Introduction to SPSS
Thursday, May 19, 2016
1:00pm – 3:00pm
RS2020

Biostatistical Consulting Service
Centre for Addiction and Mental Health
Marcos Sanches
CS 1217A
416-535-8501x36338
marcos.sanches@camh.ca

Introduction to SPSS
Thursday, May 19, 2016
1:00am-3:00pm
RS2020

Part I - Introduction to the SPSS Interface

- Starting SPSS
- Entering Data Manually into SPSS
- Importing Data from Excel
- Saving an SPSS data file
- Initial Data Checking after Importing
- The SPSS Output Window
- The Syntax Editor
- Merge Files
- Exercise

Part II – Data Exploration and Modification

- Recode
- Recode with Conditional IF
- Compute
- Select Cases
- Split File
- Frequencies
- Descriptives
- SPSS Graphs
- Exercise

Part III – Introduction to Statistics using SPSS

- Tests of Normality
- Analyses Involving Two Categorical Variables
- Analyses Involving One Categorical and One Continuous Variable
- Paired Continuous Variables
- Analyses Involving Two Continuous Variables
- Association, Causation, Models and our Research
- Exercise

Part I – Introduction to the SPSS Interface

SPSS is a statistical software package that at first glance looks a lot like Microsoft Excel. But SPSS has statistical capabilities that go far beyond what can be done in Excel. Consequently, SPSS can be a very useful tool in research activities at CAMH, with the ability to carry out many of the statistical analyses often found in publications.

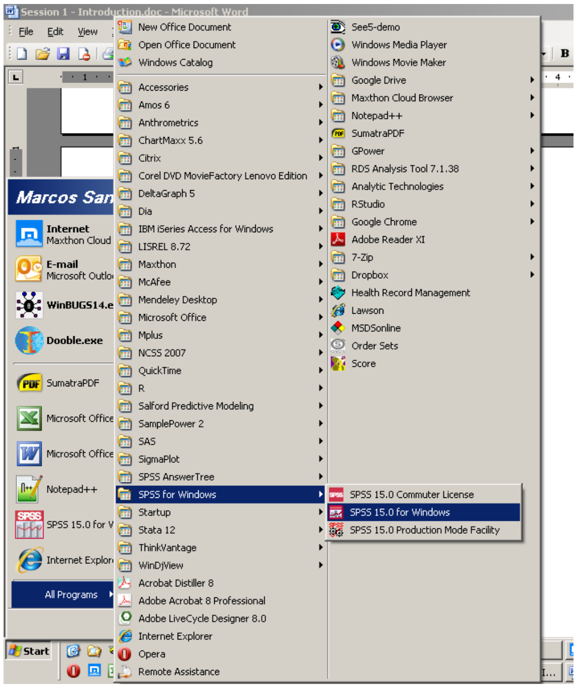
This training workshop aims to introduce users to the SPSS interface and basic operations, to the point where users should be able to carry out basic analyses and data manipulation activities. With this initial familiarity, users will have the required background to tap into other resources to go beyond the contents of this course, as SPSS can do quite advanced and complex analyses. It is important to keep in mind that statistical analysis is not only about being able to use software, it requires appropriate training in statistical theory as well in order to make decisions about study design, selecting the best model for use in analysis, and understanding assumptions and the validity of a given statistical result. For this type of assistance, the staff of the Biostatistical Consulting Service are always available and willing to discuss your research questions, statistical techniques and software and we encourage you to look for us whenever you may not feel comfortable with any statistical issues.

This training is divided into three sessions. The first session will introduce users to the SPSS interface, data entry and some aspects of data manipulation. The second session will cover more aspects of data manipulation and modification and will introduce some useful exploratory techniques. The third session will introduce users to some basic statistical tests.

This class will be interactive; during the instruction you will have the opportunity to work your own way through the tasks in SPSS, which should help to make the content make more sense. Due to the time limit, we will not be able to cover all the material in this manual during the training sessions. But it can still be used like a guide for the material that we will cover, with instructions on how to accomplish tasks in SPSS that we will cover in the training sessions.

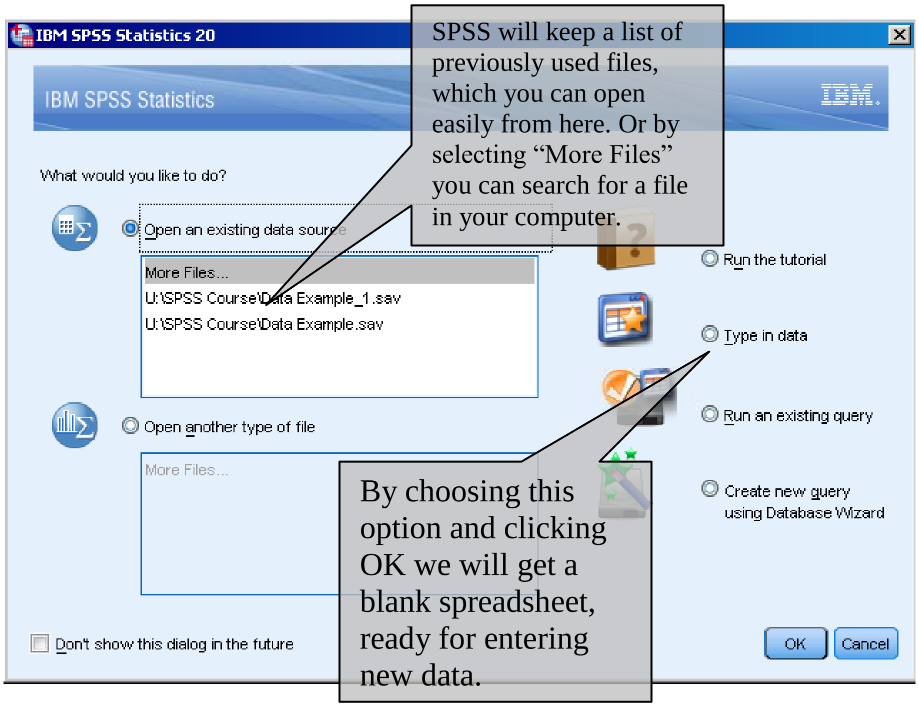
• Starting SPSS

You can start SPSS just like you would any other program by using the Start Menu. SPSS will also usually automatically open if you double click a file associated with SPSS (.sav, .sps, .spo, …).



A window will ask what you want to do with SPSS. We usually will want to open some data file or enter some data. For now, let's enter some data.

Select "Type in data" and click OK.


IBM SPSS Statistics 20
IBM SPSS Statistics
IBM.
What would you like to do?
Open an existing data source
More Files...
U:\SPSS Course\Data Example_1.sav
U:\SPSS Course\Data Example.sav
Open another type of file
More Files...
Run the tutorial
Type in data
Run an existing query
Create new query using Database Wizard
Don't show this dialog in the future
OK
Cancel

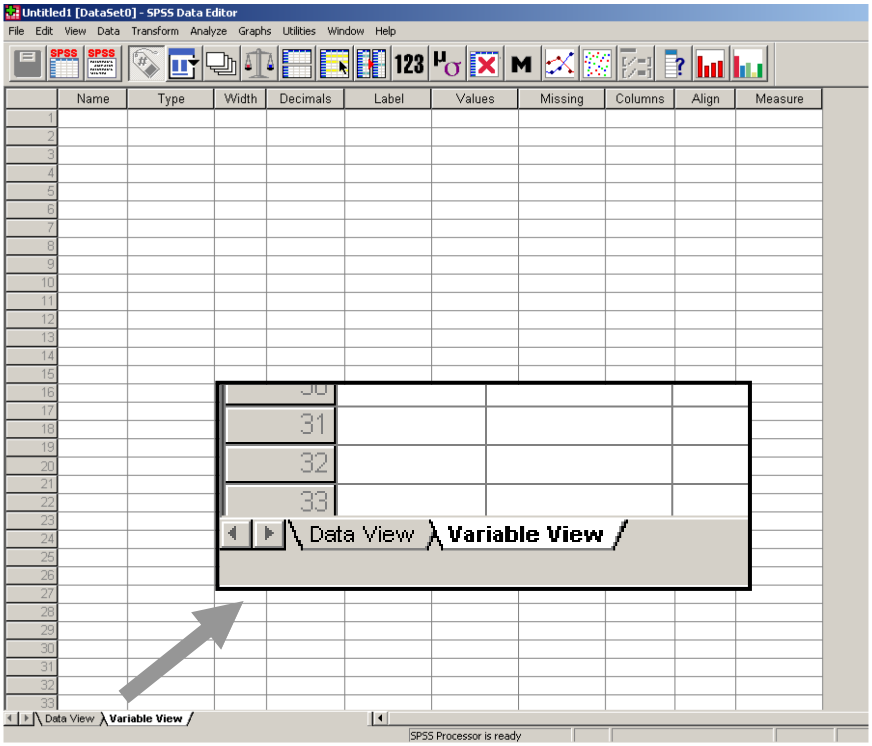
SPSS will keep a list of previously used files, which you can open easily from here. Or by selecting "More Files" you can search for a file in your computer.

By choosing this option and clicking OK we will get a blank spreadsheet, ready for entering new data.

SPSS is a spreadsheet and looks like Excel. Most often, the columns represent the variables i.e. information collected about our study participants and each line represents one individual. At the bottom we have two tabs:

- **“Data View”** allows us to see the raw data.
- **“Variable View”** shows us the characteristics of the variables.

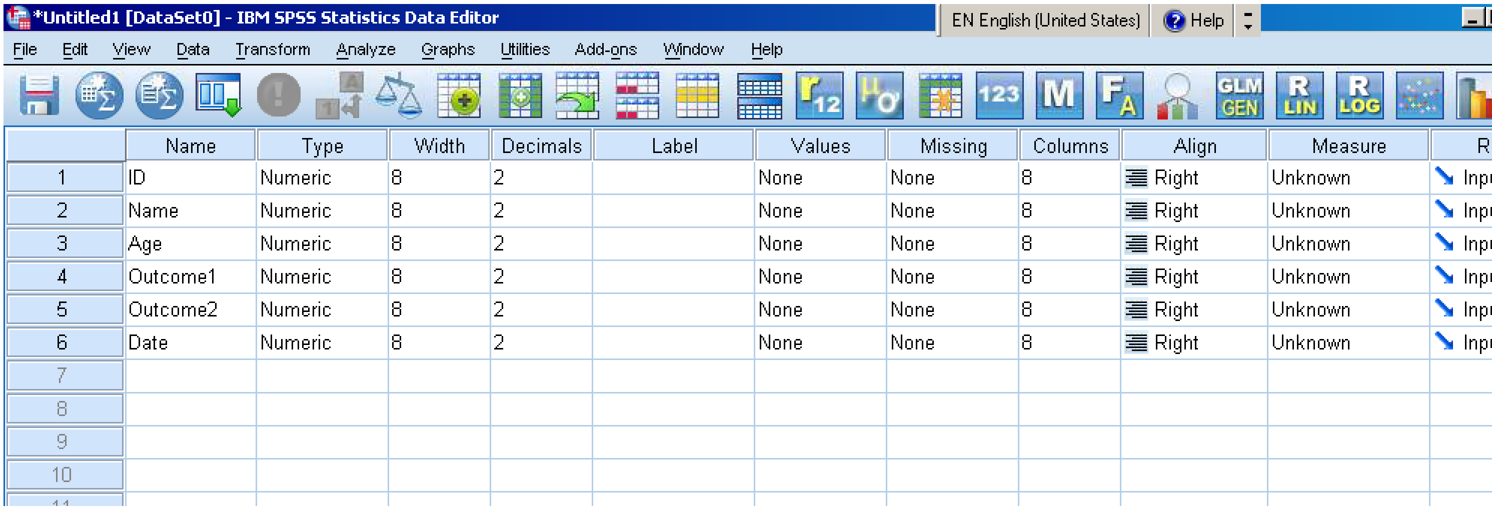
Let’s enter some data to get a feeling of how this works.



• Entering Data Manually into SPSS

The first step when entering data is to define the variable and their attributes. In the "Variable View" tab, column "Name", type the name of your variables.

"Name" of the variable is usually a short string that identifies the information that we will enter. Let's fill out the column "Name" as in the screenshot below. Notice that SPSS will automatically fill out the other columns, which we may need to change as SPSS guesses are not always correct.


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
EN English (United States)
Help
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help

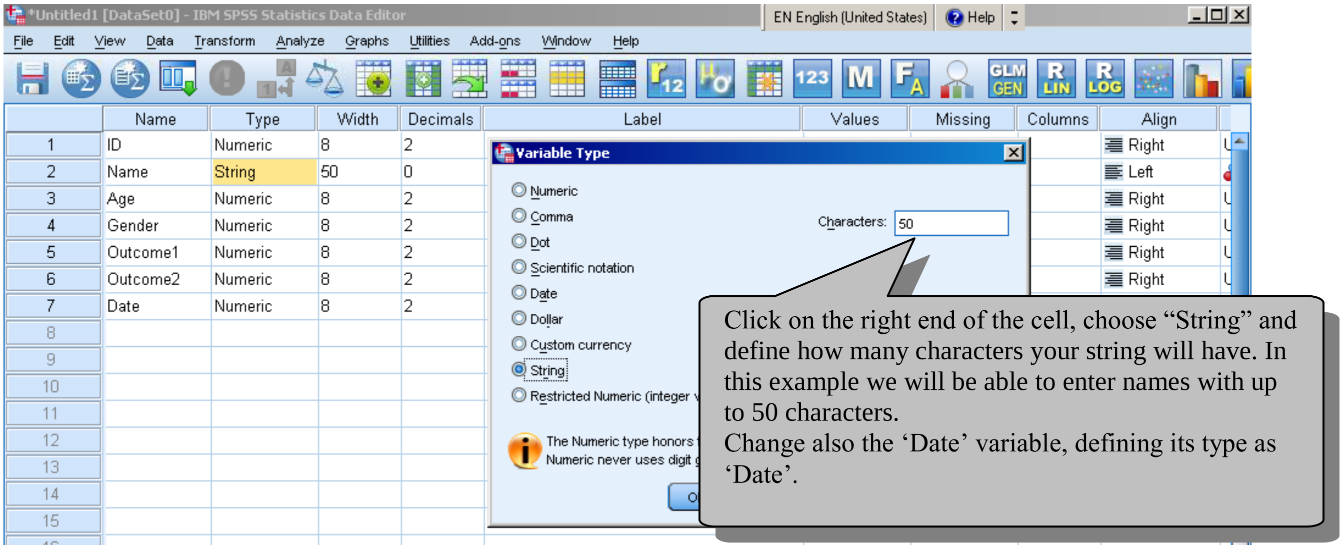
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	R
1	ID	Numeric	8	2		None	None	8	Right	Unknown	Inpu
2	Name	Numeric	8	2		None	None	8	Right	Unknown	Inpu
3	Age	Numeric	8	2		None	None	8	Right	Unknown	Inpu
4	Outcome1	Numeric	8	2		None	None	8	Right	Unknown	Inpu
5	Outcome2	Numeric	8	2		None	None	8	Right	Unknown	Inpu
6	Date	Numeric	8	2		None	None	8	Right	Unknown	Inpu
7											
8											
9											
10											

There are basically three types of variables: Numeric, String and Date. Notice that type is just the format that the data is stored in SPSS and how the data is displayed.

Numeric – Information that is composed of numbers. These can be simply numeric information, like age, or codes for non-numeric information, like gender (1 = male, 2 = female).
String – Information composed of characters, for example, the name of the person, the name of countries, the gender if you wish to enter it as 'Male' and 'Female' (or 'M' and 'F'). Notice that numbers can also be in format String because they are characters, but when in this format we cannot perform mathematical operations with them.
Date – Date information, like the birth date, the date the subject entered the trial, the date the measurement was taken.

In the field "Type" you will find many other types available, besides these three, but the ones not mentioned here are just formatting variations of these three. For example, if the variable contains numeric values that are monetary then you may choose type "Dollar". If the numeric values are too large or too small you may choose type "Scientific Notation". In both examples they are numeric; just appear in the screen with specific formatting. By default SPSS defines all types as Numeric and we need to change it according to the content of our variables.

We can do that by clicking on the right side of the cell in column "Type". Let's make the required changes.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
1	ID	Numeric	8	2					Right
2	Name	String	50	0					Left
3	Age	Numeric	8	2					Right
4	Gender	Numeric	8	2					Right
5	Outcome1	Numeric	8	2					Right
6	Outcome2	Numeric	8	2					Right
7	Date	Numeric	8	2					

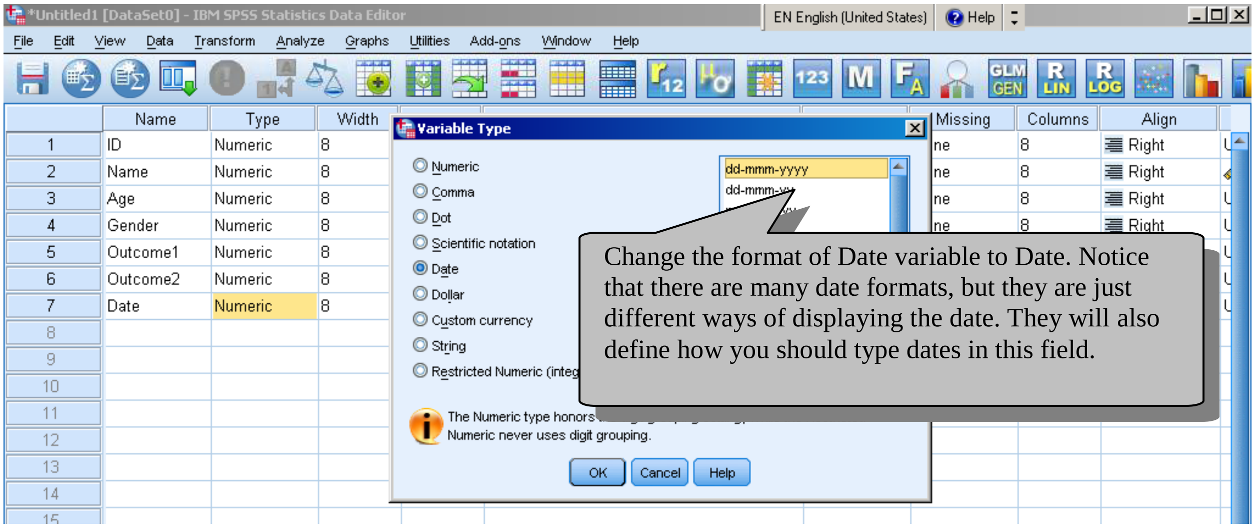
Variable Type

○ Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
○ Custom currency
● String
○ Restricted Numeric (integer w

Characters: 50

The Numeric type honors t
Numeric never uses digit g

Click on the right end of the cell, choose "String" and define how many characters your string will have. In this example we will be able to enter names with up to 50 characters.
Change also the 'Date' variable, defining its type as 'Date'.



	Name	Type	Width	Missing	Columns	Align
1	ID	Numeric	8	ne	8	Right
2	Name	Numeric	8	ne	8	Right
3	Age	Numeric	8	ne	8	Right
4	Gender	Numeric	8	ne	8	Right
5	Outcome1	Numeric	8			
6	Outcome2	Numeric	8			
7	Date	Numeric	8			

Variable Type

○ Numeric
○ Comma
○ Dot
○ Scientific notation
● Date
○ Dollar
○ Custom currency
○ String
○ Restricted Numeric (integ

dd-mmm-yyyy
dd-mmm-yy

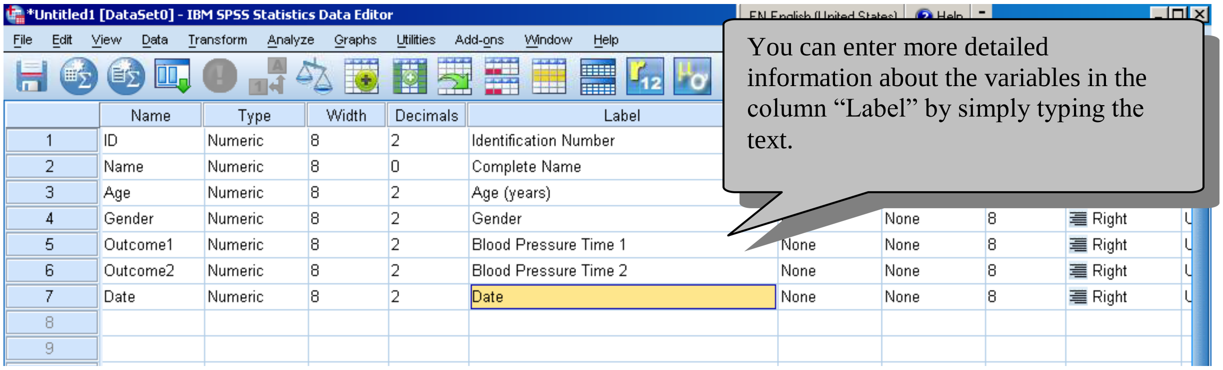
The Numeric type honors
Numeric never uses digit grouping.

OK | Cancel | Help

Change the format of Date variable to Date. Notice that there are many date formats, but they are just different ways of displaying the date. They will also define how you should type dates in this field.

You can also add Variable Labels in the cells under the column “Labels”.

Variable Labels provide a more detailed description of the variables and help us to understand what information the variable contains. It is always helpful to have labels in place as the name of the variables may be hard to decipher. You can just type in the text you wish to use.


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
EN English (United States)
Help

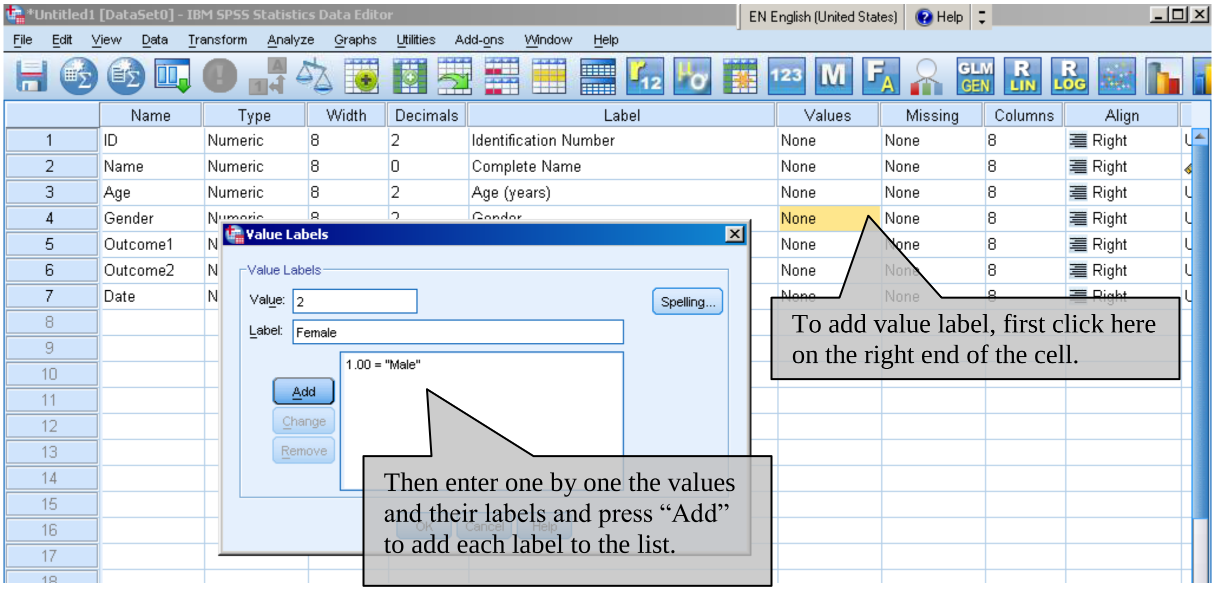
	Name	Type	Width	Decimals	Label					
1	ID	Numeric	8	2	Identification Number					
2	Name	Numeric	8	0	Complete Name					
3	Age	Numeric	8	2	Age (years)					
4	Gender	Numeric	8	2	Gender		None	8	Right	U
5	Outcome1	Numeric	8	2	Blood Pressure Time 1	None	None	8	Right	U
6	Outcome2	Numeric	8	2	Blood Pressure Time 2	None	None	8	Right	U
7	Date	Numeric	8	2	Date	None	None	8	Right	U
8										
9										

You can enter more detailed information about the variables in the column “Label” by simply typing the text.

Some data will be represented as codes. For example, the information about gender can be entered as 1 or 2, 1 being Male and 2 being Female. Age could be 1 (18 to 24), 2 (25 to 34), 3 (35 to 44) or 4 (45 or above). These are numeric values that do not have a numeric interpretation, they are just codes. In cases like these, SPSS allows us to insert ‘value labels’ for these numeric values so that we know that 1 = male and 2 = female, for example. The value labels can be entered under the column “Value”.

When looking at SPSS data you will be able to choose whether you want to see Labels (Male and Female) or Values (1 and 2).

> Click on the right end of the cell, enter the value 1 and the label Male the appropriated boxes and hit “Add”. Do again for Female.


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
EN English (United States)
Help
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
1	ID	Numeric	8	2	Identification Number	None	None	8	Right
2	Name	Numeric	8	0	Complete Name	None	None	8	Right
3	Age	Numeric	8	2	Age (years)	None	None	8	Right
4	Gender	Numeric	8	2	Gender	None	None	8	Right
5	Outcome1	N				None	None	8	Right
6	Outcome2	N				None	None	8	Right
7	Date	N				None	None	8	Right

Value Labels

Value: 2

Label: Female

1.00 = "Male"

To add value label, first click here on the right end of the cell.

Then enter one by one the values and their labels and press "Add" to add each label to the list.

In order to improve the appearance of your data, you can change the "Decimals". For example, IDNumber, Age and Gender are all numeric but they do not have any decimals, so you can enter 0 in the "Decimal" column for these variables. The Decimals column is only about formatting; it does not truncate the actual numeric value.

The Name, Type, Labels and Values columns are the most important and usually the only ones defined before entering the data. The other existing fields are:

Missing – Used to define missing values. For example, you may have blank values at Gender coded as 99 and you do not want to use the 99 in statistical analyses. You can tell SPSS that 99 is actually a missing value code by declaring it in the Missing column. This way SPSS will exclude 99 from tables, and many of the statistical analyses.
Columns – This is the width of the column in the "Data View" tab. It is just for formatting, it has no effect on the content of the variable. You can also adjust the column width by dragging the column border, as you do in Excel.
Align – Is also a formatting field that tells SPSS how to align the information in the "Data View" tab.
Measure – This defines the type of the variables and it only affects some specific analysis, specially the ones involving automatic modeling, like classification trees. Although it is usually not important to define the Measure correctly, it is very important for the analyst to know it because the Measure defines which statistical techniques are most appropriate for the variables.

These are the three variable types:

> **Scale** – Also know as Continuous, are the variables in numeric format that have a defined order and allows for measures of distance and arithmetic operations. For example, Age can be considered continuous because the distance between 1 and 2 is the same as between 4 and 5, that is, one year. CGI is not continuous because although it has an order, distances are meaningless – The distance between 1 (Normal) and 2 (Borderline) may

not be the same as between 2 (Borderline) and 3(Mildly ill). And we cannot do arithmetic operations: 1(Normal) + 2(Borderline) = 3(Mildly ill) does not makes sense. So, CGI is numeric because it is represented by a number, but not Scale.

Ordinal - These are variables like CGI, which may or may not be numeric, but which have a defined ordering. Variables like Age and Income ranges are other example of Ordinal variables.

Nominal – Variables in which the categories do not have a defined ordering, e.g. Gender, Treatment, Region. Variables of the type String will usually be Nominal, although they can be ordinal too.

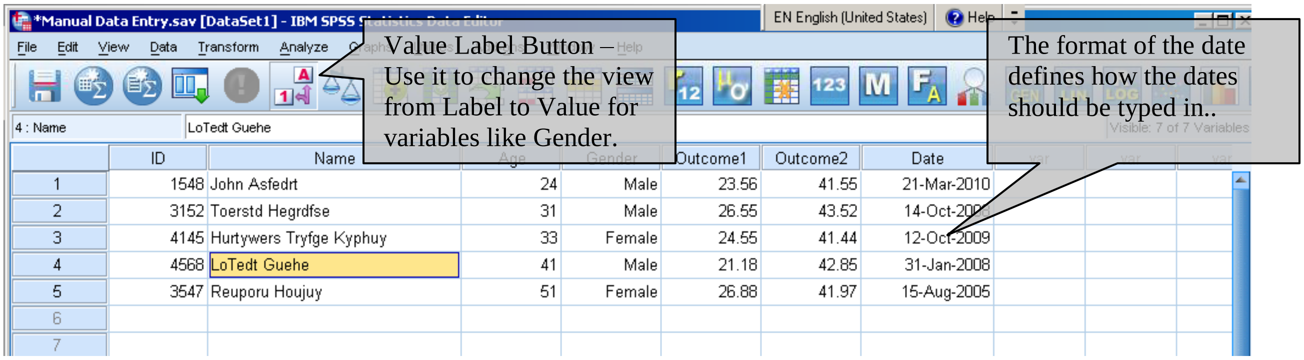
One of the requirements of regression analysis is that the dependent variable is of the type Scale. However, if your dependent variable is a Scale variable wrongly defined as Nominal, SPSS will do the Regression in the same way. SPSS does not really care about the Measure of the variable, it usually will do whichever analysis you ask for. Hence, it is up to you to know the type of variable and statistical technique most appropriate to use, SPSS will not warn you if you choose a technique that is not appropriate for your variables.

Do not confuse Type with Measure. Type is just a formatting attribute and it is not really important and a statistical sense. The Measure does not depend on Type. A Type=Numeric variable can be Measure= Scale, Ordinal or Nominal. A Type = String variable can be Measure = Ordinal or Nominal. The Measure is important for you to define the appropriate statistical analysis, so you need to know it. Defining the Measure in SPSS is not that important as long as you know it and use appropriate statistical methods for your variables, depending on their measure.

Role – This is the last field in the variable view tab. SPSS uses this information for automatic search in predictive models, which is out of the scope of this training.

At this point we have defined our variables and its attributes. If you are satisfied, go to the "Data View" tab and enter your data manually as you would do in Excel. Let's enter two or three lines of data, just to get a feeling of how it works.

Tip – You can change column width by dragging its border. You should enter 1 and 2 for gender, but you may see the labels Male and Female if "Value Labels" button is ON. This helps entering data! You should type the date obeying the type of date you defined. If a value is missing you can just leave it blank, which SPSS will show as a period "." for numeric variables.


*Manual Data Entry.sav [DataSet1] - IBM SPSS Statistics Data Editor
EN English (United States)
Help
File
Edit
View
Data
Transform
Analyze
Graphs
Help
4 : Name
LoTedt Guehe
Visible: 7 of 7 Variables
var

Value Label Button – Use it to change the view from Label to Value for variables like Gender.

The format of the date defines how the dates should be typed in..

	ID	Name	Age	Gender	Outcome1	Outcome2	Date
1	1548	John Asfedrt	24	Male	23.56	41.55	21-Mar-2010
2	3152	Toerstd Hegrdfse	31	Male	26.55	43.52	14-Oct-2008
3	4145	Hurtywers Tryfge Kyphuy	33	Female	24.55	41.44	12-Oct-2009
4	4568	LoTedt Guehe	41	Male	21.18	42.85	31-Jan-2008
5	3547	Reuporu Houjuy	51	Female	26.88	41.97	15-Aug-2005
6							
7							

After having entered the data it is important to check it. Then you can proceed with the analysis. Entering data in SPSS is not very common. Instead the data is entered in different software and imported into SPSS for analysis.

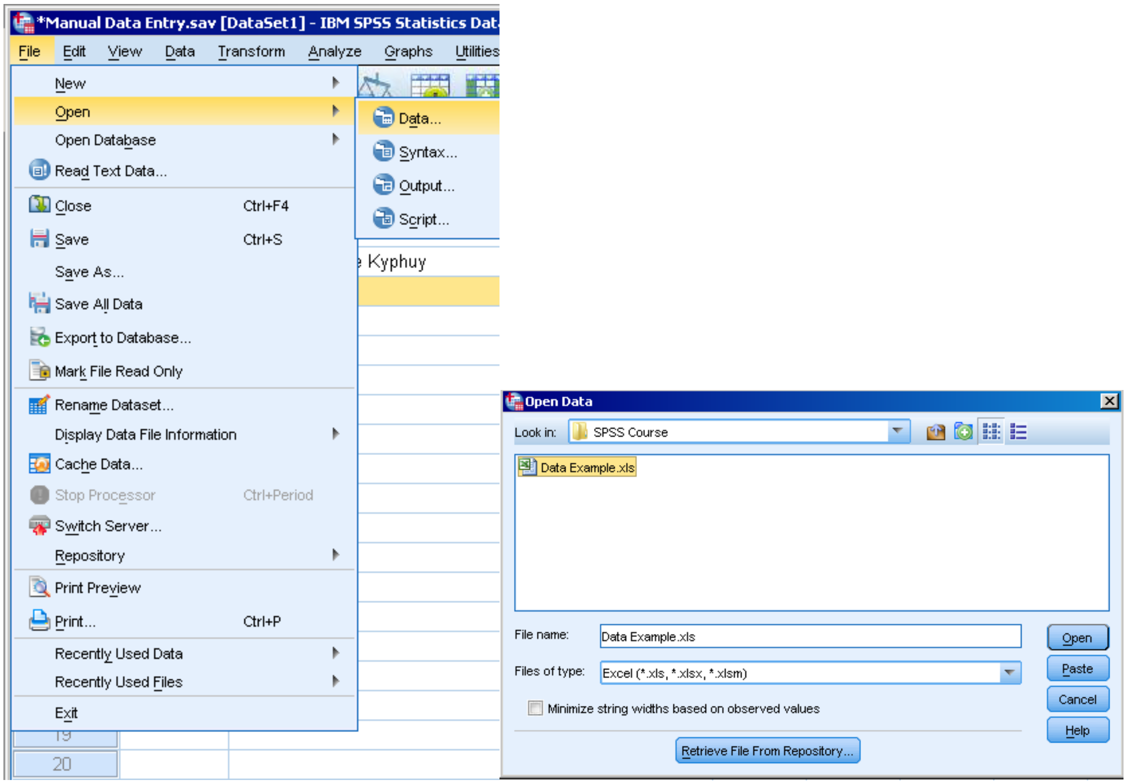
• Importing Data from Excel

SPSS can easily import data from Excel spreadsheets and many other formats. We will cover importing data from Excel given that it is the most commonly encountered format.

Tip – Make sure your data is in good shape before importing into SPSS. Many cleaning procedures are easier to carry out in Excel than in SPSS. Some examples include:

- ✓ Ensure one line of data represents one subject
- ✓ Delete unimportant variables
- ✓ Delete pivot tables, summary statistics, embedded graphs
- ✓ Make sure all values in a variable are in the same format (all numeric or text).
- ✓ Variable names are clear (with no space or strange characters).
- ✓ Make sure you have an ID variable
- ✓ Missing values have consistent codes (i.e. blank or code 99 for example).
- ✓ Exclude records that have problems.
- ✓ Always keep a copy of the original data set.

In the Menu, go to File → Open → Data. In "Files of type" box, choose "Excel" and select the excel file you want to open and click "Open".



Look in: SPSS Course

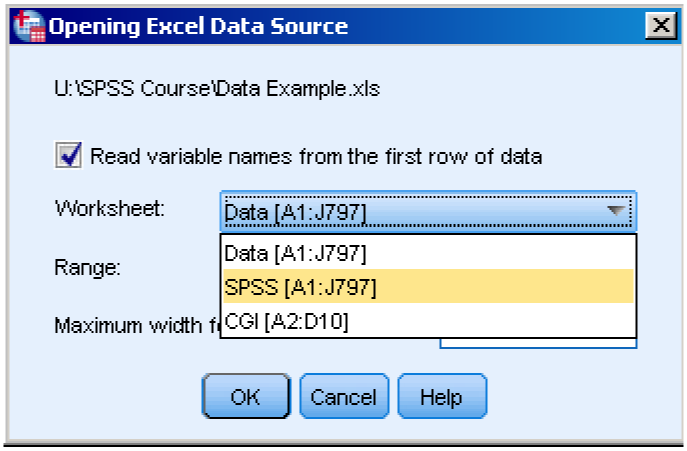
Data Example.xls

File name: Data Example.xls

Files of type: Excel (*.xls, *.xlsx, *.xlsm)

□ Minimize string widths based on observed values

An Excel file may have many worksheets and SPSS will allow you to select which spreadsheet you want to open. Choose "SPSS" worksheet and click OK.

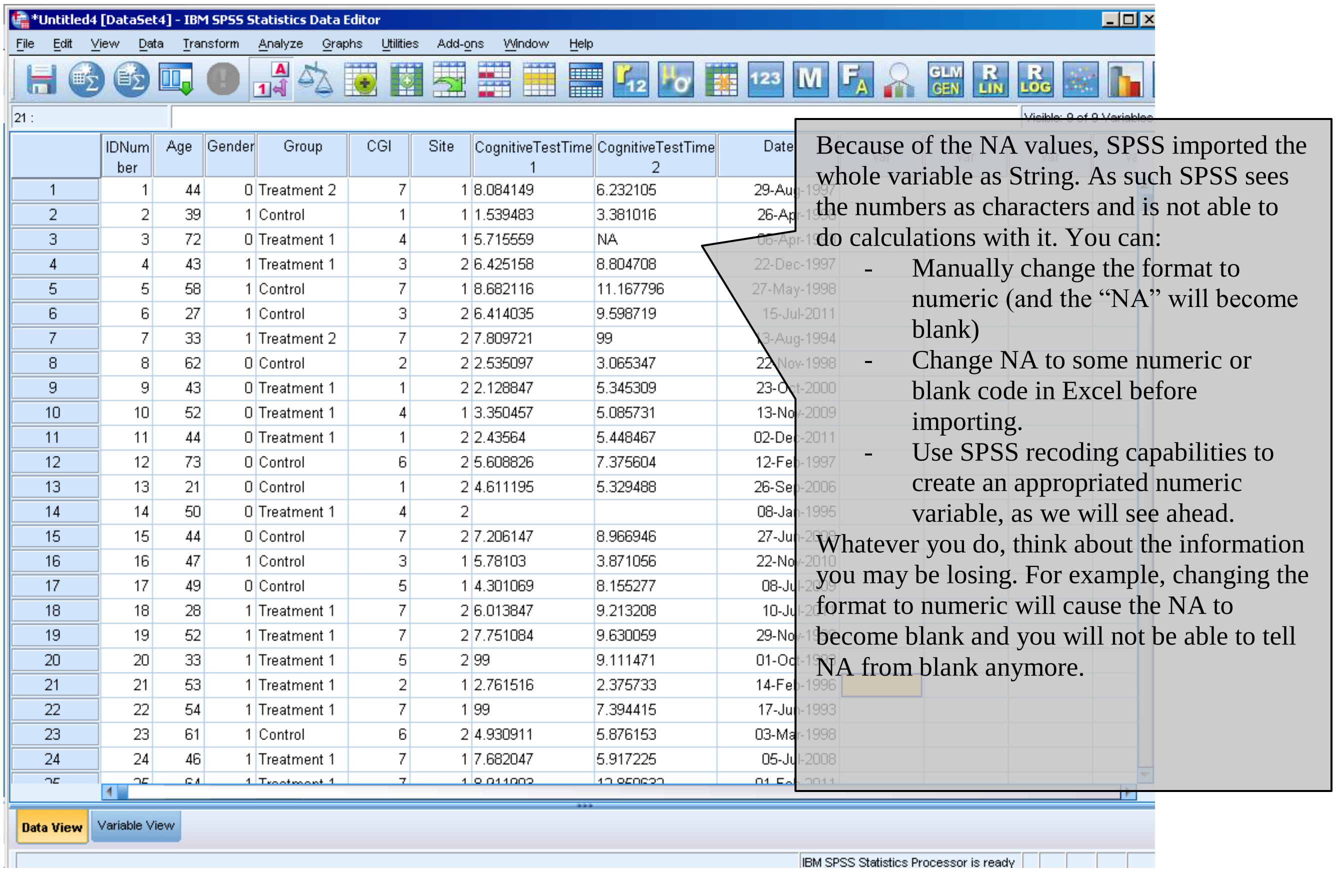


U:\SPSS Course\Data Example.xls

☑ Read variable names from the first row of data

Worksheet: Data [A1:J797]

SPSS will then open the data file. It will usually correctly identify variable types and names. <u>However, it is a good idea to check the Excel data before importing into SPSS, make sure the names of the variables are clear and the values are correct</u>. It is typically easier to make any necessary changes prior to importing into SPSS.


*Untitled4 [DataSet4] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
21 :
Visible: 9 of 9 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready

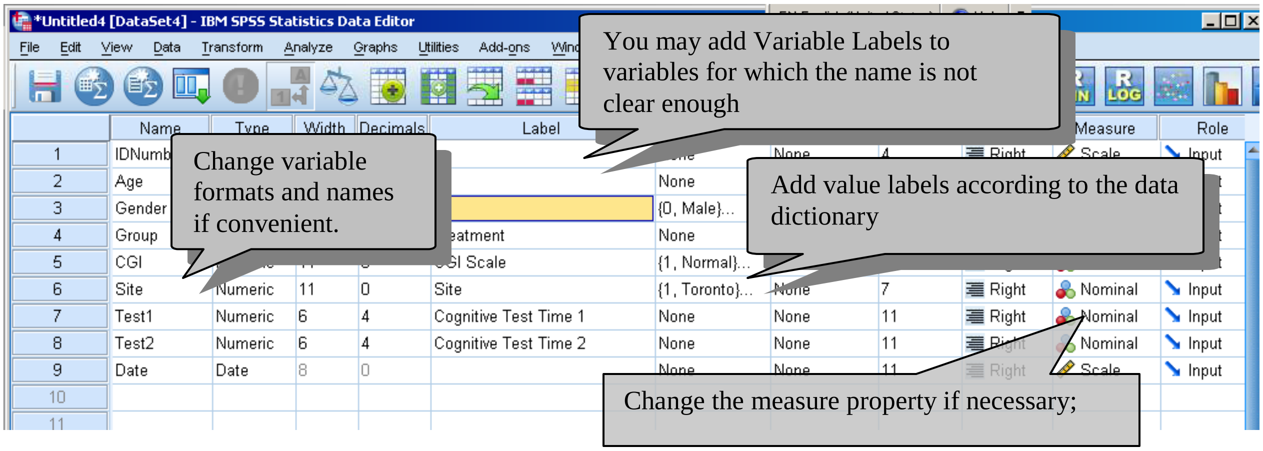
	IDNumber	Age	Gender	Group	CGI	Site	CognitiveTestTime1	CognitiveTestTime2	Date
1	1	44	0	Treatment 2	7	1	8.084149	6.232105	29-Aug-1997
2	2	39	1	Control	1	1	1.539483	3.381016	26-Apr-1998
3	3	72	0	Treatment 1	4	1	5.715559	NA	06-Apr-1999
4	4	43	1	Treatment 1	3	2	6.425158	8.804708	22-Dec-1997
5	5	58	1	Control	7	1	8.682116	11.167796	27-May-1998
6	6	27	1	Control	3	2	6.414035	9.598719	15-Jul-2011
7	7	33	1	Treatment 2	7	2	7.809721	99	13-Aug-1994
8	8	62	0	Control	2	2	2.535097	3.065347	22-Nov-1998
9	9	43	0	Treatment 1	1	2	2.128847	5.345309	23-Oct-2000
10	10	52	0	Treatment 1	4	1	3.350457	5.085731	13-Nov-2009
11	11	44	0	Treatment 1	1	2	2.43564	5.448467	02-Dec-2011
12	12	73	0	Control	6	2	5.608826	7.375604	12-Feb-1997
13	13	21	0	Control	1	2	4.611195	5.329488	26-Sep-2006
14	14	50	0	Treatment 1	4	2			08-Jan-1995
15	15	44	0	Control	7	2	7.206147	8.966946	27-Jun-2008
16	16	47	1	Control	3	1	5.78103	3.871056	22-Nov-2010
17	17	49	0	Control	5	1	4.301069	8.155277	08-Jul-2009
18	18	28	1	Treatment 1	7	2	6.013847	9.213208	10-Jul-2011
19	19	52	1	Treatment 1	7	2	7.751084	9.630059	29-Nov-1998
20	20	33	1	Treatment 1	5	2	99	9.111471	01-Oct-1993
21	21	53	1	Treatment 1	2	1	2.761516	2.375733	14-Feb-1996
22	22	54	1	Treatment 1	7	1	99	7.394415	17-Jun-1993
23	23	61	1	Control	6	2	4.930911	5.876153	03-Mar-1998
24	24	46	1	Treatment 1	7	1	7.682047	5.917225	05-Jul-2008
25	25	64	1	Treatment 1	7	1	8.011993	12.850632	01-Feb-2011

Because of the NA values, SPSS imported the whole variable as String. As such SPSS sees the numbers as characters and is not able to do calculations with it. You can:

- Manually change the format to numeric (and the "NA" will become blank)
- Change NA to some numeric or blank code in Excel before importing.
- Use SPSS recoding capabilities to create an appropriated numeric variable, as we will see ahead.

Whatever you do, think about the information you may be losing. For example, changing the format to numeric will cause the NA to become blank and you will not be able to tell NA from blank anymore.

It is always a good practice to <u>add labels to the data</u>, which facilitates understanding of each variable and their codes; thus we avoid having to refer to a data dictionary for the meaning of the variables. Excel and other types of data file don't have labels.


*Untitled4 [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Wind

	Name	Type	Width	Decimals	Label					Measure	Role
1	IDNumb						None	4	Right	Scale	Input
2	Age					None					
3	Gender					{0, Male}...					
4	Group				eatment	None					
5	CGI				GI Scale	{1, Normal}...					
6	Site	Numeric	11	0	Site	{1, Toronto}...	None	7	Right	Nominal	Input
7	Test1	Numeric	6	4	Cognitive Test Time 1	None	None	11	Right	Nominal	Input
8	Test2	Numeric	6	4	Cognitive Test Time 2	None	None	11	Right	Nominal	Input
9	Date	Date	8	0		None	None	11	Right	Scale	Input
10											
11											

You may add Variable Labels to variables for which the name is not clear enough

Change variable formats and names if convenient.

Add value labels according to the data dictionary

Change the measure property if necessary;

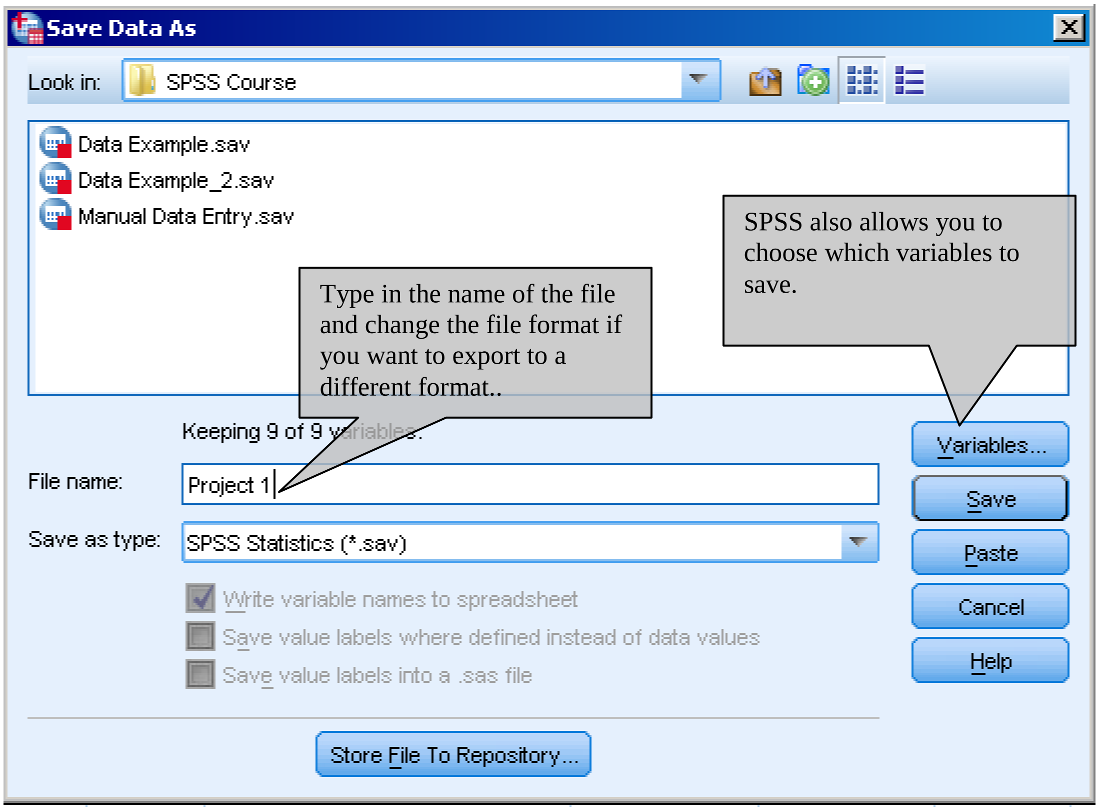
Tip: Your data should always contain a variable that uniquely identifies each subject (Identification Number (ID)). This variable will allow merging data files and possible recovery of information on subjects. It should be created when collecting the data.

Tip: Variable labels can be pasted into SPSS. Both variable and value labels may be easier assigned through the SPSS syntax, so it is a good idea to learn that.

• Saving an SPSS Data File

SPSS saves the data as any other program.

> Go to File → Save As, and a window will appear that allows you to write the name of the file to be saved. This same window also allows you to export the data to many other formats; if you want to save in a format different than SPSS, just change the file type in the box “Save as type”. Notice that SPSS data files will be saved with extension “.sav”.



Save Data As

Look in: SPSS Course

Data Example.sav
Data Example_2.sav
Manual Data Entry.sav

Keeping 9 of 9 variables.

File name: Project 1

Save as type: SPSS Statistics (*.sav)

☑ Write variable names to spreadsheet
☐ Save value labels where defined instead of data values
☐ Save value labels into a .sas file

Type in the name of the file and change the file format if you want to export to a different format..

SPSS also allows you to choose which variables to save.

Tip: Always save a copy of the original data before you do any analysis or modification. Work with a working version of the data, keeping the original untouched.

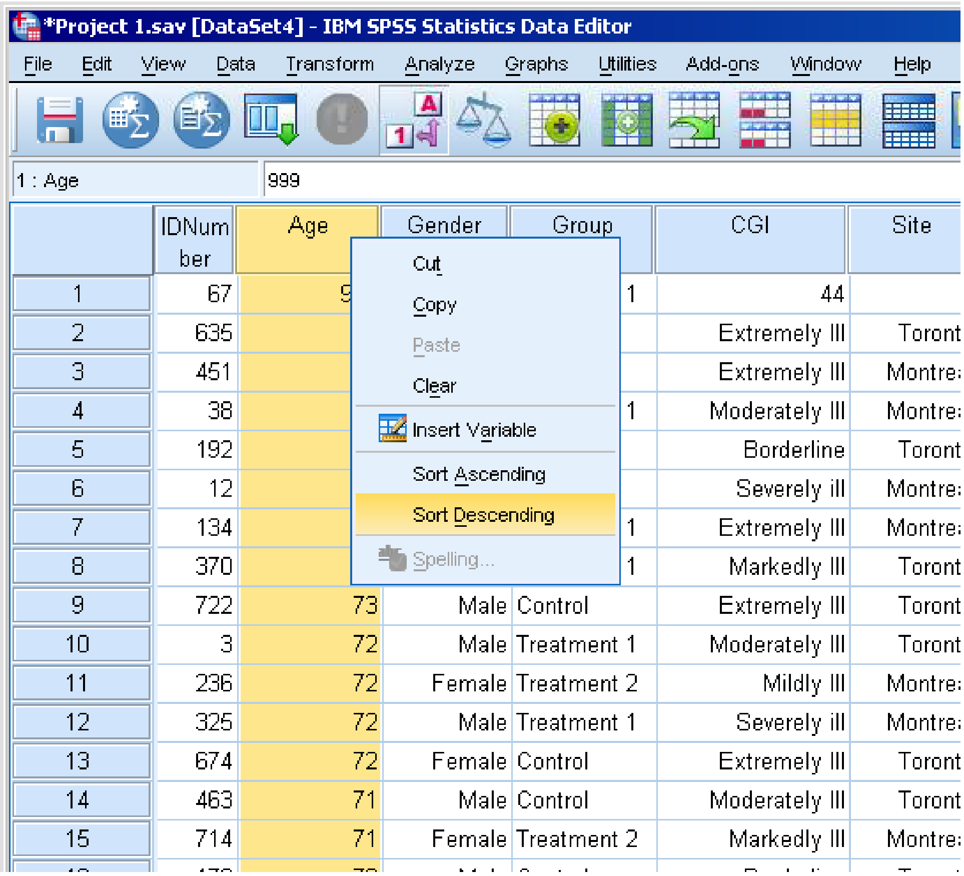
• Initial Data Checking after Importing

When opening a new data file, SPSS will keep also open any existing data files you are working on. Thus you can work with many data files in a single SPSS session. This may be both useful and confusing; as a beginner we recommend you to work with only one file at a time. So at this point you may keep only this imported data file open (saved as "Project 1.sav" above).

Upon getting the data into SPSS we should make sure it looks ok. There are a few things we can do to quickly check our data:

a) **Visually inspect the data**, **variable names, number of cases**. Look at variable values and make sure they make sense. You can navigate through SPSS data as you would in Excel and take a quick look at it.

b) **Sort the data**. By pointing to the variable name and right-clicking you will get a menu. If your data is not huge, sorting will allow you to quickly inspect the maximum and minimum of the variables and that may uncover some problems, i.e. extreme values that should not be there. For example, by sorting Age we may find subjects that are too young and should not be part of the analysis. Or some ages may be to great to be possible. Notice that the menu also allows you to insert or delete (clear) variables, and also to copy or cut variables.



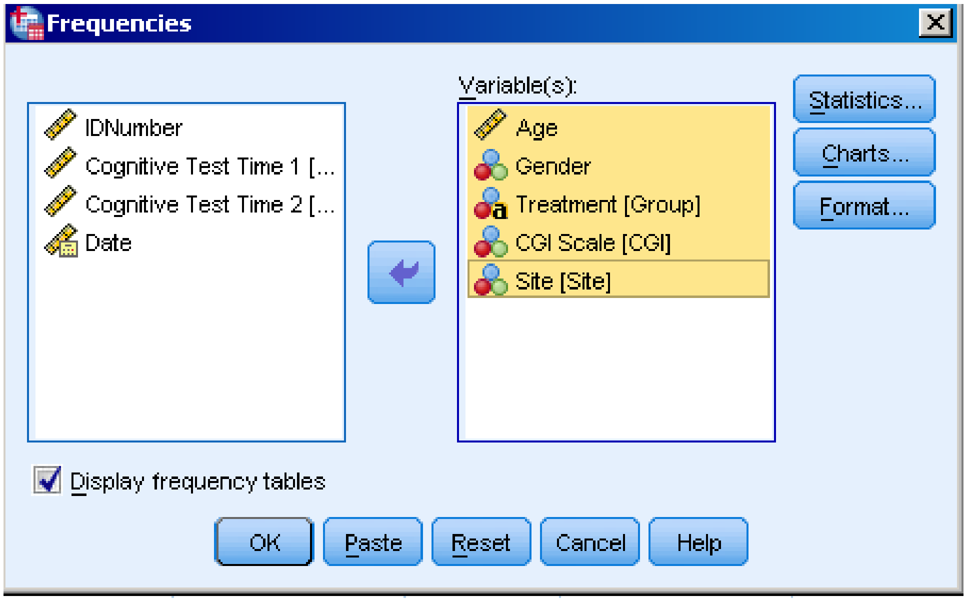
1 : Age 999

	IDNumber	Age	Gender	Group	CGI	Site
1	67	9		1	44	
2	635				Extremely Ill	Toront
3	451				Extremely Ill	Montre
4	38			1	Moderately Ill	Montre
5	192				Borderline	Toront
6	12				Severely ill	Montre
7	134			1	Extremely Ill	Montre
8	370			1	Markedly Ill	Toront
9	722	73	Male	Control	Extremely Ill	Toront
10	3	72	Male	Treatment 1	Moderately Ill	Toront
11	236	72	Female	Treatment 2	Mildly Ill	Montre
12	325	72	Male	Treatment 1	Severely ill	Montre
13	674	72	Female	Control	Extremely Ill	Toront
14	463	71	Male	Control	Moderately Ill	Toront
15	714	71	Female	Treatment 2	Markedly Ill	Montre

c) For variables that don't have too many levels, we can look at the **Frequency Tables**. It will show you the distribution of unique values in the data file. Let's create frequency tables of Age and Gender. We will go through the Frequency procedure with more detail later in this workshop.

   Click in Analysis → Descriptive Statistics → Frequencies. The window below will show up. Now you can select Age and move it to the right, by clicking in the middle button. Do the same with Gender, and click OK.

Tip – In the Frequency procedure, the box showing available variables may be displaying Variable Labels. Because labels may be long it is usually better to show Variable Names. To show names instead of labels, go to Edit -> Options and in the Tab "General" click "Display Names". This way SPSS will always show names in the analysis boxes.



IDNumber
Cognitive Test Time 1 [...
Cognitive Test Time 2 [...
Date

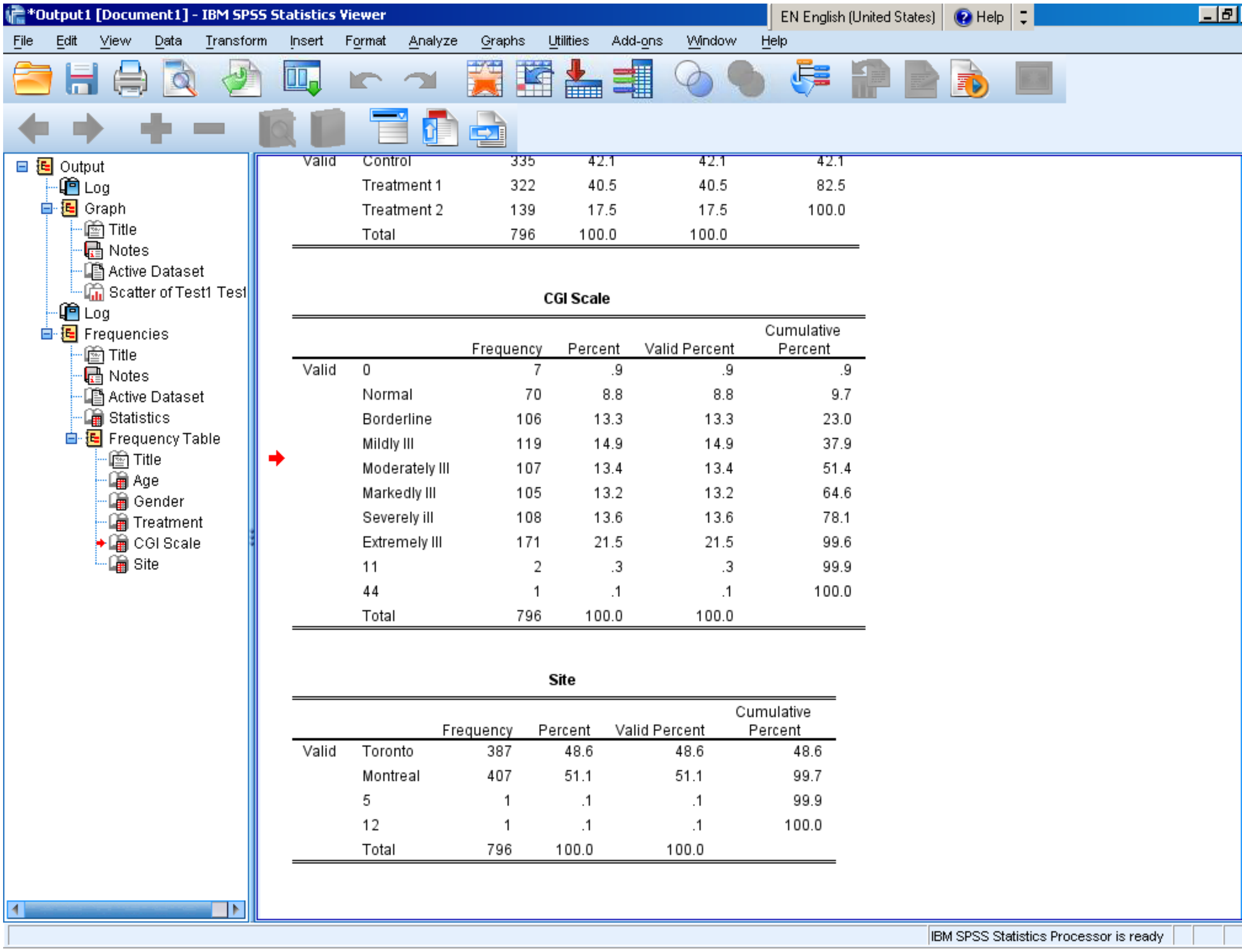
Variable(s):
Age
Gender
Treatment [Group]
CGI Scale [CGI]
Site [Site]

☑ Display frequency tables

After clicking ok, the result will appear in the SPSS Output. You will see one table for each variable you selected. On the left side you have a navigation pane which looks like Windows Explorer and helps you to navigate through your output, especially when you have a long output. We will talk more about the output soon.

The Frequency will show you the content of the variable and its distribution. For example, we can see that there are some strange values for CGI, like 0, 11 and 44. We may need to understand what is going on before doing any analysis. Maybe a 4 was incorrectly entered as 44? Is it a specific code for something like "Unknown"? The frequency table also show us basic counts, for example, 70 subjects have CGI = Normal, which is 8.8% of the subjects.

Tip – If you do not have too many variable or too many cases, you may want to include all the variables in the frequency table. This is a quick way of looking at all the variables to see if the results are what we expect. Some frequency tables may be huge and not useful (like the one for the Identification number), but you can easily delete these tables (click on it and hit the delete key) and move along.



Valid	Control	335	42.1	42.1	42.1
	Treatment 1	322	40.5	40.5	82.5
	Treatment 2	139	17.5	17.5	100.0
	Total	796	100.0	100.0	

CGI Scale

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	7	.9	.9	.9
	Normal	70	8.8	8.8	9.7
	Borderline	106	13.3	13.3	23.0
	Mildly Ill	119	14.9	14.9	37.9
	Moderately Ill	107	13.4	13.4	51.4
	Markedly Ill	105	13.2	13.2	64.6
	Severely ill	108	13.6	13.6	78.1
	Extremely Ill	171	21.5	21.5	99.6
	11	2	.3	.3	99.9
	44	1	.1	.1	100.0
	Total	796	100.0	100.0	

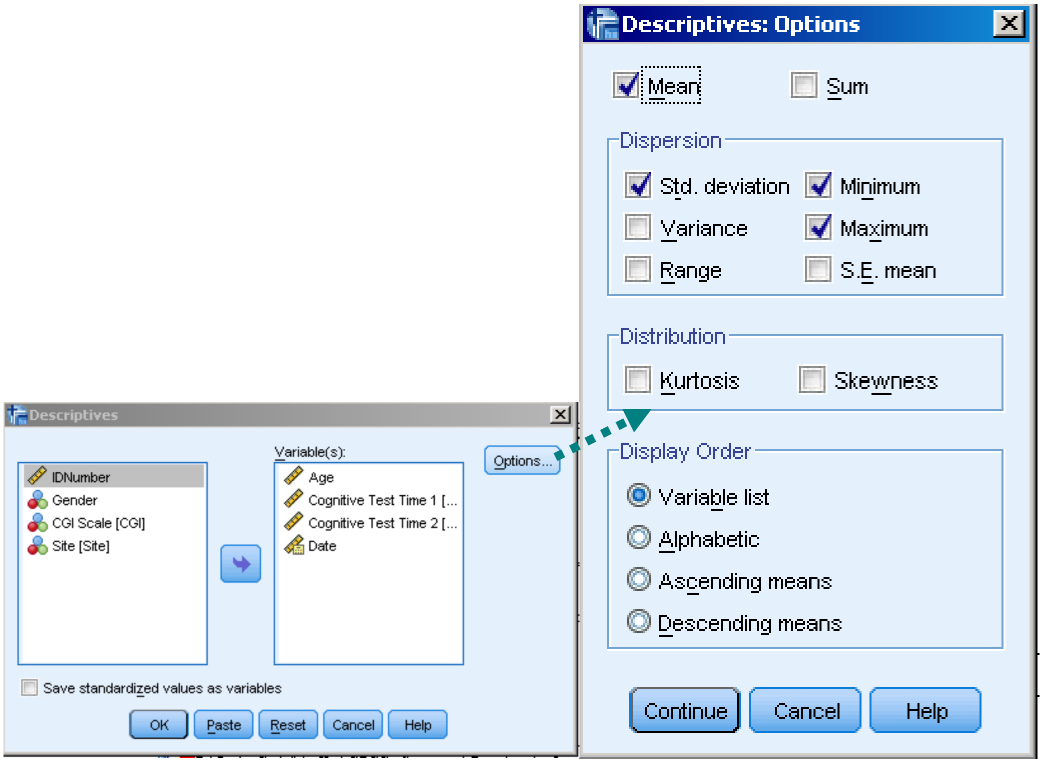
Site

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Toronto	387	48.6	48.6	48.6
	Montreal	407	51.1	51.1	99.7
	5	1	.1	.1	99.9
	12	1	.1	.1	100.0
	Total	796	100.0	100.0	

When checking the data it is important to think about whether what we see makes sense and is expected. As for missing values, it is important to think if they are expected, if they can be recovered somehow as they can be limiting factors for statistical analyses and they can introduce biases in the results.

d) **Look at Descriptive Statistics**. For variables that are **<u>continuous</u>** - those for which distance and arithmetic operations makes sense - we can look at summary statistics like the mean, variance, median, minimum, maximum, etc. These measures summarize the content of variables and can give us an idea if everything is ok. We will exemplify this procedure quickly now and go into more details further in the training.

Click in Analysis → Descriptive Statistics → Descriptive. SPSS will only show numeric variables for this analysis because it cannot be done with string variables. Remember that variables like CGI are numeric but not continuous, therefore descriptive statistics are usually not appropriate, although possible. IDNumber is also not continuous. So let's move Age and the Cognitive Tests to the "Variable" box. Then click "Options" and select the summary statistics you would like to see. Click "Continue" and "OK".



Descriptives

IDNumber
Gender
CGI Scale [CGI]
Site [Site]

Variable(s):
Age
Cognitive Test Time 1 [...
Cognitive Test Time 2 [...
Date

□ Save standardized values as variables

Descriptives: Options

☑ Mean □ Sum

Dispersion
☑ Std. deviation ☑ Minimum
□ Variance ☑ Maximum
□ Range □ S.E. mean

Distribution
□ Kurtosis □ Skewness

Display Order
● Variable list
○ Alphabetic
○ Ascending means
○ Descending means

SPSS will calculate the requested summary statistics for each of the chosen variables. You can look at this output and see if anything does not look right.

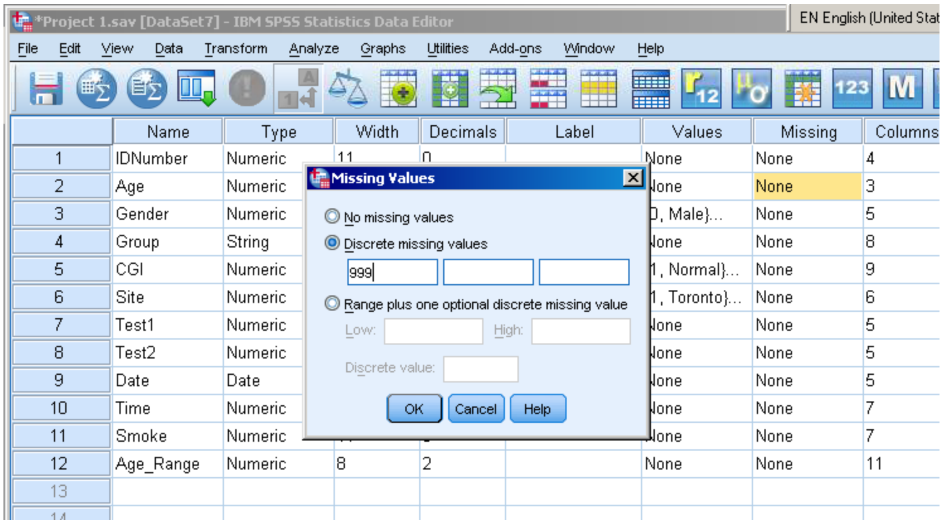
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age	796	18	999	55.39	122.239
Cognitive Test Time 1	737	1.5395	99.0000	10.038310	19.3627736
Cognitive Test Time 2	756	-1.0000	99.0000	10.724696	18.4426700
Date	793	01.01.92	08.12.11	17.02.02	
Valid N (listwise)	715				

Tip – Pay attention to means that are too large or too low. The maximum and minimum values are very useful for checking to ensure that the data is within the expected range.

Declaring Missing Values

When a variable like Age has a value like 999 it is important that we let SPSS knows that this 999 is a code for something other than age (perhaps it indicates “Unknown age” or “Refusal” or something else), that is, it is an invalid value. To declare is as missing value we go to the “Variable View” tab and click in the “Missing” column, variable Age. The Missing Values window will show up. We can then easily enter the value 999 as a discrete missing value and click “OK”. From now on SPSS will exclude 999 from statistical analyses using Age. A similar procedure can be carried out for the cognitive test variables.

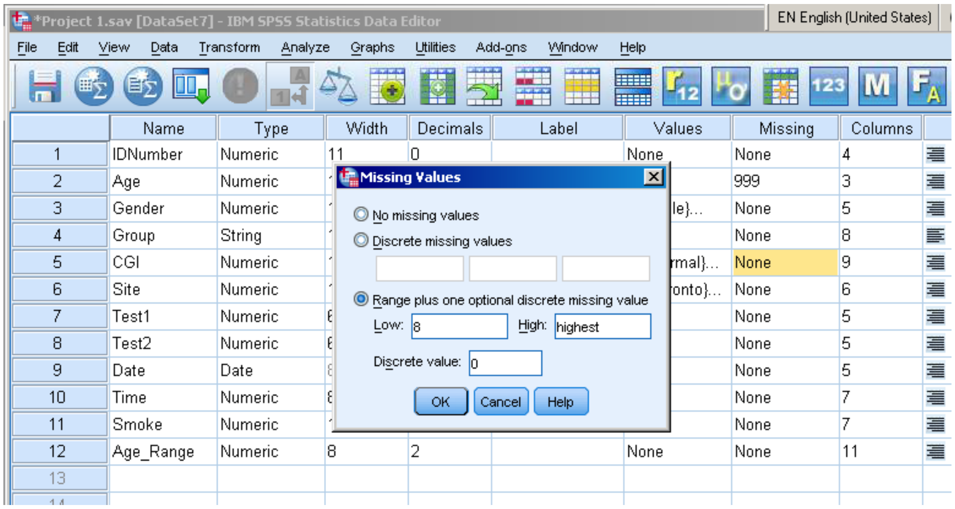


	Name	Type	Width	Decimals	Label	Values	Missing	Columns
1	IDNumber	Numeric	11	0		None	None	4
2	Age	Numeric				None	None	3
3	Gender	Numeric				0, Male}...	None	5
4	Group	String				None	None	8
5	CGI	Numeric				1, Normal}...	None	9
6	Site	Numeric				1, Toronto}...	None	6
7	Test1	Numeric				None	None	5
8	Test2	Numeric				None	None	5
9	Date	Date				None	None	5
10	Time	Numeric				None	None	7
11	Smoke	Numeric				None	None	7
12	Age_Range	Numeric	8	2		None	None	11
13								
14								

Missing Values

○ No missing values
● Discrete missing values
999
○ Range plus one optional discrete missing value
Low: High:
Discrete value:

In the case of CGI we can see more values that don’t belong to the scale range (1 to 7). The option “Range plus one optional discrete missing value” allows us to declare them as missing values. Notice that SPSS will accept values “lowest” and “highest” to declare unbounded ranges.


*Project 1.sav [DataSet7] - IBM SPSS Statistics Data Editor
EN English (United States)
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns
1	IDNumber	Numeric	11	0		None	None	4
2	Age	Numeric	1				999	3
3	Gender	Numeric	1			le}...	None	5
4	Group	String	1				None	8
5	CGI	Numeric	1			rmal}...	None	9
6	Site	Numeric	1			onto}...	None	6
7	Test1	Numeric	8				None	5
8	Test2	Numeric	8				None	5
9	Date	Date	8				None	5
10	Time	Numeric	8				None	7
11	Smoke	Numeric	1				None	7
12	Age_Range	Numeric	8	2		None	None	11
13								

Missing Values

○ No missing values

○ Discrete missing values

● Range plus one optional discrete missing value

Low: 8 High: highest

Discrete value: 0

OK Cancel Help

SPSS has two types of missing values:

- **System Missing** – blank cases, represented in the data by a period, which SPSS already knows is a missing value and will not use in any analysis.
- **User Defined Missing** – are actual values that are defined by the user as being missing, which is what we did above.

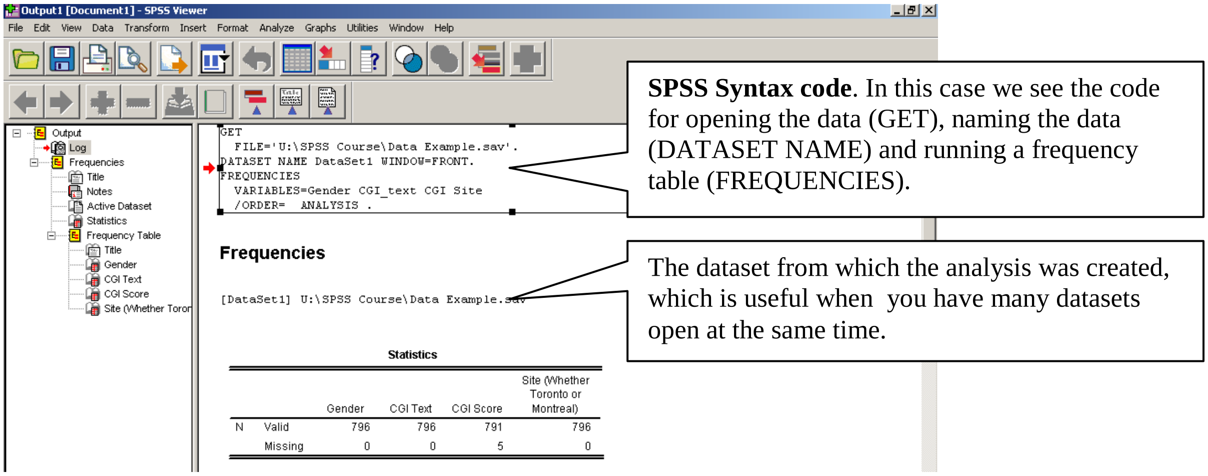
- **The SPSS Output**

As you saw, the SPSS shows the results of the analyses in the Output Window.

You can save the Output by clicking Edit → Save As and just save the Output normally. It will have the extension ".spo".

Anything in the Output Window can be copied and pasted into software like Word and Excel. Just right-click over the target object and choose the option "Copy". Then you can go to Word, Excel, Outlook, etc, and paste it there.

By default, SPSS will show the Syntax (code) of what you ran just before the analysis results. Anything you run through the menus can also be run through syntax, which is the SPSS programming language. We will talk more about that. It will also the display name of the data file from which data file the results were derived. These items are respectively the "Log" and "Active Dataset" in the left pane. You will also see the item "Notes" in the left pane, and if you double click it (it is hidden by default, so you see it only on the left pane, not on the right), you will see detailed information about the analysis. These are all important in case you need to replicate the analysis in the future, so it is a good idea to save the output with them.



```
GET
  FILE='U:\SPSS Course\Data Example.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
FREQUENCIES
  VARIABLES=Gender CGI_text CGI Site
  /ORDER=  ANALYSIS .
```

Frequencies

```
[DataSet1] U:\SPSS Course\Data Example.sav
```

Statistics

		Gender	CGI Text	CGI Score	Site (Whether Toronto or Montreal)
N	Valid	796	796	791	796
	Missing	0	0	5	0

SPSS Syntax code. In this case we see the code for opening the data (GET), naming the data (DATASET NAME) and running a frequency table (FREQUENCIES).

The dataset from which the analysis was created, which is useful when you have many datasets open at the same time.

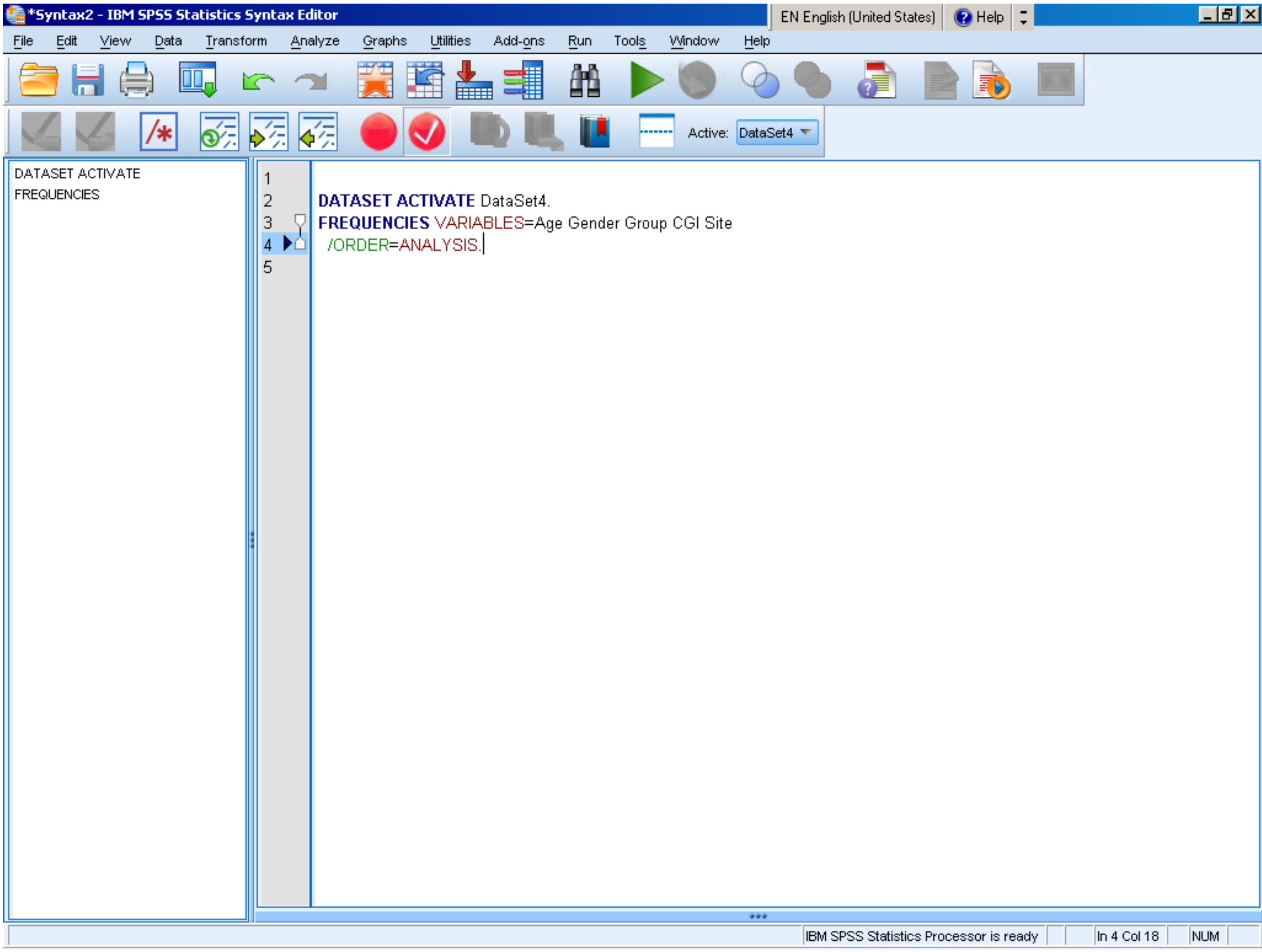
Tip – By double clicking items you can edit and format them. You can also delete items that are not of interest before saving the output, just by selecting the item and pressing "Delete". In "Insert" menu, you are able to insert new items in the output, like text explaining the tables.

The SPSS Output can also be exported, which is sometimes handy. Click File → Export and a new window will open with several options for exporting the output. If your goal is just to have the results in a Excel spreadsheet try to copy and paste the desired output table.

• The SPSS Syntax Editor

SPSS Syntax is the SPSS text editor, where you can write code to perform any task in SPSS. The code language is actually more powerful than the point-and-click menus, i.e. you can do things with syntax that you cannot with the point-and-click interface, including using macros.

> Go to Analysis → Descriptives → Frequencies and insert some variables like Gender, group, CGI in the "Variable" box. Then click "Paste" instead of the usual "OK".



```
DATASET ACTIVATE DataSet4.
FREQUENCIES VARIABLES=Age Gender Group CGI Site
  /ORDER=ANALYSIS.
```

SPSS will not run the analysis. It will instead paste the code for the analysis into the Syntax Editor.

> In order to run the analysis from the Syntax Editor, select the code you want to run and press the "Play" button. Do that with the Frequency code you just pasted.

So as you can see, another way of running the Frequency procedure is to type in the appropriate code for it, select it and run it. Experienced users often find it easier and quicker to just type in syntax instead of using the pull-down menus, in part because SPSS has shortcuts for codes. For example, the same result would be obtained if you typed `“fre gender group cgi cgi_text”.`

The code is a very important part of any analysis project because it offers a way of documenting the analyses that were conducted, as well as changes made to the data. If you always save your syntax, you will be able to easily replicate the analysis which is very important. We recommend that you use the “Paste” button regularly and save the syntax associated with everything you do.

Tip – Although SPSS syntax may seem complicated and difficult to learn at first, it gets easier with practice. It offers a significant advantage for data manipulation and it is worth keeping an open mind about learning SPSS syntax instead of relying exclusively on the point-and-click approach.

• Merge Files

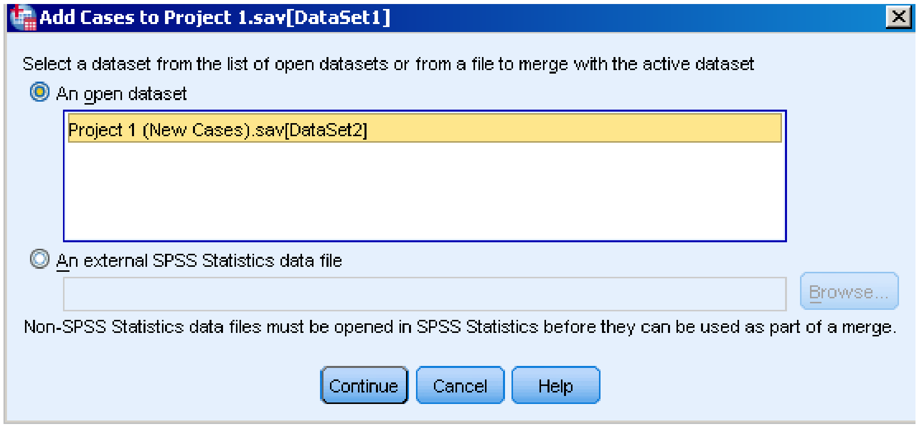
Merging is the process of adding information to a data file in form of new cases or new variables. For the purpose of introducing this topic we will consider a standard file structure in which each row of data represents information from a unique subject or study participant, and each column represents a different variable i.e. attribute or characteristic of that study participant.

• Adding new cases

The new cases should be in SPSS format. Before going through the process, there are a few things you should check:

- Check that the variable names are the same in both files. SPSS will let you change variable names in the middle of the process but it is always easier to make these modifications in advance so that SPSS can automatically pair the current variables with the variables in the new aggregate file.
- String variables need to have the same size in both current and new data. If they don't have we recommend you to change the size of the smaller string variable to match the larger one. For example, if NAME is a string variable with size 100 in the current file and 130 in the new cases then you should make the size in the current file 130.
- Assure values have the same meaning in both files. For example, if in the current file 0 = Male, 1 = Female, 98 = Not known and 99 = Refusal then these exact codes must be used in the data file with new cases. The importance of this cannot be stressed enough as adding cases with different codes may invalidate all the analyses that use the affected variable.
- Sometimes you may have variables in your new file that do not exist in the current data and vice-versa. Be aware of these variables and make a decision of whether they should be included the final data set or not.

Let's add cases to our current data file. **In the current data file**, go to Data → Merge Files → Add cases. SPSS will need to know from which file. If the file with new cases is open, you will be able to select it from the list of open datasets, as in the image below. If it is not open you will need to locate it through checking "An external SPSS Statistics data file" and clicking "Browse".


Add Cases to Project 1.sav[DataSet1]

Select a dataset from the list of open datasets or from a file to merge with the active dataset

● An open dataset

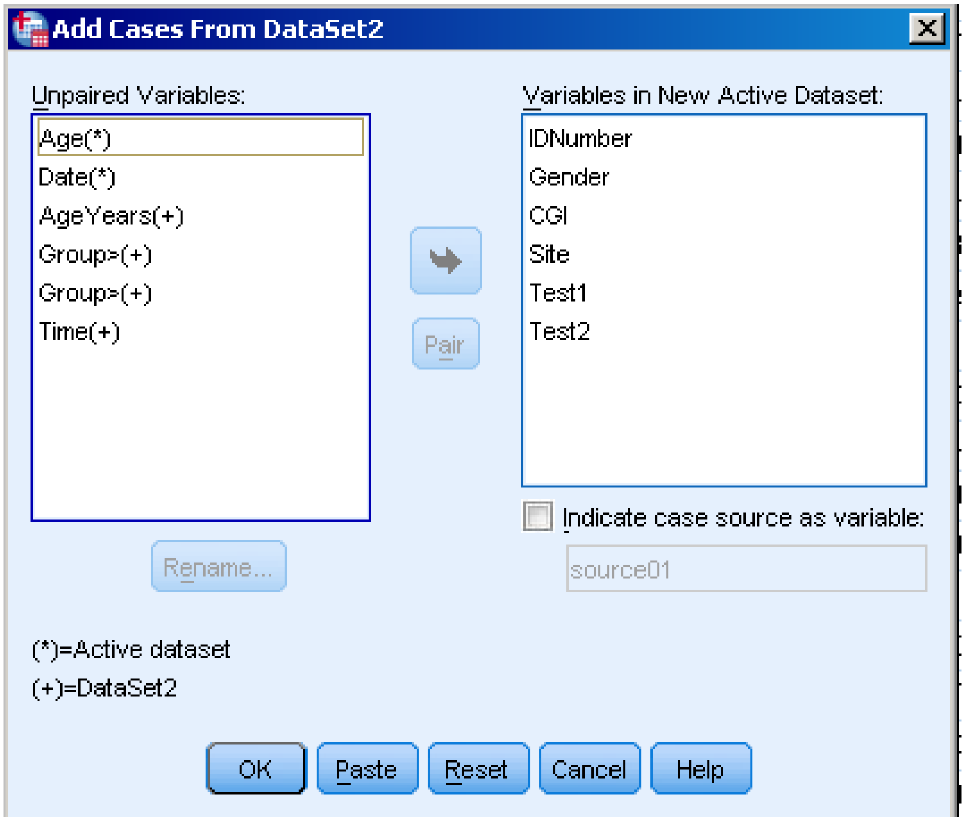
Project 1 (New Cases).sav[DataSet2]

○ An external SPSS Statistics data file

Non-SPSS Statistics data files must be opened in SPSS Statistics before they can be used as part of a merge.

Once you select the data file to be added to the current data, SPSS will compare them and show the window below where you will see:

- On the right side are the variables that were paired, that is, SPSS found that they are in both datasets and at this point they will be the only ones present in the new merged dataset.
- On the left side you have the unpaired variables. These are variables with different formatting or variables present in only one of the two files.
- “Age” and “AgeYear” are the same variable but have different names. In this case you can select one of them, hold CTRL and select the other, then press Pair. By doing that you are telling SPSS that they are the same variable and should be in the same column. You could also press “Cancel” and change the name of the variable in one of the files so that on repeating the process SPSS will automatically pair both and put them in the right box.
- “Group” is present in both files, but SPSS is not pairing them. That is because they are String variables with different lengths. <u>You should Cancel the dialogue and assure that they have the same size in both files, then repeat the process.</u>
- Variables “Date” and “Time” measure different things and should not be paired. You need to decide if you want them in the merged dataset of not. Notice that if you bring them to the new merged dataset, “Date” will be missing for the current cases and “Time” will be missing for the new cases.



Unpaired Variables:

- Age(*)
- Date(*)
- AgeYears(+)
- Group>(+)
- Group>(+)
- Time(+)

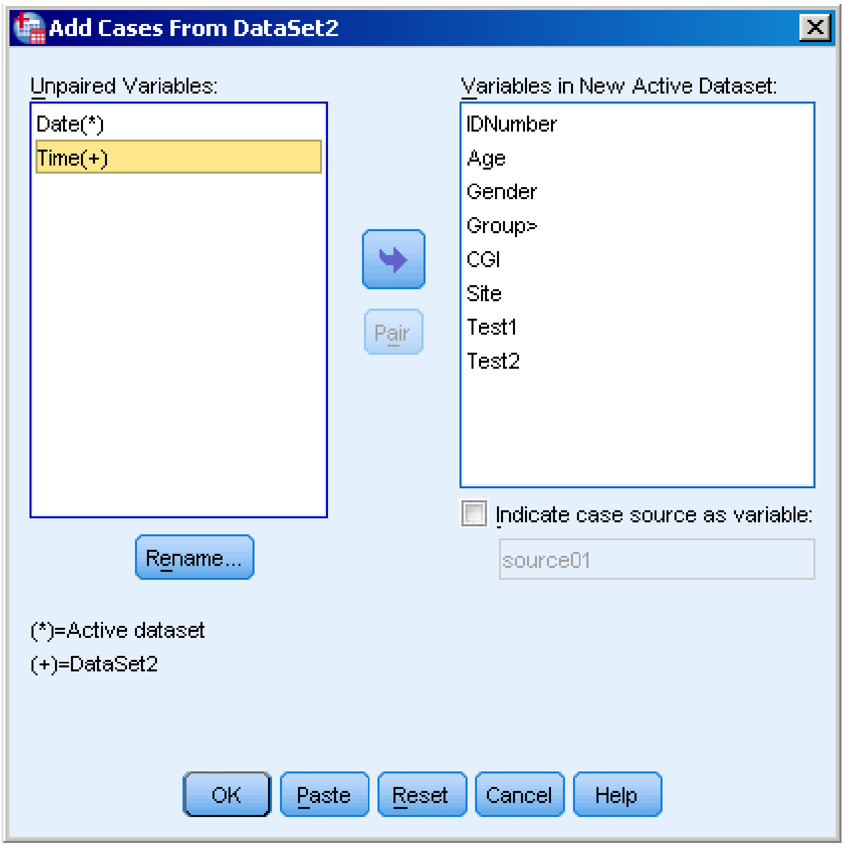
Variables in New Active Dataset:

- IDNumber
- Gender
- CGI
- Site
- Test1
- Test2

□ Indicate case source as variable: source01

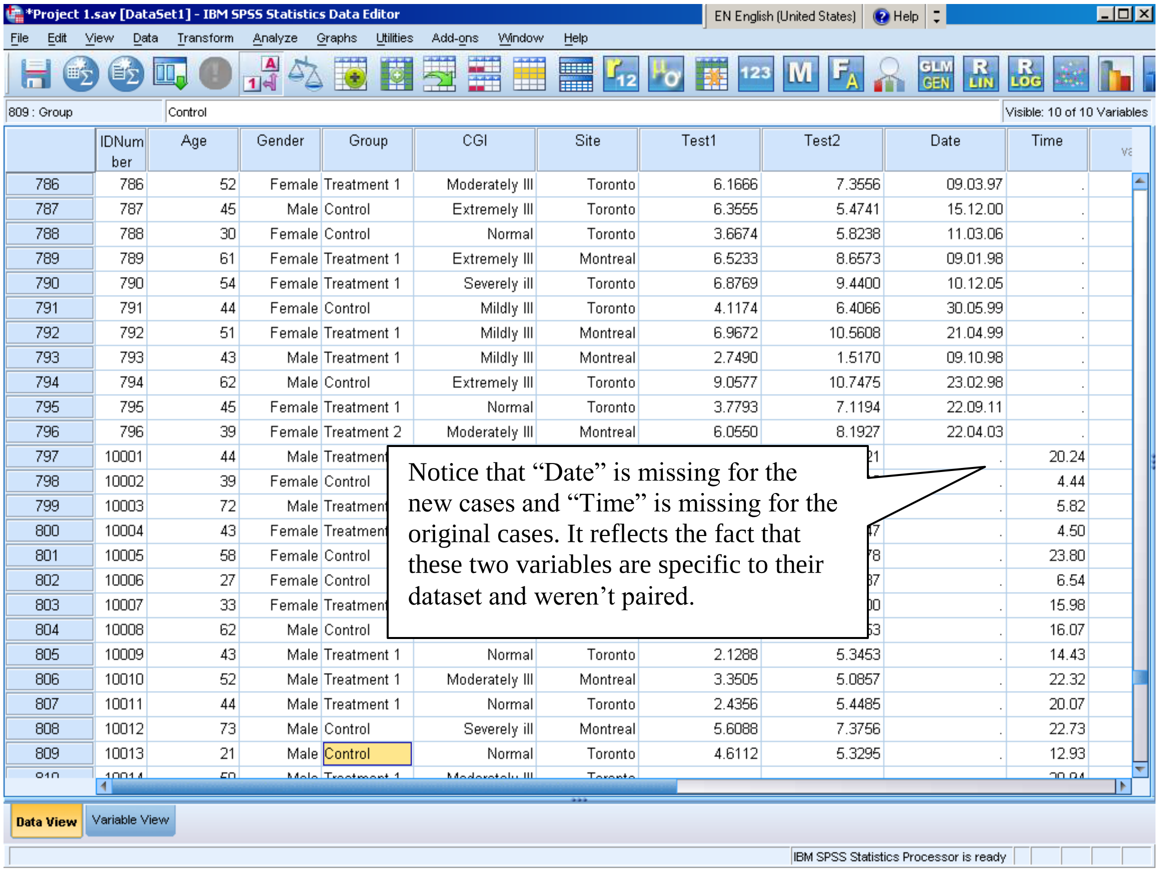
(*)=Active dataset
(+)=DataSet2

Cancel the process and changed variables “Group” and “Age” so that they are identical in both datasets. SPSS can now pair all variables except for “Date” and “Time”. In order not to lose the information you may want to select and move them to the right box. Then press “Ok”.



Below we can see that the new data is added to the bottom of the current data. We can also easily see that the variables "Date" and "Time" are missing for the cases not from their respective datasets.

Tip – If you have difficulties on seeing how adding cases works, try and do it manually, in Microsoft Excel, by copying the new data and pasting it at the bottom of the current dataset. The option of doing it manually is usually a good one if your data does not have too many variables. And remember to always check the merged dataset to ensure the merge


*Project 1.sav [DataSet1] - IBM SPSS Statistics Data Editor
EN English (United States)
Help
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
Data View
Variable View
IBM SPSS Statistics Processor is ready

809 : Group Control Visible: 10 of 10 Variables

	IDNumber	Age	Gender	Group	CGI	Site	Test1	Test2	Date	Time
786	786	52	Female	Treatment 1	Moderately Ill	Toronto	6.1666	7.3556	09.03.97	.
787	787	45	Male	Control	Extremely Ill	Toronto	6.3555	5.4741	15.12.00	.
788	788	30	Female	Control	Normal	Toronto	3.6674	5.8238	11.03.06	.
789	789	61	Female	Treatment 1	Extremely Ill	Montreal	6.5233	8.6573	09.01.98	.
790	790	54	Female	Treatment 1	Severely ill	Toronto	6.8769	9.4400	10.12.05	.
791	791	44	Female	Control	Mildly Ill	Toronto	4.1174	6.4066	30.05.99	.
792	792	51	Female	Treatment 1	Mildly Ill	Montreal	6.9672	10.5608	21.04.99	.
793	793	43	Male	Treatment 1	Mildly Ill	Montreal	2.7490	1.5170	09.10.98	.
794	794	62	Male	Control	Extremely Ill	Toronto	9.0577	10.7475	23.02.98	.
795	795	45	Female	Treatment 1	Normal	Toronto	3.7793	7.1194	22.09.11	.
796	796	39	Female	Treatment 2	Moderately Ill	Montreal	6.0550	8.1927	22.04.03	.
797	10001	44	Male	Treatment				1	.	20.24
798	10002	39	Female	Control					.	4.44
799	10003	72	Male	Treatment					.	5.82
800	10004	43	Female	Treatment				7	.	4.50
801	10005	58	Female	Control				8	.	23.80
802	10006	27	Female	Control				7	.	6.54
803	10007	33	Female	Treatment				0	.	15.98
804	10008	62	Male	Control				3	.	16.07
805	10009	43	Male	Treatment 1	Normal	Toronto	2.1288	5.3453	.	14.43
806	10010	52	Male	Treatment 1	Moderately Ill	Montreal	3.3505	5.0857	.	22.32
807	10011	44	Male	Treatment 1	Normal	Toronto	2.4356	5.4485	.	20.07
808	10012	73	Male	Control	Severely ill	Montreal	5.6088	7.3756	.	22.73
809	10013	21	Male	Control	Normal	Toronto	4.6112	5.3295	.	12.93
810	10014	50	Male	Treatment 1	Moderately Ill	Toronto				20.04

Notice that "Date" is missing for the new cases and "Time" is missing for the original cases. It reflects the fact that these two variables are specific to their dataset and weren't paired.

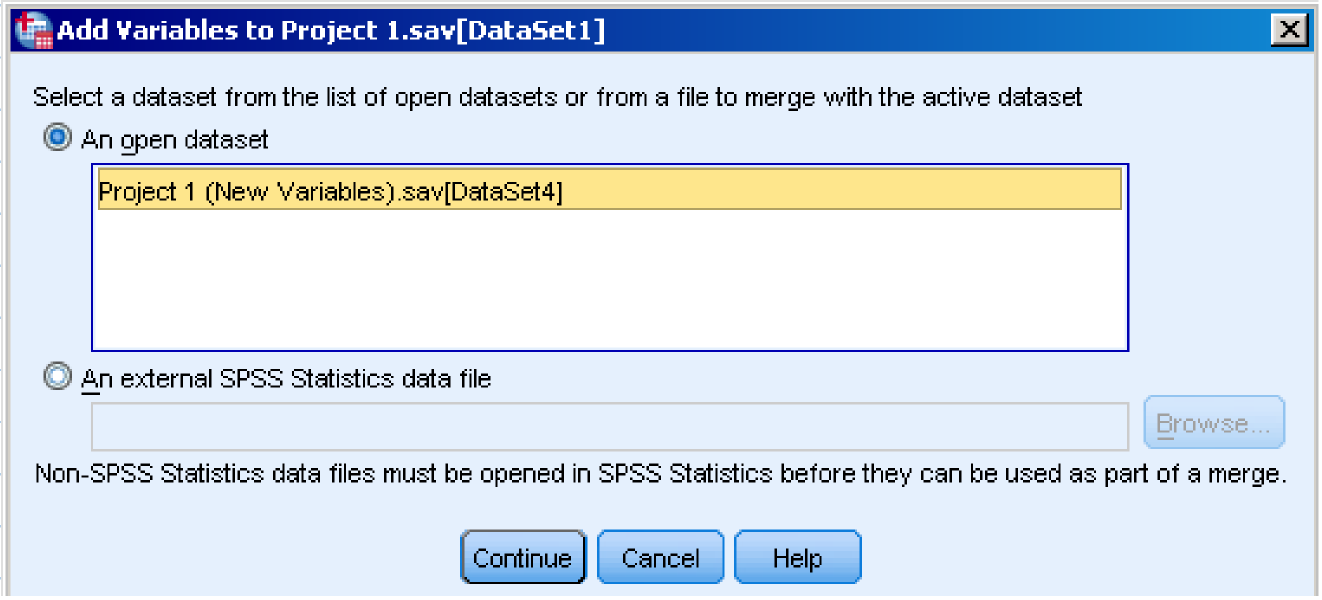
- **Adding new variables**

In this scenario we want to add new information in the form of variables to our existing dataset. Rather than updating our file by adding information about new subjects, we are adding additional characteristics, test scores, etc. about our existing group of subjects. It is *very important* that we successfully match these new characteristics to the correct subjects in our existing file. For example, suppose we would like to know whether smoking status is associated with cognitive test scores, but these two pieces of information are captured in different files. We will need to merge these files together in order to conduct this analysis. To do this:

- We need to **uniquely identify subjects** in both datasets.
  - Unique subject identification is typically accomplished by assigning each subject **an ID** number and including this variable **in both datasets**. It can be a number assigned at the start of the study or it can be some existing information, like health card number. Because it must be unique, things like the name of the person or birth date should be avoided as we cannot guarantee they will be unique.
  - Subject identification is very important. If you receive a dataset without a variable that clearly uniquely identify subjects you should create an identification number in the data and also follow up with the dataset provider to see if they have an identifier.

- We need to ensure that our unique identifyier is the same in both datasets, that is, ID = 1 refers to the same subject in both datasets. Some times identifiers are assigned to subject by different processes and the same identifier may refer to different subjects.
- We need **to understand** the type of information we have in the external (smoking) dataset. For example:
  - We may have the smoking status for every subject in our dataset and only them;
  - We may have the smoking status for every subject in our data and other subjects that are not in our data;
  - We may have smoking status for many subjects, some of whom are in our dataset. But smoking status is not available for every subject in our dataset.
  - Perhaps we have smoking status for the same subject at different time points.
- In order to merge the files, SPSS will require that both files **are sorted** by the identifier, in ascending order, which must have the same variable name in both datasets. Hence, for both datasets, right click on the identifier name and select "Sort ascending" and save the datasets. At this point you are ready to bring the smoking status variable into your study dataset.

  From the study dataset, select Data →Merge files→Add variables. Select the dataset with the smoking status (it needs to be in SPSS format) and click "Continue"..


Add Variables to Project 1.sav[DataSet1]

Select a dataset from the list of open datasets or from a file to merge with the active dataset

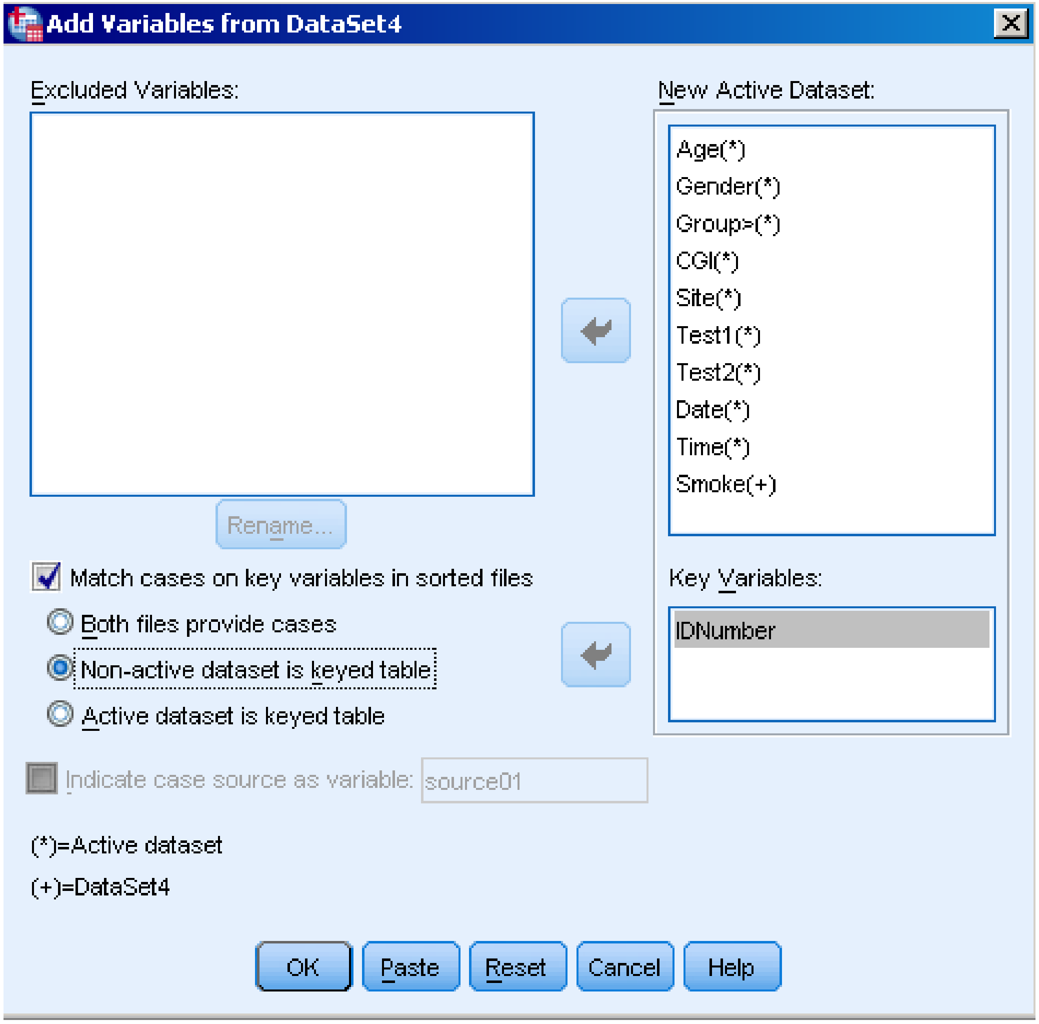
● An open dataset

Project 1 (New Variables).sav[DataSet4]

○ An external SPSS Statistics data file

Non-SPSS Statistics data files must be opened in SPSS Statistics before they can be used as part of a merge.

You will see a new window. SPSS will compare both datasets and variables that are in both datasets will be listed in the "Excluded variable" box. The subject ID variable must be there. On the right side you will see the variables that will be present in the new dataset and you can exclude some of them if you want to. Variables flagged with a "star" are the ones already in your study dataset. The ones with a "plus" sign are the variables that will be brought into your study dataset from the new dataset. In this case we can see that "Smoke" is the only new variable.



Add Variables from DataSet4

Excluded Variables:

New Active Dataset:
Age(*)
Gender(*)
Group=(*)
CGI(*)
Site(*)
Test1(*)
Test2(*)
Date(*)
Time(*)
Smoke(+)

Key Variables:
IDNumber

☑ Match cases on key variables in sorted files
○ Both files provide cases
● Non-active dataset is keyed table
○ Active dataset is keyed table

□ Indicate case source as variable: source01

(*)=Active dataset
(+)=DataSet4

The variable "IDNumber" is also called "Key variable" and should be moved to the "Key Variable" box. There are different a variable can be added:

1) If you do not move the ID variable to the "Key Variable" box, SPSS will do the merge like a "Copy & Paste" matching data from the first line of your current dataset with the first line of the new dataset, ignoring subject ID variables complete. This is NOT recommended!
2) Select "Both files provide cases" and move the ID variable to the "Key Variable" box. In this case the new dataset will have all the subjects from both datasets.

3) Select “Non-active dataset is keyed table” and move the ID variable to the “Key Variable” box. In this case SPSS will bring smoking status information for all the subjects in the study dataset that are also in the Smoking dataset. Subjects not in the smoking dataset will be retained, but with missing smoking status.
4) Select “Active dataset is keyed table” and move the ID variable to the “Key Variable” box. In this case SPSS will retain all the subjects in the smoking dataset and drop subjects that are only in the study dataset.

We want to bring smoking information to the study subjects therefore we need to select option 3) above and press OK. This will add the smoking variable to the end of the file, for the subjects for whom information is available in the smoking dataset.

Tip – In Microsoft Excel this type of merge is usually accomplished using the function “VLOOKUP”. If you feel more comfortable with Excel you may try it there first. Notice that in any case you always need to check the results of any merging procedure to ensure it was done properly.

Exercise

Import dataset “Exercise 1.xls” to SPSS. Discuss the problems you see in the dataset and how you would handle them.

Part II – Data Exploration and Modification

Data Modification

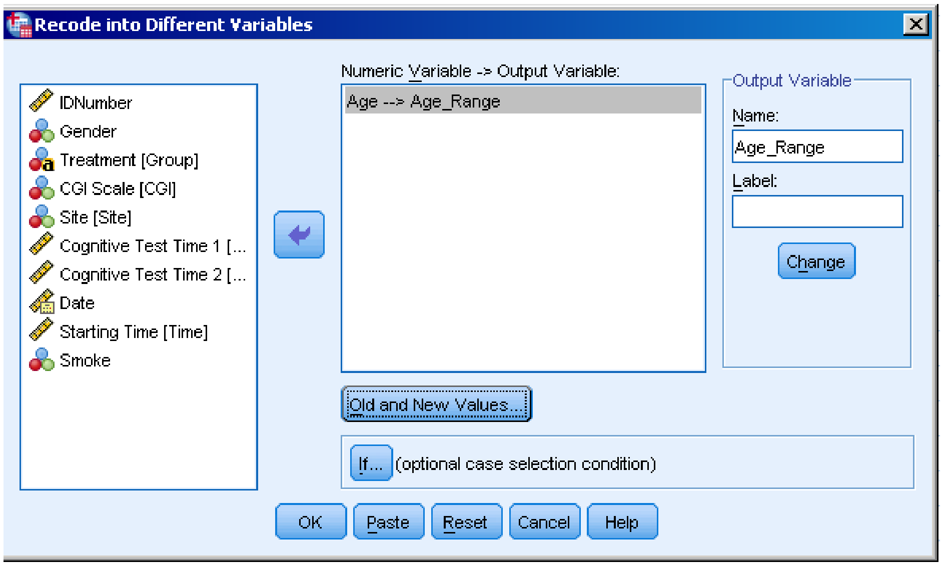
SPSS allows you to recode and transform variables and to restructure your data. Even very complex data tasks can be done in SPSS, although some knowledge of SPSS syntax may be needed. In this section we will cover some of these useful data manipulation tasks.

- ### Recode

Recoding involves changing variables in a way that is not a mathematical transformation. It could be a simple change of codes (change code of Ontario from 9 to 1), creating groups (group BC, AB and SK into West region) and creating ranges (group Age into Age range).

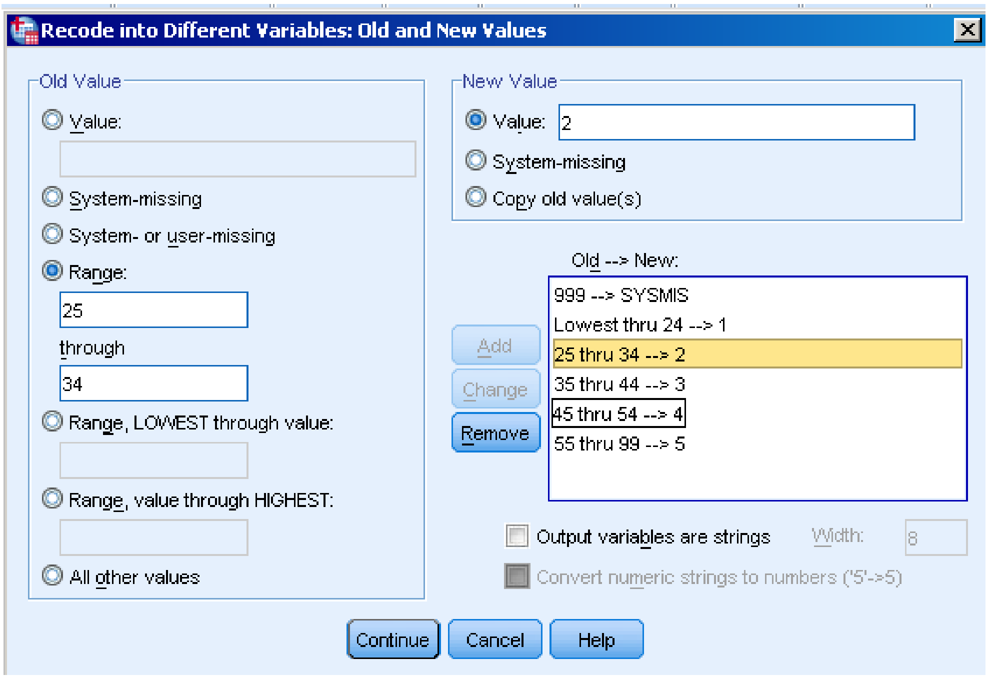
Let's recode the variable Age into age ranges to see how it works. SPSS has two recode options – in the same variable and in different variables. Recoding in the same variable will replace age by age in ranges, and you will lose the original age forever. Thus we always recommend recoding into new variables.

> Go to Transform → Recode into Different Variables. You will see the box below. Move Age to the box in the centre and give a name and label to the output variable. In the example the Age variable will be recoded into the Age_Range variable. Click Old and New Values to specify the recodification rules. You will see the second box, where you should define the Old Values (in Age) and what they will be recoded into (values that will appear in Age_Range). Follow the codification in the second box. Once you are done, hit "Continue" and "OK". You can also click "Paste" if you want to see the code for this, but in this case SPSS will not run the recode, just paste the syntax.



Numeric Variable -> Output Variable:
Age --> Age_Range

Output Variable
Name: Age_Range
Label:



Recode into Different Variables: Old and New Values

Old Value

○ Value:

○ System-missing

○ System- or user-missing

● Range:

25

through

34

○ Range, LOWEST through value:

○ Range, value through HIGHEST:

○ All other values

New Value

● Value: 2

○ System-missing

○ Copy old value(s)

Old --> New:

999 --> SYSMIS
Lowest thru 24 --> 1
25 thru 34 --> 2
35 thru 44 --> 3
45 thru 54 --> 4
55 thru 99 --> 5

□ Output variables are strings Width: 8

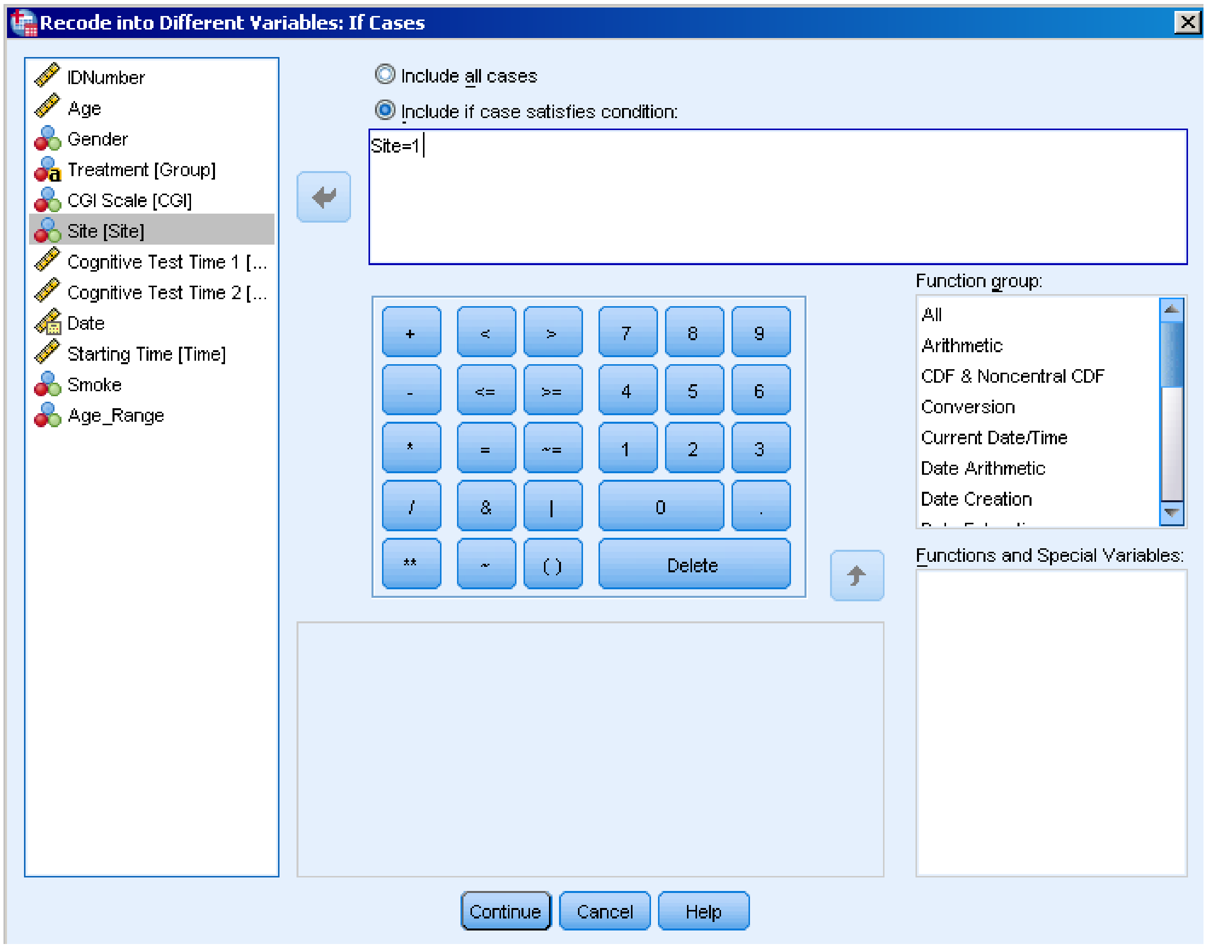
□ Convert numeric strings to numbers ('5'->5)

Notice that there are different ways of defining values to be recoded and these are on the left panel of the box above. You can define ranges like we did, an unbounded range (e.g., 50 years or more), a single value, all values not previously specified, and missing values.

- ## Recode with Conditional IF

Imagine that we need to recode Age into Age Ranges, but the ranges depend on the site (e.g., we want 18 to 24 if site = Toronto and 18 to 29 if site = Montreal. Or we may want to recode Age to Age_range, but only for Toronto and leave Montreal missing in the new age range variable). That is, we still want to create a new variable through recoding an old one, but the rules depend on a third variable.

In that case we can use conditional expressions by clicking on "If" button, in the first recoding window. A new window will appear that allows you to define the condition under which the recoding is to be applied. Assuming that we want the recoding to be done only for Toronto, we have.



Recode into Different Variables: If Cases

○ Include all cases
● Include if case satisfies condition:

Site=1

By entering `site = 1` we force the recode from Age to Age_Range to take place only for the cases recorded in Toronto. If we want a different recoding for Montreal we can repeat the process entering the recoding rules and `site = 2` in the "If" condition. If we do not repeat the process for Montreal, age range will be missing for Montreal.

Notice that these are very simple conditional operations. SPSS allows the user to define many other operations using functions and logical operators and some can be very complex. Some examples of expressions that could be used in the conditional operations include:

```
Site = 1 and Gender = 0 and group = “Control”
(Site = 1 and Gender = 1) or site = 2
Abs(test1) > 2
Sum(test1, test2) > 5
Not missing(CGI) and not missing(test1)
Substr(group,1,4) = “T”
Age > 40 and age < 50
```

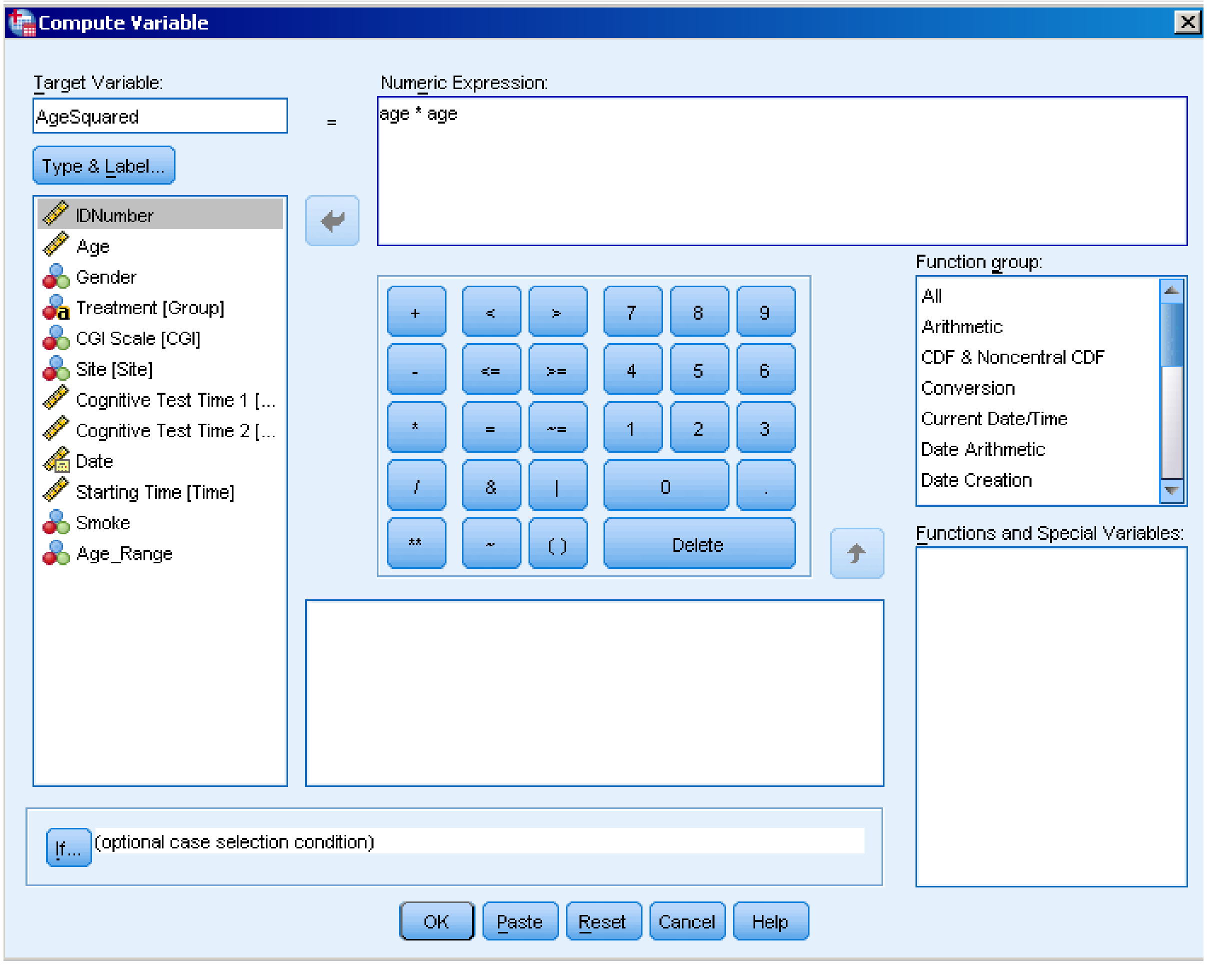
This is a very small set of examples and does not cover all the possibilities. SPSS provides a large set of functions which together with the logical and arithmetic operators can make the conditional operations very flexible.

Tip – Try to paste the recode procedure and run it from the syntax editor. This way you will have recorded the variable modifications that have been made. It will also help you to learn SPSS syntax. The syntax language for recoding is also very simple and once understood it is a quick and flexible alternative to the point-and-click menu.

• Compute

The compute command is used in situations where we need to create a new variable that is some function of one or more existing variables. For example, we may want to create a variable that is the square of Age to use in a regression model. Or maybe we would like to have a new outcome variable that is the average of the cognitive test at times 1 and 2.

Go to Transform → Compute. Let's create AgeSquared variable which is the square of Age. First, insert the name of the new variable "AgeSquared" in the Target Variable box and give it a label if you want. In the "Numeric expression" box enter the formula for Age Squared, which is Age * Age or Age ** 2, where ** stands for the exponentiation operation. Then click OK or Paste.



Compute Variable

Target Variable: AgeSquared

Numeric Expression: age * age

You can play around with formulas here. To get the average cognitive test you just enter "sum(test, test2)/2" or "(test1 + test2)/2" or "mean(test1, test2). If there were 10 time points instead of 2 you could use the keyword "to" like in "mean(test1 to test10)". This implies including in the mean calculation all variables in the data file that are in between test1 and test10.

The “If” button in the Compute window has exactly the same function as in the Recode command: it is used in case you want to transform the variable with a compute, but only for part of the data defined by a third variable.

Notice that Compute will not do calculation with values defined as missing. For example, Age = 999 will not be squared in the calculation above.

Tip – Once we get familiar with formulas, the syntax language for recode and compute is quite simple, for example:

```
Compute agesquared = age ** 2.
If group = “Control” treat = 1.
If group = “Treatment 1” treat = 2.
If group = “Treatment 2” treat = 3.
Recode group(“Control” = 1)(“Treatment 1”=2) (“Treatment 2
= 3) into treat.
Compute mean_outcome = mean(test1,test2).
Execute.
```

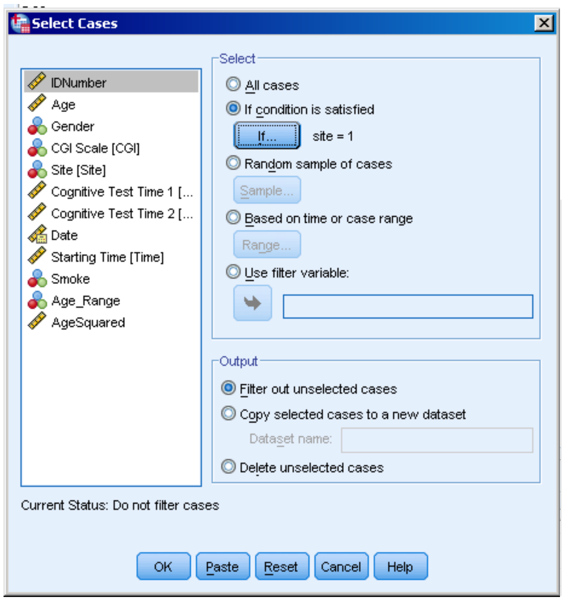
• Select Cases

Sometimes we need to run our analyses using only a subset of our data. For example, we may want to take a frequency of Group, but only among subjects in Toronto. Or we may wish to restrict our analysis to females above 45 years of age with a CGI higher than 3. These are examples were you need to select some cases according to a rule or series of rules prior to running the desired analysis. This can be accomplished in two ways:

- **Delete cases that are unwanted** so that they no longer are in the data. You will not be able to recover back the deleted cases.
- **Filter out, but not delete, the unwanted cases**. You will be able to easily bring back the filtered out cases.

You may wish to delete unwanted cases and keep only the wanted ones in the data file if you intend to perform many analyses with this subset of cases. In this case you should remember to save your original data before deleting cases. Once you delete cases, there is no undo. You may want to just filter out but not delete the unwanted cases if the goal is to do some quick exploratory analysis with your subset of cases.

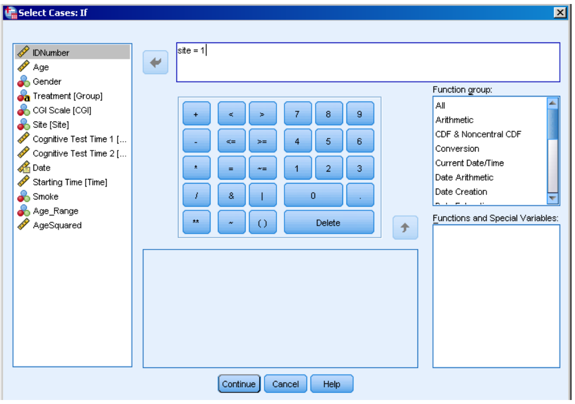
> Go to Analysis → Descriptive Statistics → Frequencies and take a Frequency of CGI. Then go to Data → Select Cases and click the "If" button. A new window will appear allowing you to define the selection rule. Enter "site = 1" and click "Continue". Back to the first window, make sure "Filter out unselected cases" in the "Output" box, is selected and click "OK". This will filter out, but not delete, unwanted cases. Now repeat the frequency of CGI and compare. To turn the Select Cases Off, check "All Cases" in the first window.


Select Cases

Select
○ All cases
● If condition is satisfied
If... site = 1
○ Random sample of cases
○ Based on time or case range
○ Use filter variable:

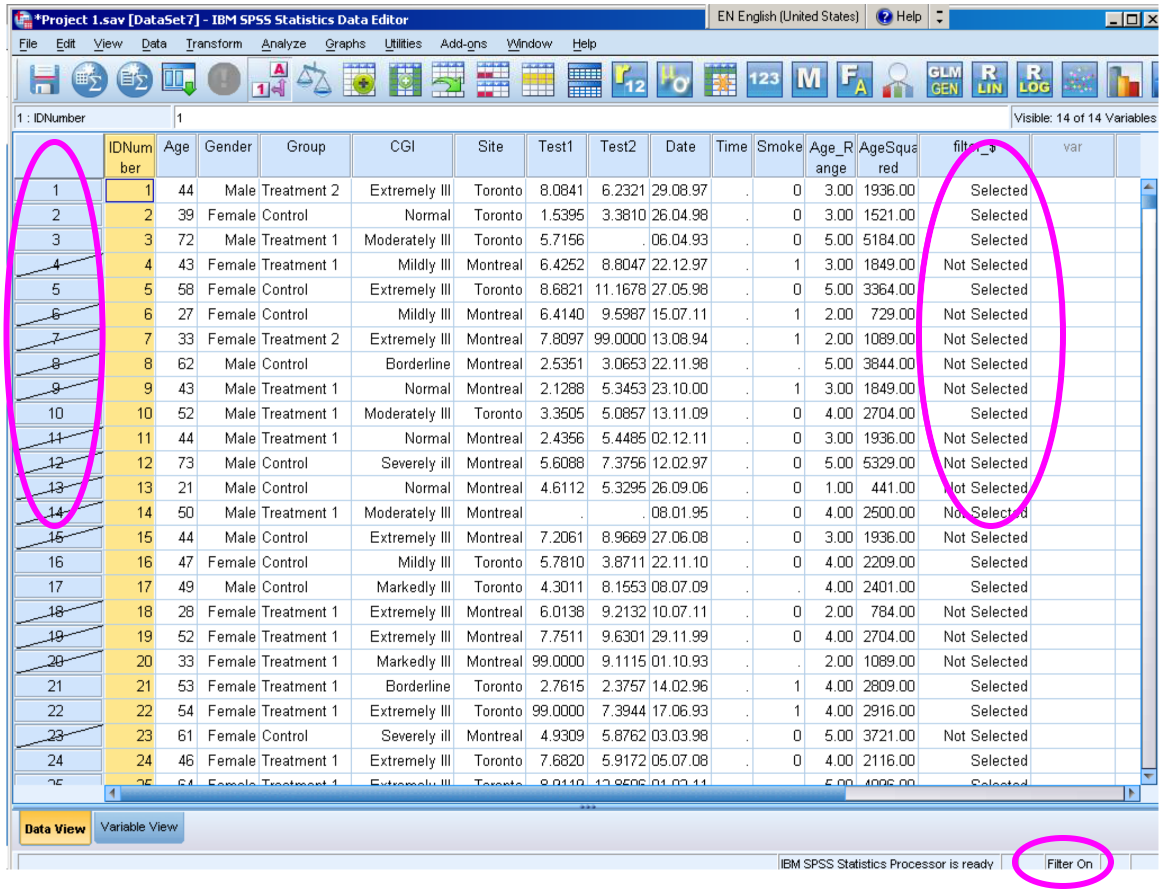
Output
● Filter out unselected cases
○ Copy selected cases to a new dataset
○ Delete unselected cases

Current Status: Do not filter cases


Select Cases: If

site = 1

This approach will filter out the unwanted cases through a flag variable, but will not delete or change the data in any way. In the Data View tab you will be able to see the new variable created by SPSS to flag the selected cases, but all the cases are still in the data file. Notice also that the case counter numbers on the leftmost column crossed for the filtered off cases. In the information bar at the bottom you will see “Filter On” reminding you that the filter is on and that any analysis will be done only among the selected subgroup of cases.



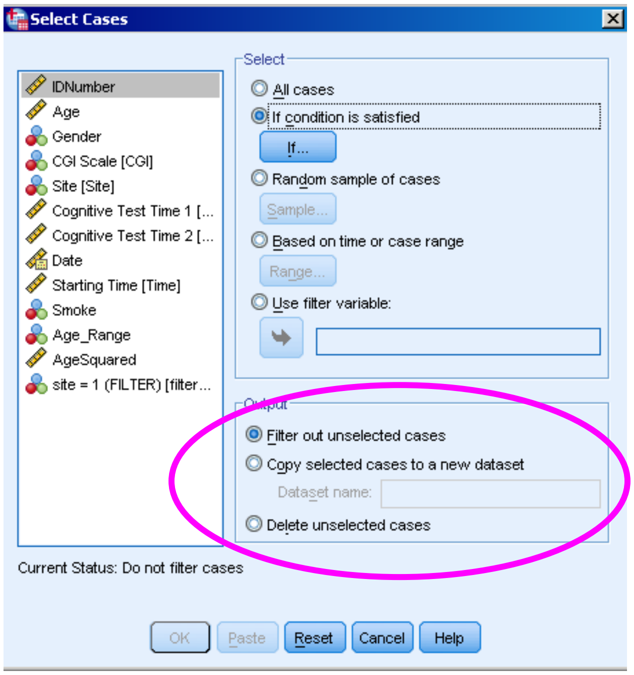
1 : IDNumber 1

Visible: 14 of 14 Variables

	IDNumber	Age	Gender	Group	CGI	Site	Test1	Test2	Date	Time	Smoke	Age_Range	AgeSquared	filter_$	var
1	1	44	Male	Treatment 2	Extremely Ill	Toronto	8.0841	6.2321	29.08.97	.	0	3.00	1936.00	Selected	
2	2	39	Female	Control	Normal	Toronto	1.5395	3.3810	26.04.98	.	0	3.00	1521.00	Selected	
3	3	72	Male	Treatment 1	Moderately Ill	Toronto	5.7156	.	06.04.93	.	0	5.00	5184.00	Selected	
~~4~~	4	43	Female	Treatment 1	Mildly Ill	Montreal	6.4252	8.8047	22.12.97	.	1	3.00	1849.00	Not Selected	
5	5	58	Female	Control	Extremely Ill	Toronto	8.6821	11.1678	27.05.98	.	0	5.00	3364.00	Selected	
~~6~~	6	27	Female	Control	Mildly Ill	Montreal	6.4140	9.5987	15.07.11	.	1	2.00	729.00	Not Selected	
~~7~~	7	33	Female	Treatment 2	Extremely Ill	Montreal	7.8097	99.0000	13.08.94	.	1	2.00	1089.00	Not Selected	
~~8~~	8	62	Male	Control	Borderline	Montreal	2.5351	3.0653	22.11.98	.	.	5.00	3844.00	Not Selected	
~~9~~	9	43	Male	Treatment 1	Normal	Montreal	2.1288	5.3453	23.10.00	.	1	3.00	1849.00	Not Selected	
10	10	52	Male	Treatment 1	Moderately Ill	Toronto	3.3505	5.0857	13.11.09	.	0	4.00	2704.00	Selected	
~~11~~	11	44	Male	Treatment 1	Normal	Montreal	2.4356	5.4485	02.12.11	.	0	3.00	1936.00	Not Selected	
~~12~~	12	73	Male	Control	Severely ill	Montreal	5.6088	7.3756	12.02.97	.	0	5.00	5329.00	Not Selected	
~~13~~	13	21	Male	Control	Normal	Montreal	4.6112	5.3295	26.09.06	.	0	1.00	441.00	Not Selected	
~~14~~	14	50	Male	Treatment 1	Moderately Ill	Montreal	.	.	08.01.95	.	0	4.00	2500.00	Not Selected	
~~15~~	15	44	Male	Control	Extremely Ill	Montreal	7.2061	8.9669	27.06.08	.	0	3.00	1936.00	Not Selected	
16	16	47	Female	Control	Mildly Ill	Toronto	5.7810	3.8711	22.11.10	.	0	4.00	2209.00	Selected	
17	17	49	Male	Control	Markedly Ill	Toronto	4.3011	8.1553	08.07.09	.	.	4.00	2401.00	Selected	
~~18~~	18	28	Female	Treatment 1	Extremely Ill	Montreal	6.0138	9.2132	10.07.11	.	0	2.00	784.00	Not Selected	
~~19~~	19	52	Female	Treatment 1	Extremely Ill	Montreal	7.7511	9.6301	29.11.99	.	0	4.00	2704.00	Not Selected	
~~20~~	20	33	Female	Treatment 1	Markedly Ill	Montreal	99.0000	9.1115	01.10.93	.	.	2.00	1089.00	Not Selected	
21	21	53	Female	Treatment 1	Borderline	Toronto	2.7615	2.3757	14.02.96	.	1	4.00	2809.00	Selected	
22	22	54	Female	Treatment 1	Extremely Ill	Toronto	99.0000	7.3944	17.06.93	.	1	4.00	2916.00	Selected	
~~23~~	23	61	Female	Control	Severely ill	Montreal	4.9309	5.8762	03.03.98	.	0	5.00	3721.00	Not Selected	
24	24	46	Female	Treatment 1	Extremely Ill	Toronto	7.6820	5.9172	05.07.08	.	0	4.00	2116.00	Selected	

IBM SPSS Statistics Processor is ready | Filter On

If you go back to Data → Select Cases, you can see the other options available in the “Output”. You can check “Delete Unselected Cases” This will delete records from Montreal and you will not be able to recover them. A third option is to “Copy selected cases to a new dataset” (available at SPSS 15 and later), which will copy Toronto cases to a new data and you will be left with two data sets: the original, with both Toronto and Montreal and the new one with only cases from Toronto. Usually the first option, filtering out unwanted cases, is enough and simpler and applicable to most cases,



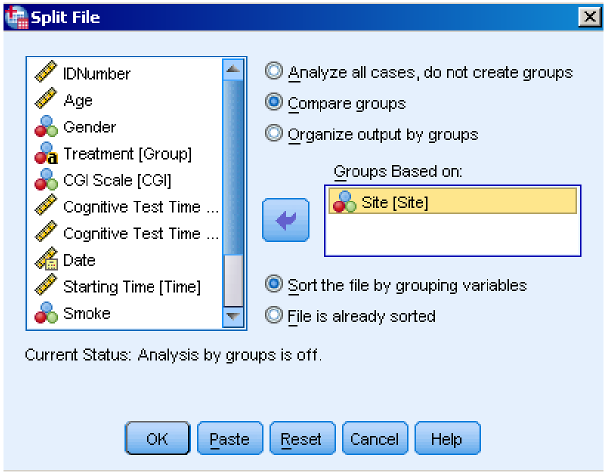
Current Status: Do not filter cases

As you can also see in the SPSS Select Cases window, it is also possible to select cases for analysis based on a random sample of cases, a range of cases (for example, from case 1 to 30) or an already existing variable (values higher than zero in the variable are selected, zero and missing are filtered out). These options are less commonly used.

• Split File

Like Select Cases, Split File allows you to run an analysis on subgroups of cases. We use Split File when we want to run the same analysis on a series of distinct subgroups of cases, defined by some variable in our dataset. For example, we may want to run a regression model for each site, or for each CGI level or for each gender. When we use Split File, the idea is that we separated the data in subgroups defined by the variable used. Every analysis will than run in each and every subgroup (i.e. a stratified analysis).

Let's Split our data by Site as if we wanted all analyses separated done in Toronto and Montreal. Go to Data → Split File. Check "Compare Groups". Move the variable "Site" to the box "Groups Based on:". Split File requires that the data is sorted by the splitting variable. If your data is not sorted or if you don't know, check the item "Sort the file by grouping variables". Now go to Analysis → Descriptive Statistics → Frequencies and take a Frequency of CGI.



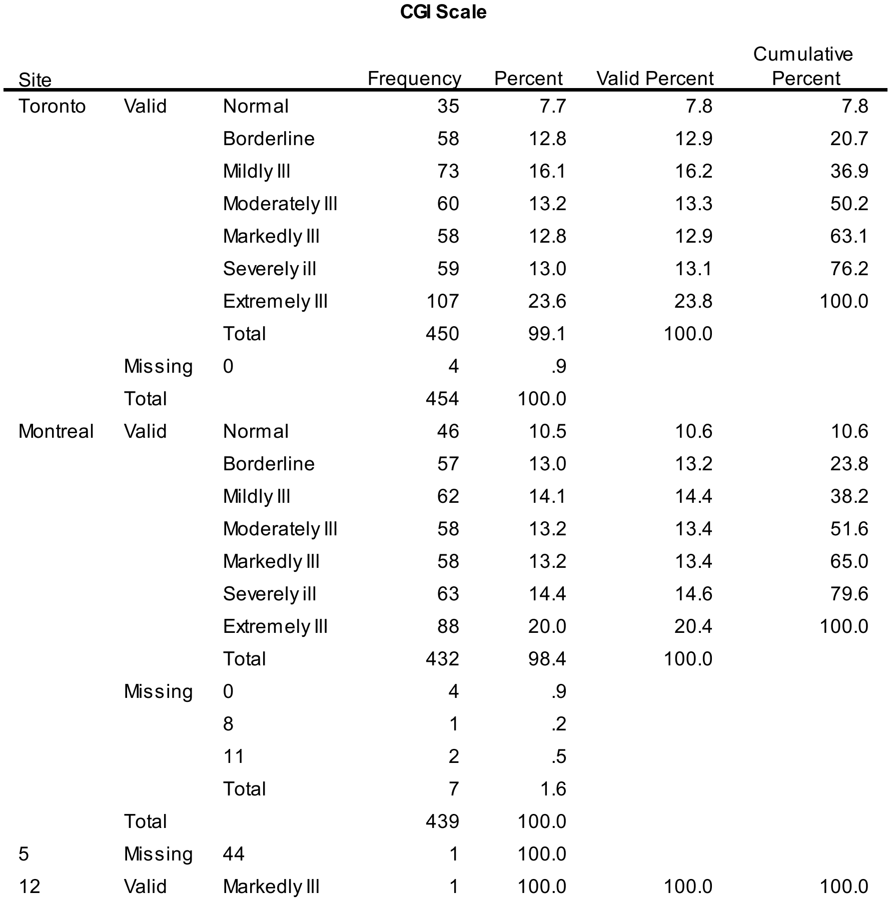
○ Analyze all cases, do not create groups
● Compare groups
○ Organize output by groups

Groups Based on:
Site [Site]

● Sort the file by grouping variables
○ File is already sorted

Current Status: Analysis by groups is off.

Notice that the frequencies are separated created for Montreal, Toronto and the missing values, but they are all in the same output table. If instead of "Compare Groups" you had selected "Organize output by groups" you would get the same result in three separate tables.



CGI Scale

Site			Frequency	Percent	Valid Percent	Cumulative Percent
Toronto	Valid	Normal	35	7.7	7.8	7.8
		Borderline	58	12.8	12.9	20.7
		Mildly Ill	73	16.1	16.2	36.9
		Moderately Ill	60	13.2	13.3	50.2
		Markedly Ill	58	12.8	12.9	63.1
		Severely ill	59	13.0	13.1	76.2
		Extremely Ill	107	23.6	23.8	100.0
		Total	450	99.1	100.0	
	Missing	0	4	.9		
	Total		454	100.0		
Montreal	Valid	Normal	46	10.5	10.6	10.6
		Borderline	57	13.0	13.2	23.8
		Mildly Ill	62	14.1	14.4	38.2
		Moderately Ill	58	13.2	13.4	51.6
		Markedly Ill	58	13.2	13.4	65.0
		Severely ill	63	14.4	14.6	79.6
		Extremely Ill	88	20.0	20.4	100.0
		Total	432	98.4	100.0	
	Missing	0	4	.9		
		8	1	.2		
		11	2	.5		
		Total	7	1.6		
	Total		439	100.0		
5	Missing	44	1	100.0		
12	Valid	Markedly Ill	1	100.0	100.0	100.0

To turn off the Split File you need to select “Analyze all cases, do not create groups” in the Split File window. You may do it now to prepare SPSS for the next section.

Exploring the Data

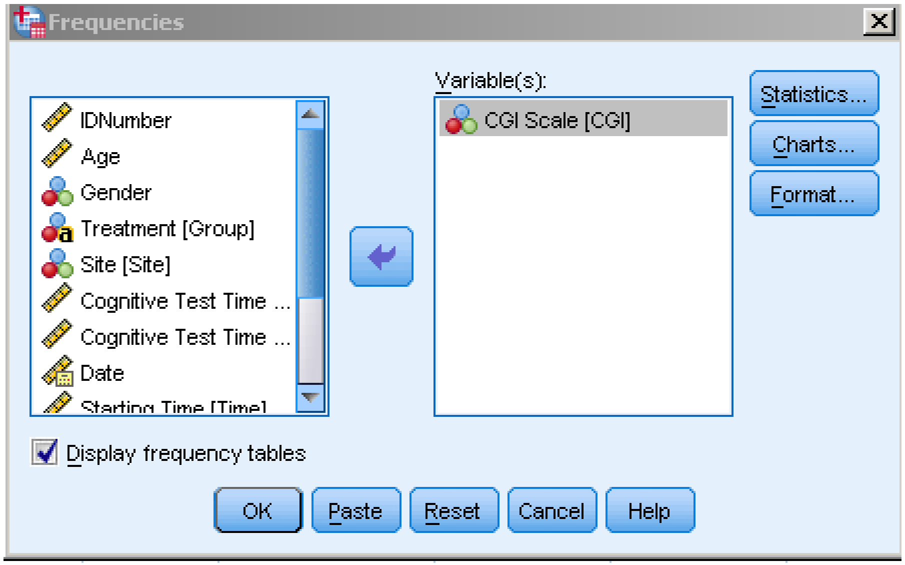
Exploratory Data Analysis (EDA) is the first step we should perform in any data analysis. It allows us to take an initial look at the data to both check and understand it. But most importantly, EDA helps us to guide our decisions relating to choice of appropriate statistical model and possible variable modifications.

Frequencies and **Descriptives** are two procedures we can use to take an initial look at our data. We have seen these two already and we will review them quickly now. Then we will also look into graphs and tables.

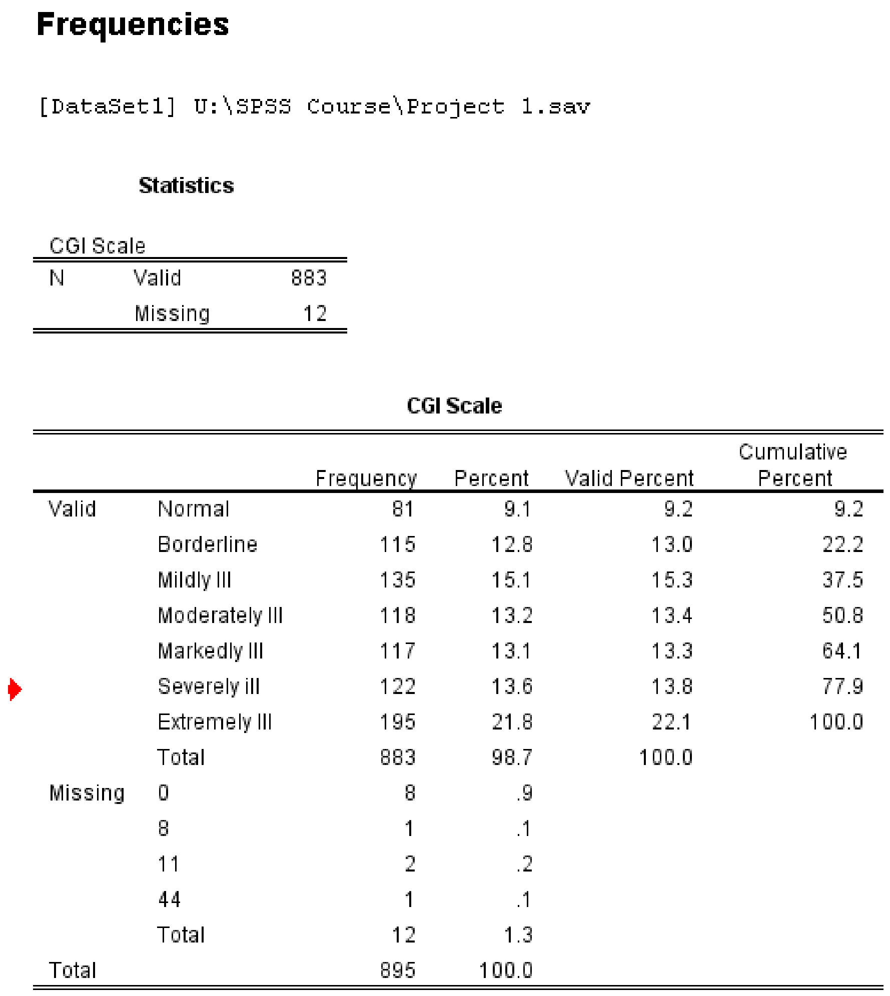
• Frequencies

Frequency tables look at the content of the variables, and are is usually used for listing their values and counts. This procedure is most useful for discrete variables that do not take on too many different values. When dealing with variables like age and cognitive test scores which are continuous, we are unlikely to want a list of all observed values (which will be a very long list), and generally prefer to summarize these variables by their mean, median and variance. This can also be accomplished by SPSS Frequencies.

Go to Analysis → Descriptive Statistics → Frequencies. Let's use the variable CGI_Text. Click Ok.



☑ Display frequency tables



Frequencies

```
[DataSet1] U:\SPSS Course\Project 1.sav
```

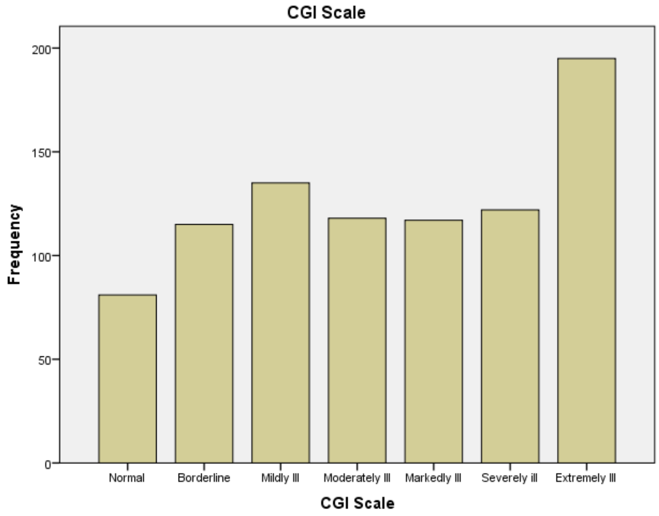
Statistics

CGI Scale		
N	Valid	883
	Missing	12

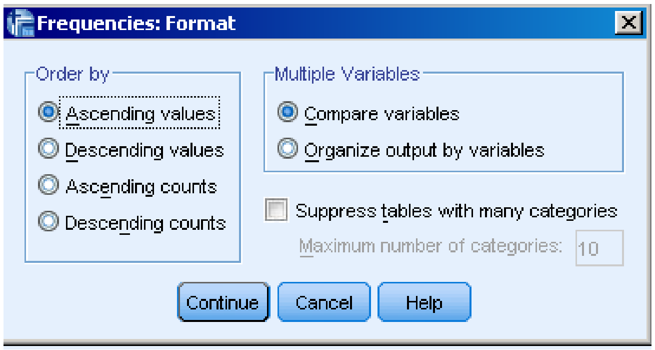
CGI Scale

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Normal	81	9.1	9.2	9.2
	Borderline	115	12.8	13.0	22.2
	Mildly Ill	135	15.1	15.3	37.5
	Moderately Ill	118	13.2	13.4	50.8
	Markedly Ill	117	13.1	13.3	64.1
	Severely ill	122	13.6	13.8	77.9
	Extremely Ill	195	21.8	22.1	100.0
	Total	883	98.7	100.0	
Missing	0	8	.9		
	8	1	.1		
	11	2	.2		
	44	1	.1		
	Total	12	1.3		
Total		895	100.0		

This is a frequency table; it shows us all the values of the variable and the number of times each one appears in our dataset. This is known as the distribution of CGI. Notice that the values defined as missing are also counted in the table, which makes the procedure very good for checking the data. By clicking “Charts” in the Frequencies dialogue you can also request a bar or pie chart (although a histogram is also available, it is not appropriate for highly discrete variables like CGI). Notice that the bar chart does not include the values defined as missing.



The “Format” button gives us a few options for organizing the frequency table. The most useful option is the ability to change how we sort the table. In large tables we may want to see the most common values at the top and we can check “Descending Counts” for this option. Notice that we can create many tables at once. In datasets with many variables it is usual to ask for a frequency of all variables and exclude tables with more than a certain number of categories.



Frequencies: Format

Order by
- ● Ascending values
- ○ Descending values
- ○ Ascending counts
- ○ Descending counts

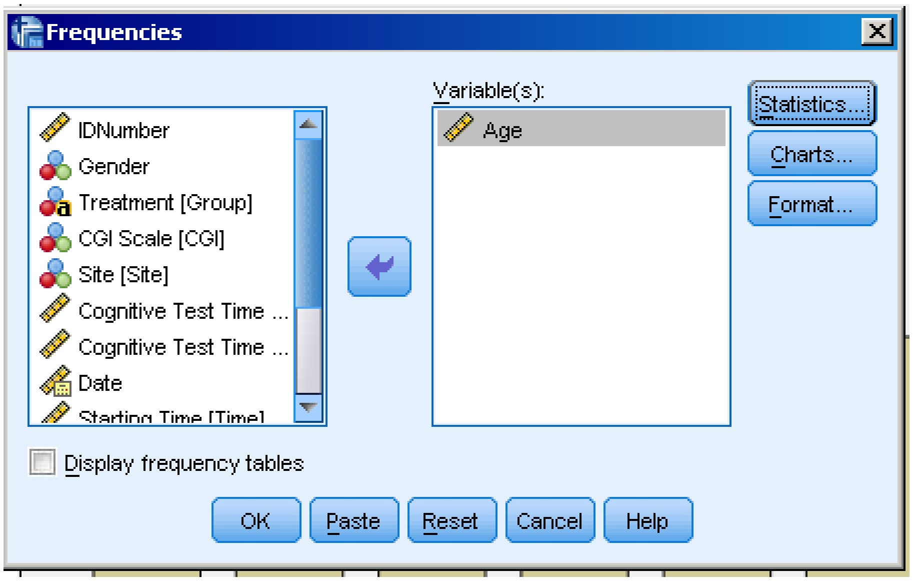
Multiple Variables
- ● Compare variables
- ○ Organize output by variables

□ Suppress tables with many categories

Maximum number of categories: 10

If we want to look at continuous variables, we can still use SPSS Frequencies menu, but in this case table of frequencies don’t make much sense because a continuous variable will result in a very long table with too many values that is difficult to interpret. Instead, we should look at summary statistics like the mean, median, standard deviation and percentiles.

> Let’s look at Age. Move the variable Age to the “Variable” box and **uncheck** the “Display frequency tables” checkbox. Click on the “Statistics” button.



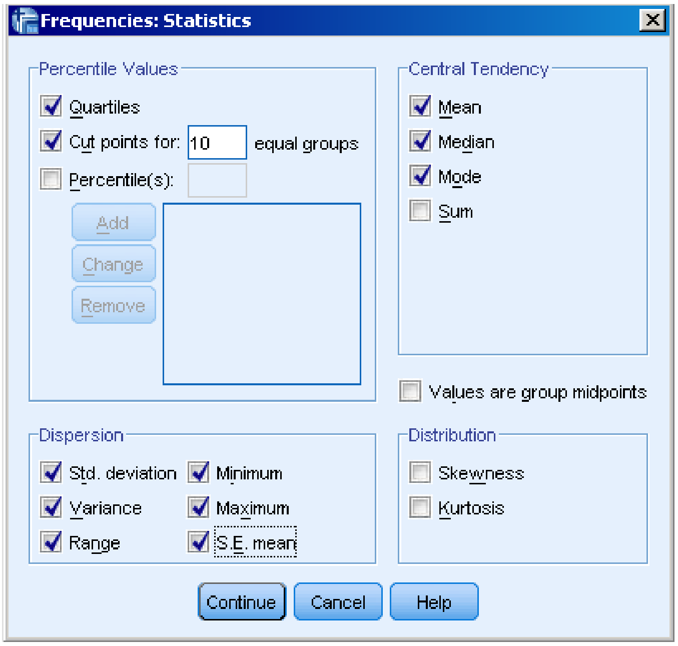
Frequencies

IDNumber
Gender
Treatment [Group]
CGI Scale [CGI]
Site [Site]
Cognitive Test Time ...
Cognitive Test Time ...
Date
Starting Time [Time]

Variable(s):
Age

□ Display frequency tables

You will see a new window that has many statistics available. Let’s check the most important ones and we will describe them below. Click “Continue” and “OK”.



Frequencies: Statistics

Percentile Values
☑ Quartiles
☑ Cut points for: 10 equal groups
□ Percentile(s):

Central Tendency
☑ Mean
☑ Median
☑ Mode
□ Sum

□ Values are group midpoints

Dispersion
☑ Std. deviation
☑ Minimum
☑ Variance
☑ Maximum
☑ Range
☑ S.E. mean

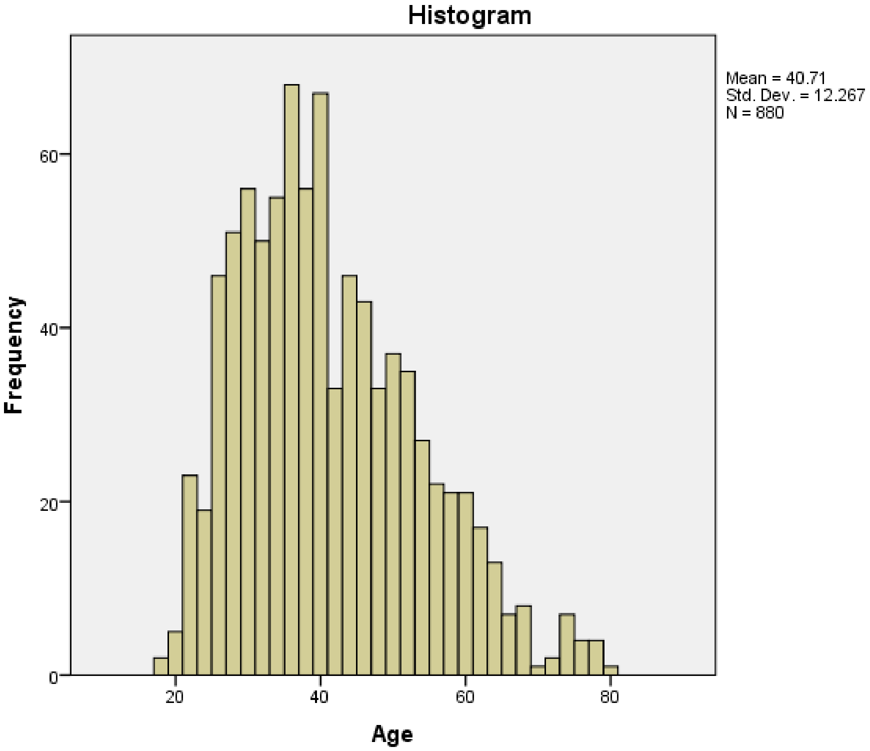
Distribution
□ Skewness
□ Kurtosis

Statistics

Age		
N	Valid	880
	Missing	15
Mean		40.71
Std. Error of Mean		.414
Median		39.00
Mode		39
Std. Deviation		12.267
Variance		150.469
Range		61
Minimum		18
Maximum		79
Sum		35823
Percentiles	10	26.00
	20	30.00
	25	31.00
	30	33.00
	40	36.00
	50	39.00
	60	42.00
	70	46.00
	75	49.00
	80	51.00
	90	58.00

- **N** – The sample size, both valid and missing.
- **Missing** – Number of missing cases.
- **Mean** – The mean of the age for this sample.
- **Std. error of the mean** – A measure of variability for the mean. Mean +/- 1.96*Std. error gives an approximate confidence interval for the mean.
- **Median** – The value that splits the sample so that 50% of the sample is above it and 50% is below.
- **Mode** – is the most often value, the value appearing highest number of times.
- **Std. Deviation** – Measure of variability of the sample. If the distribution of the values is normal then around 85% of the values will be between the mean +/- 1.96*std deviation.
- **Variance** – Is the square of the standard deviation. It is also the average of the squared difference of each value to the mean.
- **Range** – The maximum minus the minimum.
- **Maximum and Minimum** – The highest and lowest values.
- **Sum** – The sum of all the values.
- **Percentiles** – These are age values that split the sample in equal parts, In this case we requested 10 equal groups, so for example, the 10th Percentile is the value that leaves 10% of the values below it. In this case we can see that around 10% of the sample is equal or below 29 years old. 20% is below 35 years old and so on.

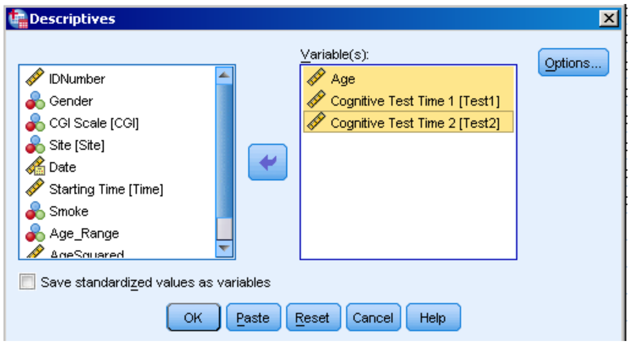
The fact that the mean and median are very close indicates that age is likely symmetric. We can verify this by creating a histogram – run the Frequencies procedure again, but now Click and "Chart" and Histogram. You can additionally ask for a normal curve. Normality is an important attribute of variables and it is an assumption underlying many statistical procedures.



• Descriptives

Typically the SPSS Frequencies procedure is used to assess the distributions of discrete variables. The Descriptives procedure is used for quickly obtaining summary statistics like themean, standard error, maximum and minimum for continuous variables in much the same way that we saw in the second part of the Frequency command above.

Go to Analysis → Descriptive Statistics → Descriptives and move Age and the cognitive tests to the "Variable(s)" box. Click in Options and select the statistics you are interested on, then "Continue".



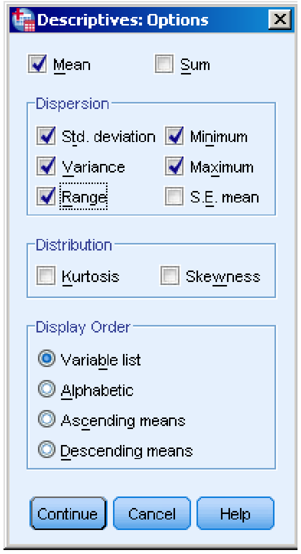
Descriptives

- IDNumber
- Gender
- CGI Scale [CGI]
- Site [Site]
- Date
- Starting Time [Time]
- Smoke
- Age_Range
- AgeSquared

Variable(s):

- Age
- Cognitive Test Time 1 [Test1]
- Cognitive Test Time 2 [Test2]

☐ Save standardized values as variables



Descriptives: Options

☑ Mean ☐ Sum

Dispersion

☑ Std. deviation ☑ Minimum
☑ Variance ☑ Maximum
☑ Range ☐ S.E. mean

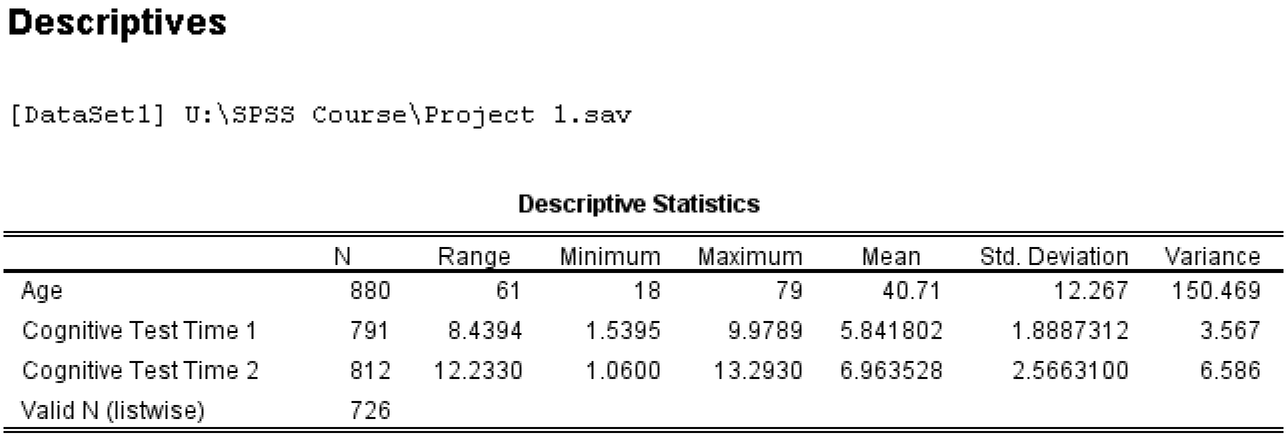
Distribution

☐ Kurtosis ☐ Skewness

Display Order

● Variable list
○ Alphabetic
○ Ascending means
○ Descending means

You will get a simple output table with all the chosen statistics for each of the selected variables. The statistics available here are the same as those available in the Frequencies procedure, but Frequencies provides us with additional options, such as the median and percentiles. The idea is that Descriptives is just a very quick way of obtaining a brief summary of our data.



Descriptives

```
[DataSet1] U:\SPSS Course\Project 1.sav
```

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Age	880	61	18	79	40.71	12.267	150.469
Cognitive Test Time 1	791	8.4394	1.5395	9.9789	5.841802	1.8887312	3.567
Cognitive Test Time 2	812	12.2330	1.0600	13.2930	6.963528	2.5663100	6.586
Valid N (listwise)	726						

Notice that Descriptives also has a box available in the first window labelled “save standardized values as variables”. If you check this box, SPSS will create a standardized version of each of the selected variables. Standardized variables are important in some statistical analyses and they are calculated as the original value minus the average divided by the standard deviation.

Tip – Standard Errors and Standard Deviations are both measures of variability but they capture different information. The Standard Error is a measure of variability of our estimates, (e.g. means and regression coefficients) and indicates how precise the estimate is when estimating the population parameter. The Standard Deviation is a descriptive measure of the current sample and helps us to understand how widely spread the values are in our sample.

• SPSS Graphs

As we have seen, the Frequency procedure offers some univariate graphs and those will often be enough in most cases. But SPSS also has an additional menu where many types of graphs can be requested, and with more options that add flexibility.

In the graph menu, you will usually see the "Legacy" option, which are graphs kept from earlier versions of SPSS and they are simple and quick to use.. Another option is "Interactive", which offers graphs that can be further customized.

Finally, the Chart Builder option is the newest implementation and likely the most flexible. But it is also the most complicated to use and it is one of the few SPSS procedures that require the variable measure (Scale, Nominal, Ordinal) to be defined correctly in the data.

In this course, our aim is to make you aware of the graphical options available in SPSS, but we will not go into detail as there are many distinct types of graphs available and based on our experience, most people do not use SPSS for graphics very often and prefer other software packages. We typically recommend that you use SPSS graphs to explore and visualize your data at first, but use another software package with advanced graphical capabilities, (even Excel is an option), to create charts suitable for reports and publications.

Exercise

Use Cognitive score at time 1 to create in the dataset a new variable with values 1 (= low cognitive score) and 2 (= high cognitive score), where these two groups will be defined as below and separately for two age groups – lower than 30 years old / 30 years old and higher.
Low = Scores lower than the average
High = Scores higher than the average.

Part III – Introduction to Statistics using SPSS

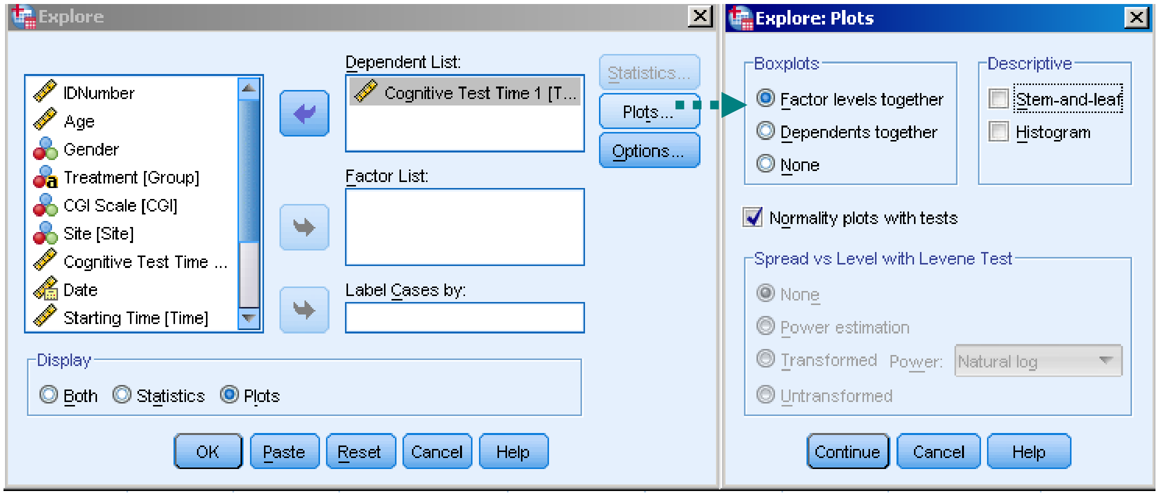
We will now introduce some basic statistical inference procedures that are often useful in analyzing research data.

• Tests of Normality

The normal (or Gaussian) distribution is bell-shaped and has many nice statistical properties. Most statistical models and tests are based on the assumption that the underlying data follows a normal distribution. If this assumption is not correct, these statistical procedures may lead to incorrect conclusions, so it is always important to check this normality assumption. There are several different ways of doing this in SPSS and we will introduce one of the tests.

Note – Statistical inference is a complex topic and we do not have time to cover it in great detail during this workshop. It is important to be aware of some basic definitions and terminology. All research questions and statistical tests can be phrased in terms of a null hypothesis and an alternative hypothesis. The null hypothesis is usually the status-quo (e.g. the treatment has no effect, there is no difference between study groups) and we assume this hypothesis is true unless we have good evidence that the alternative hypothesis is true, in which case we reject the null hypothesis in favor of the alternative hypothesis. In the case of normality tests, the null hypothesis is that the data follows a normal distribution. P-values of 0.05 or less indicate evidence against the null hypotheses (against normality). These findings should be interpreted with caution, a p-value of less than 0.05 does not necessarily mean that standard statistical techniques assuming normality cannot be applied. But if things are not clear for you, do not hesitate to contact us at the Biostatistical Consulting Service and we will be happy to help.

Go to Analyze → Descriptive Statistics → Explore. In “Dependent List” inset the Cognitive Test variable. Click “Plot” and check “Normality Plot with tests”.



Dependent List: Cognitive Test Time 1 [T...

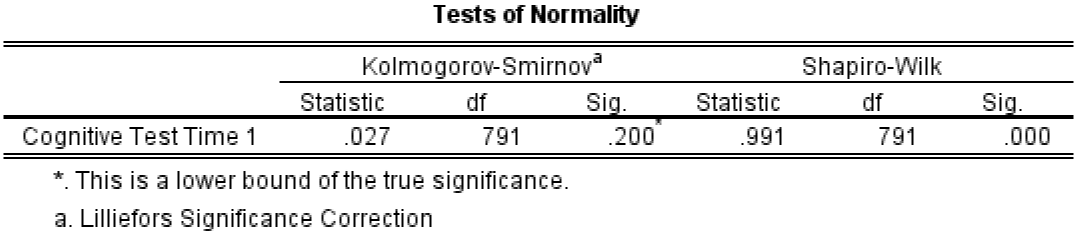
Display: ○ Both ○ Statistics ● Plots

Boxplots: ● Factor levels together ○ Dependents together ○ None

☑ Normality plots with tests

Spread vs Level with Levene Test: ● None ○ Power estimation ○ Transformed Power: Natural log ○ Untransformed

One of the output tables will be the Test of Normality, which includes two traditional tests: the Kolmogorov-Smirnov and Shapiro-Wilk test. They usually will agree with each other and in this example we have significant evidence that the distribution is not Normal. The “Sig.” column is what we call p-value which is usually significant when lower than 0.05. So in this case it is not significant thus we have no strong evidence that the data is not normally distributed.



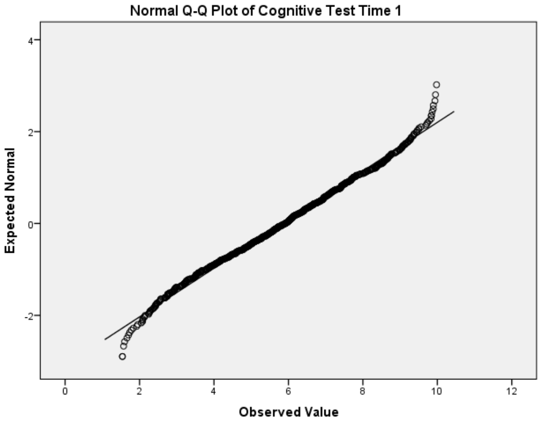
Tests of Normality

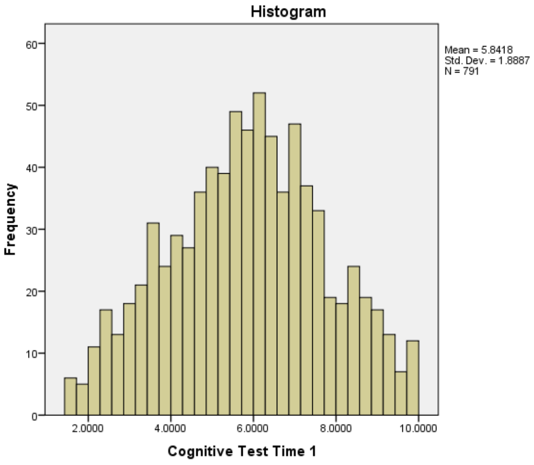
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Cognitive Test Time 1	.027	791	.200*	.991	791	.000

*. This is a lower bound of the true significance.

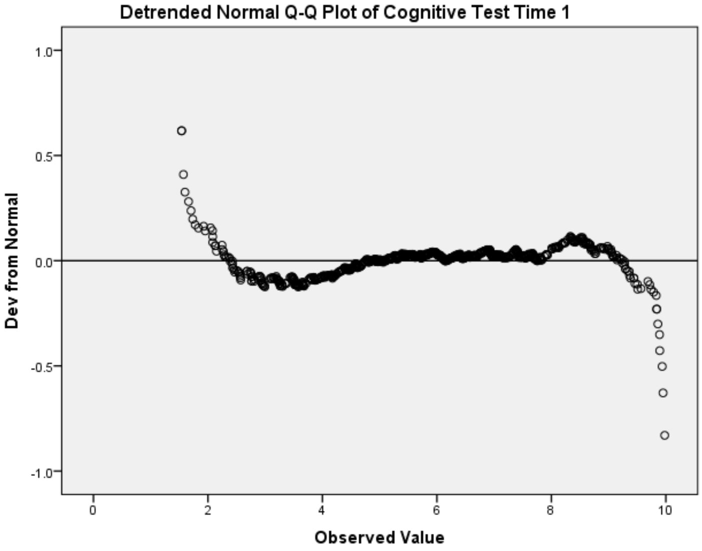
a. Lilliefors Significance Correction

We will also get a normal Q-Q plot, which is a visual tool used to assess the normality of variables. If the observations follow the straight line then the data is normal. In this case we see some departure from the straight line at the ends of the distributions, which in this case indicates that our data has lighter tails than the Normal distribution. Notice that the observed values in the tails would have to be larger in the tails to be on the line. This can also be seen in a histogram.



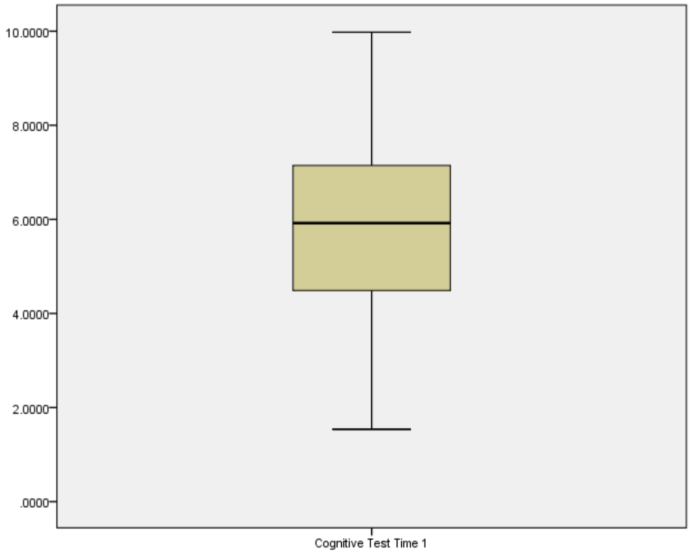


By default SPSS also creates the detrended Q-Q plot, which plots in the vertical axis the distance between the points and the straight line from the Q-Q plot. Ideally we would like the points to be all close to zero. These charts can easily uncover non-normality but departures from normality that are too small are unlikely to be of concern for statistical analyses, so it is important to look at the normality test and other diagnostic tools, not only the graphs that can be too sensitive specially with large sample size.



Finally, a box-plot is also created by the Explore procedure. It offers a nice summary of the distribution of a variable. It shows the median (middle line), first and third quartiles (edges of the box) and minimum and maximum. Box-plots will also show outliers which is defined as points outside of the usual range, usual being defined as larger than 1.5 times the height of the box from the box edge. In this case we don't have outliers because the maximum and minimum are within the edge of the box plus 1.5 times its height. This is a default definition of outliers, one that is useful for a first look at the data, however specific statistical analyses will have their own outlier detection procedures.

Great attention should be given to outliers as they can change a lot the results of entire models, specially the parametric models. The box-plot shows that the data is relatively symmetric as the distance between the median and the 1st and 3rd quartiles are similar as well as the distance between the median and the extreme points.



SPSS graphs can be formatted in many ways. You can double click the graph and formatting menus will appear. These graphs can also be created directly from the "Graph" menu, where you will also find other types of graphs.

Normality can usually be improved through transformation of the variables, and the most commonly used transformations are the log, square root and inverse transformations. You can transform the variables using Transform → Compute, as we have already seen. Transformations may not always help and extra attention is needed in the interpretation of statistical analyses that involve transformed variables.

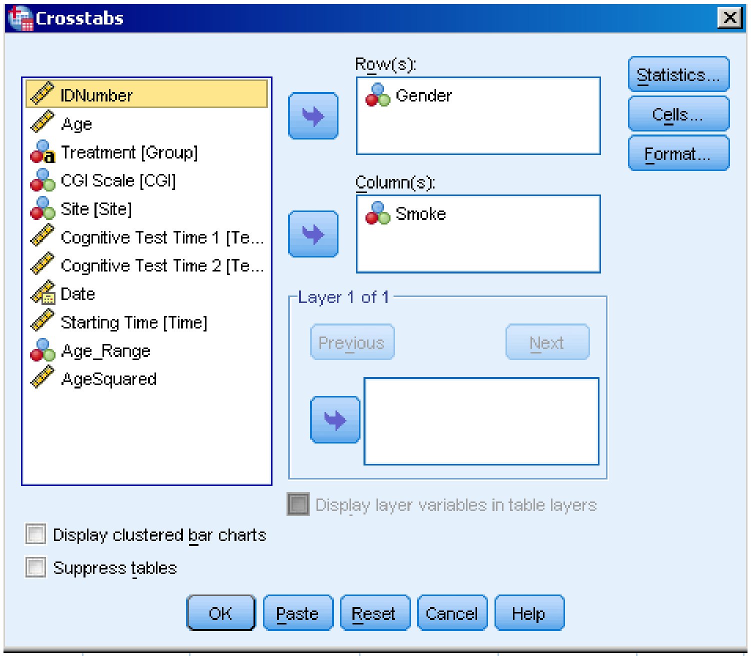
• Analyses Involving Two Categorical Variables

Contingency Tables

Contingency tables or cross-tables (i.e. crosstabs) are very commonly used to summarize the joint distribution of two or more categorical variables. You can think of it as an extension of frequency tables, where we are interested on more than one variable at a time. For example, we may want to know the proportion of the sample that are both male and are from Montreal.

In addition to the joint distribution of two or more variables, contingency tables can provide us with the conditional distributions e.g. the proportion of Males who Smokes (that is, what is the proportion of smokers, conditional on the individual being Male). The SPSS Crosstabs procedure is the quickest way to get these summary tables

> Go to Analysis → Descriptive Statistics → Crosstabs. Move Gender to the row and Smoke to the column. Click "OK". You will get the table below, which tells you the number of subjects in any Smoke X Gender combination.



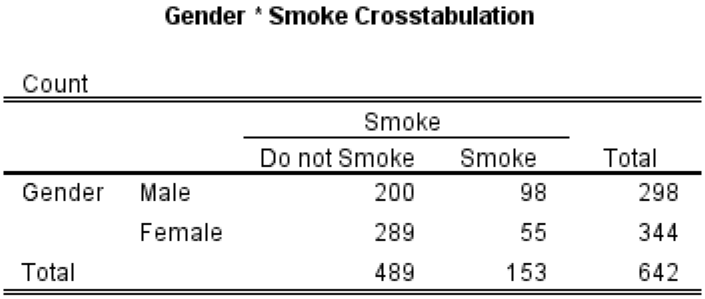
Row(s): Gender

Column(s): Smoke

☐ Display layer variables in table layers

☐ Display clustered bar charts

☐ Suppress tables

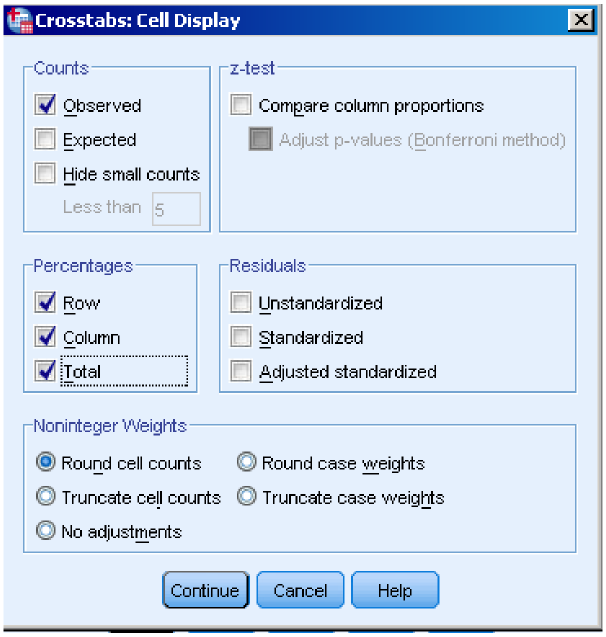


Gender * Smoke Crosstabulation

Count

		Smoke		
		Do not Smoke	Smoke	Total
Gender	Male	200	98	298
	Female	289	55	344
Total		489	153	642

Do it again. Now click the "Cells" button and check the all the "Percentages" options. You will get the much more informative table below.



Crosstabs: Cell Display

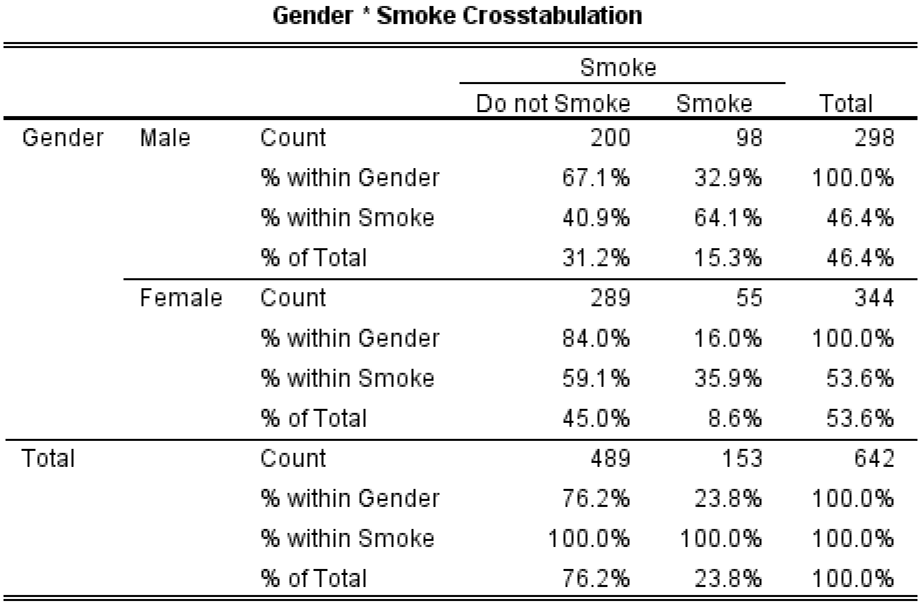
Counts
☑ Observed
□ Expected
□ Hide small counts
Less than 5

z-test
□ Compare column proportions
□ Adjust p-values (Bonferroni method)

Percentages
☑ Row
☑ Column
☑ Total

Residuals
□ Unstandardized
□ Standardized
□ Adjusted standardized

Noninteger Weights
● Round cell counts
○ Round case weights
○ Truncate cell counts
○ Truncate case weights
○ No adjustments



Gender * Smoke Crosstabulation

			Smoke		
			Do not Smoke	Smoke	Total
Gender	Male	Count	200	98	298
		% within Gender	67.1%	32.9%	100.0%
		% within Smoke	40.9%	64.1%	46.4%
		% of Total	31.2%	15.3%	46.4%
	Female	Count	289	55	344
		% within Gender	84.0%	16.0%	100.0%
		% within Smoke	59.1%	35.9%	53.6%
		% of Total	45.0%	8.6%	53.6%
Total		Count	489	153	642
		% within Gender	76.2%	23.8%	100.0%
		% within Smoke	100.0%	100.0%	100.0%
		% of Total	76.2%	23.8%	100.0%

Notice that the percentages enable you to look for associations in the table. For example, we can see that within Male group, 32.9% are smokers but a lower proportion, 16.0%, are smokers among Females. This is the distribution of Smoking Status conditional on Gender. You can also look within Smoking Status for the proportion who are Male and Female, in which case we are conditioning on smoking status.

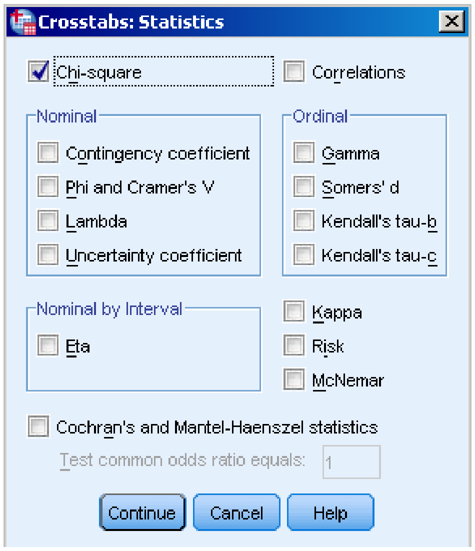
Notice that missing values (blanks and user defined) do not appear in the table. You could have the user defined missing values in the table by not defining them as missing and in order to have the blanks in the table you would need to recode the variable so that the blanks would be recoded in some code.
You can create tables with more than 2 variables by inserting variables in the "Layer" box.

Tests of Independence

In the table above we see a higher proportion of smokers among Males than among Females. This is a numeric association and indicates that the distribution of smoking status seems to depend on gender. However any association in a sample may be due to random fluctuations brought about by the sampling process. If this is the case, then in a different sample we are likely to find no association or a different association. So just by looking at the table it is difficult to say whether this is a real association (i.e., does our target population of Males really Smoke more than Females?) If we replicate the study will we find the same association again? Or it is just sampling fluctuation (i.e., our sample has more smokers among Males, but a different sample is likely to produce a different result)? A statistical test will assess the evidence that the association is real, and is too big to be due to just random fluctuation. In this case the null hypothesis is that there is no association and we accept this null hypothesis unless we have good evidence to suggest that there is an association (i.e. typically defined by a p-value less than 0.05).

One of the most commonly used tests associated with contingency tables is the chi-square test. The chi-square test is relatively robust and only requires that the table does not have cell counts that are too small (i.e. less than 5 subjects in any cell). If the table does have small counts then the likelihood ratio test, which is performed by default with the chi-square test, is a better option. Fisher's exact test is another good choice, available in SPSS for 2 by 2 tables. It is typically recommended that you exclude missing values from the table before performing the test as they often create small counts unrelated to what we want to test. You may also wish to consider re-categorizing sparsely populated variables into a smaller number of categories.

> Go to → Analysis → Descriptive Statistics → Crosstabs and now click in "Statistics". You will see many test options. Different tests are used in different situations and it is important to use an appropriate test. The chi-square test is the most common and is applicable to tables with two categorical variables, like ours. Check the Chi-square box, click "Continue" and "OK".



☑ Chi-square
□ Correlations

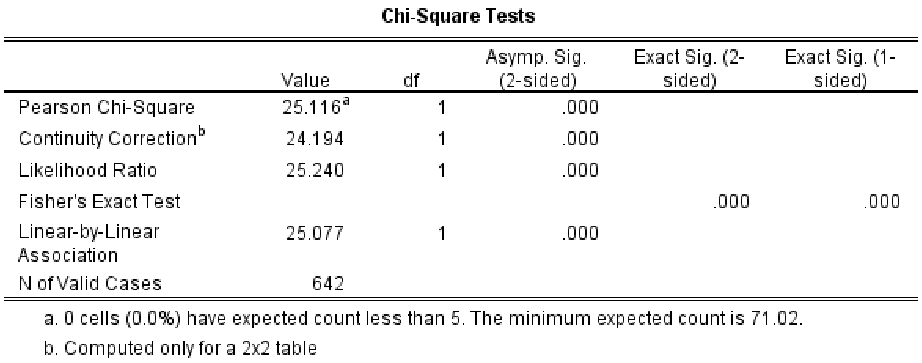
Nominal
□ Contingency coefficient
□ Phi and Cramer's V
□ Lambda
□ Uncertainty coefficient

Ordinal
□ Gamma
□ Somers' d
□ Kendall's tau-b
□ Kendall's tau-c

Nominal by Interval
□ Eta

□ Kappa
□ Risk
□ McNemar

□ Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals: 1



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	25.116[a]	1	.000		
Continuity Correction[b]	24.194	1	.000		
Likelihood Ratio	25.240	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	25.077	1	.000		
N of Valid Cases	642				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 71.02.

b. Computed only for a 2x2 table

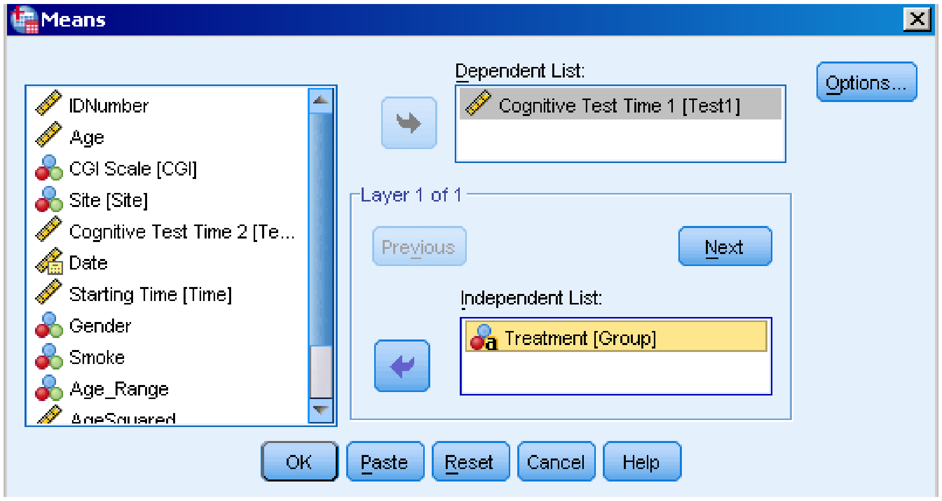
SPSS will output the chi-square test along with the likelihood ratio test and Fisher's exact test. In this case the p-value is 0.000 which is significant, meaning that we have evidence to suggest that the observed association between Gender and Smoking status is more than random fluctuation. You can see that in this case all of the test results are quite similar because our sample is large and all cell counts are of reasonable magnitude, making all of these tests appropriate. When your sample is small Fisher's exact test is generally recommended and expected to perform better than the others as it is not an asymptotic test.

• Analyses Involving One Categorical and One Continuous Variable

Table of Means

Contingency tables look at the relationship between two categorical variables. If one of the variables is continuous and the other is categorical then you can summarize the data by creating a table of means, where for each level of the categorical variable you have the mean of the continuous variable. In SPSS one of the ways we can do that is using SPSS Means.

Go to Analysis → Compare Means → Means. The "Independent List" refers to the categorical variable and the "Dependent List" refers to the continuous variable. Click "OK".



Dependent List: Cognitive Test Time 1 [Test1]

Independent List: Treatment [Group]

Report

Cognitive Test Time 1

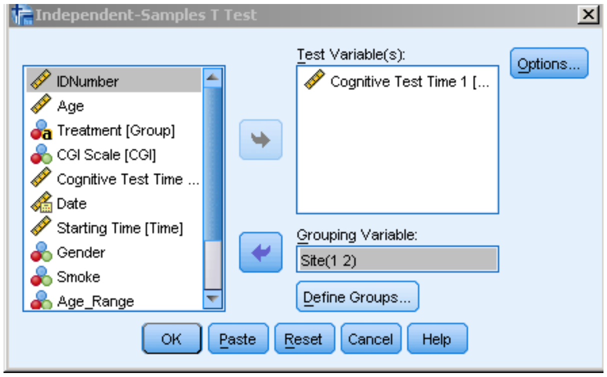
Treatment	Mean	N	Std. Deviation
Control	5.895495	323	1.8903536
Treatment 1	5.725748	326	1.8965845
Treatment 2	5.986103	142	1.8645828
Total	5.841802	791	1.8887312

SPSS Means will output a table with the means of the dependent variable by the independent variables. If you have more than one categorical variable in the "Independent List" you will get more than one table. In this case we are looking at the mean of Cognitive Test by Treatment group. When the means (or medians or even distributions) are different across groups then we say that there is an association between the two variables. Thus it can be of interest to test whether these differences are significant or not. In what follows we will go through some of the most used tests that address this question.

T-tests for Two Independent Samples

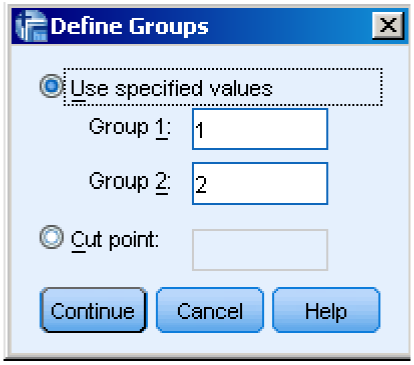
When we have only two groups a simple test for difference of means between the groups is the t-test for independent samples. It assumes that the data follows a normal distribution, the variances are the same in both groups and that the groups are independent. Site is a variable with only two groups, so it would be appropriated to use the t-test to access the significance of the difference on Cognitive Test between sites.

> Go to analysis → Compare Means → Independent-Samples t-test. Move Cognitive test 1 to the "test variable" box because it is the variable we want to test. In the "Grouping variable" enter Site. Now we need to specify which two groups we want to compare. Click in "Define Groups" and enter 1 and 2 as Groups 1 and Group 2. Click "Continue" and "Ok".



Test Variable(s): Cognitive Test Time 1 [...

Grouping Variable: Site(1 2)



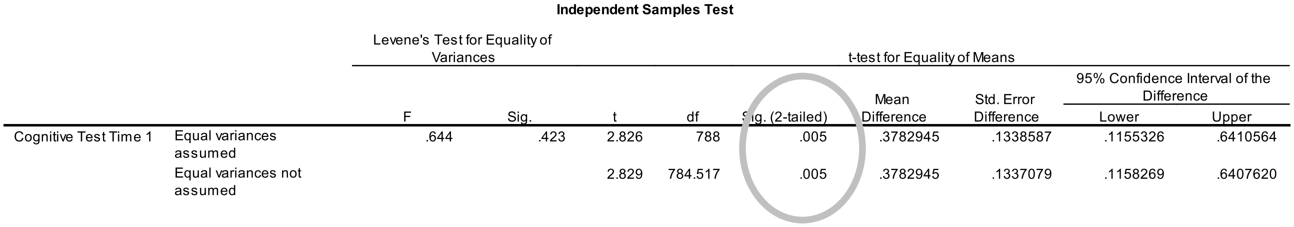
● Use specified values

Group 1: 1

Group 2: 2

○ Cut point:

The output of the t-test is a descriptive table with group means and standard errors (which we are not showing) and the table with the actual test showed below. First we see the Levene's test for equality of variances, then the results for the t-test. If the Levene's test is significant, it means that the variances of the two groups might not be equal. Equal variances is an assumption underlying the t-test. In such cases, we would read our results off the second line of the table. In this case there is no evidence that the variances are different so we can use the t-test results on the first line (p=0.423), indicating no evidence of difference in cognitive test scores across our two sites.



Independent Samples Test

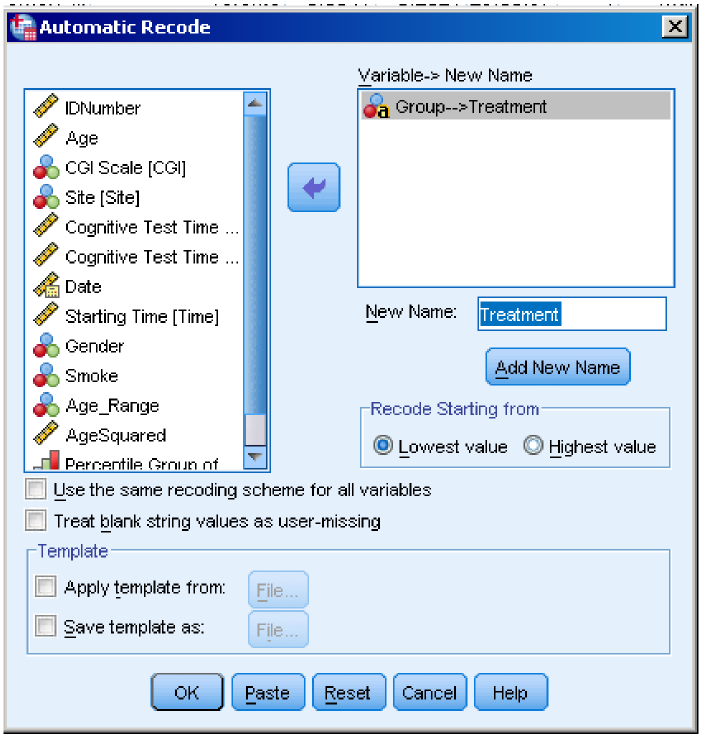
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Cognitive Test Time 1	Equal variances assumed	.644	.423	2.826	788	.005	.3782945	.1338587	.1155326	.6410564
	Equal variances not assumed			2.829	784.517	.005	.3782945	.1337079	.1158269	.6407620

Analysis of Variance (ANOVA)

In situations where the categorical variable has more than 2 levels we are usually interested in a more general test of differences between any of the group means, instead of just comparing two groups. ANOVA will test if there is any difference between groups. The null hypothesis for ANOVA is that all the groups have the same mean, so if the results of the test are significant, we reject this null hypothesis and conclude that there is some difference between the groups. This does not tell us where this difference lies. The SPSS Means (which we saw at the beginning of this section) can do a quick ANOVA, but a more complete procedure is the SPSS ANOVA. Let's test the cognitive test means by treatment group.

Before running the test we will need to transform the variable "Group" into a numeric variable, which we do very easily using SPSS Automatic Recode.

> Go to analysis → Compare Means → One-way ANOVA. Notice that you cannot see the variable "Group" among the available variables. That is because "Group" is a string variable with values that are too long. In order for "Group" to be available for t-test it needs to be recoded into another string variable with 8 or fewer characters or into a numeric variable (numeric is always preferred). SPSS make it easy to recode categorical string variables into numeric through Automatic Recode command. Go to Transform → Automatic Recode and move Group to the "Variable" box. Automatic Recode will transform Group into a numeric variable, with appropriate labels. You need to enter the name of this new variable in the box "New Name" and click "Add New Name".



Variable-> New Name: Group-->Treatment

New Name: Treatment

Recode Starting from: ● Lowest value ○ Highest value

□ Use the same recoding scheme for all variables

□ Treat blank string values as user-missing

Template

□ Apply template from:

□ Save template as:

Tip – It is a good idea to avoid string variables and use numeric variables with labels instead. This not only makes the data file smaller but also ensures that the variables are suitable for any procedure.

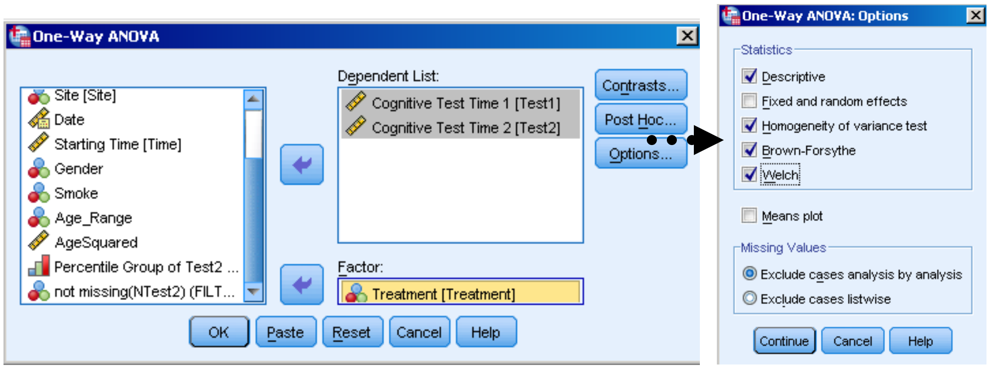
The Automatic Recode will create numeric variables equivalent to the string version. These two versions of the same variables are totally equivalent for any statistical analysis, except that some SPSS procedures may not accept the string version as we just saw with the t-test. Now we have the numeric version of “Group”, called “Treatment”, and we can see in the SPSS output that for the variable “Treatment”, 1 = Control, 2 = Treatment 1 and 3 = Treatment 2. Using this variable, let’s try the ANOVA again.

```
AUTORECODE VARIABLES=Group
  /INTO Treatment
  /PRINT.
Group into Treatment (Treatment)
Old Value     New Value  Value Label

Control               1  Control
Treatment 1           2  Treatment 1
Treatment 2           3  Treatment 2
```

ANOVA makes the same assumptions as the t-test above and they are actually equivalent when there are only two groups.

Go to Analysis → Compare Means → One-Way ANOVA. Enter Cognitive Test 1 and 2 in the “Dependent List” and Treatment in the “Factor”. Click Options. You may want to check “Descriptives” to take a look at the means. “Homogeneity of Variance test” is the Levene’s test we saw in the t-test. If this test is significant then the variances differ across groups, which is a violation of the ANOVA assumption making ANOVA results potentially incorrect. In that case, Wetch and Brown-Forsythe tests would be more appropriate as would non-parametric tests (see below) or variance-stabilizing transformations.



Dependent List:
Cognitive Test Time 1 [Test1]
Cognitive Test Time 2 [Test2]

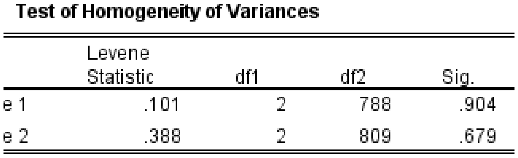
Factor:
Treatment [Treatment]

Statistics
☑ Descriptive
□ Fixed and random effects
☑ Homogeneity of variance test
☑ Brown-Forsythe
☑ Welch

□ Means plot

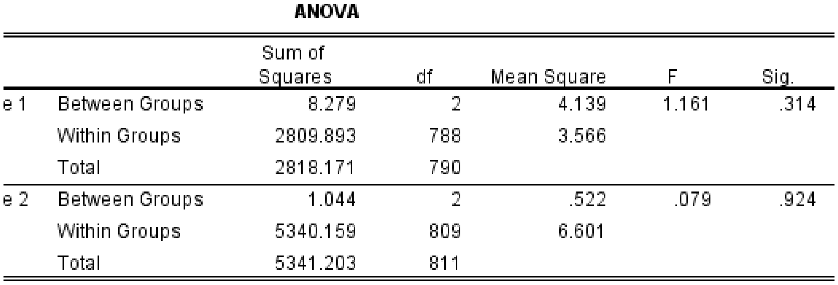
Missing Values
● Exclude cases analysis by analysis
○ Exclude cases listwise

The ANOVA is conducted for both cognitive tests. Fist there is a test of homogeneity of variance, given that this is an of the assumption of ANOVA. There is not evidence to say that the variances are not the same. We then can see that the ANOVA is not significant and the Levene’s test shows that the variances of cognitive test scores are not significantly different across groups. It is also important to test for normality, although ANOVA is usually quite robust to violations of the normality assumption. The third table showed below is to be used when the test for homogeneity of variance is significant, which is not the case here.



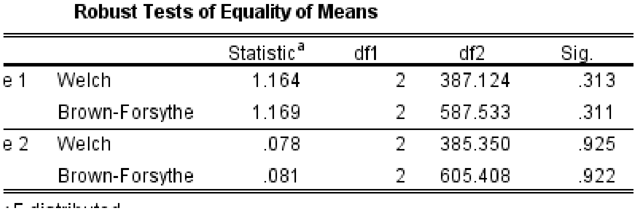
Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
e 1	.101	2	788	.904
e 2	.388	2	809	.679



ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
e 1	Between Groups	8.279	2	4.139	1.161	.314
	Within Groups	2809.893	788	3.566		
	Total	2818.171	790			
e 2	Between Groups	1.044	2	.522	.079	.924
	Within Groups	5340.159	809	6.601		
	Total	5341.203	811			

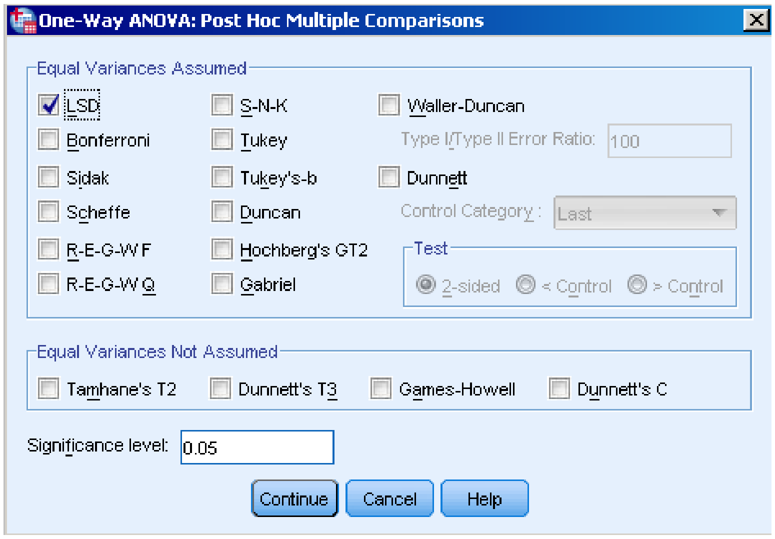


Robust Tests of Equality of Means

		Statistic[a]	df1	df2	Sig.
e 1	Welch	1.164	2	387.124	.313
	Brown-Forsythe	1.169	2	587.533	.311
e 2	Welch	.078	2	385.350	.925
	Brown-Forsythe	.081	2	605.408	.922

Post-Hoc Multiple Comparison Tests

If the results of our ANOVA are significant, we will want to know where the difference lies. A common approach is to compare all pairs of groups using post-hoc multiple comparison tests. In this case it involves 3 tests (there are 3 possible group comparisons – 1-2, 1-3 and 2-3) but depending on the case it can involve many more tests. When we set the threshold for significance at "p-value < 0.05", we are subjecting ourselves to a 5% probability of error (error = concluding the difference is significant when it is not). 5% seems reasonably small, but it also means that in a scenario where there is no difference, one in 20 tests will be incorrectly significant. Because of that, when we conduct many tests we tend to correct the threshold for the p-value for individual tests so that even with 20 tests the likelihood of in general a test being incorrectly significant is not larger than 5%. This is done in SPSS by clicking in "Post-Hoc" in the initial ANOVA window and then selecting "Bonferroni" as in the screenshot below. As you can see there are many other ways to make this adjustment, but Bonferroni is the most conservative and is generally recommended.



One-Way ANOVA: Post Hoc Multiple Comparisons

Equal Variances Assumed

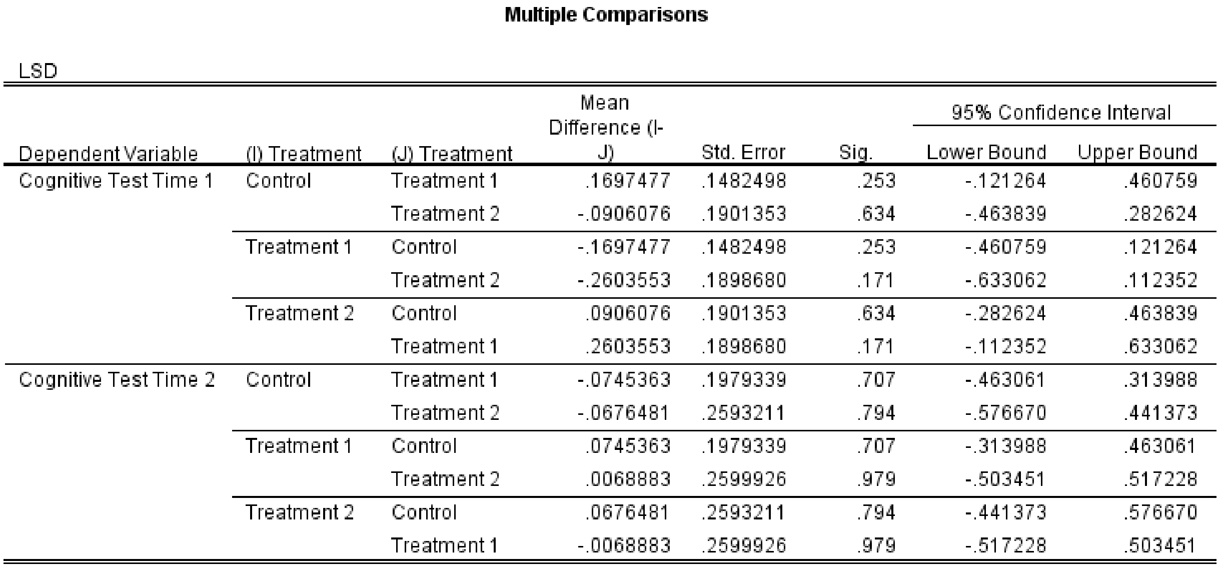
☑ LSD ☐ S-N-K ☐ Waller-Duncan
☐ Bonferroni ☐ Tukey Type I/Type II Error Ratio: 100
☐ Sidak ☐ Tukey's-b ☐ Dunnett
☐ Scheffe ☐ Duncan Control Category : Last
☐ R-E-G-W F ☐ Hochberg's GT2 Test
☐ R-E-G-W Q ☐ Gabriel ● 2-sided ○ < Control ○ > Control

Equal Variances Not Assumed

☐ Tamhane's T2 ☐ Dunnett's T3 ☐ Games-Howell ☐ Dunnett's C

Significance level: 0.05

SPSS will test all possible combination of groups and will output Bonferroni adjusted p-values for them. In this case we can see that there is no significant comparison. Notice also that in cases like this, where the ANOVA is not significant, the post-hoc test is usually not performed..



Multiple Comparisons

LSD

Dependent Variable	(I) Treatment	(J) Treatment	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Cognitive Test Time 1	Control	Treatment 1	.1697477	.1482498	.253	-.121264	.460759
		Treatment 2	-.0906076	.1901353	.634	-.463839	.282624
	Treatment 1	Control	-.1697477	.1482498	.253	-.460759	.121264
		Treatment 2	-.2603553	.1898680	.171	-.633062	.112352
	Treatment 2	Control	.0906076	.1901353	.634	-.282624	.463839
		Treatment 1	.2603553	.1898680	.171	-.112352	.633062
Cognitive Test Time 2	Control	Treatment 1	-.0745363	.1979339	.707	-.463061	.313988
		Treatment 2	-.0676481	.2593211	.794	-.576670	.441373
	Treatment 1	Control	.0745363	.1979339	.707	-.313988	.463061
		Treatment 2	.0068883	.2599926	.979	-.503451	.517228
	Treatment 2	Control	.0676481	.2593211	.794	-.441373	.576670
		Treatment 1	-.0068883	.2599926	.979	-.517228	.503451

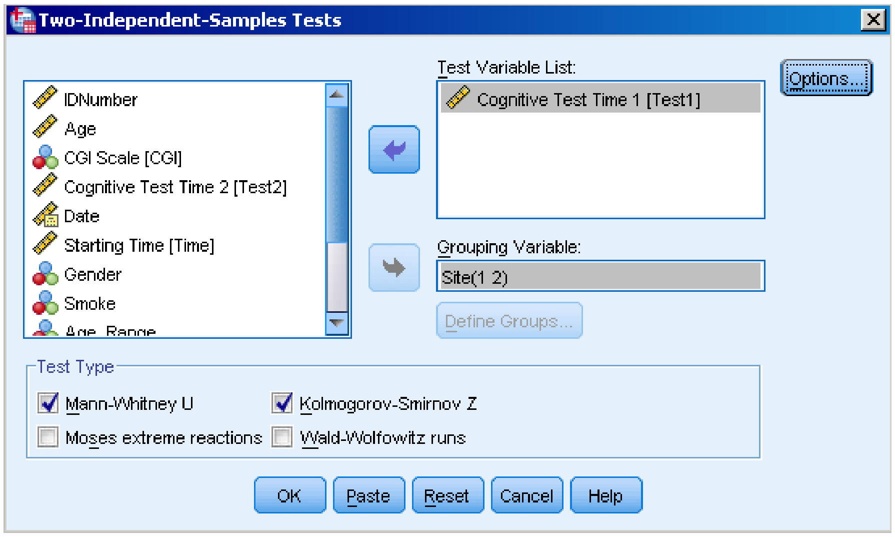
Non-parametric Tests

The ANOVA F-test and the t-test are robust to departures from normality, but these tests may not be valid with small samples and highly non normal samples, or when the variance is different across groups. In these cases it is advisable to use a non-parametric test, which does not make these assumptions and in general makes very few assumptions at all.

Non-parametric Tests for Two Independent Samples

> There are a few non-parametric tests considered equivalent to the 2 sample t-test. Go to Analysis → Non-parametric tests → Legacy Dialogue → 2 Independent samples…. Select the Cognitive test and Site variables just as you did for the t-test. There are 4 test types, select the Mann Whitney U and Kolmogorov-Smirnov tests and click "OK". The other two tests are very specific and we will not address them. Note that in this case we really do not need to use non-parametric tests because the assumptions underlying our ANOVA and t-test were valid.

You will notice that SPSS has a dialogue options for non-parametric tests that are not listed under the Legacy Dialogue option. These are options implemented more recently in SPSS and they are more iterative, but they do basically the same statistical analyses. You should be able to perform the same non-parametric test above by selecting "Independent Samples" options outside of the Legacy Dialogue and then "Compare Medians".



Test Variable List: Cognitive Test Time 1 [Test1]

Grouping Variable: Site(1 2)

Test Type

☑ Mann-Whitney U ☑ Kolmogorov-Smirnov Z

☐ Moses extreme reactions ☐ Wald-Wolfowitz runs

Of the all the non-parametric tests available to us, the **<u>Mann-Whitney U</u>** test is the most similar to the t-test. It does not compare the means of the distributions, but investigates whether the distribution of scores in one group tends to have higher values than in the other group. This test is also called Wilcoxon rank sum test. The p-value, which is the "Assymp. Sig. (2-tailed)" in the table below, is much smaller than our threshold of 0.05 so the test is significant.

a

	Cognitive Test Time 1
Mann-Whitney U	68884.500
Wilcoxon W	140137.500
Z	-2.799
Asymp. Sig. (2-tailed)	.005

Grouping Variable: Site

The **Kolmogorov-Smirnov Z** tests whether the distributions are the same. Notice that we could have the same mean or median and yet quite different distributions. Evidence of different distributions would also imply some sort of association or dependence between the cognitive test and the groups. In this case we can see that the test is also significant.

a

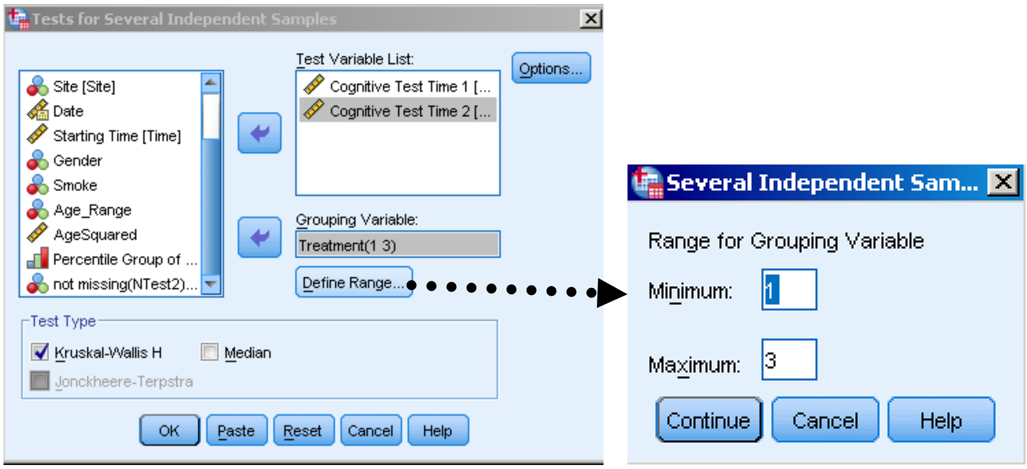
		Cognitive Test Time 1
Most Extreme Differences	Absolute	.113
	Positive	.005
	Negative	-.113
Kolmogorov-Smirnov Z		1.590
Asymp. Sig. (2-tailed)		.013

Grouping Variable: Site

Non-parametric Tests for More Than Two Independent Samples

When we have more than two independent samples we need a non-parametric test that is similar to ANOVA. The most used is the Kruskal-Wallis test, which compares the distribution of many independent samples.

> Go to Analysis → Non-Parametric Tests → Legacy Dialogue → K Independent Samples… Insert Cognitive Test at Time 1 and at time 2 in the “Test Variable List” and Treatment in the “Grouping Variable List” and you will also need to define the range of values to be used in the “Grouping Variable”. Notice that here you define a range and not only two groups like in the previous tests. After defining the groups to be analyzed, hit Continue and check the boxes for the Kruskal-Wallis test and the Median test. Click OK.



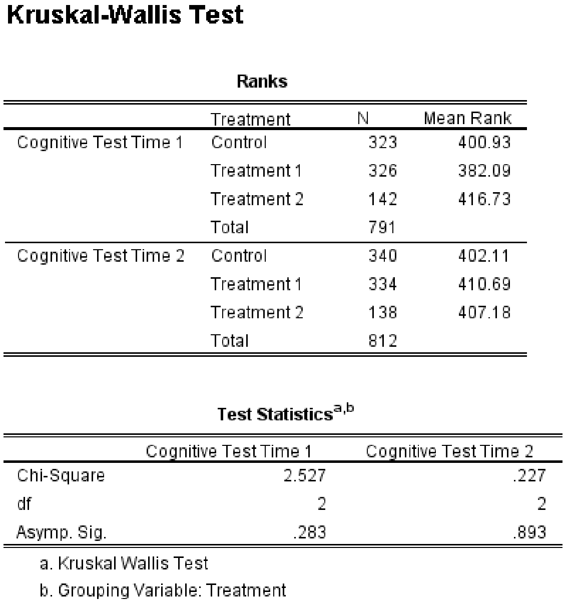
Test Variable List:
Cognitive Test Time 1 [...
Cognitive Test Time 2 [...

Grouping Variable: Treatment(1 3)

Test Type
☑ Kruskal-Wallis H
□ Median
□ Jonckheere-Terpstra

Minimum: 1
Maximum: 3

The **Kruskal-Wallis test** is similar to performing an ANOVA with ranked data. It does not assume normality or equality of variance, but it carries the assumption that the data in the different groups have the same shape. Below we can see the results of the Kruskal-Wallis test, which generates a chi-square statistic. In this example, the p-value is low, implying that the distributions are different across groups. If we wish to interpret this result further, we can use a series of Mann-Whitney U tests.



Kruskal-Wallis Test

Ranks

	Treatment	N	Mean Rank
Cognitive Test Time 1	Control	323	400.93
	Treatment 1	326	382.09
	Treatment 2	142	416.73
	Total	791	
Cognitive Test Time 2	Control	340	402.11
	Treatment 1	334	410.69
	Treatment 2	138	407.18
	Total	812	

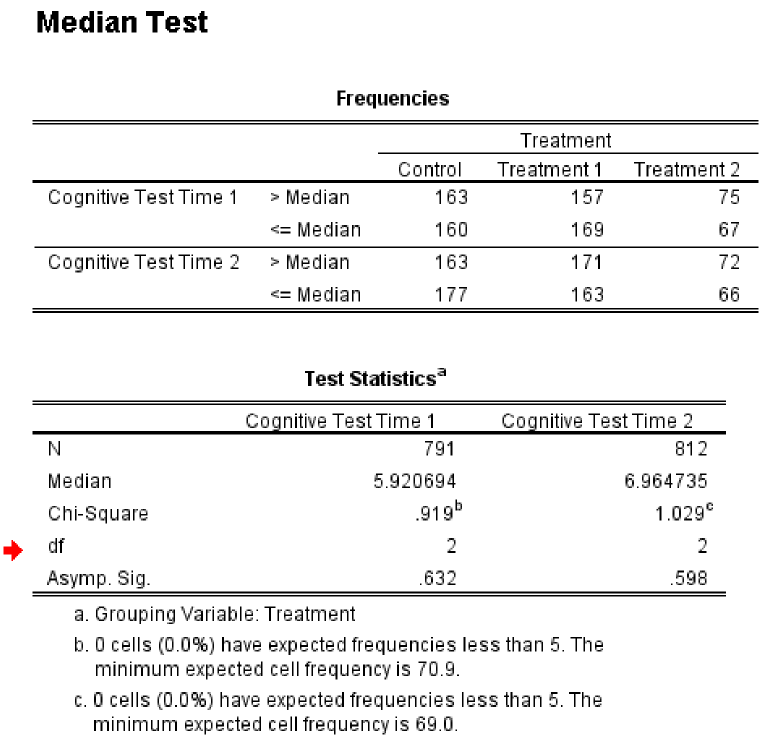
Test Statistics[a,b]

	Cognitive Test Time 1	Cognitive Test Time 2
Chi-Square	2.527	.227
df	2	2
Asymp. Sig.	.283	.893

a. Kruskal Wallis Test

b. Grouping Variable: Treatment

Similar to the Kruskal-Wallis test is the **Median test**, which can be used for two or more groups. The Median test tests if the groups from a population with the same median. It does not use all of the range of values; it only classifies values as being above or below the median so it is not as powerful as the Kruskal-Wallis test. Below we have the results from the Median test applied to the same data as the Kruskal-Wallis above, and the results are equivalent for this example.



Median Test

Frequencies

		Treatment		
		Control	Treatment 1	Treatment 2
Cognitive Test Time 1	> Median	163	157	75
	<= Median	160	169	67
Cognitive Test Time 2	> Median	163	171	72
	<= Median	177	163	66

Test Statistics[a]

	Cognitive Test Time 1	Cognitive Test Time 2
N	791	812
Median	5.920694	6.964735
Chi-Square	.919[b]	1.029[c]
df	2	2
Asymp. Sig.	.632	.598

a. Grouping Variable: Treatment

b. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 70.9.

c. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 69.0.

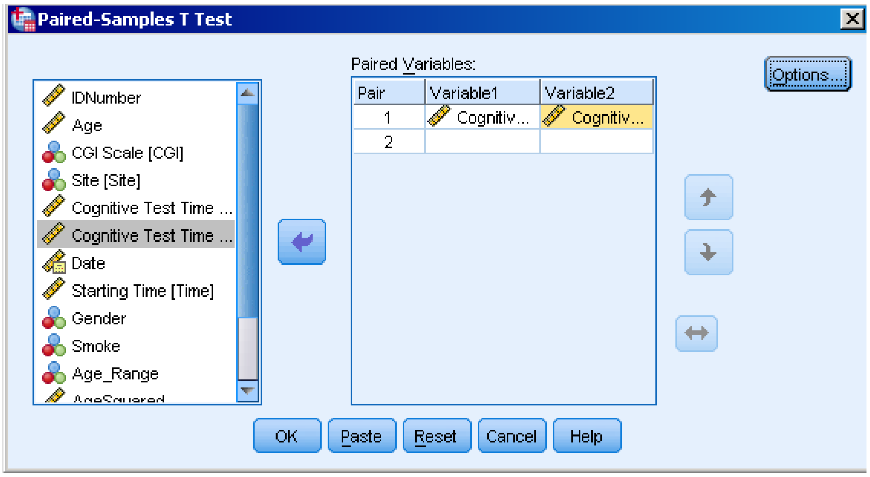
- **Paired Continuous Variables**

Sometimes our research questions may require us to compare means from samples that are not independent. This occurs most commonly when multiple samples are taken on the same group of individuals (e.g. repeated measurements over time, such as pre- and post-intervention). It can also happen in settings not involving time, for example measurements on the left and right eyes of each study participant. Given that both eyes are in the same individual the measurements are not independent. In such a case they are called matched pairs.

The Paired t-test

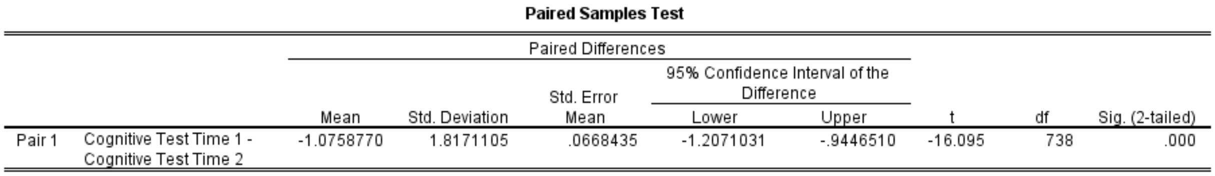
In our data, a paired t-test would be appropriate if we wanted to test if there is a difference between the cognitive test scores at time 1 and time 2. The idea of the paired t-test is that we create a third variable which is the difference "time 2 minus time 1" and test whether the mean of this new variable is zero. This is equivalent to a one-sample t-test (i.e. comparing a quantity to a fixed constant, in this case zero).

> To conduct the test in SPSS, go to Analysis → Compare Means → Paired Samples T-test, select both Cognitive test variables and move them to the "Paired Variable" Box. Then you can click OK.



Pair	Variable1	Variable2
1	Cognitiv...	Cognitiv...
2		

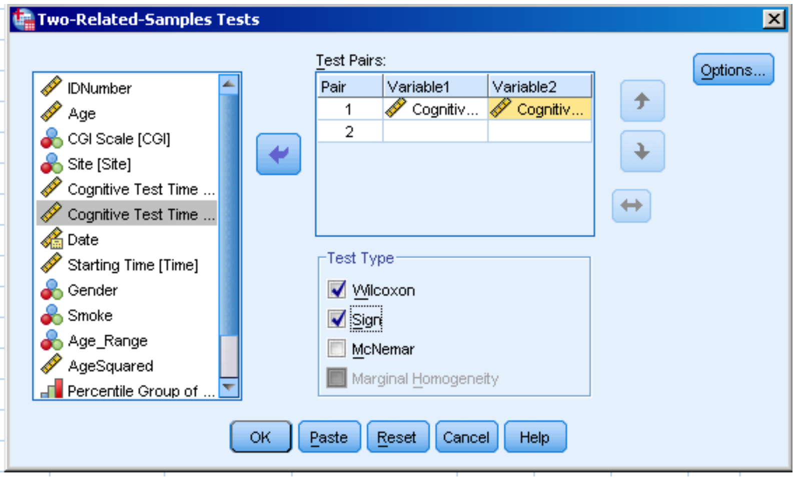
The main result of the paired test is below. In this case "Mean" refers to the mean of the difference between the cognitive test at time 1 and the test at time 2. This average difference is -1.0758 and is significant, with p-value < 0.001. It is important to pay attention to the label on the left, which indicates how the difference is being calculated. In this case, for example, the average score at time 1 is lower than the average score at time 2 hence the negative difference.



Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Cognitive Test Time 1 - Cognitive Test Time 2	-1.0758770	1.8171105	.0668435	-1.2071031	-.9446510	-16.095	738	.000

The paired t-test assumes normality of the differences. For a non-parametric equivalent test we can use the Wilcoxon Signed Rank test or the Sign Test. Both these tests can be conducted in Analysis → Non-parametric tests → Legacy Dialogue → 2 Related Samples. You can then move both cognitive tests to the “Test Pair(s) List” box, check both Wilcoxon and Sign and click OK.



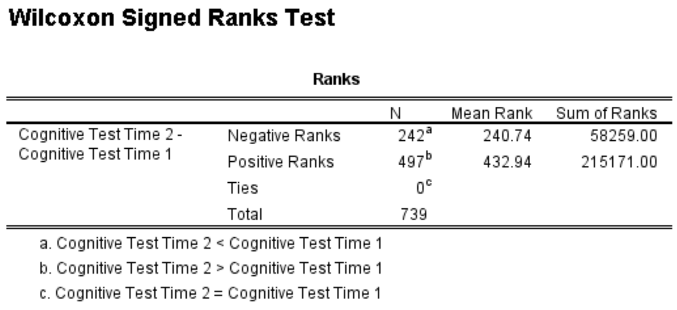
Test Pairs:

Pair	Variable1	Variable2
1	Cognitiv...	Cognitiv...
2		

Test Type
☑ Wilcoxon
☑ Sign
☐ McNemar
☐ Marginal Homogeneity

The Wilcoxon Signed Rank Test

The **Wilcoxon signed rank test** tests whether the medians of the related samples are different (or if the median of the difference is zero). The procedure will work with the rank of the difference between the two variables and return a test statistics that can be compared to the normal distribution or tabled values for the test. Below we see that we have evidence to reject the hypothesis that the medians are the same (test is significant).

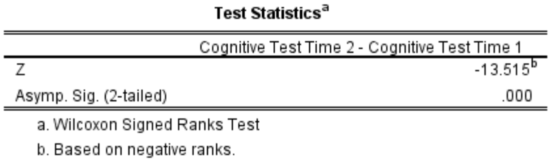


Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
Cognitive Test Time 2 - Cognitive Test Time 1	Negative Ranks	242[a]	240.74	58259.00
	Positive Ranks	497[b]	432.94	215171.00
	Ties	0[c]		
	Total	739		

a. Cognitive Test Time 2 < Cognitive Test Time 1
b. Cognitive Test Time 2 > Cognitive Test Time 1
c. Cognitive Test Time 2 = Cognitive Test Time 1



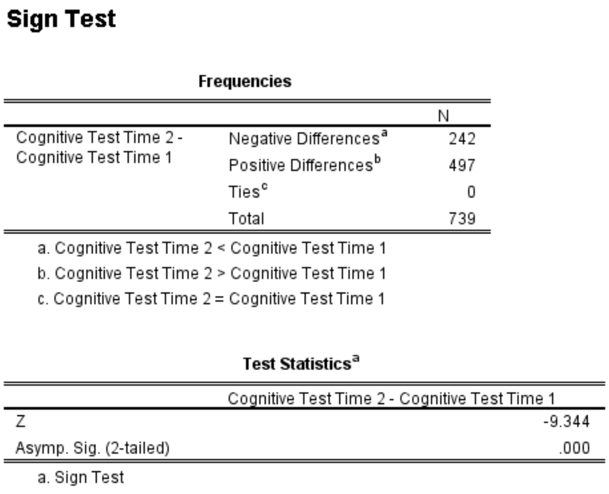
Test Statistics[a]

	Cognitive Test Time 2 - Cognitive Test Time 1
Z	-13.515[b]
Asymp. Sig. (2-tailed)	.000

a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.

The Sign Test

The **sign test** is less powerful that the Wilcoxon signed rank test and because of that the latter is recommended. The sign test will test the hypothesis that the probability of X > Y is higher than 0.5 where X is the first of the paired measure and Y is the second. Notice that for our data the test resulted not significant, while the Wilcoxon signed rank was significant, reflecting the lower power of the signed rank test.



Sign Test

Frequencies

		N
Cognitive Test Time 2 - Cognitive Test Time 1	Negative Differences[a]	242
	Positive Differences[b]	497
	Ties[c]	0
	Total	739

a. Cognitive Test Time 2 < Cognitive Test Time 1
b. Cognitive Test Time 2 > Cognitive Test Time 1
c. Cognitive Test Time 2 = Cognitive Test Time 1

Test Statistics[a]

	Cognitive Test Time 2 - Cognitive Test Time 1
Z	-9.344
Asymp. Sig. (2-tailed)	.000

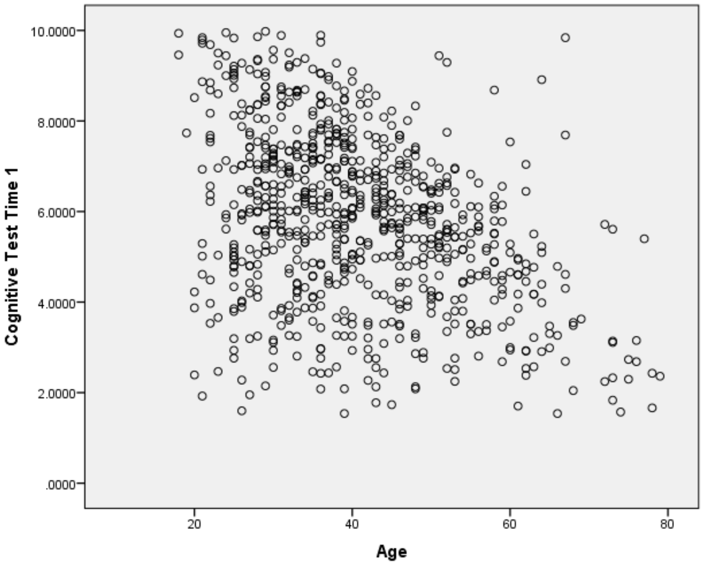
a. Sign Test

- ## Analyses Involving Two Continuous Variables

Scatter-plots

When we have two continuous variables and we want to study the relationship between them, an often useful tool is the scatter-plot. It is a good idea to use the scatter-plot as the first step in investigating the relationship between continuous variables..

> In SPSS you can easily create a scatter-plot by selecting Graphs → Legacy Dialogs → Scatter/Dot, select "Simple Scatter" and move cognitive test at time 1 to the "Y" box and Age to "X" box. Click OK.



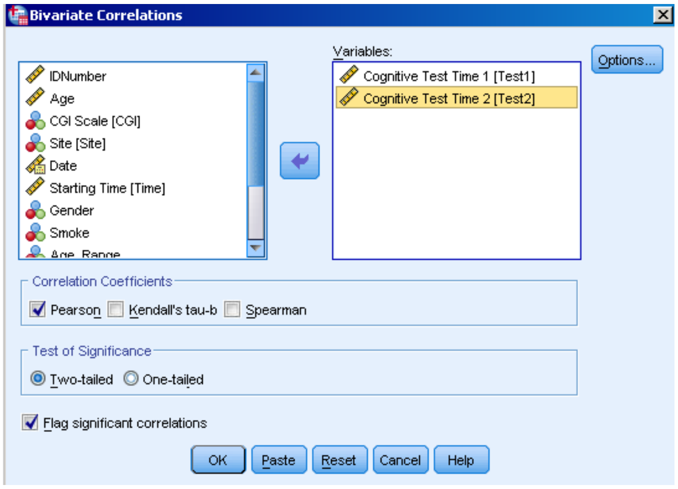
The scatter-plot shows as the age increases the result of the cognitive test tend to be lower, i.e. there is a visual evidence of a conditional relationship between test score and age, meaning that the value of age seems to provide us with some information about the value of the test at time 1. The association in this scatter-plot is negative and it seems to be linear, meaning that the higher the age the lower the cognitive test. If the association were positive then we would have a situation where the higher the age the higher the cognitive test.

Notice that scatter-plots will allow us to easily identify outliers, which are points that compared to the others seems quite unusual. Points like this can influence our statistical results and we should be concerned about them. In the plot above we see very few points where the age is higher and the cognitive test is also high. Because they are few, they seem a little unusual and it is good to ensure there is nothing wrong with these subjects (we may find that they had a different treatment, or they started the treatment earlier, or their cognitive test were mistyped, or maybe they are just ok). It is important to understand WHY the value is an outlier before deciding what to do about it. If the measurement is not an error, it should be kept in the data set. In such cases, we may wish to consider data transformations or alternate methodology.

Pearson Correlation Coefficients

The most widespread measure of association for two continuous variables is the Pearson correlation coefficient, which measures the ***linear*** association between variables.

> To calculate the Correlation Coefficient in SPSS go to Analysis → Correlate → Bivariate. Move both cognitive tests to the “Variable” box and check “Spearman”. Click OK.



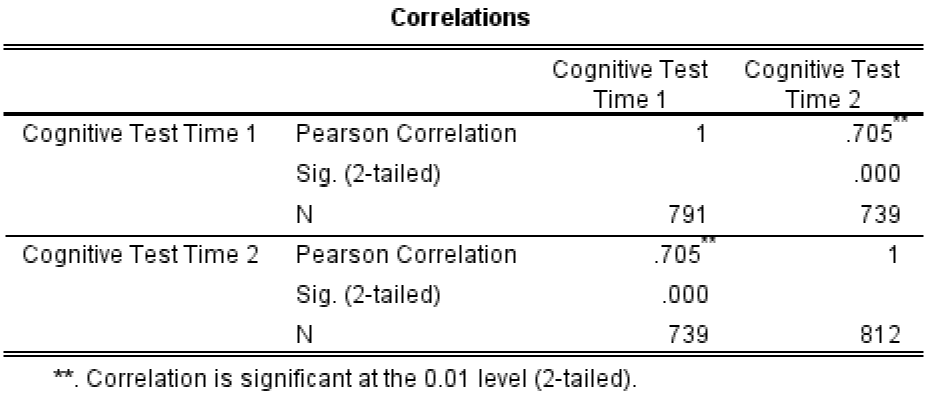
Variables:
Cognitive Test Time 1 [Test1]
Cognitive Test Time 2 [Test2]

Correlation Coefficients
☑ Pearson □ Kendall's tau-b □ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

SPSS will output the table below, which shows the Pearson correlation coefficient between both cognitive tests, which is 0.705. It also shows the p-value (sig. (2-tailed)) and the sample size used (791). The p-value tests whether the correlation coefficient is zero. In this case it is highly significant indicating that there is a significant relationship between the two variables.



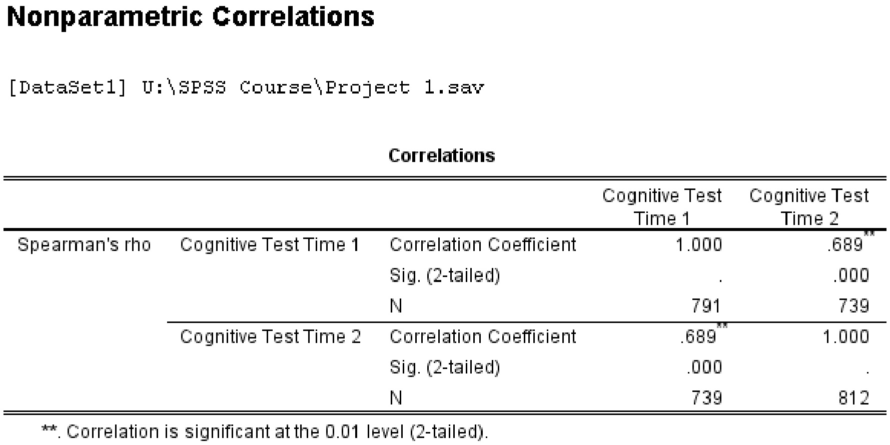
Correlations

		Cognitive Test Time 1	Cognitive Test Time 2
Cognitive Test Time 1	Pearson Correlation	1	.705**
	Sig. (2-tailed)		.000
	N	791	739
Cognitive Test Time 2	Pearson Correlation	.705**	1
	Sig. (2-tailed)	.000	
	N	739	812

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

Spearman Correlation Coefficients

The correlation coefficient is an association measure that arises naturally when both variables are normally distributed and in particular, the statistical test assumes a normal distribution. When the distributions are non-normal we can use a non-parametric equivalent, which is the Spearman correlation coefficient. We can see below that the results from the Spearman correlation coefficient are very similar to Pearson correlation coefficient.



Nonparametric Correlations

```
[DataSet1] U:\SPSS Course\Project 1.sav
```

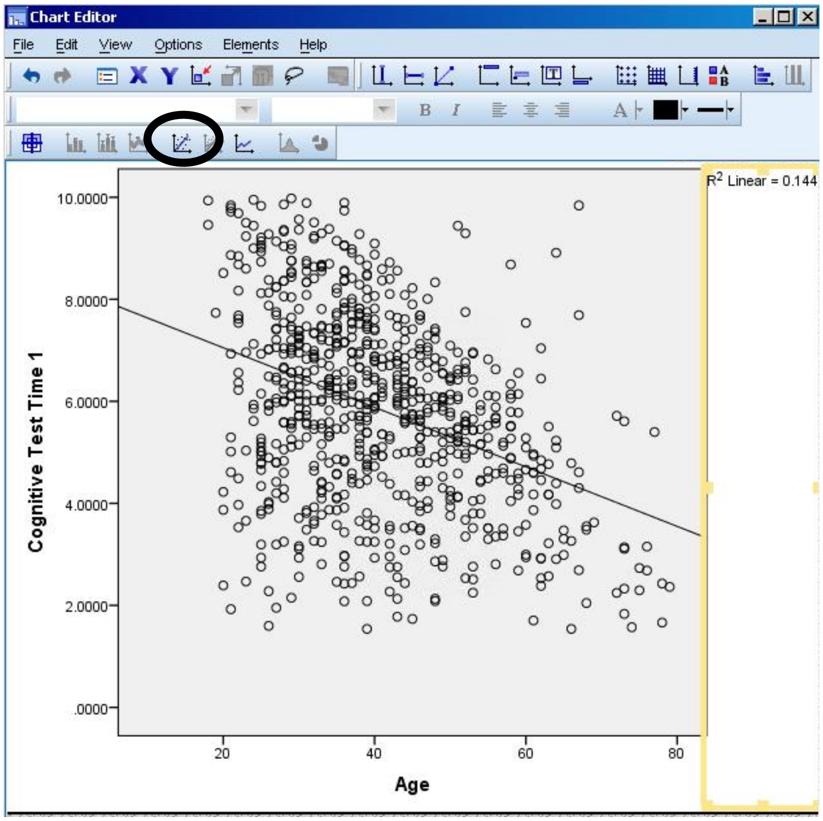
Correlations

			Cognitive Test Time 1	Cognitive Test Time 2
Spearman's rho	Cognitive Test Time 1	Correlation Coefficient	1.000	.689**
		Sig. (2-tailed)	.	.000
		N	791	739
	Cognitive Test Time 2	Correlation Coefficient	.689**	1.000
		Sig. (2-tailed)	.000	.
		N	739	812

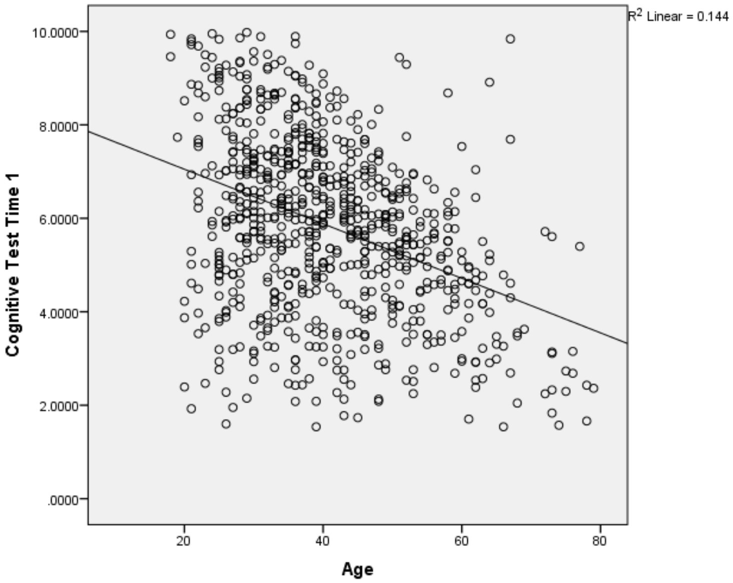
**. Correlation is significant at the 0.01 level (2-tailed).

As an exercise you may want to try and exclude the outlier that we found using the scatter-plot, and recalculate both correlation coefficients. You will see that the effect that that single subject has in the Pearson correlation coefficient is much greater than the effect it has on the Spearman coefficient. That is because the non-parametric procedures are usually robust to outliers.

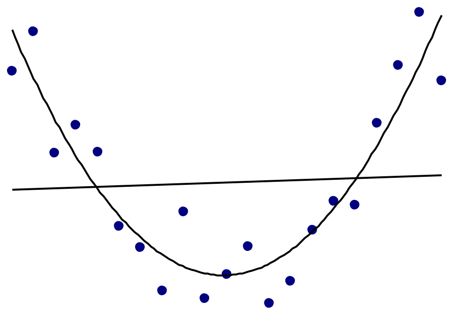
It is important to note that correlation coefficients measures the strength of a <u>Linear Association.</u> It is possible to have a strong association and yet a very low correlation coefficient if the association is not linear. The linear association in a scatter-plot is represented by the fit of a straight line through the points. In SPSS, you can double click the graph in the Output and a window will open. Then you will see a button that inserts the straight line, see screen-shot below (notice that SPSS has different types of scatter-plots and they may be different across versions, so you may not see things exactly as below).



Once you click the button you will see the option for fitting different models. Below we have the linear model.



The graph below exemplifies an association that is not linear. In this case, the simple correlation coefficient will be very low even though there is a very strong (non-linear) association between the variables. Modeling this association with a linear model is very inappropriate.



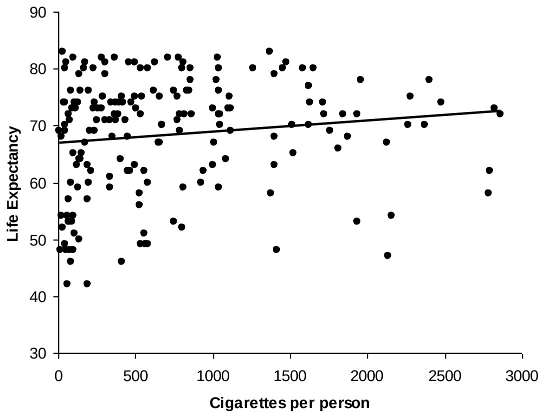
Exercise

Considere the variable CGI. Run a parametric and non-parametric test to assess the difference of Cognitive score at time 1 by CGI group. Interpret the results.

Association, Causation, Models and our Research

Association is a term that we use a lot in our daily work and it means the extent to which variables relates to each other. We just saw how to assess association across different types of variables, using different tests. In the end, the data collection process usually aims to study some sort of association, usually through multivariate modeling techniques which are beyond the scope of today's introductory course.

Association is very important if not the most important thing we look at when we do statistical analysis. It is usually our primary goal. In many cases we use it as a proxy for what we would really like to measure; causality. If smoking causes lung cancer than we expect that if we look at data we will find that smoking is associated with lung cancer. The problem is that the presence of association does not guarantee causality just as the absence of an association does not mean an absence of causality. Associations can be very misleading as proxies for causal effects. The graph below shows data from the World Health Organization on almost all the countries, their life expectancy and consumption of cigarettes per capita. The association in the graph is positive, meaning that the more cigarettes per person, the higher the life expectancy, but does that mean that countries should increase their consumption of cigarettes per person in order to increase their life expectancy? Of course not, there is association but not causation. The association comes from some third variable that we call confounder, not from causation.



Similarly, if we find that people who took a treatment have lower depression levels than those who did not take it, then we are looking at an association and we need to be careful in our interpretation because this association may not translate directly into a causal effect of the treatment on depression.

Causation is a very tricky thing and at this point the only statistical tools that can prove causality from association with very minor assumptions are randomized controlled trials (RCT). The controlled trial allows us to manipulate the treatment and observe the results, while the randomization allows us to controls external influences that can be confounded with our treatment. When we cannot use RCTs, other statistical techniques exist that are appropriated for causal inference, but none of them will give results as reliable as the RCTs.

In order to analyze RCTs data and data from other sources, we often need to go beyond measures of association like the correlation coefficient. We need to model the causal process and that often involves more sophisticated statistical tools that we will not cover here. From simple regression analysis to random effect models and structural equation models, we have a great number of statistical tools available to us that will provide us with measures of effect size and are often much more useful than simple measures of association. The correct use of statistical models requires not only the ability to use a statistical software, but also statistical knowledge of the assumptions underlying these models as well as causal inference. We recommend that you contact the Biostatistical Consulting Service if your research involves these more advanced analyses and we will always be happy to provide a customized solution to address your research questions in an appropriate manner.