

**A guide to quantitative methods** Almquist, Ashir & Brännström

This guide has been compiled by Ylva B Almquist, Sahar Ashir and Lars Brännström, all active at Stockholm University, Sweden. Go to [www.chess.su.se/methods](http://www.chess.su.se/methods) for more information and revised versions of the guide.

# CONTENT

INTRODUCTION.....	6
1. THE SPSS ENVIRONMENT.....	8
1.1 General functions.....	9
1.2 Variable view.....	11
1.3 Creating a new data set.....	13
1.4 Data view.....	14
1.5 Syntax.....	15
1.6 Output.....	18
2. BASIC STATISTICAL CONCEPTS.....	19
2.1 Study population and sampling.....	20
2.2 Measurement scales.....	24
2.3 Distributions.....	27
3. DESCRIPTIVE STATISTICS.....	32
3.1 Tables.....	33
3.2 Graphs.....	35
3.3 Measures of central tendency.....	38
3.4 Measures of variation.....	42
4. PRODUCING DESCRIPTIVE STATISTICS IN SPSS.....	43
4.1 Descriptives.....	44
4.2 Frequency table.....	49
4.3 Cross table.....	55
4.4 Bar chart.....	60
4.5 Pie chart.....	67
4.6 Histogram.....	72
4.7 Scatterplot.....	77
4.8 Edit graphs.....	82
5. OTHER USEFUL FEATURES IN SPSS.....	83
5.1 Recode.....	84
5.2 Compute.....	88
5.3 Select cases.....	92
5.4 If.....	96
5.5 Weight cases.....	99
6. STATISTICAL SIGNIFICANCE.....	102
6.1 Hypothesis testing.....	103

6.2 P-values.....	106
6.3 Confidence intervals.....	109
6.4 Discussion.....	111
7. COMPARE MEANS.....	113
7.1 T-test: independent samples.....	114
7.2 T-test: paired samples.....	121
7.3 One-way ANOVA.....	127
8. CHI SQUARE.....	134
8.1 Chi-square.....	135
9. CORRELATION ANALYSIS.....	140
9.1 Correlation analysis.....	141
10. FACTOR ANALYSIS.....	147
10.1 Factor analysis.....	148
10.2 Cronbach's alpha.....	163
11. X, Y AND Z.....	169
11.1 X and Y.....	171
11.2 Z: confounding, mediating and moderating variables.....	172
12. PREPERATIONS FOR REGRESSION ANALYSIS.....	175
12.1 What type of regression should be used?.....	176
12.2 Dummies.....	177
12.3 Standardization: z-scores.....	180
12.4 Analytical strategy.....	184
12.5 Missing data.....	186
12.6 From study sample to analytical sample.....	189
13. LINEAR REGRESSION.....	193
13.1 Simple linear regression.....	197
13.2 Multiple linear regression.....	213
13.3 Model diagnostics.....	221
14. LOGISTIC REGRESSION.....	234
14.1 Simple logistic regression.....	240
14.2 Multiple logistic regression.....	252
14.3 Model diagnostics.....	258
15. ORDINAL REGRESSION.....	274
15.1 Simple ordinal regression.....	279
15.2 Multiple ordinal regression.....	292
15.3 Model diagnostics.....	299

16. MULTINOMIAL REGRESSION .....	304
16.1 Simple multinomial regression .....	309
16.2 Multiple multinomial regression .....	323
16.3 Model diagnostics .....	331
17. INTERACTION ANALYSIS.....	334
.....	342

# INTRODUCTION

The purpose of this guide is to provide both basic understanding of statistical concepts (know-why) as well as practical tools to analyse quantitative data in SPSS (know-how). We wanted to keep the guide completely free of formulas (i.e. brain-freezing mathematical equations). In doing so, we have tried to explain everything at the most elementary level and only include aspects that are important in actual research. As such, this guide is pragmatic and research-oriented. Hopefully, you will find it useful.

This guide consists of two parts. The first part (Chapters 1-5) concerns various aspects concerning data management and descriptive statistics. Next, we discuss issues related to statistical significance (Chapter 6). The following part deals with some basic types of statistical analysis, such as t-tests, ANOVA, chi-square, correlation analysis, and factor analysis (Chapters 7-10). Then we discuss theoretical and practical dimensions of regression analysis (Chapters 11-12) before continuing into how to actually conduct regression analysis, including interaction analysis (Chapters 13-17).

There are two data materials used in the guide. The first is “SPSS\_Data1” which is mostly based on a hypothetical data set available through SPSS. The second is “SPSS\_Data2”, which is primarily based on several waves of data collection related to the Word Values Survey.

## General advice

Keep all your files for the course/project in the same main folder and use sub folders to organise the files further.

Save your files under appropriate names.

Example: "Ericsson\_Data\_Lesson1\_130603"

Keep a copy of the original file, just in case.

Example: "Ericsson\_Data\_Lesson1\_Original"

Do not forget to continuously save your file while you work with it.

Always double-check that you have spelled values and labels correctly.

If you delete or change something by mistake, it is often possible to undo the last change by holding down the Ctrl-key while pressing "z" on your key board.

Use syntax!

### *Why syntax?*

The syntax is basically a text file where you can add comments and SPSS commands. The reasons for why everyone should use the syntax function are:

- It is a way of documenting and archiving everything you have done with the data material.
- It is easy to repeat parts or all of the analysis.
- Other people involved in the data material can easily understand what you have done and how you have done it.
- It saves an enormous amount of time.

# 1. THE SPSS ENVIRONMENT

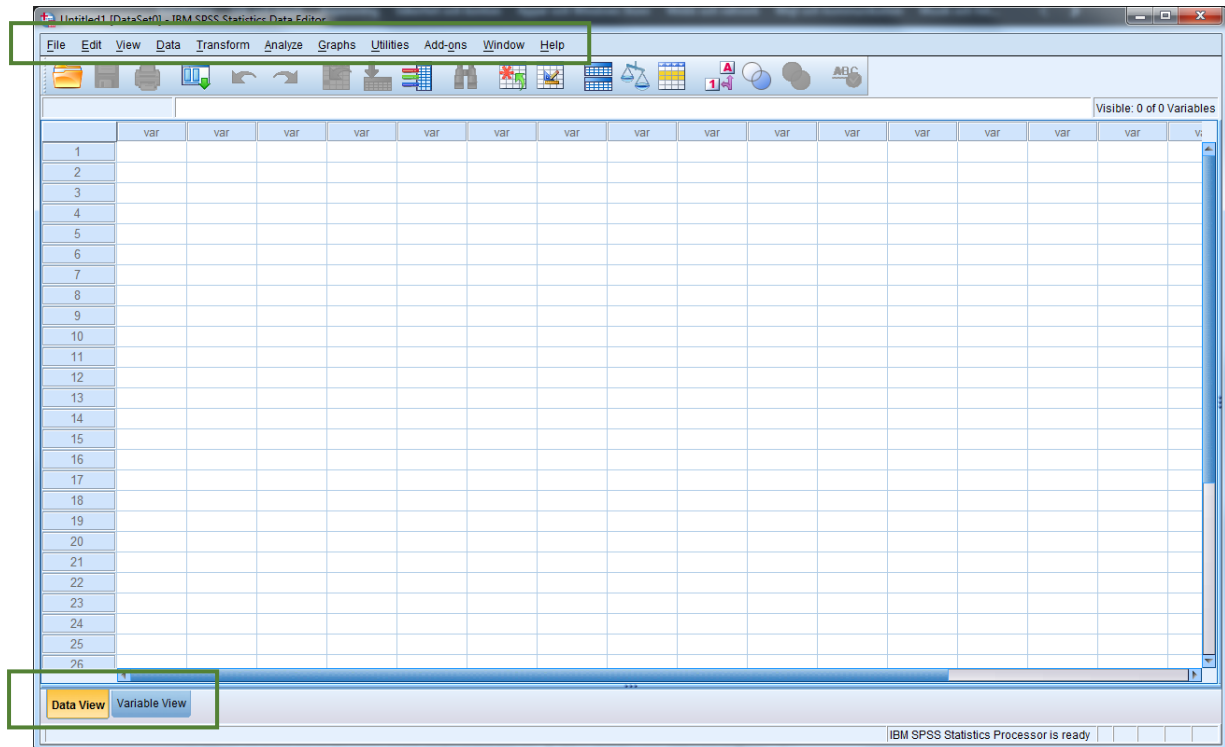
The SPSS environment may come across as rather confusing at first, but it is actually quite logical once you get a hang of it. In this part of the guide, we will discuss the following aspects of the SPSS environment:

Outline
1.1 General functions
1.2 Variable view
1.3 Creating a new data set
1.4 Data view
1.5 Syntax
1.6 Output



## 1.1 General functions

The **Menu bar** (“File”, “Edit” and so on) is located in the upper area.



In the lower left corner, two tabs are available: **Data View** and **Variable View**. When you start SPSS, **Variable View** is default.

### *File types*

SPSS uses three types of files with different functions and extensions:

Type	Extension	Content
<b>Data set</b>	.sav	Data and variables
<b>Syntax</b>	.sps	Commands and comments
<b>Output</b>	.spv	Results

## Options

The SPSS menu works similar to the menus in many other programs, such as Word or Excel. Some useful options are listed below:

Option	Description
<b>Open a file</b>	Go to <b>File\Open</b> and choose <b>Data</b> , <b>Syntax</b> or <b>Output</b> . Browse your hard drive to locate the file. Then click on <b>Open</b> .
<b>Save a file</b>	Go to <b>File\Save As</b> . Type in a descriptive name and then click <b>Save</b> . You can also choose to save only some of the variables into the new data set by clicking <b>Variables</b> (before saving) and then ticking the boxes next to the variables you want to keep.
<b>Overwrite a file</b>	Go to <b>File\Save</b> .
<b>Import a data set</b>	Go to <b>File\Open\Data</b> and choose the desired format next to <b>Files of type</b> . Browse the file and click on <b>Open</b> .

## 1.2 Variable view

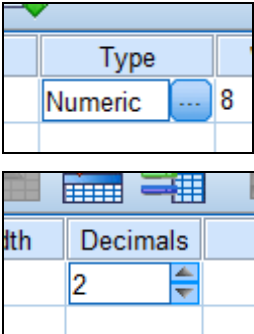
In **Variable View**, different columns are displayed. Each line corresponds to a variable. A variable is simply a quantity of something, which varies and can be measured, such as height, weight, number of children, educational level, gender and so forth.

Column	Function
<b>Name</b>	Name of the variable. It is your own choice, but make it understandable and do not use numbers or symbols as the first letter since SPSS will not accept it. Moreover, you cannot use spaces in the name. For example: "edu_level"
<b>Type</b>	Indicates the variable type. The most common is <b>Numeric</b> (only accepts numerical data, for example age or number of children) and <b>String</b> (also accepts letters, e.g. for qualitative questions). Typically, all responses in a questionnaire are transformed into numbers. For example: "Man"=0 and "Woman"=1, or "Non-smoker"=1, "Ex-smoker"=2 and "Current smoker"=3.
<b>Width</b>	Corresponds to the number of characters that is allowed to be typed in the data cell. Default for numerical and string variables is 8, which only needs to be altered if you want to type in long strings of numbers or whole sentences.
<b>Decimals</b>	Default is 2 for numerical variables and will automatically be displayed as .00 in the data view, if not otherwise specified.
<b>Label</b>	The description of the variable. Use the question that the variable is based upon or something else accurately describing the variable. For example: "What is your highest level of education?"
<b>Values</b>	Here you can add labels to each response alternative. For example: For the variable <b>gender</b> , "Men" are coded as 0 and "Women" are coded as 1. Through the option <b>Values</b> you tell SPSS to label each number according to the correct response. Next to <b>Value</b> (below <b>Value Labels</b> ), type in "0" and next to <b>Label</b> , type in "Men". Then click <b>Add</b> . Next to <b>Value</b> (below <b>Value Labels</b> ), type in "1" and next to <b>Label</b> , type in "Women". Then click <b>Add</b> .
<b>Missing</b>	By default, missing values will be coded as "." (dot) for numerical variables in the data set. For missing values in <b>String</b> variables, cells will be left blank.

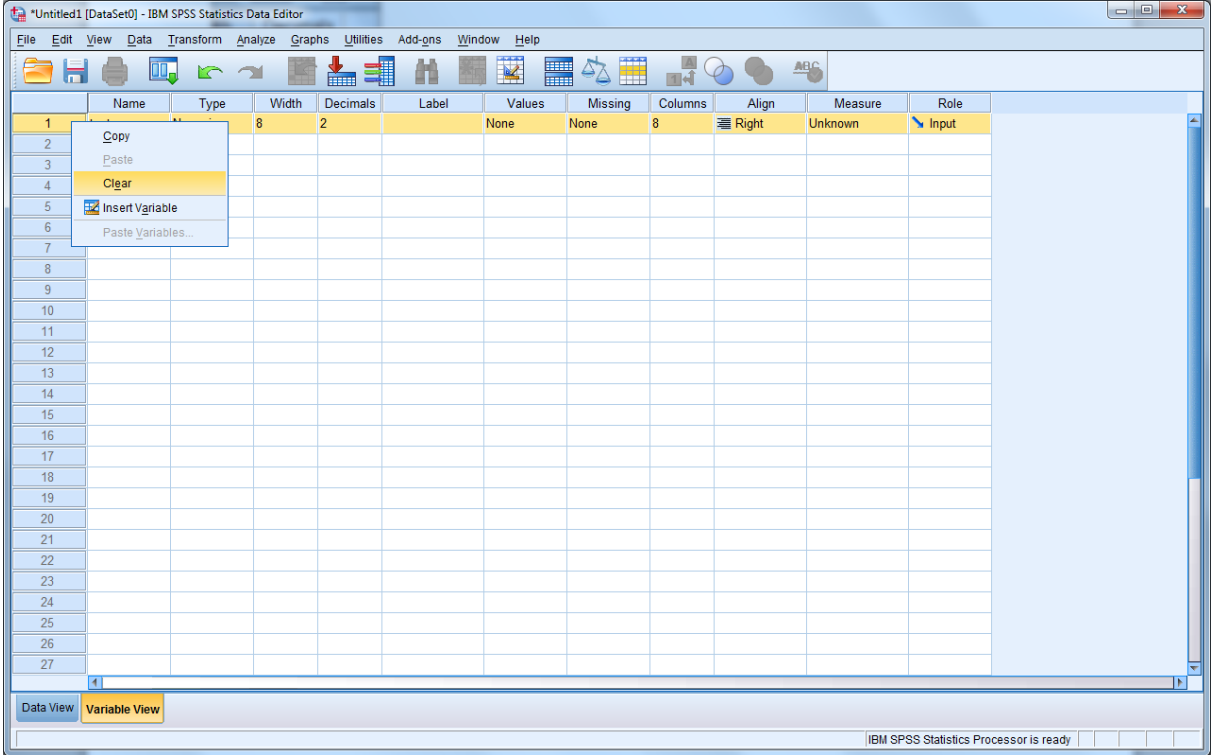
There are some additional columns in **Variable View**, but just ignore them for now. Normally, they do not need to be altered.

### Options

To alter the variable options, you may click the cells. Some columns can be typed directly into, while you need to press the arrows or dots that appeared when you click in the columns. It is often possible to use “copy and paste” here – this may efficient when you, for example, have several variables with the same **Values**.



If you want to delete a variable, select the numbered cell to the left of the variable and then right-click and choose **Clear**.



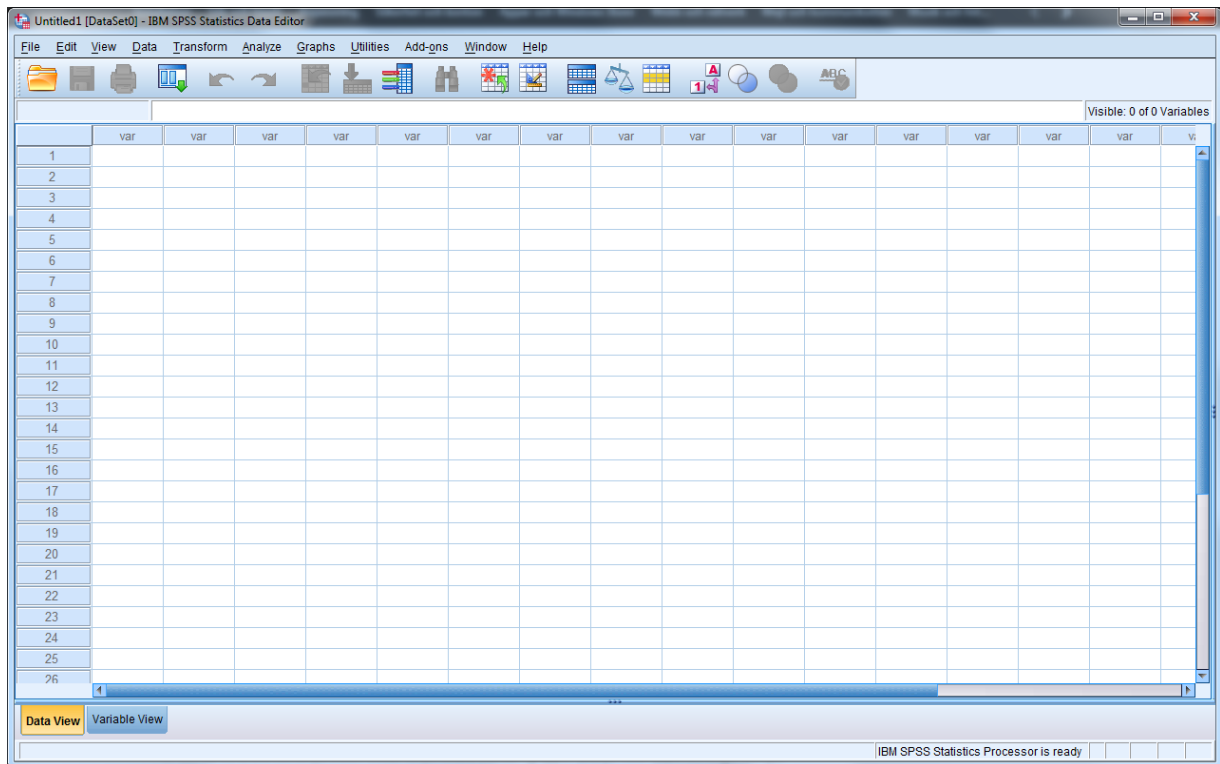
### 1.3 Creating a new data set

If you have a questionnaire, you can easily create the corresponding data structure in **Variable View** in SPSS. For example:

Name	Type	Width	Decimals	Label	Values	Missing
gender	Numeric	1	0	Gender	0=Man 1=Woman	None
srh	Numeric	1	0	Self-rated health	1=Poor 2=Fair 3=Good 4=Excellent	None
income	Numeric	10	0	Disposable income	None	None

## 1.4 Data view

Once the structure of the data set is determined, it is time to take a look at **Data View**. Access this view by clicking on the tab named **Data View** in the lower left corner.



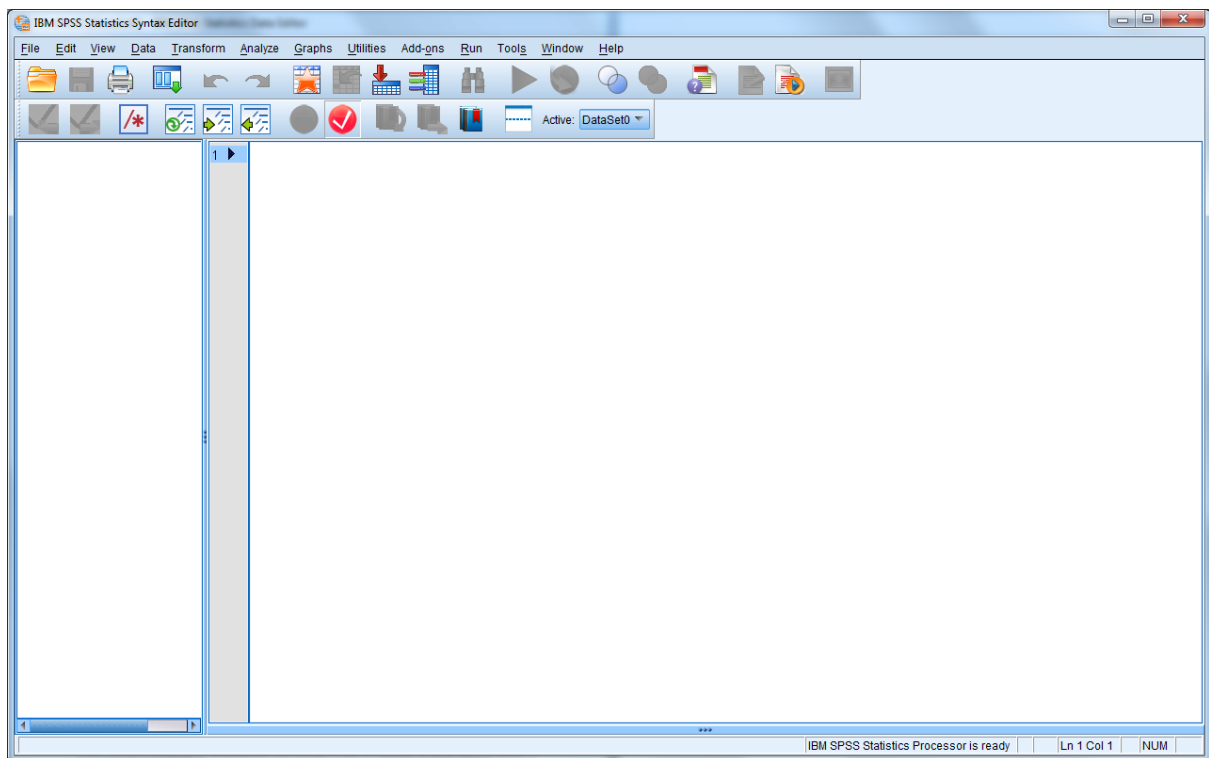
Here, each column corresponds to a variable, whereas each row corresponds to a case (most commonly an individual). It is possible to change the order of the variables by highlighting a column and “drag and drop”. You may also change the width of the column by placing the mouse over the right border of a column (next to the name of the column), pressing down the button and then “drag and drop”.

If you are creating a new data set, simply type in your data, one row (and one column) at a time. Use the left and right arrow key on your key board to move between cells.

Make sure that you have chosen the right **Type** of variable before you enter your data (i.e. **Numeric** or **String**).

## 1.5 Syntax

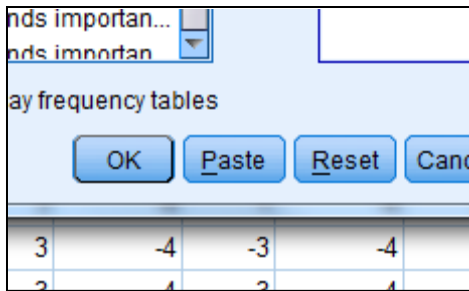
The syntax is presented in a new window called **IBM SPSS Statistics Syntax Editor**. Note that the **Menu bar** is very similar to the one available through **Data View** and **Variable View**. In order words, you do not always have to go back to **Data View** or **Variable View** to access the **Menu bar**.



There are two text areas: the big one to the right is where all commands will be displayed. What is a command? It is basically when you tell SPSS that you want it to perform a specific test or analysis or to create a table or a graph. All these features have their own commands. To the left, you will have a “short list” of the same commands.

### *How to include a command in the syntax*

There are two ways of including a command in the syntax. The first alternative is to go through the menu system. For all features provided in the **Menu bar**, you will have a button called **Paste**. If you click on this button, the correct command will be pasted into the syntax you currently have open. If you have several syntax windows open, make sure that the command is pasted into the correct one.



The second alternative is to write the command yourself. If you have never used SPSS syntax before, this is not recommended in the beginning. However, once you have created a “stock” of commands using **Paste**, you may start re-using these commands by copy, paste and then alter them directly in your syntax. When you write your own commands in the syntax, the functions will have different colours. The command will be red if it is *not* written properly.

Once you have a long list of commands in your syntax, it may be difficult to keep track of them. It is therefore highly recommended that you comment your syntax. You may add a heading above each command, and also make notes of interesting findings etc. In order to insert a comment, write an asterisk (\*) before the comment begins. End the comment with a dot/period (.). If the comment is not framed with an asterisk and a dot, it may be tangled up in the commands. You can easily double-check that the comments are correctly entered, because if they turn grey if they are.

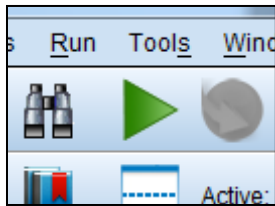
### Examples

- \* This is a frequency table of gender.
- \* Linear regression analysis of educational level (independent variable) and income (dependent variable).



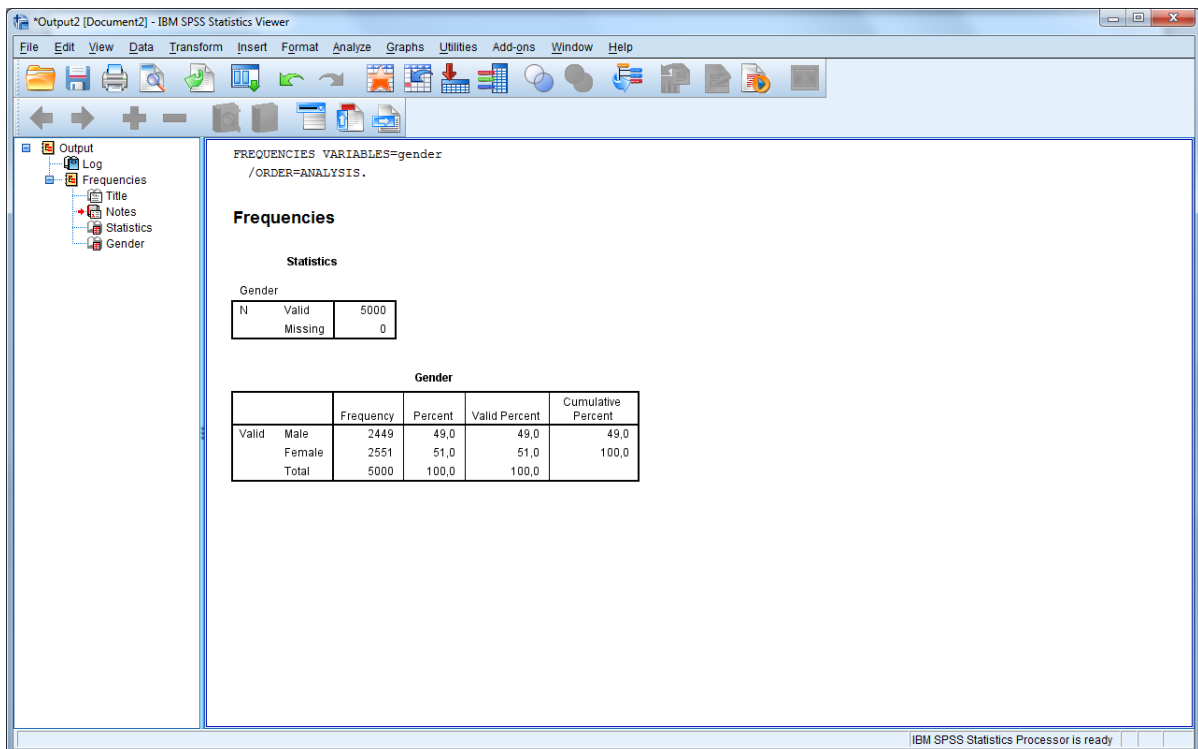
### *How to execute the command*

To execute the command, highlight all the rows for that specific command and press the big green arrow below the **Menu bar** in the syntax window. You do not have to highlight one command at a time - it is possible to execute several commands at once. If you have added comments to your syntax, make sure that they are grey (i.e. inactive), otherwise SPSS will take them for (faulty) commands.



## 1.6 Output

Everything you order in SPSS (e.g. graphs, tables, or analyses) ends up in a window called **Output**. In the area to the left, all the different steps are listed. It is possible to collapse specific steps by clicking on the box with the minus sign (and expand it again by clicking on the same box, now with a plus sign). In the area to the right, your actual output is shown. First, you see the syntax for what you have ordered, and then you get the tables or graphs related to the specific command.



The screenshot shows the IBM SPSS Statistics Viewer window. The left pane displays a tree view of the Output window, with 'Frequencies' expanded to show 'Statistics' and 'Gender'. The main pane displays the following output:

```
FREQUENCIES VARIABLES=gender
/ORDER=ANALYSIS.
```

**Frequencies**

**Statistics**

Gender

N	Valid	Missing
	5000	0

**Gender**

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid Male	2449	49,0	49,0	49,0
Female	2551	51,0	51,0	100,0
Total	5000	100,0	100,0	

IBM SPSS Statistics Processor is ready

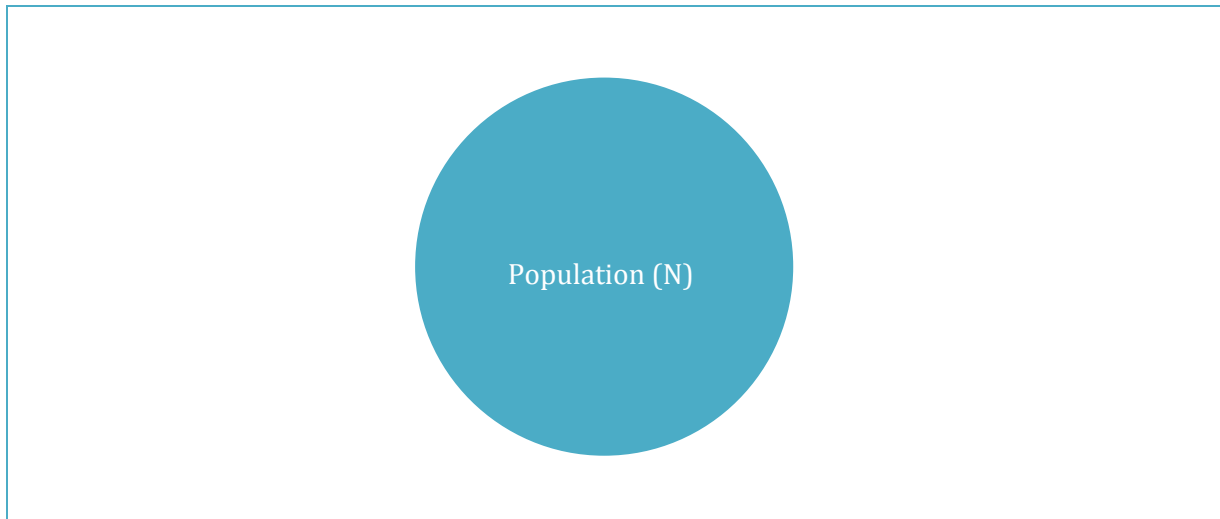
## 2. BASIC STATISTICAL CONCEPTS

The first part of this chapter is devoted to issues related to populations and samples. These are things you need to be aware of in order to make correct judgements of your data material. Before it is possible to describe the variables in the data set through the different commands in SPSS, we need to know more about the specific variables. Here, we will address two major aspects: measurement scales and distributions.

Outline
<b>2.1</b> Study population and sampling
<b>2.2</b> Measurement scales
<b>2.3</b> Distributions

## 2.1 Study population and sampling

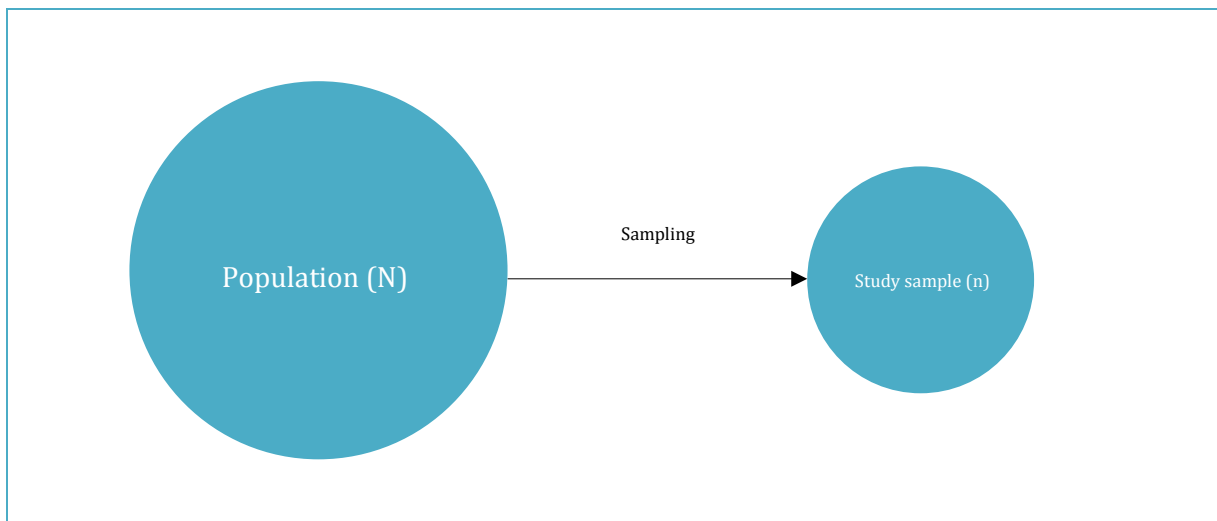
Before we start looking into our data, we need to establish what our population is, since the population is what we want to say something about. A population is often referred to by “N”.



A population can be almost anything: We have populations which are geographically defined, such as the world, a country or a city; we have age-defined populations such as teenagers, infants and elderly; and also specific groups such as women, drug addicts, teachers, master students, and so on.

## Sampling

It is seldom the case that we examine the whole population which we have chosen. Instead, we use sampling – that basically means that we take a smaller sample of the population: a study sample. A study sample is often denoted by “n”. The reasons behind sampling are primarily that it is very costly and time consuming to collect data for the entire population. However, sometimes you can include the whole population - like if you have small populations, such as one school or one hospital or one company (this is often referred to as a case study). Another example is when you use national registers (then you usually do not have to considered aspects such as time or cost since the data is already available).



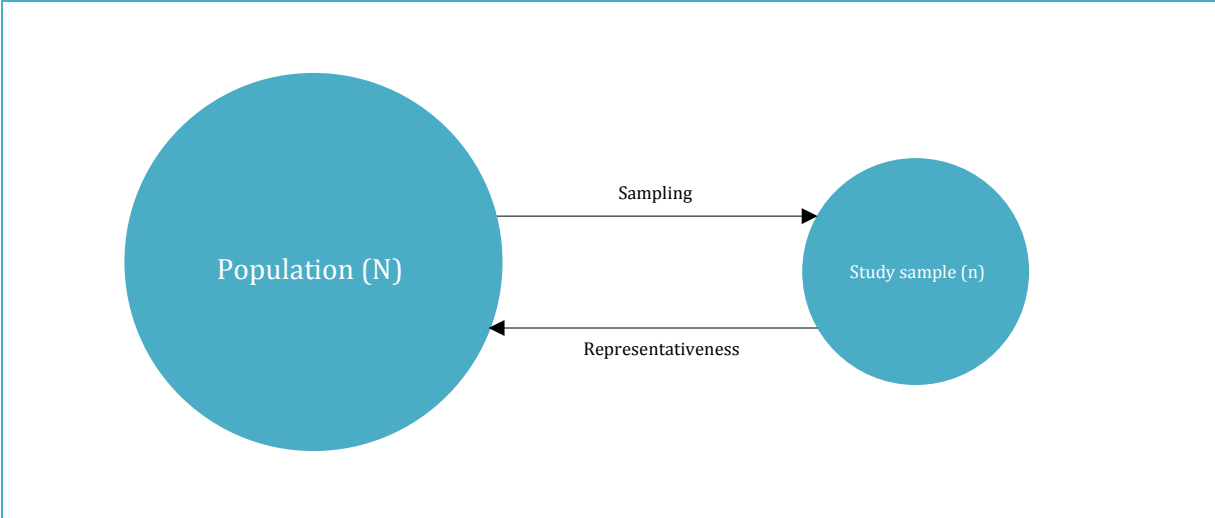
There are many different sampling techniques available. Generally, they can be categorised into two types that include several sub types:

Types of sampling	
<b>Non-probability sampling</b>	
Snowball	Finding respondents through already selected respondents
Quota	Adding suitable individuals until a certain quota is achieved
Convenience	Easy access of respondents
<b>Probability sampling</b>	
Random	Every individual has the same chance of being selected
Systematic	Sampling with intervals, e.g. every fifth of a list
Stratified	Random sampling from different groups
Clustered	Random sampling of groups, choosing all individuals from these groups

Non-probability sampling is most common in small-scale studies, marketing research, interview studies and studies like that. Snowball sampling means that you start out with some respondents and ask them to find other suitable respondents (like friends or other people they know). Quota sampling is often used in marketing research. For example, the researchers wants to have 100 respondents who have tried a new coffee brand and stands outside the store until they have found 100 persons who have bought that specific brand. Then we have convenience sampling. This is when you pick respondents who are easy to get access to, like friends, family, or members of an organisation that you are a member of yourself, and so on.

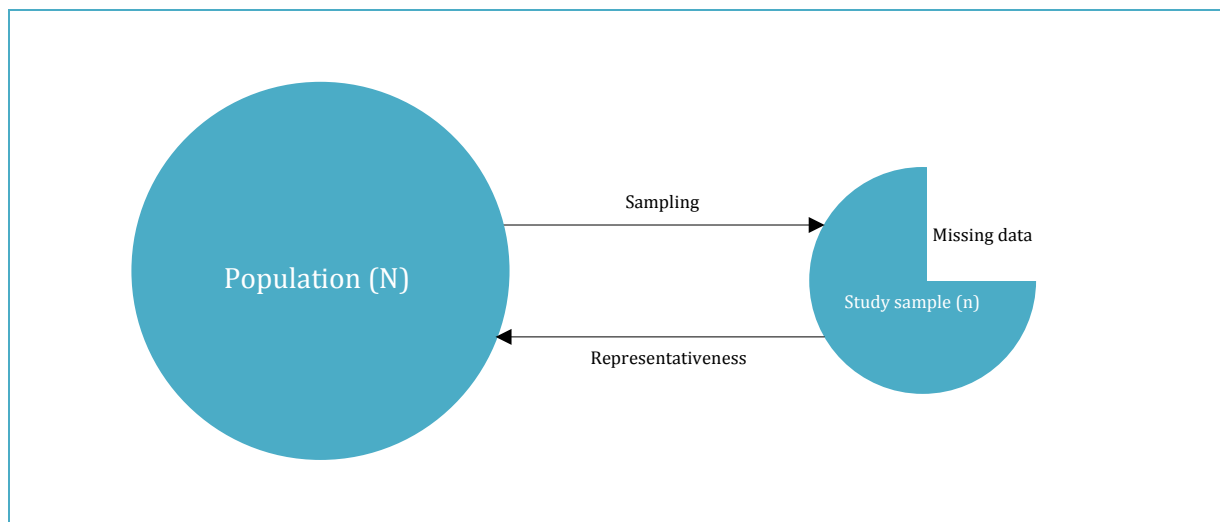
The other type of sampling is probability sampling. First, we have the random probability sampling, which postulates that every individual in the population should have the equal chance of being selected. Another procedure is the systematic sampling, where you, for example, draw every fifth or seventh from a list of people. Stratified sampling is when you draw random samples from some specific groups. For example if you want to compare labour market outcomes between Swedes and immigrants, you may not get a large enough sample of immigrants if making a random sampling. Therefore you can draw a larger random sample from the smaller group. Finally, we have clustered sampling. Perhaps you start out by drawing a random sample of schools and then select all students attending ninth grade in these schools.

Probability sampling constitutes the foundation of quantitative data analysis. Why is it so important? Well, we want our study sample to be “representative”. This means that it should have the same characteristics as our population. This is a requirement to be able to draw conclusions about the population based on the study sample (also known as generalizability).



### *Missing data: attrition and non-response*

An issue that almost all quantitative researchers deal with has to do with “missing data”. What is missing data? Well, when we have defined our population and conducted a probability sampling, we start collecting data for the individuals in our study sample – either through questionnaires or registers (or both). It is very seldom the case, however, that we get complete information for all individuals. We thus get missing data. When we use register data, missing data is commonly called “attrition”, and when we use survey data (i.e. questionnaire data), missing data is usually called “non-response”. If we have problems with missing data, we may not be able to draw conclusions about the population based on the study sample. This is discussed in further detail in Section 12.5.



## 2.2 Measurement scales

### *Types of scale*

We use a scale to make the measurements of a variable, and the characteristics of the scale determine the characteristics of the data we collect and, in turn, how we describe our data. Generally speaking, there are four measurement scales: nominal, ordinal, ratio and interval. Nominal and ordinal variables are often called “categorical” or “qualitative”, whereas ratio and interval variables are often referred to as “continuous” or “quantitative”.

Name	Type
<b>Nominal</b>	Categorical/qualitative
<b>Ordinal</b>	
<b>Ratio</b>	Continuous/quantitative
<b>Interval</b>	

It should also be noted that a nominal variable with only two categories/values is called “dichotomous” (or “binary, or “dummy”) whereas a nominal variable with more than two categories is called “polytomous”.

### *Differences between the scales*

These scales differ in three important ways: hierarchy, distance and zero point.

Checklist	
<b>Is it possible to arrange/order the values hierarchically?</b>	Yes/No
<b>Is it the same distance between the values?</b>	Yes/No
<b>Does the scale have an absolute zero point?</b>	Yes/No

What does “arrange/order the values hierarchically” mean? If we take gender as an example, it is not reasonable to say that “Man” is less or more than “Woman”. As another example, we can take ethnicity: it is not reasonable to see “Danish” as less or more than “Finnish”. For variables such as self-rated health, on the other hand, it is possible to say that “Excellent health” is better than “Good health”. Moreover, it is possible to say that the grade “A” is better than the grade “B”.



What does “distance” mean? If we take income as an example, we know that 1000 dollars are twice as much as 500 dollars, and 2000 dollars are twice as much as 1000 dollars. The same logic applies to variables such as age: it is the same distance between 2 years and 4 years as between 6 years and 8 years. Thus, having the same distance between the values means that the differences between two values are the same regardless of which part of the scale you are looking at.

What does “absolute zero point” mean? Basically, it means that the scale cannot have negative values. It is possible for the temperature to be minus 10 degrees Celsius, but is not possible to have less than zero years of schooling or having less than zero days of unemployment.

### *Examples*

Below, we can see some examples of variables on the different measurement scales.

Scale	Values	Examples
<b>Nominal</b>	Order values: No	Yes/no questions
	Same distance: No	Gender
	Absolute zero point: Not applicable	Ethnicity
<b>Ordinal</b>	Order values: Yes	Attitude questions
	Same distance: No	Self-rated health
	Absolute zero point: Not applicable	Educational level
<b>Ratio</b>	Order values: Yes	Age
	Same distance: Yes	Income
	Absolute zero point: Yes	School marks
<b>Interval</b>	Order values: Yes	Temperature (Celsius)
	Same distance: Yes	
	Absolute zero point: No	

A nominal variable is hence a variable for which the values cannot be ranked, and we do not have the same distance between the values, e.g. gender or questions that can be answered with yes or no. Ordinal variables are similar, but here the values can be ranked, such as for self-rated health: “Excellent is better than “Good”; “Good” is better than “Fair”; and “Fair” is better than “Poor”. However, for ordinal scales we do not have the same distance between the values: the “amount” of better health is not necessarily the same between “Poor” and “Fair” as between “Good” and “Excellent”. The ratio scale is similar to the ordinal scale, but here we do have the

same distance between the values: for example, we know that 10 years of schooling is twice as much as 5 years of schooling. The interval scale is similar to the ratio scale, but here we also have an absolute zero point.

*Types of values*

It is possible to distinguish between two types of values: discrete and continuous. Discrete values can only assume “whole” values, such as “Man”, “Women”, “Green”, “Car”, and “House”. Continuous values can assume any value along a scale, such as “3.5 years”, “58.3 seconds”, and “163.5 centimetres”. Note, however, that continuous variables (i.e. on a ratio or interval scale) do not necessarily have continuous values. For example, number of cars is a ratio variable but it has discrete values: while the average number of cars in a population may be 0.8, it is not correct (although many do) to say that any given individual in a population has 0.8 cars (since a car is a “whole” value”).

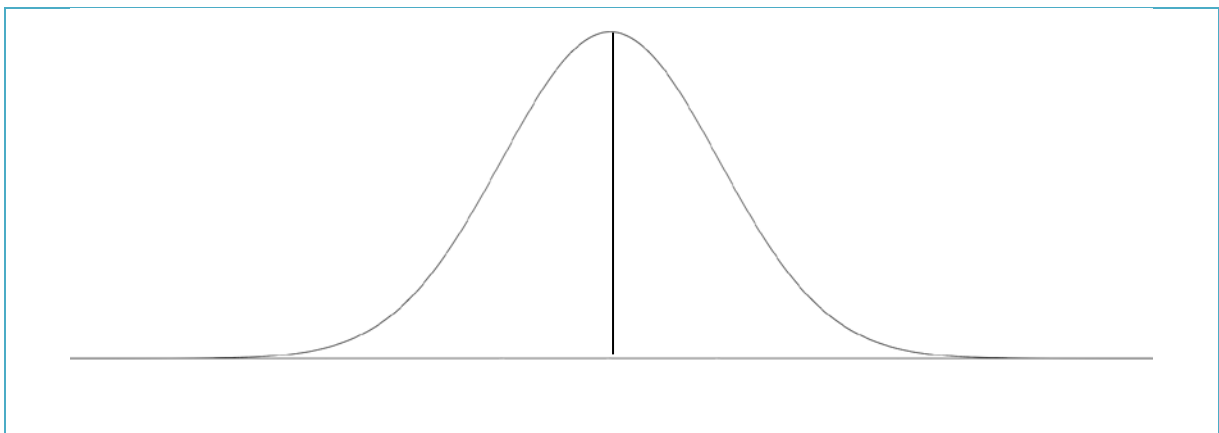
Name	Type
<b>Discrete</b>	“Whole” values
<b>Continuous</b>	Any value

## 2.3 Distributions

For continuous variables (i.e. on a ratio or interval scale) it is important to know what the distribution of values in the variable looks like.

### *Normal distributions*

One common type of distribution is the “normal distribution”. Many statistical methods are based on normal distributions.



The above figure is an example of a typical normal distribution. Here are some basic facts about the normal distribution:

#### Basic facts about normal distributions

Always bell-shaped.

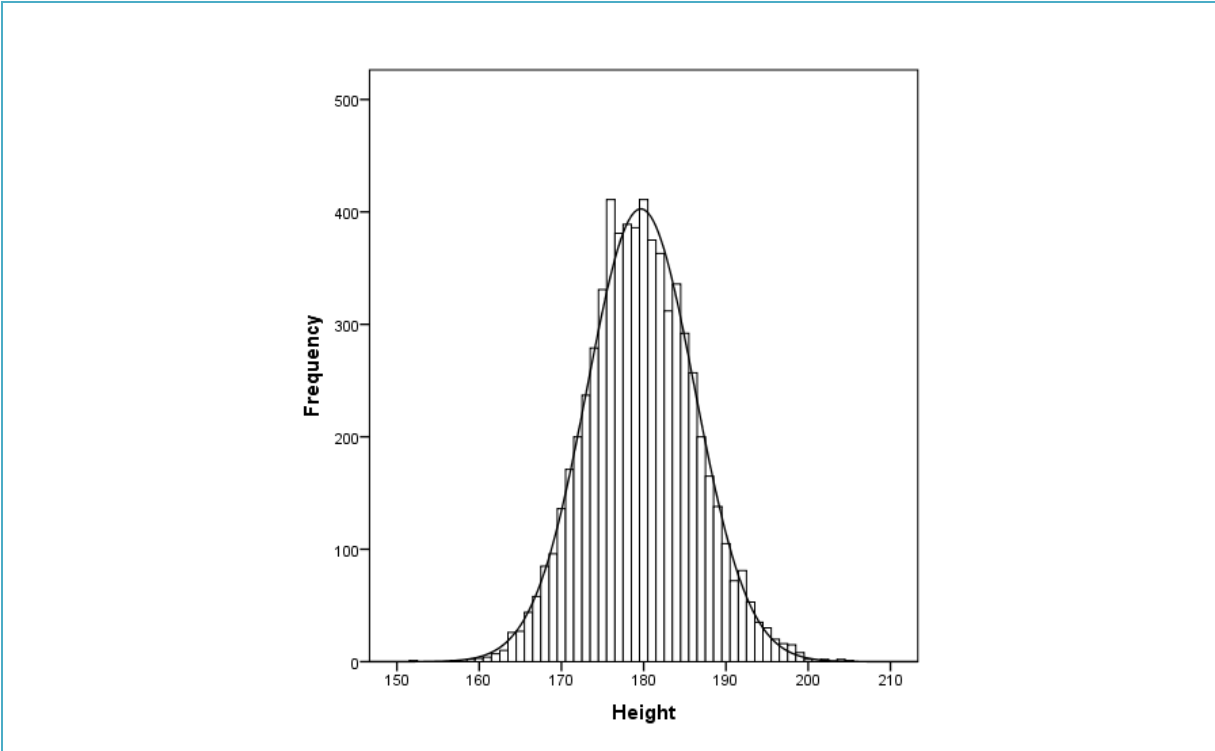
The peak always indicates the mean value.

Always symmetrical, i.e. the tails on each side of the mean are equally large. This means that 50 % of the values are on one side of the mean, and 50 % of the values are on the other side of the mean.

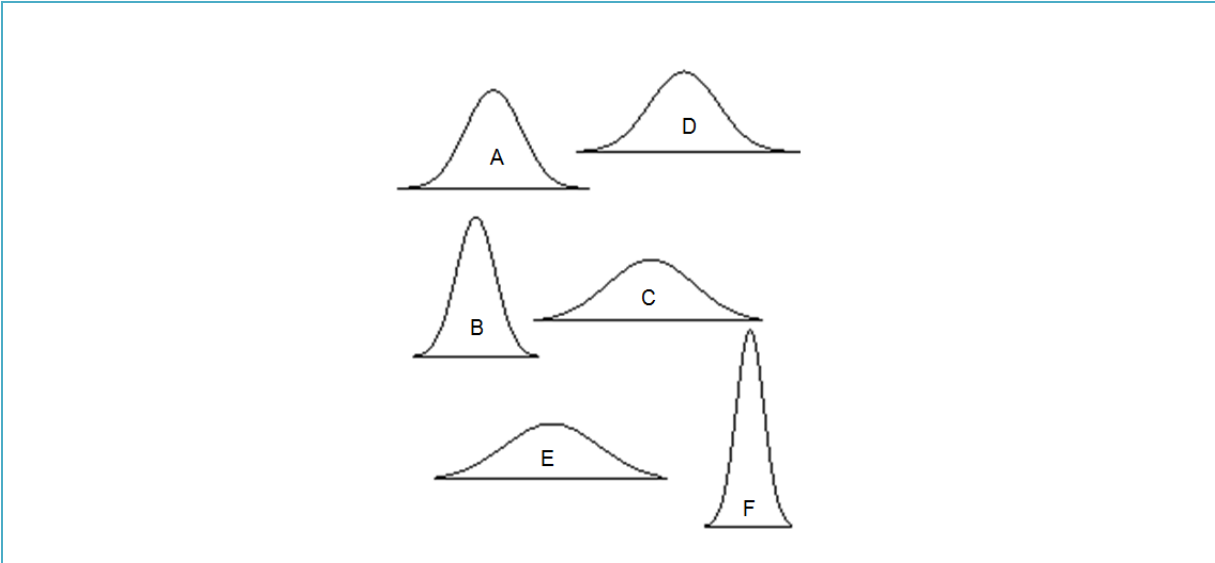
The area under the curve is always 1 (100 % of the values).

Below is an example of a (normal) distribution of height among Swedish men at the time of military service enlistment (in Swedish: “lumpen”). In this example, the mean height is about 180 centimetres. The less common a certain height gets, the smaller the area under the curve. Here, the tails are about equally large on both sides of the mean, suggesting that it is

approximately as common for individuals in the sample to be shorter than the mean as it is for them to be taller than the mean.

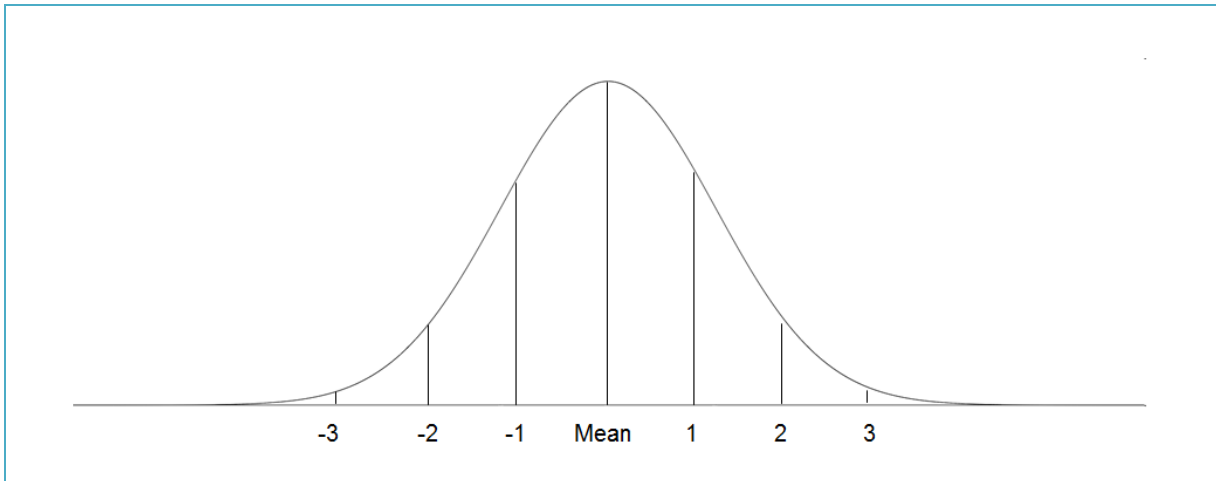


Normal distributions can look quite different. The figures below are all examples of normal distributions. The difference lies in the amount of spread of the values: because the shape of a normal distribution is not only defined by the mean value, but by the standard deviation!



But what is “standard deviation”? A simple definition is that it expresses how much variation exists from the mean for a given variable. If we have a low standard deviation, it suggests that

the individuals in our data have values close to the mean, and if we have a high standard deviation, it indicates that the values are more spread out over a large range of values.



The empirical rule of normal distributions tells us the following (see the figure above):

- 68 % of the values fall within -1 and +1 standard deviations.
- 95 % of all values fall within -2 and +2 standard deviations.
- Nearly 100 % of all values fall within -3 and +3 standard deviations.

### Example

We have collected information about weight for a sample of individuals. If the mean weight in this sample was 70 kilos and the standard deviation was 5 kilos, the empirical rule would give us the following information:

68 % of the individuals have a weight of 65-75 kilos:

Lower limit:  $70 \text{ kilos} - (5 \text{ kilos} * 1)$ ; upper limit:  $70 \text{ kilos} + (5 \text{ kilos} * 1)$

95 % of the individuals have a weight of 60-80 kilos:

Lower limit:  $70 \text{ kilos} - (5 \text{ kilos} * 2)$ ; upper limit:  $70 \text{ kilos} + (5 \text{ kilos} * 2)$

Nearly 100 % of the have a weight of 55-85 kilos:

Lower limit:  $70 \text{ kilos} - (5 \text{ kilos} * 3)$ ; upper limit:  $70 \text{ kilos} + (5 \text{ kilos} * 3)$

As long as we have information about the mean value and the standard deviation, it is possible to do the same calculation for all the normal distributions. Remember that a more pronounced

peak indicates a low standard deviation, whereas a flat distribution indicates a high standard deviation.

### *Skewed distributions*

There are other types of distribution. One very common type of distribution is the “skewed distribution”. Here are some facts about skewed distributions:

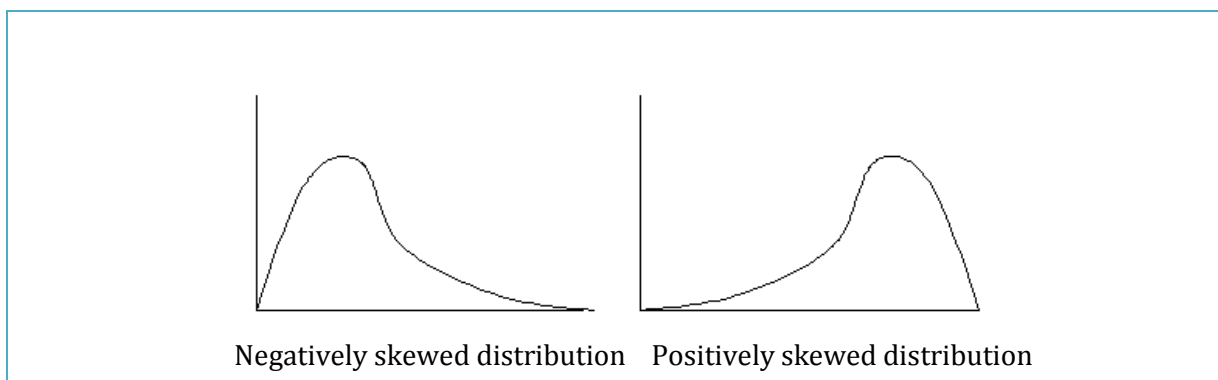
#### Basic facts about skewed distributions

Always asymmetrical = tails are different, i.e. the empirical rule does not apply

Skew can be positive (right tail longer) or negative (left tail longer)

Examples of a positively skewed distribution (like the figure to the left) are: number of hospital visits, number of days in unemployment, number of telephone calls during a day. Most individuals will have the value zero or a low value, whereas a few will have increasingly high values.

Examples of a negatively skewed distribution (like the figure to the right) are: age of retirement, or a very easy test. Most individuals will have high values, and then a few will have very low values.



The skewness of the distribution can be indicated by two types of measure: skewness and kurtosis.

#### Facts about the skewness measure

Measure of the symmetry of a distribution.

Negative skewness value = the distribution is skewed to the right (longer tail to the right)

Positive skewness value = the distribution is skewed to the left (longer tail to the left)

A perfect normal distribution has a skewness of 0

Skewness value between -2 and +2 is usually considered acceptable

#### Facts about the kurtosis measure

Measure of the shape (or the "peakedness") of a distribution

A perfect normal distribution has a kurtosis of 0 (mesokurtic distribution)

Kurtosis value above 0 = Leptokurtic distribution (sharper peak and longer/fatter tails)

Kurtosis value below 0 = Platykurtic distribution (rounder peak and shorter/thinner tails)

Kurtosis value between -2 and +2 is usually considered acceptable

### 3. DESCRIPTIVE STATISTICS

When we know about the measurement scale and the distribution of the variables in our data set, we can decide on how to best describe our variables. In this type of exploratory data analysis, we use a set of tables and graphs as well as measures of central tendency and variation. Here, we will address the following types of description:

Outline
3.1 Tables
3.2 Graphs
3.3 Measures of central tendency
3.4 Measures of variation

Two types of table will be covered: frequency table and cross table. With regard to graphs, we will discuss bar chart, pie chart, histogram and scatterplot. For measures of central tendency, the mean, mode, and median are addressed. Moreover, some examples of measures of variation will be included here, namely minimum, maximum, range, and standard deviation.

Going back to what we learnt about measurement scales and the distributions, this is generally how you should match the different types of variables with the different types of description:

Type of variable	
<b>Categorical (nominal/ordinal)</b>	Frequency table Cross table Bar chart Pie chart Mode
<b>Continuous (ratio/interval)</b>	Histogram Scatterplot Mean (if normal distribution) Median (if skewed distribution) Min Max Range Standard deviation



### 3.1 Tables

Tables are useful if one wants to see the distribution of values for categorical (nominal or ordinal) variables. Here, we will discuss frequency tables and cross tables.

A frequency table is a simple but very useful description of one variable and gives us both the frequency and various types of percentages of individuals with the different values.

Column	Content
<b>Frequency</b>	The number of individuals in the different categories.
<b>Percent</b>	The distribution of percent also taking into consideration any missing information. This means that if some individuals would have missing information about gender, the percentages in this column would be dependent upon that.
<b>Valid Percent</b>	Same as <b>Percent</b> but does not take missing into account. This column is what we primarily focus on.
<b>Cumulative percent</b>	Adds the percentages from top to bottom.

Below, compare the first table with the second table to see the differences between **Percent** and **Valid Percent** (in the second table, the information about gender has been removed for one of the individuals).

Gender					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Man	7	46,7	46,7	46,7
	Woman	8	53,3	53,3	100,0
	Total	15	100,0	100,0	

Gender					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Man	7	46,7	50,0	50,0
	Woman	7	46,7	50,0	100,0
	Total	14	93,3	100,0	
Missing	System	1	6,7		
	Total	15	100,0		

A cross table is a description of how individuals are distributed according to two variables. The table in the example below, is a cross table with the variables gender and health. Here, it is possible to see the distribution of health by gender, and vice versa.

		Gender		Total
		Man	Woman	
Health	Good	2	3	5
	Poor	5	5	10
Total		7	8	15

The first table below also includes “column percentages”, demonstrating the distribution of health according to gender. The second instead adds “row percentages”, demonstrating the distribution of gender according to health status. Note that the frequencies (i.e. the number of individuals) in the cells are the same, but the percentages are different since the focus shifts between the tables. If you find this difficult to separate in your mind, one good advice is perhaps to see where the percentages add up to 100 % in **Total** - in the rows or in the columns.

			Gender		Total
			Man	Woman	
Health	Good	Count	2	3	5
		% within Gender	28,6%	37,5%	33,3%
	Poor	Count	5	5	10
		% within Gender	71,4%	62,5%	66,7%
Total		Count	7	8	15
		% within Gender	100,0%	100,0%	100,0%

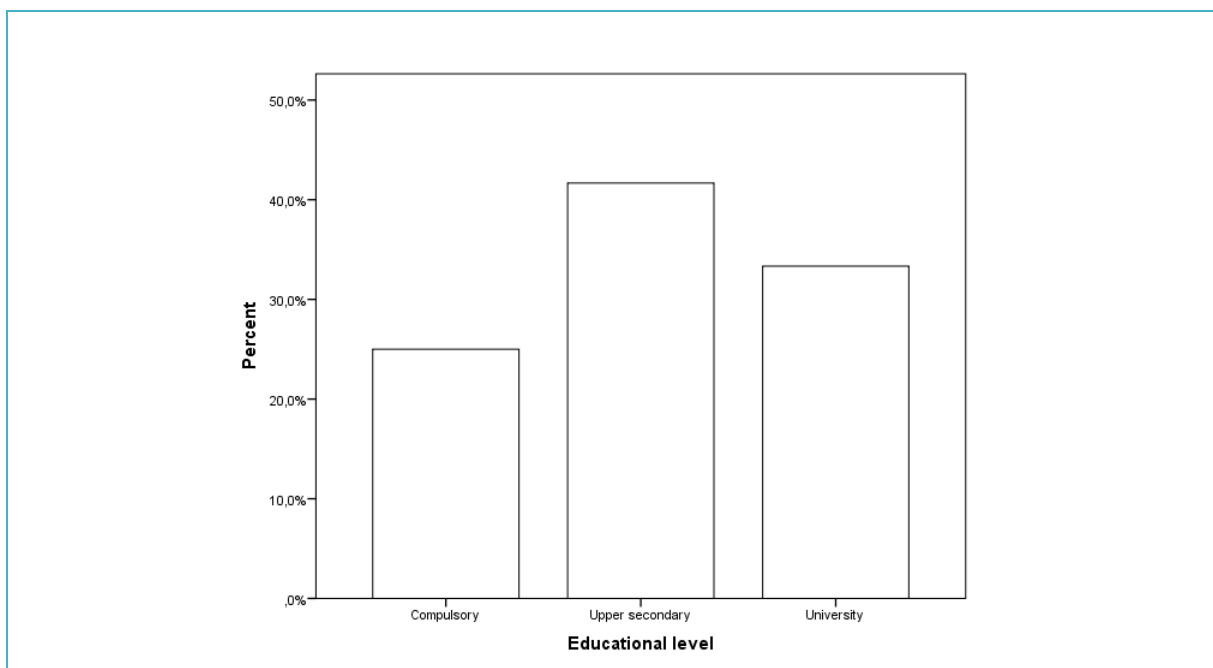
			Gender		Total
			Man	Woman	
Health	Good	Count	2	3	5
		% within Health	40,0%	60,0%	100,0%
	Poor	Count	5	5	10
		% within Health	50,0%	50,0%	100,0%
Total		Count	7	8	15
		% within Health	46,7%	53,3%	100,0%

Important to note is that if we would have individuals with missing information with regard to gender or health, these would be excluded from the cross table. Therefore, it is important to always look at the two variables used in a cross table separately (through e.g. frequency tables) as well.

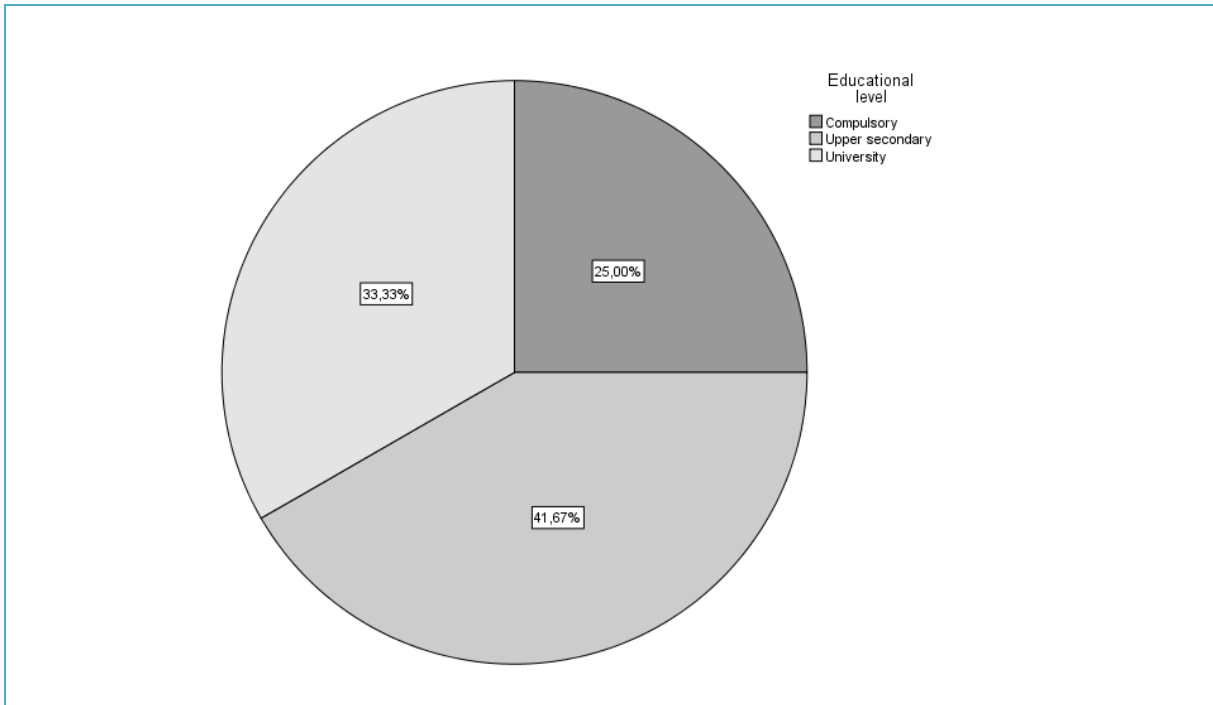
## 3.2 Graphs

For categorical variables, bar charts or pie charts may be useful. For continuous variables, we can use histograms (one variable) or scatterplots (two variables).

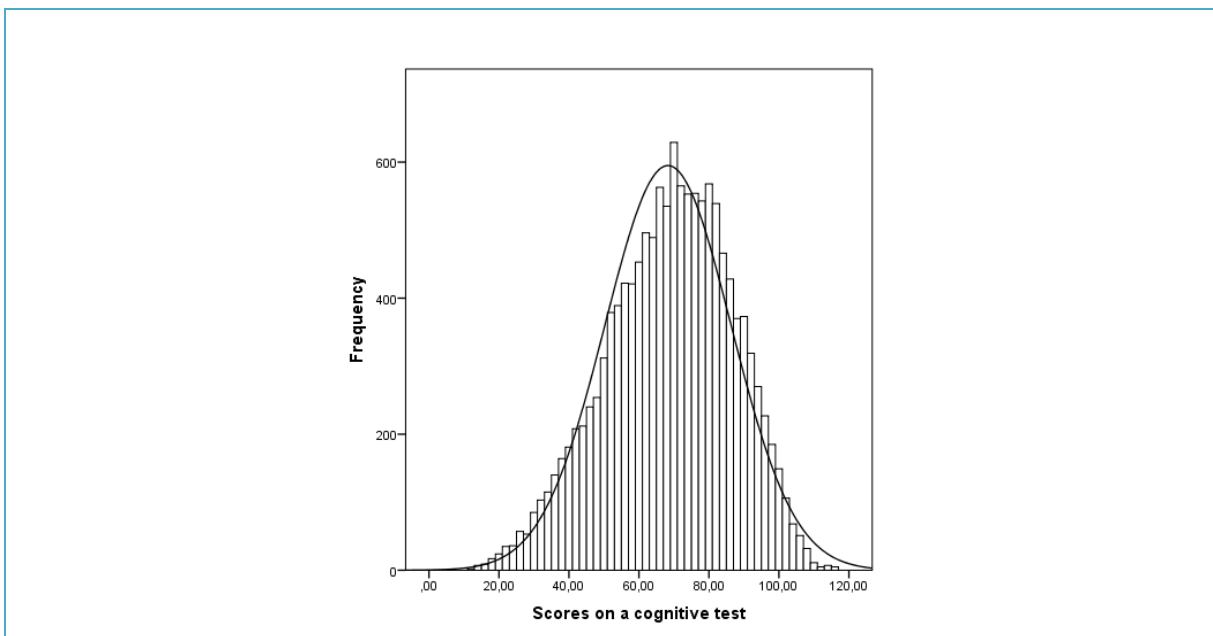
A bar chart is like an illustration of a frequency table. On the x-axis (horizontal axis) you see the different values (or categories) of the variable and on the y-axis (vertical axis) you can choose to see either the percentage of individuals in each category (like in the graph below) or the number of individuals in each category.



A pie chart can also be seen as a simple illustration of a frequency table. The slices represent the different values (or categories) of the variable and they can be specified in terms of the percentage of individuals in each category (like in the graph below) or the number of individuals in each category.

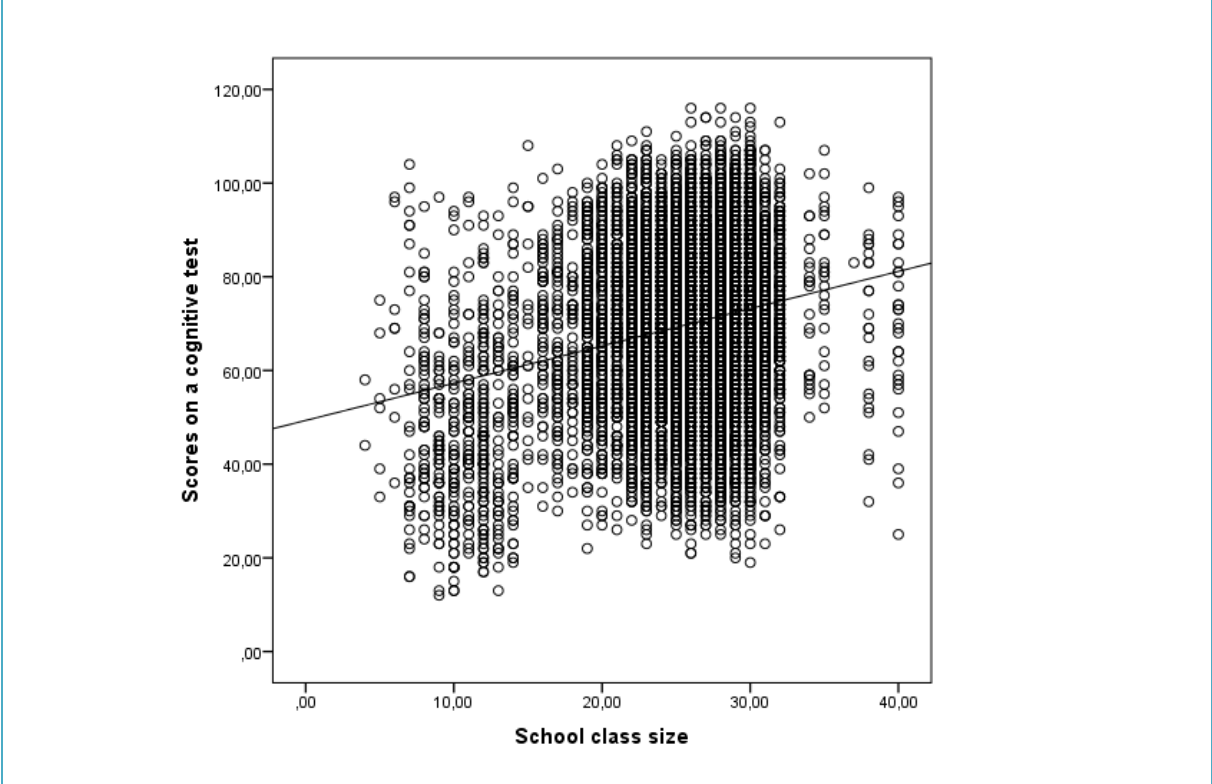


A histogram is similar to a bar chart but, unlike the bar chart, it is suitable for continuous variables. Here, a reference line representing the normal distribution curve can be added. The histogram will give us an idea about whether the distribution (of the continuous variable) is normal or skewed.



When we had two categorical variables, we could produce a cross table to see how these two variables were related. If we have two continuous variables, we may use something called a scatterplot instead. Each dot in the scatterplot represents one individual in our data. We may

also include a reference line here, to see if we have a pattern in our data: for example, if high or low values in one variable correspond to high or low values in the other variable. In the scatterplot below, we can see that increasing school class size seems to be related to increasing scores on a cognitive test, at least to some degree.



### 3.3 Measures of central tendency

Central tendency can be defined as measures of the location of the middle in a distribution. The most common types of central tendency are:

Measure	Definition
<b>Mean</b>	The average value
<b>Median</b>	The value in the absolute middle
<b>Mode</b>	The most frequently occurring value

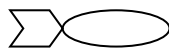
The mean is perhaps the most commonly used type of central tendency and we get it by dividing the sum of all values by the number of values.

#### Example

We have four fishes that weight:



1.1 kilos



0.8 kilos



1.1 kilos



1.0 kilos

What is the mean?

First we add the values together:  $1.1+0.8+1.1+1.0=4.0$

Then we divide the sum of the values by the number of values:  $4.0/4=1$ .

The mean is thus 1 kilo.

The median - i.e. the value in the absolute middle of the distribution - is obtained by sorting all the values from low to high and then identifying the value in the middle of the list.

**Example**

We have nine individuals who are of the following heights:



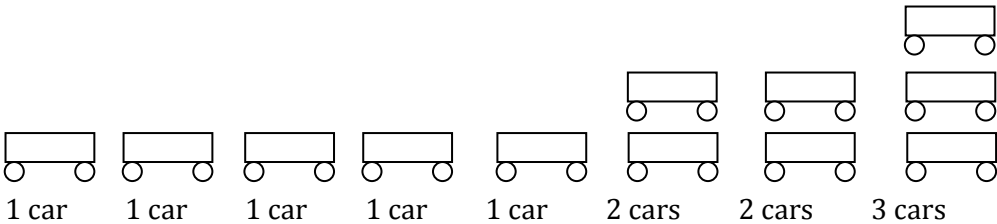
The median is thus 173 cm.

Note that when we have an odd number of values, it is easy to identify the value in the absolute middle of the distribution. When we have an even number of values, we get the median by adding the two values in the middle together and dividing the sum by 2.

The mode - or “type” - is defined as the most frequently occurring value in a distribution. Here as well, one starts by sorting responses from the lowest to the highest value and then identifies the most common value.

**Example**

We have information about the number of cars in the household:



The mode is thus 1 car (since this is the most common value).

The choice of type of central tendency is based on a) the measurement scale of the variable and b) the distribution of the variable. Generally, if the variable is categorical (nominal or ordinal), the mode is preferred. If the variable is continuous (ratio or interval), the mean or the median is

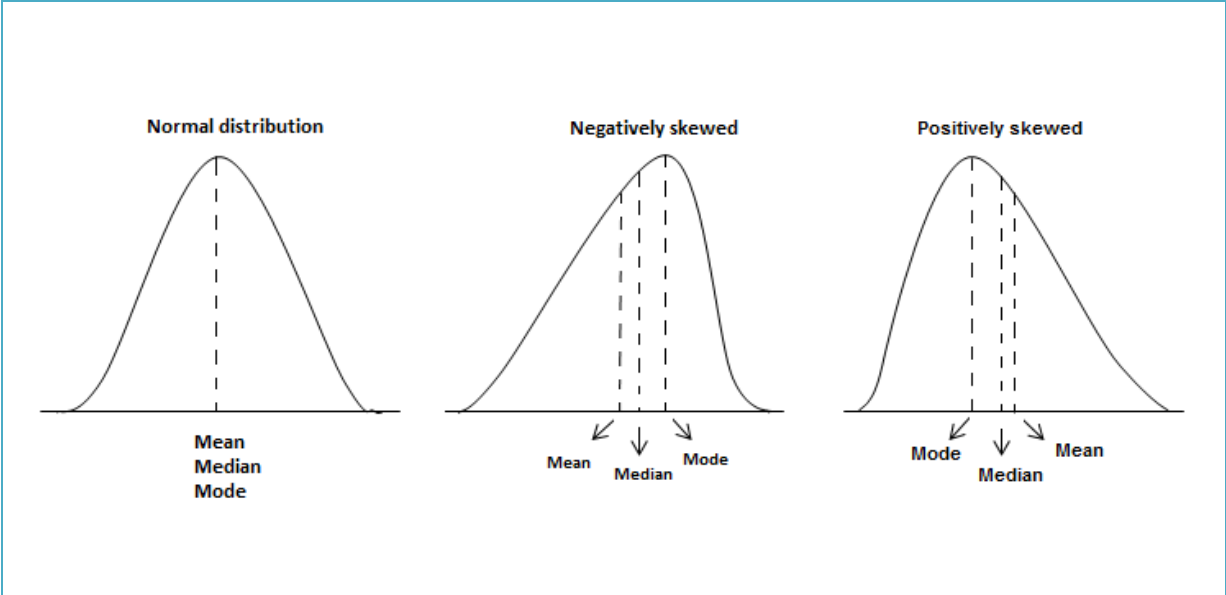
preferred. In the latter case, the mean is chosen if the variable is normally distributed and the median is chosen if the variable has a skewed distribution.

Scale	Type	Central tendency
<b>Nominal</b>	Categorical	Mode
<b>Ordinal</b>		
<b>Ratio</b>	Continuous	Normal distribution: Mean; Skewed distribution: Median
<b>Interval</b>		

Why should one not use the median or the mean for categorical variables? For nominal variables, it is easy to give an answer. Let us take country of birth as an example. In this example, the variable is coded into four categories: 1) Sweden, 2) China, 3) Canada, and 4) Norway. This is clearly a nominal variable. Since the order of the categories is random (i.e. the order of the categories does not really matter), the location of the absolute middle in the distribution would not tell us anything information about the variable: the “content” of the middle would change completely if we changed the order of the categories. Let us take gender (which is also on a nominal scale) as another example: it would not make any sense to give the mean or median of gender. For some ordinal variables, however, the median is *sometimes* used. For example, if we have five categories of occupational class, which can be ranked from lower class to upper class, it may be interesting to give the value of the median (for example, in this case, the median could be lower non-manuals which would tell us something about the distribution of values).



Why it is important to consider the distribution of the variable for continuous variables before we decide on the type of central tendency? If we take a look at the figures below, we can draw the following conclusions: if we have a perfectly normally distributed variable, the mean, median and mode would all be the same. However, if the distribution is skewed, the median would be a better description of the location of the middle in the distribution.



### 3.4 Measures of variation

Besides the mean, the median and the mode, we may use some measures of variation to describe our variables further. Here are some of the most common measures of variation:

Measure	Definition
<b>Min</b>	The lowest value
<b>Max</b>	The highest value
<b>Range</b>	The difference between the lowest and highest value
<b>Standard deviation</b>	The dispersion of values from the mean

These measures are most suitable for continuous variables (i.e. ratio or interval) but sometimes min, max, and range are used for ordinal variables as well. However, they cannot be used for nominal variables (for the same reason as why we do not use mean or median to describe nominal variables).

## 4. PRODUCING DESCRIPTIVE STATISTICS IN SPSS

This part of the guide will describe how to use SPSS to produce the various tables and graphs as well as measures of central tendency and variation previously discussed.

Outline
4.1 Descriptives
4.2 Frequency table
4.3 Cross table
4.4 Bar chart
4.5 Pie chart
4.6 Histogram
4.7 Scatterplot
4.8 Edit graphs

## 4.1 Descriptives

Quick facts	
<b>Number of variables</b>	At least one
<b>Scale of variable(s)</b>	Continuous (ratio/interval) or ordinal

The **Descriptives** function is used primarily for continuous variables (i.e. ratio/interval) but could also be used for some ordinal variables that are approximately continuous (e.g. rating measures). The SPSS function allows you to order the following statistics:

Types of statistic	
<b>Mean</b>	Mean value
<b>Sum</b>	Sum of all values
<b>Stddev</b>	Standard deviation
<b>Min</b>	Minimum (smallest) observed value
<b>Max</b>	Maximum (largest) observed value
<b>Variance</b>	Variance
<b>Range</b>	The difference between the minimum value and the maximum value
<b>Semean</b>	Standard error of the mean
<b>Kurtosis</b>	Kurtosis and standard error of kurtosis
<b>Skewness</b>	Skewness and standard error of skewness

## *Descriptives: Function*

1. Go to the **Menu bar**, choose **Analyze\Descriptive Statistics\Descriptives**.
2. This will open up a new window called **Descriptives**.
3. A small window will open, where you see two boxes.
4. In the left box, all your variables are displayed. Here you choose the variable(s) you want to get the measures of central tendency and/or variation for.  
In other words, if you want to, you can choose several variables here, and SPSS will produce descriptives for all of them.
5. Click on the small arrow between the boxes to transfer the variable(s) to the box to the right.
6. Click on **Options**.
7. Tick the boxes for the measures you want to have.
8. Click on **Continue** to close the small window.
9. Click on **OK** to close the **Descriptives** window in order to get the results in your **Output** window.

## *Descriptives: Syntax*

```
DESCRIPTIVES VARIABLES=VARNAME  
/STATISTICS=MEASURES.
```

**VARNAME**

Insert the name of the variable you want to use.

**MEASURES**

Insert the measures you want to produce.

For example:

**MEAN SUM MIN MAX RANGE STDDEV**

## Descriptives: Example

(Data: SPSS\_data1.sav)

```
DESCRIPTIVES VARIABLES=age  
/STATISTICS=MEAN MIN MAX RANGE STDDEV.
```

age

Age in years (Min=18; Max=79)

## Output

In the table called **Descriptive Statistics**, all statistics you ordered will be displayed – one type of statistic per column.

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Age in years	5000	61	18	79	46,94	17,703
Valid N (listwise)	5000					



## 4.2 Frequency table

Quick facts	
<b>Number of variables</b>	At least one (one table will be produced for each variable)
<b>Scale of variable(s)</b>	Categorical (nominal/ordinal)

This function is used primarily for categorical variables (i.e. nominal/ordinal) but can be used for any type of variable; the main concern is that the table becomes too lengthy if there are many categories/values in the variable. The **Frequencies** function does not only allow us to create a frequency table, it is also possible to produce a variety of statistics.

Types of statistic	
<b>Frequency</b>	The absolute number of observations within each category
<b>Percent</b>	The percentage of observations within each category (incl. missing)
<b>Valid Percent</b>	The percentage of observations within each category (excl. missing)
<b>Cumulative Percent</b>	Accumulated percentage across categories
<b>Mean</b>	The average value
<b>Median</b>	The value in the absolute middle
<b>Mode</b>	The most frequently occurring value
<b>Sum</b>	Sum of all values
<b>Std. deviation</b>	Standard deviation
<b>Variance</b>	Variance
<b>Range</b>	The difference between the minimum value and the maximum value
<b>Minimum</b>	Minimum (smallest) observed value
<b>Maximum</b>	Maximum (largest) observed value
<b>S.E. mean</b>	Standard error of the mean
<b>Skewness</b>	Skewness and standard error of skewness
<b>Kurtosis</b>	Kurtosis and standard error of kurtosis
<b>Quartiles</b>	Cut-off values for four groups
<b>Cut points</b>	Cut-off values for a selected number of groups
<b>Percentiles</b>	Selected cut-off values for percentiles

Besides the frequency table itself and the types of statistic listed above, the **Frequencies** function makes it possible to order some specific graphs. The default is that no graph is produced, but you may change this to include a bar chart, a pie chart, or a histogram. If you choose a histogram, you may also add a normal curve. If you do order a graph, it is recommended that you go with the option of showing percentages rather than frequencies.

Graphs	
<b>Bar chart</b>	See Section 4.4 for detailed information
<b>Pie chart</b>	See Section 4.5 for detailed information
<b>Histogram</b>	See Section 4.6 for detailed information

## *Frequency table: Function*

1. Go to the **Menu bar**, choose **Analyze\Descriptive Statistics\Frequencies**.
2. This will open up a new window called **Frequencies**.
3. A small window will open, where you see two boxes. In the left box, all your variables are displayed.
4. In the left box, you choose the variable(s) you want to get a frequency table for. In other words, you may choose several variables here, and SPSS will produce frequency tables for all of them.
5. Click on the small arrow between the boxes to transfer the variable(s) to the box to the right.
6. To order statistics for your variable, click on **Statistics**. In the new window called **Frequencies: Statistics** tick the boxes for the measures you want SPSS to show. Click on **Continue**.
7. To order a graph, click on **Charts**. In the new window called **Frequencies: Charts** tick the box for the graph you want SPSS to show. Also make sure to tick the box **Percentages**. Click on **Continue**.
8. Click on **OK**.

## Frequency table: Syntax

```
FREQUENCIES VARIABLES=VARNAME  
/STATISTICS=MEASURES  
/CHARTNAME SPECIFICATION  
/ORDER=ANALYSIS.
```

### VARNAME

Insert the name of the variable you want to use.

### MEASURES

If you want to produce statistics, insert the types of statistic here.

For example:

MEAN

or

MEAN MEDIAN STDDEV

### CHARTNAME

If you want to produce a graph, insert the name of the graph here.

For example:

BARCHART

PIECHART

HISTOGRAM

### SPECIFICATION

For bar charts and pie charts, specify if you want to display frequencies or percentages on the y-axis (i.e. vertical axis).

For example:

FREQ

PERCENT

For histograms, specify if you want to add a normal curve:

NORMAL

## Frequency table: Example

(Data: SPSS\_data1.sav)

```
FREQUENCIES VARIABLES=hometype
```

```
/STATISTICS=MODE MEDIAN
```

```
/ORDER=ANALYSIS.
```

**hometype**

Building type (1=Single family; 2=Multiple family; 3=Townhouse;  
4=Mobile home)

### Output/Step 1

The table called **Statistics** gives a summary of the variable. SPSS automatically shows the number of valid values as well as the number of missing values. If you ordered any additional statistics, they will be displayed in this table as well.

N	Valid	5000
	Missing	0
Median		2,00
Mode		1

### Output/Step 2

The next table is the actual frequency table. **Missing System** provides information about the number of individuals with missing information. **Frequency** gives the number of individuals in each category of the variable. **Percent** shows the percentage of individuals in each category of the variable, *including* missing. **Valid Percent** gives the percentage of individuals in each category of the variable, *excluding* missing (this is the type of percentage we most often report). **Cumulative Percent** adds up the percentages from the first category to the second, from the second to the third, and so on.

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid Single family	2246	44,9	44,9	44,9
Multiple family	1554	31,1	31,1	76,0
Townhouse	908	18,2	18,2	94,2
Mobile home	292	5,8	5,8	100,0
Total	5000	100,0	100,0	

### 4.3 Cross table

Quick facts	
<b>Number of variables</b>	Two (it is possible to split/panel the table by a third variable)
<b>Scale of variable(s)</b>	Categorical (nominal/ordinal)

This function is used primarily for categorical variables (i.e. nominal/ordinal) but can be used for any type of variable; the main concern is that the table becomes too complex and difficult to interpret if there are many categories/values in the variables used. Moreover, it is possible to add a chi-square to the cross table (for more information about chi-square, see Chapter 8).

## *Cross table: Function*

1. Go to the **Menu bar**, choose **Analyze\Descriptive Statistics\Crosstabs**.
2. A small window will open, where you see one big box and three small boxes. In the left box, all your variables are displayed.
3. Here, you choose two variables: one to be the **Row** variable, and one to be the **Column** variable. It is your choice which variable is row and which is column, but it is recommended that you make that choice depending on how you want to interpret your crosstable. If you are unsure, try both ways to see which the better choice is.
4. Move your variables to the **Row** and **Column** boxes by using the arrows.
5. If you stop here and press **OK**, you will only have a crosstable that displays frequencies. In most cases, you also want to see percentages.
6. To do this, click on **Cells**. Tick the boxes for **Row** and/or **Column**.
7. Click on **Continue**.
8. Click on **OK** to close the **Crosstabs** window in order to get the results in your **Output** window.



## Cross table: Syntax

```
CROSSTABS  
  /TABLES=VARNAME1 BY VARNAME2  
  /FORMAT=AVALUE TABLES  
  /CELLS=COUNT PERCENTAGES  
  /COUNT ROUND CELL.
```

**VARNAME1**                      Insert the name of the first variable you want to use.  
                                    This variable will be chosen for rows.

**VARNAME2**                      Insert the name of the second variable you want to use.  
                                    This variable will be chosen for columns.

**PERCENTAGES**                    Specify which type/types of percentages you want to see.  
                                    For example:  
                                    **ROW COLUMN TOTAL**

## Cross table: Example

(Data: SPSS\_data1.sav)

CROSSTABS

/TABLES=**union** BY **gender**

/FORMAT=AVALUE TABLES

/CELLS=COUNT ROW COLUMN

/COUNT ROUND CELL.

**gender**

Gender (0=Man; 1=Woman)

**union**

Union member (0=No; 1=Yes)

**Output**

This is a cross table of the variables **union** and **gender**. **Count** allows us to count the number of times that a particular combination of the two variables occurs. The **% within Union member** shows the distribution of **gender** among those who are not a union member as well as the distribution of **gender** among those who are a union member (each row adds up to 100 %). The **% within Gender** shows the distribution of **union** among men as well as the distribution of **union** among women (each column adds up to 100 %).

			Gender		Total
			Male	Female	
Union member	No	Count	2085	2168	4253
		% within Union member	49,0%	51,0%	100,0%
		% within Gender	85,1%	85,0%	85,1%
	Yes	Count	364	383	747
		% within Union member	48,7%	51,3%	100,0%
		% within Gender	14,9%	15,0%	14,9%
Total	Count	2449	2551	5000	
	% within Union member	49,0%	51,0%	100,0%	
	% within Gender	100,0%	100,0%	100,0%	

## 4.4 Bar chart

Quick facts	
<b>Number of variables</b>	Simple bar chart: one (it is possible to split/panel the chart by a second variable) Clustered bar chart: two (it is possible to split/panel the chart by a third variable)
<b>Scale of variable(s)</b>	Categorical (nominal/ordinal)

The bar chart is useful primarily for categorical variables (i.e. nominal/ordinal) but can be used for any type of variable as long as there are not too many values for each variable. There are two useful types of bar chart: the simple bar chart and the clustered bar chart. The simple bar chart allows for an illustration of one variable, but it is possible to split it by a second variable. The clustered bar chart uses two variables and shows how the values of these two variables cluster together (here as well, it is possible to split the chart by another variable).

### **Simple bar chart**

1. Go to the **Menu bar**, choose **Graphs\Legacy Dialogs\Bar**.
2. A small window will open, where you click on **Define**.
3. A new window called **Define Simple Bar: Summaries for Groups of Cases** will open.
4. Tick the option **% of cases**.
5. In the left box, all your variables are displayed. Here, you select the variable you want SPSS to show a bar chart for. Click on the arrow next to **Category Axis**.
6. Click on **OK** to close the **Define Simple Bar: Summaries for Groups of Cases** window in order to get the results in your **Output** window.

### **Clustered bar chart**

1. Go to the **Menu bar**, choose **Graphs\Legacy Dialogs\Bar**.
2. A small window will open, where you choose the option **Clustered**.
3. Click on **Define**.
4. A new window called **Define Clustered Bar: Summaries for Groups of Cases** will open.
5. Tick the option **% of cases**.
6. In the left box, all your variables are displayed. Here, you select the variable you want SPSS to show a bar chart for. Click on the arrow next to **Category Axis**.
7. In the left box, you select the variable you want the previous variable to be clustered by. Click on the arrow next to **Define Clusters by**.
8. Click on **OK** to close the **Define Simple Bar: Summaries for Groups of Cases** window in order to get the results in your **Output** window.

## Bar chart: Syntax

### Simple bar chart

```
GRAPH  
/BAR(SIMPLE)=PCT BY VARNAME.
```

### Clustered bar chart

```
GRAPH  
/BAR(GROUPED)=COUNT BY VARNAME1 BY VARNAME2.
```

**VARNAME**                      Insert the name of the variable you want to use.

**VARNAME1**                    Insert the name of the variable you want to use as the main variable.

**VARNAME2**                    Insert the name of the variable you want to use as your grouping/clustering variable

*Bar chart: Example (simple bar chart)*

(Data: SPSS\_data1.sav)

GRAPH

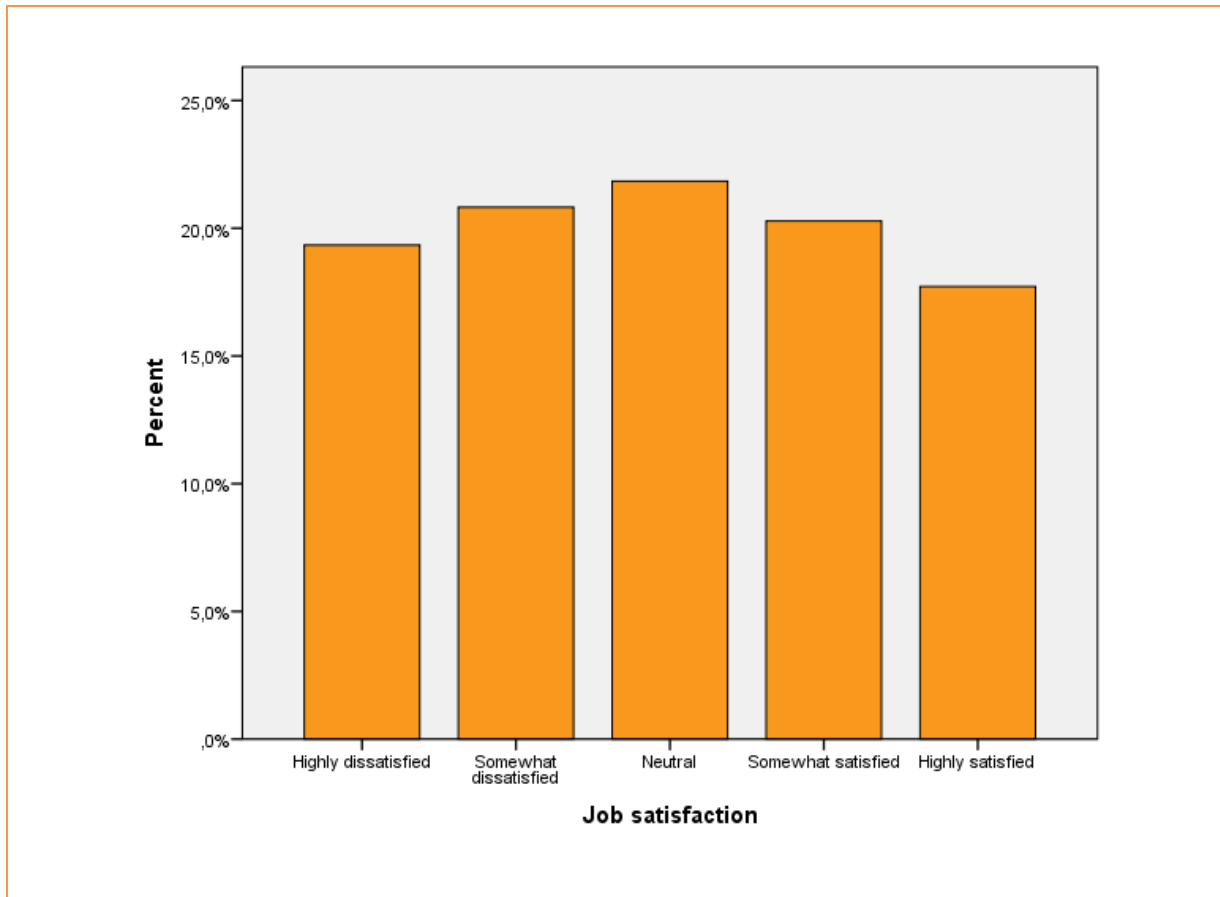
/BAR(SIMPLE)=PCT BY **jobsat**.

**jobsat**

Job satisfaction (1=Highly dissatisfied; 2=Somehwat dissatisfied;  
3=Neutral; 4=Somewhat satisfied; 5=Highly satisfied)

## Output

This is a bar chart of the variable `jobsat`. The x-axis (horizontal axis) represents the values of the variable. The y-axis (vertical axis) represents the proportion of individuals.



Note: To include the exact percentage for each bar, double-click on the graph in the **Output** window. This opens a new window called **Chart Editor**. Click on **Elements** in this window's **Menu bar** and then click on **Show Data Labels**. A new window called **Properties** is opened; click on **Close**. Finally, close the **Chart Editor** to save the changes to the **Output** window.



*Bar chart: Example (clustered bar chart)*

(Data: SPSS\_data1.sav)

GRAPH

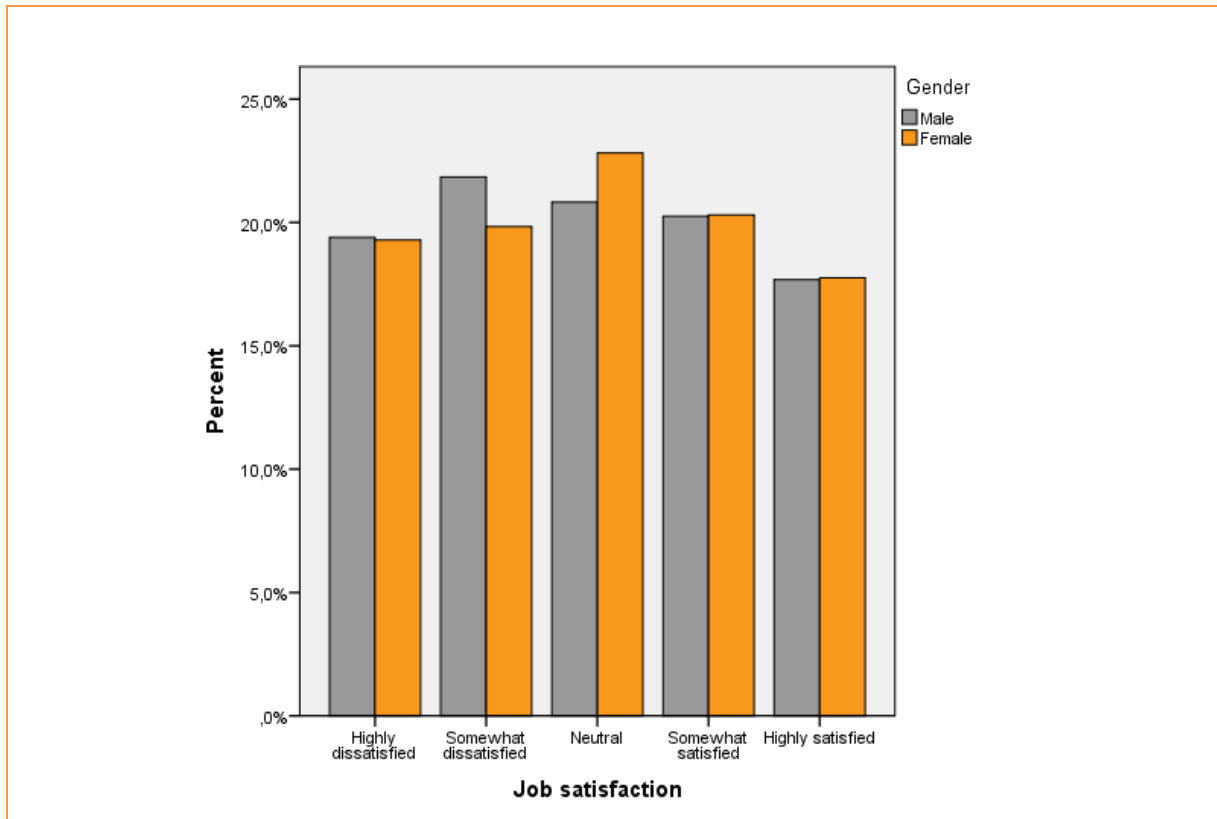
/BAR(GROUPED)=PCT BY **jobsat** BY **gender**.

**jobsat** Job satisfaction (1=Highly dissatisfied; 2=Somehwat dissatisfied; 3=Neutral; 4=Somewhat satisfied; 5=Highly satisfied)

**gender** Gender (0=Man; 1=Woman)

## Output

This is a bar chart of the variable `jobsat`, grouped by the variable `gender`. The x-axis (horizontal axis) represents the values of the first variable (`jobsat`). The bars have different colours depending on the value of the grouping variable (`gender`). The y-axis (vertical axis) represents the proportion of individuals.



Note: To include the exact percentage for each bar, double-click on the graph in the **Output** window. This opens a new window called **Chart Editor**. Click on **Elements** in this window's **Menu bar** and then click on **Show Data Labels**. A new window called **Properties** is opened; click on **Close**. Finally, close the **Chart Editor** to save the changes to the **Output** window.

## 4.5 Pie chart

Quick facts	
<b>Number of variables</b>	One (it is possible to split/panel the chart by a second variable)
<b>Scale of variable(s)</b>	Categorical (nominal/ordinal)

This function is used only for categorical variables (i.e. nominal/ordinal) with relatively few categories – otherwise the pie chart will get too complex.

### *Pie chart: Function*

1. Go to the **Menu bar**, choose **Graphs\Legacy Dialogs\Pie**.
2. A small window will open, where you click on **Define**.
3. A new window called **Define Pie: Summaries for Groups of Cases** will open.
4. Tick the option **% of cases**.
5. In the left box, all your variables are displayed. Here, you select the variable you want SPSS to show a pie chart for. Click on the arrow next to **Define slices by**.
6. Click on **OK** to close the **Define Pie: Summaries for Groups of Cases** window in order to get the results in your **Output** window.

## *Pie chart: Syntax*

```
GRAPH
```

```
/PIE=PCT BY VARNAME.
```

**VARNAME**

Insert the name of the variable you want to use.

## *Pie chart: Example*

(Data: SPSS\_data1.sav)

GRAPH

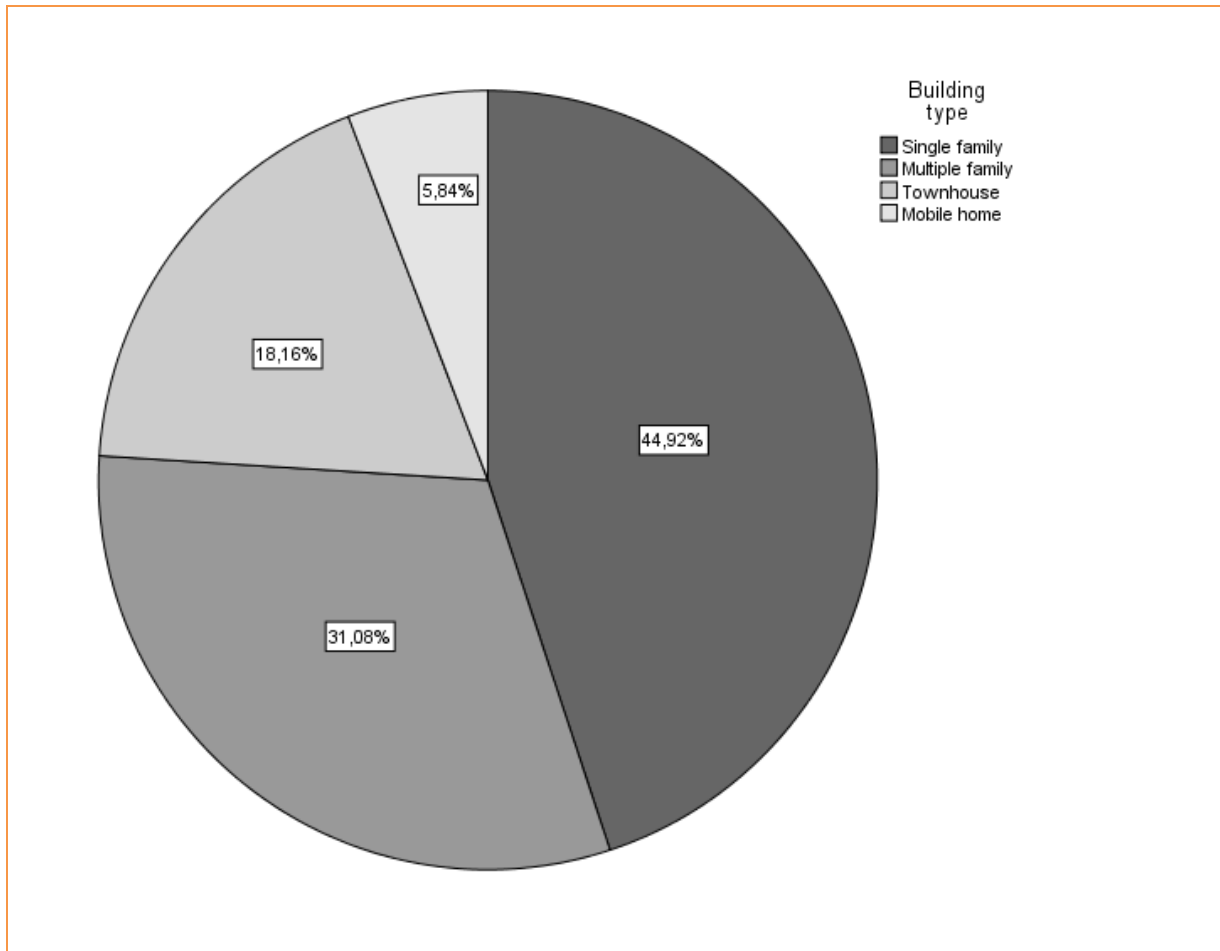
/PIE=PCT BY **hometype**.

**hometype**

Building type (1=Single family; 2=Multiple family; 3=Townhouse;  
4=Mobile home)

## Output

This is a pie chart of the variable **hometype**. It is a circular diagram, divided into slices, where each slice represents the proportion of a specific value of the variable.



Note: To include the exact percentage for each slice, double-click on the graph in the **Output** window. This opens a new window called **Chart Editor**. Click on **Elements** in this window's **Menu bar** and then click on **Show Data Labels**. A new window called **Properties** is opened; click on **Close**. Finally, close the **Chart Editor** to save the changes to the **Output** window.

## 4.6 Histogram

### Quick facts

<b>Number of variables</b>	One (it is possible to split/panel the chart by a second variable)
<b>Scale of variable(s)</b>	Continuous (ratio/interval)

This function is used to illustrate the distribution of continuous variables (i.e. ratio/interval). It is possible to include a normal curve in the chart in order to see how the data adheres to a normal distribution.



## *Histogram: Function*

1. Go to the **Menu bar**, choose **Graphs\Legacy Dialogs\Histogram**.
2. A new window called **Histogram** will open.
3. In the left box, all your variables are displayed. Here, you select the variable you want SPSS to show a histogram for. Click on the arrow next to **Variable**.
4. Tick the option **Display normal curve** to include a normal curve fitted to the data.
5. Click on **OK** to close the **Histogram** window in order to get the results in your **Output** window.

## *Histogram: Syntax*

```
GRAPH  
/HISTOGRAM(NORMAL)=VARNAME.
```

**VARNAME**

Insert the name of the variable you want to use.

## *Histogram: Example*

(Data: SPSS\_data1.sav)

GRAPH

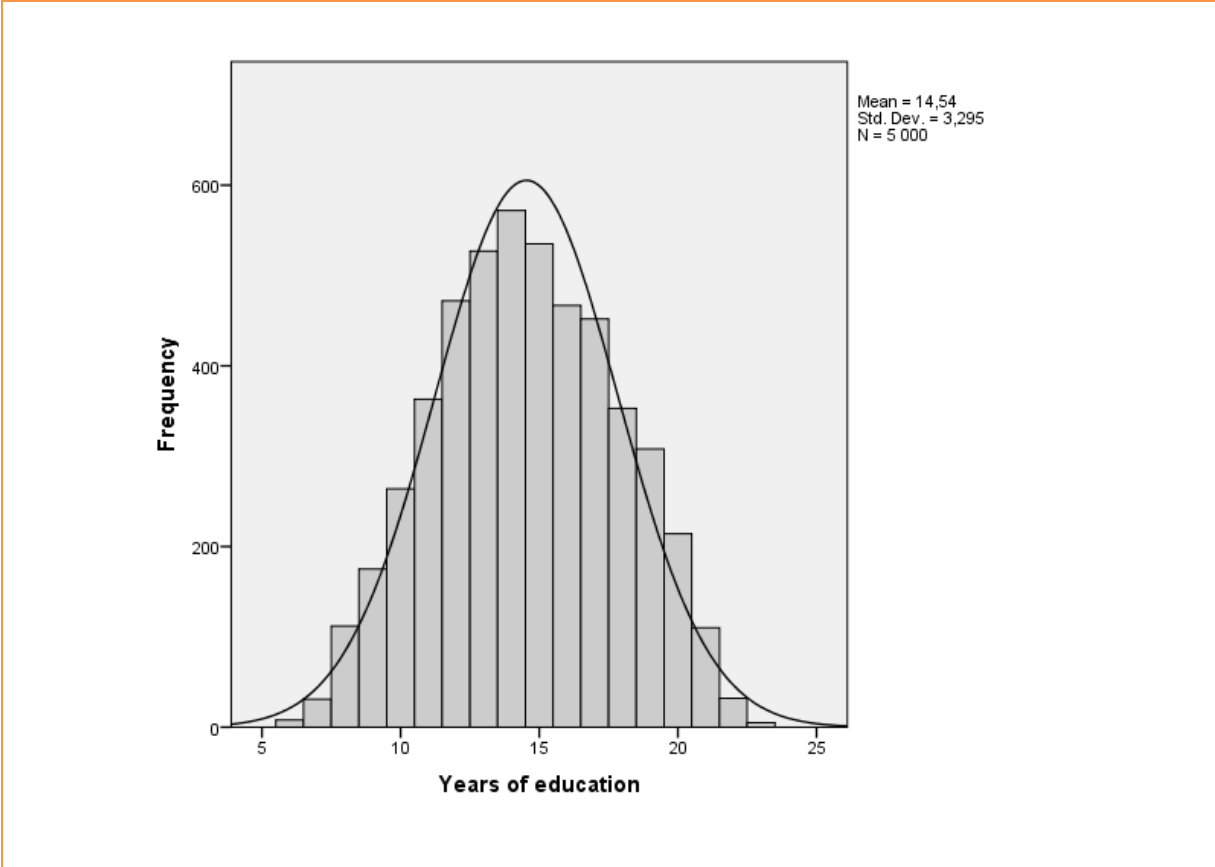
/HISTOGRAM(NORMAL)=ed.

ed

Years of education (Min=6; Max=23)

**Output**

This is a histogram of the variable **ed**. The x-axis (horizontal axis) represents the values of the variable. The y-axis (vertical axis) represents the number of individuals. The black line displays the normal curve. SPSS automatically gives you the mean, standard deviation and the total number of cases for the variable.

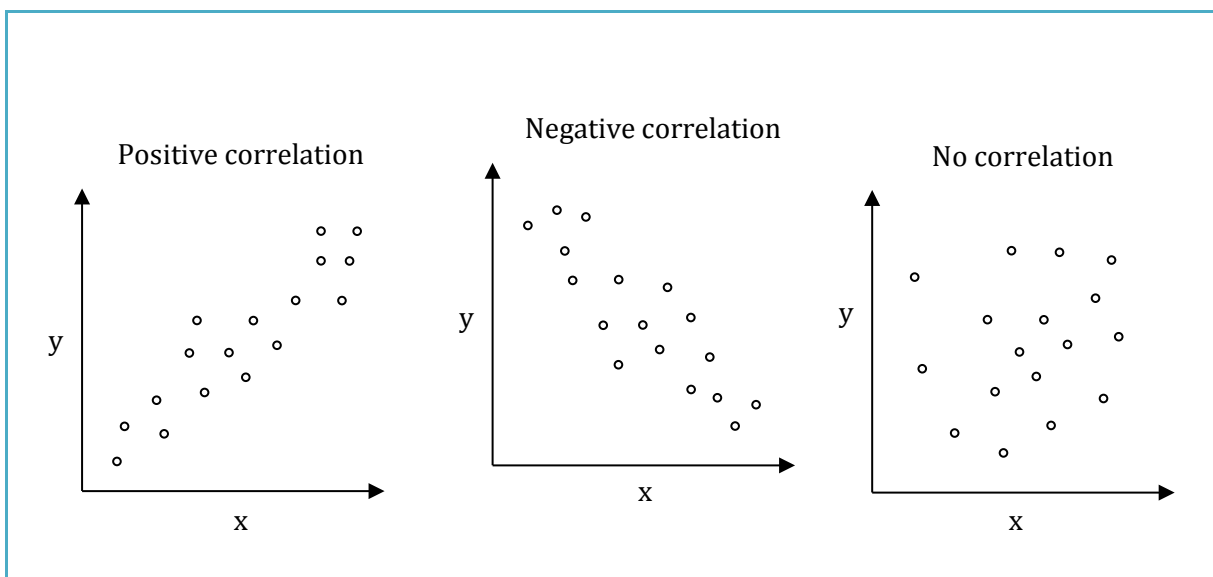


## 4.7 Scatterplot

### Quick facts

<b>Number of variables</b>	Two (it is possible to split/panel the chart by a third variable)
<b>Scale of variable(s)</b>	Continuous (ratio/interval)

This function is used to illustrate how two continuous variables co-vary – or “correlate” – in their pattern of values (see Chapter 9 for more detailed information about correlation). If increasing values of one variable correspond to increasing values of another variable, it is called a positive correlation. If increasing values of one variable correspond to decreasing values of another variable, we have a negative correlation. In the graph below, different types of correlation are presented. The letter “x” stands for x-axis (horizontal axis) and the letter “y” stands for y-axis (vertical axis).



## *Scatterplot: Function*

1. Go to the **Menu bar**, choose **Graphs\Legacy Dialogs\Scatter/Dot**.
2. A small window will open, where you click on **Define**.
3. A new window called **Simple Scatterplot** will open.
4. In the left box, all your variables are displayed. Move the variables you want to use to the **Y Axis** and **X Axis** by highlighting them and clicking on the arrow next to the axis you want them on.
5. Click on **OK** to close the **Simple Scatterplot** window in order to get the results in your **Output** window.

## Scatterplot: Syntax

```
GRAPH  
/SCATTERPLOT(BIVAR)=XVAR WITH YVAR  
/MISSING=LISTWISE.
```

**XVAR**                      Insert the name of the first variable you want to use.  
This variable will be chosen for the x-axis.

**YVAR**                      Insert the name of the second variable you want to use.  
This variable will be chosen for the y-axis.

## Scatterplot: Example

(Data: SPSS\_data1.sav)

GRAPH

/SCATTERPLOT(BIVAR)=age WITH pets

/MISSING=LISTWISE.

age

Age in years (Min=18; Max=79)

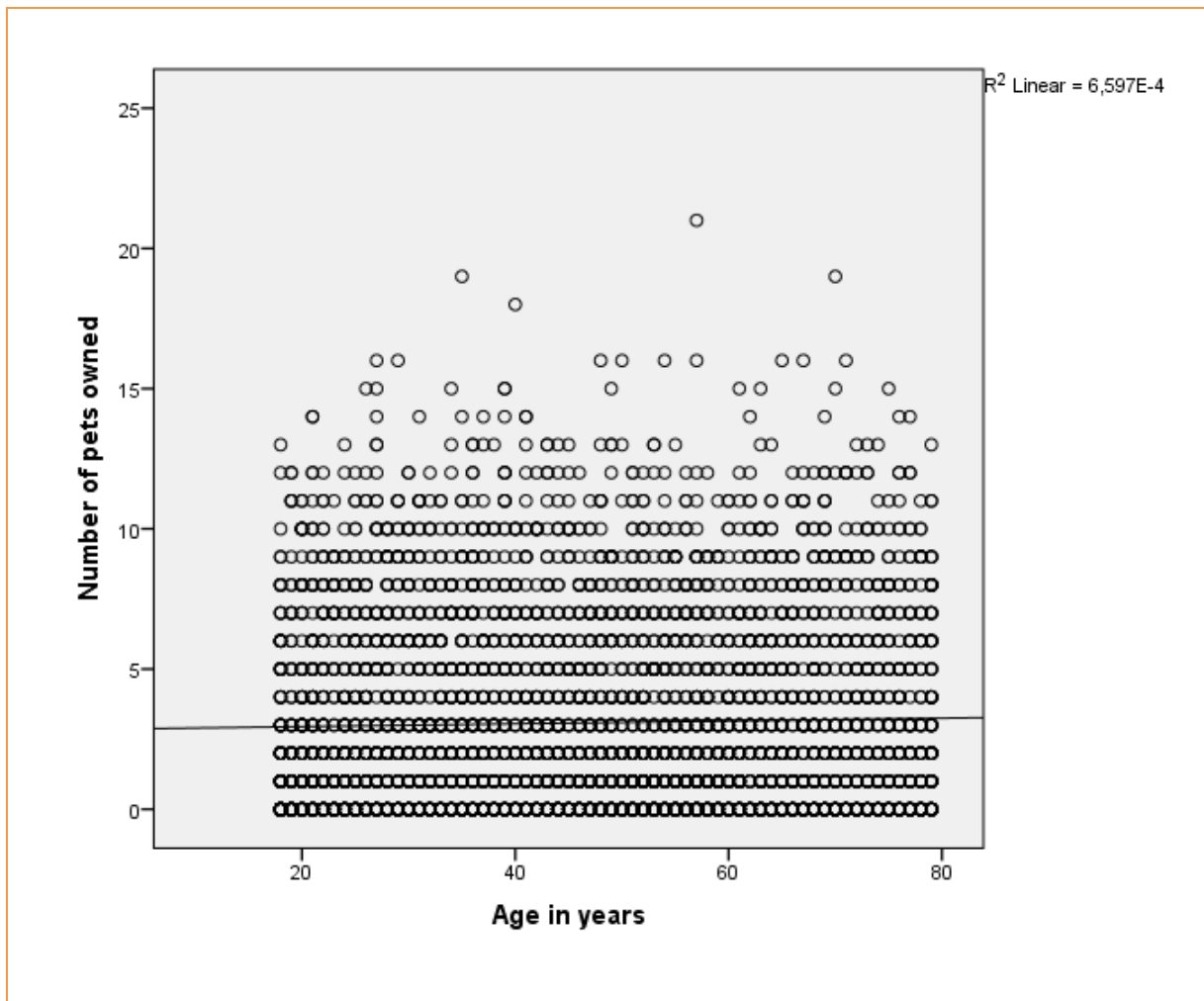
pets

Number of pets owned (Min=0; Max=21)



## Output

This is a scatterplot of the variables **age** and **pets**. The x-axis (horizontal axis) represents the values of the variable **age**. The y-axis (vertical axis) represents the values of the variable **pets**. Each dot represents one individual. As can be seen from this scatterplot, these two variables do not co-vary. If they would have co-varied, we would typically have seen an oval shape (tilted to the left or to the right), where low values for **age** would correspond to either low or high values for **pets** (and vice versa).



Note: To include a reference line, double-click on the graph in the **Output** window. This opens a new window called **Chart Editor**. Click on **Elements** in this window's **Menu bar** and then click on **Fit Line at Total**. A new window called **Properties** is opened; click on **Close**. Finally, close the **Chart Editor** to save the changes to the **Output** window.

## 4.8 Edit graphs

As already touched upon, it is possible to modify various elements of the graphs you produce in SPSS. In the **Output** window, just double-click on a graph of choice. This will open the **Chart Editor**. For example, the **Chart Editor** allows you to change the size and colours of the graph, as well as to add data labels (e.g. percentages) and reference lines. Try it out!

# 5. OTHER USEFUL FEATURES IN SPSS

This part of the guide will describe other features in SPSS that you may need to use when you work with your data material.

<b>5.1</b> Recode
<b>5.2</b> Compute
<b>5.3</b> Select cases
<b>5.4</b> If
<b>5.5</b> Weight cases

## 5.1 Recode

The recoding command is useful in a lot of different situations. For example: if you have a variable that is continuous and you want to create categories, if you have a categorical variable where you want to combine categories, if you want to change a binary variable (i.e. a variable with only two values, such as gender) to have the opposite coding, or if you want to change any value(s) into missing.

## Recode: Function

Imagine that we have a continuous variable with information about age that we want to change into age groups.

1. Go to the **Menu bar**, choose **Transform\Recode into different variables**.
2. A new window called **Recode into Different Variables** will open.
3. In the left box, all your variables are displayed. You choose the variable that you want to recode and use the arrow to move it to the right box.
4. Below **Output variable**, you specify the **Name** of the new variable.  
If the old variable was called **age**, the new could be named **agecat**.
5. Also specify the **Label**, for example: "Age category".
6. Click on **Change**.
7. Click on **Old and New Values**, whereby a new window called **Recode into Different Variables: Old and New Values** will open.
8. Here, the basic principle is very easy: you let SPSS know what the old values are, and then what you want the new values to be.
9. In the example of **age** (that ranges from 18 to 79), we choose the following age groups: 18-24; 25-34; 35-49; 50-64; >65 (65 or older).  
Below **Range** you write "18" through "24" and next to **Value** (below **New Value**) you write "1". Then click on **Add**.  
Below **Range** you write "25" through "34" and next to **Value** (below **New Value**) you write "2". Then click on **Add**.  
Below **Range** you write "35" through "49" and next to **Value** (below **New Value**) you write "3". Then click on **Add**.  
Below **Range** you write "50" through "64" and next to **Value** (below **New Value**) you write "4". Then click on **Add**.  
Below **Range** you write "65" through "79" and next to **Value** (below **New Value**) you write "5". Then click on **Add**.
10. Click on **Continue** and then **OK**.
11. Now you will have a new categorical variable based on your continuous variable.

## Recode: Syntax

```
RECODE VARNAME (OLD AND NEW VALUES) INTO VARNAME_NEW.  
EXECUTE.
```

**VARNAME**                      Insert the name of the old variable.

**VARNAME\_NEW**                Insert the name you want for the new variable.

**(OLD AND NEW VALUES)**    Specify how you want to transform the values.

Some examples:

(1 thru 3=1) (4 thru 6=2) (7=3)

or

(LOWEST thru 10=0) (11 thru HIGHEST=1)

or

(0=1) (1=2) (2=3) (3=4) (ELSE=SYSMIS)

or

(99=SYSMIS) (ELSE=COPY)

## *Recode: Example*

(Data: SPSS\_data1.sav)

```
RECODE age (18 thru 24=1) (25 thru 34=2) (35 thru 49=3) (50 thru 64=4) (65 thru 79=5)  
INTO agecat.  
EXECUTE.
```

age

Age in years (Min=18; Max=79)

## 5.2 Compute

Compute is another very useful command that works just like an ordinary calculator. For example, you can use it when you want to add, subtract, multiply or divide the values of one or more variables.



## Compute: Function

Imagine that we have one variable indicating how many saltwater fish people own (`pets_saltfish`) and one variable that indicates how many freshwater fish people own (`pets_freshfish`). We want to create a variable indicating how many fish people own in total (`pets_fish`), i.e. the sum of these two fish variables.

1. Go to the **Menu bar**, choose **Transform\Compute Variable**.
2. A new window called **Compute Variable** will open.
3. In the left box (below **Target Variable**), you chose the name of the new variable, for example `pets_fish`.
4. In the right box (below **Numeric Expression**), you write your formula. In this case, you would write "`pets_saltfish+pets_freshfish`".
5. Click on **OK**.

## Compute: Syntax

```
COMPUTE VARNAME_NEW=EXPRESSION.  
EXECUTE.
```

**VARNAME\_NEW**

Insert the name you want for the new variable.

**EXPRESSION**

State how the new variable should be created.

Some examples:

`(income_year1+income_year2+income_year3)/3`

or

`2005-age`

or

`height_cm*0,01`

## Compute: Example

(Data: SPSS\_data1.sav)

```
COMPUTE pets_fish=pets_saltfish+pets_freshfish.  
EXECUTE.
```

**pets\_saltfish**

Number of saltwater fish owned (Min=0; Max=8)

**pets\_freshfish**

Number of freshwater fish owned (Min=0) (Max=16)

### 5.3 Select cases

Sometimes you may want to select only a specific part of the data set. For example, you may want to get descriptive statistics for men and women separately, or for a certain age range, or for only non-smokers. In that case, we may use a command called **Select Cases**.

## Select cases: Function

As an example, we may take the variable **gender**. Let us pretend we want to limit our descriptive statistics only to women.

1. Go to the **Menu bar**, choose **Data\Select Cases**.
2. A new window called **Select Cases** will open.
3. Tick the box called **If condition is satisfied**.
4. Click on the button called **If**.
5. A new window called **Select Cases: If** will open.
6. In the left box, all your variables are displayed. You choose the variable that you want to select cases for and use the arrow to move it to the right box.
7. In this specific example, we choose the variable **gender** (for which we want to select only women). In the right box, we thus write "**gender=1**" (since women has the value 1 here).
8. Click on **Continue** and then **OK**. Now, only women are selected (and men are temporarily filtered out from the data set).
9. If you want to "re-select" all cases, you carry out the following steps:
10. Go to the **Menu bar**, choose **Data\Select Cases**.
11. A new window called **Select Cases** will open.
12. Tick the box called **All cases**.
13. Click on **OK**.

When you select cases, a new "filter" variable will be created in your data set. If you delete it, the selection will disappear. You can double-check that the selection works by opening up the **Data view** and see that the case numbers for cases you "un-selected" are crossed over.

The selection will be applied to everything you do from the point you select cases and onwards, until you remove the selection. In other words, all tables and graphs will be based only on the selected individuals until you remove (or change) the selection.

## Select cases: Syntax

### Apply selection

USE ALL.

COMPUTE FILTERNAME=(VARNAME=SPECIFICATION).

FILTER BY FILTERNAME.

EXECUTE.

### Remove selection

FILTER OFF.

USE ALL.

EXECUTE.

**VARNAME**

Insert the name of the variable you want to filter on.

**SPECIFICATION**

Specify which value/values of this variable you want to keep.

Some examples:

(good\_health=1)

or

(age>17)

or

(gender=1 & smoking=0)

**FILTERNAME**

Specify what you want to call the new variable that indicates which values to keep.

## Select cases: Example

(Data: SPSS\_data1.sav)

### Apply selection

USE ALL.

COMPUTE **onlywomen**=(**gender**=1).

FILTER BY **onlywomen**.

EXECUTE.

### Remove selection

FILTER OFF.

USE ALL.

EXECUTE.

**gender**

Gender (0=Man; 1=Woman)

## 5.4 If

The **If** command is very similar to the **Recode** command. Just like **Recode**, **If** can be used to create new variables – but unlike **Recode**, you can also use **If** to condition your data. Put differently, it means you can construct a new variable (or change an existing one) given certain properties of one or more other variables.

### Example

Suppose we ask ten individuals a couple of questions about their smoking behaviour. The first question is: “Do you smoke?” (smoke) and the possible responses are: 0=No or 1=Yes. The second question is: “How many packs of cigarettes do you usually smoke per week?” (number\_packs) and the response options are: 1=Less than one pack; 2=1 pack; 3=2 packs; 4=3-4 packs; 5=5-7 packs; 6=8 or more packs. This is what our data looks like:

	smoke	number_packs
1	0	1
2	1	5
3	0	.
4	1	7
5	0	.
6	0	.
7	0	1
8	1	2
9	0	.
10	0	.

The problem we have is that some of the individuals, who answered that they did not smoke in the first question, also gave an answer to the second question (individuals 1 and 7). We want to change their value in number\_packs to missing (missing is illustrated by a “dot”). That is when we use the **If** command. By just writing the following in our syntax, we fix this problem:

```
IF (smoke=0) number_packs=$SYSMIS.  
EXECUTE.
```

	smoke	number_packs
1	0	.
2	1	5
3	0	.
4	1	7
5	0	.
6	0	.
7	0	.
8	1	2
9	0	.
10	0	.



## If: Syntax

```
IF (CONDITION) OUTCOME.  
EXECUTE.
```

**(CONDITION)** Specify how you want to condition your data.

Some examples:

`(education=3)`

or

`(gender=0 & smoke=1)`

or

`(age>17 & age<66)`

or

`SYSMIS(gender)`

**OUTCOME** Specify how you want result to turn out.

Some examples:

`health=$SYSMIS`

or

`working_age=1`

or

`(gender=1 & smoking=0)`

## If: Example

(Data: SPSS\_data1.sav)

We have two variables. The first variable (**marital**) contains information about marital status and the second variable (**spoused**) contains information about how many years of schooling the spouse has. Not all individuals have information for the variable **spoused** – it is primarily missing if the individual does not have a spouse (i.e. unmarried), but in some cases the individual is married but did not provide any information about the spouse's education. So, we want to separate these two types of missing by recoding the missing information in **spoused** to the value -1 upon the condition that **marital** has the value 0 (i.e. unmarried).

```
IF (marital=0) spoused=-1.
```

```
EXECUTE.
```

**marital**

Marital status (0=Unmarried; 1=Married)

**spoused**

Spouse's years of education (Min=0; Max=24)

## 5.5 Weight cases

Optimally, our study sample should be a miniature of the population we are interested in (see Section 2.1). This is seldom the case, due to missing data. Missing data may cause some groups in the study sample to be smaller (i.e. underrepresented) or bigger (i.e. overrepresented) than they actually are. If such problems occur, we cannot draw reliable conclusions from our data. However, we can try to correct for the lack of representativeness by using “weights”. This is a rather common strategy in surveys. If you use an existing survey material, there are usually already designed weight variables in the data set. The guide will not discuss exactly how weights are calculated, just why and how they are used in SPSS.

### Example

Suppose we have a sample of 1,000 adults who have answered a questionnaire about their health. Now we want to collect health data from all children age 10-18 living in these adult persons' household. This gives us a sample of 2,500 children. However, children who live in the same household are likely to experience similar living conditions – they do not adequately represent the population of children age 10-18 – and we want to adjust for that. Thus, we use a weight variable that takes into account this particular sampling of children.

To get “un-weighted” cases, remember to de-activate the **Weight cases** function in SPSS.

## *Weight cases: Syntax*

### **Activate the weight**

WEIGHT BY **VARNAME**.

### **De-activate the weight**

WEIGHT OFF.

**VARNAME**

Insert the name of the weight variable.

## *Weight cases: Example*

(Data: SPSS\_data2.sav)

The four-digit weight variable in the data material we use corrects the sample to reflect national distributions of key variables. We apply the weight before we run any analyses based on the data.

WEIGHT BY `weight_var`.

## 6. STATISTICAL SIGNIFICANCE

Outline
6.1 Hypothesis testing
6.2 P-values
6.3 Confidence intervals
6.4 Discussion

## 6.1 Hypothesis testing

Quantitative research is commonly about examining relationships between variables (see Chapter 11 for a more detailed discussion about those issues). Assuming that all is done correctly, data analysis will give us information about the direction of the relationship (i.e. is the relationship negative or positive) and the effect size (i.e. how strong the relationship is). These are the two most important outcomes of data analysis, but it is not uncommon that research inquiry instead focuses on a third point: statistical significance. Statistical significance can be seen as an indicator of the reliability of the results – although that is important indeed, it is not what exclusively should guide which findings we focus on and which we discard. A fourth issue that needs to be taken into account is whether the findings have any practical or clinical importance – in order words; do they matter? We therefore suggest the following priority list when it comes to how results from data analysis should be interpreted and valued:

Priority list	
<b>1. Direction</b>	Is the relationship positive or negative?
<b>2. Effect size</b>	Is the relationship strong or weak?
<b>3. Statistical significance</b>	Is the relationship reliable?
<b>4. Practical importance</b>	Is the relationship relevant?

### *Hypotheses*

Let us return to the matter of statistical significance: what is it really? Well, for example, if we find that cats are smarter than dogs, we want to know whether this difference is “real”. Hypothesis testing is how we may answer that question. We start by converting the question into two hypotheses:

Hypotheses		
<b>Null hypothesis</b>	( $H_0$ )	There is no difference
<b>Alternative hypothesis</b>	( $H_1$ )	There is a difference

There is no “law” saying that the null hypothesis is always “no difference” and the alternative hypothesis is always “difference”. However, for the null hypothesis, precedence is commonly given to the “simpler” (or more “conservative” or “normative”) hypothesis. Here, it is generally

simpler to claim that there is no difference in intelligence between cats and dogs than to say that there is a difference.

*Outcomes*

There are two possible outcomes of hypothesis testing:

Outcomes of hypothesis testing	
<b>Reject <math>H_0</math> in favour of <math>H_1</math></b>	Suggests that the alternative hypothesis <i>may</i> be true (but it does not prove it)
<b>Do not reject <math>H_0</math></b>	Suggests that there is not sufficient evidence against $H_0$ in favour of $H_1$ (but it does not prove that the null hypothesis is true)

Note that we are never able to decide from hypothesis testing that we should reject or accept  $H_1$ . However, rejecting  $H_0$  may lead us to suggest that  $H_1$  might be accepted.

*Errors*

There are two types of error that may occur in hypothesis testing: a type I error occurs when the null hypothesis is rejected although it is true, whereas a type II error occurs when the null hypothesis is not rejected although it is false. In the example of cats and dogs, a type I error would thus occur if we concluded that there is a difference in the intelligence between cats and dogs although that is not true. A type II error, on the other hand, would occur if we concluded that there is no difference in intelligence when in fact there is.

Type I and type II errors			
		<u>Conclusion</u>	
		Reject $H_0$ in favour of $H_1$	Do not reject $H_0$
<u>“Truth”</u>	$H_0$	<i>Type 1 error</i>	<i>Right decision</i>
	$H_1$	<i>Right decision</i>	<i>Type II error</i>

Type I errors are generally considered to be more serious than type II errors. Type II errors are often due to small sample sizes.



### *Statistical hypothesis testing*

Conducting a statistical hypothesis test is easy to do in statistical software such as SPSS. These tests give us a probability value (p-value) that can help us decide whether or not the null hypothesis should be rejected. See Section 6.2 for a further discussion about the p-value.

## 6.2 P-values

The probability value – or p-value – helps us decide whether or not the null hypothesis should be rejected. There are some common misunderstandings about p-values:

The p-value is <i>not</i> ...
... the probability that the null hypothesis is true
... the probability that the alternative hypothesis is false
... the probability of the occurrence of a type I error (falsely rejecting $H_0$ )
... the probability that replicating the experiment would yield the same conclusion
... the probability that the finding is a “fluke”
... an indicator of the size of the effect or importance of the findings
... determining the significance level

Using the p-value to make this decision, it must first be decided what probability value we find acceptable. This is often referred to “the significance level”. If the p-value is below this level, it means that we can reject the null hypothesis in favour of the alternative hypothesis, and if the p-value is above this level, it means that we cannot reject the null hypothesis. The smaller the p-value, the more convincing is the rejection of the null hypothesis.

### *Significance levels*

The significance level is set by the individual researcher – in that sense, it is quite arbitrary – but there are some levels that are widely used (asterisks are often used to illustrate these levels):

Significance levels		
p<0.05	Statistically significant at the 5 % level	*
p<0.01	Statistically significant at the 1 % level	**
p<0.001	Statistically significant at the 0.1 % level	***

It should be noted that p<0.10 – statistical significance at the 10 % level – is also a commonly used significance level in some fields of research.

Let us return to the example of differences in intelligence between cats and dogs. For instance, if we find a difference in intelligence between these types of animal, and the p-value is below 0.05,

we may thus state that the null hypothesis (i.e. no difference) is rejected at the 5 % significance level. The p-value does not, however, state whether the difference is small or big, or whether cats or dogs represent the smarter type of animal (in order to state such things, one would have to look at the direction and the effect size).

It should be noted that the p-value is affected by the sample size, which means that a smaller sample size often translates to a larger p-value, For example, if you have a data material of 100 individuals, the effect size has to be quite large (e.g. large income differences income between men and women) in order to get small p-values. Conversely, larger sample size makes it easier to find small p-values. For example, if you analyse a data material containing the entire population of a country, even tiny differences are likely to have small p-values. In other words, the size of the sample influences the chances of rejecting the null hypothesis.

## *Practical importance*

As stated earlier in this section, statistical significance – determined by the p-value – is *not* the same as effect size or practical/clinical importance (i.e. if it “matters”). We can use couple of examples to illustrate the differences:

### **Example 1**

A pharmaceutical company has developed a drug to cure obesity. During tests of this drug, it appears as migraine could be one of the side effects of taking this drug. The null hypothesis would here be that there are no differences in the risk of migraine between people who had taken the drug and those who have not. The alternative hypothesis would then be that there are differences. When we run the analysis on this data material, we see that those who have taken the drug have ten times the risk of migraine, but the p-value is above the 5 % level (i.e.  $p > 0.05$ ). Thus, we cannot reject the null hypothesis. The difference is however large and is likely to have significant impact on people’s lives. It could moreover be the case that a type II error has occurred here due to a small sample size.

### **Example 2**

In the second example, researchers have gathered data on coffee consumption and happiness among 100,000 company employees. The null hypothesis would here be that there are no differences in happiness between people who drink coffee and those who do not. The alternative hypothesis would be that there are differences. The analysis suggests that there is a tiny difference in happiness between those who drink coffee and those who do not, to the advantage of the coffee drinkers. The p-value is below 0.05 which suggests that the null hypothesis can be rejected at the 5 % level. However, the difference is very small and the results may not be very useful.

### 6.3 Confidence intervals

Confidence intervals (CI) are closely related to the concept of statistical hypothesis testing, but they are more informative than p-values since they do not only suggest whether we should reject  $H_0$  or not, they also provide the range of plausible values.

#### *The “unknown population parameter”*

Before we get into the discussion about confidence intervals, we need to address the concept of “unknown population parameter”. A parameter tells us something about a population (while a “statistic” tells us something about a sample). The population parameter is thus basically a measure of any given population. Examples of population parameters are: the mean height of Swedish men, the average intelligence score in 12-year olds, or the mean number of children among highly educated people. The parameter is a fixed value; it does not vary. We seldom have information about the entire population, generally only for a part of it (i.e. a sample). In that case, the population parameter is unknown. Simply put, a confidence interval is a range that includes the “unknown population parameter”.

#### *Limits and levels*

The interval has an upper and a lower bound (i.e. confidence limits). Similar to p-values, confidence intervals have “confidence levels” that indicate how certain we can be that the interval includes the true population parameter. Confidence intervals are typically stated at the 5 % level. A 95 % confidence level would thus mean that if we replicated a certain analysis in 100 samples of the population, we would expect that 95 % of the intervals would include the true population parameter. Thus, strictly speaking, it is *not* correct to say that “with 95 % probability, the true population parameter lies within this interval” (because the parameter either is or is not within the interval).

#### *Confidence and precision*

When discussing confidence intervals, it is important to be aware of the tension between precision and certainty: better precision means being less confident, whereas more confidence means less precision. As previously stated, confidence is reflected by the confidence level we choose; logically, a higher confidence level means more confidence. The higher the confidence

level we choose, the wider the interval gets – and the wider the interval is, the less the precision we get.

### Confidence versus precision

Higher confidence level = wider confidence interval = less precision

Lower confidence level = slimmer confidence interval = more precision

However, it is important to know that the width of the confidence interval is also affected by the sample size: the larger the sample size, the slimmer the interval (which means better precision).

Let us take an example to sum up what has been said about confidence intervals so far: We have gathered data on all sociology students at Stockholm University and find that their mean age is 26 years. Instead of highlighting this relatively non-informative fact, we can calculate the confidence interval (at the 5 % level). In this case, it is 22-30. Therefore we could make the more informative statement that: “with 95 % confidence, we conclude that the mean age of sociology students is 22 to 30 years”.

The most common application for confidence intervals as a way of significance testing is when we are interested in the difference between two samples. For example: the difference in the mean income between men and women, or the difference in the percentage of daily smokers among individuals with a lower level of education versus those with a higher level of education. In this case, we may look at the “overlap” between the confidence intervals estimated for each sample. Suppose that we have an upcoming election and just got the results from the latest poll. There are two parties in the race: the green party and the yellow party. The results from the poll show that the green party got 42 % of the votes and the confidence interval is 40-44 (at the 5 % level). The yellow party got 58 % of the votes and the confidence interval is 54-62 (at the 5 % level). What does this tell us? First of all, we can conclude that the yellow party has a greater share of votes. Looking at the two confidence intervals, we see that the intervals do not overlap. Why is that important? Well, remember that *all* values in a confidence interval are plausible. Hence, if the confidence intervals do not overlap, it means that the estimates (in this case: the share of votes) are indeed different given the chosen confidence level (in this case: at the 5 % level). However, it should be emphasized that while non-overlap always mirrors a significant difference, overlap is not always the same as a non-significance difference.

## 6.4 Discussion







Now you are maybe wondering; should I use p-values or confidence intervals? Almost all disciplines would recommend using both because they capture several dimensions. In the following, the advantages and disadvantages of p-values and confidence intervals will be described and discussed.

P-value is an important part of research, most likely the heart of it. The p-value is based on “yes-or no”-questions in which it shows how much evidence we have against the null hypothesis. P-values are much clearer than confidence intervals and it helps the researcher to make quick judgments about his research. Another advantage with the p-value is that it can give the difference from a previous specified statistical level. Unfortunately there are misconceptions about the p-value among researchers and many disciplines rely on them to draw conclusions rather than understanding the background. One of the common mistakes among researchers is that they do not further analyze their data in order to ensure that the p-value is not affected by other factors. Moreover, p-values cannot alone permit any direct statements about the direction or size of difference. In order to make those decisions, one must always look at the confidence intervals.

A confidence interval informs the researcher about the power of the study and whether the data is compatible, it also shows the likelihood of the null hypothesis being true and that in turn tells us how much confidence we have in our findings. The width of the confidence interval indicates the precision of the point estimates, in which a small interval indicates a more precise estimate, while a wide interval indicates a less precise estimate. The precision is related to the sample size and power in which it tells us that the larger sample size we have, the greater, the more precise estimates we have. The intervals are useful when having small sample sizes. Normally, small studies fail to find statistically significant treatments, when including point estimates with wide intervals that include the null value may be consistent and significant. The intervals provide the researcher an understanding of the sample size. This can also be a disadvantage when having large data because it produces statistically significant results even if the difference between the groups is small. Another advantage with the confidence interval is that it can provide means of analysis for studies that seek to describe and explain, rather than make decisions about treatments effects. A disadvantage with the confidence interval is that it captures several elements at the time, in which it may not give precise information like the p-values.

As mentioned, a majority of disciplines recommend including both p-values and confidence intervals because they capture information in different dimensions. Neither p-values nor confidence intervals can prevent biases or other problems but the combination of them provides a more flexible approach and highlights new perspectives on the data. Confidence intervals permit us to draw several conclusions at the same time and they are more informative about sample sizes and point estimates. They are also useful in studies when we have small sample sizes. But they are not as precise as p-values when it comes to accepting and rejecting the null hypothesis. Thus, when we combine them together we can be more certain.

The figure below shows the advantages and disadvantages when interpreting and drawing conclusions with the help of p-values and confidence intervals.

P-values versus confidence intervals		
	P-values	Confidence intervals
Accept/reject		
Degree of support		
Estimate and uncertainty		



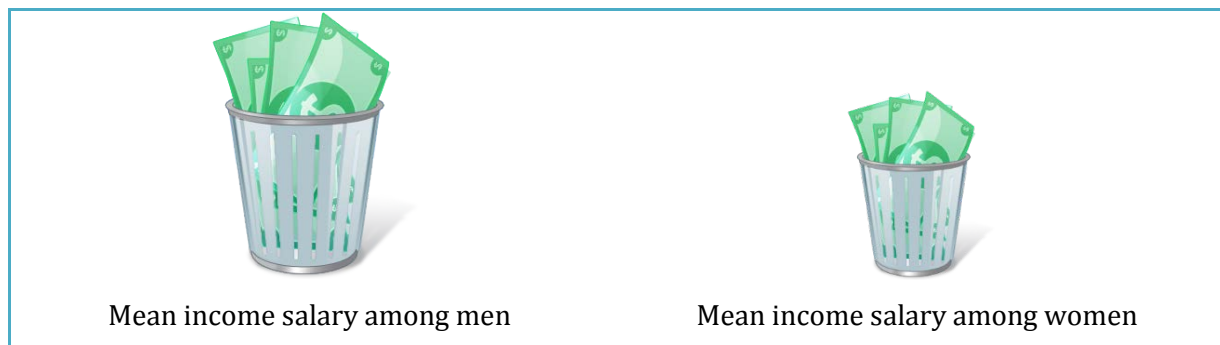
## 7. COMPARE MEANS

Outline
7.1 T-test: independent samples
7.2 T-test: paired samples
7.3 One-way ANOVA

## 7.1 T-test: independent samples

Quick facts	
<b>Number of variables</b>	One independent (x) One dependent (y)
<b>Scale of variable(s)</b>	Independent: categorical with two values (binary) Dependent: continuous (ratio/interval)

The independent samples t-test is a method for comparing the mean of one variable between two (unrelated) groups. For example, you may want to see if the income salary of teachers differs between men and women, or if the score of a cognitive test differs between children who have parents with low versus high education.



## Assumptions

First, you have to check your data to see that the assumptions behind the independent samples t-test hold. If your data “passes” these assumptions, you will have a valid result. However, “real world” data is often a bit complicated, and it is not uncommon that at least one of the assumptions is violated. While you most commonly will be able to conduct the t-test anyway, it is important to be aware of the possible problems.

Checklist	
<b>Continuous dependent variable</b>	Your dependent variable should be continuous (i.e. interval/ratio). For example: Income, height, weight, number of years of schooling, and so on. Although they are not really continuous, it is still very common to use ratings as continuous variables, such as: “How satisfied with your income are you?” (on a scale 1-10) or “To what extent do you agree with the previous statement?” (on a scale 1-5).
<b>Two unrelated categories in the independent variable</b>	Your independent variable should be categorical and consist of only two groups. Unrelated means that the two groups should be mutually excluded: no individual can be in both groups. For example: men vs. women, employed vs. unemployed, low-income earner vs. high-income earner, and so on.
<b>No outliers</b>	An outlier is an extreme (low or high) value. For example, if most individuals have a test score between 40 and 60, but one individual has a score of 96 or another individual has a score of 1, this will distort the test.

### *T-test: independent samples: Function*

1. Go to the **Menu bar**, choose **Analyze\Compare Means\Independent-Samples T Test**.
2. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable and transfer it to the box called **Test Variable(s)**.
3. Then you choose the variable you want as your independent variable and transfer it to the box called **Grouping Variable**.
4. Click on **Define Groups...**
5. Specify which values the two categories in the independent variable have.
6. Click on **Continue**.
7. Click on **OK**.

## *T-test: independent samples: Syntax*

```
T-TEST GROUPS=INDEPVAR(G1 G2)
/MISSING=ANALYSIS
/VARIABLES=DEPVAR
/CRITERIA=CI(.95).
```

**INDEPVAR**

Name of the categorical independent variable

**(G1 G2)**

Specify which values the two categories the independent variable have.

For example:

**(0 1)**

or

**(1 2)**

**DEPVAR**

Name of the continuous dependent variable

## *T-test: independent samples: Example*

(Data: SPSS\_data1.sav)

```
T-TEST GROUPS=retire(0 1)
```

```
/MISSING=ANALYSIS
```

```
/VARIABLES=hourstv
```

```
/CRITERIA=CI(.95).
```

retire

Retired (0=No; 1=Yes)

hourstv

Hours spent watching TV last week (Min=0; Max=36)

### Output/Step 1

The table called **Group Statistics** sums the statistics for the variables in the t-test. Here, it can be interesting to look at each groups' mean value. As can be seen, those who are not retired have a slightly higher mean value for hours spent watching TV: 19.89 compared to 18.21 for those who are retired.

	Retired	N	Mean	Std. Deviation	Std. Error Mean
Hours spent watching TV last week	No	4268	19,89	4,745	,073
	Yes	732	18,21	7,018	,259

## Output/Step 2

The difference between the not-retired group's and retired group's mean value is tested in the next table (**Independent Samples Test**).

The first value of interest is the **Levene's Test for Equality of Variances**. This test indicates which row of the table you are supposed to focus on. Look at the column called **Sig.** If Levene's test *is* significant at the 5 % level ( $p < 0.05$ ) then you focus on the row for **Equal variances not assumed**. If the test *is not* significant ( $p > 0.05$ ), you focus on the row for **Equal variances assumed**.

Once you have established which row you are supposed to look at, the second value of interest is the column called **Sig. (2-tailed)**. A significant p-value here means that the difference between the groups is statistically significant. In this case, the p-value is 0.000 which means that the difference we found in the table above – which showed that those who are not retired watch more TV than those who are retired – is statistically significant (at the 0.1 % level). However, the difference is rather small.

		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	
									Lower	Upper
Hours spent watching TV last week	Equal variances assumed	108,569	,000	8,170	4998	,000	1,680	,206	1,277	2,083
	Equal variances not assumed			6,238	849,229	,000	1,680	,269	1,152	2,209

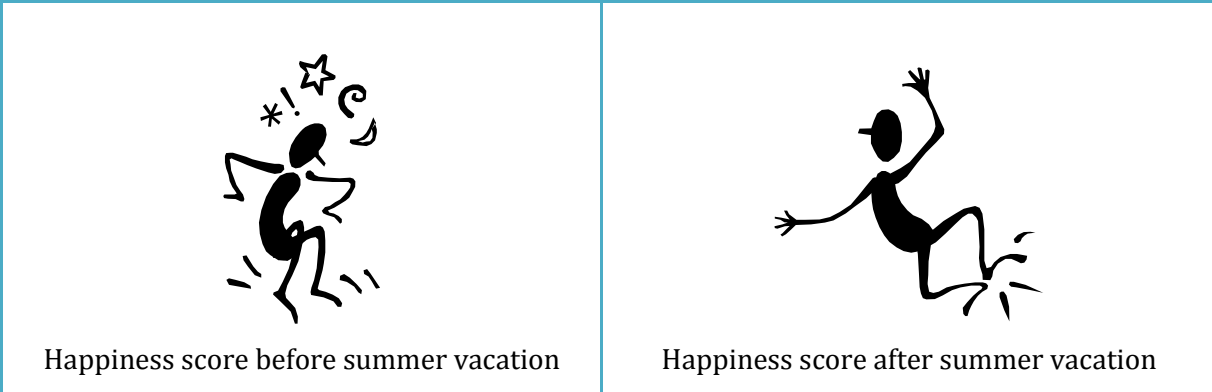


## 7.2 T-test: paired samples

Quick facts	
Number of variables	Two (reflecting repeated measurement points)
Scale of variable(s)	Continuous (ratio/interval)

A dependent or “paired” samples t-test is used to see the difference or change between two measurement points. For example, you could apply this test to see if the staff’s job satisfaction has improved after their boss has taken a course in “socio-emotional skills” compared to before, or if the rate of cigarette smoking in certain schools has declined since the introduction of a new intervention programme.

For the independent samples t-test, you were supposed to have two groups for which you compared the mean. For the paired samples t-test, you instead have two measurements of the same variable, and you look at whether there is a change from one measurement point to the other.



## Assumptions

First, you have to check your data to see that the assumptions behind the paired samples t-test hold. If your data “passes” these assumptions, you will have a valid result. However, “real world” data is often a bit complicated, and it is not uncommon that at least one of the assumptions is violated. While you most common will be able to conduct the t-test anyway, it is important to be aware of the possible problems.

Checklist	
<b>Continuous variables</b>	Your two variables should be continuous (i.e. interval/ratio). For example: Income, height, weight, number of years of schooling, and so on. Although they are not really continuous, it is still very common to use ratings as continuous variables, such as: “How satisfied with your income are you?” (on a scale 1-10) or “To what extent do you agree with the previous statement?” (on a scale 1-5).
<b>Two measurement points</b>	Your two variables should reflect one single phenomenon, but this phenomenon is measured at two different time points for each individual.
<b>Normal distribution</b>	Both variables need to be approximately normally distributed. Use a histogram to check (see Section 4.6).
<b>No outliers in the comparison between the two measurement points</b>	For example, if one individual has an extremely low value at the first measurement point and an extremely high value at the second measurement point (or vice versa), this will distort the test. Use a scatterplot to check (see Section 4.7).

### *T-test: paired samples: Function*

1. Go to the **Menu bar**, choose **Analyze\Compare Means\Paired Samples T Test**.
2. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable and transfer it to the box called **Paired variables**.
3. Then you choose the variable you want as your independent variable and transfer it to the box called **Paired variable**.
4. Click on **OK**.

*T-test: paired samples: Syntax*

```
T-TEST PAIRS=MEASURE1 WITH MEASURE2 (PAIRED)  
/CRITERIA=CI(.9500)  
/MISSING=ANALYSIS.
```

**MEASURE1**                      Insert the name of the variable containing information about the first measurement point.

**MEASURE2**                      Insert the name of the variable containing information about the second measurement point.

## *T-test: paired samples: Example*

(Data: SPSS\_data1.sav)

```
T-TEST PAIRS=unempl_03 WITH unempl_05 (PAIRED)
```

```
/CRITERIA=CI(.9500)
```

```
/MISSING=ANALYSIS.
```

**unempl\_03**

Unemployment days in 2003 (ranges from 0 to 365)

**unempl\_05**

Unemployment days in 2005 (ranges from 0 to 365)

### Output/Step 1

The table called **Paired Samples Statistics** shows the statistics for the variables. For example, it shows the mean value for each of the two measurement points. In the current example, we see that the mean number of unemployment days is lower in 2003 (mean=8.12) than in 2005 (mean=11.31).

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 Unemployment days in 2003	8,12	4971	23,286	,330
Unemployment days in 2005	11,31	4971	44,103	,626

### Output/Step 2

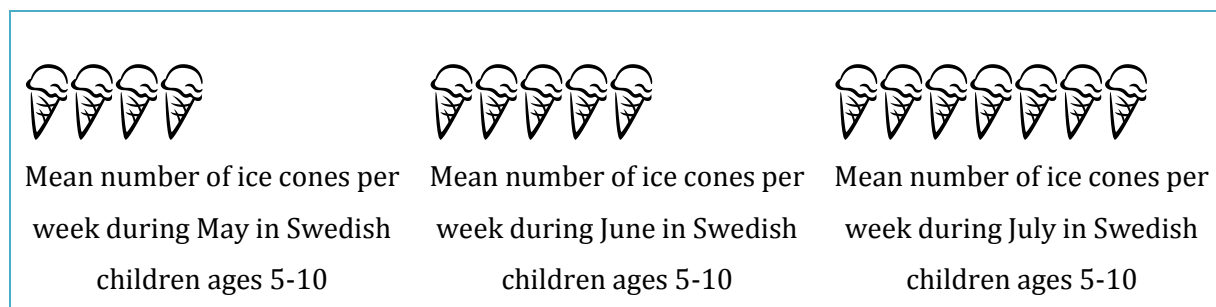
The table called **Paired Samples Test** shows the results from the actual t-test. The first column – **Mean** – shows that the mean difference between unemployment days in 2003 and unemployment days in 2005 is -3.190 (this difference is actually just derived from taking 11.31 minus 8.12). The last column – **Sig. (2-tailed)** – shows the p-value for this difference. If the p-value is smaller than 0.05, the test suggests that there is a statistically significant difference (at the 5 % level). Thus, here we can conclude that there is a statistically significant increase in unemployment days from 2003 to 2005.

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 Unemployment days in 2003 - Unemployment days in 2005	-3,190	43,017	,610	-4,386	-1,994	-5,228	4970	,000

### 7.3 One-way ANOVA

Quick facts	
<b>Number of variables</b>	One independent (x) One dependent (y)
<b>Scale of variable(s)</b>	Independent: categorical (nominal/ordinal) Dependent: continuous (ratio/interval)

The one-way ANOVA is very similar to the independent samples t-test. The difference is that the one-way ANOVA allows you to have more than two categories in your independent variable. For example, you can compare how many cups of coffee people drink per day depending on if they have a low-stress, medium-stress, or high-stress job. Or you can see if the number of days of paternity leave differs between fathers in Sweden, Denmark, Norway and Finland. However, it is important to remember that the one-way ANOVA does not tell you exactly which groups are different from each other; it only tells you that at least two of the groups differ in terms of the outcome.



## Assumptions

First, you have to check your data to see that the assumptions behind the one-way ANOVA hold. If your data “passes” these assumptions, you will have a valid result. However, “real world” data is often a bit complicated, and it is not uncommon that at least one of the assumptions is violated. While you most common will be able to conduct the test anyway, it is important to be aware of the possible problems.

Checklist	
<b>Continuous dependent variable</b>	Your dependent variable should be continuous (i.e. interval/ratio). For example: Income, height, weight, number of years of schooling, and so on. Although they are not really continuous, it is still very common to use ratings as continuous variables, such as: “How satisfied with your income are you?” (on a scale 1-10) or “To what extent do you agree with the previous statement?” (on a scale 1-5).
<b>Two or more unrelated categories in the independent variable</b>	Your independent variable should be categorical (i.e. nominal or ordinal) and consist of two or more groups. Unrelated means that the groups should be mutually excluded: no individual can be in more than one of the groups. For example: low vs. medium vs. high educational level; liberal vs. conservative vs. socialist political views; or poor vs. fair, vs. good vs. excellent health; and so on.
<b>No outliers</b>	An outlier is an extreme (low or high) value. For example, if most individuals have a test score between 40 and 60, but one individual has a score of 96 or another individual has a score of 1, this will distort the test.



## *One-way ANOVA: Function*

1. Go to the **Menu bar**, choose **Analyze\Compare Means\One-way ANOVA**.
2. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable and transfer it to the box called **Dependent list**.
3. You also choose the variable you want as your independent variable and transfer it to the box called **Factor**.
4. Go to the box **Option**. Tick the boxes called **Descriptive**, **Homogeneity of variance test**, **Brown-Forsythe**, **Welch** and **Means Plot**.
5. Click on **Continue** and then on **OK**.

## *One-way ANOVA: Syntax*

```
ONEWAY DEPVAR BY INDEPVAR  
/STATISTICS OPTIONS  
/MISSING ANALYSIS.
```

<b>DEPVAR</b>	Name of the continuous dependent variable It is possible to list more than one dependent variable
<b>INDEPVAR</b>	Name of the categorical independent variable
<b>OPTIONS</b>	List the options you want to be included For example: <b>DESCRIPTIVES</b>

## One-way ANOVA: Example

(Data: SPSS\_data1.sav)

```
ONEWAY income BY bfast  
/STATISTICS DESCRIPTIVES HOMOGENEITY WELCH  
/MISSING ANALYSIS.
```

**income** Household income in thousands of dollars (Min=9; Max=1073)

**bfast** Preferred breakfast (1=Energy bar; 2=Oatmeal; 3=Cereal)

### Output/Step 1

The table called **Paired Samples Statistics** shows the descriptive statistics for the variables, including the mean value of the dependent variable (**income**) for each category of the independent variable (**bfast**).

Descriptives								
Household income in thousands								
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Energy bar	1554	46,7651	39,96608	1,01383	44,7765	48,7537	9,00	424,00
Oatmeal	1549	60,7366	61,95588	1,57419	57,6488	63,8244	9,00	780,00
Cereal	1897	57,1687	60,04177	1,37854	54,4651	59,8723	9,00	1073,00
Total	5000	55,0406	55,54475	,78552	53,5006	56,5806	9,00	1073,00

### Output/Step 2

The table called **Test of Homogeneity of Variances** shows the results from a Levene's test for testing the assumption of equal variances. Put differently, this test is concerned with whether or not the variances of the categories are different from one another. The column called **Sig.** shows the p-value for this test. If the p-value is larger than 0.05, we can use the results from the standard ANOVA test. However, if the p-value is smaller than 0.05, it means that the assumption of homogeneity of variance is violated and we cannot trust the standard ANOVA results. Instead, we focus on the results from the Welch ANOVA. Note, however, than both these tests produce so-called F statistics.

Test of Homogeneity of Variances			
Household income in thousands			
Levene Statistic	df1	df2	Sig.
46,458	2	4997	,000

### Output/Step 3

If the p-value for the Levene's test had been larger than 0.05 in this example, we would have focused on the table called **ANOVA**. The column called **Sig.** shows whether or not we have a statistically significant difference in our dependent variable (**income**) between the categories of the independent variable (**bfast**). A p-value that is smaller than 0.05 would suggest that there is indeed a statistically significant difference (at the 5 % level).

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	165270,934	2	82635,467	27,064	,000
Within Groups	15257739,82	4997	3053,380		
Total	15423010,76	4999			

### Output/Step 4

Since the p-value for the Levene's test in this example was smaller than 0.05, we instead focus on the next table: **Robust Tests of Equality of Means**. The same principle as for the standard ANOVA test applies here: if the p-value (in the column called **Sig.**) is smaller than 0.05, it means that we have a statistically significant difference (at the 5 % level) in our dependent variable (**income**) between the categories of the independent variable (**bfast**).

	Statistic <sup>a</sup>	df1	df2	Sig.
Welch	35,404	2	3209,671	,000

a. Asymptotically F distributed.

## 8. CHI SQUARE

### Outline

#### 8.1 Chi-square

## 8.1 Chi-square

Quick facts	
<b>Number of variables</b>	Two
<b>Scale of variable(s)</b>	Categorical (nominal/ordinal)

There are two different forms of the chi-square test: a) The multidimensional chi-square test, and b) The goodness of fit chi-square test. It is the first form that will be covered in this part of the guide. The second form is discussed in Section 14.3.

The multidimensional chi-square test assesses whether there is a relationship between two categorical variables. For example, you want to see if young women smoke more than young men. The variable gender has two categories (men and women) and, in this particular case, the variable smoking consists of the categories: no smoking, occasional smoking and frequent smoking. The multidimensional chi-square test can be thought of as a simple crosstable where the distribution of these two variables is displayed:

	<i>No smoking</i>	<i>Occasional smoking</i>	<i>Frequent smoking</i>
<i>Men (age 15-24)</i>	85 %	10 %	5 %
<i>Women (age 15-24)</i>	70 %	20 %	10 %

### *Assumptions*

First, you have to check your data to see that the assumptions behind the chi-square test hold. If your data “passes” these assumptions, you will have a valid result.

Checklist	
<b>Two or more unrelated categories in both variables</b>	Both variables should be categorical (i.e. nominal or ordinal) and consist of two or more groups. Unrelated means that the groups should be mutually excluded: no individual can be in more than one of the groups. For example: low vs. medium vs. high educational level; liberal vs. conservative vs. socialist political views; or poor vs. fair, vs. good vs. excellent health; and so on.

## *Chi-square: Function*

1. Go to the **Menu bar**, choose **Analyze\Descriptive Statistics\Crosstabs**.
2. A small window will open, where you see one big box and three small boxes. In the left box, all your variables are displayed.
3. Here, you choose two variables: one to be the **Row** variable, and one to be the **Column** variable.
4. Move your variables to the **Row** and **Column** boxes by using the arrows.
5. Click on **Statistics**.
6. Tick the box for **Chi-square**.
7. Click on **Continue**.
8. Tick the box called **Suppress tables** located below the box containing all variables.
9. Click on **OK**.



## Chi-square: Syntax

```
CROSSTABS  
  /TABLES= VARNAME1 BY VARNAME2  
  /FORMAT=NOTABLES  
  /STATISTICS=TYPE  
  /COUNT ROUND CELL.
```

**VARNAME1**                      Insert the name of the first variable you want to use.  
This variable will be chosen for rows.

**VARNAME2**                      Insert the name of the second variable you want to use.  
This variable will be chosen for columns.

**/FORMAT=NOTABLES**            Insert this if you do not want SPSS to produce a cross table (in that  
case you will only get the results from the chi-square test.

**TYPE**                            Choose the type of statistics you want to produce.  
For example:  
**CHISQ**

## Chi-square: Example

(Data: SPSS\_data1.sav)

CROSSTABS

/TABLES=bfast BY gender

/FORMAT=NOTABLES

/STATISTICS=CHISQ

/COUNT ROUND CELL.

bfast

Preferred breakfast (1=Energy bar; 2=Oatmeal; 3=Cereal)

gender

Gender (0=Man; 1=Woman)

## Output

The table called **Chi-Square Tests** shows the results from the chi-square test for the variables **bfast** and **gender**. Here, we look at the row called Pearson Chi-Square and the column **Asymp. Sig. (2-sided)** to see the p-value for the test. A p-value smaller than 0.05 indicates that there is a statistically significant association (at the 5 % level) between the two variables in the test, whereas a p-value larger than 0.05 suggests that there is not a statistically significant association. Since the p-value in this example is 0.891, we can conclude that **bfast** and **gender** are not associated with one another at a statistically significant level.

**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	,230 <sup>a</sup>	2	,891
Likelihood Ratio	,230	2	,891
Linear-by-Linear Association	,168	1	,682
N of Valid Cases	5000		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 758,70.

## 9. CORRELATION ANALYSIS

### Outline

#### 9.1 Correlation analysis

## 9.1 Correlation analysis

Quick facts	
<b>Number of variables</b>	Two or more
<b>Scale of variable(s)</b>	Continuous (ratio/interval)

A correlation analysis tests the relationship between two continuous variables in terms of: a) how strong the relationship is, and b) in what direction the relationship goes. The strength of the relationship is given as a coefficient (the Pearson product-moment correlation coefficient, or simply Pearson's  $r$ ) which can be anything between -1 and 1. But how do we know if the relationship is strong or weak? This is not an exact science, but here is one rule of thumb:

Strength		
Negative	Positive	
-1	1	<b>Perfect</b>
-0.9 to -0.7	0.7 to 0.9	<b>Strong</b>
-0.6 to -0.4	0.4 to 0.6	<b>Moderate</b>
-0.3 to -0.1	0.1 to 0.3	<b>Weak</b>
0	0	<b>Zero</b>

Thus, the coefficient can be negative or positive. These terms, "negative" and "positive", are not the same as good and bad (e.g. excellent health or poor health; high income or low income). They merely reflect the direction of the relationship.

Direction	
<b>Negative</b>	As the values of Variable 1 increases, the values of Variable 2 <i>decreases</i>
<b>Positive</b>	As the values of Variable 1 increases, the values of Variable 2 <i>increases</i>

Note however that correlation analysis does not imply anything about causality: Variable 1 does not *cause* Variable 2 (or vice versa). The correlation analysis only says something about the degree to which the two variables co-vary.

## Assumptions

First, you have to check your data to see that the assumptions behind the correlation analysis hold. If your data “passes” these assumptions, you will have a valid result.

Checklist	
<b>Two continuous variables</b>	Both variables should be continuous (i.e. interval/ratio). For example: Income, height, weight, number of years of schooling, and so on. Although they are not really continuous, it is still rather common to use ratings as continuous variables, such as: “How satisfied with your income are you?” (on a scale 1-10) or “To what extent do you agree with the previous statement?” (on a scale 1-5).
<b>Linear relationship between the two variables</b>	There needs to be a linear relationship between your two variables. You can check this by creating a scatterplot (described in Section 4.7).
<b>No outliers</b>	An outlier is an extreme (low or high) value. For example, if most individuals have a test score between 40 and 60, but one individual has a score of 96 or another individual has a score of 1, this will distort the test.

## *Correlation analysis: Function*

1. Go to the **Menu bar**, choose **Analyze\Correlate\Bivariate**.
2. A new window called **Bivariate Correlations** will open.
3. In the left box, all your variables are displayed. Highlight the variables for which you want to test the correlation, and then transfer them to **Variables**.
4. Click on **OK**.

Note that it is possible to include more than two variables, and in that case you will get correlation coefficients for each pair of variables.

## *Correlation analysis: Syntax*

```
CORRELATIONS  
/VARIABLES=VARNAMES  
/PRINT=TWOTAIL NOSIG  
/MISSING=PAIRWISE.
```

**VARNAMES**

Insert the names of at least two variables



## Correlation analysis: Example

(Data: SPSS\_data1.sav)

CORRELATIONS

/VARIABLES=jobsat age

/PRINT=TWOTAIL NOSIG

/MISSING=PAIRWISE.

jobsat	Job satisfaction (1=Highly dissatisfied; 2=Somewhat dissatisfied; 3=Neutral; 4=Somewhat satisfied; 5=Highly satisfied)
age	Age (Min=18; Max=79)

## Output

The table called **Correlations** shows the results from the correlation analysis of the two variables **jobsat** and **age**. Note that the table is a matrix, meaning that it can be read from left to right *or* from the top to the bottom. Here, we focus on the row called **Pearson Correlation**. Of course, the correlation between **jobsat** and **jobsat** is perfect ( $r=1$ ), and so is the correlation between **age** and **age** ( $r=1$ ). Now, remember that an increasing value for **age** means that the individual is older, but an increasing value for **jobsat** means that the individual is less satisfied with his or her job. The correlation coefficient for **jobsat** and **age** is 0.413, which is positive. In other words: as the age increases, job satisfaction decreases. Concerning the strength of the correlation, 0.413 can be said to be moderate.

**Sig. (2-tailed)** shows the p-value for the correlation. A p-value smaller than 0.05 suggests that the correlation is statistically significant (at the 5 % level). SPSS also includes a significance test at the 1 % level, indicated by the asterisks (\*\*) presented next to the correlation coefficient.

		Job satisfaction	Age in years
Job satisfaction	Pearson Correlation	1	,413**
	Sig. (2-tailed)		,000
	N	5000	5000
Age in years	Pearson Correlation	,413**	1
	Sig. (2-tailed)	,000	
	N	5000	5000

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

# 10. FACTOR ANALYSIS

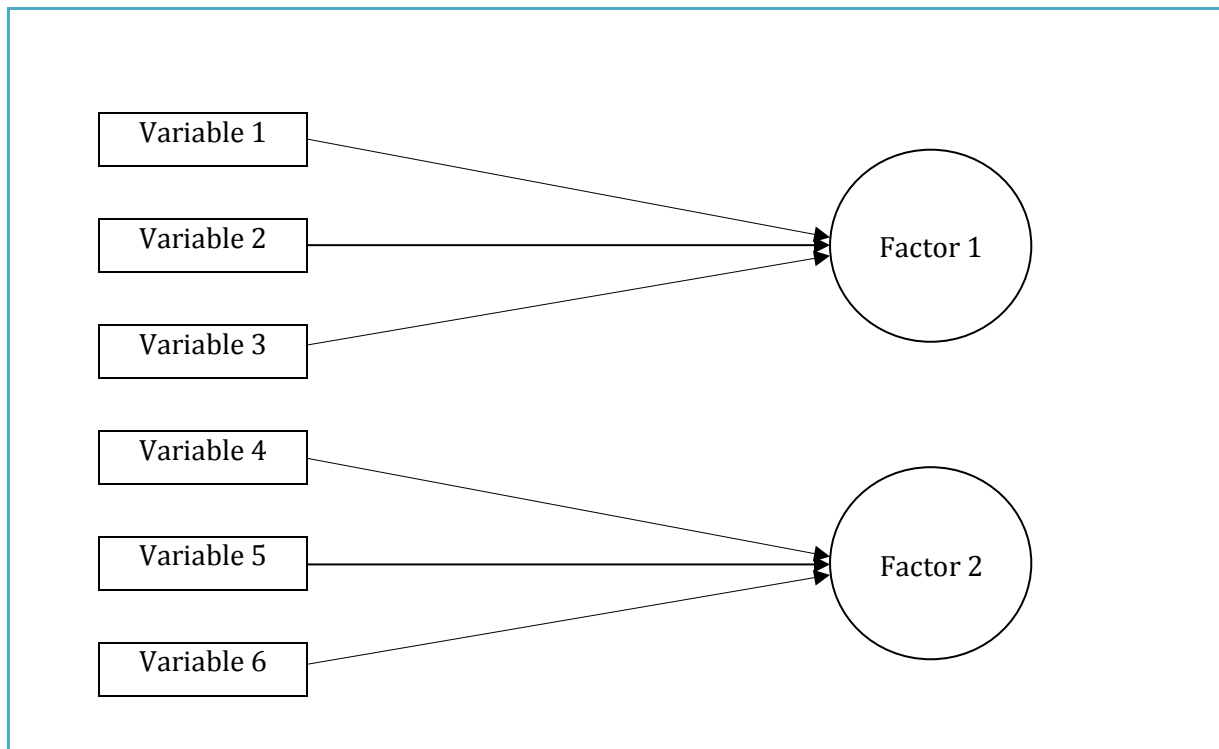
Outline
10.1 Factor analysis
10.2 Cronbach's alpha

## 10.1 Factor analysis

### Quick facts

<b>Number of variables</b>	Two or more
<b>Scale of variable(s)</b>	Continuous (ratio/interval) or approximately continuous

There are two general types of factor analysis: exploratory factor analysis and confirmatory factor analysis. It is the first type that will be covered here. The main feature of exploratory factor analysis (hereafter referred to as “factor analysis”) is that it enables us to investigate the underlying structure in the pattern of correlations between a number of variables (often referred to as “items”). If we have a large number of variables, we can thus investigate if these variables represent a smaller number of factors – or “dimensions”. Before getting into factor analysis in more detail, just a cautionary note: it is easy to let the data guide all your decisions, but do not forget what theory tells you before making these decisions!



## Assumptions

First, you have to check your data to see that the assumptions behind the factor analysis hold. If your data “passes” these assumptions, you will have a valid result.

Checklist	
<b>Ratio/interval/ordinal variables</b>	Your variables should be continuous (i.e. interval/ratio) or ordinal (but still approximately continuous). For example: Income, height, weight, number of years of schooling, or ratings.
<b>Linear associations</b>	The variables in the factor analysis should be associated with each other in a linear fashion (use scatter plots to check, see Section 4.7).
<b>Sample size</b>	Factor analysis requires rather large samples. However, recommendations on this topic vary greatly. Some recommendations highlight the absolute sample size (here, lower limits range from $n=100$ to $n=500$ ) whereas others says that subject-to-variable ratio is important (and here, ratios from 2:1 to 20:1 are suggested).
<b>No outliers</b>	An outlier is an extreme (low or high) value. For example, if most individuals have a test score between 40 and 60, but one individual has a score of 96 or another individual has a score of 1, this will distort the test.

Suppose that we have asked a bunch of individuals, six questions about their health. We conduct a factor analysis to see how many dimensions these questions reflect: do all questions reflect only one dimension (namely “health”) or can they be categorised into two or more dimensions (i.e. different types of health)?

## Pre-tests

First, we have to find out if it is a good idea to conduct a factor analysis at all. Here, you may use two tests to help you decide:

Factor analysis or not?
Kaiser-Meyer-Olkin Measure of Sampling Adequacy
The Bartlett's Test of Sphericity

The test called Kaiser-Meyer-Olkin Measure of Sampling Adequacy (in short: the KMO test) reflects the sum of partial correlations relative to the sum of correlations. It varies between 0 and 1, where a value closer to 1 is better. It has been suggested to use 0.5 as a minimum requirement. Thus, if the value is lower than 0.5, factor analysis may be inappropriate.

The Bartlett's Test of Sphericity tests the hypothesis that the correlation matrix is an identity matrix; if it is an identity matrix then there would be no correlations between the variables (and that is obviously not very good). Thus, the test needs to be statistically significant (i.e.  $p < 0.05$ ) so we are able to reject this hypothesis. If we cannot reject it, it would be inappropriate to conduct a factor analysis.

## *Number of factors*

So, suppose that we find that it is suitable to conduct a factor analysis. How do we ascertain how many factors/dimensions there are in our data? Well, there are several different ways to do this.

<b>Determining the number of factors</b>	
<b>Eigenvalue&gt;1</b>	Eigenvalues are indicators of the variance explained by a factor. Using the rule “eigenvalue is greater than one” is very common. The reasoning behind this rule is that a factor should account for at least as much variance as any single variable. Thus, the average of all eigenvalues is one, and the factor analysis should thus extract factors that have an eigenvalue greater than this average value.
<b>Scree plot</b>	In a scree plot, factors have their eigenvalues plotted alongside the y-axis (i.e. vertical axis) in the order of magnitude. Factors explaining large amounts of variance appear to the left, whereas factors explaining little variance are aligned to the right. The somewhat weird task is here to “locate the elbow”. This means to identify the number of factors stated before the line starts becoming flat.
<b>Explained variance</b>	This refers to the amount of variation that the factor explains. Some suggest that 90 % should be used as a cut-off point, whereas others go as low as 50 %.

## *Factor loadings*

Once we have decided on the number of factors, we retrieve the “factor loadings”. A factor loading is basically a correlation coefficient (see Chapter 9) and, thus, it varies between -1 and +1 (where a value closer to -1 or +1 indicates a stronger correlation). Factor loadings are given for each variable, for each factor separately. In other words, a factor loading shows how strongly a certain variable correlates with the given factor. There are no exact rules for deciding on when a loading is strong enough, but one suggested rule of thumb is below -0.5 or above 0.5. However, sometimes a variable has strong loadings for more than one factor (called “cross-loading”). This can for example happen if you have not extracted enough factors, or if the factors are correlated. Sometimes a variable has weak loadings for all factors; this may suggest that this variable is weakly related to all other variables or that you need to explore an additional factor (or maybe even exclude this specific variable).



## Rotation

A factor analysis has the most interpretative value when: 1) Each factor loads strongly on only one factor; 2) Each factor shows at least three strong loadings; 3) Most loadings are either high or low; and 4) We get a “simple” factor structure. Rotation is a way of maximizing high loadings and minimizing low loadings so that we get the simplest factor structure possible. There are two main types of rotation:

Rotation	
<b>Orthogonal</b>	Assumes that the factors are uncorrelated Examples of sub types: equamax, quartimax and varimax
<b>Oblique</b>	Assumes that the factors are correlated Examples of sub types: direct oblimin and promax

Thus, the orthogonal rotation is based on the assumption that the factors are not correlated to each other, i.e. that the different factors represent different unrelated dimensions of what you are examining. This is not always the case. For example, if you have several variables measuring health, and find one factor that reflects physical health and another one reflecting psychological health, it may not be reasonable to assume that physical and psychological health two unrelated dimension. In that case, you need to change the type of rotation to oblique.

## Factor analysis: Function

1. Go to the **Menu bar**, choose **Analyze\Dimension Reduction\Factor**.
2. A new window called **Factor Analysis** will open.
3. In the left box, all your variables are displayed. Highlight the variables that you want to include in the analysis, and then transfer them to **Variables**.
4. To order **Kaiser-Meyer-Olkin Measure of Sampling Adequacy** and **Bartlett's Test of Sphericity**: click on **Descriptives**, tick the box called **KMO** and **Bartlett's test of sphericity**, and then click on **Continue**.
5. To order a scree plot: click on **Extraction**, tick the box called **Scree plot**, and then click on **Continue**.
6. If you for some reason want to tell SPSS exactly how many factors you want, go to **Extraction**, tick the box called **Fixed number of factors**, state the preferred number of factors, and then click on **Continue**.
7. To change the rotation of the factor analysis: click on **Rotation**, tick the box for the preferred type of rotation (**None**; **Varimax**; **Direct Oblimin**; **Quartimax**; **Equamax**; or **Promax**), and then click on **Continue**. Note that you will only get rotated factor loadings if SPSS identifies *at least* two factors.
8. Click on **OK**.

## Factor analysis: Syntax

```
FACTOR
/VARIABLES VARLIST1
/MISSING LISTWISE
/ANALYSIS VARLIST2
/PRINT INITIAL TESTNAME EXTRACTION ROTATION
/PLOT PLOTNAME
/CRITERIA MINEIGEN(1) ITERATE(25)
/EXTRACTION PC
/CRITERIA ITERATE(25)
/ROTATION ROTATIONTYPE
/METHOD=CORRELATION.
```

VARLIST1	List all the variables you want to analyse
VARLIST2	List all the variables you want to analyse (same as VARLIST1)
TESTNAME	Name of the test you want to order For example: KMO (KMO and Bartlett's test of sphericity)
ROTATION	Add this here if you want to rotate your analysis
PLOTNAME	Name of the plot you want to order For example: EIGEN (Scree plot)
/CRITERIA ITERATE(25)	Add this here if you are rotating your analysis
ROTATIONTYPE	Name of the type of rotation you want to use For example: VARIMAX

## Factor analysis: Example

(Data: SPSS\_data2.sav)

FACTOR

```
/VARIABLES imp_ideas imp_rich imp_secure imp_good imp_help imp_success imp_risk  
imp_behave imp_environ imp_trad  
/MISSING LISTWISE  
/ANALYSIS imp_ideas imp_rich imp_secure imp_good imp_help imp_success imp_risk  
imp_behave imp_environ imp_trad  
/PRINT INITIAL KMO EXTRACTION ROTATION  
/PLOT EIGEN  
/CRITERIA MINEIGEN(1) ITERATE(25)  
/EXTRACTION PC  
/CRITERIA ITERATE(25)  
/ROTATION VARIMAX  
/METHOD=CORRELATION.
```

imp_ideas	Important to think up new ideas
imp_rich	Important to be rich
imp_secure	Important living in secure surroundings
imp_good	Important to have a good time
imp_help	Important to help the people
imp_success	Important being very successful
imp_risk	Important with adventure and taking risks
imp_behave	Important to always behave properly
imp_environ	Important looking after the environment
imp_trad	Important with tradition

(1=Very much like me; 2=Like me; 3=Somewhat like me; 4=Not like me; 5=Not at all like me)

### Output/Step 1

The first step is to look at the table called **KMO and Bartlett's Test**. The estimate in the first row is the result from the **KMO test**. The value here is 0.792, which suggests that the factor analysis is appropriate. The estimate in the last row is the result from the **Bartlett's Test of Sphericity**. The p-value here is 0.000, which also suggests that factor analysis is appropriate.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,792
Bartlett's Test of Sphericity	Approx. Chi-Square	102321,410
	df	45
	Sig.	,000

## Output/Step 2

The second step is to look at the table called **Total Variance Explained**. There is one row for each factor-solution (called **Component**). With regard to **Initial Eigenvalues** and the column called **Total**, the results show a suitable cut-off between two and three factors since the three-factor solution has an eigenvalue below 1. Thus, this suggests that a two-factor solution is preferable. Looking at the column called **Cumulative %**, the results show that the two-factor solution together explain approximately 46 % of the variance. That is not a high percentage but we may conclude that it is acceptable.

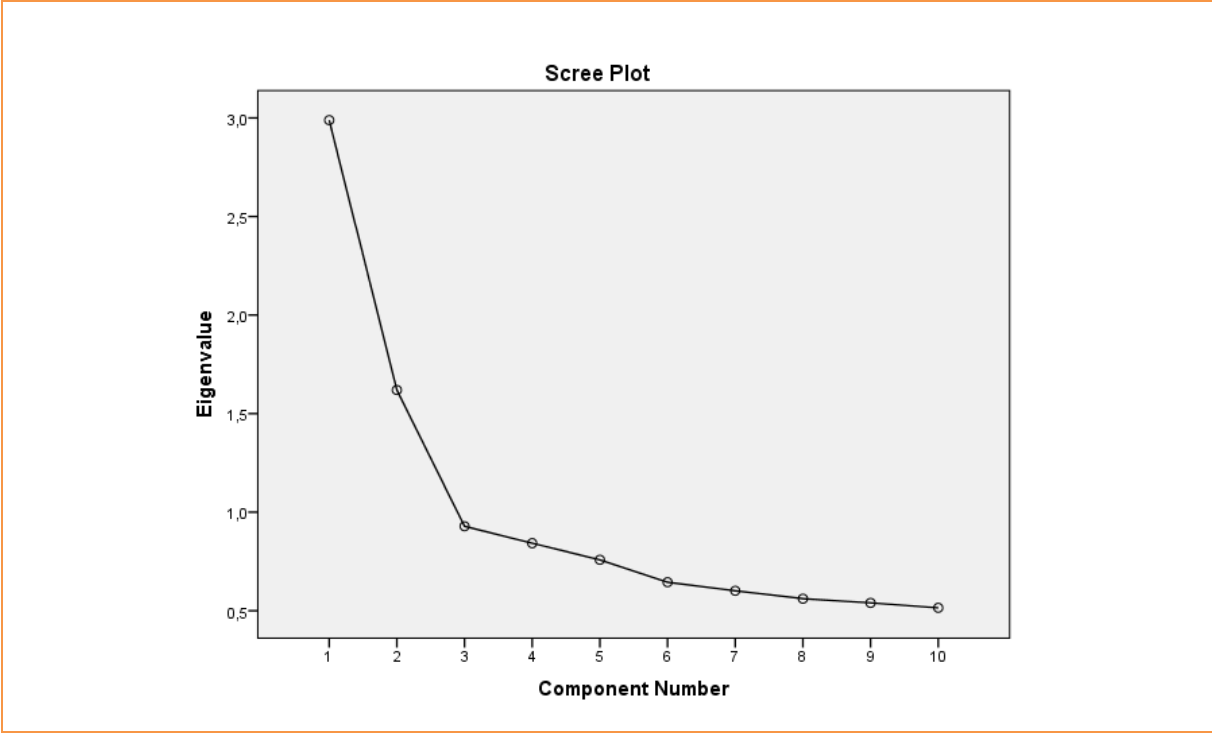
One important thing to know is that SPSS automatically chooses to proceed with the solution that consists of the most factors that has an eigenvalue greater than one (in this case, the two-factor solution).

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	2,989	29,887	29,887	2,989	29,887	29,887	2,449	24,495	24,495
2	1,620	16,197	46,084	1,620	16,197	46,084	2,159	21,589	46,084
3	,929	9,286	55,370						
4	,843	8,427	63,797						
5	,758	7,582	71,379						
6	,645	6,447	77,826						
7	,601	6,014	83,840						
8	,561	5,613	89,453						
9	,540	5,400	94,853						
10	,515	5,147	100,000						

Extraction Method: Principal Component Analysis.

**Output/Step 3**

The third step is to look at the graph called **Scree Plot**. Eigenvalues are presented alongside the y-axis and the number of the component (i.e. factor) is presented alongside the x-axis. Here we can see that the eigenvalues are getting much flatter (i.e. the elbow!) starting at the three-factor solution. Thus, once again it can be argued that we should go with the two-factor solution.



#### Output/Step 4

The fourth step is to look at the table called **Rotated Component Matrix**. This table shows the factor loading for each variable and for each of the two factors separately.

	Component	
	1	2
Important to think up new ideas	,205	,591
Important to be rich	,038	,701
Important living in secure surroundings	,624	,164
Important to have a good time	,053	,593
Important to help the people	,665	,151
Important being very successful	,345	,632
Important with adventure and taking risks	-,046	,709
Important to always behave properly	,703	,089
Important looking after the environment	,682	,076
Important with tradition	,701	-,020

Extraction Method: Principal Component Analysis.  
Rotation Method: Varimax with Kaiser Normalization.  
a. Rotation converged in 3 iterations.

We identify for with factor each variable has the higher loading, we can conclude that the two factors contain the following variables:

#### **Factor 1**

- Important living in secure surroundings (security)
- Important to help the people (benevolence)
- Important to always behave properly (conformity)
- Important looking after the environment (universalism)
- Important with tradition (tradition)

#### **Factor 2**

- Important to think up new ideas (self-direction)
- Important to be rich (power)



Important to have a good time (hedonism)

Important being very successful (achievement)

Important with adventure and taking risks (stimulation)

The ten variables used in this factor analysis actually stem from a theory of human values, developed by Schwartz. According to this theory, the variables should be categorised in the following way:

Conservation: security, tradition, and conformity

Openness to change: self-direction, stimulation, and hedonism

Self-enhancement: power and achievement

Self-transcendence: benevolence and universalism

If we compare the theoretical categories with the factors derived from factor analysis, we actually see that the Factor 1 includes all variables theoretically associated with conservation and self-transcendence, whereas Factor 2 includes all variables theoretically associated with openness to change and self-enhancement. What do we do with this information then? Well, we need to examine possible reasons as to why the factor analysis did not reveal the same factors as the theory proposes. If we find no apparent problems with the empirics (e.g. missing data, problems with the questionnaire itself, etc.) we may suggest that the theory needs to be modified. At least it is important to discuss the differences between the theory and the empirics.

Sometimes, we do not have a clear theory guiding the factor analysis and, thus, we have no a priori understanding about which factors that are reasonable to expect. In that case, it is common practice to focus on a factor solution with good properties (i.e. clear factor structure and high factor loadings). It is always a trade-off between theory and empirics: if theory has precedence over empirics, we may be more disposed to accept lower factor loadings.

### *A note on composite measures (indices)*

A common reason for conducting a factor analysis is that we want to make a composite measure (i.e. an index) of a set of variables. If these variables all fall into one single factor (with acceptable factor loadings), that is safe to do. If they clearly fall into different factors, a single index for these variables *should not* be calculated. Possible solutions could be to exclude variables not fitting into the factor of choice. Another solution is to create more than one index, reflecting different dimensions of the concept you focusing on.

## 10.2 Cronbach's alpha

Quick facts	
<b>Number of variables</b>	Two or more
<b>Scale of variable(s)</b>	Continuous (ratio/interval) or approximately continuous

The common application of the Cronbach's alpha is when we have a composite measure – i.e. an index – and want to see if the variables included in the index actually reflect the same underlying factor/dimension. Formally speaking, the Cronbach's alpha is a measure of internal consistency; how closely related a number of items are as a group. The coefficient ranges between 0 and 1. A high alpha value indicates that items measure an underlying factor. However, it is not a statistical test but a test of reliability/consistency.

One important thing to note is that the Cronbach's alpha is affected by the number of variables: including a higher number of variables automatically increases the alpha value to some extent.

### *Rule of thumb*

There are many rules of thumb with regard to what is considered a good or bad alpha value. Generally, an alpha value of at least 0.7 is considered acceptable.

Alpha values	
<b>Between 0.7 and 1.0</b>	Acceptable
<b>Below 0.7</b>	Not acceptable

## *Cronbach's alpha: Function*

1. Go to the **Menu bar**, choose **Analyze\Scale\Reliability Analysis**.
2. A new window called **Reliability Analysis** will open.
3. In the left box, all your variables are displayed. Highlight the variables that you want to include in the analysis, and then transfer them to **Items**.
4. If you want to see what would happen to the alpha value if you would exclude any specific variable, click on **Statistics**. Tick the box called **Scale if item deleted**, and then click on **Continue**.
5. Click on **OK**.

## *Cronbach's alpha: Syntax*

```
RELIABILITY  
/VARIABLES=VARLIST  
/SCALE('ALL VARIABLES') ALL  
/MODEL=ALPHA  
/SUMMARY=TOTAL.
```

**VARLIST**                      List all the variables you want to analyse

**/SUMMARY=TOTAL**            Add this if you want to see how the alpha value changes if a certain variable was excluded

## Cronbach's alpha: Example

(Data: SPSS\_data2.sav)

RELIABILITY

/VARIABLES=imp\_secure imp\_help imp\_behave imp\_environ imp\_trad

/SCALE('ALL VARIABLES') ALL

/MODEL=ALPHA

/SUMMARY=TOTAL.

imp_secure	Important living in secure surroundings
imp_help	Important to help the people
imp_behave	Important to always behave properly
imp_environ	Important looking after the environment
imp_trad	Important with tradition

(1=Very much like me; 2=Like me; 3=Somewhat like me; 4=Not like me; 5=Not at all like me)

### Output/Step 1

The first step is to look at the table called **Reliability Statistics**. Here, you see the alpha value. In this case, it is 0.716.

Cronbach's Alpha	N of Items
.716	5

## Output/Step 2

The second step is to look at the table called **Item-Total Statistics**. In the column for **Cronbach's Alpha if Item Deleted**, you see what happens to the alpha value if any of the specific variables would be excluded from the test. In this specific case, the exclusion of any variable would actually decrease the alpha value – hence, it is better to keep all the variables in the index. If the results had been the opposite (i.e. the alpha value would increase if a certain variable was excluded), we may have considered removing that variable from the index – but any such decisions should always be evaluated against what is stipulated by the theory you use.

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Important living in secure surroundings	9,72	13,264	,444	,681
Important to help the people	9,85	13,919	,470	,672
Important to always behave properly	9,57	12,426	,508	,654
Important looking after the environment	9,71	13,505	,478	,668
Important with tradition	9,64	12,401	,480	,668



## 11. X, Y AND Z

### Outline

11.1 X and Y

11.2 Z: confounding, mediating and moderating variables

We talk a lot about variables in this guide, because variables are the cornerstones of quantitative data materials and quantitative data analysis. Other terms are sometimes used instead of “variables” – such as “indicators”, “measures” or “items” – but they basically reflect the same thing.

### *Associations*

In many types of analysis – such as when we compare means or conduct regression analysis – we are interested in the association between two (or more) variables. The term “association”, or “relationship”, reflects the assumption that the variables are related to one another in some way. Basically, that means that the variables are correlated.

### *Effects*

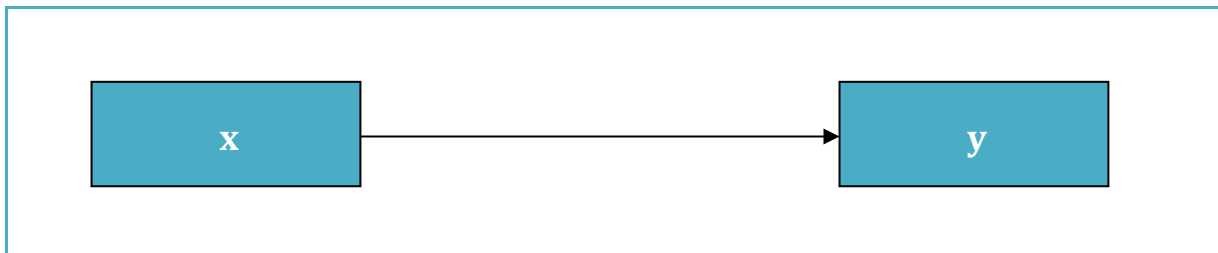
We often assume that the one variable has an “effect” on another variable. Here, we are talking about statistical effect, *not* causal effect. In other words, while we may find that one of the variables has a statistical effect on the other variable, it does not mean that the first variable *causes* the second variable. A phrase commonly used in statistics to capture this is: “correlation does not imply causation”.

## *X, Y and Z*

Variables play different roles in the analysis. Researchers often use various terms to distinguish between these roles. Here, we will try to shed some light on the terms that are used.

Variables	
<b>X</b>	Independent variable; Exposure; Predictor
<b>Y</b>	Dependent variable; Outcome
<b>Z</b>	Covariate; Confounder; Mediator; Moderator; Effect modifier

## 11.1 X and Y



If read about a variable being “independent”, an “exposure”, or a “predictor” – what does that mean? Basically, it means that you think that this variable has an (statistical) effect on another variable. For the sake of simplicity, let us just call this type of variable “x”. The other variable – the one that x is assumed to affect – is called “dependent” variable or “outcome”. Again, to make it simpler, we can call it “y”.

### Some examples

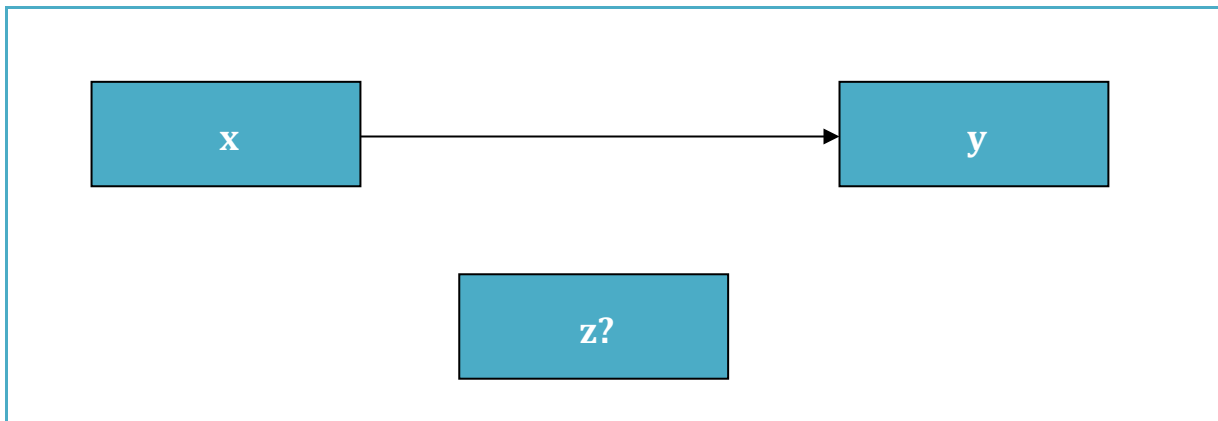
Smoking (x) -> Lung cancer (y)

Unemployment (x) -> Low income (y)

Yoga lessons (x) -> Lower stress levels (y)

The examples presented above may suggest that it is easy to know which variable is x and which is y, but this is not always the case. Sometimes we deal with more complex issues, such as the association between health and educational attainment: does a lower educational attainment (x) lead to worse health (y) or does poor health (x) result in lower educational attainment (y)? In cases like that you need to think about that is more reasonable, and what previous literature and theory would say about the issue at hand.

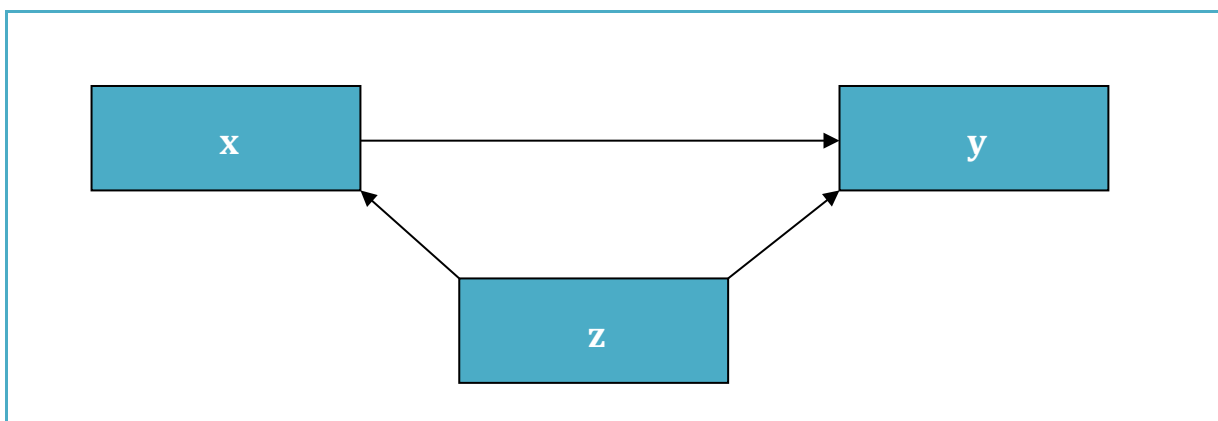
## 11.2 Z: confounding, mediating and moderating variables



The association – between x and y – that we are most interested in is often called “main association”. This is the focus of our analysis. However, sometimes there are other variables that we might find important for this main association. Strictly speaking, those variables are also called “x” (or “covariates”) but for clarity we will label them “z”. There are three important types of z-variables that are common in data analysis:

Types of “z”	
<b>Confounder</b>	Both x and y are affected by z
<b>Mediator</b>	A part of the association between x and y goes through z
<b>Moderator</b>	Z affects the association between x and y

### *Confounding variables*

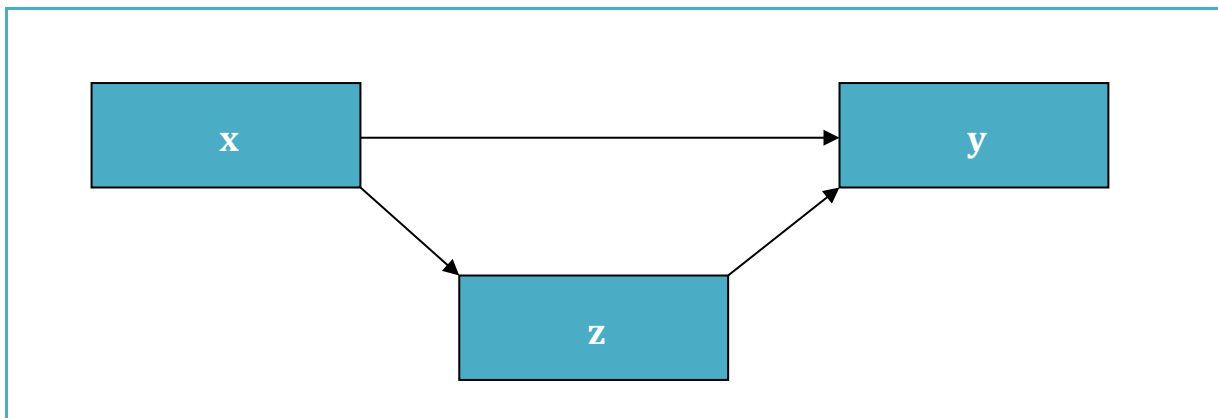


A confounder is a variable that influences both the x-variable and the y-variable and, thus, makes you think that there is an actual relationship between x and y (but it is due to z). Put differently,

the confounder distorts the analysis. Suppose that we find that people who consume a lot of coffee (x) have an increased risk of lung cancer (y). A probable confounder could be cigarette smoking (z): smokers drink more coffee and have greater risk of lung cancer.

In data analysis, we commonly want to get rid of the confounding effects – in that context, we often talk about “controlling” or “adjusting” for confounders.

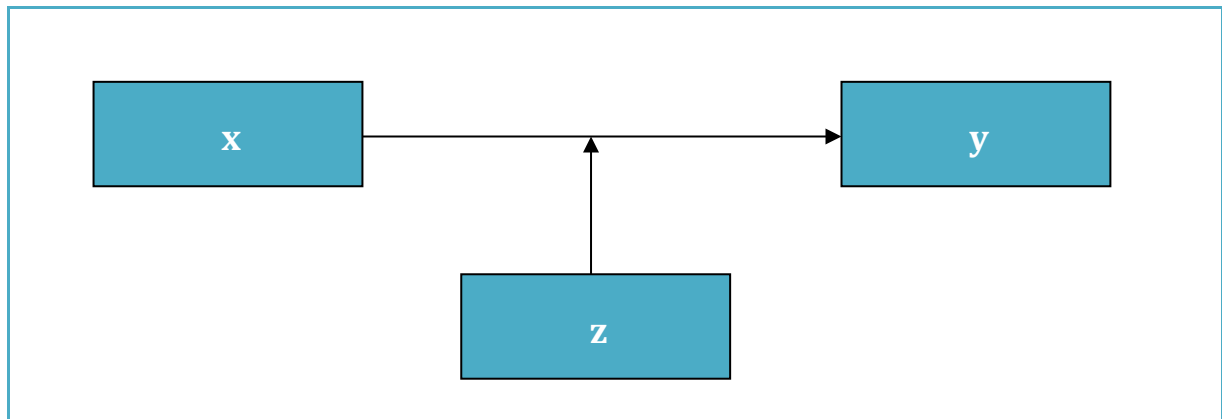
### *Mediating variables*



A mediator is a variable that is influenced by the x-variable and influences the y-variable. Thus, some (it could be a little or a lot) of the effect of x on y is mediated through z. For example, let us say that we are interested in the association between parents’ educational attainment (x) and children’s success on the labour market (y). It could be reasonable to assume that the educational attainment of the parents (x) influences children’s own educational attainment (z), which in turn affects their following success on the labour market (y).

In data analysis, we often talk about “explaining” an association by the inclusion of certain mediating variables. Particularly when one has a data material that consists of information collected across several points in time (i.e. longitudinal or life course data), it is common to talk about mediation as “pathways” or “mechanisms”. It should however be highlighted that mediation analysis has become increasingly criticized (for statistical reasons that we will not discuss here).

### *Moderating (or effect modifying) variables*



A moderator (or effect modifier) is a variable that influences the very association between the x-variable and the y-variable. Thus, the association between x and y looks different depending on the value of z. Suppose that we are interested in the association between unemployment (x) and ill-health (y). Here, it could be reasonably to assume that men's and women's health is affected differently by unemployment – in that case, gender would be a moderating variable (z).

In data analysis, moderating variables are examined through something called “interaction analysis” (see Chapter 17).

## 12. PREPERATIONS FOR REGRESSION ANALYSIS

Outline
12.1 What type of regression should be used?
12.2 Dummies
12.3 Standardization: z-scores
12.4 Analytical strategy
12.5 Missing data
12.6 From study sample to analytical sample

## 12.1 What type of regression should be used?

There are many different types of regression analysis. Some of the most common types are included in this guide: linear, logistic, ordinal regression, multinomial and Poisson. Which one you should choose depends on your outcome ( $y$ ).

Outcome ( $y$ )	Type of regression
Nominal with two categories, i.e. dichotomous (binary)	<b>Logistic regression</b>
Nominal with more than two categories, i.e. polytomous	<b>Multinomial regression</b>
Ordinal	<b>Ordinal regression</b>
Continuous (ratio/interval)	<b>Linear regression</b>

Just to underline it one more time: the type of regression you choose depends on your outcome ( $y$ ). However, your  $x$ -variable(s) can take on any form – they can be categorical (i.e. nominal/ordinal) or continuous (i.e. ratio/interval). If you include only one  $x$ -variable in your regression analysis, this is called “simple” (or “bivariate”) regression analysis. If you include two or more  $x$ -variables in your regression analysis, this is called “multiple” regression analysis. In multivariate regression analysis, it is possible to mix different types of  $x$ -variables: you can thus use both categorical and continuous  $x$ -variables.



## 12.2 Dummies

When we conduct regression analysis – regardless of the type – we can only include x-variables that are continuous (ratio/interval) or binary (i.e. they consist of only two values). A binary variable is sometimes called “dichotomous”, “binomial” or “dummy”. If we have a categorical variable with more than two values, such as in the example below, we need to “trick” the regression analysis to correctly analyse those variables. To do this, we create one dummy for each category of the variable:

	<i>Categories</i>	<i>Dummy</i>
Educational attainment	1=Compulsory	1=Compulsory, 0=Other
	2=Upper secondary	1=Upper secondary, 0=Other
	3=University	1=University, 0=Other

In the regression analysis, all dummies for the specific variable should be included as x-variables, *except one*. The dummy that you exclude – and it is your own choice which one you exclude – will be the “reference category”. The other dummies will be compared to the dummy that is excluded. Creating dummies is *much* easier using syntax than through the menus, but if you want to use the menus, you need to go through [Recode Into Different Variables](#) separately for each dummy.

### *A note on the choice of reference category*

There are many different ways of choosing a reference category:

Choosing a reference category
The largest category, because we want a stable group to compare the other categories to
The group in the middle, to represent the average
The “best off” category – if increasing values of the outcome is more negative
The “worst off” category – if increasing values of the outcome is more positive
Note: never choose a very small group – you may end up with very strange estimates!

## Dummies: Syntax

### For each dummy separately

```
RECODE VARNAME (OLD AND NEW VALUES) INTO DUMMYNAME.  
EXECUTE.
```

**VARNAME**

Insert the name of the original variable

For example:

`income`

**DUMMYNAME**

Insert the name of the dummy variable

For example:

`income_low`

or

`income_middle`

or

`education_high`

**(OLD AND NEW VALUES)**

Specify how you want to transform the values

Some examples:

`(1=1) (2=0) (3=0)`



### 12.3 Standardization: z-scores

The standard score – or the z-score – is very useful when we have continuous (ratio/interval) variables with different normal distributions. For example, if we have one variable called income (measured as annual household income in Swedish crowns) and another variable called years of schooling (measured as the total number of years spent in the educational system), these variables obviously have very different distributions. Suppose we want to compare which one – income or years of schooling – has a larger statistical effect on our outcome. That is not possible using the variables we have. The solution is to standardize (i.e. calculate z-scores for) these two variables so that they are comparable.

Z-scores are expressed in terms of standard deviations from the mean. What we do is that we take a variable and “rescale” it so that it has a mean of 0 and a standard deviation of 1. Each individual’s value on the standardized variable indicates its difference from the mean of the original (unstandardized) variable in number of standard deviations. A value of 1.5 would thus suggest that this individual has a value that is 1½ standard deviations above the mean, whereas a value of -2 would suggest that this individual has a value that is 2 standard deviations below the mean.

### *Standardization: z-scores: Function*

1. Go to the **Menu bar**, choose **Analyze\Descriptive Statistics\Descriptives**.
2. This will open up a new window called **Descriptives**.
3. A small window will open, where you see two boxes.
4. In the left box, all your variables are displayed. Here you choose the variable(s) you want to standardize.
5. Click on the small arrow between the boxes to transfer the variable(s) to the box to the right.
6. Tick the box called **Save standardized values as variables**.
7. Click on **OK** to generate a standardized version of the variable(s).

## *Standardization: z-scores: Syntax*

```
DESCRIPTIVES VARIABLES=VARNAME  
/SAVE.
```

**VARNAME**

Insert the name of the variable you want to use.

**/SAVE**

Insert this if you want to produce z-scores for your variable.

## Standardization: z-scores: Example

(Data: SPSS\_data1.sav)

```
DESCRIPTIVES VARIABLES=unempl_03 age  
/SAVE.
```

**unempl\_03**

Unemployment days in 2003 (ranges from 0 to 365)

**age**

Age in years (Min=18; Max=79)

## 12.4 Analytical strategy

Regression analysis is of course about data, but it is also about design. The way in which you think your variables are related needs to be translated into an “analytical strategy” (or “modelling strategy”). A good way to start is to make a drawing with boxes and arrows: each variable is put into one box and then you put simple-headed or double-headed arrows between the boxes to illustrate how the variables are associated to one another.

A good analytical strategy should reflect the aim of the study. Suppose we are interested in the association between children’s cognitive ability and educational attainment in adulthood. To examine this association is thus the aim of the study. We think that this association may be confounded by parents’ educational attainment and mediated by children’s school marks. Moreover, we suspect that the association may look different depending on the child’s gender. The research questions can thus be formulated as:

Question 1. Is children’s cognitive ability associated with educational attainment in adulthood?

Question 2. If so, is this association confounded by parents’ educational attainment?

Question 3. To what extent is the association between children’s cognitive ability and educational attainment in adulthood mediated by school marks in childhood?

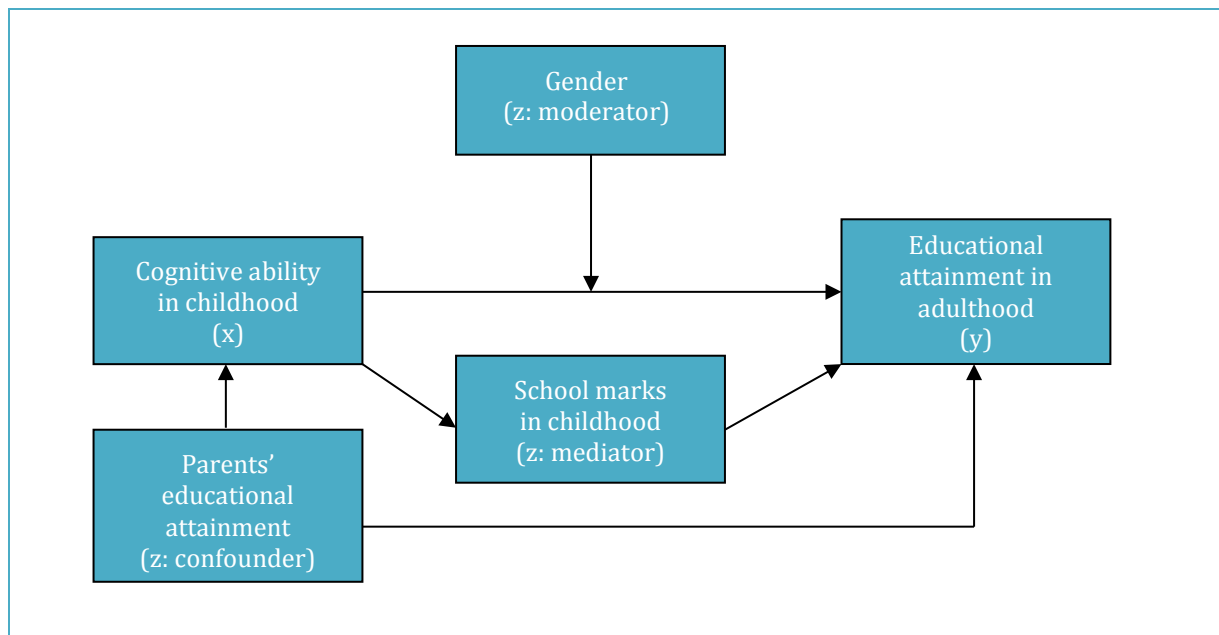
Question 4. Is there any gender difference in the association between children’s cognitive ability and educational attainment in adulthood?

Accordingly, these are the variables we need to include in our analysis:

x	Cognitive ability in childhood
y	Educational attainment in adulthood
z (confounder)	Parents’ educational attainment
z (mediator)	School marks in childhood
z (moderator)	Child’s gender



And this is how we may choose to illustrate our analytical strategy:



Often, we want to break down our analysis in different steps – or “regression models”. Each model commonly reflects one research question. In the present example, we would have a whole set of models that would include different variables:

- Model 1 Cognitive ability in childhood (x) and educational attainment in adulthood (y)
- Model 2 Cognitive ability in childhood (x), educational attainment in adulthood (y), and parents’ educational attainment (z)
- Model 3 Cognitive ability in childhood (x), educational attainment in adulthood (y), and school marks in childhood (z)
- Model 4 Cognitive ability in childhood (x), educational attainment in adulthood (y), and gender (z)

## 12.5 Missing data

As we discussed earlier (see Section 2.1), it is common that some individuals in our data material have missing information for one or more of the variables. Missing data is sometimes called “attrition” (particularly in register data) and sometimes “non-reponse” (particularly in survey/questionnaire data). Attrition or non-response can be external or internal:

External or internal?	
<b>External</b>	For some reason, the individuals are not included in the register data (they have immigrated, died, moved, are imprisoned, etc.) or do not participate in the survey (they decline, are too sick, cannot be reached, etc.).
<b>Internal</b>	For some reason, the individual has no information for a specific variable or a set of variables (they missed a page of the questionnaire, they refuse to answer specific questions, etc.).

As shown above, there are many reasons for missing data. If the missingness is problematic or not, depend on which type of missing data we have. In statistical analysis, there are three types of missing data:

Types of missing data	
<b>MCAR</b>	Missing Completely At Random: The probability of missing data is unrelated to both observed and unobserved data; it is completely by chance alone
<b>MAR</b>	Missing At Random: The probability of missing data is unrelated to unobserved data but may be related to observed data
<b>NMAR</b>	Missing Not At Random: The probability of missing data is related to unobserved data

This was probably a bit confusing – let us exemplify the differences between MCAR, MAR and NMAR. Suppose we examine the distribution of income in the Swedish population. If missing data were MCAR, it means that the missingness is unrelated to both observed data (e.g. gender, employment status) and unobserved data (e.g. lower income does not influence the risk of missingness). If missing data were MAR, it would mean that missingness could be related to other variables in the data set, but the probability of missingness is not increased by certain

values of the variable itself (e.g. individuals having lower incomes). Finally, if individuals who had certain values of the variables itself were more likely to be missing, we would have MNAR.

### *How may we deal with missing data?*

There are several ways of assessing whether missingness is MCAR, MAR or MNAR, but we are not going to get into advanced statistical stuff here. The most important advice is that you have to know your data well: produce descriptive statistics for your study variables to see the extent of missingness in the data material. Obviously, if you have a small number of individuals in your data material, a couple of missing values would have more serious consequences than if you have a couple of missing values in a data material based on the total population of a country.

A sound strategy to map out and illustrate potential problems with missingness is first to find out anything you can about the reasons for external attrition. Why are some individuals not included in your data set? Is it likely that they similar in any important way or is the missingness due to technical reasons? Then you get into the issue of internal attrition. Analysing internal attrition is simply called “attrition analysis” or “non-response analysis”. What you do here is to pick one or more variables for which all individuals in the study sample has information, such as gender, age, or some other socio-demographic variable. Produce descriptive statistics (choice of type of descriptive statistics depends on the measurement scale) for those variables, *for all individuals in the study sample*. Then you produce descriptive statistics for the same variables, but now *only for the individuals in the analytical sample* (Section 12.6 describes how to define an analytical sample).

For example, our study sample contains 10,000 individuals. Approximately 49 % are men and 51 % are women. The mean age is 45 years. Due to missing data on some of the variables we want to include in our analysis, our analytical sample is reduced to 9,451 individuals. In this sample, 46 % are men and 54 % are women. The mean age is 47 years. You can illustrate this in a simple descriptive table:

	Study sample (n=10,000)	Analytical sample (n=9,451)
<b>Gender</b>		
Man	49 %	46 %
Woman	51 %	54 %
<b>Age (mean)</b>	45 years	47 years

If we thus compare the distribution of gender and age in the study sample with the distribution of gender and age in the analytical sample, we can conclude that women and older individuals are more likely to be included in our analysis. This is information that could be important to have when we interpret our results.

## 12.6 From study sample to analytical sample

This section is an attempt to connect the two previous sections. It is like this: we often split our analysis in different steps or models. Thus, different models include different sets of variables; and different variables have different amount of missing data. The total number of individuals may thus vary across models, and this makes it difficult to compare the results between the models. In other words, we must ensure that all our analyses – and all steps of analysis – are based on the same individuals. These individuals represent our “analytical sample” (or “effective sample”). Put differently: our analytical sample is defined as only those individuals who have valid information (i.e. no missing) for all variables we use in our analysis.

It is good to first check the amount of missing data for each of the variables included in the analysis, to see if any certain variable is particularly problematic in terms of missingness. If a variable has serious problems with missingness, it could be wise to exclude it from the analysis (but it depends on how important the variable is to you).

The analytical sample should not only be the basis for regression analysis, but all other statistical tests and descriptive statistics should also be based on the analytical sample. Moreover, make sure to state the total number of individuals in the heading of each table and each figure. It could look something like this:

### Some examples

Table 1. Descriptive statistics for all study variables (n=9,451).

Figure 5. Histogram of annual income (n=9,451).

Table 3. The association between educational attainment and mortality. Results from logistic regression analysis, separately for men (n=4,701) and women (n=4,750).

### *The “pop” variable*

It is rather easy to define an analytical sample in SPSS. First, you need to determine exactly which variables are included in the analysis (i.e. all variable you *use*, not all variables in the data material). They should be properly recoded as you want them, and all missing values should be coded as actually missing in SPSS (**Recode**, see Section 5.1). Second, you create a “pop” variable – “pop” stands for population – through the **Compute** (see Section 5.2). Third, you make a frequency table of the pop variable. The individuals who have the highest value on this variable constitute your analytical sample. Finally, you use this pop variable to restrict everything you do from that point onwards, to include only those with the desired value on the pop variable (**Select cases**, see Section 5.3).

## *From study sample to analytical sample: Syntax*

```
COMPUTE NAME=nvalid(VARLIST).  
EXECUTE.
```

**NAME**                      Insert the name you want to give the variable indicating the analytical sample.

**VARLIST**                      Insert the names of all study variables, separate them by comma.  
For example:  
*gender, age, marital\_status, income*

Note: SPSS will calculate the number of valid values with this command. If you include a variable list with four variables, the pop variable will range between zero and four. Individuals with the value four will have valid information for all variables and thus constitute your analytical sample (and subsequently, you select only these individuals through the select-cases command). If you would have included nine variables, the pop variable will range between zero and nine. In that case, individuals with the value nine would have constituted your analytical sample.

*From study sample to analytical sample: Example*

(Data: SPSS\_data1.sav)

```
COMPUTE pop=nvalid(gender, age, marital, jobsat, unempl_03, unempl_05).  
EXECUTE.
```



## 13. LINEAR REGRESSION

Outline
13.1 Simple linear regression
13.2 Multiple linear regression
13.3 Model diagnostics

Before you begin, make sure that you have defined your analytical sample correctly (see Section 12.6).

## Introduction

Linear regression is used when  $y$  is continuous (ratio/interval; see Section 2.2). If you have only one  $x$ , it is called “simple” linear regression, and if you have more than one  $x$ , it is called “multiple” linear regression. Regardless of whether you are doing a simple or a multiple regression, the  $x$ -variables can be categorical (nominal/ordinal) and/or continuous (ratio/interval).

Key information from linear regression	
<b>1. Direction</b>	
Negative	Minus sign before the B coefficient
Positive	No minus sign before the B coefficient
<b>2. Effect size</b>	
B coefficient	How much does $y$ change for every one-unit increase in $x$ ?
<b>3. Statistical significance</b>	
P-value	$p < 0.05$ Statistically significant at the 5 % level $p < 0.01$ Statistically significant at the 1 % level $p < 0.001$ Statistically significant at the 0.1 % level
95 % Confidence intervals	Interval includes 0: Statistically significant at the 5 % level Interval does not include 0: Statistically non-significant at the 5 % level

## *B coefficients (B)*

A linear regression analysis describes the linear association between  $x$  and  $y$ . The effect that  $x$  has on  $y$  is estimated through a “Beta coefficient” – or “B coefficient”. The B coefficient is interpreted in the following way: “for every one-unit increase in  $x$ ,  $y$  increases/decreases by [the B coefficient]”. Accordingly, if you get a negative B coefficient (below 0), you say: “for every one-unit increase in  $x$ ,  $y$  decreases by [the B coefficient]”, and if you get a positive B coefficient (above 0), you say: “for every one-unit increase in  $x$ ,  $y$  increases by [the B coefficient]”. What the B coefficient actually stands for – and whether we can say that an effect is small or big – depends on the values of  $x$  and  $y$ .

### *P-values and confidence intervals*

In linear regression analysis you can of course get information about statistical significance, in terms of both p-values and confidence intervals. The p-values and the confidence intervals will give you partly different information, *but*: they are not contradictory. If the p-value is below 0.05, the 95 % confidence interval will not include 0 (statistical significance at the 5 % level), and if the p-value is above 0.05, the 95 % confidence interval will include 0 (statistical non-significance at the 5 % level).

Note that when you look at the p-value, you can rather easily distinguish between the significance levels (i.e. you can directly say whether you have statistical significance at the 5 % level, the 1 % level, or the 0.1 % level). When it comes to confidence intervals, SPSS will by default choose 95 % level confidence intervals (i.e. statistical significance at the 5 % level). For some analyses, it is however possible to change the confidence level for the intervals. For example, you may instruct SPSS to show 99 % confidence intervals instead.

### *R-Squared*

You also get information about something called “R-Squared” or “R<sup>2</sup>”. This term refers to amount of the variation in y that is explained by the inclusion of the x-variable. The R<sup>2</sup> value ranges between 0 and 1 – a higher value means a higher amount of explained variation. Generally speaking, the higher the R<sup>2</sup> values, the better the model fits the data (i.e. the model has better predictive ability).

### *Simple versus multiple regression models*

The difference between simple and multiple regression models, is that in a multiple regression each x-variable’s effect on y is estimated while taking into account the other x-variables’ effects on y. We then say that these other x-variables are “held constant”, or “adjusted for”, or “controlled for”. Because of this, multiple regression analysis is a way of dealing with the issue of “confounding” variables, and to some extent also “mediating” variables (see Section 11.2).

It is highly advisable to run a simple linear regression for each of the x-variables before including them in a multiple regression. Otherwise, you will not have anything to compare the adjusted B coefficients with (i.e. what happened to the B coefficients when other x-variables

were included in the analysis). Including multiple x-variables in the same model usually (but not always) means that the associations are reduced in strength – which would of course be expected if the x-variables overlapped in their effect on y.

### *Define your analytical sample*

Before you begin, make sure that you have defined your analytical sample correctly (see Section 12.6).

## 13.1 Simple linear regression

Quick facts	
<b>Number of variables</b>	One dependent (y) One independent (x)
<b>Scale of variable(s)</b>	Dependent: continuous (ratio/interval) and normally distributed Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

## *Simple linear regression: Function*

1. Go to the **Menu bar**, choose **Analyze\Regression\Linear**.
2. A new window called **Linear Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Then you choose the variable you want as your independent variable (x) and transfer it to the box called **Independent(s)**.
5. Click on **Statistics**.
6. Tick the box for **Confidence Intervals**.
7. Click on **Continue**.
8. Click on **OK** to get the results in your **Output** window.

## *Simple linear regression: Syntax*

```
REGRESSION  
/MISSING LISTWISE  
/STATISTICS COEFF OUTS CI(95) R ANOVA  
/CRITERIA=PIN(.05) POUT(.10)  
/NOORIGIN  
/DEPENDENT DEPVAR  
/METHOD=ENTER INDEPVAR.
```

**DEPVAR**                      Name of the dependent variable.

**INDEPVAR**                    Name of the independent variable.

## *Simple linear regression with a continuous x*

### **Example 1**

Suppose we want to examine the association between unemployment days ( $x$ ) and income ( $y$ ) by means of a simple linear regression analysis. Unemployment days are measured as the total number of days in unemployment during a year, and ranges from 0 to 365. Income is measured in thousands of Swedish crowns per month and ranges between 20 and 40 thousands of Swedish crowns. Let us say that we get a B coefficient that is -0.13. That would mean that for each one-unit increase in unemployment days, income would decrease by 0.13. Given the values of our variables, we can conclude that for each additional day in unemployment, monthly income would decrease by 130 SEK on average.

### **Example 2**

In another example, we may examine the association between intelligence scores ( $x$ ) and years of schooling ( $y$ ). Intelligence scores are measured by a series of tests that render various amounts of points, and ranges between 20 and 160 points. Years of schooling is measured as the total number of years spent in the educational system and ranges from 9 to 20 years. Here, we get a B coefficient that is 0.08. Given the values of our variables, we can conclude that for each additional point on the intelligence variable, the number of years spent in the educational system increases by 0.08 on average (corresponding to approximately one month).



## Simple linear regression with a continuous x: Example

(Data: SPSS\_data1.sav)

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS CI(95) R ANOVA

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT **hourstv**

/METHOD=ENTER **age**.

**hourstv**

Hours spent watching TV last week (Min=0; Max=36)

**age**

Age in years (Min=18; Max=79)

### Output/Step 1

The first step is to look at the table called **Model Summary**. Here you focus on the column for **Adjusted R Square**, which reflects the so-called “R<sup>2</sup>”. The value here reflects how much of the variation in the dependent variable (**hourstv**) that is explained by the variation in the independent variable (**age**). Just move the decimal two places to the right to be able to interpret the R<sup>2</sup> value as percentages. Accordingly, we see that age explains 0.1 % of the variation in hours spent watching TV (that is obviously very little).

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,036 <sup>a</sup>	,001	,001	5,171

a. Predictors: (Constant), Age in years

## Output/Step 2

The second step is to look at the table called **Coefficients**. Start focusing on the column called **B**; this is the B coefficient. The B coefficient in this example is -0.011, which first of all means we have a negative association between **age** and **hourstv**. Based on what we know about the values of these two variables, we can conclude the following: for every one year increase in age, the number of hours spent watching TV decreases by 0.011 hours.

The column called **Sig.** shows the p-value. It is 0.010, which means that the association between **age** and **hourstv** is statistically significant (at the 1 % level). Then we look at the part of the table called **95 % Confidence Interval for B**. This gives us the lower confidence limit (**Lower Bound**) and the upper confidence limit (**Upper Bound**). In the present example, the lower limit is -0.019 whereas the upper limit is -0.003. The interval does not include the null value (which is always  $x=0$  in linear regression) and, thus, the results are statistically significant (at the 5 % level).

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		
	B	Std. Error	Beta			Lower Bound	Upper Bound	
1	(Constant)	20,145	,207	97,194	,000	19,739	20,552	
	Age in years	-,011	,004	-,036	-,2574	,010	-,019	-,003

a. Dependent Variable: Hours spent watching TV last week

## *Simple linear regression with a binary x*

### **Example 1**

Suppose we want to examine the association between gender ( $x$ ) and income ( $y$ ) by means of a simple linear regression analysis. Gender has the values 0=Man and 1=Woman. Income is measured in thousands of Swedish crowns per month and ranges between 20 and 40 thousands of Swedish crowns. Let us say that we get a B coefficient that is -1.3. That would mean that for every one-unit increase in gender – i.e. changing from the category of men to the category of women – income would decrease by 1.3. Given the values of our variables, we can conclude that women have 1300 SEK less in monthly income compared to men.

### **Example 2**

Suppose we want to examine the association between having small children ( $x$ ) and the number of furry pets ( $y$ ) by means of a simple linear regression analysis. Having small children is measured as either 0=No or 1=Yes. The number of furry pets is measured as the number of cats, dogs or other furry animals living in the household, and ranges between 0 and 10. Let us say that we get a B coefficient that is 0.98. That would mean that for every one-unit increase in having small children – i.e. going from not having small children to having small children – the number of furry pets would increase by 0.98. Given the values of our variables, we can conclude that those who have small children have almost one more furry pet compared to those who do not have small children.

## Simple linear regression with a binary x: Example

(Data: SPSS\_data1.sav)

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS CI(95) R ANOVA

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT **hourstv**

/METHOD=ENTER **marital**.

**hourstv**

Hours spent watching TV last week (Min=0; Max=36)

**marital**

Marital status (0=Unmarried; 1=Married)

### Output/Step 1

The first step is to look at the table called **Model Summary**. Here you focus on the column for **Adjusted R Square**, which reflects the so-called “R2”. The value here reflects how much of the variation in the dependent variable (**hourstv**) that is explained by the variation in the independent variable (**marital**). Just move the decimal two places to the right to be able to interpret the R2 value as percentages. Accordingly, we see that marital status explains 0.1 % of the variation in hours spent watching TV (that is obviously not much).

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,039 <sup>a</sup>	,002	,001	5,171

a. Predictors: (Constant), Marital status

## Output/Step 2

The second step is to look at the table called **Coefficients**. Start focusing on the column called **B**; this is the B coefficient. The B coefficient in this example is 0.403, which first of all means we have a positive association between **marital** and **hourstv**. Based on what we know about the values of these two variables, we can conclude the following: the number of hours spent watching TV is 0.403 higher among those who are married compared to those who are unmarried.

The column called **Sig.** shows the p-value. It is 0.006, which means that the association between **marital** and **hourstv** is statistically significant (at the 1 % level). Then we look at the part of the table called **95 % Confidence Interval for B**. This gives us the lower confidence limit (**Lower Bound**) and the upper confidence limit (**Upper Bound**). In the present example, the lower limit is 0.116 and the upper limit is 0.690. The interval does not include the null value (which is always  $x=0$  in linear regression) and, thus, the results are statistically significant (at the 5 % level).

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		
	B	Std. Error	Beta			Lower Bound	Upper Bound	
1	(Constant)	19,449	,102	190,268	,000	19,249	19,650	
	Marital status	,403	,146	,039	2,753	,006	,116	,690

a. Dependent Variable: Hours spent watching TV last week

## *Simple linear regression with a categorical x (dummies)*

### **Example 1**

We want to investigate the association between educational attainment ( $x$ ) and income ( $y$ ) by means of a simple linear regression analysis. Educational attainment has the values: 1=Compulsory education, 2=Upper secondary education, and 3=University education. Income is measured in thousands of Swedish crowns per month and ranges between 20 and 40 thousands of Swedish crowns. Since our  $x$ -variable is categorical with more than two categories, we have to create dummies (one 0/1 coded variable for each category). We choose those with compulsory education as our reference category, meaning that this dummy will not be included in the analysis. We will thus get one  $B$  coefficient for upper secondary education and one  $B$  coefficient for university education. Each of these should be compared to the reference category. Let us say that we get a  $B$  coefficient for upper secondary education that is 2.1 and we get a  $B$  coefficient for university education that is 3.4. First of all, we can conclude – based on the direction of the estimates – that higher educational attainment is associated with higher income. More specifically, those with upper secondary education have (on average) 2100 SEK higher income compared to those with compulsory education, and those with university education have (on average) 3400 SEK higher income compared to those with compulsory education.



### Example 2

Suppose we are interested in the association between family type ( $x$ ) and children's average school marks ( $y$ ). Family type has three categories: 1=Two-parent household, 2=Joint custody, and 3=Single-parent household. Children's average school marks range from 1 to 5. Since our  $x$ -variable is categorical with more than two categories, we have to create dummies (one 0/1 coded variable for each category). We choose children living in a two-parent household as our reference category, meaning that this dummy will not be included in the analysis. We will thus get one B coefficient for joint custody and one B coefficient for single-parent household. Each of these should be compared to the reference category. The analysis results in a B coefficient of -0.1 for joint custody and a B coefficient of -0.9 for single-parent household. That would mean that children living in joint custody families have a 0.1 point lower score for average school marks compared to those living in two-parent households. Moreover, children living in single-parent households have a 0.9 point lower score for average school marks compared to those living in two-parent households.

## Simple linear regression with a categorical x (dummies): Example

(Data: SPSS\_data1.sav)

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS CI(95) R ANOVA

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT **income**

/METHOD=ENTER **agecat\_2534 agecat\_3549 agecat\_5064 agecat\_6579.**

<b>income</b>	Household income in thousands of dollars (Min=9; Max=1073)
<b>agecat_1824</b>	(0=No; 1=Yes)
<b>agecat_2534</b>	(0=No; 1=Yes)
<b>agecat_3549</b>	(0=No; 1=Yes)
<b>agecat_5064</b>	(0=No; 1=Yes)
<b>agecat_6579</b>	(0=No; 1=Yes)

### Output/Step 1

The first step is to look at the table called **Model Summary**. Here you focus on the column for **Adjusted R Square**, which reflects the so-called “R2”. The value here reflects how much of the variation in the dependent variable (**income**) that is explained by the dummies of the independent variable (**agecat**). Just move the decimal two places to the right to be able to interpret the R2 value as percentages. Accordingly, we see that the dummies of **agecat** explain 12.1 % of the variation in household income.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,349 <sup>a</sup>	,122	,121	52,06483

a. Predictors: (Constant), agecat\_6579, agecat\_2534, agecat\_5064, agecat\_3549

## Output/Step 2

The second step is to look at the table called **Coefficients**. Start focusing on the column called **B**; here we have the B coefficients. The dummy **agecat\_1824** was chosen as the reference category and is thus not included. The other B coefficients should be compared to that category. The B coefficient for **agecat\_2534** is 14.972 which means that individuals aged 25-34 have (on average) almost 15000 dollars more in household income compared to individuals in ages 18 to 24. The B coefficient for **agecat\_3549** is 35.909, meaning that individuals aged 35-49 have (on average) almost 36000 dollars more in household income compared to individuals in ages 18 to 24. The B coefficient for **agecat\_5064** is 60.033, which suggests that individuals aged 50-64 have (on average) roughly 60000 dollars more in household income compared to individuals in ages 18 to 24. Finally, the B coefficient for **agecat\_6579** is 27.094, which suggests that individuals aged 65-79 have (on average) roughly 27000 dollars more in household income compared to individuals in ages 18 to 24.

The column called **Sig.** shows the p-value for each category of **agecat**. All of them are 0.000, which means that the differences between the reference category and each of the remaining categories are statistically significant (at the 0.1 % level). Then we look at the part of the table called **95 % Confidence Interval for B**. This gives us the lower confidence limits (**Lower Bound**) and the upper confidence limits (**Upper Bound**). In the present example, the intervals do not include the null value (which is always  $x=0$  in linear regression) and, thus, the results are statistically significant (at the 5 % level).

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	23,355	2,086		11,196	,000	19,265	27,444
	agecat_2534	14,972	2,723	,103	5,499	,000	9,634	20,310
	agecat_3549	35,909	2,555	,280	14,054	,000	30,900	40,918
	agecat_5064	60,033	2,570	,462	23,357	,000	54,994	65,072
	agecat_6579	27,094	2,635	,198	10,281	,000	21,928	32,261

a. Dependent Variable: Household income in thousands

## 13.2 Multiple linear regression

Quick facts	
<b>Number of variables</b>	One dependent (y) At least two independent (x)
<b>Scale of variable(s)</b>	Dependent: continuous (ratio/interval) and normally distributed Independent: categorical (nominal/ordinal) and/or continuous (ratio/interval)

## *Multiple linear regression: Function*

1. Go to the **Menu bar**, choose **Analyze\Regression\Linear**.
2. A new window called **Linear Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Then you choose the variables you want as your independent variables (x) and transfer them to the box called **Independent(s)**.
5. Click on **Statistics**.
6. Tick the box for **Confidence Intervals**.
7. Click on **Continue**.
8. Click on **OK** to get the results in your **Output** window.

## *Multiple linear regression: Syntax*

```
REGRESSION  
/MISSING LISTWISE  
/STATISTICS COEFF OUTS CI(95) R ANOVA  
/CRITERIA=PIN(.05) POUT(.10)  
/NOORIGIN  
/DEPENDENT DEPVAR  
/METHOD=ENTER INDEPVAR.
```

**DEPVAR**                      Name of the dependent variable.

**INDEPVAR**                      List the names of the independent variables.

## Multiple linear regression

### Example

Suppose we are interested to see if the number of furry pets ( $y$ ) is related to having small children ( $x$ ), residential area ( $x$ ), and income ( $x$ ). The number of furry pets is measured as the number of cats, dogs or other furry animals living in the household, and ranges between 0 and 10. Having small children is measured as either 0=No or 1=Yes. Residential area has the values 1=Metropolitan, 2=Smaller city, and 3=Rural. Income is measured as the yearly household income from salary in thousands of SEK (ranges between 100 and 700 SEK).

When we do our linear regression, we can include the variables “having children” and “income” as they are, since the first is binary and the other is continuous. The remaining variable – residential area – is categorical with more than two values and therefore dummies must be used. We create one dummy for each category of “residential area”. The first dummy (metropolitan) is chosen as the reference category and is thus excluded from the analysis.

In the regression analysis, we get a B coefficient for having small children that is 0.51. That means that the number of furry pets is higher among those who have small children. This association is adjusted for residential area and income. With regards to residential area, we get a B coefficient for “smaller city” of 2.02 and the B coefficient for “rural” is 4.99. That suggests, firstly, that the number of furry pets is higher (about two more pets, on average) among individuals living in smaller cities compared to metropolitan areas. Secondly, the number of furry pets is much higher (almost five more pets, on average) among individuals living in rural areas compared to metropolitan areas. This association is adjusted for having small children and income. Finally, the B coefficient for income is -0.1. This suggests that for every one-unit increase in income (i.e. for every additional one thousand SEK), the number of furry pets decrease by 0.1. This association is adjusted for having small children and residential area.



## Multiple linear regression: Example

(Data: SPSS\_data1.sav)

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS CI(95) R ANOVA

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT **hourstv**

/METHOD=ENTER **income marital agecat\_2534 agecat\_3549 agecat\_5064 agecat\_6579.**

<b>hourstv</b>	Hours spent watching TV last week (Min=0; Max=36)
<b>income</b>	Household income in thousands of dollars (Min=9; Max=1073)
<b>marital</b>	Marital status (0=Unmarried; 1=Married)
<b>agecat_1824</b>	(0=No; 1=Yes) Reference category
<b>agecat_2534</b>	(0=No; 1=Yes)
<b>agecat_3549</b>	(0=No; 1=Yes)
<b>agecat_5064</b>	(0=No; 1=Yes)
<b>agecat_6579</b>	(0=No; 1=Yes)

### Output/Step 1

The first step is to look at the table called **Model Summary**. Here you focus on the column for **Adjusted R Square**, which reflects the so-called “R2”. The value here reflects how much of the variation in the dependent variable (**hourstv**) that is explained by the independent variables (**income**, **marital**, dummies of **agecat**). Just move the decimal two places to the right to be able to interpret the R2 value as percentages. Accordingly, we see that the independent variables explain 0.8 % of the variation in hours spent watching TV.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,096 <sup>a</sup>	,009	,008	5,153

a. Predictors: (Constant), agecat\_6579, Marital status, Household income in thousands of dollars, agecat\_2534, agecat\_3549, agecat\_5064

## Output/Step 2

The second step is to look at the table called **Coefficients**. Start focusing on the column called **B**; here we have the B coefficients. With regard to income, we can see that the B coefficient is 0.005. This means that for every one-unit increase in household income (i.e. every additional one thousands of dollars in household income), the number of hours watching TV last week increases by 0.005. The column called **Sig.** shows the p-value for income; it is 0.000, which means that the association between **income** and **hourstv** is statistically significant (at the 0.1 % level). Then we look at the part of the table called **95 % Confidence Interval for B**. This gives us the lower confidence limits (**Lower Bound**) and the upper confidence limits (**Upper Bound**). For income, the interval does not include the null value (which is always  $x=0$  in linear regression) and, thus, the results are statistically significant (at the 5 % level). The conclusion here is that there is a statistically significant association between income and TV watching – where higher income is related to more TV watching – also when marital status and age is adjusted for.

The B coefficient for marital status is 0.391, suggesting that married individuals watched more TV compared to unmarried individuals (we can conclude this since married have the higher value in the variable and hence the unmarried are automatically chosen as the reference category). The column called **Sig.** shows the p-value for marital status; it is 0.007, which means that the association between **marital** and **hourstv** is statistically significant (at the 1 % level). Then we look at the part of the table called **95 % Confidence Interval for B**. This gives us the lower confidence limits (**Lower Bound**) and the upper confidence limits (**Upper Bound**). For marital status, the interval does not include the null value (which is always  $x=0$  in linear regression) and, thus, the results are statistically significant (at the 5 % level). The conclusion here is that there is a statistically significant association between marital status and TV watching – where married individuals watch more TV than unmarried individuals – also when income and age is adjusted for.

Then we have one B coefficient for each of the dummies of age. The dummy **agecat\_1824** was chosen as the reference category and is thus not included. The other B coefficients should be compared to that category. The B coefficient for **agecat\_2534** is 0.889 which means that individuals aged 25-34 watch more TV compared to those aged 18-24. The B coefficient for **agecat\_3549** is 0.532 which means that individuals aged 35-49 watch more TV compared to those aged 18-24. The B coefficient for **agecat\_5064** is 0.136 which means that individuals aged 50-64 watch more TV compared to those aged 18-24. Finally, the B coefficient for **agecat\_6579** is -0.169 which means that individuals aged 65-79 watch less TV compared to those aged 18-24.

The column called **Sig.** shows the p-value for each category of **agecat**. The two first dummies have p-values of 0.001 and 0.039, which means that they are statistically significantly different (at the 0.1 % level and the 5 % level respectively) from the reference category. The two latter dummies have p-values greater than 0.05, which that there are no statistically significant differences from the reference category. Then we look at the part of the table called **95 % Confidence Interval for B**. This gives us the lower confidence limits (**Lower Bound**) and the upper confidence limits (**Upper Bound**). For **agecat**, the intervals do not include the null value (which is always  $x=0$  in linear regression) for the first two dummies and, thus, these results are statistically significant. For the other two dummies, the intervals do include the null value and, thus, these results are not statistically significant. The conclusion here is that there is a partly statistically significant association between age and TV watching: the individuals watching the most TV are those in ages 25-34, but also those in ages 35-49 watch a lot of TV. TV watching is less common among those below the age of 25 as well as 50 years or older. These results are adjusted for income and marital status.

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

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	18,888	,220		85,905	,000	18,457	19,319
	Household income in thousands of dollars	,005	,001	,055	3,634	,000	,002	,008
	Marital status	,391	,146	,038	2,680	,007	,105	,677
	agecat_2534	,889	,270	,066	3,290	,001	,359	1,419
	agecat_3549	,532	,258	,044	2,062	,039	,026	1,037
	agecat_5064	,136	,268	,011	,507	,612	-,390	,661
	agecat_6579	-,169	,264	-,013	-,642	,521	-,686	,348

a. Dependent Variable: Hours spent watching TV last week

### 13.3 Model diagnostics

First, it should be emphasised that a regression model generally has the aim to predict or “forecast” the value of  $y$ , depending on the values of the  $x$ -variables. Linear regression is concerned with finding the best-fitting straight line through the data points. Imagine that we make a scatterplot of two continuous variables; then a line is chosen so that it comes as close to all of the data points as possible. The best-fitting line is called a regression line.

The regression line has an “intercept” (or “constant”) and a “slope”. The intercept is where the regression line strikes the  $y$ -axis when the value of the  $x$ -variable(s) is 0. The slope is basically the steepness of the line; i.e. how much  $y$  changes when  $x$  increases.

The regression model thus gives us predicted values of  $y$  across the values of the  $x$ -variable(s). Of course, there is generally a difference between what the model predicts and what the individuals’ actual (observed) values are. This difference is called “residual” and is calculated as the observed value minus the predicted value.

Often, “error” is used instead of “residual”, and although these terms are closely related, they are not the exact same thing: an error is the difference between the observed value and the population mean (and the population mean is typically unobservable), whereas a residual is the difference between the observed value and the sample mean (and the sample mean is observable).

## Assumptions

Before we can trust the results from our linear regression analysis to be valid, we need to assess our model to check that it does not violate any of the fundamental assumptions of linear regression.

Checklist	
<b>No outliers</b>	Outliers are individuals who do not follow the overall pattern of data.
<b>Homoscedasticity</b>	The variance around the regression line should be constant across all values of the x-variable(s).
<b>Normality</b>	The residuals for our x-variables should be normally distributed.
<b>Linearity</b>	The effect of x on y should be linear.
<b>No multicollinearity</b>	Multicollinearity may occur when two or more x-variables that are included simultaneously in the model are strongly correlated with each another.

## Types of diagnostics

Model diagnostics	
<b>1. Scatterplot</b>	Check for linearity and outliers. Applies to continuous (ratio/interval) x-variables.
<b>2. Histogram</b>	Check for normality. The histogram is based on the residuals, not the actual values. Applies to both simple and multiple linear regression analysis.
<b>3. Residual plot</b>	Check for linearity and homoscedasticity. The plot is based on the residuals, not the actual values. Applies to both simple and multiple linear regression analysis.
<b>4. Normal probability plot</b>	Check for normality. Applies to both simple and multiple linear regression analysis.
<b>a. P-P plot</b>	Check for normality, based on actual values (better at detecting anomalies in the middle of the distribution).
<b>b. Q-Q plot</b>	Check for normality, based on cumulative probabilities (better at detecting anomalies at the tails of the distribution).
<b>5. Correlation analysis</b>	Check for multicollinearity. Applies to continuous (ratio/interval) and some ordinal x-variables.

Scatterplots are described elsewhere in this guide (see Section 4.7) and so is correlation analysis (Chapter 9). The remainder of the current section will be divided into two parts: the first will deal with diagnostics in terms of histograms, residual plots and P-P plots, whereas the second deals with Q-Q plots.

## *Histograms, residual plots and P-P plots: Function*

1. Go to the **Menu bar**, choose **Analyze\Regression\Linear**.
2. A new window called **Linear Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Then you choose the variables you want as your independent variables (x) and transfer them to the box called **Independent(s)**.
5. Click on **Statistics**.
6. Tick the box for **Confidence Intervals**.
7. Click on **Plots**.
8. In the left box, click on **\*ZRESID** and transfer it to the box called **Y:**
9. In the left box, click on **\*ZPRED** and transfer it to the box called **X:**
10. Tick the boxes called **Histogram** and **Normal probability plot**.
11. Click on **Continue**.
12. Click on **OK** to get the results in your **Output** window.



## *Histograms, residual plots and P-P plots: Syntax*

```
REGRESSION  
/MISSING LISTWISE  
/STATISTICS COEFF OUTS CI(95) R ANOVA  
/CRITERIA=PIN(.05) POUT(.10)  
/NOORIGIN  
/DEPENDENT DEPVAR  
/METHOD=ENTER INDEPVARs  
/SCATTERPLOT=(*ZRESID ,*ZPRED)  
/RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```

DEPVAR	Name of the dependent variable.
INDEPVAR	List the names of the independent variables.
/SCATTERPLOT= (*ZRESID ,*ZPRED)	Generates a residual plot.
/RESIDUALS HISTOGRAM (ZRESID) NORMPROB(ZRESID)	Generates a histogram and a P-P plot.

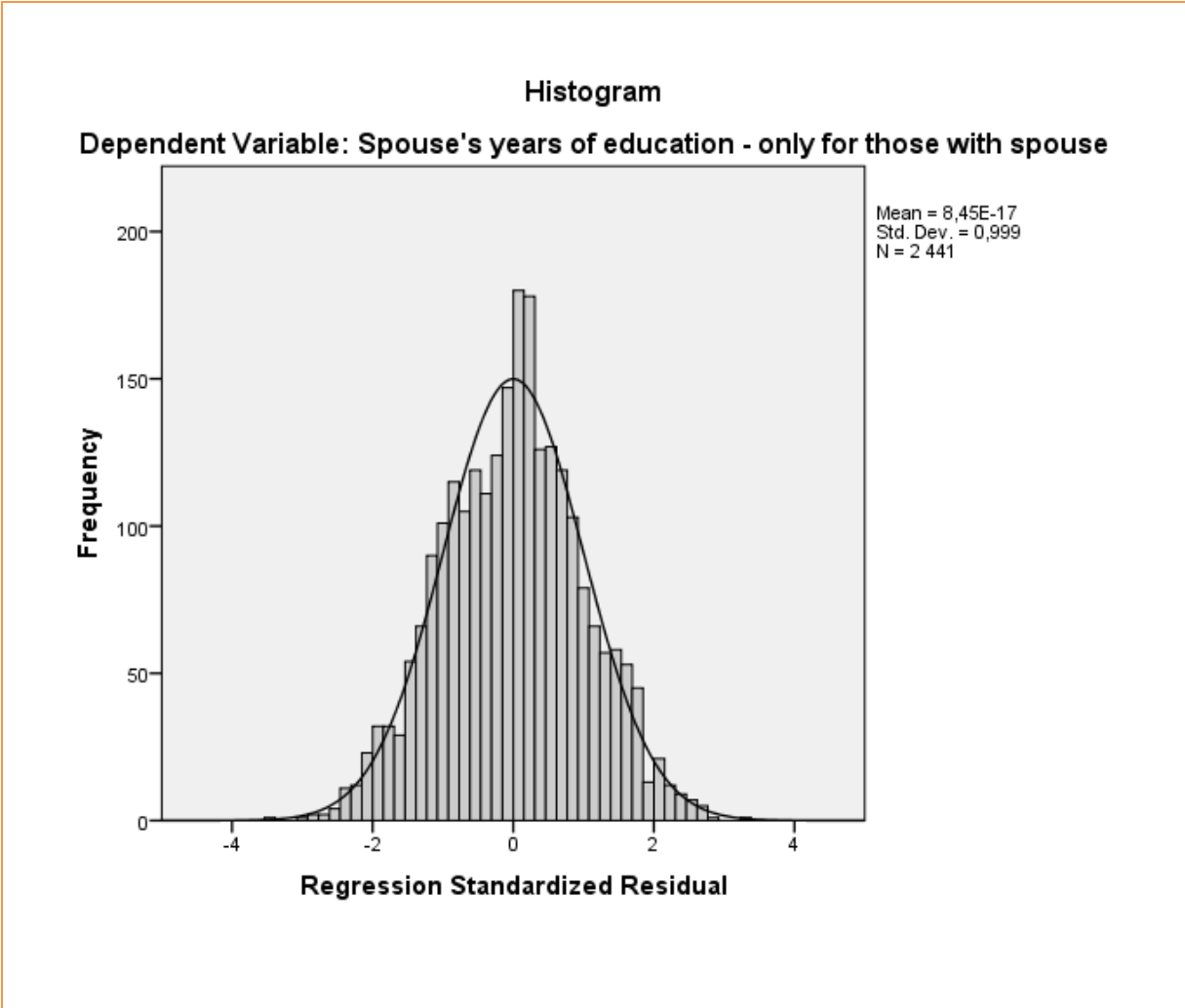
## Histograms, residual plots and P-P plots: Example

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT spoused2sel
/METHOD=ENTER gender ed hometype_singlefam hometype_townhouse
hometype_mobilehome
/SCATTERPLOT=(*ZRESID,*ZPRED)
/RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```

<b>spoused2sel</b>	Spouse's years of education, only for those with a spouse (Min=0; Max=24)
<b>gender</b>	Gender (0=Man; 1=Woman)
<b>hometype_singlefam</b>	Building type: single family (0=No; 1=Yes)
<b>hometype_multiplefam</b>	Building type: multiple family (0=No; 1=Yes) Reference category
<b>hometype_townhouse</b>	Building type: multiple family (0=No; 1=Yes)
<b>hometype_mobilehome</b>	Building type: mobile home (0=No; 1=Yes)

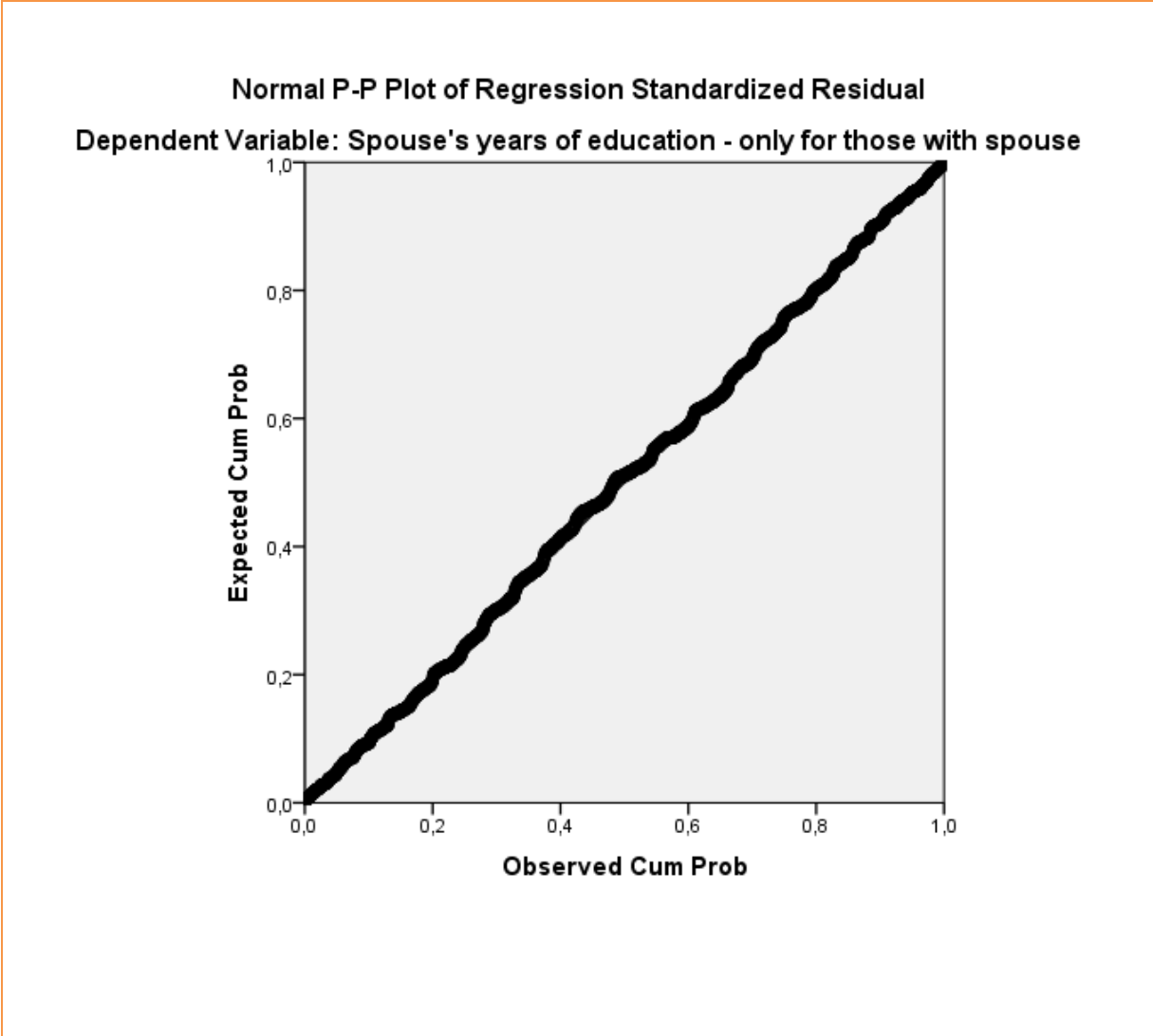
**Output/Step 1**

The first step is to look at the graph called **Histogram**. It shows how the residuals are distributed. They should follow the normal curve on the diagram. In the present example, this histogram looks rather normally distributed.



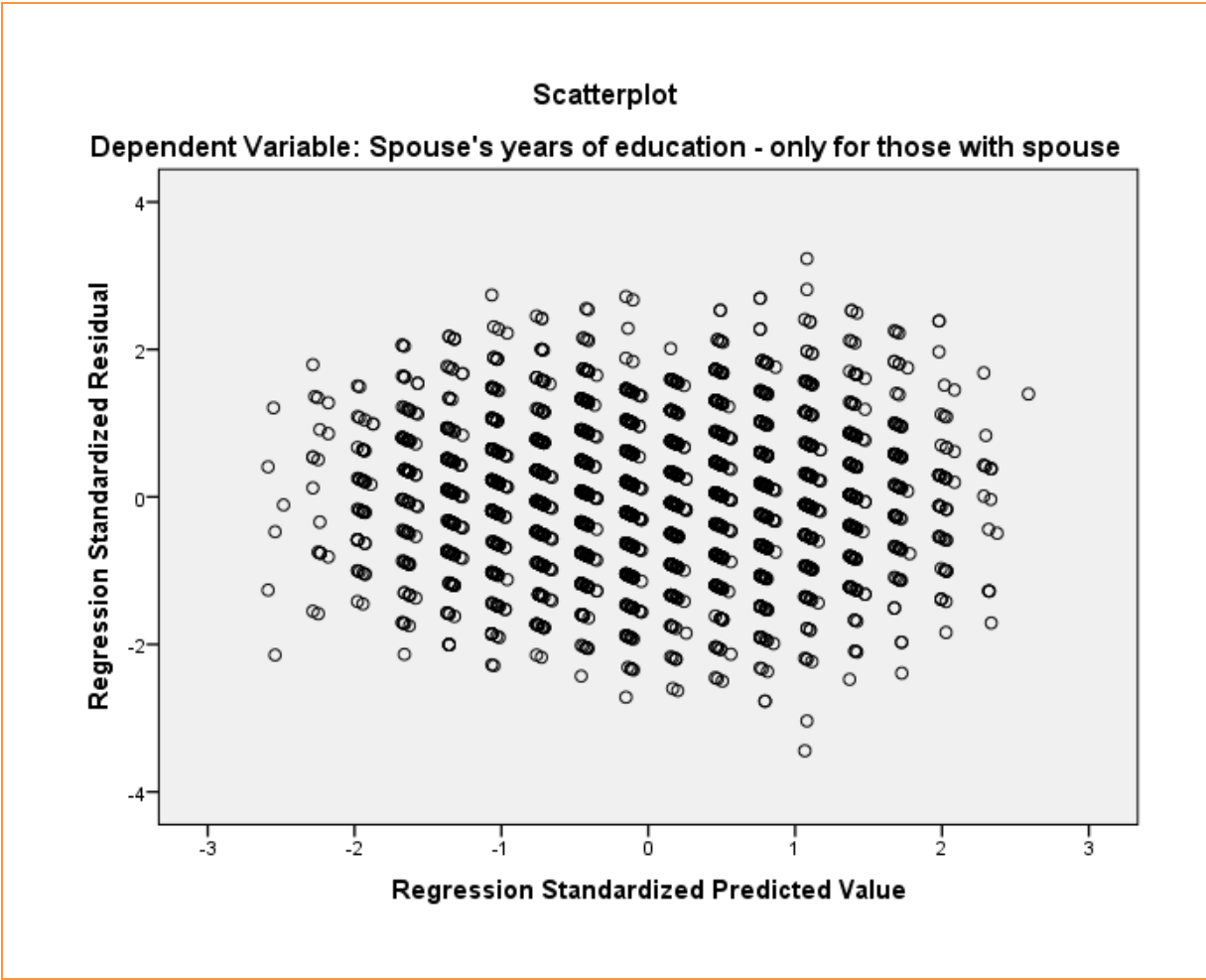
**Output/Step 2**

The second step is to look at the graph called **Normal P-P Plot of Regression Standardized Residual**. The distribution should follow a diagonal line across the plot if the residuals are normally distributed. The P-P plot below looks good; the residuals are normally distributed which was already indicated from the histogram.



**Output/Step 3**

The third step is to look at the graph called **Scatterplot**. The residuals should be equally distributed around 0. Systematic patterns usually indicate some problem with the model, like curve linearity or heteroscedasticity. Problematic patterns can look like cones or curves where the variation in the residuals is not constant over the whole range of values. For example, a cone-shaped form on the residual plot has low variation in the residuals on low values, and high variation in the residuals on high values, or reversed.



## Q-Q plots: Function

### Step 1

1. Go to the **Menu bar**, choose **Analyze\Regression\Linear**.
2. A new window called **Linear Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Then you choose the variables you want as your independent variables (x) and transfer them to the box called **Independent(s)**.
5. Click on **Statistics**.
6. Tick the box for **Confidence Intervals**.
7. Click on **Save**.
8. Under Residuals, tick the box called **Standardized**.
9. Click on **Continue**.
10. Click on **OK** to generate a new variable that consists of the standardized residuals from the regression analysis.

### Step 2

1. Go to the **Menu bar**, choose **Analyze\Descriptive Statistics\Q-Q Plots**.
2. In the left box, highlight the new variable you just generated, and transfer it to the box called **Variables**:
3. Click on **OK**.

## Q-Q plots: Syntax

### Part 1

```
REGRESSION  
/MISSING LISTWISE  
/STATISTICS COEFF OUTS CI(95) R ANOVA  
/CRITERIA=PIN(.05) POUT(.10)  
/NOORIGIN  
/DEPENDENT DEPVAR  
/METHOD=ENTER INDEPVAR  
/SAVE ZRESID.
```

### Part 2

```
PLOT  
/VARIABLES=VARNAME  
/NOLOG  
/NOSTANDARDIZE  
/TYPE=Q-Q  
/FRACTION=BLOM  
/TIES=MEAN  
/DIST=NORMAL.
```

<b>DEPVAR</b>	Name of the dependent variable.
<b>INDEPVAR</b>	List the names of the independent variables.
<b>/SAVE ZRESID</b>	Saves the standardized residuals.
<b>VARNAME</b>	Name of the variable containing the standardized residuals.

## Q-Q plots: Example

### Part 1

```
REGRESSION  
/MISSING LISTWISE  
/STATISTICS COEFF OUTS CI(95) R ANOVA  
/CRITERIA=PIN(.05) POUT(.10)  
/NOORIGIN  
/DEPENDENT spoused2sel  
/METHOD=ENTER gender ed hometype_singlefam hometype_townhouse  
hometype_mobilehome  
/SAVE ZRESID.
```

### Part 2

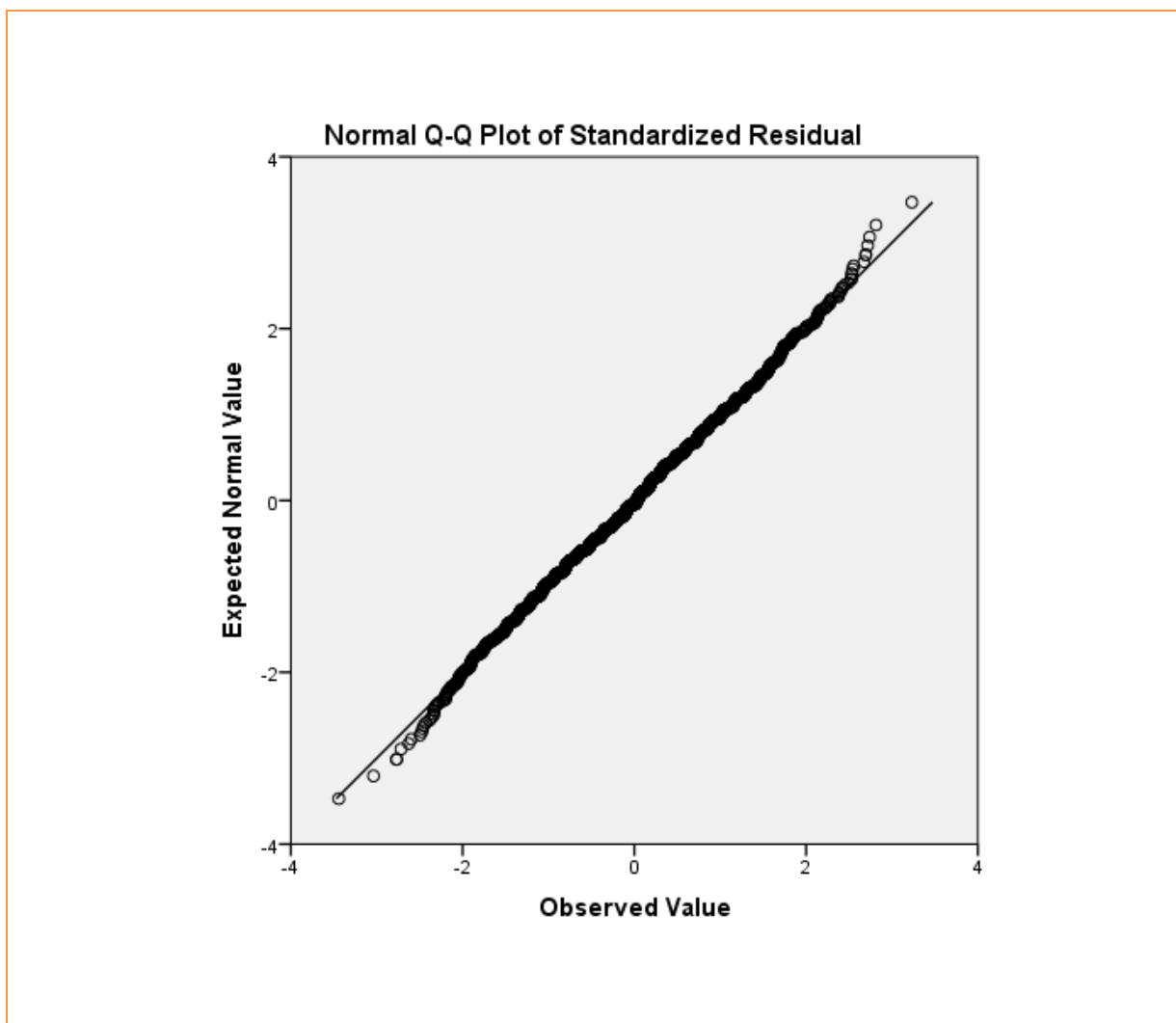
```
P PLOT  
/VARIABLES=ZRE_1  
/NOLOG  
/NOSTANDARDIZE  
/TYPE=Q-Q  
/FRACTION=BLOM  
/TIES=MEAN  
/DIST=NORMAL.
```

<i>spoused2sel</i>	Spouse's years of education, only for those with a spouse (Min=0; Max=24)
<i>gender</i>	Gender (0=Man; 1=Woman)
<i>hometype_singlefam</i>	Building type: single family (0=No; 1=Yes)
<i>hometype_multiplefam</i>	Building type: multiple family (0=No; 1=Yes) Reference category
<i>hometype_townhouse</i>	Building type: multiple family (0=No; 1=Yes)
<i>hometype_mobilehome</i>	Building type: mobile home (0=No; 1=Yes)



## Output

Look at the graph called **Normal Q-Q Plot of Standardized Residual**. If the observed values and the expected values coincide completely, the dots would follow the diagonal line and the variable tested would then be completely normally distributed. This is however very rare and some deviation is always present, but the principle is that the closer to the line the dots are the more normally distributed the variable is. For example the diagram above show a normally distributed variable, deviation at the ends is almost inevitable. It is more problematic if the dots are distributed in a wider s-shaped pattern and deviate from the diagonal over the whole range of values.



## 14. LOGISTIC REGRESSION

Outline
14.1 Simple logistic regression
14.2 Multiple logistic regression
14.3 Model diagnostics

## Introduction

Logistic regression is used when  $y$  is categorical with only two outcomes, i.e. dichotomous/binary (see section 2.2). If you have only one  $x$ , it is called “simple” logistic regression, and if you have more than one  $x$ , it is called “multiple” logistic regression. Regardless of whether you are doing a simple or a multiple regression, the  $x$ -variables can be categorical (nominal/ordinal) and/or continuous (ratio/interval).

Key information from logistic regression	
<b>1. Direction</b>	
Negative	Odds ratio below 1
Positive	Odds ratio above 1
<b>2. Effect size</b>	
Odds ratio	The odds of the outcome being a case divided by the odds that the outcome is a non-case, for every one-unit increase in $x$
<b>3. Statistical significance</b>	
P-value	$p < 0.05$ Statistically significant at the 5 % level $p < 0.01$ Statistically significant at the 1 % level $p < 0.001$ Statistically significant at the 0.1 % level
95 % Confidence intervals	Interval includes 1: Statistically significant at the 5 % level Interval does not include 1: Statistically non-significant at the 5 % level

## *Odds ratio (OR)*

A logistic regression is thus based on the fact that the outcome has only two possible values: 0 or 1. Often, 1 is used to denote a “case” whereas 0 is then a “non-case”. What a “case” or “non-case” means depends on how the hypothesis is formulated.

### **Example 1a**

We want to investigate the association between educational attainment ( $x$ ) and employment ( $y$ ). Our hypothesis is that educational attainment is positively associated with employment (i.e. higher educational attainment = more likely to be employed).

Coding of employment: 0=Unemployment (non-case); 1=Employment (case)

### **Example 1b**

We want to investigate the association between educational attainment ( $x$ ) and unemployment ( $y$ ). Our hypothesis is that educational attainment is negatively associated with unemployment (i.e. higher educational attainment = less likely to be unemployed).

Coding of employment: 0=Employment (non-case); 1=Unemployment (case)

Logistic regression is used to predict the “odds” of being a “case” based on the values of the  $x$ -variable(s). Just as for linear regression analysis, we get a coefficient (log odds) that shows the effect of  $x$  on  $y$ . However, because logistic regression is based on other assumptions than linear regression, we cannot interpret these coefficients very easily. Instead we focus on something called the “odds ratio” (“OR”). We can get the odds ratio by taking the “exponent” of the coefficient: “ $\exp(B)$ ”.

The odds ratio is interpreted in the following way: “for every one-unit increase in  $x$ ,  $y$  increases/decreases by [the odds ratio]”. Accordingly, if you get a negative OR (below 1), you say: “for every one-unit increase in  $x$ ,  $y$  decreases by [the OR]”, and if you get a positive OR (above 1), you say: “for every one-unit increase in  $x$ ,  $y$  increases by [the OR]”. Unlike linear regression, where the null value (i.e. value that denotes no difference) is 0, the null value for logistic regression is 1. Also note that an OR can never be negative – it can range between 0 and infinity. What the OR actually stands for – and whether we can say that an effect is small or big – depends on the values of  $x$  and  $y$ .

### *Is the odds ratio the same as risk, probability, chance or likelihood?*

The simple answer is no. These terms are *not* the same (but the more uncommon the outcome is, the closer odds ratios and risks become). Still, it is very common that odds ratios are interpreted in terms of risks, probabilities, chances or likelihoods. For example, suppose we examine the association between unemployment (0=Employed; 1=Unemployed) and alcohol abuse (0=No; 1=Yes) and we get an OR for unemployment that is 2.01. It is very tempting to interpret this as the unemployed having twice the risk of alcohol abuse compared to the employed. Or, if we investigate the association between school marks (ranges between 10 and 20, higher score=better marks) and continuation to university education (0=No; 1=Yes), and get an OR for school marks that is 1.20, many would say that the chance of university education increases by 20 % for every one-unit increase of school marks.

To avoid all of these problems, here is one suggestion: if you do not have to give any exact numbers, then it is all right to say that some individuals have higher or lower odds/odds ratio/risk/probability compared to other individuals. However, if you want to give exact numbers to exemplify, always use the actual OR.

#### **Some examples**

The results suggest that women (OR=0.84) are less likely than men to subscribe to a daily newspaper.

Based on logistic regression analysis, it may be concluded that individuals with more behavioural problems in childhood have a greater risk of drug abuse in adulthood (OR=1.49).

There is a negative association between educational attainment and number of children: the higher the educational attainment, the lower the number of children (OR=0.90).

Individuals living in urban areas (OR=0.33) are less likely compared to those living in rural areas to own a horse.

### *P-values and confidence intervals*

In logistic regression analysis you can of course get information about statistical significance, in terms of both p-values and confidence intervals. The p-values and the confidence intervals will give you partly different information, *but*: they are not contradictory. If the p-value is below

0.05, the 95 % confidence interval will not include 1 (statistical significance at the 5 % level), and if the p-value is above 0.05, the 95 % confidence interval will include 1 (statistical non-significance at the 5 % level).

Note that when you look at the p-value, you can rather easily distinguish between the significance levels (i.e. you can directly say whether you have statistical significance at the 5 % level, the 1 % level, or the 0.1 % level). When it comes to confidence intervals, SPSS will by default choose 95 % level confidence intervals (i.e. statistical significance at the 5 % level). For some analyses, it is however possible to change the confidence level for the intervals. For example, you may instruct SPSS to show 99 % confidence intervals instead.

### *R-Squared*

In contrast to linear regression, “R-Squared” or “R<sup>2</sup>” is not very usable (again, because of the assumptions behind logistic regression). You will, however, get a value for the so-called “Nagelkerke R Square” which is similar to the R-squared.

### *Simple versus multiple regression models*

The difference between simple and multiple regression models, is that in a multiple regression each x-variable’s effect on y is estimated while taking into account the other x-variables’ effects on y. We then say that these other x-variables are “held constant”, or “adjusted for”, or “controlled for”. Because of this, multiple regression analysis is a way of dealing with the issue of “confounding” variables, and to some extent also “mediating” variables (see Section 11.2).

It is highly advisable to run a simple logistic regression for each of the x-variables before including them in a multiple regression. Otherwise, you will not have anything to compare the adjusted odds ratios with (i.e. what happened to the OR when other x-variables were included in the analysis). Including multiple x-variables in the same model usually (but not always) means that the associations are reduced in strength – which would of course be expected if the x-variables overlapped in their effect on y.

### *Define your analytical sample*

Before you begin, make sure that you have defined your analytical sample correctly (see Section 12.6).

# 14.1 Simple logistic regression

Quick facts	
<b>Number of variables</b>	One dependent (y) One independent (x)
<b>Scale of variable(s)</b>	Dependent: binary Independent: categorical (nominal/ordinal) or continuous (ratio/interval)



## *Simple logistic regression: Function*

1. Go to the **Menu bar**, choose **Analyze\Regression\Binary Logistic**.
2. A new window called **Logistic Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Then you choose the variable you want as your independent variable (x) and transfer it to the box called **Covariates**.
5. Click on **Options**.
6. Tick the box for **CI for exp(B)**.
7. Click on **Continue**.
8. Click on **OK** to get the results in your **Output** window.

## *Simple logistic regression: Syntax*

```
LOGISTIC REGRESSION VARIABLES DEPVAR  
/METHOD=ENTER INDEPVAR  
/PRINT=CI(95)  
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

**DEPVAR**                      Name of the dependent variable.

**INDEPVAR**                    Name of the independent variable.

## *Simple logistic regression with a continuous x*

### **Example 1**

Suppose we want to examine the association between unemployment days ( $x$ ) and mortality ( $y$ ) by means of a simple logistic regression analysis. Unemployment days are measured as the total number of days in unemployment during a year, and ranges from 0 to 365. Mortality has the values 0=Alive and 1=Dead. Let us say that we get an OR that is 1.67. That would mean that we have a positive association: the higher the number of unemployment days, the higher the risk of dying.

### **Example 2**

In another example, we may examine the association between intelligence scores ( $x$ ) and drug use ( $y$ ). Intelligence scores are measured by a series of tests that render various amounts of points, and ranges between 20 and 160 points. Drug use has the values 0=No and 1=Yes. Here, we get an OR of 0.91. We can thus conclude that the risk of using drugs decrease for every one-unit increase in intelligence scores.

## Simple logistic regression with a continuous x: Example

(Data: SPSS\_data1.sav)

```
LOGISTIC REGRESSION VARIABLES active
```

```
/METHOD=ENTER age
```

```
/PRINT=CI(95)
```

```
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

active

Active lifestyle (0=No; 1=Yes)

age

Age in years (Min=18; Max=79)

## Output

Look at the table called **Variables in the Equation**. The column called **Exp(B)** shows the odds ratio (OR) for the variable **age**. The OR is 0.972, which means that we have a negative association between **age** and **active**. In other words, for every one-unit increase in age (i.e. one additional lived year), the likelihood of having an active lifestyle decreases.

The column called **Sig.** shows the p-value. Here, the p-value is 0.000 which means that the association between **age** and **active** is statistically significant (at the 0.1 % level). The part of the table called **95 % C.I. for EXP(B)** gives us the lower confidence limits (**Lower**) and the upper confidence limits (**Upper**). The interval does not include the null value (which is always  $x=1$  in logistic regression) and, thus, the results are statistically significant (at the 5 % level).

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 <sup>a</sup>								
age	-,029	,002	284,284	1	,000	,972	,969	,975
Constant	1,202	,084	205,740	1	,000	3,326		

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

## *Simple logistic regression with a binary x*

### **Example 1**

Suppose we want to examine the association between gender ( $x$ ) and alcohol abuse ( $y$ ) by means of a simple logistic regression analysis. Gender has the values 0=Man and 1=Woman, whereas alcohol abuse has the values 0=No and 1=Yes. Now, we get an OR of 0.66. This would mean that women are less likely to abuse alcohol compared to men.

### **Example 2**

Here we want to examine the association between having small children ( $x$ ) and owning a pet ( $y$ ) by means of a simple logistic regression analysis. Having small children is measured as either 0=No or 1=Yes. Owning a pet has the values 0=No and 1=Yes. Let us say that we get an OR that is 1.49. We can hereby conclude that it is more common to own a pet in families with small children compared to families without small children.

## Simple logistic regression with a binary x: Example

(Data: SPSS\_data1.sav)

```
LOGISTIC REGRESSION VARIABLES marital
```

```
/METHOD=ENTER active
```

```
/PRINT=CI(95)
```

```
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

marital

Marital status (0=Unmarried; 1=Married)

active

Active lifestyle (0=No; 1=Yes)

## Output

Look at the table called **Variables in the Equation**. The column called **Exp(B)** shows the odds ratio (OR) for the variable **active**. The OR is 0.987, which means that we have a negative association between **active** and **marital**. In other words, those who have an active lifestyle are less likely to be married compared to those who do not have an active lifestyle.

The column called **Sig.** shows the p-value. Here, the p-value is 0.823 which means that the association between **active** and **marital** is not statistically significant. The part of the table called **95 % C.I. for EXP(B)** gives us the lower confidence limits (**Lower**) and the upper confidence limits (**Upper**). The interval includes the null value (which is always  $x=1$  in logistic regression) and, thus, the results are not statistically significant.

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 <sup>a</sup> active	-,013	,057	,050	1	,823	,987	,884	1,103
Constant	-,041	,039	1,135	1	,287	,960		

a. Variable(s) entered on step 1: active.



## *Simple logistic regression with a categorical x (dummies)*

### **Example 1**

We want to investigate the association between educational attainment (x) and mortality (y) by means of a simple logistic regression analysis. Educational attainment has the values: 1=Compulsory education, 2=Upper secondary education, and 3=University education. Mortality has the values 0=No and 1=Yes. Since our x-variable is categorical with more than two categories, we have to create dummies (one 0/1 coded variable for each category). We choose those with compulsory education as our reference category, meaning that this dummy will not be included in the analysis. We will thus get one odds ratio for upper secondary education and one odds ratio for university education. Each of these should be compared to the reference category. Let us say that we get an OR for upper secondary education that is 0.82 and we get an OR for university education that is 0.69. We can thus conclude – based on the direction of the estimates – that higher educational attainment is associated with a lower mortality risk.

### **Example 2**

Suppose we are interested in the association between family type (x) and children's average school marks (y). Family type has three categories: 1=Two-parent household, 2=Joint custody, and 3=Single-parent household. Children's average school marks are categorised into 0=Above average and 1=Below average. Since our x-variable is categorical with more than two categories, we have to create dummies (one 0/1 coded variable for each category). We choose children living in a two-parent household as our reference category, meaning that this dummy will not be included in the analysis. We will thus get one odds ratio for joint custody and one odds ratio for single-parent household. Each of these should be compared to the reference category. The analysis results in an OR of 1.02 for joint custody and an OR of 1.55 for single-parent household. That would mean that children living in family types other than two-parent households are more likely to have school marks below average.

## Simple logistic regression with a categorical x (dummies): Example

(Data: SPSS\_data1.sav)

```
LOGISTIC REGRESSION VARIABLES ownpc
```

```
/METHOD=ENTER edcat_no edcat_somecoll edcat_colldeg edcat_postgrad
```

```
/PRINT=CI(95)
```

```
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

ownpc	Owns computer (0=No; 1=Yes)
edcat_no	Did not complete high school (0=No; 1=Yes)
edcat_highsc	High school degree (0=No; 1=Yes) Reference group
edcat_somecoll	Some college (0=No; 1=Yes)
edcat_colldeg	College degree (0=No; 1=Yes)
edcat_postgrad	Post-undergraduate degree (0=No; 1=Yes)

## Output

Look at the table called **Variables in the Equation**. The column called **Exp(B)** shows the odds ratios (OR) for the dummies of the variable **edcat**. The OR is 0.380 for the dummy **edcat\_no** (Did not complete high school), 2.253 for **edcat\_somecoll** (Some college), 4.864 for **edcat\_colldeg** (College degree), and 8.756 for **edcat\_postgrad** (Post-undergraduate degree). This means that those who do not complete high school are less likely to own a computer compared to those with a high school degree, whereas those who have some college education or more are more likely to own a computer compared to those with a high school degree. We can thus see a clear positive association between **edcat** and **ownpc**: the higher the educational attainment, the higher the likelihood of owning a computer.

The column called **Sig.** shows the p-value. Here, the p-values are 0.000 which means that the association between **edcat** and **ownpc** is statistically significant (at the 0.1 % level). The part of the table called **95 % C.I. for EXP(B)** gives us the lower confidence limits (**Lower**) and the upper confidence limits (**Upper**). The intervals do not include the null value (which is always  $x=1$  in logistic regression) and, thus, the results are statistically significant (at the 5 % level).

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 <sup>a</sup>								
edcat_no	-,968	,086	126,260	1	,000	,380	,321	,450
edcat_somecoll	,812	,088	86,054	1	,000	2,253	1,897	2,674
edcat_colldeg	1,582	,099	254,469	1	,000	4,864	4,004	5,907
edcat_postgrad	2,170	,195	124,347	1	,000	8,756	5,980	12,821
Constant	,195	,051	14,853	1	,000	1,216		

a. Variable(s) entered on step 1: edcat\_no, edcat\_somecoll, edcat\_colldeg, edcat\_postgrad.

## 14.2 Multiple logistic regression

Quick facts	
<b>Number of variables</b>	One dependent (y) At least two independent (x)
<b>Scale of variable(s)</b>	Dependent: binary Independent: categorical (nominal/ordinal) and/or continuous (ratio/interval)

## *Multiple logistic regression: Function*

1. Go to the **Menu bar**, choose **Analyze\Regression\Binary Logistic**.
2. A new window called **Logistic Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Then you choose the variables you want as your independent variables (x) and transfer them to the box called **Covariates**.
5. Click on **Options**.
6. Tick the box for **CI for exp(B)**.
7. Click on **Continue**.
8. Click on **OK** to get the results in your **Output** window.

## *Multiple logistic regression: Syntax*

```
LOGISTIC REGRESSION VARIABLES DEPVAR  
/METHOD=ENTER INDEPVAR  
/PRINT=CI(95)  
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

**DEPVAR**                      Name of the dependent variable.

**INDEPVAR**                    List the names of the independent variables.

## *Multiple logistic regression*

### **Example**

Suppose we are interested to see if having a pet (y) is related to having small children (x), residential area (x), and income (x). Having a pet has the values 0=No and 1=Yes. Having small children is measured as either 0=No or 1=Yes. Residential area has the values 1=Metropolitan, 2=Smaller city, and 3=Rural. Income is measured as the yearly household income from salary in thousands of SEK (ranges between 100 and 700 SEK).

When we do our logistic regression, we can include the variables “having children” and “income” as they are, since the first is binary and the other is continuous. The remaining variable – residential area – is categorical with more than two values and therefore dummies must be used. We create one dummy for each category of “residential area”. The first dummy (metropolitan) is chosen as the reference category and is thus excluded from the analysis.

In the regression analysis, we get an OR for having small children that is 1.30. That means that those who have small children are more likely to also have a pet. This association is adjusted for residential area and income. With regards to residential area, we get an OR for “smaller city” of 1.78 and the OR for “rural” is 4.03. This suggests that those who live in a smaller city are more likely to have a pet, and so are those living in rural areas. These results are adjusted for having small children and income. Finally, the OR for income is 0.93. This suggests that for every one-unit increase in income (i.e. for every additional one thousand SEK), the likelihood of having a small pet decreases. This association is adjusted for having small children and residential area.

## Multiple logistic regression: Example

(Data: SPSS\_data1.sav)

LOGISTIC REGRESSION VARIABLES **active**

/METHOD=ENTER **age marital bfast\_oatmeal bfast\_cereal**

/PRINT=CI(95)

/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).

<b>active</b>	Active lifestyle (0=No; 1=Yes)
<b>age</b>	Age in years (Min=18; Max=79)
<b>marital</b>	Marital status (0=Unmarried; 1=Married)
<b>bfast_energy</b>	Preferred breakfast: Energy bar (0=No; 1=Yes) Reference group
<b>bfast_oatmeal</b>	Preferred breakfast: Oatmeal (0=No; 1=Yes)
<b>bfast_cereal</b>	Preferred breakfast: Cereal (0=No; 1=Yes)



## Output

Look at the table called **Variables in the Equation**. The column called **Exp(B)** shows the odds ratios (OR) for the variables. The OR is 0.981 for **age**. This means that for every one-unit increase in age (i.e. one additional year of living), the odds of having an active lifestyle decreases. This association is adjusted for **marital** and **bfast**. The OR is 1.153 for **marital**, meaning that those who are married are more likely to have an active lifestyle compared to those who are married. This association is adjusted for **age** and **bfast**. With regard to **bfast**, we have included two dummies in the model (**bfast\_energy** is the reference category). The OR for **bfast\_oatmeal** is 0.349 and the OR for **bfast\_cereal** is 0.362. This means that those who prefer oatmeal or cereal for breakfast are less likely to have an active lifestyle compared to those who prefer an energy bar. These results are adjusted for **age** and **marital**. The column called **Sig.** shows the p-value. Here, the p-value for **age** and **bfast** is 0.000 which means that we have statistically significant associations with active lifestyle (at the 0.1 % level). The p-value for **marital** is 0.020, suggesting that also marital status has a statistically significant association with active lifestyle (at the 5 % level). Remember that all these association are mutually adjusted. The part of the table called **95 % C.I. for EXP(B)** gives us the lower confidence limits (**Lower**) and the upper confidence limits (**Upper**). None of the intervals include the null value (which is always  $x=1$  in logistic regression) and, thus, the results are statistically significant (at the 5 % level).

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 <sup>a</sup>								
age	-,019	,002	82,675	1	,000	,981	,977	,985
marital	,142	,061	5,453	1	,020	1,153	1,023	1,299
bfast_oatmeal	-1,052	,092	131,821	1	,000	,349	,292	,418
bfast_cereal	-1,016	,074	189,880	1	,000	,362	,313	,418
Constant	1,393	,100	195,228	1	,000	4,028		

a. Variable(s) entered on step 1: age, marital, bfast\_oatmeal, bfast\_cereal.

## 14.3 Model diagnostics

### *Goodness of fit*

The assumptions behind logistic regression are different from linear regression. For example, we do not need to assume linearity, homoscedasticity or normality. Multicollinearity should however be avoided (i.e. strong correlations between the x-variables in multiple logistic regression – check with correlation analysis; see Chapter 9). Most importantly, the model should fit the data. There are several tests to determine “goodness of fit” or, put differently, if the estimated model (i.e. the model with one or more x-variables) predicts the outcome better than the null model (i.e. a model without any x-variables). Below, some of these tests are discussed: classification tables, the Hosmer and Lemeshow test and the ROC curve.

Estimate the goodness of fit
<b>Classification tables</b>
<b>The Hosmer and Lemeshow test</b>
<b>ROC curve</b>

Before going into the specific tests, we need to address the issues of “sensitivity” and “specificity”. By comparing the cases and non-cases predicted by the model with the cases and non-cases actually present in the outcome, we can draw a conclusion about the proportion of correctly predicted cases (sensitivity) and the proportion of correctly classified non-cases (specificity).

Sensitivity and specificity			
		<u>Estimated model</u>	
		<b>Non-case</b>	<b>Case</b>
<u>“Truth”</u>	<b>Non-case</b>	<i>True negative</i>	<i>False positive</i>
	<b>Case</b>	<i>False negative</i>	<i>True positive</i>

## *Classification tables*

A classification table is similar to the table about sensitivity and specificity, only that it is based on the model that you run, and thus gives you the frequency of true negatives, false positives, false negatives, and true positives, as well as the overall percentages of cases and non-cases that are correctly predicted by the estimated model. Note that a classification table is automatically produced by SPSS and appears in the standard output.

## *The Hosmer and Lemeshow test*

This test is a type of a chi-square test. It indicates the extent to which the estimated model provides a better fit to the data (i.e. better predictive power) than the null model. The test will produce a p-value: if the p-value is above 0.05 (statistically non-significant) the estimated model has adequate fit, and if the p-value is below 0.05 (statistically significant) the estimated model does not adequately fit the data.

## *ROC curve*

The ROC curve is a graph that shows how well the estimated model predicts cases (sensitivity) and non-cases (specificity). What we are interested in here is the “area under the curve” (AUC). The AUC ranges between 0.5 and 1.0. The nearer the AUC is to 1, the better the predictive power. On the other hand, a value of 0.5 suggests that we may just flip a coin to decide on whether the outcome is a case or non-case. Here are some commonly used cut-off points when it comes to AUC:

Area under the curve (AUC)	
<b>0.5-0.6</b>	Fail
<b>0.6-0.7</b>	Poor
<b>0.7-0.8</b>	Fair
<b>0.8-0.9</b>	Good
<b>0.9-1.0</b>	Excellent

### *Some remarks on model fit*

A general comment about model fit: if the main interest was to identify the best model to predict a certain outcome, that would solely guide which x-variables we put into the analysis. For example, we would exclude x-variables that do not contribute to the model's predictive ability. However, research is often guided by theory and by the interest of examining associations between variables. If we thus have good theoretical reasons for keeping a certain x-variable or sticking to a certain model, we should most likely do that (but still, the model should not fit the data horribly). Model diagnostics will then be a way of showing others the potential problems with the model we use.

## *Classification tables: Function*

1. Go to the **Menu bar**, choose **Analyze\Regression\Binary Logistic**.
2. A new window called **Logistic Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Then you choose the variables you want as your independent variables (x) and transfer them to the box called **Covariates**.
5. Click on **Options**.
6. Tick the box for **CI for exp(B)**.
7. Click on **Continue**.
8. Click on **OK** to get the results in your **Output** window.

## *Classification tables: Syntax*

```
LOGISTIC REGRESSION VARIABLES DEPVAR  
/METHOD=ENTER INDEPVAR  
/PRINT=CI(95)  
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

**DEPVAR**                      Name of the dependent variable.

**INDEPVAR**                      List the names of the independent variables.

## Classification tables: Example

(Data: SPSS\_data1.sav)

LOGISTIC REGRESSION VARIABLES **active**

/METHOD=ENTER **age marital bfast\_oatmeal bfast\_cereal**

/PRINT=CI(95)

/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).

<b>active</b>	Active lifestyle (0=No; 1=Yes)
<b>age</b>	Age in years (Min=18; Max=79)
<b>marital</b>	Marital status (0=Unmarried; 1=Married)
<b>bfast_energy</b>	Preferred breakfast: Energy bar (0=No; 1=Yes) Reference group
<b>bfast_oatmeal</b>	Preferred breakfast: Oatmeal (0=No; 1=Yes)
<b>bfast_cereal</b>	Preferred breakfast: Cereal (0=No; 1=Yes)

### Output/Step 1

Look at the table called **Classification Table** (note: not the first but the second one; below **Block 1: Method = Enter**). The combination No+No (n=2053) contains the true negative, the combination No+Yes (n=612) contains the false positive, the combination Yes+No (n=1134) contains the false negative, and the combination Yes+Yes (n=1201) contains the true positive. The overall percentage of cases and non-cases that is correctly classified by the estimated model is 65.1 %.

Observed			Predicted		
			Active lifestyle		Percentage Correct
			No	Yes	
Step 1	Active lifestyle	No	2053	612	77,0
		Yes	1134	1201	51,4
Overall Percentage					65,1

a. The cut value is ,500

### Output/Step 2

The next step is to look at the other table called **Classification Table** (note: not the second but the first one; below **Block 0: Method = Enter**). The overall percentage of cases and non-cases that is correctly classified by the null model is 53.3 %. In other words, the estimated model did a better job of predicting the outcome than the null model.

Observed			Predicted		
			Active lifestyle		Percentage Correct
			No	Yes	
Step 0	Active lifestyle	No	2665	0	100,0
		Yes	2335	0	,0
Overall Percentage					53,3

a. Constant is included in the model.  
b. The cut value is ,500



### *The Hosmer and Lemeshow test: Function*

1. Go to the **Menu bar**, choose **Analyze\Regression\Binary Logistic**.
2. A new window called **Logistic Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Then you choose the variables you want as your independent variables (x) and transfer them to the box called **Covariates**.
5. Click on **Options**.
6. Tick the box for **CI for exp(B)**.
7. Click on **Continue**.
8. Click on **OK** to get the results in your **Output** window.

## *The Hosmer and Lemeshow test: Syntax*

```
LOGISTIC REGRESSION VARIABLES DEPVAR  
/METHOD=ENTER INDEPVAR  
/PRINT=GOODFIT CI(95)  
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

<b>DEPVAR</b>	Name of the dependent variable.
<b>INDEPVAR</b>	List the names of the independent variables.
<b>GOODFIT</b>	Include the Hosmer-Lemeshow goodness-of-fit test

## The Hosmer and Lemeshow test: Example

(Data: SPSS\_data1.sav)

LOGISTIC REGRESSION VARIABLES **active**

/METHOD=ENTER **age marital bfast\_oatmeal bfast\_cereal**

/PRINT=GOODFIT CI(95)

/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).

<b>active</b>	Active lifestyle (0=No; 1=Yes)
<b>age</b>	Age in years (Min=18; Max=79)
<b>marital</b>	Marital status (0=Unmarried; 1=Married)
<b>bfast_energy</b>	Preferred breakfast: Energy bar (0=No; 1=Yes) Reference group
<b>bfast_oatmeal</b>	Preferred breakfast: Oatmeal (0=No; 1=Yes)
<b>bfast_cereal</b>	Preferred breakfast: Cereal (0=No; 1=Yes)

## Output

Look at the table called **Hosmer and Lemeshow test**. The column that is labelled **Sig.** shows the p-value for the test. If the p-value is above 0.05 (statistically non-significant) the estimated model has adequate fit, and if the p-value is below 0.05 (statistically significant) the estimated model does not adequately fit the data. In the current example, we have a p-value of 0.130. This suggests that the estimated model has adequate fit.

**Hosmer and Lemeshow Test**

Step	Chi-square	df	Sig.
1	12,510	8	,130

## ROC curve: Function

### Step 1

1. Go to the **Menu bar**, choose **Analyze\Regression\Binary Logistic**.
2. A new window called **Logistic Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Then you choose the variables you want as your independent variables (x) and transfer them to the box called **Covariates**.
5. Click on **Save**.
6. Tick the box for **Probabilities**.
7. Click on **Continue**.
8. Click on **OK** to generate a new variable that contains the predicted probabilities.

### Step 2

1. Go to the **Menu bar**, choose **Analyze\ROC Curve**.
2. A new window called **ROC Curve** will open.
3. In the left box, all your variables are displayed. You choose the variable you just generated (the one with the predicted probabilities) and transfer it to the box called **Test Variable**.
4. Choose your dependent variable (y) and transfer it to the box called **State Variable**.
5. In the box called **Value of State Variable**, write the value that signifies a case (commonly a non-case has the value 0 and a case has the value 1; then write 1).
6. Tick the box called **With diagonal reference line**.
7. Click on **OK** to get the results in your **Output** window.

## ROC curve: Syntax

### Step 1

```
LOGISTIC REGRESSION VARIABLES DEPVAR  
/METHOD=ENTER INDEPVAR  
/SAVE=PRED  
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

### Step 2

```
ROC PRED BY DEPVAR (CASEVALUE)  
/PLOT=CURVE(REFERENCE)  
/CRITERIA=CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)  
/MISSING=EXCLUDE.
```

<b>DEPVAR</b>	Name of the dependent variable.
<b>INDEPVAR</b>	List the names of the independent variables.
<b>/SAVE=PRED</b>	Generate a new variable containing the predicted probabilities
<b>PRED</b>	Name of the variable containing the predicted probabilities
<b>(CASEVALUE)</b>	The value of the dependent variable that signifies a “case”

## ROC curve: Example

(Data: SPSS\_data1.sav)

### Part 1

LOGISTIC REGRESSION VARIABLES **active**

/METHOD=ENTER **age marital bfast\_oatmeal bfast\_cereal**

/SAVE=PRED

/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).

### Part 2

ROC **PRE\_1** BY **active** (1)

/PLOT=CURVE(REFERENCE)

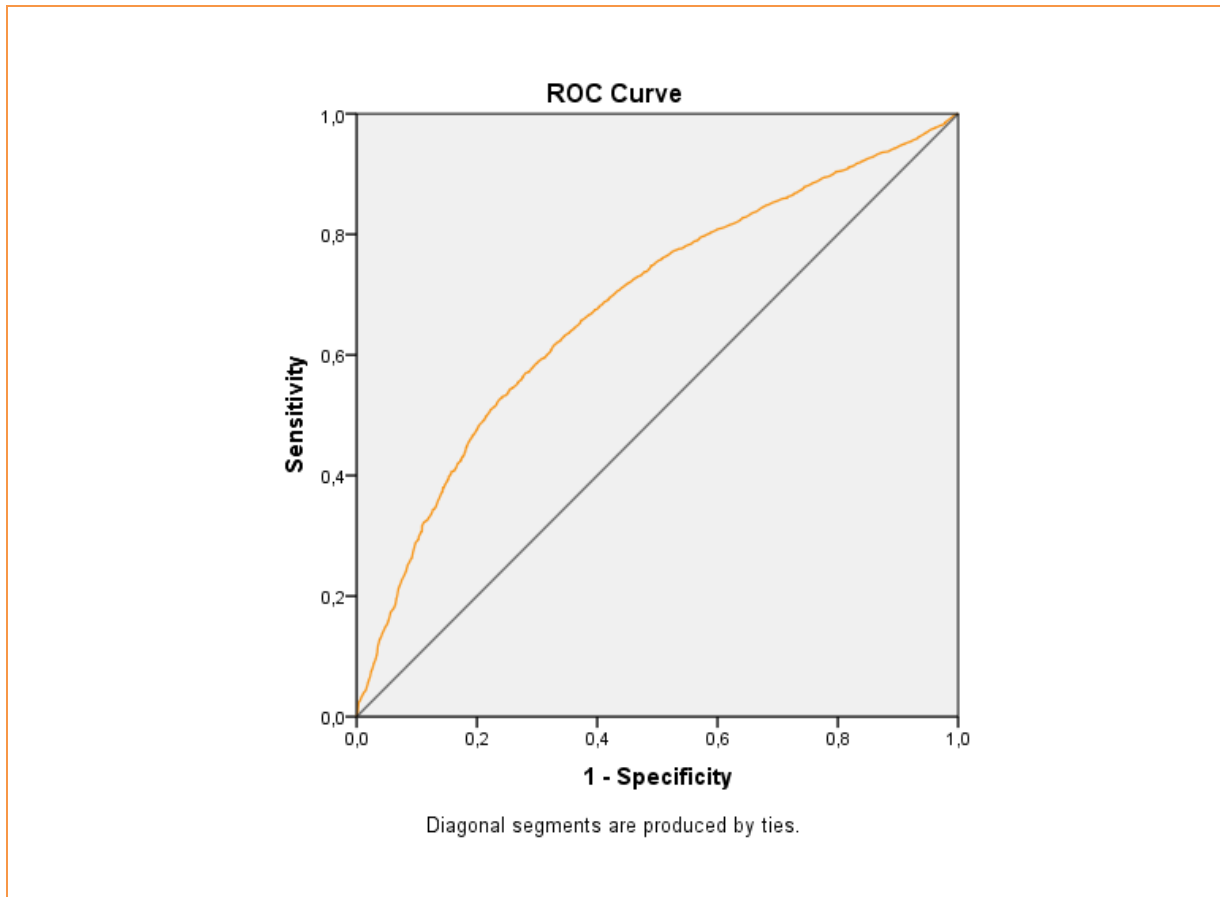
/CRITERIA=CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)

/MISSING=EXCLUDE.

<b>active</b>	Active lifestyle (0=No; 1=Yes)
<b>age</b>	Age in years (Min=18; Max=79)
<b>marital</b>	Marital status (0=Unmarried; 1=Married)
<b>bfast_energy</b>	Preferred breakfast: Energy bar (0=No; 1=Yes) Reference group
<b>bfast_oatmeal</b>	Preferred breakfast: Oatmeal (0=No; 1=Yes)
<b>bfast_cereal</b>	Preferred breakfast: Cereal (0=No; 1=Yes)
<b>PRE_1</b>	Predicted probabilities

## Output/Step 1

Look at the graph called **ROC Curve**. The closer to the upper left corner the curve is, the better the predictions are. If the curve follows the corner perfectly all predictions are correct, if the line follows the diagonal line we could just as well flip a coin.





**Output/Step 2**

Look at the table called **Area Under the Curve**. The AUC ranges between 0.5 and 1.0. The nearer the AUC is to 1, the better the predictive power. On the other hand, a value of 0.5 suggests that we may just flip a coin to decide on whether the outcome is a case or non-case. A value of 0.681, as we have here, suggests rather poor predictive power.

<b>Area Under the Curve</b>	
Test Result Variable(s): Predicted probability	
Area	
	.681
The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.	

## 15. ORDINAL REGRESSION

Outline
15.1 Simple ordinal regression
15.2 Multiple ordinal regression
15.3 Model diagnostics

## Introduction

Ordinal regression is used when  $y$  is ordinal (see section 2.2). If you have only one  $x$ , it is called “simple” ordinal regression, and if you have more than one  $x$ , it is called “multiple” ordinal regression. Regardless of whether you are doing a simple or a multiple regression, the  $x$ -variables can be categorical (nominal/ordinal) and/or continuous (ratio/interval).

Key information from ordinal regression	
<b>1. Direction</b>	
Negative	Odds ratio below 1
Positive	Odds ratio above 1
<b>2. Effect size</b>	
Odds ratio	The odds of the outcome being lower or higher, for every one-unit increase in $x$
<b>3. Statistical significance</b>	
P-value	$p < 0.05$ Statistically significant at the 5 % level $p < 0.01$ Statistically significant at the 1 % level $p < 0.001$ Statistically significant at the 0.1 % level
95 % Confidence intervals	Interval includes 1: Statistically significant at the 5 % level Interval does not include 1: Statistically non-significant at the 5 % level

## *Odds ratios (OR)*

An ordinal regression is thus based on the fact that the outcome consists of three or more categories that are possible to rank (i.e. ordered categories):

### **Some examples**

Educational attainment (1=Compulsory; 2=Upper secondary; 3=University)

School marks (1=Low; 2=Average; 3=High)

Self-rated health (1=Excellent; 2=Good; 3=Fair; 4=Poor)

Statement: "Eurovision Song Contest is entertaining" (1=Strongly agree; 2=Agree; 3=Neither agree nor disagree; 4=Disagree; 5=Strongly disagree)

Ordinal regression is used to predict the "odds" of having a lower or a higher value for your dependent variable ( $y$ ), based on the values of the  $x$ -variable(s). Just as for linear regression analysis, we get a coefficient (log odds) that shows the effect of  $x$  on  $y$ . However, because ordinal regression is based on other assumptions than linear regression, we cannot interpret these coefficients very easily. Instead we focus on something called the "odds ratio" ("OR"). We can get the odds ratio by taking the "exponent" of the coefficient: " $\exp(B)$ ".

The OR is interpreted in the following way: "for every one-unit increase in  $x$ ,  $y$  increases/decreases by [the OR]". Accordingly, if you get a negative OR (below 1), you say: "for every one-unit increase in  $x$ ,  $y$  decreases by [the OR]", and if you get a positive OR (above 1), you say: "for every one-unit increase in  $x$ ,  $y$  increases by [the OR]". Unlike linear regression, where the null value (i.e. value that denotes no difference) is 0, the null value for ordinal regression is 1. Also note that an OR can never be negative – it can range between 0 and infinity. What the OR actually stands for – and whether we can say that an effect is small or big – depends on the values of  $x$  and  $y$ .

### *Is the odds ratio the same as risk, probability, chance or likelihood?*

The simple answer is no. These terms are *not* the same (but the more uncommon the outcome is, the closer odds ratios and risks become). Still, it is very common that odds ratios are interpreted in terms of risks, probabilities, chances or likelihoods. For example, suppose we examine the association between unemployment (0=Employed; 1=Unemployed) and alcohol consumption

(1=None; 2=Moderate; 3=Excessive) and we get an OR for unemployment that is 1.36. It is very tempting to interpret this as the unemployed having 36 % higher risk of drinking more alcohol compared to the employed. Or, if we investigate the association between school marks (ranges between 10 and 20, higher score=better marks) and educational attainment (1=Low; 2=Medium; 3=High), and get an OR for school marks that is 1.09, many would say that the chance of higher educational attainment increases by 9 % for every one-unit increase of school marks.

To avoid all of these problems, here is one suggestion: if you do not have to give any exact numbers, then it is all right to say that some individuals have higher or lower odds/odds ratio/risk/probability compared to other individuals. However, if you want to give exact numbers to exemplify, always use the actual OR.

### *P-values and confidence intervals*

In ordinal regression analysis you can of course get information about statistical significance, in terms of both p-values and confidence intervals. The p-values and the confidence intervals will give you partly different information, *but*: they are not contradictory. If the p-value is below 0.05, the 95 % confidence interval will not include 1 (statistical significance at the 5 % level), and if the p-value is above 0.05, the 95 % confidence interval will include 1 (statistical non-significance at the 5 % level).

Note that when you look at the p-value, you can rather easily distinguish between the significance levels (i.e. you can directly say whether you have statistical significance at the 5 % level, the 1 % level, or the 0.1 % level). When it comes to confidence intervals, SPSS will by default choose 95 % level confidence intervals (i.e. statistical significance at the 5 % level). For some analyses, it is however possible to change the confidence level for the intervals. For example, you may instruct SPSS to show 99 % confidence intervals instead.

### *R-Squared*

In contrast to linear regression, “R-Squared” or “R<sup>2</sup>” is not very usable (again, because of the assumptions behind ordinal regression). You will, however, get a value for the so-called “Nagelkerke R Square” which is similar to the R-squared.

### *Simple versus multiple regression models*

The difference between simple and multiple regression models, is that in a multiple regression each x-variable's effect on y is estimated while taking into account the other x-variables' effects on y. We then say that these other x-variables are "held constant", or "adjusted for", or "controlled for". Because of this, multiple regression analysis is a way of dealing with the issue of "confounding" variables, and to some extent also "mediating" variables (see Section 11.2).

It is highly advisable to run a simple ordinal regression for each of the x-variables before including them in a multiple regression. Otherwise, you will not have anything to compare the adjusted odds ratios with (i.e. what happened to the OR when other x-variables were included in the analysis). Including multiple x-variables in the same model usually (but not always) means that the associations are reduced in strength – which would of course be expected if the x-variables overlapped in their effect on y.

### *Define your analytical sample*

Before you begin, make sure that you have defined your analytical sample correctly (see Section 12.6).

# 15.1 Simple ordinal regression

Quick facts	
<b>Number of variables</b>	One dependent (y) One independent (x)
<b>Scale of variable(s)</b>	Dependent: ordinal Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

## *Simple ordinal regression: Function*

### **Part 1**

1. Go to the **Menu bar**, choose **Analyze\Regression\Ordinal**.
2. A new window called **Ordinal Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Then you choose the variable you want as your independent variable (x). If it is categorical (i.e. nominal/ordinal with more than two values), you transfer it to the box called **Factor(s)** and if it is continuous (i.e. ratio/interval) or binary, you transfer it to the box called **Covariate(s)**.
5. Click on **OK** to get the results in your **Output** window.

### **Part 2**

As of SPSS v.15, you will not get odds ratios directly in your output – you will only something called **Estimate** (which is the log odds). An easy solution is to calculate the odds ratios as well as the 95 % confidence intervals in Excel, based on the **Estimate** (log odds) and **Std. Error** (standard error) you get from SPSS.



## Simple ordinal regression: Syntax

### Part 1

PLUM **DEPVAR** BY **INDEPVAR1** WITH **INDEPVAR2**

/CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5) PCONVERGE(1.0E-6) SINGULAR(1.0E-8)

/LINK=LOGIT

/PRINT=FIT PARAMETER SUMMARY.

### Part 2

Do the following in Excel based on the output you get from SPSS:

1. Open Excel
2. To get the OR, write in one cell:  
=EXP(*Estimate*)
3. To get the lower end of the confidence interval, write in one cell:  
=EXP(*Estimate*-1,96\**Std. Error*)
4. To get the upper end of the confidence interval, write in one cell:  
=EXP(*Estimate*+1,96\**Std. Error*)

**DEPVAR** Name of the dependent variable.

**BY INDEPVAR1** Name of the categorical (i.e. nominal/ordinal with more than two categories) independent variable.\*

or

**WITH INDEPVAR2** Name of the binary or continuous independent variable.

\* If you make dummies of your categorical variable, you include them as **INDEPVAR2** instead.

## *Simple ordinal regression with a continuous x*

### **Example 1**

Suppose we want to examine the association between unemployment days ( $x$ ) and self-rated health ( $y$ ) by means of a simple ordinal regression analysis. Unemployment days are measured as the total number of days in unemployment during a year, and ranges from 0 to 365. Self-rated health has the values 1=Poor; 2=Fair; and 3=Good. Let us say that we get an OR that is 0.93. That would mean that we have a negative association: the higher the number of unemployment days, the lower the odds (or likelihood) of having good health.

### **Example 2**

In another example, we may examine the association between intelligence scores ( $x$ ) and the amount of books read per month ( $y$ ). Intelligence scores are measured by a series of tests that render various amounts of points, and ranges between 20 and 160 points. Book reading has the values 1=0 books; 2=1-3 books; and 3=4 or more books. Here, we get an OR of 1.81. We can thus conclude that higher intelligence scores are associated with more reading of books.

## Simple ordinal regression with a continuous x: Example

(Data: SPSS\_data2.sav)

PLUM **health** WITH **income**

/CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5) PCONVERGE(1.0E-6) SINGULAR(1.0E-8)

/LINK=LOGIT

/PRINT=FIT PARAMETER SUMMARY.

**health** Self-rated health (1=Very good; 2=Good; 3=Fair; 4=Poor)

**income** Income scale 1 (Lowest step) to 10 (Highest step)

## Output

Look at the table called **Parameter Estimates**. The column called **Estimate** shows the coefficient for the variable **income**. Note that these coefficients are the log relative risks, and thus the null value is 0 (like in linear regression). The coefficient is -0.205, which means that we have a negative association between **income** and **health**. In other words, for every one-unit increase in income (i.e. moving from one income step to a higher), the risk of poor health decreases. The column called **Sig.** shows the p-value. Here, the p-value is 0.000 which means that the association between **income** and **health** is statistically significant (at the 0.1 % level).

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[health = 1]	-2,180	,020	11354,819	1	,000	-2,220	-2,140
	[health = 2]	-,127	,018	50,120	1	,000	-,163	-,092
	[health = 3]	1,749	,022	6340,268	1	,000	1,706	1,792
Location	income	-,205	,004	3369,466	1	,000	-,212	-,198

Link function: Logit.

If we want to calculate the OR and the 95 % confidence intervals related to this OR, we do the following in Excel:

1. To get the OR, write in one cell:  
 $=\text{EXP}(-0,205)$
2. To get the lower end of the confidence interval, write in one cell:  
 $=\text{EXP}(-0,205-1,96*0,004)$
3. To get the lower end of the confidence interval, write in one cell:  
 $=\text{EXP}(-0,205+1,96*0,004)$

This gives us the OR=0.81 and the 95 % CI=0.81-0.82. This confirms what the **Estimate** already told us. The interval does not include the null value (which is always  $x=1$  in ordinal regression) and, thus, the results are statistically significant (at the 5 % level).

## *Simple ordinal regression with a binary x*

### **Example 1**

Suppose we want to examine the association between gender (x) and educational attainment (y) by means of a simple ordinal regression analysis. Gender has the values 0=Man and 1=Woman, whereas educational attainment has the values 1=Low, 2=Medium, and 3=High. Now, we get an OR of 1.62. This would mean that women have higher educational attainment compared to men.

### **Example 2**

Here we want to examine the association between having small children (x) and number of pets (y) by means of a simple ordinal regression analysis. Having small children is measured as either 0=No or 1=Yes. Number of pets has the values 1=No pet, 2=1-2 pets, and 3=3 or more pets. Let us say that we get an OR that is 1.29. We can hereby conclude that families with small children own more pets than families without small children.

## Simple ordinal regression with a binary x: Example

(Data: SPSS\_data2.sav)

PLUM **health** WITH **gender**

/CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5) PCONVERGE(1.0E-6) SINGULAR(1.0E-8)

/LINK=LOGIT

/PRINT=FIT PARAMETER SUMMARY.

**health**

Self-rated health (1=Very good; 2=Good; 3=Fair; 4=Poor)

**gender**

Gender (0=Man; 1=Woman)

## Output

Look at the table called **Parameter Estimates**. The column called **Estimate** shows the coefficient for the variable **gender**. Note that these coefficients are the log relative risks, and thus the null value is 0 (like in linear regression). The coefficient is 0.185, which means that we have a positive association between **gender** and **health**. In other words, for every one-unit increase in gender (i.e. moving from man to woman), the risk of poor health increases. The column called **Sig.** shows the p-value. Here, the p-value is 0.000 which means that the association between **gender** and **health** is statistically significant (at the 0.1 % level).

	Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Threshold [health = 1]	-1,031	,012	7628,143	1	,000	-1,055	-1,008
[health = 2]	,924	,012	6297,260	1	,000	,901	,947
[health = 3]	2,750	,018	23583,558	1	,000	2,715	2,785
Location gender	,185	,015	158,936	1	,000	,157	,214

Link function: Logit.

If we want to calculate the OR and the 95 % confidence intervals related to this OR, we do the following in Excel:

1. To get the OR, write in one cell:  
`=EXP(0,185)`
2. To get the lower end of the confidence interval, write in one cell:  
`=EXP(0,185-1,96*0,015)`
3. To get the upper end of the confidence interval, write in one cell:  
`=EXP(0,185+1,96*0,015)`

This gives us the OR=1.20 and the 95 % CI=1.17-1.24. This confirms what the **Estimate** already told us. The interval does not include the null value (which is always  $x=1$  in ordinal regression) and, thus, the results are statistically significant (at the 5 % level).

## *Simple ordinal regression with a categorical x (dummies)*

### **Example 1**

We want to investigate the association between educational attainment (x) and happiness (y) by means of a simple ordinal regression analysis. Educational attainment has the values: 1=Compulsory education, 2=Upper secondary education, and 3=University education. Happiness has the values 1=Happy, 2=Neither happy not unhappy; 3=Unhappy. Since our x-variable is categorical with more than two categories, we have to create dummies (one 0/1 coded variable for each category). We choose those with compulsory education as our reference category, meaning that this dummy will not be included in the analysis. We will thus get one OR for upper secondary education and one OR for university education. Each of these should be compared to the reference category. Let us say that we get an OR for upper secondary education that is 0.87 and we get an OR for university education that is 0.66. We can thus conclude that higher educational attainment is associated with less unhappiness (or more happiness).

### **Example 2**

Suppose we are interested in the association between family type (x) and adolescent smoking (y). Family type has three categories: 1=Two-parent household, 2=Joint custody, and 3=Single-parent household. Adolescent smoking has the values 1=No, 2=Occasionally, and 3=Frequently. Since our x-variable is categorical with more than two categories, we have to create dummies (one 0/1 coded variable for each category). We choose adolescents living in a two-parent household as our reference category, meaning that this dummy will not be included in the analysis. We will thus get one OR for joint custody and one OR for single-parent household. Each of these should be compared to the reference category. The analysis results in an OR of 1.33 for joint custody and an OR of 3.01 for single-parent household. That would mean that adolescents living in family types other than two-parent households smoke more.



## Simple ordinal regression with a categorical x (dummies): Example

(Data: SPSS\_data2.sav)

```
PLUM health WITH age_3cat_younger age_3cat_older  
  /CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5) PCONVERGE(1.0E-  
6) SINGULAR(1.0E-8)  
  /LINK=LOGIT  
  /PRINT=FIT PARAMETER SUMMARY.
```

health	Self-rated health (1=Very good; 2=Good; 3=Fair; 4=Poor)
age_3cat_younger	Age intervals: 15-29 (0=No; 1=Yes)
age_3cat_middle	Age intervals: 30-49 (0=No; 1=Yes) Reference category
age_3cat_older	Age intervals: 50-98 (0=No; 1=Yes)

## Output

Look at the table called **Parameter Estimates**. The column called **Estimate** shows the coefficients for the dummies of the variable **age\_3cat**. Note that these coefficients are the log relative risks, and thus the null value is 0 (like in linear regression). The coefficient for **age\_3cat\_younger** is -0.368 and the coefficient for **age\_3cat\_older** is 0.807. This suggests that we have an overall positive association between **age\_3cat** and **health**: the older the age, the greater the risk of ill-health (or, to be more exact, that the younger individuals have better health compared to the reference category and that the older individuals have worse health compared to the reference category). The column called **Sig.** shows the p-values. Here, both p-values are 0.000 which means that the association between **age\_3cat** and **health** is statistically significant (at the 0.1 % level).

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[health = 1]	-1,038	,013	6422,174	1	,000	-1,063	-1,012
	[health = 2]	1,009	,013	6118,718	1	,000	,984	1,034
	[health = 3]	2,893	,019	23070,444	1	,000	2,856	2,931
Location	age_3cat_younger	-,368	,018	408,337	1	,000	-,404	-,333
	age_3cat_older	,807	,018	2053,541	1	,000	,772	,842

Link function: Logit.

If we want to calculate the OR and the 95 % confidence intervals related to this OR, we do the following in Excel:

1. To get the OR for **age\_3cat\_younger**, write in one cell:  
=EXP(-0,368)
2. To get the lower end of the confidence interval, write in one cell:  
=EXP(-0,368-1,96\*0,018)
3. To get the upper end of the confidence interval, write in one cell:  
=EXP(-0,368+1,96\*0,018)
4. To get the OR for **age\_3cat\_older**, write in one cell:  
=EXP(0,807)
5. To get the lower end of the confidence interval, write in one cell:  
=EXP(0,807-1,96\*0,018)
6. To get the upper end of the confidence interval, write in one cell:  
=EXP(0,807+1,96\*0,018)

For `age_3cat_younger`, this gives us the OR=0.69 and the 95 % CI=0.67-0.72. For `age_3cat_older`, we get the OR=2.24 and the 95 % CI=2.16-2.32. All of this confirms what the `Estimate` already told us. The intervals do not include the null value (which is always  $x=1$  in ordinal regression) and, thus, the results are statistically significant (at the 5 % level).

## 15.2 Multiple ordinal regression

Quick facts	
<b>Number of variables</b>	One dependent (y) At least two independent (x)
<b>Scale of variable(s)</b>	Dependent: ordinal Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

## Multiple ordinal regression: Function

### Part 1

1. Go to the **Menu bar**, choose **Analyze\Regression\Ordinal**.
2. A new window called **Ordinal Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Then you choose the variables you want as your independent variables (x). If they are categorical (i.e. nominal/ordinal with more than two values), you transfer them to the box called **Factor(s)** and they are continuous (i.e. ratio/interval) or binary, you transfer them to the box called **Covariate(s)**. It is possible to include variables of both types at the same time.
5. Click on **OK** to get the results in your **Output** window.

### Part 2

As of SPSS v.15, you will not get odds ratios directly in your output – you will only something called **Estimate** (which is the log odds). An easy solution is to calculate the odds ratios as well as the 95 % confidence intervals in Excel, based on the **Estimate** (log odds) and **Std. Error** (standard error) you get from SPSS.

## Multiple ordinal regression: Syntax

### Part 1

PLUM **DEPVAR** BY **INDEPVAR1** WITH **INDEPVAR2**

/CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5) PCONVERGE(1.0E-6) SINGULAR(1.0E-8)

/LINK=LOGIT

/PRINT=FIT PARAMETER SUMMARY.

### Part 2

Do the following in Excel based on the output you get from SPSS:

5. Open Excel
6. To get the OR, write in one cell:  
=EXP(*Estimate*)
7. To get the lower end of the confidence interval, write in one cell:  
=EXP(*Estimate*-1,96\**Std. Error*)
8. To get the upper end of the confidence interval, write in one cell:  
=EXP(*Estimate*+1,96\**Std. Error*)

**DEPVAR** Name of the dependent variable.

**BY INDEPVAR1** List the names of the categorical (i.e. nominal/ordinal with more than two categories) independent variables.\*

or

**WITH INDEPVAR2** List the names of the binary or continuous independent variables.

\* If you make dummies of your categorical variables, you include them as **INDEPVAR2** instead.

## *Multiple ordinal regression*

### **Example**

Suppose we are interested to see if alcohol consumption ( $y$ ) is related to having small children ( $x$ ), residential area ( $x$ ), and income ( $x$ ). Alcohol consumption has the values 1=None/low, 2=Medium, 3=High. Having small children is measured as either 0=No or 1=Yes. Residential area has the values 1=Metropolitan, 2=Smaller city, and 3=Rural. Income is measured as the yearly household income from salary in thousands of SEK (ranges between 100 and 700 SEK).

When we do our ordinal regression, we can include the variables “having children” and “income” as they are, since the first is binary and the other is continuous. The remaining variable – residential area – is categorical with more than two values and therefore dummies must be used. We create one dummy for each category of “residential area”. The first dummy (metropolitan) is chosen as the reference category and is thus excluded from the analysis.

In the regression analysis, we get an OR for having small children that is 0.65. That means that those who have small children drink less alcohol. This association is adjusted for residential area and income. With regards to residential area, we get an OR for “smaller city” of 1.32 and the OR for “rural” is 2.44. This suggests that those who live in a smaller city drink more alcohol, and so do those living in rural areas. These results are adjusted for having small children and income. Finally, the OR for income is 0.95. This suggests that for every one-unit increase in income (i.e. for every additional one thousand SEK), the consumption of alcohol decreases. This association is adjusted for having small children and residential area.

## Multiple ordinal regression: Example

(Data: SPSS\_data2.sav)

```
PLUM health WITH income gender age_3cat_younger age_3cat_older
  /CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5) PCONVERGE(1.0E-
6) SINGULAR(1.0E-8)
  /LINK=LOGIT
  /PRINT=FIT PARAMETER SUMMARY.
```

health	Self-rated health (1=Very good; 2=Good; 3=Fair; 4=Poor)
income	Income scale 1 (Lowest step) to 10 (Highest step)
gender	Gender (0=Man; 1=Woman)
age_3cat_younger	Age intervals: 15-29 (0=No; 1=Yes)
age_3cat_middle	Age intervals: 30-49 (0=No; 1=Yes) Reference category
age_3cat_older	Age intervals: 50-98 (0=No; 1=Yes)



## Output

Look at the table called **Parameter Estimates**. The column called **Estimate** shows the coefficients for the independent variables. Note that these coefficients are the log relative risks, and thus the null value is 0 (like in linear regression). The coefficient for **income** is -0.195, which means that we have a negative association between **income** and **health**. In other words, for every one-unit increase in income (i.e. moving from one income step to a higher), the risk of poor health decreases. This association is adjusted for **gender** and **age\_3cat**. The coefficient for **gender** is 0.187, which means that we have a positive association between **gender** and **health**. In other words, for every one-unit increase in gender (i.e. moving from man to woman), the risk of poor health increases. This association is adjusted for **income** and **age\_3cat**. Then we have the dummies for age\_3cat: the coefficient for **age\_3cat\_younger** is -0.371 and the coefficient for **age\_3cat\_older** is 0.757. This suggests that we have an overall positive association between **age\_3cat** and **health**: the older the age, the greater the risk of ill-health (or, to be more exact, that the younger individuals have better health compared to the reference category and that the older individuals have worse health compared to the reference category). This association is adjusted for **income** and **gender**.

The column called **Sig.** shows the p-values. Here, all the p-values are 0.000 which means that the mutually adjusted associations analyzed here are statistically significant (at the 0.1 % level).

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[health = 1]	-1,963	,024	6493,798	1	,000	-2,011	-1,915
	[health = 2]	,179	,023	61,726	1	,000	,134	,223
	[health = 3]	2,120	,026	6453,907	1	,000	2,068	2,172
Location	income	-,195	,004	3003,156	1	,000	-,202	-,188
	gender	,187	,016	142,460	1	,000	,156	,217
	age_3cat_younger	-,371	,019	370,762	1	,000	-,409	-,334
	age_3cat_older	,757	,019	1619,585	1	,000	,720	,793

Link function: Logit.

If we want to calculate the odds ratios and the 95 % confidence intervals related to these odds ratios, we do the following in Excel:

1. To get the OR for **income**, write in one cell:  
=EXP(-0,195)

2. To get the lower end of the confidence interval, write in one cell:  
=EXP(-0,195-1,96\*0,004)
3. To get the upper end of the confidence interval, write in one cell:  
=EXP(-0,195+1,96\*0,004)
4. To get the OR for **gender**, write in one cell:  
=EXP(0,187)
5. To get the lower end of the confidence interval, write in one cell:  
=EXP(0,187-1,96\*0,016)
6. To get the upper end of the confidence interval, write in one cell:  
=EXP(0,187+1,96\*0,016)
7. To get the OR for **age\_3cat\_younger**, write in one cell:  
=EXP(-0,371)
8. To get the lower end of the confidence interval, write in one cell:  
=EXP(-0,371-1,96\*0,019)
9. To get the upper end of the confidence interval, write in one cell:  
=EXP(-0,371+1,96\*0,019)
10. To get the OR for **age\_3cat\_older**, write in one cell:  
=EXP(0,757)
11. To get the lower end of the confidence interval, write in one cell:  
=EXP(0,757-1,96\*0,019)
12. To get the upper end of the confidence interval, write in one cell:  
=EXP(0,757+1,96\*0,019)

These are the results:

**income**: OR=0.82, 95 % CI=0.82-0.83

**gender**: OR=1.21, 95 % CI=1.17-1.24

**age\_3cat\_younger**: OR=0.69, 95 % CI=0.66-0.72

**age\_3cat\_older**: OR=2.13, 95 % CI=2.05-2.21

All of this confirms what the **Estimates** already told us. The intervals do not include the null value (which is always  $x=1$  in ordinal regression) and, thus, the results are statistically significant (at the 5 % level).

## 15.3 Model diagnostics

### *Parallel lines*

The assumptions behind ordinal regression are different from linear regression. For example, we do not need to assume linearity, homoscedasticity or normality. Multicollinearity should however be avoided (i.e. strong correlations between the x-variables in multiple ordinal regression – check with correlation analysis; see Chapter 9). One critical thing that we need to consider is called the “proportional odds assumption” or the “parallel lines assumptions”. The assumption behind ordinal regression analysis is that the coefficients that describe the relationship between, for example, the lowest versus all higher categories of the outcome variable are the same as those that describe the relationship between the next lowest category and all higher categories, and so on. Because the relationships between all pairs of categories are assumed to be the same, we only get one estimate for each x-variable.

#### **The parallel lines assumption**

**The effect of x is the same for each pair of categories in y**

If the parallel lines assumption is violated, we should consider another type of analysis: either we could use multinomial regression (see Chapter 16) or we could change our ordinal outcome into a binary version and use logistic regression instead (see Chapter 14).

### *Parallel lines assumption: Function*

1. Go to the **Menu bar**, choose **Analyze\Regression\Ordinal**.
2. A new window called **Ordinal Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Then you choose the variables you want as your independent variables (x). If they are categorical (i.e. nominal/ordinal with more than two values), you transfer them to the box called **Factor(s)** and they are continuous (i.e. ratio/interval) or binary, you transfer them to the box called **Covariate(s)**. It is possible to include variables of both types at the same time.
5. Click on **Output**.
6. Tick the box called **Test of parallel lines**.
7. Click on **Continue**.
8. Click on **OK** to get the results in your **Output** window.

## Parallel lines assumption: Syntax

```
PLUM DEPVAR BY INDEPVAR1 WITH INDEPVAR2  
/CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5) PCONVERGE(1.0E-  
6) SINGULAR(1.0E-8)  
/LINK=LOGIT  
/PRINT=FIT PARAMETER SUMMARY TPARALELL.
```

**DEPVAR** Name of the dependent variable.

**BY INDEPVAR1** Name of the categorical (i.e. nominal/ordinal with more than two categories) independent variables.\*

or

**WITH INDEPVAR2** Name of the binary or continuous independent variables.

**TPARALELL** Order a test of the parallel lines assumption

\* If you make dummies of your categorical variables, you include them as **INDEPVAR2** instead.

## Parallel lines assumption: Example

(Data: SPSS\_data2.sav)

```
PLUM health WITH income gender age_3cat_younger age_3cat_older  
/CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5) PCONVERGE(1.0E-  
6) SINGULAR(1.0E-8)  
/LINK=LOGIT  
/PRINT=FIT PARAMETER SUMMARY TPARALELL.
```

health	Self-rated health (1=Very good; 2=Good; 3=Fair; 4=Poor)
income	Income scale 1 (Lowest step) to 10 (Highest step)
gender	Gender (0=Man; 1=Woman)
age_3cat_younger	Age intervals: 15-29 (0=No; 1=Yes)
age_3cat_middle	Age intervals: 30-49 (0=No; 1=Yes) Reference category
age_3cat_older	Age intervals: 50-98 (0=No; 1=Yes)

## Output

Look at the table called **Test of Parallel Lines**. The column called **Sig.** shows the p-value for the test of the parallel lines assumption. If the p-value is above 0.05, the test is statistically non-significant and the assumption has not been violated. If the p-value is below 0.05, the test is statistically significant and the assumption has been violated. In the current example, the p-value is 0.000 and therefore the model violates the parallel lines assumption – therefore, we should consider using another type of regression analysis.

**Test of Parallel Lines<sup>a</sup>**

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	1891,983			
General	1487,964	404,019	8	,000

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

## 16. MULTINOMIAL REGRESSION

Outline
16.1 Simple multinomial regression
16.2 Multiple multinomial regression
16.3 Model diagnostics



## Introduction

Multinomial regression is used when  $y$  is nominal with more than two categories, i.e. polytomous (see Section 2.2). However, it is a good idea not to have too many categories because the interpretation quickly gets quite messy (if you have more than 5-6, try to collapse some of the categories).

If you have only one  $x$ , it is called “simple” multinomial regression, and if you have more than one  $x$ , it is called “multiple” multinomial regression. Regardless of whether you are doing a simple or a multiple regression, the  $x$ -variables can be categorical (nominal/ordinal) and/or continuous (ratio/interval).

Key information from multinomial regression	
<b>1. Direction</b>	
Negative	Relative risk ratio below 1
Positive	Relative risk ratio above 1
<b>2. Effect size</b>	
Relative risk ratio	The relative risk of the outcome being one category divided by the relative risk that the outcome is the baseline category, for every one-unit increase in $x$
<b>3. Statistical significance</b>	
P-value	$p < 0.05$ Statistically significant at the 5 % level $p < 0.01$ Statistically significant at the 1 % level $p < 0.001$ Statistically significant at the 0.1 % level
95 % Confidence intervals	Interval includes 1: Statistically significant at the 5 % level Interval does not include 1: Statistically non-significant at the 5 % level

## *Relative risk ratios (RRR)*

The most complicated part about the multinomial regression is that we decide on a reference category in the outcome variable as well (for linear, logistic and ordinal regression, we only had to deal with reference categories for the x-variables). To make it easier to distinguish between reference categories in x on the one hand, and in y on the other hand, the following text will continue to talk about “reference category” when x-variables are concerned but use “reference level” with regard to the y-variable.

All right, so our outcome should have a reference level – what does that mean? Let us use an example:

### **Example**

We want to investigate the association between gender (x) and preferred ice-cream flavour (y). Gender has the values 0=Man and 1=Women. Preferred ice-cream flavour has the values: 1=Vanilla, 2=Chocolate, 3=Strawberry. We choose the first category (vanilla) as our reference level. When we run the multinomial regression analysis, we will get two relative risk ratios; one for the risk of the outcome being chocolate instead of vanilla depending on the values of the x-variable, and one for the outcome being strawberry instead of vanilla depending on the values of the x-variable.

So, basically multinomial regression is used to predict the “relative risk” of being a “case” based on the values of the x-variable(s). Just as for linear regression analysis, we get a coefficient (log relative risk) that shows the effect of x on y. However, because multinomial regression is based on other assumptions than linear regression, we cannot interpret these coefficients very easily. Instead we focus on something called the “relative risk ratio” (“RRR”). We can get the relative risk ratio by taking the “exponent” of the coefficient: “exp(B)”.

So what is this about “relative risk ratios” – what happened to odds ratios? Well, these are two different statistical concepts but they are very similar. As previously discussed in e.g. Chapters 14 and 15, the OR is the odds of the outcome being a case divided by the odds of the outcome being a non-case, for every one-unit increase in x. The RRR is the risk of the outcome being a case given a certain value of x, divided by the risk of the outcome being a case given another

value of  $x$ . This may sound complicated, but the take-home message is that odds ratios and relative risk ratios are interpreted in the similar way.

The RRR is interpreted in the following way: “for every one-unit increase in  $x$ ,  $y$  increases/decreases by [the RRR]”. Accordingly, if you get a negative RRR (below 1), you say: “for every one-unit increase in  $x$ ,  $y$  decreases by [the RRR]”, and if you get a positive RRR (above 1), you say: “for every one-unit increase in  $x$ ,  $y$  increases by [the RRR]”. However – and this is important – in multinomial regression we have to deal with the fact that our outcome has a reference level to take into consideration. Therefore, the RRR in multinomial regression analysis is interpreted in the following way: “for every one-unit increase in  $x$ , the risk of  $y$  being [Category  $y_2$ ] compared to  $y$  being [Category  $y_1$ ; reference level] increases/decreases by [the RRR]”, and for every one-unit increase in  $x$ , the risk of  $y$  being [Category  $y_3$ ] compared to  $y$  being [Category  $y_1$ ; reference level] increases/decreases by [the RRR]”, and so on.

Unlike linear regression, where the null value (i.e. value that denotes no difference) is 0, the null value for multinomial regression is 1. Also note that a RRR can never be negative – it can range between 0 and infinity. What the RRR actually stands for – and whether we can say that an effect is small or big – depends on the values of  $x$  and  $y$ .

### *P-values and confidence intervals*

In multinomial regression analysis you can of course get information about statistical significance, in terms of both p-values and confidence intervals. The p-values and the confidence intervals will give you partly different information, *but*: they are not contradictory. If the p-value is below 0.05, the 95 % confidence interval will not include 1 (statistical significance at the 5 % level), and if the p-value is above 0.05, the 95 % confidence interval will include 1 (statistical non-significance at the 5 % level).

Note that when you look at the p-value, you can rather easily distinguish between the significance levels (i.e. you can directly say whether you have statistical significance at the 5 % level, the 1 % level, or the 0.1 % level). When it comes to confidence intervals, SPSS will by default choose 95 % level confidence intervals (i.e. statistical significance at the 5 % level). For some analyses, it is however possible to change the confidence level for the intervals. For example, you may instruct SPSS to show 99 % confidence intervals instead.

## *R-Squared*

In contrast to linear regression, “R-Squared” or “R<sup>2</sup>” is not very usable (again, because of the assumptions behind multinomial regression). You will, however, get a value for the so-called “Nagelkerke R Square” which is similar to the R-squared.

## *Simple versus multiple regression models*

The difference between simple and multiple regression models, is that in a multiple regression each x-variable’s effect on y is estimated while taking into account the other x-variables’ effects on y. We then say that these other x-variables are “held constant”, or “adjusted for”, or “controlled for”. Because of this, multiple regression analysis is a way of dealing with the issue of “confounding” variables, and to some extent also “mediating” variables (see Section 11.2).

It is highly advisable to run a simple multinomial regression for each of the x-variables before including them in a multiple regression. Otherwise, you will not have anything to compare the adjusted relative risk ratios with (i.e. what happened to the RRR when other x-variables were included in the analysis). Including multiple x-variables in the same model usually (but not always) means that the associations are reduced in strength – which would of course be expected if the x-variables overlapped in their effect on y.

## *Define your analytical sample*

Before you begin, make sure that you have defined your analytical sample correctly (see Section 12.6).

## 16.1 Simple multinomial regression

Quick facts	
<b>Number of variables</b>	One dependent (y) One independent (x)
<b>Scale of variable(s)</b>	Dependent: nominal (with more than two categories) Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

### *Simple multinomial regression: Function*

1. Go to the **Menu bar**, choose **Analyze\Regression\Multinomial Logistic**.
2. A new window called **Multinomial Logistic Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Click on **Reference Category** to choose which category of your outcome you want to have as your reference level: **First Category** (lowest value), **Last Category** (highest value) or **Custom** (any value you want).
5. Then you choose the variable you want as your independent variable (x). If it is categorical (i.e. nominal/ordinal with more than two values), you transfer it to the box called **Factor(s)** and if it is continuous (i.e. ratio/interval) or binary, you transfer it to the box called **Covariate(s)**.
6. Click on **OK** to get the results in your **Output** window.

## Simple multinomial regression: Syntax

```
NOMREG DEPVAR (BASE=LAST ORDER=ASCENDING) BY INDEPVAR1 WITH INDEPVAR2
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001)
SINGULAR(0.00000001)
/MODEL
/STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT=INCLUDE
/PRINT=PARAMETER SUMMARY LRT CPS STEP MFI.
```

**DEPVAR** Name of the dependent variable.

**BY INDEPVAR1** Name of the categorical (i.e. nominal/ordinal with more than two categories) independent variable.\*

or

**WITH INDEPVAR2** Name of the binary or continuous independent variable.

\* If you make dummies of your categorical variable, you include them as **INDEPVAR2** instead.

## *Simple multinomial regression with a continuous x*

### **Example 1**

Suppose we want to examine the association between unemployment days ( $x$ ) and type of hospitalization ( $y$ ) by means of a simple multinomial regression analysis. Unemployment days are measured as the total number of days in unemployment during a year, and ranges from 0 to 365. Type of hospitalization has the values 1=No hospitalization, 2=Out-patient care, and 3=In-patient care. We choose “no hospitalization” as our reference level. Let us say that we get a RRR for unemployment days and out-patient care that is 2.88. That would mean that for every one-unit increase of employment days, the risk of experiencing out-patient care compared to no hospitalization increases. Moreover, we get a RRR for unemployment days and in-patient care that is 4.02. This would suggest that for every one-unit increase of employment days, the risk of experiencing in-patient care compared to no hospitalization increases.

### **Example 2**

In another example, we may examine the association between intelligence scores ( $x$ ) and the preferred type of books ( $y$ ). Intelligence scores are measured by a series of tests that render various amounts of points, and ranges between 20 and 160 points. Preferred type of books has the values 1=Fiction, 2=Non-fiction, 3=Comic books. We choose “fiction” as our reference level. Here, we get a RRR of 1.40 for intelligence scores and non-fiction, meaning that for every one-unit increase of intelligence, the likelihood of preferring non-fiction books increases. For intelligence scores and comic books, the RRR is 0.92. This suggests that for every one-unit increase of intelligence, the likelihood of preferring comic books decreases.



## Simple multinomial regression with a continuous x: Example

(Data: SPSS\_data1.sav)

```
NOMREG bfast (BASE=FIRST ORDER=ASCENDING) WITH age
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001)
SINGULAR(0.00000001)
/MODEL
/STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT=INCLUDE
/PRINT=PARAMETER SUMMARY LRT CPS STEP MFI.
```

**bfast** Preferred breakfast (1=Energy bar; 2=Oatmeal; 3=Cereal)  
**age** Age in years (Min=18; Max=79)

## Output

Look at the table called **Parameter Estimates**. The column called **Exp(B)** shows the relative risk ratio (RRR) for the variable **age**. The RRR for **Oatmeal** is 1.116, which means that the likelihood of preferring oatmeal over an energy bar increases for every one-unit increase in age. The RRR is 1.026 for **Cereal**, which means that the likelihood of preferring cereal over an energy bar also increases for every one-unit increase in age. The column called **Sig.** shows the p-value. Here, both p-values are 0.000 which means that the results are statistically significant (at the 0.1 % level). Then we look at the part of the table called **95 % Confidence Interval for Exp (B)**. This gives us the lower confidence limit (**Lower Bound**) and the upper confidence limit (**Upper Bound**). The intervals do not include the null value (which is always  $x=1$  in multinomial regression) and, thus, the results are statistically significant (at the 5 % level).

Preferred breakfast <sup>a</sup>	B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
							Lower Bound	Upper Bound
Oatmeal Intercept	-5,498	,171	1036,475	1	,000			
age	,110	,003	1162,981	1	,000	1,116	1,109	1,124
Cereal Intercept	-,838	,099	71,228	1	,000			
age	,026	,002	121,674	1	,000	1,026	1,022	1,031

a. The reference category is: Energy bar.

## *Simple multinomial regression with a binary x*

### **Example 1**

Suppose we want to examine the association between gender ( $x$ ) and political views ( $y$ ) by means of a simple multinomial regression analysis. Gender has the values 0=Man and 1=Woman, whereas political views has the values 1=Conservative, 2=Centre, and 3=Liberal. The category “centre” is chosen as the reference level. Now, we get a RRR of 0.82 for conservative, which means that women are less likely to be conservative than centre compared to men. The RRR for liberal is 1.39, suggesting that women are more likely to be liberal than centre compared to men.

### **Example 2**

Here we want to examine the association between having small children ( $x$ ) and the type of pet owned ( $y$ ) by means of a simple multinomial regression analysis. Having small children is measured as either 0=No or 1=Yes. Type of pet owned has the values 1=No pet, 2=Cat, 3=Dog, and 4=Other type of pet. The category “no pet” is chosen as the reference level. Let us say that we get a RRR for cat that is 1.50. This means that those who have small children are more likely to own a cat than no pet at all, compared to those who do not have small children. The RRR for dog is 1.75, suggesting that those who have small children are more likely to own a dog than no pet at all, compared to those who do not have small children. Moreover, the RRR for “other type of pet” is 1.96, which tells us that those who have small children are more likely to own “other type of pet” than no pet at all, compared to those who do not have small children.

## Simple multinomial regression with a binary x: Example

(Data: SPSS\_data1.sav)

```
NOMREG bfast (BASE=FIRST ORDER=ASCENDING) WITH gender
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001)
SINGULAR(0.00000001)
/MODEL
/STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT=INCLUDE
/PRINT=PARAMETER SUMMARY LRT CPS STEP MFI.
```

**bfast** Preferred breakfast (1=Energy bar; 2=Oatmeal; 3=Cereal)  
**gender** Gender (0=Man; 1=Woman)

## Output

Look at the table called **Parameter Estimates**. The column called **Exp(B)** shows the relative risk ratio (RRR) for the variable **gender**. The RRR for **Oatmeal** is 1.030, which means that women are more likely than men to prefer oatmeal over an energy bar. The RRR for **Cereal** is 1.029, which means that women are more likely than men to also prefer cereal over an energy bar. The column called **Sig.** shows the p-value. Here, both p-values are above 0.05 which means that the results are not statistically significant. Then we look at the part of the table called **95 % Confidence Interval for Exp (B)**. This gives us the lower confidence limit (**Lower Bound**) and the upper confidence limit (**Upper Bound**). The intervals include the null value (which is always  $x=1$  in multinomial regression) and, thus, the results are not statistically significant (at the 5 % level).

Preferred breakfast <sup>a</sup>	B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
							Lower Bound	Upper Bound
Oatmeal Intercept	-,018	,051	,129	1	,720			
gender	,030	,072	,172	1	,678	1,030	,895	1,186
Cereal Intercept	,185	,049	14,325	1	,000			
gender	,029	,068	,179	1	,672	1,029	,900	1,177

a. The reference category is: Energy bar.

## *Simple multinomial regression with a categorical x (dummies)*

### **Example 1**

We want to investigate the association between educational attainment (x) and building type (y) by means of a simple multinomial regression analysis. Educational attainment has the values: 1=Compulsory education, 2=Upper secondary education, and 3=University education. Building type has the values 1=Apartment, 2=Town house, and 3=Villa. We choose “apartment” as our reference level. Since our x-variable is categorical with more than two categories, we have to create dummies (one 0/1 coded variable for each category). We choose those with compulsory education as our reference category, meaning that this dummy will not be included in the analysis. The RRR for upper secondary education in combination with town house is 2.01, meaning that those with upper secondary education are more likely to live in a town house than an apartment, compared to those with compulsory education. The RRR for upper secondary education in combination with villa is 1.32, meaning that those with upper secondary education are more likely to live in a villa than an apartment, compared to those with compulsory education. For university education in combination with town house, the RRR is 0.95, suggesting that those who have university education are less likely to live in a town house than an apartment compared to those with compulsory education. Finally, the RRR for university education in combination with villa is 3.44, meaning that those with university education are more likely to live in a villa than an apartment, compared to those with compulsory education.

### Example 2

Suppose we are interested in the association between family type (x) and adolescent health behaviour (y). Family type has three categories: 1=Two-parent household, 2=Joint custody, and 3=Single-parent household. Adolescent health behaviour has the values 1=No smoking or alcohol consumption, 2=Smoking, 3=Alcohol consumption, 4=Both smoking and alcohol consumption. We choose the category no smoking or alcohol consumption as the reference level. Since our x-variable is categorical with more than two categories, we have to create dummies (one 0/1 coded variable for each category). We choose adolescents living in a two-parent household as our reference category, meaning that this dummy will not be included in the analysis. The RRR for the combination joint custody and smoking is 1.20, meaning that adolescents living in joint custody are more likely to smoke than not to smoke or drink alcohol compared to those living in a two-parent household. The RRR for the combination single-parent household and smoking is 1.49, meaning that adolescents living in single-parent household are more likely to smoke than not to smoke or drink alcohol compared to those living in a two-parent household. The RRR for the combination joint custody and alcohol consumption is 1.00, meaning that adolescents living in joint custody are as likely to drink alcohol as not to smoke or drink alcohol compared to those living in a two-parent household. The RRR for the combination single-parent household and alcohol consumption is 2.02, meaning that adolescents living in single-parent household are more likely to drink alcohol than not to smoke or drink alcohol compared to those living in a two-parent household. The RRR for the combination joint custody and both smoking and alcohol consumption is 1.55, meaning that adolescents living in joint custody are more likely to both smoke and drink alcohol than not to smoke or drink alcohol compared to those living in a two-parent household. The RRR for the combination single-parent household and both smoking and alcohol consumption is 4.45, meaning that adolescents living in single-parent household are more likely to both smoke and drink alcohol than not to smoke or drink alcohol compared to those living in a two-parent household.

## Simple multinomial regression with a categorical x (dummies): Example

(Data: SPSS\_data1.sav)

```
NOMREG bfast (BASE=FIRST ORDER=ASCENDING) WITH edcat_no edcat_somcoll
edcat_colldeg edcat_postgrad
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001)
SINGULAR(0.00000001)
/MODEL
/STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT=INCLUDE
/PRINT=PARAMETER SUMMARY LRT CPS STEP MFI.
```

<b>bfast</b>	Preferred breakfast (1=Energy bar; 2=Oatmeal; 3=Cereal)
<b>edcat_no</b>	Did not complete high school (0=No; 1=Yes)
<b>edcat_highsc</b>	High school degree (0=No; 1=Yes) Reference group
<b>edcat_somcoll</b>	Some college (0=No; 1=Yes)
<b>edcat_colldeg</b>	College degree (0=No; 1=Yes)
<b>edcat_postgrad</b>	Post-undergraduate degree (0=No; 1=Yes)



## Output

Look at the table called **Parameter Estimates**. The column called **Exp(B)** shows the relative risk ratios (RRR) for the dummies of the variable **edcat**. The first part of the table concerns educational differences in **Energy bar** versus **Oatmeal**. The RRR for the combination **edcat\_no** and **Oatmeal** is 1.568, which means that those who did not complete high school are more likely to prefer oatmeal over an energy bar compared to those with a high school degree. The RRR is for the combination **edcat\_somcoll** and **Oatmeal** is 0.922, which means that those who have some college education are less likely to prefer oatmeal over an energy bar compared to those with a high school degree. The RRR for the combination **edcat\_colldeg** and **Oatmeal** is 1.032, which means that those who have a college degree are more likely to prefer oatmeal over an energy bar compared to those with a high school degree. The RRR for the combination **edcat\_postgrad** and **Oatmeal** is 0.955, which means that those who have a post-undergraduate degree are less likely to prefer oatmeal over an energy bar compared to those with a high school degree.

The second part of the table concerns educational differences in **Energy bar** versus **Cereal**. The RRR for the combination **edcat\_no** and **Cereal** is 1.049, which means that those who did not complete high school are more likely to prefer cereal over an energy bar compared to those with a high school degree. The RRR for the combination **edcat\_somcoll** and **Cereal** is 1.039, which means that those who have some college education are more likely to prefer cereal over an energy bar compared to those with a high school degree. The RRR for the combination **edcat\_colldeg** and **Cereal** is 1.075, which means that those who have a college degree are more likely to prefer cereal over an energy bar compared to those with a high school degree. The RRR for the combination **edcat\_postgrad** and **Cereal** is 1.047, which means that those who have a post-undergraduate degree are more likely to prefer cereal over an energy bar compared to those with a high school degree.

The column called **Sig.** shows the p-values. Here, most p-values are above 0.05 which means that the results are not statistically significant. The only significant difference is for **edcat\_no** and **Oatmeal** (p=0.000) meaning that those who did not complete high school are significantly more likely to prefer oatmeal over an energy bar compared to those with a high school degree (at the 0.1 % level). Then we look at the part of the table called **95 % Confidence Interval for Exp (B)**. This gives us the lower confidence limit (**Lower Bound**) and the upper confidence limit (**Upper Bound**). Most intervals – with the exception of **edcat\_no** and **Oatmeal** – include the null value (which is always  $x=1$  in multinomial regression) and, thus, the overall results are not statistically significant.

**Parameter Estimates**

Preferred breakfast <sup>a</sup>	B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
							Lower Bound	Upper Bound
Oatmeal	Intercept	-,082	,064	1,642	1	,200		
	edcat_no	,450	,103	18,984	1	,000	1,568	1,281 1,919
	edcat_somecoll	-,081	,104	,603	1	,437	,922	,753 1,131
	edcat_colldeg	,032	,100	,101	1	,751	1,032	,848 1,256
	edcat_postgrad	-,046	,150	,096	1	,757	,955	,711 1,281
Cereal	Intercept	,163	,060	7,321	1	,007		
	edcat_no	,048	,103	,217	1	,641	1,049	,857 1,284
	edcat_somecoll	,039	,096	,162	1	,687	1,039	,861 1,255
	edcat_colldeg	,072	,094	,594	1	,441	1,075	,894 1,292
	edcat_postgrad	,046	,139	,109	1	,741	1,047	,798 1,374

a. The reference category is: Energy bar.

## 16.2 Multiple multinomial regression

Quick facts	
<b>Number of variables</b>	One dependent (y) At least two independent (x)
<b>Scale of variable(s)</b>	Dependent: nominal (with more than two categories) Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

## *Multiple multinomial regression: Function*

1. Go to the **Menu bar**, choose **Analyze\Regression\Multinomial Logistic**.
2. A new window called **Multinomial Logistic Regression** will open.
3. In the left box, all your variables are displayed. You choose the variable you want to have as your dependent variable (y) and transfer it to the box called **Dependent**.
4. Click on **Reference Category** to choose which category of your outcome you want to have as your reference level: **First Category** (lowest value), **Last Category** (highest value) or **Custom** (any value you want).
5. Then you choose the variables you want as your independent variables (x). If they are categorical (i.e. nominal/ordinal with more than two values), you transfer them to the box called **Factor(s)** and they are continuous (i.e. ratio/interval) or binary, you transfer them to the box called **Covariate(s)**. It is possible to include variables of both types at the same time.
6. Click on **OK** to get the results in your **Output** window.

## Multiple multinomial regression: Syntax

```
NOMREG DEPVAR (BASE=LAST ORDER=ASCENDING) BY INDEPVAR1 WITH INDEPVAR2
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001)
SINGULAR(0.00000001)
/MODEL
/STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT=INCLUDE
/PRINT=PARAMETER SUMMARY LRT CPS STEP MFI.
```

**DEPVAR** Name of the dependent variable.

**BY INDEPVAR1** List the names of the categorical (i.e. nominal/ordinal with more than two categories) independent variables.\*

or

**WITH INDEPVAR2** List the names of the binary or continuous independent variables.

\* If you make dummies of your categorical variables, you include them as **INDEPVAR2** instead.

## *Multiple multinomial regression*

### **Example**

Suppose we are interested to see if smoking ( $y$ ) is related to having small children ( $x$ ), residential area ( $x$ ), and income ( $x$ ). Smoking has the values 1=Non-smoker, 2=Former smoker, and 3=Current smoker. The category “non-smoker” is chosen as the reference level. Having small children is measured as either 0=No or 1=Yes. Residential area has the values 1=Metropolitan, 2=Smaller city, and 3=Rural. Income is measured as the yearly household income from salary in thousands of SEK (ranges between 100 and 700 SEK). When we do our multinomial regression, we can include the variables “having children” and “income” as they are, since the first is binary and the other is continuous. The remaining variable – residential area – is categorical with more than two values and therefore dummies must be used. We create one dummy for each category of “residential area”. The first dummy (metropolitan) is chosen as the reference category and is thus excluded from the analysis.

In the regression analysis, we get a RRR of 1.19 for having small children and former smoker, suggesting that those who have small children are more likely to be former smokers than non-smokers compared to those who do not have small children. Then we get a RRR of 0.77 for having small children and being a current smoker, which means that those who have small children are less likely to be current smokers than non-smokers compared to those who do not have small children. These results are adjusted for residential area and income. The RRR for living in a smaller city and being a former smoker is 2.09, which suggests that those who live in a smaller city are more likely to be former smokers than non-smokers compared to those who live in a metropolitan area. The RRR for living in a smaller city and being a current smoker is 3.71, which suggests that those who live in a smaller city are more likely to be current smokers than non-smokers compared to those who live in a metropolitan area. The RRR for living in an urban area and being a former smoker is 3.59, which suggests that those who live in an urban area are more likely to be former smokers than non-smokers compared to those who live in a metropolitan area. The RRR for living in an urban area and being a current smoker is 5.01, which suggests that those who live in an urban area are more likely to be current smokers than non-smokers compared to those who live in a metropolitan area. These results are adjusted for having small children and income. With regard to income, the RRR of being a former smoker is 0.93, suggesting that for every one-unit increase in income, the risk of being a former smoker decreases. The RRR of being a current smoker is 0.78, which

means that for every one-unit increase in income, the risk of being a current smoker also decreases. These results are adjusted for having small children and residential area.

## Multiple multinomial regression: Example

(Data: SPSS\_data1.sav)

```
NOMREG bfast (BASE=FIRST ORDER=ASCENDING) WITH age gender edcat_no edcat_somecoll
edcat_colldeg edcat_postgrad
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001)
SINGULAR(0.00000001)
/MODEL
/STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT=INCLUDE
/PRINT=PARAMETER SUMMARY LRT CPS STEP MFI.
```

<b>bfast</b>	Preferred breakfast (1=Energy bar; 2=Oatmeal; 3=Cereal)
<b>age</b>	Age in years (Min=18; Max=79)
<b>gender</b>	Gender (0=Man; 1=Woman)
<b>edcat_no</b>	Did not complete high school (0=No; 1=Yes)
<b>edcat_highsc</b>	High school degree (0=No; 1=Yes) Reference group
<b>edcat_somecoll</b>	Some college (0=No; 1=Yes)
<b>edcat_colldeg</b>	College degree (0=No; 1=Yes)
<b>edcat_postgrad</b>	Post-undergraduate degree (0=No; 1=Yes)



## Output

Look at the table called **Parameter Estimates**. The column called **Exp(B)** shows the mutually adjusted relative risk ratios for the variables **gender**, **age**, and **edcat**. The first part of the table concerns differences between **Energy bar** and **Oatmeal**. The RRR for **gender** and **Oatmeal** is 1.059, which means that women are more likely to prefer oatmeal over energy bars. For **age** and **Oatmeal**, the RRR is 1.117, suggesting that for every one-unit increase in age, individuals are more likely to prefer oatmeal over energy bars. The RRR for the combination **edcat\_no** and **Oatmeal** is 0.966, which means that those who did not complete high school are less likely to prefer oatmeal over an energy bar compared to those with a high school degree. The RRR for the combination **edcat\_somecoll** and **Oatmeal** is 0.859, which means that those who have some college education are less likely to prefer oatmeal over an energy bar compared to those with a high school degree. The RRR for the combination **edcat\_colldeg** and **Oatmeal** is 1.086, which means that those who have a college degree are more likely to prefer oatmeal over an energy bar compared to those with a high school degree. The RRR for the combination **edcat\_postgrad** and **Oatmeal** is 0.939, which means that those who have a post-undergraduate degree are less likely to prefer oatmeal over an energy bar compared to those with a high school degree.

The second part of the table concerns educational differences in **Energy bar** versus **Cereal**. For **gender** and **Cereal**, the RRR is 1.042. This means that women are more likely to prefer cereal over energy bars. The RRR for **age** and **Cereal** is 1.027, suggesting that for every one-unit increase in age, individuals are more likely to prefer cereal over energy bars. The RRR for the combination **edcat\_no** and **Cereal** is 0.911, which means that those who did not complete high school are less likely to prefer cereal over an energy bar compared to those with a high school degree. The RRR for the combination **edcat\_somecoll** and **Cereal** is 1.007, which means that those who have some college education are about as likely to prefer cereal over an energy bar compared to those with a high school degree. The RRR for the combination **edcat\_colldeg** and **Cereal** is 1.028, which means that those who have a college degree are more likely to prefer cereal over an energy bar compared to those with a high school degree. The RRR for the combination **edcat\_postgrad** and **Cereal** is 0.959, which means that those who have a post-undergraduate degree are less likely to prefer cereal over an energy bar compared to those with a high school degree.

The column called **Sig.** shows the p-values. Here, most p-values are above 0.05 which means that the results are not statistically significant (at the 5 % level). However, **age** is the exception (p=0.000): increased age is associated at a statistically significant level (the 0.1 % level) with preferring both oatmeal and cereal over energy bars. Then we look at the part of the table called

95 % Confidence Interval for Exp (B). This gives us the lower confidence limit (Lower Bound) and the upper confidence limit (Upper Bound). Most intervals – with the exception of the ones concerning age – include the null value (which is always x=1 in multinomial regression) and, thus, the overall results are not statistically significant (at the 5 % level).

Parameter Estimates									
Preferred breakfast <sup>a</sup>		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
Oatmeal	Intercept	-5,518	,189	856,805	1	,000			
	gender	,058	,088	,431	1	,511	1,059	,892	1,258
	age	,110	,003	1148,380	1	,000	1,117	1,110	1,124
	edcat_no	-,035	,127	,074	1	,785	,966	,753	1,239
	edcat_somecoll	-,152	,128	1,411	1	,235	,859	,669	1,104
	edcat_colldeg	,082	,122	,456	1	,500	1,086	,855	1,379
	edcat_postgrad	-,063	,180	,122	1	,727	,939	,661	1,335
Cereal	Intercept	-,859	,115	56,318	1	,000			
	gender	,041	,070	,345	1	,557	1,042	,909	1,194
	age	,026	,002	122,413	1	,000	1,027	1,022	1,032
	edcat_no	-,094	,106	,785	1	,376	,911	,740	1,120
	edcat_somecoll	,007	,098	,005	1	,941	1,007	,832	1,220
	edcat_colldeg	,028	,095	,084	1	,771	1,028	,853	1,239
	edcat_postgrad	-,042	,141	,089	1	,766	,959	,728	1,263

a. The reference category is: Energy bar.

### 16.3 Model diagnostics

The assumptions behind multinomial regression are different from linear regression. For example, we do not need to assume linearity, homoscedasticity or normality. Multicollinearity should however be avoided (i.e. strong correlations between the x-variables in multiple ordinal regression – check with correlation analysis; see Chapter 9). It is not very easy to evaluate model fit with multinomial regression. One alternative is to carry out a bunch of separate logistic regressions. For example, if the outcome was smoking and that variable had the categories 1=Non-smoker, 2=Former smoker, and 3=Current smoker, we could make two separate logistic regressions. The first would have the outcome 0=Non-smoker and 1=Former smoker, whereas the second would have the outcome 0=Non-smoker and 1=Current smoker. Then we could use the same type of model diagnostics as for logistic regression (see Section 14.3). Another option to assess the model’s goodness of fit – or basically its predictive power – is to use a “likelihood ratio test” (which is a type of chi-square test). The likelihood ratio test gives an answer to whether the estimated model (i.e. the model with one or more x-variables) predicts the outcome better than the null model (i.e. a model without any x-variables).

#### Estimate the goodness of fit

<b>Likelihood ratio test</b>	Does the estimated model predict the outcome better than the null model?
------------------------------	--

## Likelihood ratio test: Example

(Data: SPSS\_data1.sav)

```
NOMREG bfast (BASE=FIRST ORDER=ASCENDING) WITH gender age edcat_no edcat_somecoll
edcat_colldeg edcat_postgrad
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001)
SINGULAR(0.00000001)
/MODEL
/STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT=INCLUDE
/PRINT=PARAMETER SUMMARY LRT CPS STEP MFI.
```

<b>bfast</b>	Preferred breakfast (1=Energy bar; 2=Oatmeal; 3=Cereal)
<b>age</b>	Age in years (Min=18; Max=79)
<b>gender</b>	Gender (0=Man; 1=Woman)
<b>edcat_no</b>	Did not complete high school (0=No; 1=Yes)
<b>edcat_highsc</b>	High school degree (0=No; 1=Yes) Reference group
<b>edcat_somecoll</b>	Some college (0=No; 1=Yes)
<b>edcat_colldeg</b>	College degree (0=No; 1=Yes)
<b>edcat_postgrad</b>	Post-undergraduate degree (0=No; 1=Yes)

## Output

Look at the table called **Model Fitting Information**. The column called **Sig.** shows the p-value for the likelihood ratio test. If the p-value is above 0.05, the test is statistically non-significant and the estimated model fits the data better than the empty model. If the p-value is below 0.05, the test is statistically significant and this means that the estimated model does not fit the data better. In the current example, the p-value is 0.000 and therefore we can say that the estimated model fits the data rather poorly.

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	4996,455			
Final	3010,464	1985,991	12	,000

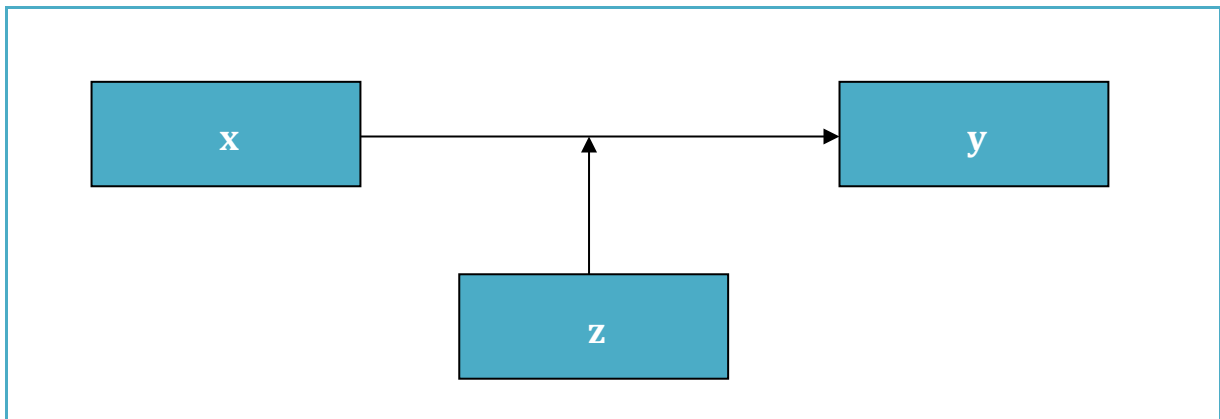
## 17. INTERACTION ANALYSIS

### Outline

**17.1** Interaction analysis for linear regression

**17.2** Interaction analysis for logistic regression

## Introduction



A moderator (or effect modifier) is a variable ( $z$ ) that influences the very association between the  $x$ -variable and the  $y$ -variable. Thus, the association between  $x$  and  $y$  looks different depending on the value of  $z$ .

### Some examples

We want to examine the association between social support ( $x$ ) and mental health ( $y$ ). We think that the association may be moderated by gender ( $z$ ). For example, we may expect social support to be more important for mental health among women than among men.

We are interested in the association between mother's educational attainment ( $x$ ) and babies' birth weight ( $y$ ). It is reasonable that mother's smoking ( $z$ ) affects that association: there may be an association between  $x$  and  $y$  if the mother smokes, but no association between  $x$  and  $y$  if the mother does not smoke.

## Main effects and interaction effects

In order to carry out an interaction analysis, we first we need to decide on the type of regression analysis that fits our outcome of choice – it could be any type (e.g. linear, logistic, ordinal, multinomial, or any other type). When we have decided that, we need a minimum of three independent variables.

Independent variables	
<b>x</b>	The variable we are mainly interested in with regard to its effect on y. Called “main effect term”.
<b>y</b>	The variable we suspect may modify the effect of x on y. Called “main effect term”.
<b>x*y</b>	The product of x and z – or the x-variable times the z-variable. Called “interaction effect term” or simple “interaction term”. Create the interaction with the <b>Compute</b> command (see Section 5.2).

Based on the interaction analysis, we get information on whether or not there is a statistically significant interaction (i.e. if the z-variable modifies the effect of x on y or not). We also get information on in which direction the interaction effect goes, i.e. what it means.

## Measurement scales

There are some important things to consider before carrying out the interaction analysis, such as the measurement scale of the independent variables (see Section 2.2). Generally, it is easier to interpret interaction terms based on the following combinations:

Combinations of variables
One binary x * one binary z
One ordinal/ratio/interval x * one binary z
One binary x * one ordinal/ratio/interval z
One ordinal/ratio/interval x * one ordinal/ratio/interval z



In other words: avoid nominal variables with more than two categories! If you combine two ordinal/ratio/interval variables, make sure that x-variable and the z-variable goes in the same direction. This means that higher values in both variables should be either “better” *or* “worse” in relation to the outcome.

### *Interpretation*

The most complicated part about interaction analysis is the interpretation. It is important that you keep track how your variables are coded, if you want to say something about what the interaction means.

#### **Example**

We want to examine the association between social support (x) and happiness (y). We think that the association may be moderated by gender (z). The following hypotheses are formulated: 1) Those with higher levels of social support are more likely to be happy, 2) Women are more likely to be happy, and 3) Social support is more strongly associated with happiness among women than among men.

Since the outcome is binary (0=Not happy and 1=Happy), we choose logistic regression analysis. Social support ranges between 0 and 10, where higher values reflect higher levels of social support. Gender has the values 0=Man and 1=Women.

To begin with, we examine the association between x and y: the odds ratio for social support is 1.20, which confirmed our first hypothesis. Next, we examine the association between z and y: the odds ratio for gender is 1.17, which confirms the second hypothesis. Finally, we include x and z as well as the interaction term (i.e.  $x*z$ ) in a new logistic regression. The interaction term has an odds ratio of 1.45, which means that the combination of having higher levels of social support and being a woman is associated with increasing chances of being happy.

If the interpretation of the interaction analysis is difficult, you may improve your understanding by doing a separate regression analysis for each category of the z-variable (this is of course only possible if you have a rather large dataset and not too many categories in your z-variable). This is called “stratified” analyses. We can go back to the example to illustrate this:

### Example

We want to examine the association between social support (x) and happiness (y). We think that the association may be moderated by gender (z). The following hypotheses are formulated: 1) Those with higher levels of social support are more likely to be happy, 2) Women are more likely to be happy, and 3) Social support is more strongly associated with happiness among women than among men.

Since the outcome is binary (0=Not happy and 1=Happy), we choose logistic regression analysis. Social support ranges between 0 and 10, where higher values reflect higher levels of social support. Gender has the values 0=Man and 1=Women.

To begin with, we examine the association between x and y *among men only*: the odds ratio for social support is 1.04. Next, we examine the association between x and y *among women only*: the odds ratio for social support is 1.76. Thus, we now see that we have a stronger effect of social support on happiness among women than among men (just like the interaction analysis said).

Remember, however: stratified analyses are perhaps easier to understand, but if you want to say that differences between strata (i.e. categories of the z-variable) are statistically significant, you should do a proper interaction analysis.

## Interaction analysis for linear regression: Example

(Data: SPSS\_data1.sav)

### Part 1

```
COMPUTE active_gender=active*gender.  
EXECUTE.
```

### Part 2

```
REGRESSION  
  /MISSING LISTWISE  
  /STATISTICS COEFF OUTS CI(95) R ANOVA  
  /CRITERIA=PIN(.05) POUT(.10)  
  /NOORIGIN  
  /DEPENDENT hourstv  
  /METHOD=ENTER active gender active_gender.
```

hourstv	Hours spent watching TV last week (Min=0; Max=36)
active	Active lifestyle (0=No; 1=Yes)
gender	Gender (0=Man; 1=Woman)
active_gender	Interaction term: active*gender

## Output

Look at the table called **Coefficients**. The column called **B** shows the B coefficients. We focus on the B coefficient for the interaction term, **active\_gender**. Here, it is 0.211. This means that the combination of being active (since this category has the higher value) and being a woman (since this category has the higher value) is associated with more TV watching. The column called **Sig.** shows the p-value. For the interaction term, **active\_gender**, the p-value is 0.471 which means that the interaction is not statistically significant (at least not at the 5 % level). Then we look at the part of the table called **95 % Confidence Interval for B**. This gives us the lower confidence limit (**Lower Bound**) and the upper confidence limit (**Upper Bound**). For the interaction term, **active\_gender**, the lower limit is -0.364 whereas the upper limit is -0.787. The interval includes the null value (which is always  $x=0$  in linear regression) and, thus, the results are not statistically significant.

To sum up, there is no statistically significant interaction between having an active lifestyle and gender with regard to TV watching. Put differently, gender does not moderate the association between active lifestyle and hours watching TV.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	19,719	,142		138,533	,000	19,440	19,998
	Active lifestyle	-,045	,210	-,004	-,215	,830	-,456	,366
	Gender	-,201	,201	-,019	-1,003	,316	-,594	,192
	active_gender	,211	,293	,017	,720	,471	-,364	,787

a. Dependent Variable: Hours spent watching TV last week

## Interaction analysis for logistic regression: Example

(Data: SPSS\_data1.sav)

### Part 1

```
COMPUTE age_gender=age*gender.  
EXECUTE.
```

### Part 2

```
LOGISTIC REGRESSION VARIABLES active  
/METHOD=ENTER age gender age_gender  
/PRINT=CI(95)  
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

active	Active lifestyle (0=No; 1=Yes)
age	Age in years (Min=18; Max=79)
gender	Gender (0=Man; 1=Woman)
age_gender	Interaction term: age*gender

## Output

Look at the table called **Coefficients**. The column called **Exp(B)** shows the odds ratios (OR). We focus on the OR for the interaction term, **age\_gender**. Here, it is 1.016. This means that the combination of being older (since increased values reflect older age) and being a woman (since this category has the higher value) increased the likelihood of being active. The column called **Sig.** shows the p-value. For the interaction term, **age\_gender**, the p-value is 0.000 which means that the interaction is statistically significant (at the 0.1 % level). Then we look at the part of the table called **95 % C.I. for EXP(B)**. This gives us the lower confidence limit (**Lower**) and the upper confidence limit (**Upper**). For the interaction term, **age\_gender**, the lower limit is 1.009 whereas the upper limit is 1.023. The interval does not include the null value (which is always  $x=1$  in logistic regression) and, thus, the results are statistically significant.

To sum up, there is a statistically significant interaction between having age and gender with regard to having an active lifestyle. Put differently, gender does moderate the association between age and active lifestyle: age has a larger positive effect on active lifestyle among women than among men.

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 <sup>a</sup>								
age	-,037	,003	214,767	1	,000	,964	,959	,968
gender	-,686	,169	16,527	1	,000	,504	,362	,701
age_gender	,016	,003	21,680	1	,000	1,016	1,009	1,023
Constant	1,566	,124	159,448	1	,000	4,787		

a. Variable(s) entered on step 1: age, gender, age\_gender.

