.161	.038	.289	4.255	.000	.894	1.118	
lsqrft	.720	.092	.613	7.842	.000	.672	1.487	
bdrms2	.029	.016	.646	1.825	.072	.033	30.432	
bdrms	-.206	.131	-.570	-1.576	.119	.031	31.836	
colonial	.068	.045	.104	1.519	.133	.876	1.142	

a. Dependent Variable: lprice

Before you even consider the significance of the individual regressors you must investigate whether the design criteria are satisfied. We now consider the design criterion that needs attention.

## 15.1 Test of design criteria

SPSS can be used to evaluate the design criteria's given by the regression model:

D1. Zero mean:  $E(\epsilon_i) = 0$  for all  $i$ .

D2. Homoskedasticity:  $\text{var}(\epsilon_i) = \sigma^2$  for all  $i$ .

D3. Mutually uncorrelated: and  $\epsilon_j$  uncorrelated for all  $i \neq j$ .

D4. Uncorrelated with  $x_1, \dots, x_k$ :  $\epsilon_i$  and  $x_1, \dots, x_k$  are uncorrelated for all  $i$  and  $j$ .

D5. Normality:  $\epsilon_i \sim \text{i.i.d.} - N(0, \sigma^2)$  for all  $i$ .

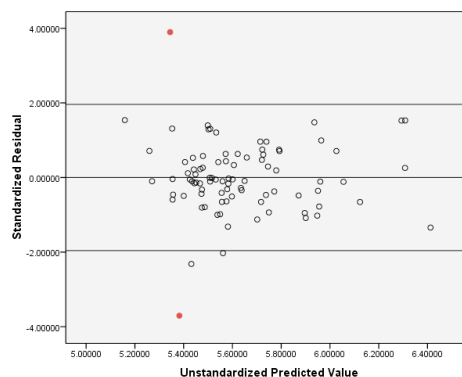
### 15.1.1 Zero mean: $E(\epsilon_i) = 0$ for all $i$ .

Always satisfied with a constant term in the model. Intuitively, the constant term equals the fixed portion of the dependent variable that cannot be explained by the independent variables, whereas the error term equals the stochastic portion of the unexplained value of the response.

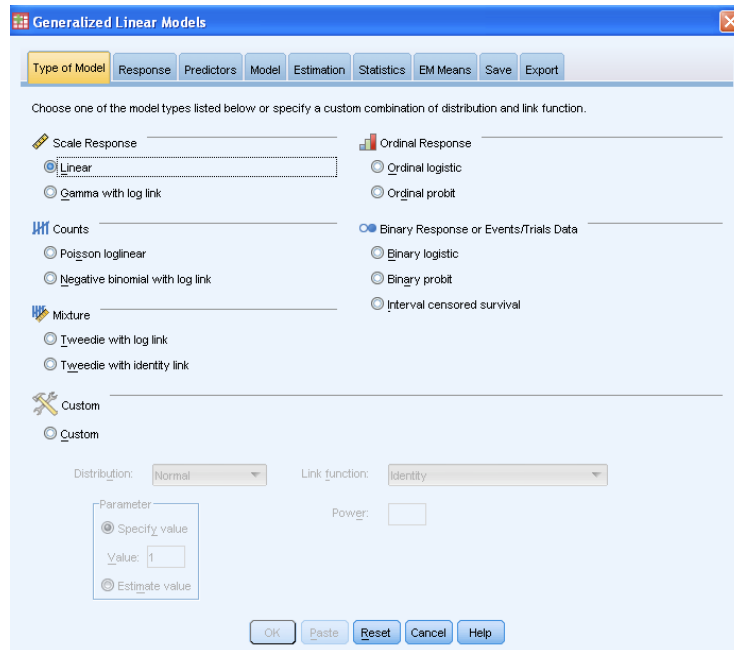
### 15.1.2 Homoscedasticity: $\text{var}(\epsilon_i) = \sigma^2$ for all $i$ .

One should make scatter plots of the residuals against the regressors. This can be quite lengthy, so a shortcut is just to plot the residuals or the squared residuals against the predicted values. If the residual variation changes as we move along the horizontal axis then we should be concerned. However, in most cases we need not worry about heteroscedasticity if we mechanically use robust standard errors. In general it is useful to compute both the OLS and the robust standard error. If they are the same then nothing is lost in using the robust one; if they differ then you should use the more reliable ones that allow for heteroskedasticity.

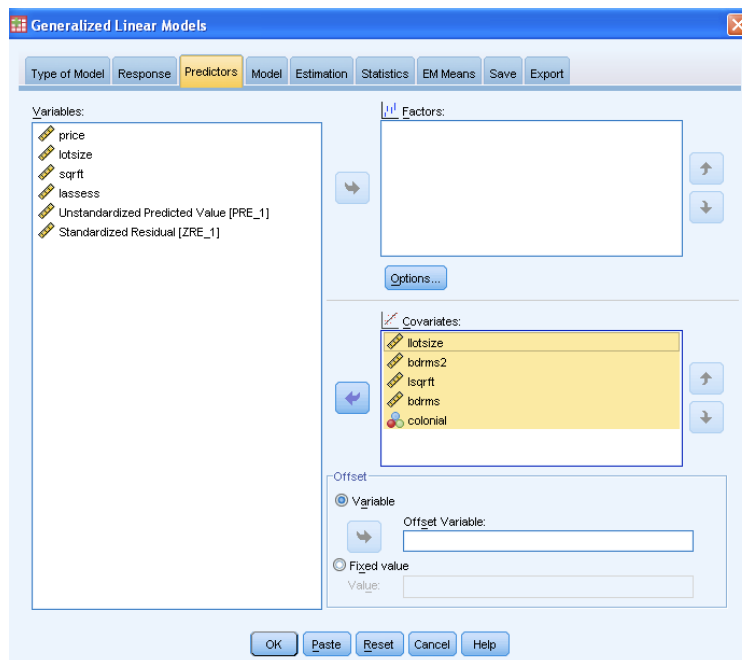
First we plot the residuals or the squared residuals against the predicted values,  $\hat{y}_i$ . Graphs => Legacy Dialogs => Scatter/Dot



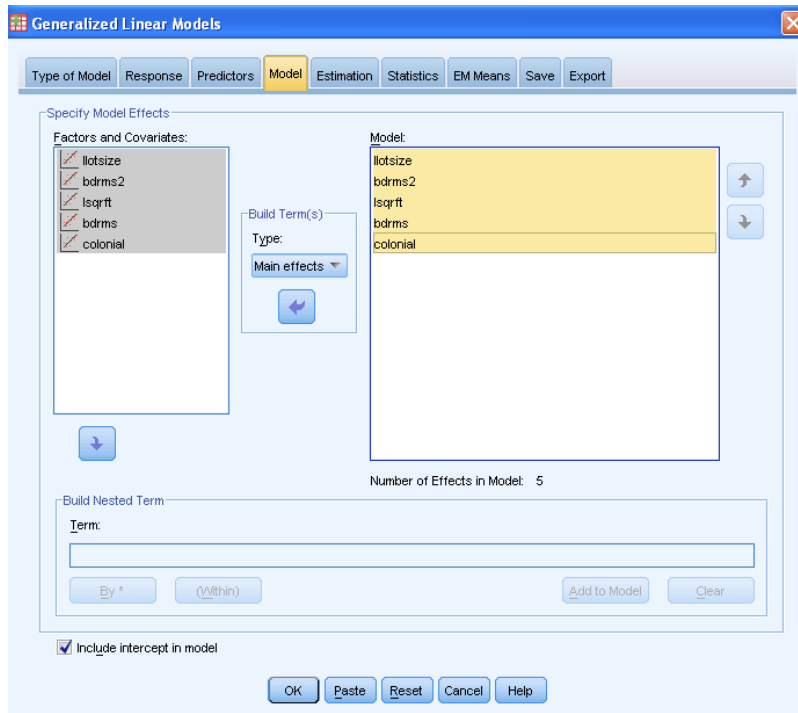
If we ignore the two outlying observations (marked with red in both plots) then there are no sign of heteroskedasticity. Still, these outliers could affect the results both in terms of the estimated coefficients and their precision. To further check for heteroskedasticity we compute the robust standard errors below. This can be done under Analyze => Generalized Linear Models => Generalized Linear Models. The following window will appear.



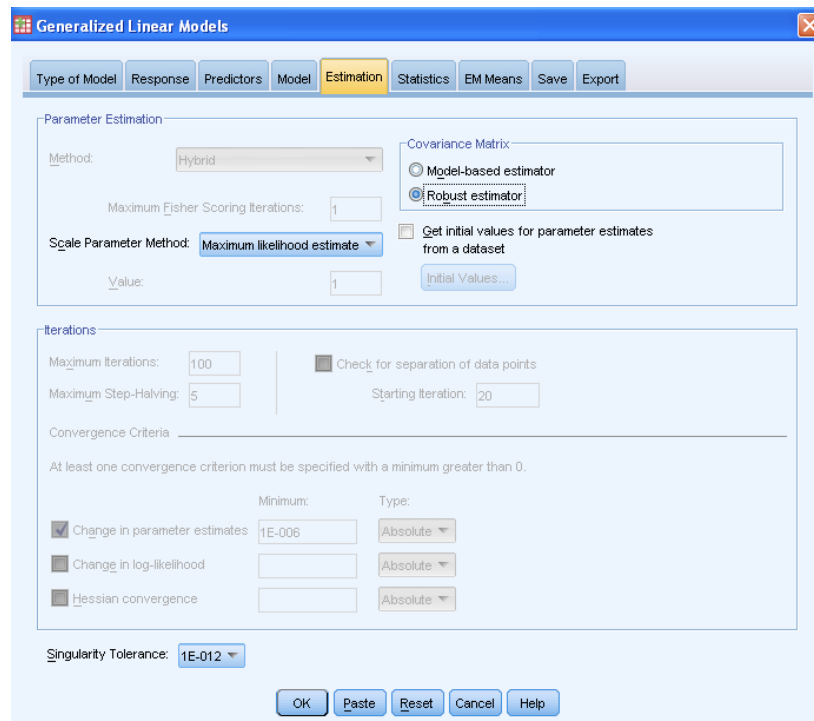
In Type of Model we have to make sure that Linear is ticked. This is normally done by default. Next we have to go to the Response tab.



Lprice has be moved to the dependent variable, this simply done by marking lprice and then click on the arrow.



In the Predictors tab all the explanatory variables has be moved to the Covariates box.



In the Model tab we define the model, this means we have to move all the explanatory variables to the model. The type has to be Main effects.

In the Estimation tab, the Robust estimator has to be ticked.

In the Statistics tab, we only want the Parameter estimates to be ticked. It can be found under Print.

Now we are ready to perform the test, this is done by pressing ok. The following output appears.

**Parameter Estimates**

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-.947	.7559	-2.429	.534	1.570	1	.210
colonial	.068	.0472	-.024	.161	2.081	1	.149
llothesize	.161	.0401	.082	.240	16.116	1	.000
lsqrft	.720	.0984	.527	.913	53.512	1	.000
bdrms	-.206	.1404	-.481	.069	2.147	1	.143
bdrms2	.029	.0176	-.006	.063	2.615	1	.106
(Scale)	1 <sup>a</sup>						

Dependent Variable: lprice

Model: (Intercept), colonial, llothesize, lsqrft, bdrms, bdrms2

a. Fixed at the displayed value.

In this table we are only interested in the standard errors (the coefficient estimates should be identical to those we got above, if they are not there is a bug somewhere), below we present them next to the ones from the OLS output:

	B	Std. Error	HRSE
Constant	-0.9470	0.6790	0.7559
lloftsize	0.1610	0.0380	0.0472
lsqrft	0.7200	0.0920	0.0401
bdrms	-0.2060	0.1310	0.0984
bdrms2	0.0290	0.0160	0.1404
colonial	0.0680	0.0450	0.0176

Since there are virtually no difference, the design criteria Homoskedasticity:  $\text{var}(\epsilon_i) = \sigma^2$  for all  $i$ , is fulfilled.

### 15.1.3 Mutually uncorrelated: and $\epsilon_j$ uncorrelated for all $i \neq j$

This assumption is typically only problematic in connection with timeseries data, since we normally only work with cross-sectional data, this assumption is fulfilled. Independence in the error term will be fulfilled if the data is collected randomly (so the sampling procedure should be the main focus, since there is no natural order of the observations).

### 15.1.4 Uncorrelated with $x_1, \dots, x_k$ : $\epsilon_i$ and $x_j$ are uncorrelated for all $i$ and $j$ .

Again one should make scatterplots of the residuals against the regressors or the predicted values,  $\hat{y}_i$ . If we find some kind of systematic pattern then we should try to expand the model to account for this. Possible solutions are to include omitted regressors or to alter the functional form. Another approach is to use so-called instrumental variables, a technic that we will not work with. The scatter/dot has already been created 15.1.1.2. But it looked as follows.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-.947	.7559	-2.429	.534	1.570	1	.210
colonial	.068	.0472	-.024	.161	2.081	1	.149
lloftsize	.161	.0401	.082	.240	16.116	1	.000
lsqrft	.720	.0984	.527	.913	53.512	1	.000
bdrms	-.206	.1404	-.481	.069	2.147	1	.143
bdrms2	.029	.0176	-.006	.063	2.615	1	.106
(Scale)	1 <sup>a</sup>						

Dependent Variable: lprice

Model: (Intercept), colonial, lloftsize, lsqrft, bdrms, bdrms2

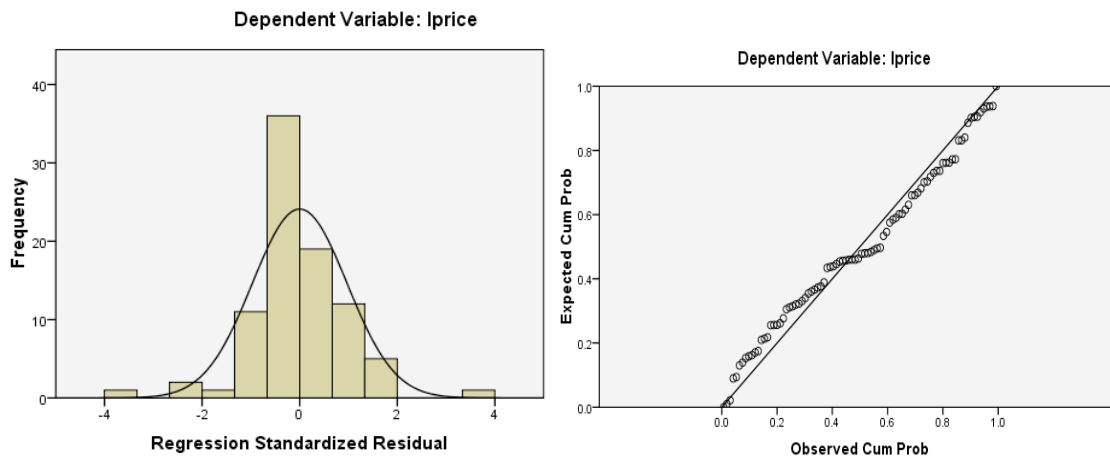
a. Fixed at the displayed value.

As you can see, it doesn't look like there is any pattern. Therefore the assumption is fulfilled. This is though the quick and dirty way to do it. The right approach is to make scatterplots of the residuals against the explanatory variables.

### 15.1.5 Normality: $\epsilon_i \sim \text{i.i.d.} - N(0, \sigma^2)$ for all $i$ .

If we have more than 100 observations, then we rarely care. The CLT ensures that in large samples the coefficient estimates are approximately normally distributed, and this holds for almost any choice of distribution for  $\epsilon_i$ . Still, it is very simple to make a histogram, P-P plots etc. that compares the residual distribution to the normal distribution. In small samples where the normal assumption fails to apply, we simply state this and note that the conclusion is to be taken lightly.

Since we already have created a pp-plot and histogram for  $Y$ . (This was created in the plots option, when we did test, see 15.1). The graphs we get looks as follows.



As it can be seen there is normality and the assumption is therefore considered fulfilled.

**Step 5:** Now with a model that approximately satisfies the design criterion, we can progress to simplify the model. There only seem to be serious multicollinearity to the bedroom terms. This can be seen in the output we had earlier in this chapter, but to make it easier for you, we have it shown below.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-.947	.679		-1.395	.167		
	llotsize	.161	.038	.289	4.255	.000	.894	1.118
	lsqrft	.720	.092	.613	7.842	.000	.672	1.487
	bdrms2	.029	.016	.646	1.825	.072	.033	30.432
	bdrms	-.206	.131	-.570	-1.576	.119	.031	31.836
	colonial	.068	.045	.104	1.519	.133	.876	1.142

a. Dependent Variable: lprice

We have to look at the collinearity statistics and then at VIF (Variance Inflation Factor). The basic rule is that, if VIF is higher than 5. This is the case with the variables  $\text{bdrms}^2$  and  $\text{bdrms}$ .

Before we do anything with the variables  $\text{bdrms}^2$  and  $\text{bdrms}$ , we will remove the variable  $\text{colonial}$  from the model. (It has the highest significance level). By re-estimating the model we get the following output.

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-.941	.684		-1.375	.173		
	lotsize	.162	.038	.291	4.254	.000	.895	1.118
	lsqft	.709	.092	.604	7.690	.000	.676	1.479
	bdrms2	.024	.016	.552	1.572	.120	.034	29.502
	bdrms	-.159	.128	-.442	-1.247	.216	.033	30.102

a. Dependent Variable: lprice

Before you can consider the significance of the individual regressors you normally have to investigate whether the design criteria are satisfied. In this case we will ignore it, and consider the design criteria's as fulfilled.

Now, the question is whether we should remove the bedroom terms. This is a hard question to answer based on the p-values only. The best approach we have is to estimate models with and without  $\text{bdrms}_i$  and/or  $\text{bdrms}_i^2$  and compare the adjusted  $R^2$  from these models. The adjusted  $R^2$  for the model:

$$\text{lprice}_i = \beta_0 + \beta_1 \text{lotsize}_i + \beta_2 \text{lsqft}_i + \beta_3 \text{bdrms}_i + \beta_4 \text{bdrms}_i^2 + (\text{equation 1})$$

is 0.637. The adjusted  $R^2$  can be seen in the model summary.

**Model Summary<sup>b</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.808 <sup>a</sup>	.653	.637	.18301

a. Predictors: (Constant), bdrms, lotsize, lsqft, bdrms2

b. Dependent Variable: lprice

Excluding the variable  $\text{bdrms}^2$  we get

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-1.297	.651		-1.992	.050		
	lotsize	.168	.038	.301	4.388	.000	.903	1.107
	lsqft	.700	.093	.597	7.540	.000	.679	1.473
	bdrms	.037	.028	.102	1.342	.183	.730	1.370

a. Dependent Variable: lprice



**Model Summary<sup>b</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.802 <sup>a</sup>	.643	.630	.18460

a. Predictors: (Constant), bdrms, llotsize, lsqft

b. Dependent Variable: lprice

and the adjusted  $R^2$  for the model is 0.630. In the model it is clear that bdrms is not significant, if we remove it we get

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-1.640	.602		-2.725	.008		
	llotsize	.168	.038	.302	4.380	.000	.903	1.107
	lsqft	.762	.081	.650	9.425	.000	.903	1.107

a. Dependent Variable: lprice

**Model Summary<sup>b</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.797 <sup>a</sup>	.635	.627	.18547

a. Predictors: (Constant), lsqft, llotsize

b. Dependent Variable: lprice

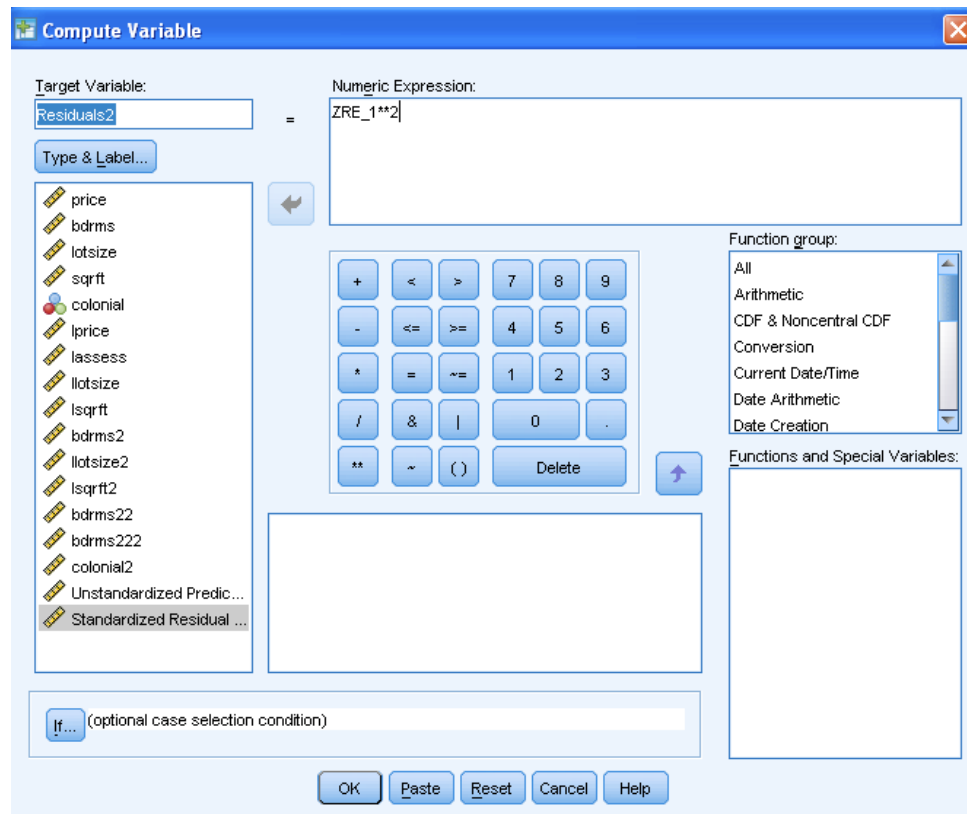
and the adjusted  $R^2$  for the model is 0,627. So comparing to this to the model with both bdrms and bdrms<sup>2</sup> we only lost about 1 percent of explanatory power. This is very little so it seems correct to remove bdrms and bdrms<sup>2</sup>.

**Step 6:** We are left with a constant elasticity model. A 1 % increase in llotsize increases the price of a house by 0.17 %. A 1 % increase in lsqft increases the price of a house by 0.76 %.

## 15.2 Further Topics

### 15.2.1 LM test for Heteroscedasticity

There is said to be heteroscedasticity if the assumption about constant variance for the residuals is broken. Meaning  $\text{var}(e_i)$  is not constant. The test is performed by testing whether the residuals raised to the power of 2 are related to any of the variables in the model. This means that you should make a regression model where you try to explain the squared residuals by different transformations of the original explanatory variables. To be able to do this you need to make new variables in the dataset one more time. In our example these would be:  $(\text{residual})^2$  and the explanatory variables raised to the power of 2 (Your weight 2) by the following transformation: Select Transform => Compute and a window like the one below will show:



In the field Target variable you enter the name of the new variable – Residuals2. The variable ZRE\_1 is raised to the power of 2 and you press “OK”. As explained earlier this needs to be done for all the explanatory variables and residuals.

These new variables need to be included into a regression model one more time with the following look:

$$e^2 = \beta_0 + \beta_1 * \text{llotsize} + \beta_2 * \text{llotsize}^2 + \beta_3 * \text{lsqft} + \beta_4 * \text{lsqft}^2 + \beta_5 * \text{bdrms}^2 + \beta_6 * \text{bdrms}^2 + \beta_7 * \text{bdrms} + \beta_8 * \text{bdrms}^2 + \beta_9 * \text{colonial} + \beta_{10} * \text{colonial}^2 + \epsilon_i$$

The model needs to be estimated to get the  $R^2$ -value since it is to be used in the  $n * R^2$  test. This number will then be evaluated in the  $\chi^2$ -distribution with k-degrees-of-freedom where k is the number of explanatory variables in the regression model. In our example k equals 10.

### 15.2.1.1 Output

**Model Summary<sup>a</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.280 <sup>a</sup>	.067	-.027	2.29352

The  $R^2$  is 0,067, we now have multiply it with n, in this case 88. We then get a  $X^2_{\text{obs}} = 5,896$ , since this is lesser than the critical value on 19,31. This means that we cannot reject that there is homoscedasticity.

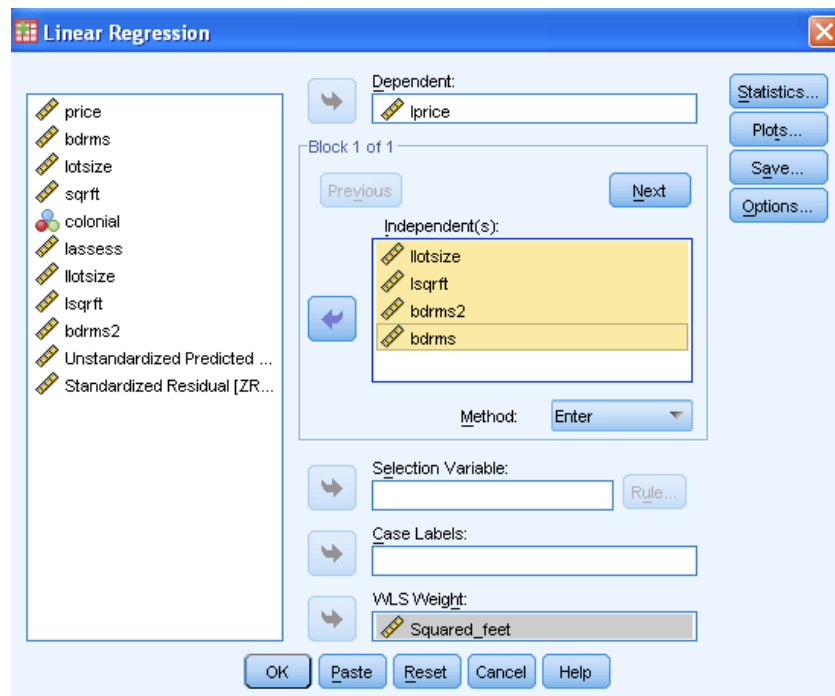
### 15.2.2 WLS

WLS (weighted least squares) is used when there is heteroscedasticity and one knows which variable that causes the problem.

In the following it will be assumed that it is the square feet variable that causes the heteroscedasticity and the variable `lsqrft` will therefore be used as weight in the WLS model. First step is to make a variable that can be used as square feet in this example the weight should be:

$$\frac{1}{lsqrft} \text{ (can be made in the compute menu).}$$

Now it is possible to make the WLS model. This is done in the same way as normal OLS regression by choosing Analyze => Regression => Linear. The only change is that under WLS Weight the variable that one just constructed should be added as shown below (the variable is in this example called `Squared_feet`).



The model can be used to evaluate if the different independent variables are significant, as the heteroscedasticity has been taken into consideration. One should remember that the interpretation of the new parameter estimates should be based on the OLS model, as the WLS models are only used to evaluate the size of the parameter estimates and whether these are significant.

## 16. Logistic Regression

When one has a binary variable as dependent variable, it is not possible to use the normal linear regression as described in the previous section. A binary variable only has two values (0 and 1) this could for example indicate yes/no or male/female. As the value only has two possible values a logistical transformation is made so the result is representing a probability. The following example will be based on the dataset logistic.sav and can be found in the down-loaded zip folder (see top of document)

This dataset concerns unemployment in 1990. In the regression model one wishes to find out how the probability of being unemployed in 1990 depends on a number of variables. The model is:

$$\text{Ln}\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{province} + \beta_3 \text{edu}$$

Where  $\pi$  is the probability of being unemployed in 1990 i.e. getting a value of 1.

### 16.1 The procedure

In SPSS the model can be made under: Analyze => Regression => Binary Logistic. And the below shown window should appear. The dependent variable should be put in the box Dependent, in this example the variable unem90, which has the value 1 when you are unemployed in 1990 and 0 if you had a job in 1990. The independent variable should be put in the box Covariates, in this example age, province and education. If one wants to include an interaction term this can be done by marking two variables at the same time and then pressing the button >a\*b>.



## 16.2 The output

The test gives the below shown output. The first table "Omnibus Tests..." is a test of the whole model like the F-test in linear regression. By looking in the row Model one can see the chi-square value, degrees of freedom and the p-value.

**Omnibus Tests of Model Coefficients**

		Chi-square	df	Sig.
Step 1	Step	161,093	3	,000
	Block	161,093	3	,000
	Model	161,093	3	,000

**Model Summary**

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2125,827 <sup>a</sup>	,074	,111

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

**Variables in the Equation**

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	age	-,044	,005	70,423	1	,000	,957
	province	,700	,119	34,538	1	,000	2,013
	educatio	-,169	,022	57,977	1	,000	,845
	Constant	1,844	,349	27,894	1	,000	6,322

a. Variable(s) entered on step 1: age, province, educatio.

In the last table the coefficients for each of the variables can be seen and under the column sig. the p-value is shown and one can see that all the variables are significant and we get the following model:

$$\text{Ln}\left(\frac{\pi}{1-\pi}\right) = 1.844 - 0.044\text{age} + 0.7\text{ province} - 0.169\text{edu}$$

The last column Exp(B) shows the odds ratio for each variable which is the change in the odds when the independent variable is changed by one unit.

## 17. Test for Homogeneity and independence<sup>5</sup>

The two different tests are applied, when you want to test the interaction between a number of data sets of nominal scale. The purpose of both tests is to test whether you can determine, that the outcome in one group or category is determined by the outcome in another group or category.

The best way to get a general view of the dataset is to make a table of frequencies. On the basis of this table, the test is similar to examine whether there is a connection between the count in the rows and columns.

Both tests are nonparametric, and can be solved by Analyze=> Nonparametric test in SPSS.

### 17.1 Difference between the tests

There are a number of differences between the two tests. First of all, the test for independence focuses on 2 variables in one sample. For instance the independence of sex on a specific education (this is the example we use below).

The test for homogeneity, focuses on whether the proportion of one variable is equal to 2 or more different groups/samples. One example of this is the interaction between the results of several different independent surveys.

The two tests have different assumptions, these will be dealt with in section 16.5.

The differences are mainly theoretically, when put into practice there is no difference, since both the  $\chi^2$ -observer and the SPSS procedure are similar. The only practical difference is therefore the tested hypothesis. The difference is as follows:

$H_0$  : No difference among the machines

$H_1$  : At least two are not equal

### 17.2 Construction of the dataset

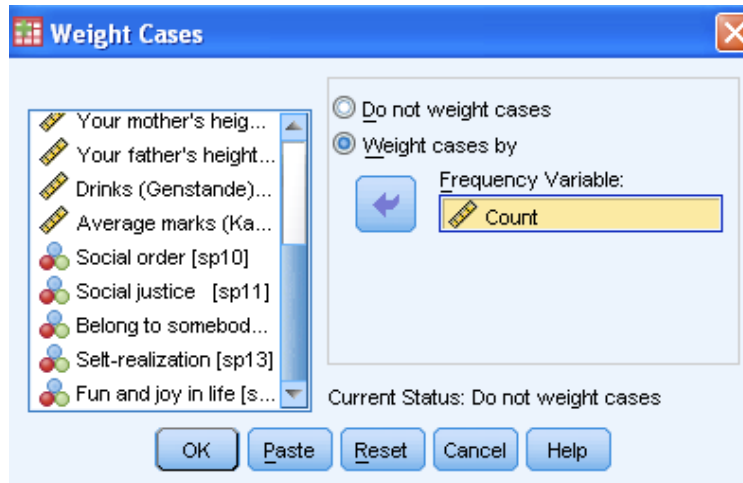
Prior to running the test in SPSS, it is important to ensure that the dataset is "built" right in the data view. If the dataset is not constructed in one of the following two ways, the output will be wrong.

In the example to the left, each respondent is shown as a separate row, which means that the number of respondents equals the number of rows in the dataset. On the picture to the right, each row is equal to the different possible outcomes a respondent can belong to.

	hold	sp01		hold	sp01	Count
1	BA int	Male	1	HA1-6	Female	40.00
2	HA1-6	Female	2	HA1-6	Male	103.00
3	HA7-10,dat	Male	3	HA7-10,dat	Female	40.00
4	BSc B	Male	4	HA7-10,dat	Male	50.00
5	HA1-6	Male	5	BA int	Female	55.00
6	BSc B	Female	6	BA int	Male	67.00
7	BA int	Female	7	HA jur	Female	64.00
8	BA int	Female	8	HA jur	Male	56.00
9	HA1-6	Male	9	BSc B	Female	45.00
10	HA jur	Male	10	BSc B	Male	42.00
11	BSc B	Male				

<sup>5</sup> Keller (2005) ch. 16, E281 ch. 12.1.1 and 12.1.2, and E282 ch. 7.

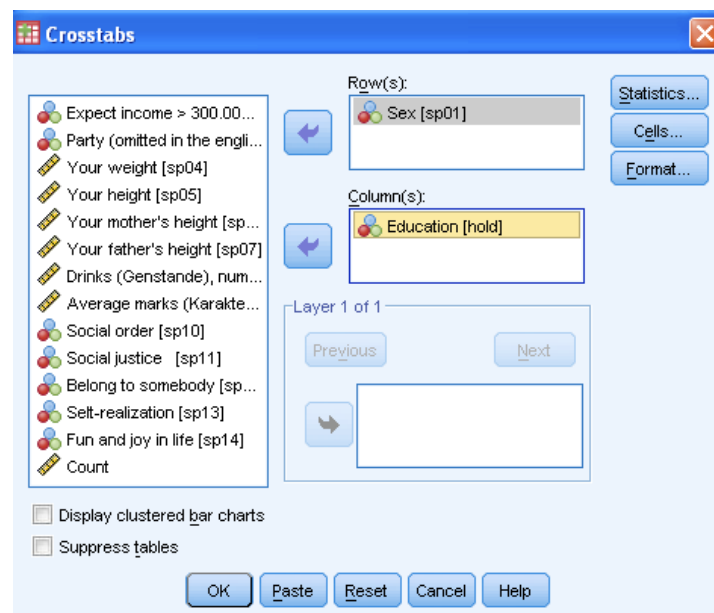
The variable count states the number of respondents in each group. If your dataset is constructed as in the first (left) example, you are ready to start the actual analysis. If, on the other hand, your dataset is constructed like on the picture to the right, you need to weight the dataset by the count variable. Weighting the dataset can be done by selecting Data => Weight cases and the following screen will appear:



Choose Weight cases by and click the Count variable into the Frequency variable field. The dataset is now prepared for the test.

## 17.3 Running the tests

We will now show an example of a test for independence, which uses data from the survey Rus98\_eng.sav. The purpose of the test is to examine whether there is any kind of interaction between the sex of the students and their chosen education.



In SPSS you can choose Analyze=> Descriptive Statistics=> Crosstabs and the above screen appears. The (2) relevant variables must be moved to either Row(s) or Column(s) respectively. It is of no relevance for the analysis, which variables are in Rows and Columns.

After selecting the relevant variables, you need to make some different selections. This is done with the buttons to the right:

- If statistics is chosen the screen below will appear. On this screen you have the opportunity to choose which test statistics you would like to have in your output. Both tests uses the following  $\chi^2$ -observerator:

$$\chi_{(r-1)(c-1)}^2 \sim \sum_{i=1}^r \sum_{j=1}^c \frac{O_{ij} - E_{ij}}{E_{ij}}$$

You can choose this observatory by marking Chi-square and then press 'Continue'. If the 'Cells...' button is chosen, you can select which informations /statistics needed in the output. The screen plot below shows the different options. The most commonly selected are Ob-served and Expected, if these are chosen, the respective values are shown on the output. These two values are used to compute the  $\chi^2$ -observerator. Furthermore Standardized in the Residual box (compare to  $\pm 1,96$ ) has to be selected. When the desired options are selected, press 'Continue'.



SPSS will now return to the 'Crosstabs statistics' dialog box, and if you have made all the desired selections, press 'OK' and SPSS will start the analysis.

## 17.4 Output

If the analysis is run, with the selections shown above, the output will look like the one below. The screen plot below will be basis for further interpretation.

Sex * Education Crosstabulation								
			Education					Total
			HA1-6	HA7-10,dat	BA int	HA jur	BSc B	
Sex	Female	Count	53	47	36	27	10	173
		Expected Count	58,9	54,8	26,2	21,7	11,4	173,0
		Std. Residual	-,8	-1,0	1,9	1,1	-,4	
	Male	Count	102	97	33	30	20	282
		Expected Count	96,1	89,2	42,8	35,3	18,6	282,0
		Std. Residual	,6	,8	-1,5	-,9	,3	
Total		Count	155	144	69	57	30	455
		Expected Count	155,0	144,0	69,0	57,0	30,0	455,0

The contingency table contains the information chosen in the previous section. The output contains another table, which is shown below. This table contains the test statistics, being the  $\chi^2$ -observerator and the p-value. Of the different values, always focus on the Pearson Chi-Square value.

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10,992 <sup>a</sup>	4	,027
Likelihood Ratio	10,806	4	,029
Linear-by-Linear Association	3,000	1	,083
N of Valid Cases	455		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 11,41.

In this analysis we get a test value of 10,992 which equals the sum of the squared standardized residuals;  $(-0,8)^2 + (-1,0)^2 + 1,9^2 + \dots = 10,992$

The corresponding p-value, that is  $P(\chi_4^2 > 10,992)$  is 0,027.

On basis of the normal  $\alpha$ -level of 0,05, the  $H_0$  hypothesis is rejected, and the conclusion is therefore that there is dependence between sex and the selected education.

When doing either a test for independence or a test for homogeneity, it is important to continue the analysis, to establish what caused the conclusion. In this case to find out where the dependence is. To evaluate this, we focus on the standardized residuals from the first table. The standardized residuals are calculated by the following formula:

$$SR_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

That is the difference between the observed and expected values divided by the square root of the expected value.

By looking at the contingency table, we find the largest value to be 1,9 with the girls on BA (int). This indicates that there are significantly more girls on this education, compared to a situation where there is independence. The second largest value is -1,5, this value shows that there are fewer boys on BA (int) than expected, if  $H_0$  was true.

The final conclusion of our analysis is that between the two observed variables, there is a certain degree of dependence. The primary reason for this dependence is the high number of girls on BA (int) and thereby the less number of boys. There is though none of the dependencies which are significant, because all the std. residuals  $< |1,96|$ . This is quite unusual because the rejection of  $H_0$  often means, that there exist at least one significant dependency where the std.res  $> |1,96|$ .

## 17.5 Assumptions

One assumption for using a  $\chi^2$ -observer is that the distribution can be approximated to this. If the two different non parametric test has to be approximated to this distribution, one of the assumptions is that the expected value in each cell  $E_{ij}$  is greater than 5. Whether this assumption is fulfilled can be seen from the contingency table in the previous section. In this example, the smallest expected value is 11,4, and the assumption is therefore approved.

If the assumption is not approved, that is if there are cells with an expected value less than 5, the approximation cannot be accepted. If this happens, you need to merge some of the different classes, so that the assumption will be accepted. For further information on how to merge different classes see section 4 on data processing.

The last assumption for the two tests is that the variables follow either a k-dimensional hypergeometric distribution or a multinomial distribution. These two distributions are similar to respectively the hyper geometric – and the binomial distribution, they just have more than two possible outcomes.

## 18. Factor

The following presentation and interpretation of Factor Analysis of the results are based on:

- "Videregående dataanalyse med SPSS og AMOS", Niels Blunch 1. udgave 2000, Sy-stime. Chp. 6, p. 124-155
- "Analyse af markedsdata", 2 rev. Udgave 2000, Systime. Chp. 3, p. 87-118
- Hair et.al. (2006): Multivariate Analysis 6th Ed., Pearson, kap. 3

### 18.1 Introduction

Factor analysis can be divided into component analysis as well as exploratory and confirmative factor analysis. The three types of analysis can be used on the same data set and builds on different mathematical models. Component analysis and exploratory factor analysis still produce relatively similar results. In this manual only component analysis is described.

In component analysis the original variables are transformed into the same number of new variables, so-called principal components, by linear transformation. The principal components have the characteristic that they are uncorrelated, which is why they are suitable for further data processing such as regression analysis.

Furthermore, the principal components are calculated so that the first component carries the bulk of information (explains most variance), the second component carries second-most information and so forth. For this reason component analysis is often used to reduce the number of variables/components so that the last components with the least information are disregarded. Henceforth the task is to discover what the new components correspond to, which is exemplified below.

### 18.2 Example

Data: Faktoranalysedata.sav from the downloaded zip folder (see top of document)

The following example is based on an evaluation of a course at the Aarhus School of Business. 162 students attending the lecture were asked to fill out a questionnaire containing various questions regarding the assessment of the course. Each of the assessments were based on a five point Likert-scale, where 1 is "No, not at all" and 5 is "Yes, absolutely".

The questions in the survey were:

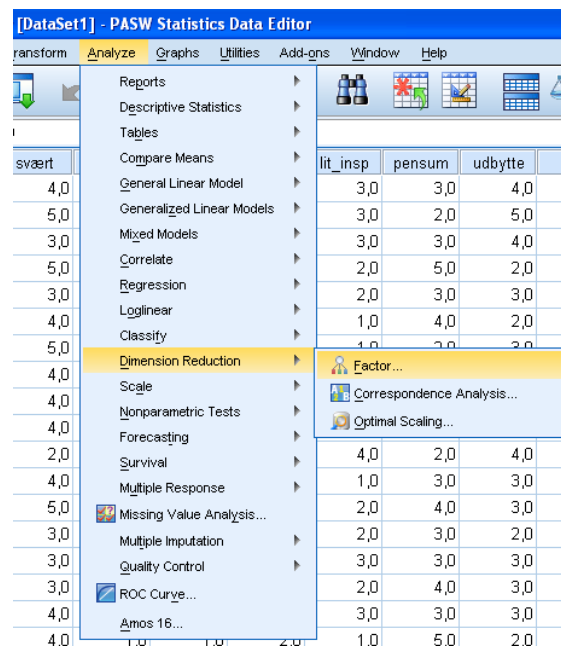
Question	Label name
Has this course met your expectations?	Met expectations
Was this course more difficult than your other courses?	More difficult than other courses
Was there a reasonable relationship between the amount of time spent and the benefit derived?	Relationship time/Benefit
Have you found the course interesting?	Course Interesting
Have you found the textbooks suitable for the course	Textbooks suitable
Was the literature inspiring?	Literature inspiring
Was the curriculum too extensive?	Curriculum too extensive
Has your overall benefit from the course been good?	Overall benefit

The original data consists of more assessments, but these have not been included in the example.

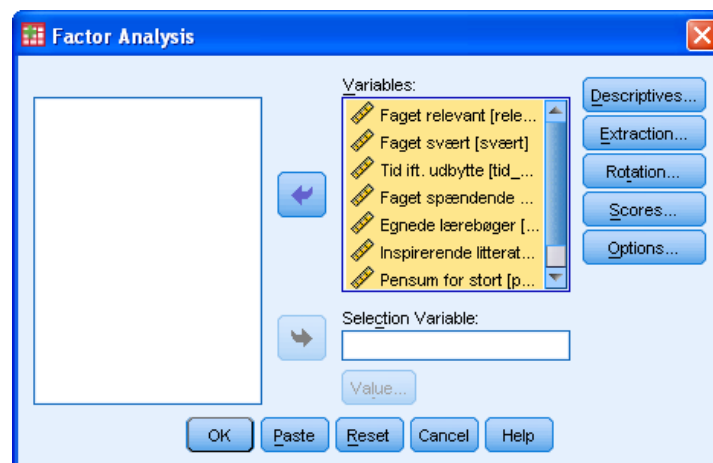
The purpose is now, based on the survey, to carry out a component analysis with a view to reduce the number of assessment criteria in a smaller number of components. Furthermore, the new components should be examined with a view to name them.

### 18.3 Implementation of the analysis

Component analysis is a method, which is used exclusively for uncovering latent factors from manifest variables in a data set. Since these fewer factors usually form the basis of further analysis, the component analysis is to be found in the following menu:



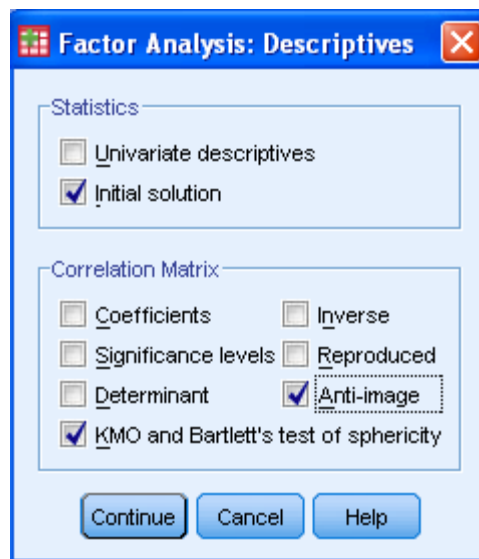
This results in the dialogue box shown below. The variables to be included in the component analysis are marked in the left-hand window, where all numeric variables in the data set are listed, and moved to the Variables window by clicking the arrow. In this case all the variables are chosen.



It has now been specified, which variables SPSS should base the analysis on. However, a more definite method for performing the component analysis has yet to be chosen. This is done by means of the Descriptives, Extraction, Rotation, Scores and Options buttons. These are de-scribed individually below.

### 18.3.1 Descriptives

By clicking the Descriptives button the following dialogue box appears:



In short the purpose of the individual options is as follows:

#### Statistics

- Univariate descriptives includes the mean, standard deviation and the number of useful observations for each variable.
- Initial solution includes initial communalities, eigenvalues, and the percentage of variance explained.

#### Correlation Matrix

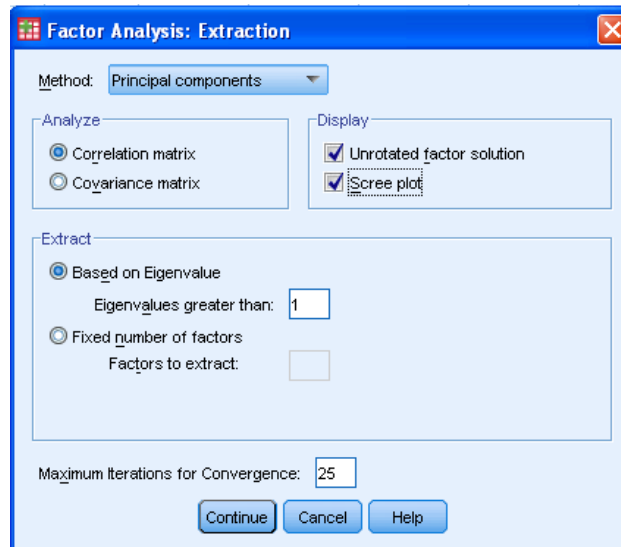
- Here it is possible to get information about the correlation matrix, among other thing the appropriateness to perform factor analysis on the data set.

In this example Initial solution is chosen, because a display of the explained variance for the suggested factors of the component analysis is desired. At the same time this is the most wide-ly used method. Anti-Image, KMO and Bartlett's test of sphericity are checked as well to analyze the appropriateness of the data analysis on the given data set.

With regards to the tests selected above, it may be useful to add a few comments. The Anti-Image provides the negative values of the partial correlations between the variables. Anti-Image ought therefore to be low indicating that the variables do not differ too much from the other variables. KMO and Bartlett provide as previously mentioned measures for the appropriateness as well. As a rule of thumb one could say that KMO ought to attain values of at least 0.5 and preferably above 0.7 to indicate that the data is suitable for a factor analysis. Equivalently the Bartlett's test should be significant, indicating that significant correlations exist be-tween the variables.

### 18.3.2 Extraction

By clicking the Extraction button the following dialogue box appears:



This is where the component analysis in itself is managed. In this example it has been chosen to use the Principal components method for the component analysis. This is chosen in the Method drop-down-box.

Since the individual variables of this example are scaled very differently, it has been chosen to base the analysis on the Correlation matrix. Cf. a standardization is carried out.

A display of the un-rotated factor solution is wanted in order to compare this with the rotated solution. Therefore, Unrotated factor solution is activated in Display.

Since the last components do not explain very much of the variance in a data set, it is standard practice to ignore these. This results in a bit of lost information (variance) in the data set, but in return a more simple output is obtained for further analysis. In addition, it makes the interpretation of the data easier.

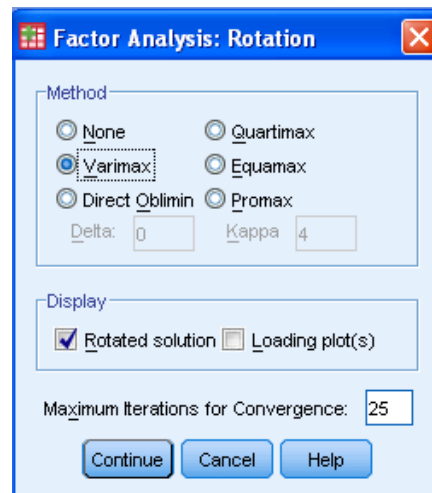
The excluded components are treated as “noise” in the data set. The important question is just how many components to exclude without causing too much loss of information. The following rules of thumb for choosing the right number of components for the analysis apply:

- Scree plot: By selecting this option in Display a graphical illustration of the variance of the components appears. A typical feature of this graph is a break on the curve. This curve break forms the basis of a judgment of the right number of components to include.
- Kaiser's criterion: Components with an eigenvalue of more than 1 are included. This can be observed from the Total Variance Explained table or the scree plot shown in section 8.4. However, these are only guidelines. The actual number of chosen factors is subjective, and it depends strongly on the data set and the characteristics of the further analysis.

If the user wants to carry out the component analysis based on a specific number of factors, this number can be specified in Number of factors in Extract. Last but not least it is possible to specify the maximum number of iterations. Default is 25. This option is not relevant for this ex-ample, but for the Maximum Likelihood method it could be relevant.

### 18.3.3 Rotation

Clicking the Rotation button results in the following dialogue box:



In brief, rotation of the solution is a method, where the axes of the original solution are rotated in order to obtain an easier interpretation of the found components. In other words it is assured that the individual variables are highly correlated with a small proportion of the components, while being low correlated with the remaining components.

In this example Varimax is chosen as the rotation method, since it ensures that the components of the rotated solution are uncorrelated. The remaining methods will not be further described here.

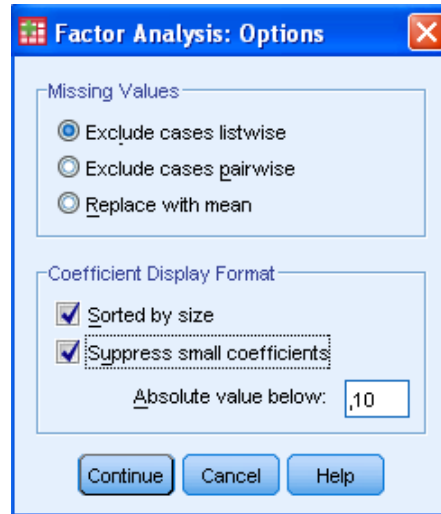
A display of the rotated solution has been chosen in Display. Loading plots enables a display of the solution in a three-dimensional plot of the first three components. If the solution consists of only two components, the plot will be two-dimensional instead. With the 12 assessment criteria in this example, a three-dimensional plot looks rather confusing, and it has therefore been ignored here.

### 18.3.4 Scores

Now the solution of the component analysis has been rotated, which should have resulted in a clearer picture of the results. However, there are still two options to bear in mind before the analysis is carried out. By selecting Scores it is possible to save the factor scores, which is sensible if they are to be used for other analyses such as profile analysis, etc. These factor scores will be added to the data set as new variables with default names provided by SPSS. In this example this option has not been chosen, as no further analysis including these scores is to be performed.

### 18.3.5 Options

Treatment of missing values in the data set is managed in the Options dialogue box shown below:



Missing values can be treated as follows:

- Exclude cases listwise excludes observations that have missing values for any of the variables.
- Exclude cases pairwise excludes observations with missing values for either or both of the pair of variables in computing a specific statistic.
- Replace with mean replaces missing values with the variable mean.

In this example Exclude cases listwise has been chosen in order to exclude variables with missing values. With regard to the output of the component analysis there are two options:

- Sorted by size sort's factor loading and structure matrices so that variables with high loadings on the same factor appear together. The loadings are sorted in descending order.
- Suppress absolute values less than makes it possible to control the output so that coefficients with absolute values less than a specified value (between 0 and 1) are not shown. This option has no effect on the analysis, but ensures a good overview of the variables in their respective factors. In this analysis it has been chosen that no values below 0.1 is shown in the output

## 18.4 Output

When the various settings in the dialogue boxes have been specified, SPSS performs the component analysis. After clicking OK a rather comprehensive output is produced, of which the most relevant outputs are commented below.

**KMO and Bartlett's Test**

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,791
Bartlett's Test of Sphericity	Approx. Chi-Square	413,377
	df	28
	Sig.	,000



Anti-image Matrices

		Faget relevant	Faget svært	Tid ift. udbytte	Faget spændende og interessant	Egnede lærebøger	Inspirerende litteratur	Pensum for stort	Samlede udbytte
Anti-image Covariance	Faget relevant	,497	,070	-,172	-,221	-,032	,080	-,005	-,076
	Faget svært	,070	,876	-,025	-,079	-,010	,114	-,152	,050
	Tid ift. udbytte	-,172	-,025	,579	-,009	,019	-,061	,081	-,147
	Faget spændende og interessant	-,221	-,079	-,009	,499	,005	-,111	,027	-,104
	Egnede lærebøger	-,032	-,010	,019	,005	,526	-,263	-,087	-,108
	Inspirerende litteratur	,080	,114	-,061	-,111	-,263	,467	,084	-,054
	Pensum for stort	-,005	-,152	,081	,027	-,087	,084	,886	,019
	Samlede udbytte	-,076	,050	-,147	-,104	-,108	-,054	,019	,486
Anti-image Correlation	Faget relevant	,763 <sup>a</sup>	,106	-,321	-,443	-,063	,166	-,008	-,155
	Faget svært	,106	,728 <sup>a</sup>	-,036	-,120	-,015	,178	-,172	,077
	Tid ift. udbytte	-,321	-,036	,843 <sup>a</sup>	-,017	,035	-,117	,114	-,278
	Faget spændende og interessant	-,443	-,120	-,017	,806 <sup>a</sup>	,010	-,231	,041	-,211
	Egnede lærebøger	-,063	-,015	,035	,010	,746 <sup>a</sup>	-,531	-,127	-,214
	Inspirerende litteratur	,166	,178	-,117	-,231	-,531	,735 <sup>a</sup>	,130	-,114
	Pensum for stort	-,008	-,172	,114	,041	-,127	,130	,751 <sup>a</sup>	,029
	Samlede udbytte	-,155	,077	-,278	-,211	-,214	-,114	,029	,875 <sup>a</sup>

a. Measures of Sampling Adequacy(MSA)

SPSS has now generated the tables for the KMO and Bartlett's test as well as the Anti-Image matrices. KMO attains a value of 0.791, which apparently seems to satisfy the criteria mentioned above. Equivalently the Bartlett's test attains a probability value of 0.000. Similarly high correlations are primarily found on the diagonal of the Anti-Image matrix (marked with an "a"). This confirms our thesis of an underlying structure of the variables.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,499	43,740	43,740	3,499	43,740	43,740	2,526	31,574	31,574
2	1,114	13,922	57,662	1,114	13,922	57,662	1,853	23,167	54,741
3	1,031	12,887	70,549	1,031	12,887	70,549	1,265	15,807	70,549
4	,755	9,434	79,983						
5	,547	6,835	86,818						
6	,401	5,013	91,831						
7	,383	4,792	96,624						
8	,270	3,376	100,000						

Extraction Method: Principal Component Analysis.

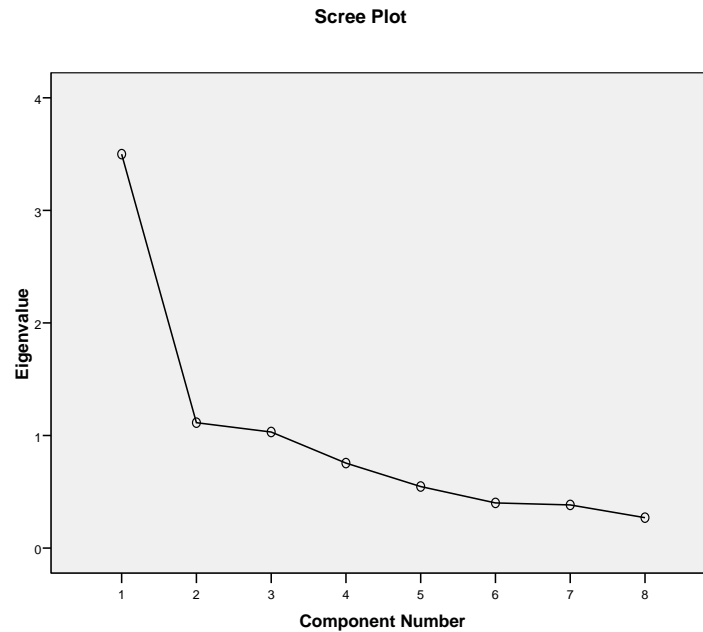
As can be seen from the output shown above, SPSS has produced the table Total Variance Explained, which takes account of both the rotated and unrotated solution.

(1) Initial Eigenvalues displays the calculated eigenvalues as well as the explained and accumulated variance for each of the 8 components.

(2a) Extraction Sums of Squared Loadings displays the components, which satisfy the criterion that has been chosen in section Extraction (Kaiser's criterion was chosen in section 8.3.2). In this case there are three components with an eigenvalue above 1. These three components together explain 70.549% of the total variation in the data. The individual contributions are 43.740%, 13.922%, 12.887% of the variation for component 1, 2 and 3 respectively. These are the results for the un-rotated solution.

In (2b) Rotation Sums of Squared Loadings similar information to that of (2a) can be found, except that these are the results for the Varimax rotated solution. As shown the sum of the three variances is the same both before and after the rotation. However, there has been a shift in the relationship between the three components, as they contribute more equally to the variation in the rotated solution.

The Scree plot below is an illustration of the variance of the principal components:



After the inclusion of the third component the factors no longer have eigenvalues above 1, and consequently the curve flattens. Normally the Scree plot will exhibit a break on the curve, which confirms how many components to include. The graphical depiction could therefore be better in the current example however it is decided to include three components that satisfy the Kaisers Criteria in the further analysis. It is therefore reasonable to treat the remaining components as "noise". This agrees with the previous output of the explained variance in the table Total Variance Explained.

**Component Matrix<sup>a</sup>**

	Component		
	1	2	3
Samlede udbytte	,816		
Faget spændende og interessant	,762	,286	-,111
Inspirerende litteratur	,730	-,257	,430
Faget relevant	,723	,330	-,328
Tid ift. udbytte	,722	,187	-,278
Egnede lærebøger	,672		,592
Faget svært	-,340	,710	
Pensum for stort	-,331	,547	,540

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

(3a) Component Matrix displays the principal component loadings for the un-rotated solution. This table shows the coefficients of the variables in the un-rotated solution. For example it can be observed that the correlation between the variable Overall benefit and component 1 is 0.829, i.e. very high. The un-rotated solution does not form an optimal picture of the correlations, however. Therefore, it is a good idea to rotate the solution in hope of clearer results.

(3b) Rotated Component Matrix displays the principal component loadings in the same way, but as the name reveals this is for the rotated solution.

**Rotated Component Matrix<sup>a</sup>**

	Component		
	1	2	3
Faget relevant	,854		
Faget spændende og interessant	,771	,282	
Tid ift. udbytte	,764	,153	-,161
Samlede udbytte	,669	,462	-,123
Egnede lærebøger	,226	,871	
Inspirerende litteratur	,261	,819	-,214
Pensum for stort	-,226		,799
Faget svært		-,303	,731

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

As previously mentioned, the purpose of rotating the solution is to make some variables correlate highly with one of the components – i.e. to enlarge the large coefficients (loadings) and to reduce the small coefficients (loadings). Whether a variable is highly or lowly correlated is subjective, but one rule of thumb is that correlations below 0.4 are considered low.

By means of output (3b) Rotated Component Matrix it is possible to try to join the most similar assessment criteria in three different groups:

- Component 1: For this group the variables Met expectations, Relationship time/Benefit, Course Interesting and Overall benefit are important. Therefore it will be appropriate to categorize component 1 as course benefit.
- Component 2: The variable Textbooks suitable and Literature inspiring are correlating highly with component 2 and therefore it is named quality of the literature.
- Component 3: The variables More difficult than other courses and Curriculum too extensive all have values above 0.4 compared to the third component. Thus could be named difficulty of the course.

This concludes the example. It is worth mentioning that if these results were to be used in later analyses, all factor scores should be used. As mentioned, these scores can be calculated by selecting the property called Scores. SPSS will then calculate a score for each respondent based on each of the components.

## 19. Cluster analysis

The following analysis and interpretation of Cluster analysis is based on the following literature:

- "Videregående data-analyse med SPSS og AMOS", Niels Blunch 1. udgave 2000, Systime. Chp. 1, p. 3 – 29.

### 19.1 Introduction

Cluster analysis is a multivariate procedure for detecting groupings in the data where there is no clarity. Cluster analysis is often applied as an explorative technique. The purpose of a cluster analysis is to divide the units of the analysis into smaller clusters so that the observations in the cluster are homogenous and the observations in the other clusters in one way or the other are different from these.

In contrast to the discriminant analysis, the groups in the cluster analysis are unknown from the outset. This means that the groups are not based on pre-defined characteristics. Instead the groups are created on basis of the characteristics of the data material. It is important to note that cluster analysis should be used with extreme caution since SPSS will always find a structure no matter if there is one present or not. The choice of method should therefore always be carefully considered and the output should be critically examined before the result of the analysis is employed.

Cluster analysis in SPSS is divided into hierarchical and K-means clustering (non-hierarchical analysis) Examples of both types of analyses are found below.

### 19.2 Hierarchical analysis of clusters

In the hierarchical method, clustering begins by finding the closest pair of objects (cases or variables) according to a distance measure and combining them to form a cluster. This algorithm starts with each case (or variable) in a separate cluster and combines clusters one step at a time until all data is placed in a cluster. The method is called hierarchical because it does not allow single objects to change cluster once they have been joined; this requires that the entire cluster be changed. As indicated, the method can be used for both cases and variables.

#### 19.2.1 Example

Data set: Hierarkiskdata.sav from the downloaded zip folder (see top of document).

Data on a number of economic variables is registered for 15 randomly chosen countries. By means of hierarchical cluster analysis it is now possible to find out which countries are similar with regard to the variables in question. In other words the task is to divide the countries into relevant clusters. The chosen countries are:

- Argentina, Austria, Bangladesh, Bolivia, Brazil, Chile, Denmark, The Dominican Republic, India, Indonesia, Italy, Japan, Norway, Paraguay and Switzerland.

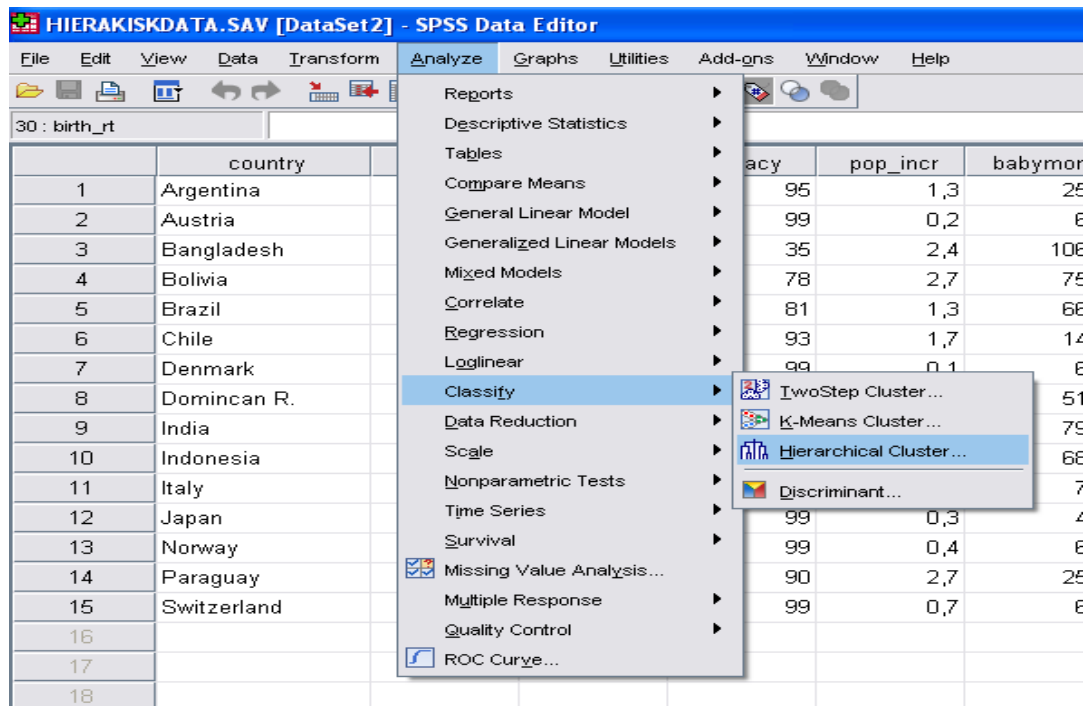
The following variables are included in the analysis:

- urban (number of people living in cities), lifeexpf (average life span for women), literacy (number of people that can read), pop\_incr (increase in population in % per year), babymort (mortality for newborns per 1000 newborns), calories (daily calorie intake), birth\_rt (birth rate per capita), death\_rt (death rate per 1000 inhabitants), log\_gdp (the logarithm to BNP per inhabitant), b\_to\_d (birth rate in proportion to mortality), fertility (average number of babies), log\_pop (the logarithm of a number of inhabitants).

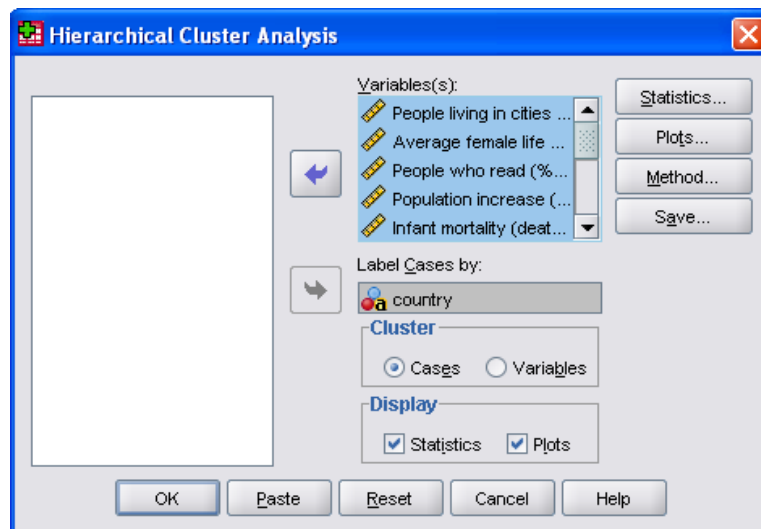
Both the hierarchical and the non-hierarchical method will be described in the following paragraphs. However it should be noted that for both analysis methods the cluster structure is made based on cases. As it will be mentioned later, it is possible to conduct the hierarchical analysis based on variables, however this is beyond the scope of the manual.

### 19.2.2 Implementation of the analysis

Choose the following commands in the menu bar:



Hereby the following dialogue box will appear:



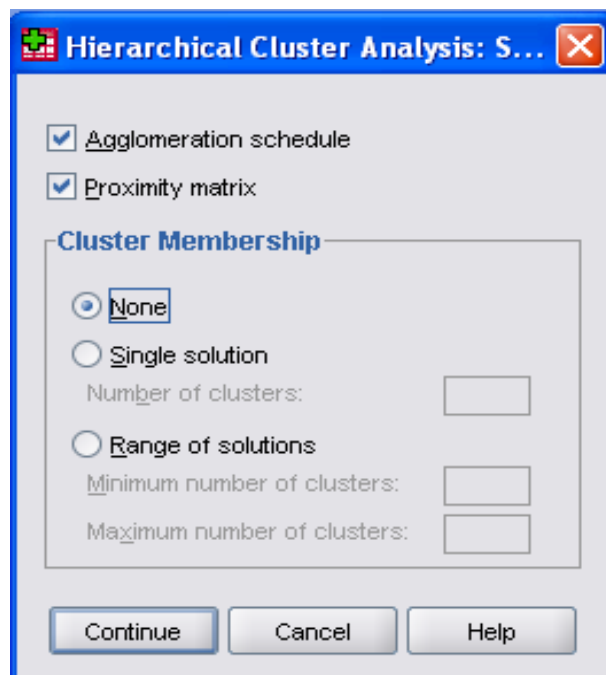
The relevant variables are chosen and added to the Variable(s) window (see below). Under Cluster it is chosen whether the analysis has to do with observations or variables, i.e. if the user wants to group variables or observations into clusters. Since this example has to do with observations, Cases are chosen.

Provided that the hierarchical analysis is based on cases, such as the above standing example, it is very important that the chosen variable is defined as a string (in variable view – type).

Under Display it is possible to include a display of plots and/or statistics. The Statistics and Methods buttons must be chosen before the analysis can be carried through. When these dialogue boxes are activated it is possible to click the Save button. Subsequently, the above-mentioned buttons will be described:

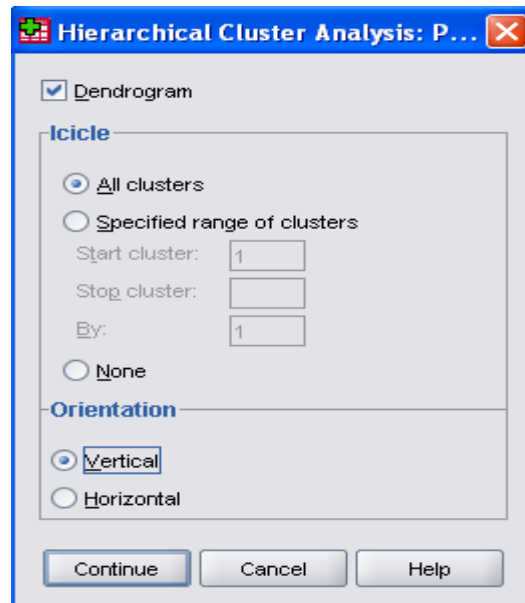
#### 19.2.2.1 Statistics

- Choosing *Agglomeration schedule* helps identify which clusters are combined in each iterative step.
- *Proximity matrix* shows how the distances or the similarities between the units (cases or variables) appear in the output.
- *Cluster membership* shows how every single observations- or variables cluster membership appear in the output.



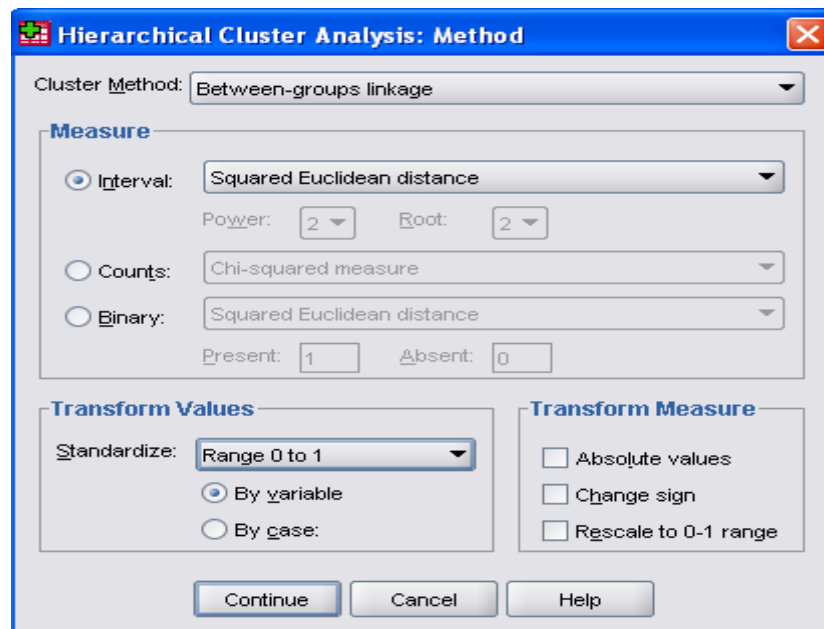
### 19.2.2.2 Plots

A dendrogram or an icicle plot can be chosen here. In this example as well as in the following, A dendrogram will be chosen, since the interpretation of the cluster analysis is often based on this.



### 19.2.2.3 Method

In this dialogue box the methods on which the cluster analysis is based are indicated. In Cluster method a number of methods can be chosen. The most important are de-scribed below:



- *Nearest neighbour (single linkage)*: The clusters are created by minimizing the distance between pairs of clusters.
- *Furthest neighbour (complete linkage)*: The distance between two clusters is calculated as the distance between the two elements (one from each cluster) that are furthest apart.
- *Between-groups linkage (average linkage)*: The distance between two clusters is calculated as the average of all distance combinations (which is given by all distances between elements from the two clusters).
- *Ward's method (Ward's algorithm)*: For each cluster a stepwise minimize action of the sum of the squared distances to the center (the centroid) within each cluster is carried out.

The result of the cluster analysis depends heavily on the chosen method. As a consequence, one should try different methods before completing the analysis. In this example *Between-groups linkage* has been chosen.

*Measure* provides the user with the choice of different distance measures depending on the type of data in question. The type of the data in question should be stated (interval data, frequency table or binary variables). In this example *interval* has been chosen.

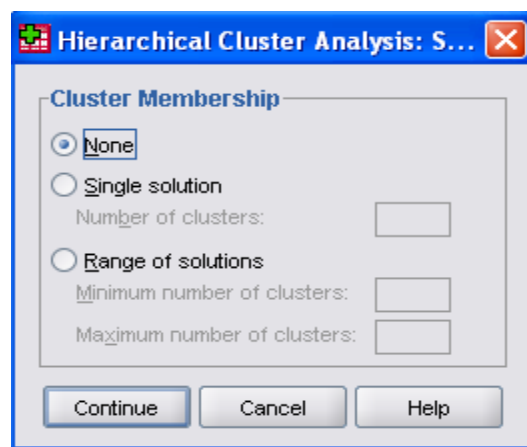
After this, the distance measure must be specified in the *interval* drop-down box. A thorough discussion of these is beyond the scope of this manual, although several are relevant. It should be noted, however, that *Pearson's correlation* is often used in connection with variables, while *Squared Euclidean distance* is often used in connection with cases. The latter is default in SPSS, and it is chosen in the example as well, since observations are considered here. When using *Pearson's correlation* one should be aware that larger distance measures mean less distance (measures range from -1 to 1).

*Transform values* is used for standardizing data before the calculation of the distance matrix, which is often relevant for the cluster analysis. This is due to the fact that variables with high values contribute more than variables with low values to the calculation of distances. Consequently, this matter should be carefully considered. In this case, the values are transformed to a scale ranging from 0 to 1.

*Transform Measures* is used for transforming data after the calculation of the distance matrix, which is not relevant in this case.

#### 19.2.2.4 Save

This dialogue box gives the user the opportunity to save cluster relationships for observations or variables as new variables in the data set. This option is ignored in the example.





### 19.2.3 Output

After the specification of the options the analysis is carried out. This results in a number of outputs, of which three parts are commented on and interpreted below.

(1) *Proximity matrix* shows the distances that have been calculated in the analysis, in this case the calculated distances between the countries. For example, the distance between Norway and Denmark is 0.154 when using the Squared Euclidean distance. The cluster analysis itself is carried out on the basis of this distance matrix. As can be seen, the matrix is symmetric:

Case	Squared Euclidean Distance														
	1:Argentina	2:Austria	3:Bangladesh	4:Bolivia	5:Brazil	6:Chile	7:Denmark	8:Dominican R.	9:India	10:Indonesia	11:Italy	12:Japan	13:Norway	14:Paraguay	15:Switzerland
1:Argentina	,000	1,008	4,883	2,233	,442	,423	1,024	,936	3,164	1,353	,877	,836	,694	2,286	,840
2:Austria	1,008	,000	7,671	4,809	1,885	1,895	,184	2,667	5,653	2,746	,191	,742	,135	4,659	,128
3:Bangladesh	4,883	7,671	,000	1,398	3,032	5,109	8,498	3,024	,484	1,727	7,886	7,867	7,655	4,089	7,830
4:Bolivia	2,233	4,809	1,398	,000	1,590	1,883	5,276	,788	1,367	1,261	5,157	4,817	4,318	1,370	4,600
5:Brazil	,442	1,885	3,032	1,590	,000	,921	2,106	,787	1,665	,547	1,611	1,452	1,717	2,572	1,837
6:Chile	,423	1,895	5,109	1,883	,921	,000	2,081	,446	3,543	1,721	1,771	1,217	1,341	1,309	1,441
7:Denmark	1,024	,184	8,498	5,276	2,106	2,081	,000	3,125	6,437	3,488	,355	,994	,154	5,243	,342
8:Dominican R.	,936	2,667	3,024	,788	,787	,446	3,125	,000	2,160	,911	2,772	2,286	2,231	,920	2,311
9:India	3,164	5,653	,484	1,367	1,665	3,543	6,437	2,160	,000	,739	5,479	5,270	5,684	3,344	5,763
10:Indonesia	1,353	2,746	1,727	1,261	,547	1,721	3,488	,911	,739	,000	2,657	2,639	2,872	2,320	2,759
11:Italy	,877	,191	7,886	5,157	1,611	1,771	,355	2,772	5,479	2,657	,000	,312	,311	4,883	,236
12:Japan	,836	,742	7,867	4,817	1,452	1,217	,994	2,286	5,270	2,639	,312	,000	,634	4,245	,561
13:Norway	,694	,135	7,655	4,318	1,717	1,341	,154	2,231	5,684	2,872	,311	,634	,000	4,013	,112
14:Paraguay	2,286	4,659	4,089	1,370	2,572	1,309	5,243	,920	3,344	2,320	4,883	4,245	4,013	,000	3,952
15:Switzerland	,840	,128	7,830	4,600	1,837	1,441	,342	2,311	5,763	2,759	,236	,561	,112	3,952	,000

This is a dissimilarity matrix

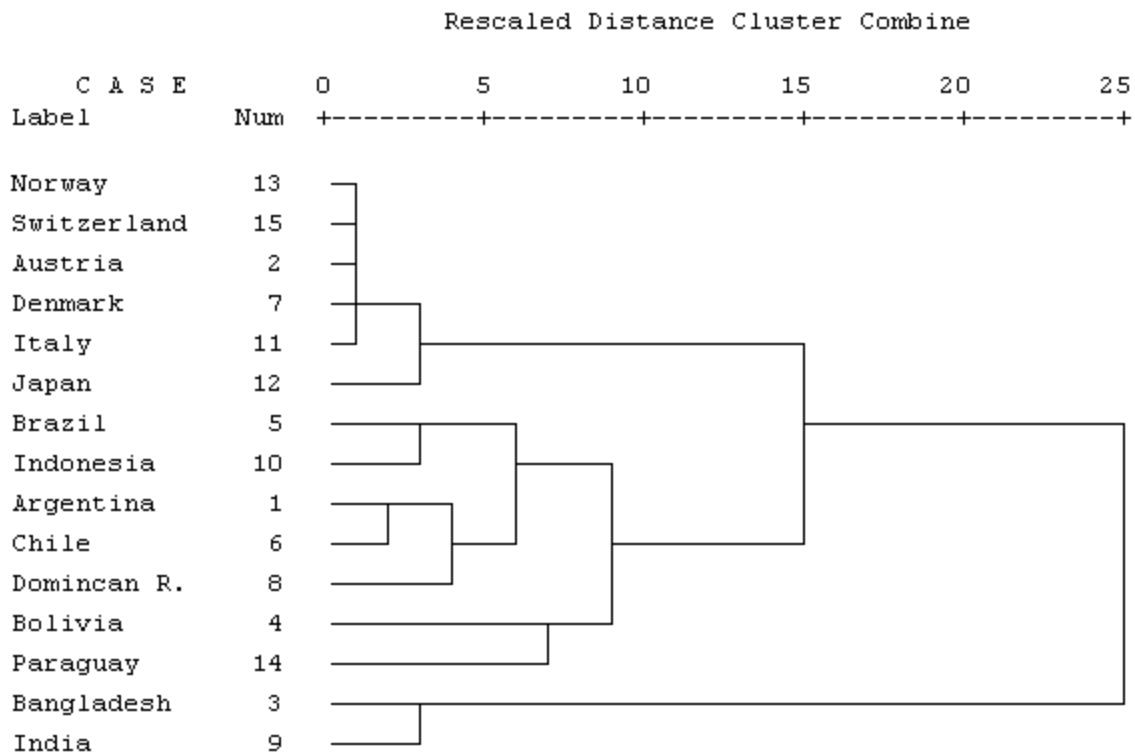
The next output is (2) *Agglomeration Schedule*, which shows when the individual observations or clusters are combined. In the example observation 13 (Norway) is first combined with observation 15 (Switzerland), where after observation 2 (Austria) is tied to the very same cluster. This procedure continues until all observations are combined. The *Coefficients* column shows the distance between the observations/clusters that are combined.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	13	15	,112	0	0	2
2	2	13	,132	0	1	3
3	2	7	,227	2	0	4
4	2	11	,273	3	0	8
5	1	6	,423	0	0	9
6	3	9	,484	0	0	14
7	5	10	,547	0	0	10
8	2	12	,649	4	0	13
9	1	8	,691	5	0	10
10	1	5	1,023	9	7	12
11	4	14	1,370	0	0	12
12	1	4	1,716	10	11	13
13	1	2	2,718	12	8	14
14	1	3	4,651	13	6	0

The (3) *Dendrogram* shown below is a simple way of presenting the cluster analysis:

\* \* \* \* \* H I E R A R C H I C A L C L U S T E R A N A L Y S I S \* \* \* \* \*

Dendrogram using Average Linkage (Between Groups)



The dendrogram indicates that there are three clusters in the data set. These are:

- Norway, Denmark, Austria, Switzerland, Italy and Japan (i.e. the European countries and Japan).
- Brazil, Argentina, Chile, The Dominican Republic, Bolivia, Paraguay and Indonesia (i.e. the South American countries and Indonesia).
- India and Bangladesh (i.e. the Asian countries).

The cluster structure is found by making a sectional elevation through the dendrogram, where the distance is relatively long. In the example above the sectional elevation is made between 10 and 15. This structure is very much in accordance with the prior expectations one might have regarding the economic variables in question.

### 19.3 K-means cluster analysis (Non-hierarchical cluster analysis)

Contrary to the hierarchical cluster analysis, the non-hierarchical cluster analysis does allow observations to change from one cluster to another while the analysis is proceeding. The method is based on an iterative procedure where every single observation is grouped into a number of clusters until the relation, the variance between the clusters and the variance within the clusters, is maximized. The procedure itself will not be explained further here. It is noted, however, that the user has the opportunity to choose clusters as well as so-called cluster-centres, which will often be an advantage.

The non-hierarchical cluster analysis is often used in large data sets since the method is capable of treating more observations than the hierarchical method. Apart from this the method is only used in analysis regarding observations and not analysis regarding variables.

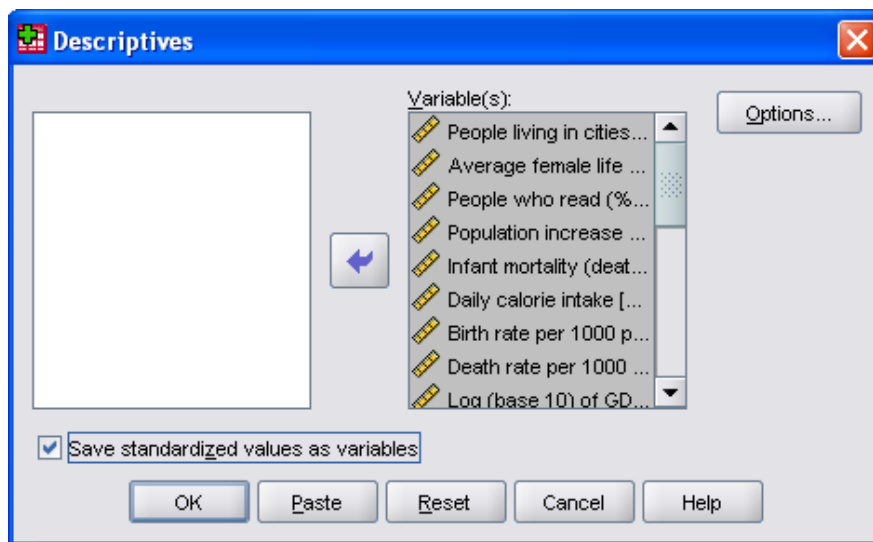
#### 19.3.1 Example

Data: K-meansdata.sav can be found in the downloaded zip folder (see top of document). The analysis is based on the same example as the hierarchical cluster analysis in section 3.2.1.

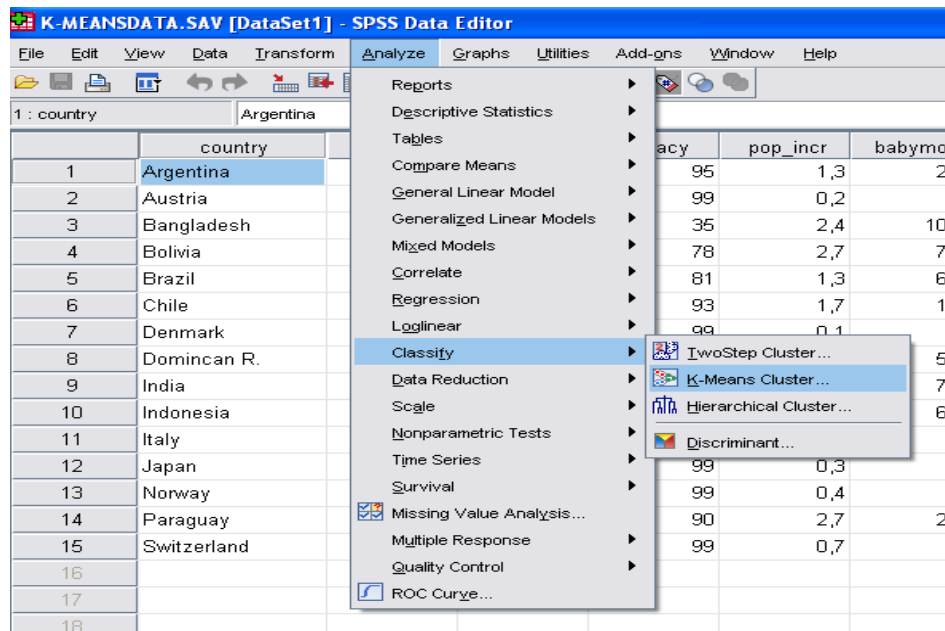
#### 19.3.2 Implementation of the analysis

Before the cluster analysis is carried out it is necessary to standardize the variables from the previous example. As mentioned earlier this is done to prevent large values from swinging too much weight on the analysis. When using the k-means method it is not possible to make any transformations during the analysis, as described in the hierarchical method. Therefore, this must be done prior to the analysis.

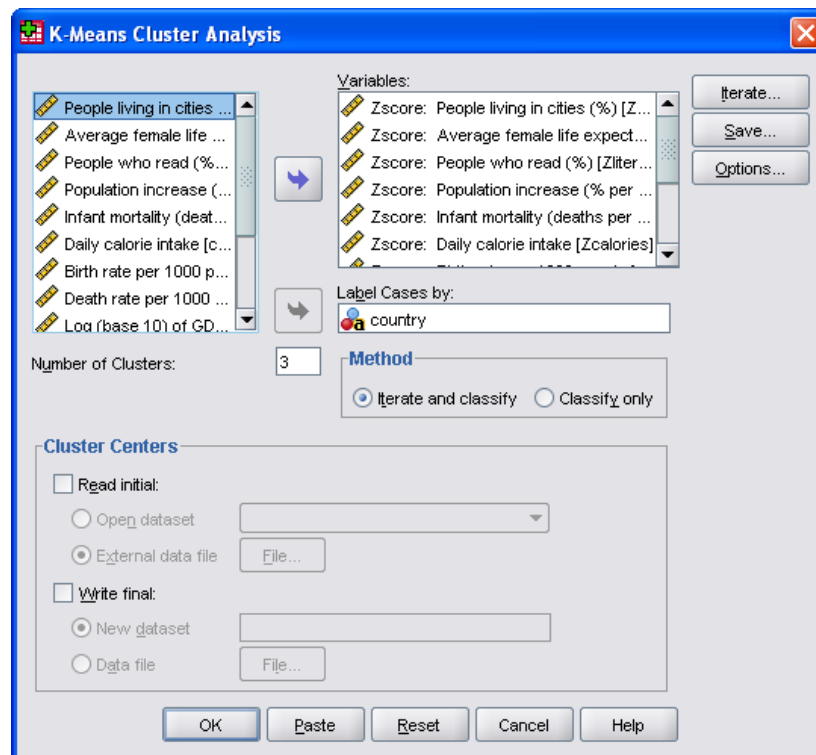
Choose Analyze – Descriptive statistics – Descriptives from the menu bar. This results in a dialogue box, which should be completed as follows:



After this it is time to carry out the analysis on the basis of the new standardized variables. Choose the following from the menu bar:



This results in the following dialogue box:



Here all the standardized variables are moved to the *Variables* window. Since the names of the countries are meant to be included in the output, the variable *Country* is moved to the *Label Cases by* window.

As opposed to the hierarchical cluster analysis it is not a requirement for the non-hierarchical cluster analysis that the variable used to group the clusters by (in this case *country*) is defined as a string.

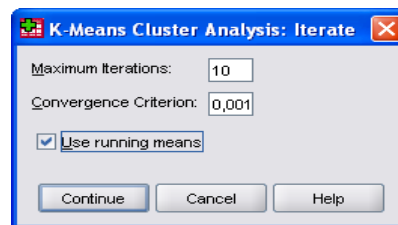
*Method* is set to *Iterate* and *classify* as default, since the classification is to be performed on the basis of iterations.

In that connection it must be decided how many clusters are wanted as a result of the iteration process. This is chosen in the *Number of Clusters* window. The number of clusters depends on the data set as well as on preliminary considerations. From the result of the hierarchical analysis there is reason to believe that the number of clusters is three, so this is the number that has been entered. If one has detailed knowledge of the number and characteristics of expected clusters, it is possible to define cluster centers in *Centers*.

In addition to the options described above there is a number of ways, in which the user can influence the iteration procedure and the output. These are dealt with below.

#### 19.3.2.1 Iterate

Clicking the *Iterate* button results in the following dialogue box with three changeable settings:



*Maximum iteration* indicates the number of iterations in the algorithm, which ranges between 1 and 999. Default is 10 and this will be used in this example as well.

*Convergence Criterion* defines the limit, at which the iteration is ended. The value must be between 0 and 1. In this example a value of 0.001 is used, which means that the iteration ends when none of the cluster centers move by more than 0.1 percent of the shortest distance between two clusters.

Selecting *Use running means* is the mean values will be recalculated each time an observation has been assigned to a cluster. By default this only happens when all observations have been assigned to the clusters.

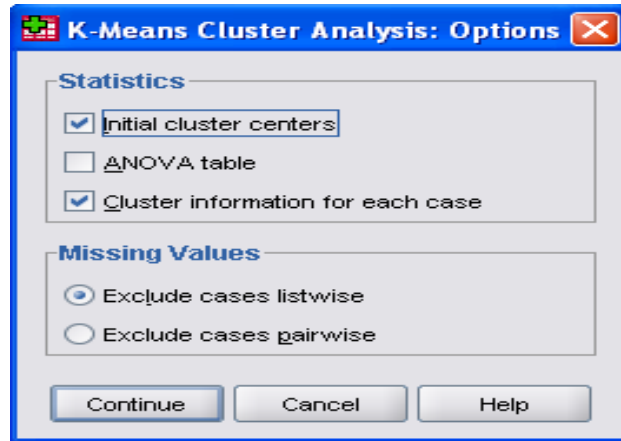
#### 19.3.2.2 Save

Under *Save* it is possible to save the cluster number in a new variable in the data set. In addition it is possible to create a new variable that indicates the distance between the specific observation and the mean value of the cluster. See the dialogue box below:



### 19.3.2.3 Options

Clicking *Options* gives the user the following possibilities:



*Initial cluster centers* are the initial estimates of the cluster mean values (before any iterations).

*ANOVA table* carries out an F-test for the cluster variables. However, since the observations are assigned to clusters on the basis of the distance between the clusters, it is not advisable to use this test for the null hypothesis that there is no difference between the clusters.

*Cluster information for each case* ensures that a cluster number is assigned to each observation – i.e. information about the cluster relationship. Furthermore, it produces information about the distance between the observation and the mean value of the cluster.

Choosing *exclude cases listwise* in the Missing values section means, that a respondent will be excluded from the calculations of the clusters, if there is a missing value for any of the variables.

*Exclude cases pairwise* will classify a respondent in the closest cluster on the basis of the remaining variables. Therefore, a respondent will always be classified, unless there are missing values for all variables.

### 19.3.3 Output

The analysis results in a number of outputs, of which the most relevant will be interpreted and commented on the following page.

**Cluster Membership**

Ca...	country	Cluster	Distance
1	Argentina	2	1,787
2	Austria	2	1,386
3	Bangladesh	3	2,154
4	Bolivia	3	2,370
5	Brazil	2	3,075
6	Chile	1	1,677
7	Denmark	2	1,786
8	Dominican R.	1	1,244
9	India	3	1,438
10	Indonesia	3	2,106
11	Italy	2	1,150
12	Japan	2	1,921
13	Norway	2	1,126
14	Paraguay	1	2,120
15	Switzerland	2	1,257

The output (1) *Cluster Membership* shows, as the name indicates the countries' relation to the clusters. From the output it is obvious that, except for a few changes, the cluster structure is identical to that of the hierarchical method. Distance is a measure of how far the individual country is located compared to the center of its cluster. The output shows that Brazil is the country, which is located furthest away from its cluster center with a distance of 3,075.

The next output to be commented on is (2) *Final Cluster Centers*. From this output the mean value of the standardized variables for each of the three clusters appear. It is therefore possible to compare the clusters on the basis of each variable.

**Final Cluster Centers**

	Cluster		
	1	2	3
Zscore: People living in cities (%)	,18741	,59085	-1,32225
Zscore: Average female life expectancy	,18159	,60833	-1,35285
Zscore: People who read (%)	,18983	,57653	-1,29543
Zscore: Population increase (% per year)	,84337	-,78296	,93340
Zscore: Infant mortality (deaths per 1000 live births)	-,17953	-,59600	1,32665
Zscore: Daily calorie intake	-,52205	,75216	-1,11277
Zscore: Birth rate per 1000 people	,67384	-,78614	1,06691
Zscore: Death rate per 1000 people	-1,58610	,37670	,43618
Zscore: Log (base 10) of GDP_CAP	-,42130	,77500	-1,23403
Zscore: Birth to death ratio	1,45708	-,68799	,28317
Zscore: Fertility: average number of kids	,39893	-,70822	1,11725
Zscore: Log (base 10) of Population	-,71600	-,14993	,83687

The last output, which will be commented on is (3) *Distances between Final Cluster Centers*.

**Distances between Final Cluster Centers**

Cluster	1	2	3
1		4,301	4,321
2	4,301		5,818
3	4,321	5,818	

This output indicates the individual distances between the clusters. In this example cluster 1 and 2 are most alike. A closer look at output (2) is required in order to explore the differences and similarities in depth.



## 20. Non-Parametric tests

### 20.1 Cochran's Test

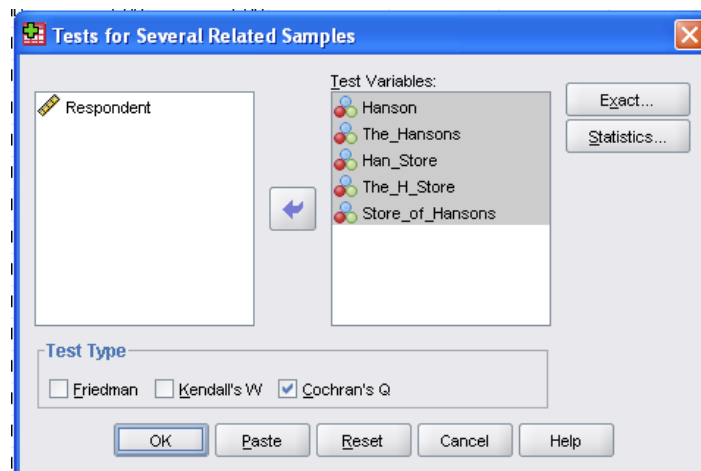
The Cochran test calculates the Cochran's test statistic. All variables to be tested must have a binary outcome (0/1). The hypothesis for the test looks like the following.

$H_0$  : No difference in use of name for product

$H_1$  : Difference in use of name for product

To illustrate the use of this test we use a dataset, where a number of different respondents have made a valuation of 6 different names for a clothing store, each respondent had to evaluate whether he liked (1) or disliked (0) the name. The dataset CochranTest.sav for this test can be found in the downloaded zip folder (see top of document).

Selecting Analyze -> Non Parametric Tests -> K Related samples can run the test. Choose your settings as shown below and press OK.



Remember to tick the Cochran's test.

**Frequencies**

	Value	
	0	1
Hanson	11	9
The_Hansons	11	9
Han_Store	12	8
The_H_Store	8	12
Store_of_Hansons	9	11

**Test Statistics**

N	20.000
Cochran's Q	2.118 <sup>a</sup>
df	4.000
Asymp. Sig.	.714

a. 1 is treated as a success.

The first table is a table of frequencies showing how the answers from the respondents are distributed. The second table above shows the result of the Cochran's test. In this case we get a Q value of 2,118, and a p-value of 0,714. On the basis of these values we cannot reject the  $H_0$  hypothesis, and conclude there is no difference in use of name.

## 20.2 Friedman's Test

Friedman's test is used when e.g. a group of respondents have evaluated an effect of different products (interdependence between observations). The hypothesis looks like the following:

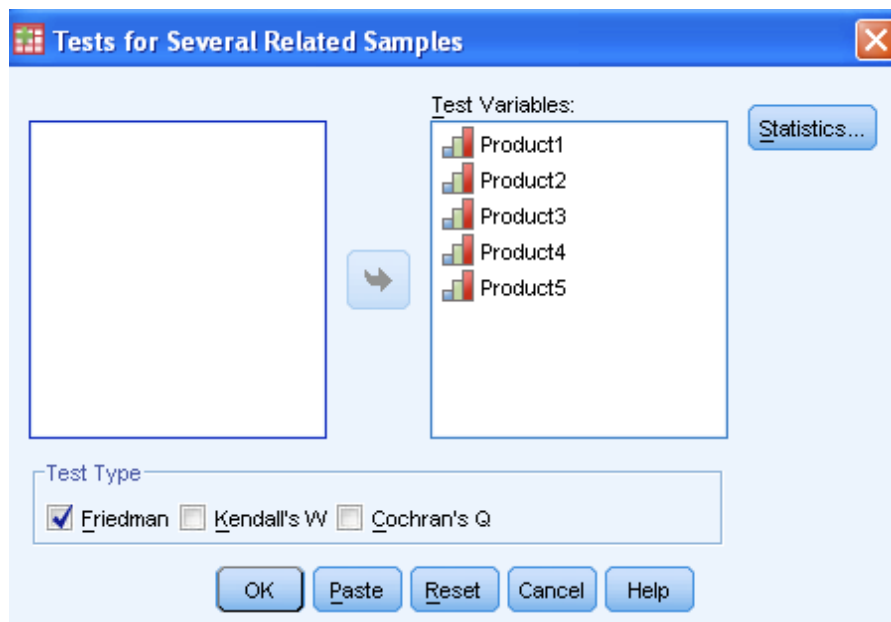
$H_0$  : *Ranking independent of product*

$H_1$  : *Ranking dependent of product*

In contrast to Cochran's test, it is not required for the variables to have a binary outcome. Instead the Friedman's test is based on a re-coding of the variables, so they become ranked. In this way the lowest value for each observation row (for each respondent) will be given the value 1 while the second lowest value will be given the value 2 etc. After recoding, the sum of the evaluation for each product (column) is calculated. In the end a test will be made for differences between the product's sums.

The example used in this test is in many ways similar to the one used in Cochran, but here the respondents have valued the different products on a scale from 1-5.

The dataset FriedsmannTest can be found in the downloaded zip folder (see top of document) To run the test open Analyze -> Non Parametric Tests -> Legacy Dialogs -> K Related Samples and type in the settings as shown below.



As seen above products1, products2, products3, products4 and products5 have to be the test variables. Remember to tick the Friedman test.

## Friedman Test

Ranks

	Mean Rank
Product1	4.60
Product2	3.30
Product3	2.65
Product4	1.60
Product5	2.85

Test Statistics<sup>a</sup>

N	10.000
Chi-Square	20.787
df	4.000
Asymp. Sig.	.000

a. Friedman Test

The output above shows a statistic of 20,787, which gives a p-value of 0%. On the basis of this we reject the H0 hypothesis and conclude that at least two of the products differ in their treatment effect.

Since the data is Ordinal we cannot do a bonferroni. Instead we have to create a syntax. The syntax will look as follows.

```

*friedman.sps - PASW Statistics Syntax Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Run Tools Window Help
compute
compute
compute
/* No changes below ...
exe.
delete variables
1 compute alpha=0.05.
2 compute k=5.
3 compute b=10.
4 /* No changes below */
5 compute fried=idf.srange(1-alpha,k,999999)*sqr(b*k*(k+1)/12).
6 exe.
7 delete variables k TO b.
  
```

- At compute alpha we have to choose our significant level. In this case it is 0,05.
- At compute k we have type in how many groups there is in the test. In this case 5.
- At compute b the value has to be 10 since we in this case have 10 respondents.

When the play button has been pressed a new variable in the dataset will appear. In this case we get 19,29. We use this value to find significant difference between groups. Before this can be done, the sums for each group have to be calculated. It can be done under Analyze -> Descriptive Statistics -> Descriptive. The following output appears.

Descriptive Statistics		
	N	Sum
Product1	10	43.00
Product2	10	29.50
Product3	10	21.50
Product4	10	13.00
Product5	10	28.00
Valid N (listwise)	10	

It can now be seen that there is a significant difference between Product1 and Product4, since there is a difference greater than 19,29. Product1 and Product3 is as well significant different.

### 20.3 Kruskal Wallis Test

The Kruskal Wallis' test is used when analyzing for differences e.g. between different machines. Under normal conditions a one factor ANOVA test would be used. But if the assumptions of the test variable are not met see, we often use Kruskal Wallis.

In the following example, an expert in taste has made a valuation of 3 different machines, with different configurations, and given each machine a value from 0-100 on the basis of there performance. The dataset "Kruskall Wallis.sav" can be found in the downloaded zip folder (see top of document)

We want to test whether there is a difference in the respondent's opinion on the 3 different machines. The hypothesis looks like the following:

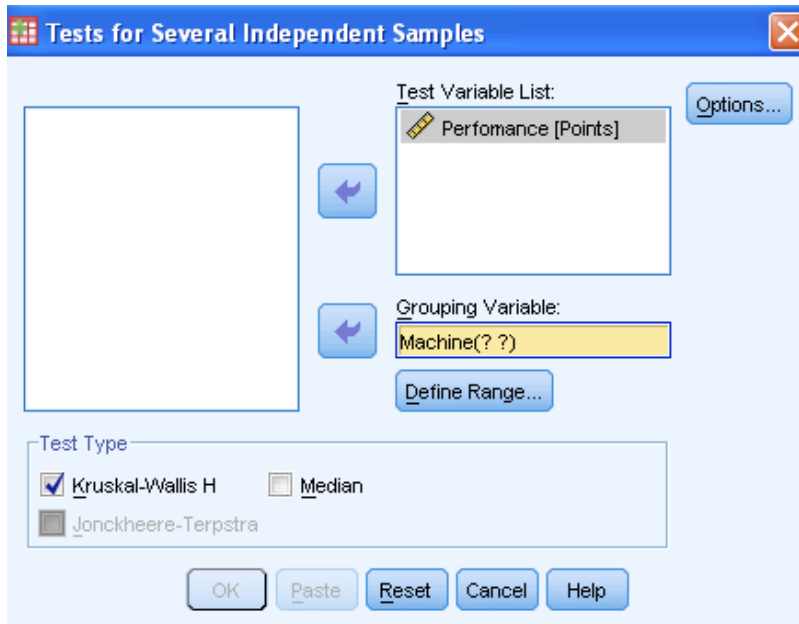
$$H_0 : \text{No difference among the machines}$$

$$H_1 : \text{At least two are not equal}$$

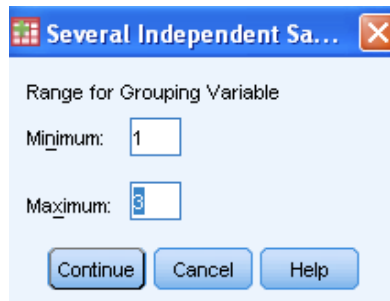
One of the conditions for running the test is that the dependent variable is of an ordinal scale. The valuation of the different products on the 0-100 scale must therefore be ranked, so the lowest score gets the value 1, the second lowest the value 2 and so forth.

The test is done in the following way. Go to Analyze -> Non Parametric Tests -> Legacy Dialogs -> K Independent samples.

Arrange your window similar to the one shown below.



In define groups you put in the minimum and maximum value as shown below.



### Kruskal-Wallis Test

Ranks			
	Machine	N	Mean Rank
Performance	Machine 1	7	16.36
	Machine 2	7	7.43
	Machine 3	7	9.21
	Total	21	

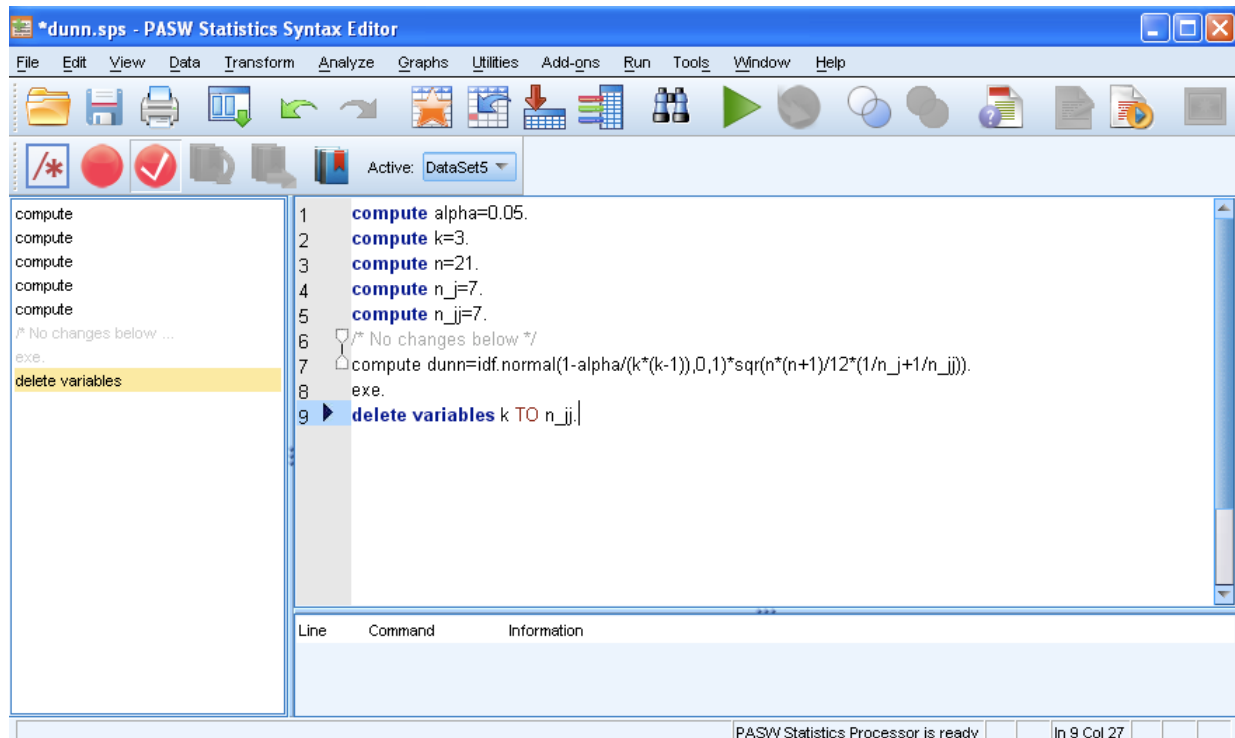
Test Statistics <sup>a, b</sup>	
	Performance
Chi-Square	8.165
df	2
Asymp. Sig.	.017

a. Kruskal Wallis Test

b. Grouping Variable: Machine

The test results are shown above and the p-value is 0,005, which means rejection of  $H_0$  and differences between the PH-values exist, although we are not able to tell which PH-value that differ from each other in taste.

We have to create a syntax to be able to tell that. It will look like the one below. Before it will work the text has to be marked and the play button has to be pressed. It is very important that the dot at every line does not get deleted.



- At compute alpha we have to choose our significant level. In this case it is 0,05.
- At compute k we have type in how many groups there is in the test, in this case 3.
- At compute n the value has to be 21 since we in this case have  $3 \times 7 = 21$  respondents.
- At compute  $n_j$  and compute  $n_{ij}$  the value has to be 7, since there are 7 values for each machine. If it there where different numbers of respondents in the group, then a dunn had to be performed for two groups at time.

When the play button has been pressed a new value in the dataset will appear. In this case we get 7,94.. We use this value to find significant difference between groups. In Kruskal Wallis the difference between the mean ranks has to be calculated, and if the groups have a difference more than 7,94, then they are significant different. In this case Machine 1 is significant better than machine 2.

THIS GUIDE HAS BEEN PRODUCED BY

#### ANALYTICS GROUP



Analytics Group, a division comprised of student instructors under AU IT, primarily offers support to researchers and employees.

Our field of competence is varied and covers questionnaire surveys, analyses and processing of collected data etc. AG also offers teaching assistance in a number of analytical resources such as SAS, SPSS and Excel by hosting courses organised by our student assistants. These courses are often an integrated part of the students' learning process regarding their specific academic area which ensures the coherence between these courses and the students' actual educational requirements.

In this respect, AG represents the main support division in matters of analytical software.

#### ADVANCED MULTIMEDIA GROUP



Advanced Multimedia Group is a division under AU IT supported by student instructors. Our primary objective is to convey knowledge to relevant user groups through manuals, courses and workshops.

Our course activities are mainly focused on MS Office, Adobe CS and CMS. Furthermore we engage in e-learning activities and auditive and visual communication of lectures and classes. AMG handles video assignments based on the recording, editing and distribution of lectures and we carry out a varied range of ad hoc assignments requested by employees.

In addition, AMG offers solutions regarding web development and we support students' and employees' daily use of typo3.

PLEASE ADDRESS QUESTIONS OR COMMENTS REGARDING THE CONTENTS OF THIS GUIDE TO

[ANALYTICS@ASB.DK](mailto:ANALYTICS@ASB.DK)