INTRODUCTION TO SPSS

PART I

INTRODUCTION

Background

This handbook is designed to introduce **SPSS for Windows**. It assumes familiarity with Microsoft windows and standard windows-based office productivity software such as word processing and spreadsheets.

SPSS for Windows is a popular and comprehensive data analysis package containing a multitude of features designed to facilitate the execution of a wide range of statistical analyses. It was developed for the analysis of data in the social sciences - SPSS means Statistical Package for Social Science. It is well suited to analysing data from surveys and database.

The practical uses dataset from a cross-sectional survey of respiratory function and dust levels amongst foundry workers. The object of the survey data was to determine whether the dust levels found in the foundries have any effect on the respiratory function.

Acquiring the DATA

A number of datasets have been created to enable you to work through this guide. These can be found online or via the **'Shared Data'** folder. To access click the Start button in the bottom left hand corner and type - **shared data** – and press enter, the window explorer will open and then double click: **mhs** > **health methodology course data** >

We suggest you **copy and paste foundry.sav, foundry.xls, and foundrysyn.SPS** to your **desktop**.

To access the data online click the link:
http://research.bmh.manchester.ac.uk/biostatistics/teaching/statisticalsupport
and download the relevant SPSS handouts and above datasets to your desktop. You may at some point be asked for your username and password.

Note: *for further information this booklet where possible will link you to a relevant Youtube video explaining the technique discussed.*

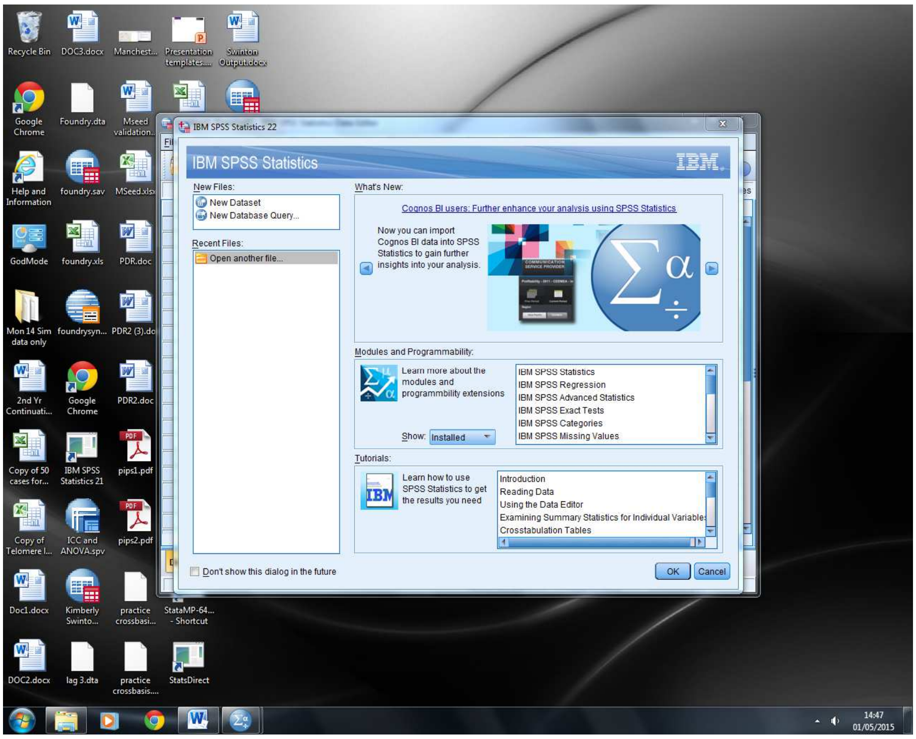
Starting SPSS

After logging on to Windows 7, the user will be presented with a screen containing a number of different icons. Start SPSS by clicking the **Start** button then selecting

All Programs ⟶ **IBM SPSS Statistics** ⟶ **IBM SPSS Statistics 23.0**

Then the **SPSS 23.0 for Windows 7** screen will appear called **Untitled – SPSS Data Editor** (shown below). In the middle of the **Data Editor** screen you can see another window with the following options -

- New Files – Create a new dataset
- Recent Files – Open a previously used dataset
- What's New – Learn about new features in SPSS 23.0
- Modules and Programmability – Links to help menus for advanced users
- Tutorials – Beginners guides to features in SPSS 23.0



IBM SPSS Statistics 22

Cognos BI users: Further enhance your analysis using SPSS Statistics

Now you can import Cognos BI data into SPSS Statistics to gain further insights into your analysis.

Learn more about the modules and programmbility extensions

Show: Installed

Learn how to use SPSS Statistics to get the results you need

☐ Don't show this dialog in the future

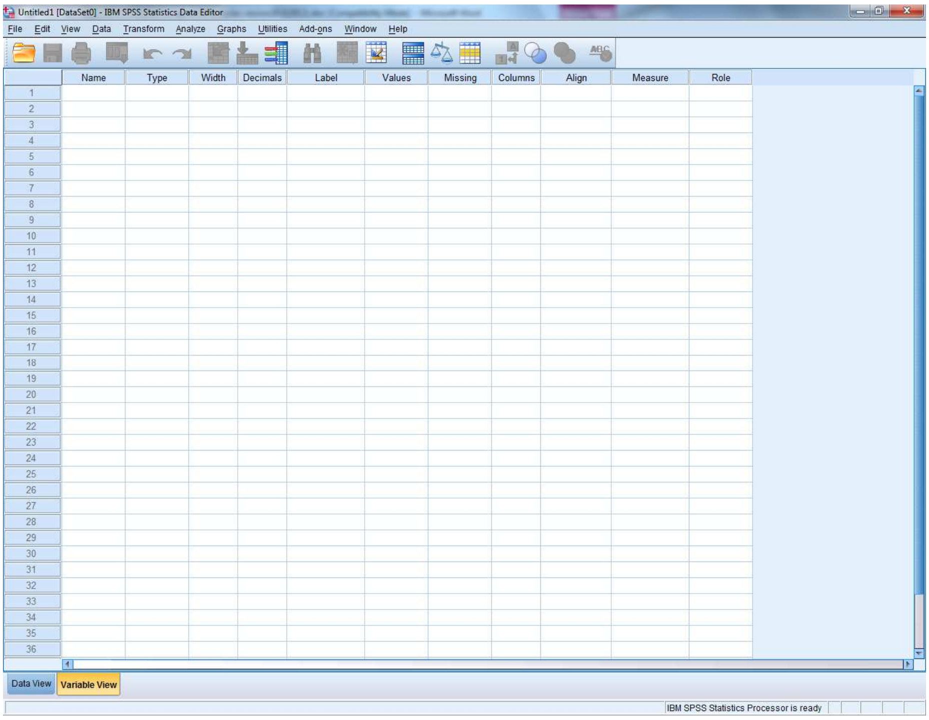
Click, the **New Dataset within the New Files** option, to get a blank SPSS data screen and the maximise your SPSS window.

Data Entry

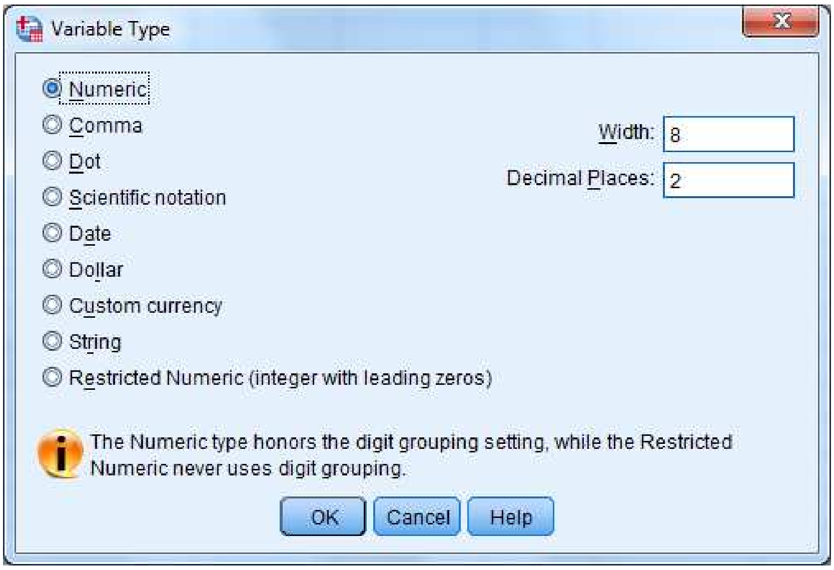
The SPSS Data Editor screen looks like a spreadsheet but there are some important differences. Each row represents the data for a case. A case could be a patient or a laboratory specimen. It could also be a set of results for a patient at a particular time. Each column represents a variable. A variable could be the answer to a question or any other piece of information recorded on each case. Before you enter any data in the spreadsheet you have to create a variable for the information you have collected. You must define a variable for each question in your data set you plan to analyse.

Defining Variables

If you look at the left hand corner at the bottom of the SPSS Data Editor screen, you will see two small tabs labelled: **Data View** and **Variable View**. To create a new variable click on **Variable View** and the following screen will appear.


Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Name
Type
Width
Decimals
Label
Values
Missing
Columns
Align
Measure
Role
Data View
Variable View
IBM SPSS Statistics Processor is ready

Each row describes the attributes of one variable. Begin by entering a variable name in the **Name** column. A variable name can be up to 64 characters long, must contain no spaces, and should be something meaningful. It is best to stick to alphanumeric characters and start with a letter. Once you have entered a name, SPSS defines the variable type as **Numeric**. You may need to change the variable type, to e.g. **String** if you wanted to use text such as names, or to **Date** if you want to enter dates. To do this, click on the cell within the **Type** column. A little combo button will appear on the right hand side, click the button and the following screen will appear.



Variable Type

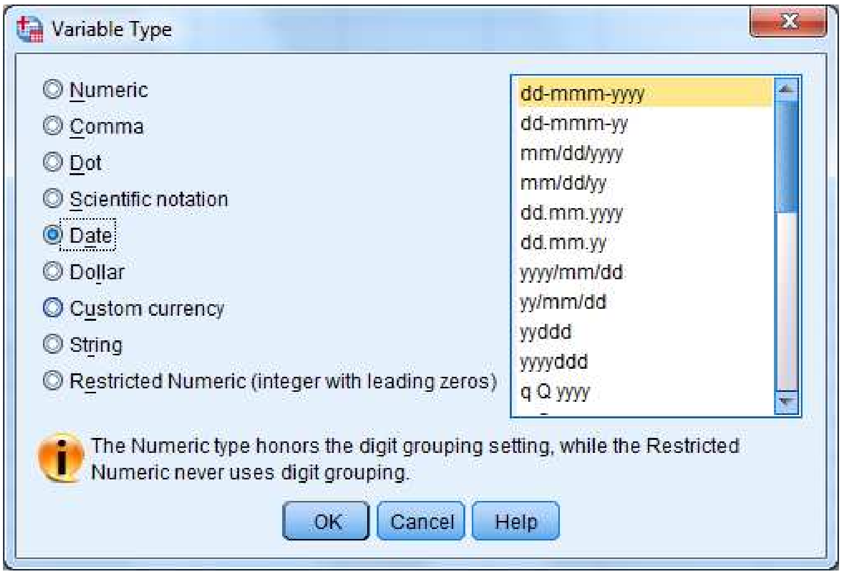
● Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
○ Custom currency
○ String
○ Restricted Numeric (integer with leading zeros)

Width: 8

Decimal Places: 2

The Numeric type honors the digit grouping setting, while the Restricted Numeric never uses digit grouping.

You will usually be working with one of **Numeric**, **Date** or **String** type of data. For Numeric variables you may want to change the decimal places. If the data are integers (whole numbers) such as age in complete years you could alter the decimal places to zero. If the numbers you are planning to enter are very small (0.00072) or you require a high level of precision (21.7865) you may want to increase the number of decimal places. Usually there is no need to change the width from 8, note that width must be larger than the number of decimal places. For a date variable it is best to use a 4 digit year (dd.mm.yyyy)



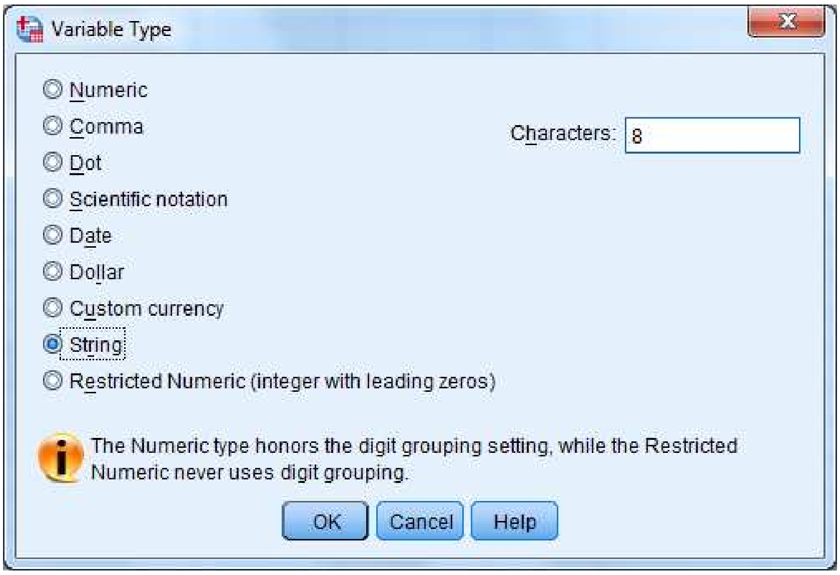
Variable Type

○ Numeric
○ Comma
○ Dot
○ Scientific notation
● Date
○ Dollar
○ Custom currency
○ String
○ Restricted Numeric (integer with leading zeros)

dd-mmm-yyyy
dd-mmm-yy
mm/dd/yyyy
mm/dd/yy
dd.mm.yyyy
dd.mm.yy
yyyy/mm/dd
yy/mm/dd
yyddd
yyyyddd
q Q yyyy

The Numeric type honors the digit grouping setting, while the Restricted Numeric never uses digit grouping.

With text strings you are given the option to change the number of characters



○ Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
○ Custom currency
● String
○ Restricted Numeric (integer with leading zeros)

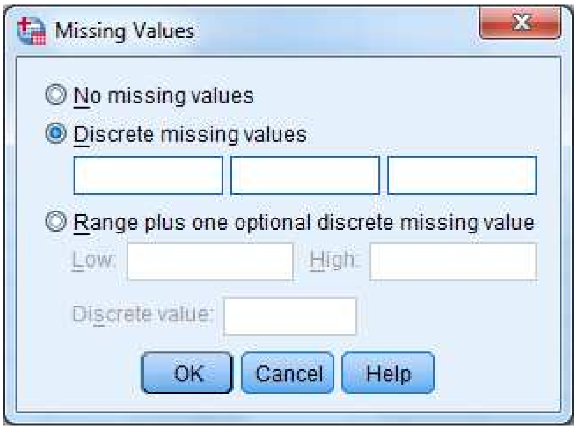
Characters: 8

The Numeric type honors the digit grouping setting, while the Restricted Numeric never uses digit grouping.

Where possible you are strongly advised to use numerical coding rather than strings as this makes statistical analysis easier. If you are entering string data that is longer than 8 characters, you will need to increase the Width from the default of eight. To be able to fully display the string in the **data view** window you may need to increase the numbers of columns in the **variable view** window.

The column missing in the variable view window allows you to define codes that identify a missing value. You can have several values allowing you to distinguish between types of missing data due to the respondent forgetting to answer rather than say not applicable or refused to answer. For example, a code of **-88** could indicate not applicable, and **-99** could indicate the respondent had missed a question out. If a value is defined as a missing value code for a particular variable, subjects with that code will be dropped from the analysis of that variable.

To set up missing value codes for a variable, click on a cell followed by the grey square within the **Missing** column as you did with **Type**. Click **Discrete missing values** and enter the values to represent missing in the boxes below (Up to 3 can be entered). To complete the entry press **OK**



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

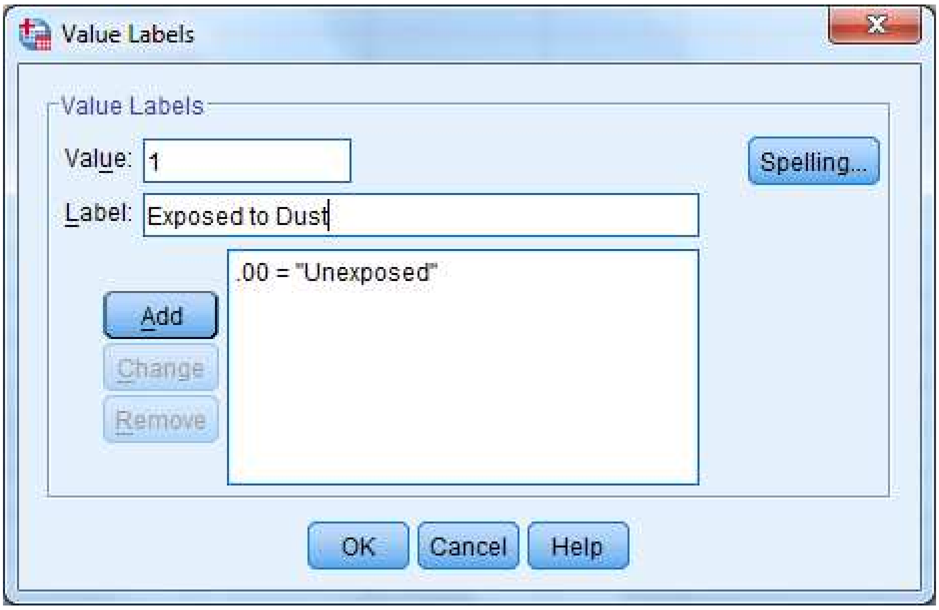
Variable and Value Labels

There are two types of labels in SPSS. A **variable label**, given to a variable gives a clearer description of the variable and will be displayed on the statistical output such as graphs and tables.

The second, a **value label** allows you to describe each of the values in a variable. These labels will be displayed on tables improving readability. For example, *Exposure group* in the following practical has two values "Unexposed" and "Exposure to dust" which are coded as "0" and "1". The label option in the variable view window also allows you to define labels for missing values.

To define a variable label click the cell within a ***Label*** column screen and enter your description of the variable**.**

To define **Value Labels** - click the cell of the **value** column and then the click on the combo button to the right, then enter the **Value** and its associated label then press ***Add***. The added label will then appear in the window below.



Value: 1

Label: Exposed to Dust

.00 = "Unexposed"

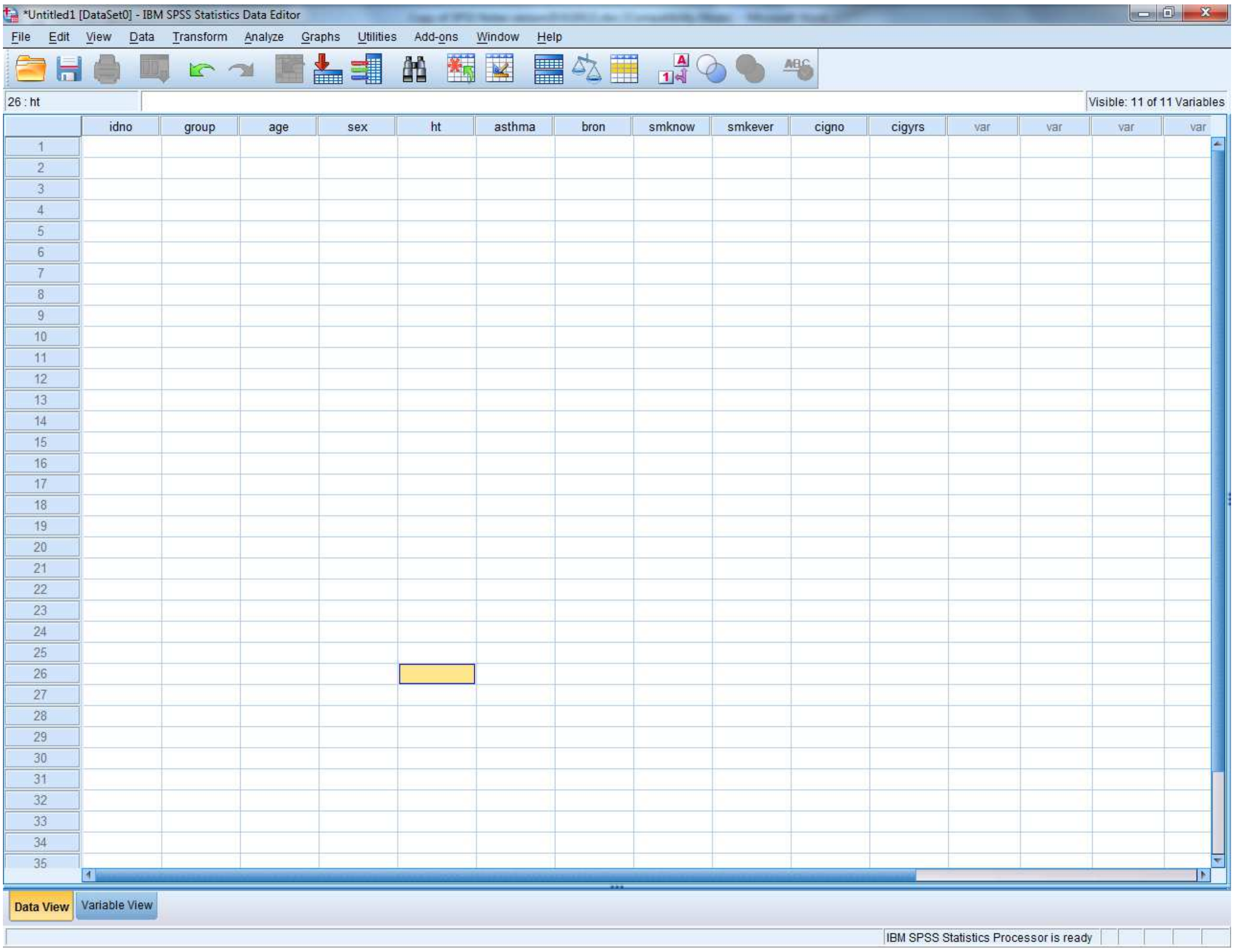
Once you have entered all the value labels for a variable press **OK.**

Exercise The table below lists the example variables from the foundry study. Set-up the following variables

Variable Name	Description (Variable Label)	Missing Data Code	Value Labels for each code
idno	Identification No		
group	Exposure Group		1 = Exposed to dust 0 = Unexposed
age	Age at assessment		
sex			0 = female 1 = male
ht	Height in cms		
asthma	Ever had asthma		0 = No 1 = Yes 2 = Don’t Know
bron	Ever had Bronchitis		0 = No 1 = Yes 2 = Don’t Know
smknow	Do you smoke now		1 = Yes 0 = No
smkever	Have you ever smoked		0 = No 1 = Ex smoker 2 = Current smoker
cigno	No of cigarettes per day	-88	
cigyrs	No of years smoked	-88	

Entering Data

When you finish creating all the variables, you enter the **Data View** and the following screen with all the variable names at the top of the spreadsheet.


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
26 : ht
Visible: 11 of 11 Variables
idno group age sex ht asthma bron smknow smkever cigno cigyrs var var var var
1–35
Data View
Variable View
IBM SPSS Statistics Processor is ready

You can now enter the data as you would in an excel spreadsheet. To make an entry in a particular cell on the spreadsheet use the mouse to move the cursor to select that cell and type in the value. The value will appear in the cell. Click on the mouse, press enter or use the cursor keys to enter that value.

If you attempt to enter data of the wrong type into a variable (for example text into a numeric variable) the data will not be accepted. If incorrect data is entered, it can be overtyped or deleted.

Video Tutorial – Setting up a dataset and entering data

https://www.youtube.com/watch?v=MoKDcPpRa_0

Exercise

The data below are some variables from the foundry study for which you have just entered the variable codes. If you leave a gap in any cell in the worksheet, **SPSS** will put a dot (**.**) and treat it as missing data. To enter the cases, either type the number corresponding to the value label or alternatively display the **Value Labels** of the coded values. These are displayed by using choosing value labels button from the second row of options at the top of either the Data view or Variable View window.

Idno	group	age	Sex	Ht	asthma	bron	smknow	smkever	cigno	cigyrs
1001	Exp.	49	Female	175	No	No	Yes	Curr	20	31
1002	Exp.	46	Female	168	Yes	No	Yes	Curr	20	11
1003	Non	34	Female	180	No	No	No	Never		
1004	Non	34	Male	180	No	No	Yes	Curr	25	16

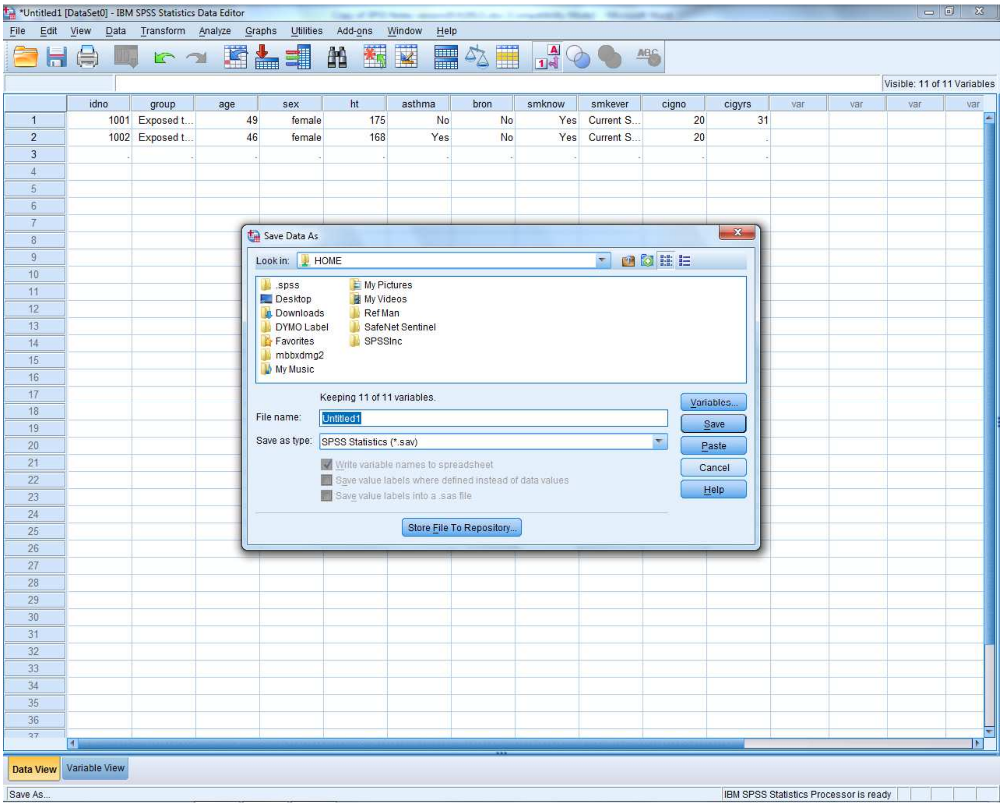
FILE MANAGEMENT

Saving an SPSS for Windows 7 File

Once you have entered some data you should save the file. It is good practice to save data at regular intervals during data entry just in case.

To save the data you have just entered, click the **File** at the top left corner of the screen and then the **Save As...** sub-option.

Something similar to the following screen will appear:


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
Visible: 11 of 11 Variables
Data View
Variable View
Save As...
IBM SPSS Statistics Processor is ready

	idno	group	age	sex	ht	asthma	bron	smknow	smkever	cigno	cigyrs	var	var	var	var
1	1001	Exposed t...	49	female	175	No	No	Yes	Current S...	20	31				
2	1002	Exposed t...	46	female	168	Yes	No	Yes	Current S...	20	.				
3	.	.	.	.	.	.	.	.	.	.	.				

Save Data As

Look in: HOME

.spss
Desktop
Downloads
DYMO Label
Favorites
mbbxdmg2
My Music
My Pictures
My Videos
Ref Man
SafeNet Sentinel
SPSSInc

Keeping 11 of 11 variables.

File name: Untitled1

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

Save a copy of the current **SPSS for Windows 7** file on your P: Drive or your pen drive, under **Drives:** click on ⯆ in the **Look in** window to generate a list of the drives.

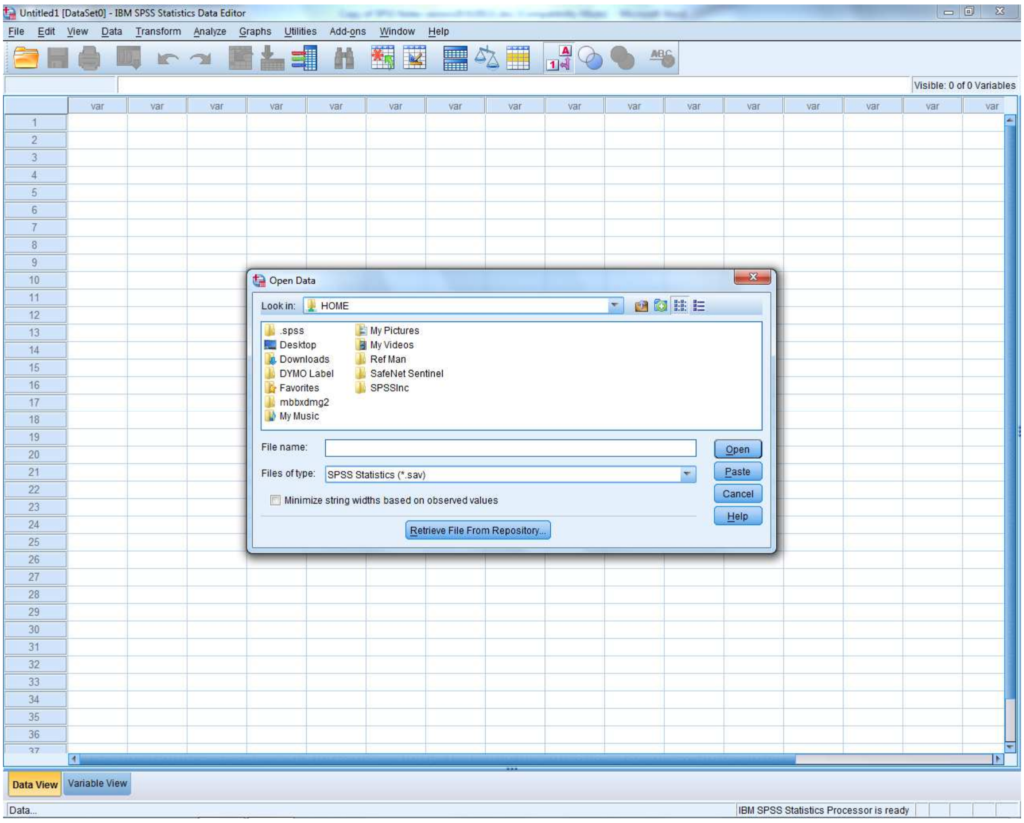
Click on the up/down-arrows to move to the **relevant pen drive** and enter a suitable name in the **File name** window. By default SPSS will add the file extension **.sav** in order to help identify the file as a SPSS data file. Finally, click on the **Save** button.

Backing Up Your Data

It is good practice to save data on different disks and also several names as data entry progresses (e.g. **mydata1 mydata2** etc). To make a backup copy of your data repeat the **Save Data As** procedure.

Retrieving Data Files

Retrieving an SPSS for Windows 7 File is essentially the reverse of the save process. Click on the **File** option, then the **Open** sub-option followed by the **Data** option. Something similar to the following screen will appear. Then retrieve the required file from the saved location.



Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

Open Data

Look in: HOME

Files of type: SPSS Statistics (*.sav)

□ Minimize string widths based on observed values

We can also open a data file when we as start an SPSS session (see above).

DESCRIPTIVE STATISTICS

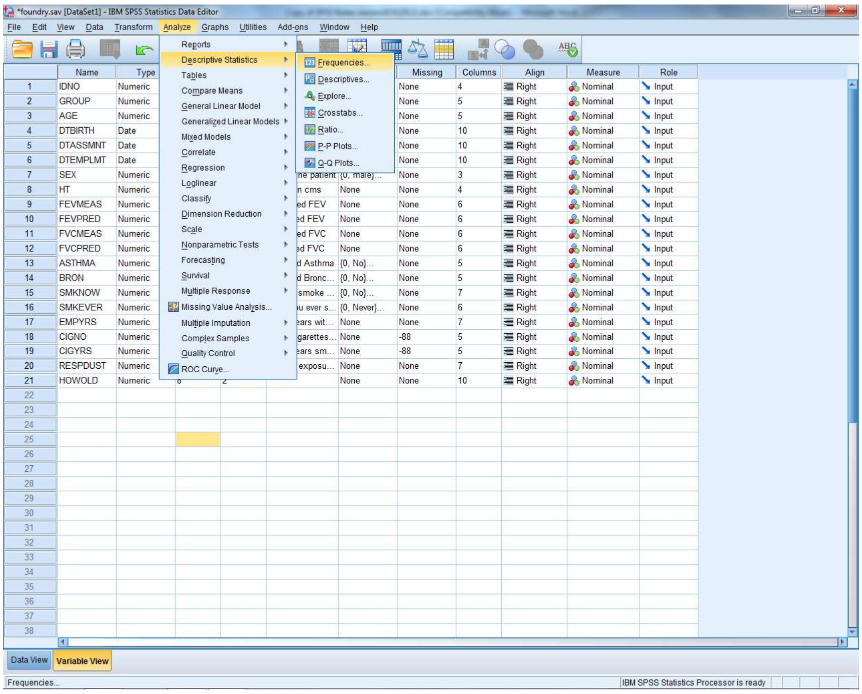
For the next stage you need to retrieve the data file **foundry.sav** which contains the fully labelled dataset you saved earlier to your desktop (see page 2). The open your data in SPSS as you would in any other package click **File, Open, Data** and retrieve your data from your workspace.

The first step in data analysis is to generate descriptive statistics. This will give us a feel for the data. It will also help identify any inconsistencies that may be in the data. This is sometimes called data cleaning. Techniques that are commonly used to do this include:

- Frequency Analyses
- Descriptive Statistics
- Cross-tabulations
- Plots

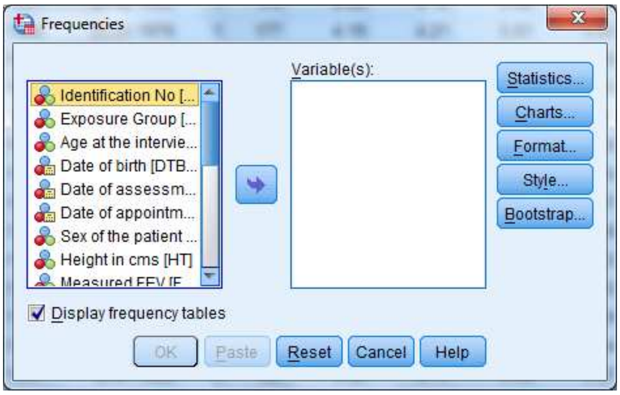
Frequency Tables

Carrying out a frequencies analysis on variables is the first step when checking for data errors, click on **Analyze** and choose the **Descriptive Statistics** option and then choose **Frequencies.** Move the variables of interest into the **Variables** box on the right-hand side, and then click **Statistics** to select some summary statistics such as range, maximum, minimum, mean and median, which will help you look for errors.



	Name	Type					Missing	Columns	Align	Measure	Role
1	IDNO	Numeric					None	4	Right	Nominal	Input
2	GROUP	Numeric					None	5	Right	Nominal	Input
3	AGE	Numeric					None	5	Right	Nominal	Input
4	DTBIRTH	Date					None	10	Right	Nominal	Input
5	DTASSMNT	Date					None	10	Right	Nominal	Input
6	DTEMPLMT	Date					None	10	Right	Nominal	Input
7	SEX	Numeric			he patient	{0, male}...	None	3	Right	Nominal	Input
8	HT	Numeric			n cms	None	None	4	Right	Nominal	Input
9	FEVMEAS	Numeric			ed FEV	None	None	6	Right	Nominal	Input
10	FEVPRED	Numeric			ed FEV	None	None	6	Right	Nominal	Input
11	FVCMEAS	Numeric			ed FVC	None	None	6	Right	Nominal	Input
12	FVCPRED	Numeric			ed FVC	None	None	6	Right	Nominal	Input
13	ASTHMA	Numeric			d Asthma	{0, No}...	None	5	Right	Nominal	Input
14	BRON	Numeric			d Bronc...	{0, No}...	None	5	Right	Nominal	Input
15	SMKNOW	Numeric			smoke ...	{0, No}...	None	7	Right	Nominal	Input
16	SMKEVER	Numeric			u ever s...	{0, Never}...	None	6	Right	Nominal	Input
17	EMPYRS	Numeric			ears wit...	None	None	7	Right	Nominal	Input
18	CIGNO	Numeric			garettes...	None	-88	5	Right	Nominal	Input
19	CIGYRS	Numeric			ears sm...	None	-88	5	Right	Nominal	Input
20	RESPDUST	Numeric			exposu...	None	None	7	Right	Nominal	Input
21	HOWOLD	Numeric	8	2		None	None	10	Right	Nominal	Input

The following screen will appear.



☑ Display frequency tables

To select the variable to perform a frequency table for example the Exposure group variable, click on its name in the left hand list and then press ►. Finally click on **OK** and the following output is then generated in the output window.

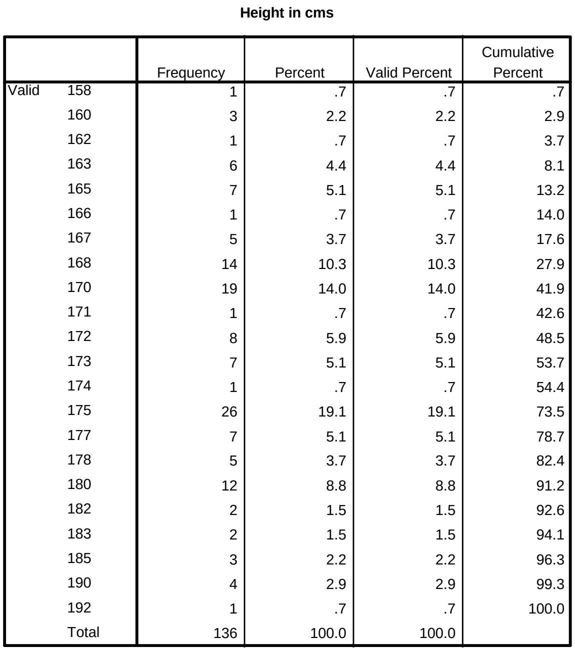
Exposure Group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Unexposed	63	46.3	46.3	46.3
	Exposure to Dust	73	53.7	53.7	100.0
	Total	136	100.0	100.0	

To return to the data editor click on **Window** and take the data editor option from the list. With the frequency table you can have a list of summary statistics as well. Click **Analyze, Descriptive Statistics, and Frequencies**. Press reset and then bring the variable (say, **ht**) to the **Variable(s)** window, click on **Statistics** option and select some summary statistics. Click **Continue** and **OK** button.

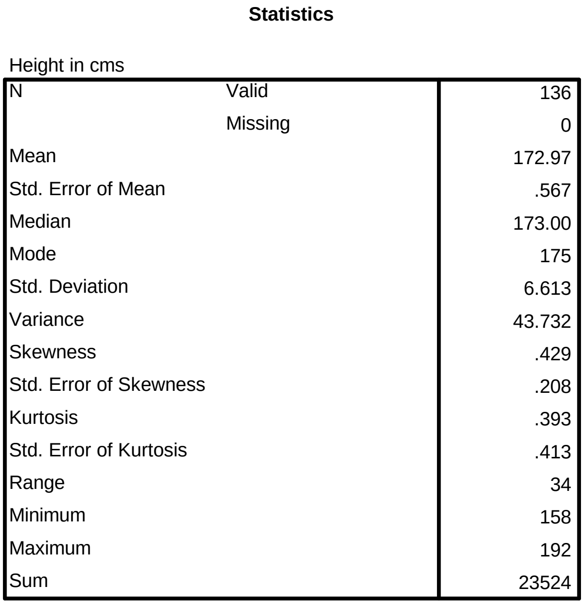
Once the **OK** button is pressed the results are automatically produced in an **Output window**, if the screen does not appear then the Output window may already exist but is located in the background. All results including the can be copied into word processing documents by clicking on the table and performing a standard copy and paste procedure.

Output from Frequencies with some summary statistics



Height in cms

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	158	1	.7	.7	.7
	160	3	2.2	2.2	2.9
	162	1	.7	.7	3.7
	163	6	4.4	4.4	8.1
	165	7	5.1	5.1	13.2
	166	1	.7	.7	14.0
	167	5	3.7	3.7	17.6
	168	14	10.3	10.3	27.9
	170	19	14.0	14.0	41.9
	171	1	.7	.7	42.6
	172	8	5.9	5.9	48.5
	173	7	5.1	5.1	53.7
	174	1	.7	.7	54.4
	175	26	19.1	19.1	73.5
	177	7	5.1	5.1	78.7
	178	5	3.7	3.7	82.4
	180	12	8.8	8.8	91.2
	182	2	1.5	1.5	92.6
	183	2	1.5	1.5	94.1
	185	3	2.2	2.2	96.3
	190	4	2.9	2.9	99.3
	192	1	.7	.7	100.0
	Total	136	100.0	100.0	



Statistics

Height in cms

N	Valid	136
	Missing	0
Mean		172.97
Std. Error of Mean		.567
Median		173.00
Mode		175
Std. Deviation		6.613
Variance		43.732
Skewness		.429
Std. Error of Skewness		.208
Kurtosis		.393
Std. Error of Kurtosis		.413
Range		34
Minimum		158
Maximum		192
Sum		23524

Exercise Using the frequencies options find out

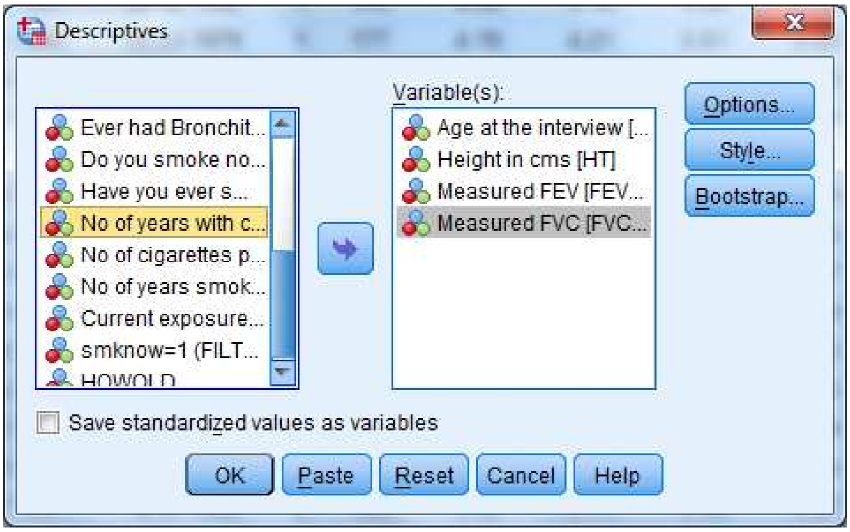
- what proportion of the foundry workers were exposed to dust?
- what proportions had ever suffered from bronchitis?
- what proportion had ever smoked?
- what proportion smoked more than 40 cigarettes per day?

Video Tutorial – Frequency Tables & Descriptive Statistics

https://www.youtube.com/watch?v=XrfQfEwjZA4

Descriptives

The descriptives command in SPSS is useful for summarizing quantitative data. To use this click on the **Analyse** tile choose the **Descriptive Statistics** option and then choose **descriptives.** Move the variables of interest into the **Variables** box on the right-hand side. As with the frequencies command we can obtain descriptive statistics for several variables at once. In the panel below we have chosen some of the quantitative variables in the foundry data set.



__**Exercise**__ Use the **descriptive** procedure to determine

- the current mean exposure to dust per day
- the mean number of cigarettes smoked per day

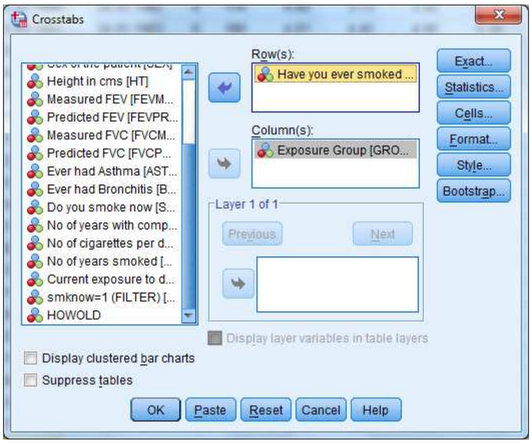
For mean number of cigarettes per day you may get a negative answer. Check the missing value codes and redo.

Cross-tabulation

To examine the relationship between two categorical variables, a two way Frequency Table can be used. This is called a cross-tabulation. Click on **Analyze** then **Descriptive Statistics** and then **Crosstabs.** The screen below appears. Suppose we wished to examine how smoking status related to exposure. We could examine this by a cross-tabulation of the variables **group** and **smkever**.

Select the smoking status variable **smkever** labelled **Have you ever smoked** in the source list then click ▶ by the **Row(s)** box to make this the row variable

Select **group** labelled **Exposure Group** in the source list and click ▶ by the **Column's** box to select the column variable. Finally press **OK**



Row(s): Have you ever smoked ...

Column(s): Exposure Group [GRO...

The following result appears when the two frequency table has been completed.

Have you ever smoked * Exposure Group Crosstabulation

Count

		Exposure Group		
		Unexposed	Exposure to Dust	Total
Have you ever smoked	Never	24	20	44
	Ex Smoker	19	19	38
	Curr. Smoker	20	34	54
Total		63	73	136

Two way frequency tables are more informative if they include percentages. To add percentages to the table select **Cells** from the Crosstabs screen. On pressing **Cells,** the following screen appears. Column, row, or total percentages can be selected by clicking the appropriate box. Whilst it is tempting to click all three this will make the output confusing. For the table above column percentages are the most useful as they will allow us to compare the smoking status of non-exposed and exposed subjects. By clicking column we get the resulting table.

Have you ever smoked * Exposure Group Crosstabulation

			Exposure Group		
			Unexposed	Exposure to Dust	Total
Have you ever smoked	Never	Count	24	20	44
		% within Exposure Group	38.1%	27.4%	32.4%
	Ex Smoker	Count	19	19	38
		% within Exposure Group	30.2%	26.0%	27.9%
	Curr. Smoker	Count	20	34	54
		% within Exposure Group	31.7%	46.6%	39.7%
Total		Count	63	73	136
		% within Exposure Group	100.0%	100.0%	100.0%

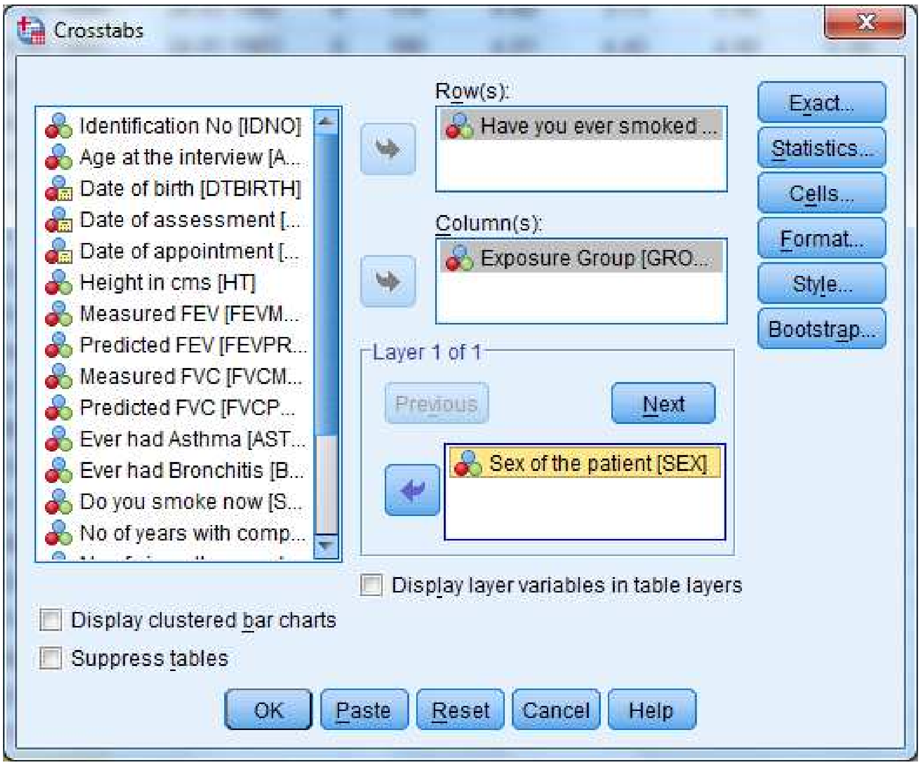
Video Tutorial – Two-way crosstabulation with percentages

Simple Cross tabulation - https://www.youtube.com/watch?v=ZOGwysV9ZQY

Adding percentages - https://www.youtube.com/watch?v=ByluYl5LncQ

Three-way tables

You may need to do comparisons on three variables. To do this, choose **Analyze** then **Descriptive Statistics** and then **Crosstabs.** Then the following screen appears. To create a three dimensional table instead of a two dimensional table, click on a variable and move using ▶ to layer 1 of 1 box.


Row(s): Have you ever smoked ...
Column(s): Exposure Group [GRO...
Layer 1 of 1: Sex of the patient [SEX]
□ Display layer variables in table layers
□ Display clustered bar charts
□ Suppress tables

If we add the variable sex we will now get separate tables for men and women giving the following output.

Have you ever smoked * Exposure Group * Sex of the patient Crosstabulation

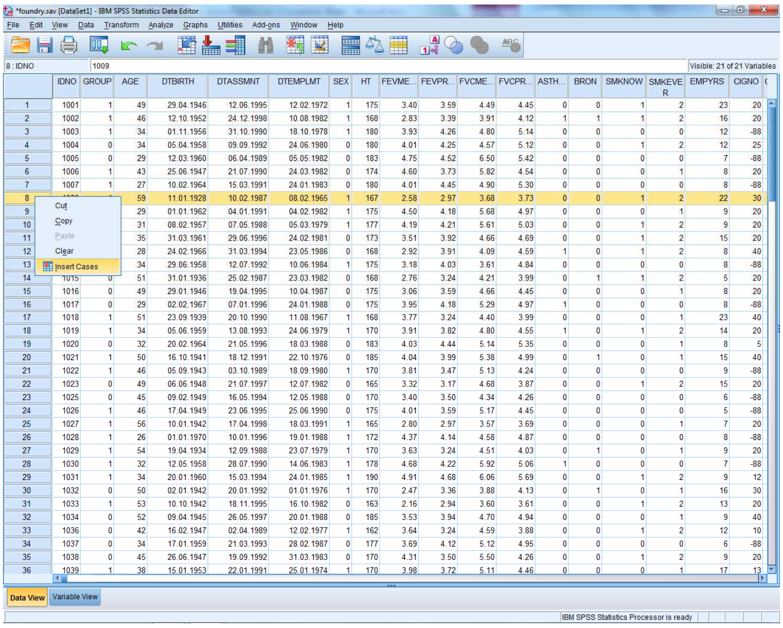
Sex of the patient				Exposure Group		Total
				Unexposed	Exposure to Dust	Unexposed
male	Have you ever smoked	Never	Count	14	6	20
			% within Exposure Group	42.4%	20.0%	31.7%
		Ex Smoker	Count	7	7	14
			% within Exposure Group	21.2%	23.3%	22.2%
		Curr. Smoker	Count	12	17	29
			% within Exposure Group	36.4%	56.7%	46.0%
	Total		Count	33	30	63
			% within Exposure Group	100.0%	100.0%	100.0%
female	Have you ever smoked	Never	Count	10	14	24
			% within Exposure Group	33.3%	32.6%	32.9%
		Ex Smoker	Count	12	12	24
			% within Exposure Group	40.0%	27.9%	32.9%
		Curr. Smoker	Count	8	17	25
			% within Exposure Group	26.7%	39.5%	34.2%
	Total		Count	30	43	73
			% within Exposure Group	100.0%	100.0%	100.0%

EDITING AND MODIFYING THE DATA

Having done some preliminary analysis we may need to change the data. There are some useful functions for modifying data files.

Inserting Data

You may have noticed that idno 1008 was missing. **To insert** it, either click **Edit** then **Insert Case** or right click on the sidebar (immediately before IDNO 1009) and click **Insert Case** and a new blank row is added as shown below.


*foundry.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
8 : IDNO
1009
Visible: 21 of 21 Variables
Cut
Copy
Paste
Clear
Insert Cases
Data View
Variable View
IBM SPSS Statistics Processor is ready

	IDNO	GROUP	AGE	DTBIRTH	DTASSMNT	DTEMPLMT	SEX	HT	FEVME...	FEVPR...	FVCME...	FVCPR...	ASTH...	BRON	SMKNOW	SMKEVER	EMPYRS	CIGNO
1	1001	1	49	29.04.1946	12.06.1995	12.02.1972	1	175	3.40	3.59	4.49	4.45	0	0	1	2	23	20
2	1002	1	46	12.10.1952	24.12.1998	10.08.1982	1	168	2.83	3.39	3.91	4.12	1	1	1	2	16	20
3	1003	1	34	01.11.1956	31.10.1990	18.10.1978	1	180	3.93	4.26	4.80	5.14	0	0	0	0	12	-88
4	1004	0	34	05.04.1958	09.09.1992	24.06.1980	0	180	4.01	4.25	4.57	5.12	0	0	1	2	12	25
5	1005	0	29	12.03.1960	06.04.1989	05.05.1982	0	183	4.75	4.52	6.50	5.42	0	0	0	0	7	-88
6	1006	1	43	25.06.1947	21.07.1990	24.03.1982	0	174	4.60	3.73	5.82	4.54	0	0	0	1	8	20
7	1007	1	27	10.02.1964	15.03.1991	24.01.1983	0	180	4.01	4.45	4.90	5.30	0	0	0	0	8	-88
8			59	11.01.1928	10.02.1987	08.02.1965	1	167	2.58	2.97	3.68	3.73	0	0	1	2	22	30
9			29	01.01.1962	04.01.1991	04.02.1982	1	175	4.50	4.18	5.68	4.97	0	0	0	1	9	20
10			31	08.02.1957	07.05.1988	05.03.1979	1	177	4.19	4.21	5.61	5.03	0	0	1	2	9	20
11			35	31.03.1961	29.06.1996	24.02.1981	0	173	3.51	3.92	4.66	4.69	0	0	1	2	15	20
12			28	24.02.1966	31.03.1994	23.05.1986	0	168	2.92	3.91	4.09	4.59	1	0	1	2	8	40
13			34	29.06.1958	12.07.1992	10.06.1984	1	175	3.18	4.03	3.61	4.84	0	0	0	0	8	-88
14	1015	0	51	31.01.1936	25.02.1987	23.03.1982	0	168	2.76	3.24	4.21	3.99	0	1	1	2	5	20
15	1016	0	49	29.01.1946	19.04.1995	10.04.1987	0	175	3.06	3.59	4.66	4.45	0	0	0	1	8	20
16	1017	0	29	02.02.1967	07.01.1996	24.01.1988	0	175	3.95	4.18	5.29	4.97	1	0	0	0	8	-88
17	1018	1	51	23.09.1939	20.10.1990	11.08.1967	1	168	3.77	3.24	4.40	3.99	0	0	0	1	23	40
18	1019	1	34	05.06.1959	13.08.1993	24.06.1979	1	170	3.91	3.82	4.80	4.55	1	0	1	2	14	20
19	1020	0	32	20.02.1964	21.05.1996	18.03.1988	0	183	4.03	4.44	5.14	5.35	0	0	0	1	8	5
20	1021	1	50	16.10.1941	18.12.1991	22.10.1976	0	185	4.04	3.99	5.38	4.99	0	1	0	1	15	40
21	1022	1	46	05.09.1943	03.10.1989	18.09.1980	1	170	3.81	3.47	5.13	4.24	0	0	0	0	9	-88
22	1023	0	49	06.06.1948	21.07.1997	12.07.1982	0	165	3.32	3.17	4.68	3.87	0	0	1	2	15	20
23	1025	0	45	09.02.1949	16.05.1994	12.05.1988	0	170	3.40	3.50	4.34	4.26	0	0	0	0	6	-88
24	1026	1	46	17.04.1949	23.06.1995	25.06.1990	0	175	4.01	3.59	5.17	4.45	0	0	0	0	5	-88
25	1027	1	56	10.01.1942	17.04.1998	18.03.1991	1	165	2.80	2.97	3.57	3.69	0	0	0	1	7	20
26	1028	1	26	01.01.1970	10.01.1996	19.01.1988	1	172	4.37	4.14	4.58	4.87	0	0	0	0	8	-88
27	1029	1	54	19.04.1934	12.09.1988	23.07.1979	1	170	3.63	3.24	4.51	4.03	0	1	0	1	9	20
28	1030	1	32	12.05.1958	28.07.1990	14.06.1983	1	178	4.68	4.22	5.92	5.06	1	0	0	0	7	-88
29	1031	1	34	20.01.1960	15.03.1994	24.01.1985	1	190	4.91	4.68	6.06	5.69	0	0	1	2	9	12
30	1032	0	50	02.01.1942	20.01.1992	01.01.1976	1	170	2.47	3.36	3.88	4.13	0	1	0	1	16	30
31	1033	1	53	10.10.1942	18.11.1995	16.10.1982	0	163	2.16	2.94	3.60	3.61	0	0	1	2	13	20
32	1034	0	52	09.04.1945	26.05.1997	20.01.1988	0	185	3.53	3.94	4.70	4.94	0	0	0	1	9	40
33	1036	0	42	16.02.1947	02.04.1989	12.02.1977	1	162	3.64	3.24	4.59	3.88	0	0	1	2	12	10
34	1037	0	34	17.01.1959	21.03.1993	28.02.1987	0	177	3.69	4.12	5.12	4.95	0	0	0	0	6	-88
35	1038	0	45	26.06.1947	19.09.1992	31.03.1983	0	170	4.31	3.50	5.50	4.26	0	0	1	2	9	20
36	1039	1	38	15.01.1953	22.01.1991	25.01.1974	1	170	3.98	3.72	5.11	4.46	0	0	0	1	17	13

You can insert the following case (idno 1008) in the blank line

Variable	Value	Variable	Value
Idno	1008	Asthma	0
Group	1	Bron	0
Sex	1	Smknow	1
Ht	180	Smkever	2
Fevmeas	4.01	Cigno	30
Fevpred	4.45	Cigsyrs	20
Fvcmeas	4.90	Empyrs	10
Fvcpred	5.30	Respdust	2.04

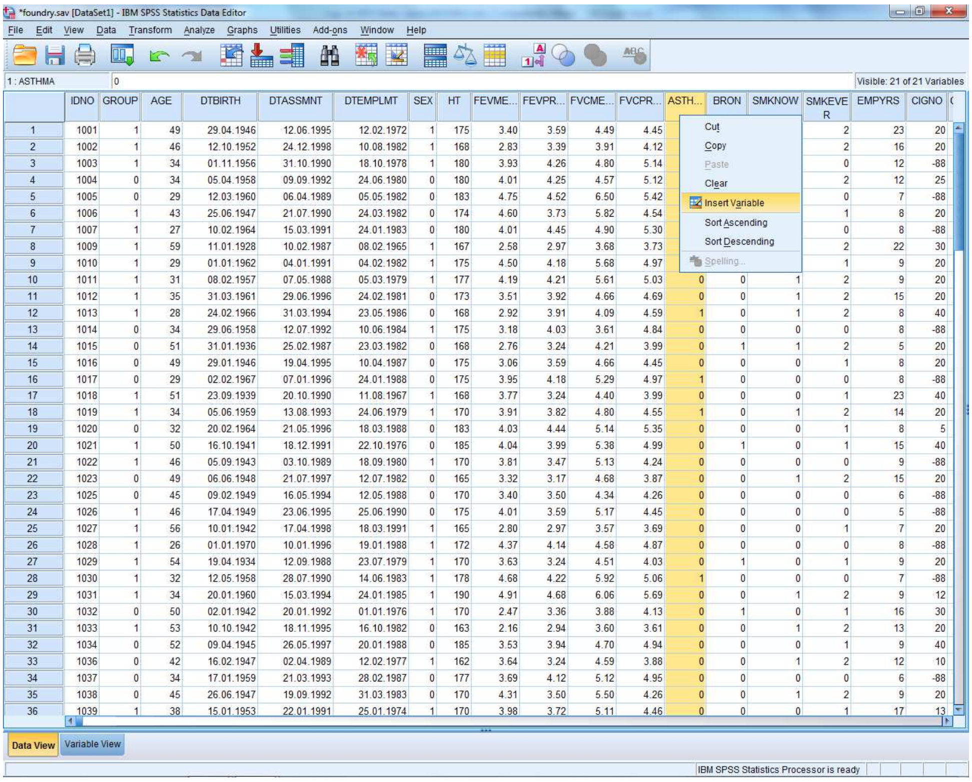
Deleting A Case

To delete a case, right click on the row number on the far left of the Data Editor to highlight the row containing the case. Press the **Clear** button (alternatively, click on the **Edit** option on the menu bar then click on the **Clear** option) and the case is deleted and the cases below move up to fill the gap.

<u>Exercise</u> Delete case no 1008

Inserting A Variable

To insert a variable into the middle of the data, click on the variable after the position at which you wish the variable to appear and then click on **Data** then **Insert Variable**. A blank column is inserted before the selected variable shown here.


*foundry.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Cut
Copy
Paste
Clear
Insert Variable
Sort Ascending
Sort Descending
Spelling...
Data View
Variable View

1 : ASTHMA 0

Visible: 21 of 21 Variables

	IDNO	GROUP	AGE	DTBIRTH	DTASSMNT	DTEMPLMT	SEX	HT	FEVME...	FEVPR...	FVCME...	FVCPR...	ASTH...	BRON	SMKNOW	SMKEVER	EMPYRS	CIGNO
1	1001	1	49	29.04.1946	12.06.1995	12.02.1972	1	175	3.40	3.59	4.49	4.45				2	23	20
2	1002	1	46	12.10.1952	24.12.1998	10.08.1982	1	168	2.83	3.39	3.91	4.12				2	16	20
3	1003	1	34	01.11.1956	31.10.1990	18.10.1978	1	180	3.93	4.26	4.80	5.14				0	12	-88
4	1004	0	34	05.04.1958	09.09.1992	24.06.1980	0	180	4.01	4.25	4.57	5.12				2	12	25
5	1005	0	29	12.03.1960	06.04.1989	05.05.1982	0	183	4.75	4.52	6.50	5.42				0	7	-88
6	1006	1	43	25.06.1947	21.07.1990	24.03.1982	0	174	4.60	3.73	5.82	4.54				1	8	20
7	1007	1	27	10.02.1964	15.03.1991	24.01.1983	0	180	4.01	4.45	4.90	5.30				0	8	-88
8	1009	1	59	11.01.1928	10.02.1987	08.02.1965	1	167	2.58	2.97	3.68	3.73				2	22	30
9	1010	1	29	01.01.1962	04.01.1991	04.02.1982	1	175	4.50	4.18	5.68	4.97				1	9	20
10	1011	1	31	08.02.1957	07.05.1988	05.03.1979	1	177	4.19	4.21	5.61	5.03	0	0	1	2	9	20
11	1012	1	35	31.03.1961	29.06.1996	24.02.1981	0	173	3.51	3.92	4.66	4.69	0	0	1	2	15	20
12	1013	1	28	24.02.1966	31.03.1994	23.05.1986	0	168	2.92	3.91	4.09	4.59	1	0	1	2	8	40
13	1014	0	34	29.06.1958	12.07.1992	10.06.1984	1	175	3.18	4.03	3.61	4.84	0	0	0	0	8	-88
14	1015	0	51	31.01.1936	25.02.1987	23.03.1982	0	168	2.76	3.24	4.21	3.99	0	1	1	2	5	20
15	1016	0	49	29.01.1946	19.04.1995	10.04.1987	0	175	3.06	3.59	4.66	4.45	0	0	0	1	8	20
16	1017	0	29	02.02.1967	07.01.1996	24.01.1988	0	175	3.95	4.18	5.29	4.97	1	0	0	0	8	-88
17	1018	1	51	23.09.1939	20.10.1990	11.08.1967	1	168	3.77	3.24	4.40	3.99	0	0	0	1	23	40
18	1019	1	34	05.06.1959	13.08.1993	24.06.1979	1	170	3.91	3.82	4.80	4.55	1	0	1	2	14	20
19	1020	0	32	20.02.1964	21.05.1996	18.03.1988	0	183	4.03	4.44	5.14	5.35	0	0	0	1	8	5
20	1021	1	50	16.10.1941	18.12.1991	22.10.1976	0	185	4.04	3.99	5.38	4.99	0	1	0	1	15	40
21	1022	1	46	05.09.1943	03.10.1989	18.09.1980	1	170	3.81	3.47	5.13	4.24	0	0	0	0	9	-88
22	1023	0	49	06.06.1948	21.07.1997	12.07.1982	0	165	3.32	3.17	4.68	3.87	0	0	1	2	15	20
23	1025	0	45	09.02.1949	16.05.1994	12.05.1988	0	170	3.40	3.50	4.34	4.26	0	0	0	0	6	-88
24	1026	1	46	17.04.1949	23.06.1995	25.06.1990	0	175	4.01	3.59	5.17	4.45	0	0	0	0	5	-88
25	1027	1	56	10.01.1942	17.04.1998	18.03.1991	1	165	2.80	2.97	3.57	3.69	0	0	0	1	7	20
26	1028	1	26	01.01.1970	10.01.1996	19.01.1988	1	172	4.37	4.14	4.58	4.87	0	0	0	0	8	-88
27	1029	1	54	19.04.1934	12.09.1988	23.07.1979	1	170	3.63	3.24	4.51	4.03	0	1	0	1	9	20
28	1030	1	32	12.05.1958	28.07.1990	14.06.1983	1	178	4.68	4.22	5.92	5.06	1	0	0	0	7	-88
29	1031	1	34	20.01.1960	15.03.1994	24.01.1985	1	190	4.91	4.68	6.06	5.69	0	0	1	2	9	12
30	1032	0	50	02.01.1942	20.01.1992	01.01.1976	1	170	2.47	3.36	3.88	4.13	0	1	0	1	16	30
31	1033	1	53	10.10.1942	18.11.1995	16.10.1982	0	163	2.16	2.94	3.60	3.61	0	0	1	2	13	20
32	1034	0	52	09.04.1945	26.05.1997	20.01.1988	0	185	3.53	3.94	4.70	4.94	0	0	0	1	9	40
33	1036	0	42	16.02.1947	02.04.1989	12.02.1977	1	162	3.64	3.24	4.59	3.88	0	0	1	2	12	10
34	1037	0	34	17.01.1959	21.03.1993	28.02.1987	0	177	3.69	4.12	5.12	4.95	0	0	0	0	6	-88
35	1038	0	45	26.06.1947	19.09.1992	31.03.1983	0	170	4.31	3.50	5.50	4.26	0	0	1	2	9	20
36	1039	1	38	15.01.1953	22.01.1991	25.01.1974	1	170	3.98	3.72	5.11	4.46	0	0	0	1	17	13

IBM SPSS Statistics Processor is ready

Deleting A Variable

To delete a variable, click on its column name at the top of the Data Editor to highlight the column containing the variable. Then press the **Delete** button. The variable is deleted and the variables to the right move to the left to fill the gap. Now delete the variable you just created.

Moving A Variable

Insert a blank variable as mentioned above in the required position. Click on the name of the variable to be moved (This highlights the column), **Edit** and **Cut.** Click on the name of the blank variable and **Edit** then **Paste**.

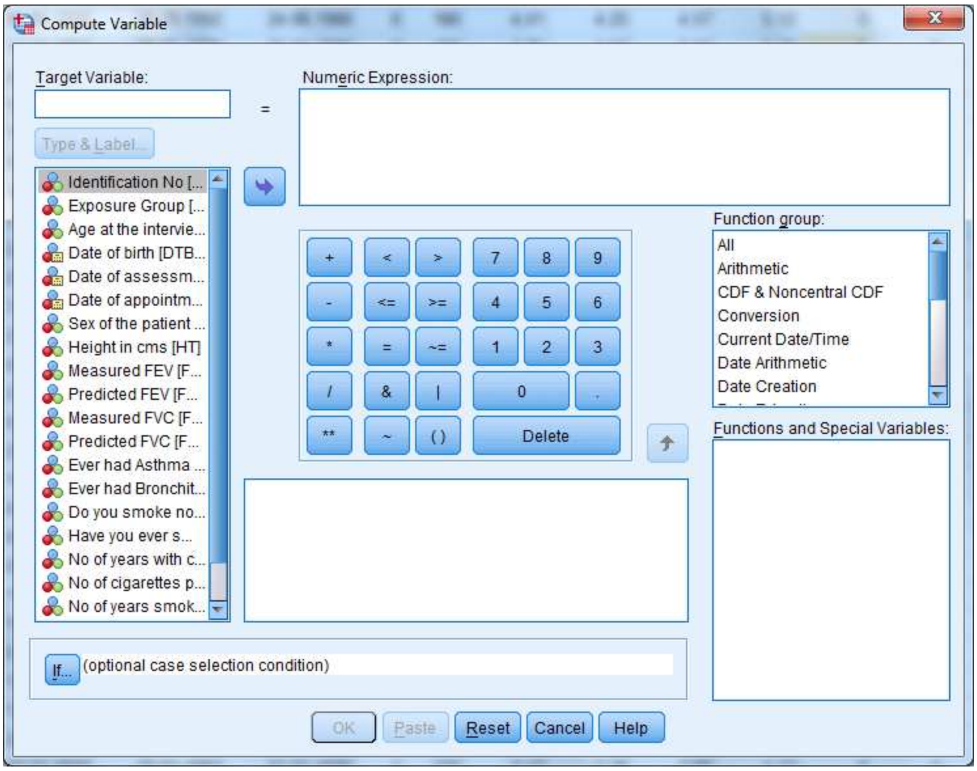
PART II

CONSTRUCTING NEW VARIABLES

Sometimes we need to compute new variables from the data entered. For example in the foundry data set we might want to compute the ratio of the measured to predicted fev. Alternatively, we might want to group ages into bands. SPSS has procedures to construct a new variable from existing variables.

Computing a New Variable

For the foundry worker data we shall compute the variable **fevratio** defined as **fevmeas/fevpred.** Click **Transform** then **Compute** and the following screen appears:-



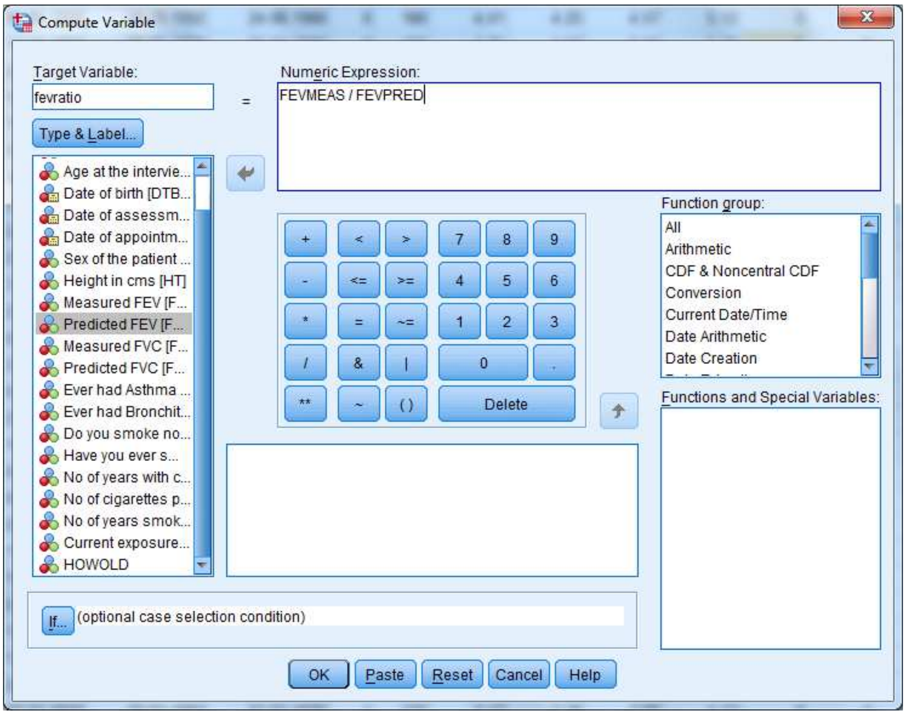
Enter the name **fevratio** in **Target variable** window. If the variable is new, click on **Type & Label** to define the type and variable label. To build up mathematical expression which will create the new variable you can choose variables from the left hand box then click ▶ to move them to the **numeric expression** window. You can choose any of the keys on the calculator pad in the centre or any of the functions from the built-in functions box followed by.▲

Select the function using up ▲ and down ▼ arrow key from the Built in function window and then click on the button ▲ .The expression will appear in the **Numeric Expression** window

These are the functions on the calculator pad are defined as follows.

Operator	Mnemonic form	Description	Operator	Mnemonic form	Description
+		Addition	>=	GE	Greater Than Or Equal To
-		Subtraction	=	EQ	Equals
*		Multiplication	~=	NE	Not Equals
/		Division	&	AND	Logical And
**		Power Of	|	OR	Logical Or
<	LT	Less Than	( )		Parentheses
>	GT	Greater Than	~	NOT	Logical Not
<=	LE	Less Than Or Equal To			

To compute **fevratio** we move **fevmeas** and **fevpred** into the **numeric expression** window. You can also type a formulae into the numeric expression window. This is illustrated below.



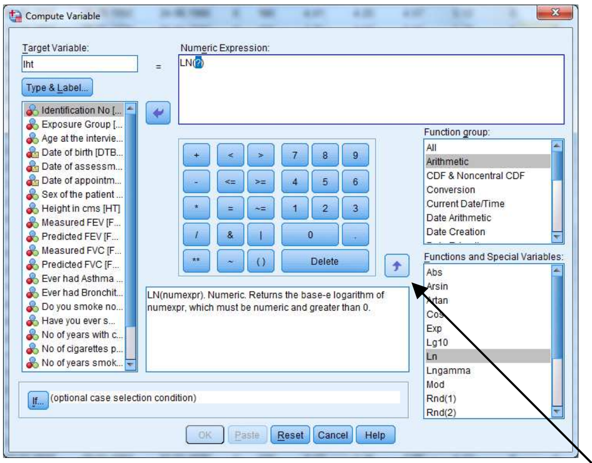
Target Variable: fevratio

Numeric Expression: FEVMEAS / FEVPRED

Once the expression is complete press **OK.**

Computing a New Variable by using built-in Functions

In the **Compute** procedure there is a built in functions window which can be used to create a new variable or to transform the values of an existing variable. Transformations such as the square root, or the logarithm, are easily made. Suppose you wish to do a log transformation of the variable called height (**ht**) from the **foundry** data set. First click **Transform** from menu bar and then choose **Compute** from drop down menu, then you get the compute window.



Target Variable: lht

Numeric Expression: LN(?)

LN(numexpr). Numeric. Returns the base-e logarithm of numexpr, which must be numeric and greater than 0.

Function group: Arithmetic

Functions and Special Variables: Ln

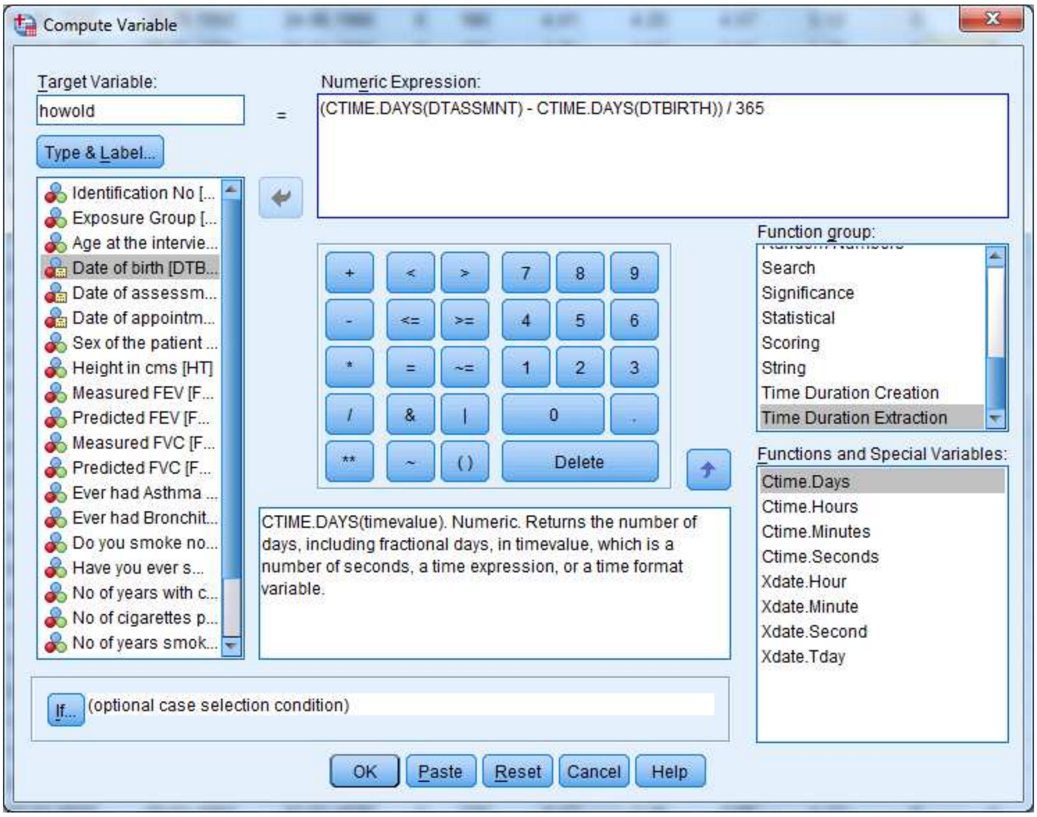
Type a name, say **lht**, in the target variable window. Click on the arrow on the right of the **Functions** box to scroll up and down through the functions. Select **Arithmetic** followed by **Ln** function in the **Functions and Special Variables** box for natural log and click on **Functions** : ▲ , this will put the function with a **?** in parentheses in the window named **Numeric Expression.** Then select the variable to replace **?** i.e. **ht** by clicking ▶ and then press **OK** button. Then a new variable **lht** will be created (located at the end of the variable list). Having carried out a transformation it is important to check the result. For example, taking a log of a negative value creates a missing value. Other commonly used transformation functions are **LG10, SQRT, ABS, TRUNC** etc.

Video Tutorial – Creating a new variable – log transformion (2min onwards)

https://www.youtube.com/watch?v=xZCOyQ92X9g

Computing Duration of Time Difference by built-in Functions

In the same data set there are some variables (date of birth, date of assessment etc) which are stored in date format. One is able to calculate the time difference (in days) by using the functions **Ctime.Days** The age of the patients on the date of assessment can be calculated from the date of birth and assessment date. As before click **Transform** from menu bar and then **Compute** from drop down menu, you then get the compute window. In the target variable window type a name say **howold**, then select the functions group **Time Duration Extraction** followed by **Ctime.Days** in the Functions and Special Variables window using the up and down arrow keys, click on **Functions** : ▲ , this will put the function with a **?** in parentheses in the box named **Numeric Expression.** Then select the variable to replace **?** i.e. date of assessment by clicking ▶ . Perform the same procedure for date of birth. You can then compute the difference **Time** (in days), then you have to divide the whole thing by 365 (number of days in quarterly leap year) to get **howold** in years. Below is the example.



Target Variable: howold

Numeric Expression: (CTIME.DAYS(DTASSMNT) - CTIME.DAYS(DTBIRTH)) / 365

CTIME.DAYS(timevalue). Numeric. Returns the number of days, including fractional days, in timevalue, which is a number of seconds, a time expression, or a time format variable.

Whenever you compute a new variable from existing data it is important to check that what you have created is sensible. You also need to check that missing values have not been converted into none missing values. Using the **Data view** tab check the value of **howold**.

Exercise

- Calculate the duration of the patients in the employment and compare with the values of employment **days** provided in the data set.
- Calculate the duration of the patients in the employment and compare with the values of employment **years** provided in the data set.

Recoding a value

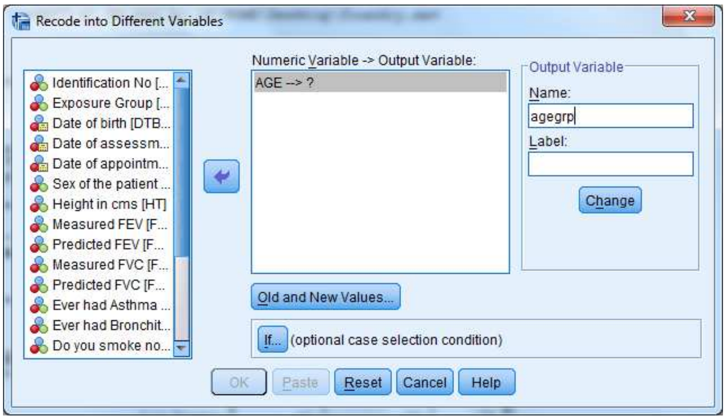
To assist in data analyses you often need to group a continuous variable (e.g. age) into categories

To do this select **Transform** then **Recode.** Two options are now given

- Into Same Variables
- Into Different Variables

The first option leads to potentially valuable information being overwritten. It is usually best to use the second option as it is then possible to check whether the recode has worked correctly by comparing the new and old version.

Having chosen the second option the following screen will appear. First choose an input variable from the list on the left hand side then press ▶ Then enter the name of the variable for the recoded data under Output Variable Name and press **Change.**



Recode into Different Variables

- Identification No [...
- Exposure Group [...
- Date of birth [DTB...
- Date of assessm...
- Date of appointm...
- Sex of the patient ...
- Height in cms [HT]
- Measured FEV [F...
- Predicted FEV [F...
- Measured FVC [F...
- Predicted FVC [F...
- Ever had Asthma ...
- Ever had Bronchit...
- Do you smoke no...

Numeric Variable -> Output Variable: AGE --> ?

Output Variable

Name: agegrp

Label:

Now press **Old and New Values** and the following screen appears.

Suppose we wish to recode age into bands <30, 30-39, 40-49, 50+

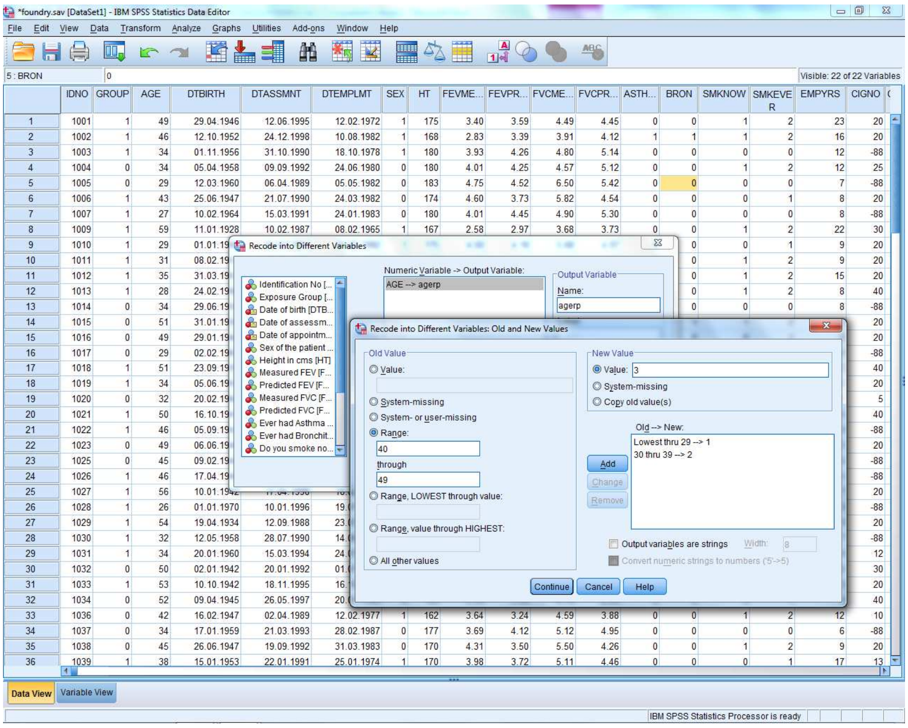
Click on **Range Lowest Through** and enter 29 into the box then click on value under **New value** and enter 1 and finally press **Add.**

Click on **Range** then enter 30 and 39. Then click on **New Value** and enter 2 and finally press **Add.**

Click on **Range** then enter 40 and 49. Then click on **New Value** and enter 3 and finally press **Add.**

Finally click on **Range Through highest** enter 50 then click on **New Value** and enter 4 and finally press **Add.**

Once you have specified all the **OLD -> New** recodes, click on **Continue** then **OK** on the **Recode into Different Variables screen**. The following shows an example of setting up a recoded value.



Numeric Variable -> Output Variable: AGE --> agerp

Output Variable Name: agerp

Old Value: Range: 40 through 49

New Value: Value: 3

Old --> New:
Lowest thru 29 --> 1
30 thru 39 --> 2

*foundry.sav [DataSet1] - IBM SPSS Statistics Data Editor
Recode into Different Variables
Recode into Different Variables: Old and New Values
Add
Change
Remove
Continue
Cancel
Help
Data View
Variable View
IBM SPSS Statistics Processor is ready

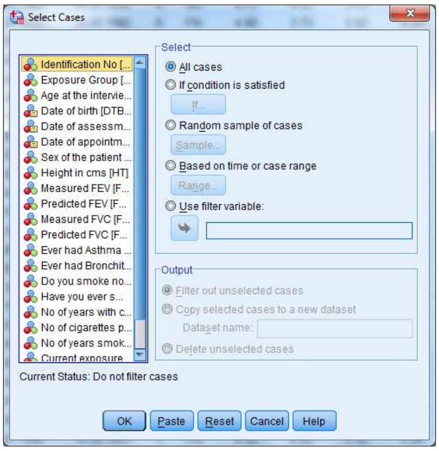
After recoding a variable it is usually advisable to run case summaries to compare the old and new values

<u>Video Tutorial – Recoding a variable</u>

https://www.youtube.com/watch?v=47GslKRT8Ck

Selecting a Subset of the Data

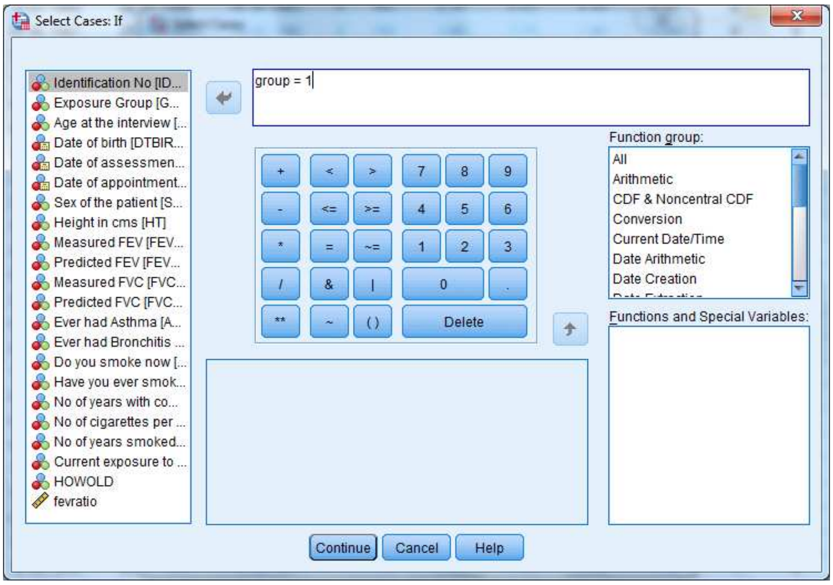
In addition to analysing the full set of data, you may want to analyse a subset. If, for example**,** you want to perform an analysis on exposed cases only, click on the **Data** option at the top of the **Data View** screen, then on the **Select Cases** option and the following screen will appear:



Select: ● All cases

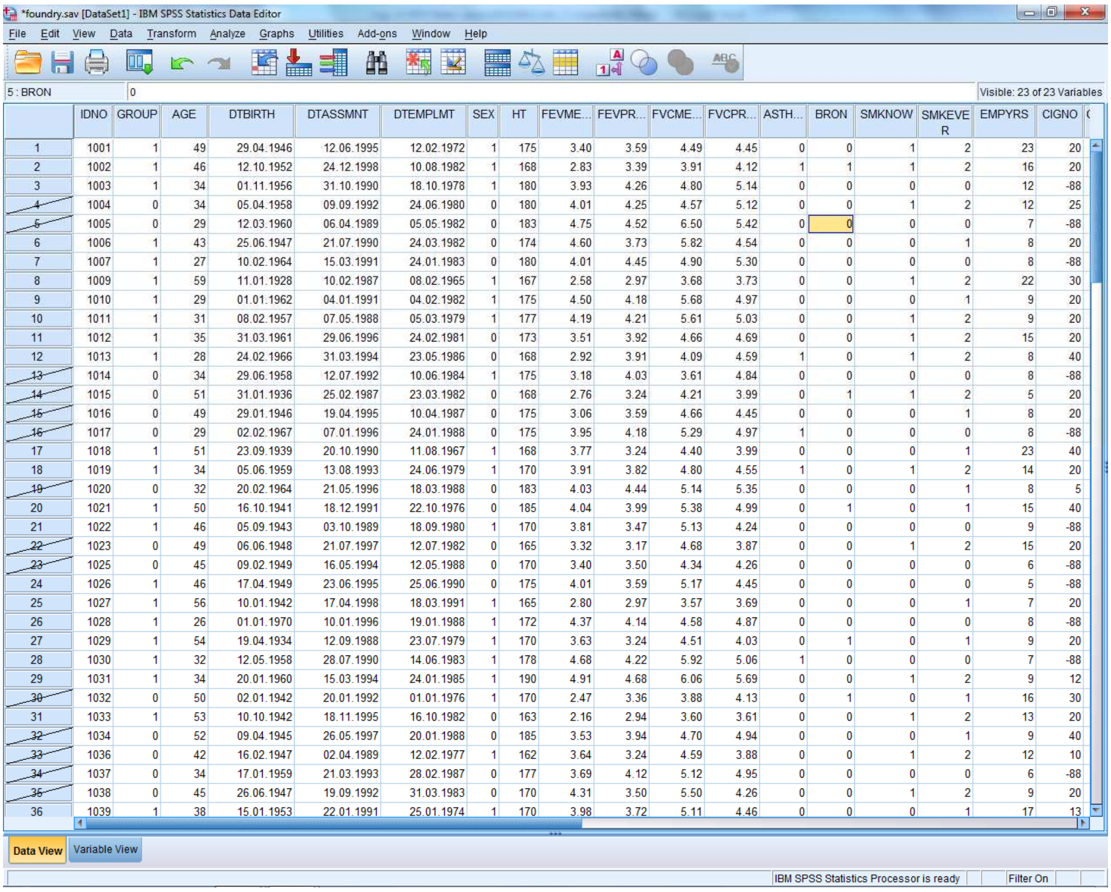
Current Status: Do not filter cases

To make the selection, click in the circle with the **If Condition is Satisfied** box, then click the **If...** button. The following panel will then appear. (group = 1 has been entered in the box provided to select the exposed cases),



group = 1

Click on the **Continue** tile at the bottom of the screen. Once you have returned to the main Select Cases screen, click on the OK button. The effect of the above filter on the data is shown below. Please note the / on the left hand side showing the records which have been excluded. To remove the filter click on **Data** then **Select Cases** and **Select all cases**.


*foundry.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
5 : BRON
0
Visible: 23 of 23 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Filter On

	IDNO	GROUP	AGE	DTBIRTH	DTASSMNT	DTEMPLMT	SEX	HT	FEVME...	FEVPR...	FVCME...	FVCPR...	ASTH...	BRON	SMKNOW	SMKEVER	EMPYRS	CIGNO
1	1001	1	49	29.04.1946	12.06.1995	12.02.1972	1	175	3.40	3.59	4.49	4.45	0	0	1	2	23	20
2	1002	1	46	12.10.1952	24.12.1998	10.08.1982	1	168	2.83	3.39	3.91	4.12	1	1	1	2	16	20
3	1003	1	34	01.11.1956	31.10.1990	18.10.1978	1	180	3.93	4.26	4.80	5.14	0	0	0	0	12	-88
~~4~~	1004	0	34	05.04.1958	09.09.1992	24.06.1980	0	180	4.01	4.25	4.57	5.12	0	0	1	2	12	25
~~5~~	1005	0	29	12.03.1960	06.04.1989	05.05.1982	0	183	4.75	4.52	6.50	5.42	0	0	0	0	7	-88
6	1006	1	43	25.06.1947	21.07.1990	24.03.1982	0	174	4.60	3.73	5.82	4.54	0	0	0	1	8	20
7	1007	1	27	10.02.1964	15.03.1991	24.01.1983	0	180	4.01	4.45	4.90	5.30	0	0	0	0	8	-88
8	1009	1	59	11.01.1928	10.02.1987	08.02.1965	1	167	2.58	2.97	3.68	3.73	0	0	1	2	22	30
9	1010	1	29	01.01.1962	04.01.1991	04.02.1982	1	175	4.50	4.18	5.68	4.97	0	0	0	1	9	20
10	1011	1	31	08.02.1957	07.05.1988	05.03.1979	1	177	4.19	4.21	5.61	5.03	0	0	1	2	9	20
11	1012	1	35	31.03.1961	29.06.1996	24.02.1981	0	173	3.51	3.92	4.66	4.69	0	0	1	2	15	20
12	1013	1	28	24.02.1966	31.03.1994	23.05.1986	0	168	2.92	3.91	4.09	4.59	1	0	1	2	8	40
~~13~~	1014	0	34	29.06.1958	12.07.1992	10.06.1984	1	175	3.18	4.03	3.61	4.84	0	0	0	0	8	-88
~~14~~	1015	0	51	31.01.1936	25.02.1987	23.03.1982	0	168	2.76	3.24	4.21	3.99	0	1	1	2	5	20
~~15~~	1016	0	49	29.01.1946	19.04.1995	10.04.1987	0	175	3.06	3.59	4.66	4.45	0	0	0	1	8	20
~~16~~	1017	0	29	02.02.1967	07.01.1996	24.01.1988	0	175	3.95	4.18	5.29	4.97	1	0	0	0	8	-88
17	1018	1	51	23.09.1939	20.10.1990	11.08.1967	1	168	3.77	3.24	4.40	3.99	0	0	0	1	23	40
18	1019	1	34	05.06.1959	13.08.1993	24.06.1979	1	170	3.91	3.82	4.80	4.55	1	0	1	2	14	20
~~19~~	1020	0	32	20.02.1964	21.05.1996	18.03.1988	0	183	4.03	4.44	5.14	5.35	0	0	0	1	8	5
20	1021	1	50	16.10.1941	18.12.1991	22.10.1976	0	185	4.04	3.99	5.38	4.99	0	1	0	1	15	40
21	1022	1	46	05.09.1943	03.10.1989	18.09.1980	1	170	3.81	3.47	5.13	4.24	0	0	0	0	9	-88
~~22~~	1023	0	49	06.06.1948	21.07.1997	12.07.1982	0	165	3.32	3.17	4.68	3.87	0	0	1	2	15	20
~~23~~	1025	0	45	09.02.1949	16.05.1994	12.05.1988	0	170	3.40	3.50	4.34	4.26	0	0	0	0	6	-88
24	1026	1	46	17.04.1949	23.06.1995	25.06.1990	0	175	4.01	3.59	5.17	4.45	0	0	0	0	5	-88
25	1027	1	56	10.01.1942	17.04.1998	18.03.1991	1	165	2.80	2.97	3.57	3.69	0	0	0	1	7	20
26	1028	1	26	01.01.1970	10.01.1996	19.01.1988	1	172	4.37	4.14	4.58	4.87	0	0	0	0	8	-88
27	1029	1	54	19.04.1934	12.09.1988	23.07.1979	1	170	3.63	3.24	4.51	4.03	0	1	0	1	9	20
28	1030	1	32	12.05.1958	28.07.1990	14.06.1983	1	178	4.68	4.22	5.92	5.06	1	0	0	0	7	-88
29	1031	1	34	20.01.1960	15.03.1994	24.01.1985	1	190	4.91	4.68	6.06	5.69	0	0	1	2	9	12
~~30~~	1032	0	50	02.01.1942	20.01.1992	01.01.1976	1	170	2.47	3.36	3.88	4.13	0	1	0	1	16	30
31	1033	1	53	10.10.1942	18.11.1995	16.10.1982	0	163	2.16	2.94	3.60	3.61	0	0	1	2	13	20
~~32~~	1034	0	52	09.04.1945	26.05.1997	20.01.1988	0	185	3.53	3.94	4.70	4.94	0	0	0	1	9	40
~~33~~	1036	0	42	16.02.1947	02.04.1989	12.02.1977	1	162	3.64	3.24	4.59	3.88	0	0	1	2	12	10
~~34~~	1037	0	34	17.01.1959	21.03.1993	28.02.1987	0	177	3.69	4.12	5.12	4.95	0	0	0	0	6	-88
~~35~~	1038	0	45	26.06.1947	19.09.1992	31.03.1983	0	170	4.31	3.50	5.50	4.26	0	0	1	2	9	20
36	1039	1	38	15.01.1953	22.01.1991	25.01.1974	1	170	3.98	3.72	5.11	4.46	0	0	0	1	17	13

Note In order to return to the complete data set for further analyses you need to return to the select cases option and click the all cases button.

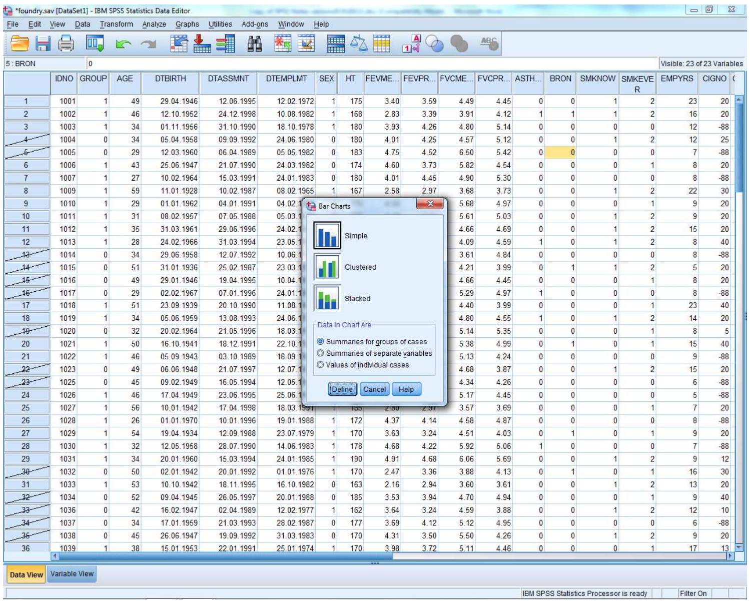
GRAPHICS

SPSS will produce good quality high- resolution statistical graphics. We will look at Bar Charts, Histograms, and Scatter Plots with regression lines directly from the data. Please note, that sometimes it is easier in Excel to create bar charts using the frequencies.

Bar Charts

Bar Charts can only be produced for categorical variables e.g. Ever smoked Asthma etc

To produce a Bar Chart click **Graphs, Legacy Dialogs** then **Bar** and the following screen appears.



*foundry.sav [DataSet1] - IBM SPSS Statistics Data Editor

5 : BRON 0

Visible: 23 of 23 Variables

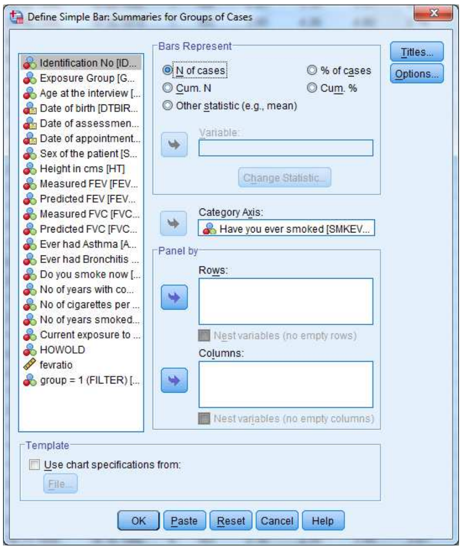
	IDNO	GROUP	AGE	DTBIRTH	DTASSMNT	DTEMPLMT	SEX	HT	FEVME...	FEVPR...	FVCME...	FVCPR...	ASTH...	BRON	SMKNOW	SMKEVER	EMPYRS	CIGNO
1	1001	1	49	29.04.1946	12.06.1995	12.02.1972	1	175	3.40	3.59	4.49	4.45	0	0	1	2	23	20
2	1002	1	46	12.10.1952	24.12.1998	10.08.1982	1	168	2.83	3.39	3.91	4.12	1	1	1	2	16	20
3	1003	1	34	01.11.1956	31.10.1990	18.10.1978	1	180	3.93	4.26	4.80	5.14	0	0	0	0	12	-88
~~4~~	1004	0	34	05.04.1958	09.09.1992	24.06.1980	0	180	4.01	4.25	4.57	5.12	0	0	1	2	12	25
~~5~~	1005	0	29	12.03.1960	06.04.1989	05.05.1982	0	183	4.75	4.52	6.50	5.42	0	0	0	0	7	-88
6	1006	1	43	25.06.1947	21.07.1990	24.03.1982	0	174	4.60	3.73	5.82	4.54	0	0	0	1	8	20
7	1007	1	27	10.02.1964	15.03.1991	24.01.1983	0	180	4.01	4.45	4.90	5.30	0	0	0	0	8	-88
8	1009	1	59	11.01.1928	10.02.1987	08.02.1965	1	167	2.58	2.97	3.68	3.73	0	0	1	2	22	30
9	1010	1	29	01.01.1962	04.01.1991	04.02.1					5.68	4.97	0	0	0	1	9	20
10	1011	1	31	08.02.1957	07.05.1988	05.03.1					5.61	5.03	0	0	1	2	9	20
11	1012	1	35	31.03.1961	29.06.1996	24.02.1					4.66	4.69	0	0	1	2	15	20
12	1013	1	28	24.02.1966	31.03.1994	23.05.1					4.09	4.59	1	0	1	2	8	40
~~13~~	1014	0	34	29.06.1958	12.07.1992	10.06.1					3.61	4.84	0	0	0	0	8	-88
~~14~~	1015	0	51	31.01.1936	25.02.1987	23.03.1					4.21	3.99	0	1	1	2	5	20
~~15~~	1016	0	49	29.01.1946	19.04.1995	10.04.1					4.66	4.45	0	0	0	1	8	20
~~16~~	1017	0	29	02.02.1967	07.01.1996	24.01.1					5.29	4.97	1	0	0	0	8	-88
17	1018	1	51	23.09.1939	20.10.1990	11.08.1					4.40	3.99	0	0	0	1	23	40
18	1019	1	34	05.06.1959	13.08.1993	24.06.1					4.80	4.55	1	0	1	2	14	20
~~19~~	1020	0	32	20.02.1964	21.05.1996	18.03.1					5.14	5.35	0	0	0	1	8	5
20	1021	1	50	16.10.1941	18.12.1991	22.10.1					5.38	4.99	0	1	0	1	15	40
21	1022	1	46	05.09.1943	03.10.1989	18.09.1					5.13	4.24	0	0	0	0	9	-88
~~22~~	1023	0	49	06.06.1948	21.07.1997	12.07.1					4.68	3.87	0	0	1	2	15	20
~~23~~	1025	0	45	09.02.1949	16.05.1994	12.05.1					4.34	4.26	0	0	0	0	6	-88
24	1026	1	46	17.04.1949	23.06.1995	25.06.1					5.17	4.45	0	0	0	0	5	-88
25	1027	1	56	10.01.1942	17.04.1998	18.03.1991	1	165	2.80	2.97	3.57	3.69	0	0	0	1	7	20
26	1028	1	26	01.01.1970	10.01.1996	19.01.1988	1	172	4.37	4.14	4.58	4.87	0	0	0	0	8	-88
27	1029	1	54	19.04.1934	12.09.1988	23.07.1979	1	170	3.63	3.24	4.51	4.03	0	1	0	1	9	20
28	1030	1	32	12.05.1958	28.07.1990	14.06.1983	1	178	4.68	4.22	5.92	5.06	1	0	0	0	7	-88
29	1031	1	34	20.01.1960	15.03.1994	24.01.1985	1	190	4.91	4.68	6.06	5.69	0	0	1	2	9	12
~~30~~	1032	0	50	02.01.1942	20.01.1992	01.01.1976	1	170	2.47	3.36	3.88	4.13	0	1	0	1	16	30
31	1033	1	53	10.10.1942	18.11.1995	16.10.1982	0	163	2.16	2.94	3.60	3.61	0	0	1	2	13	20
~~32~~	1034	0	52	09.04.1945	26.05.1997	20.01.1988	0	185	3.53	3.94	4.70	4.94	0	0	0	1	9	40
~~33~~	1036	0	42	16.02.1947	02.04.1989	12.02.1977	1	162	3.64	3.24	4.59	3.88	0	0	1	2	12	10
~~34~~	1037	0	34	17.01.1959	21.03.1993	28.02.1987	0	177	3.69	4.12	5.12	4.95	0	0	0	0	6	-88
~~35~~	1038	0	45	26.06.1947	19.09.1992	31.03.1983	0	170	4.31	3.50	5.50	4.26	0	0	1	2	9	20
36	1039	1	38	15.01.1953	22.01.1991	25.01.1974	1	170	3.98	3.72	5.11	4.46	0	0	0	1	17	13

Bar Charts

Data in Chart Are
- ● Summaries for groups of cases
- ○ Summaries of separate variables
- ○ Values of individual cases

IBM SPSS Statistics Processor is ready | Filter On

Click on **Simple** and then **Define** and the next screen will appear. Click **No of Cases**, then move your chosen variable from the left hand list to the **Categorical Axis** and press **OK.**



Define Simple Bar: Summaries for Groups of Cases

Identification No [ID...
Exposure Group [G...
Age at the interview [...
Date of birth [DTBIR...
Date of assessmen...
Date of appointment...
Sex of the patient [S...
Height in cms [HT]
Measured FEV [FEV...
Predicted FEV [FEV...
Measured FVC [FVC...
Predicted FVC [FVC...
Ever had Asthma [A...
Ever had Bronchitis ...
Do you smoke now [...
No of years with co...
No of cigarettes per ...
No of years smoked...
Current exposure to ...
HOWOLD
fevratio
group = 1 (FILTER) [...

Bars Represent
- ● N of cases
- ○ Cum. N
- ○ Other statistic (e.g., mean)
- ○ % of cases
- ○ Cum. %

Variable:

Category Axis: Have you ever smoked [SMKEV...

Panel by

Rows:

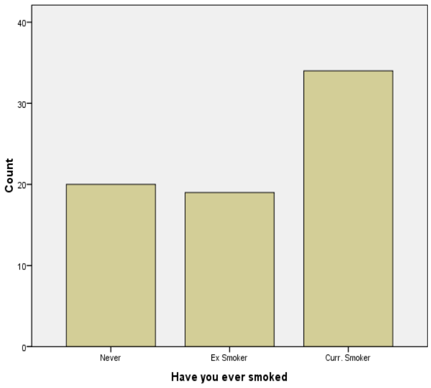
□ Nest variables (no empty rows)

Columns:

□ Nest variables (no empty columns)

Template

□ Use chart specifications from:



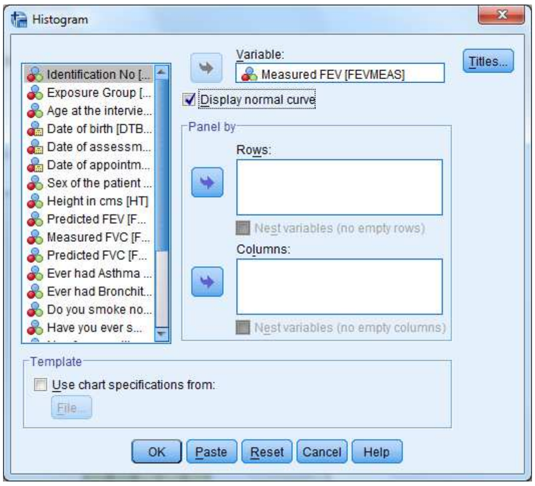
Video Tutorial – Bar Chart (specifically 3min+)

https://www.youtube.com/watch?v=0NeaD1Mojp0

Histograms

At this point it is a good idea to return the select cases back to all data, by **Data, Select Cases**, then **All Cases** followed by **ok**.

Histograms are produced for interval variables e.g. age. To produce a histogram click on **Graphs, Legacy Dialogs** then **Histogram** and the following screen appears.

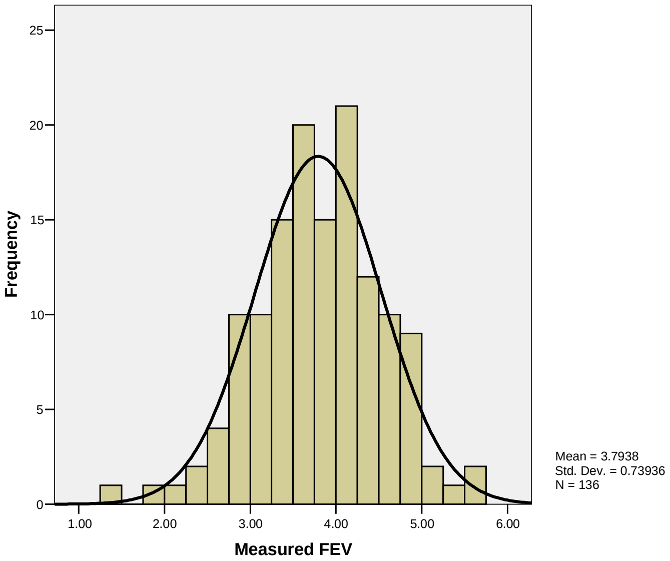


Variable: Measured FEV [FEVMEAS]

☑ Display normal curve

Click on the required variable, in this case FEV, in the left hand side list and press ➡ then press **OK.** If you require a normal curve to be drawn on the graph click on **Display normal curve.**

This is the Histogram produced for measured FEV.



<u>Video Tutorial – Histogram (specifically 2min 50)</u>

https://www.youtube.com/watch?v=LFGT0WqY5d4

Scatter Plots

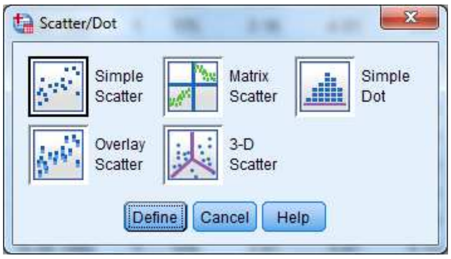
Scatter plots show the joint behaviour of two interval variables. If you want to decide whether two interval variables are related in any way you should first draw a scatter plot.

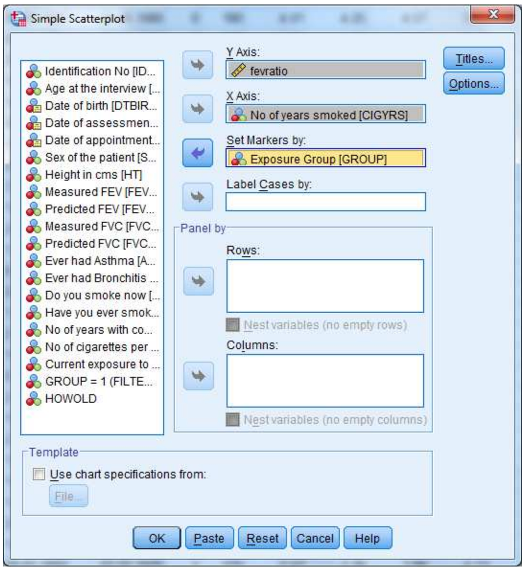
Scatter plots have 2 axes:

- the value of the dependent or response variable on the y axis.
- the value of the independent variable on the horizontal axis.

To run a scatter plot click **Graphs – Legacy Dialogs – Scatter/dot** and the following will appear.

Click on **Simple scatter** and then select variables



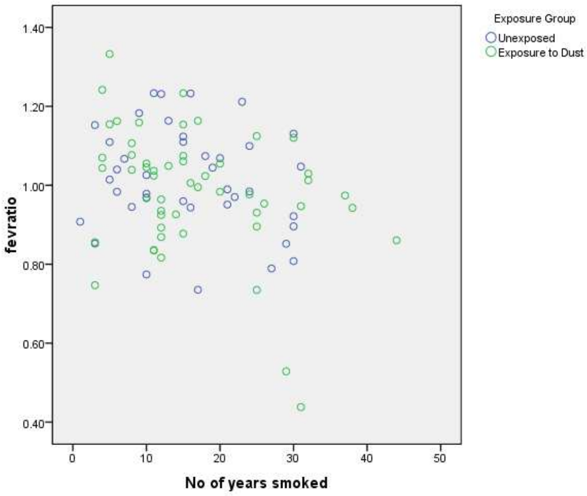


Y Axis: fevratio

X Axis: No of years smoked [CIGYRS]

Set Markers by: Exposure Group [GROUP]

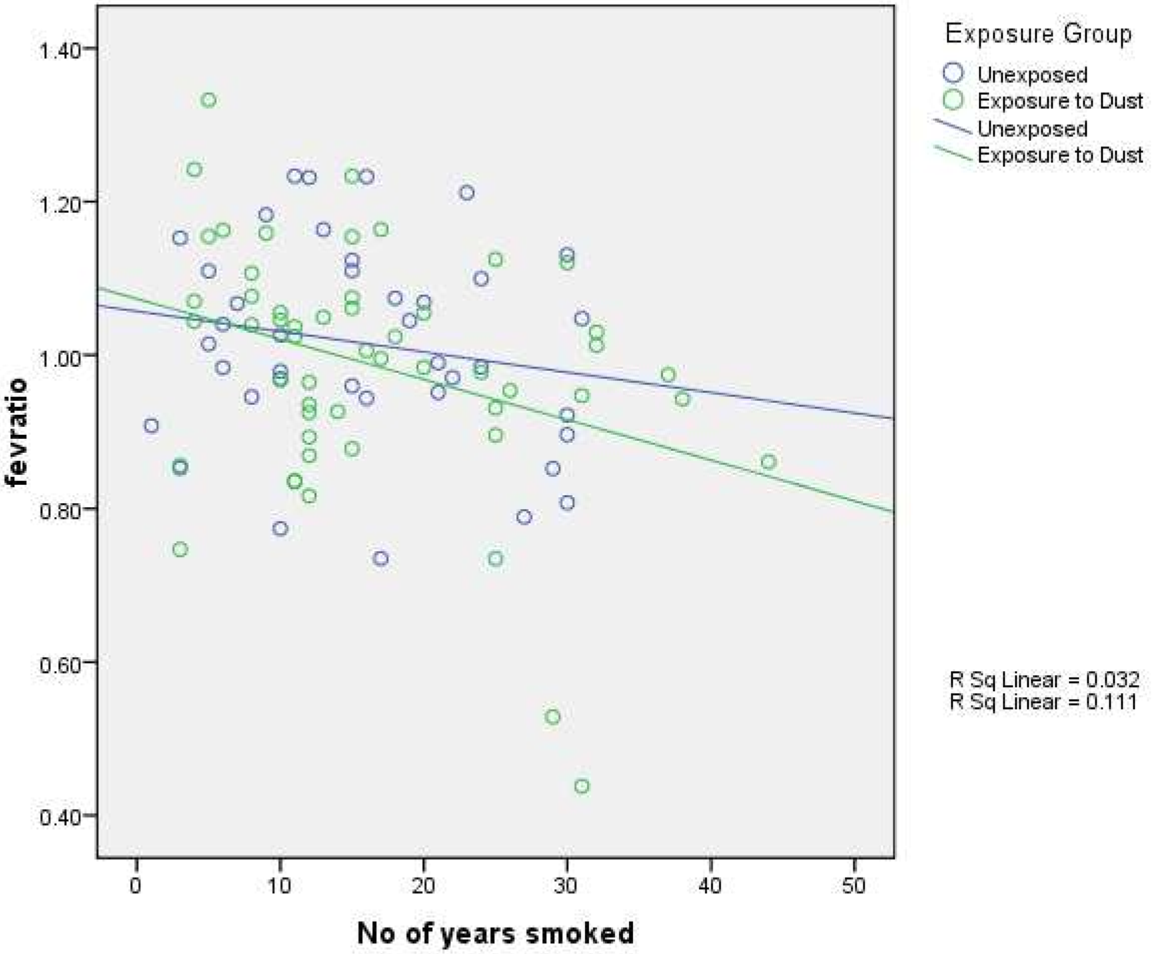
The above selection produces the following graph



Plotting a Regression Line on a Scatter Plot

To fit a line of regression, double left click on the graph. This moves the graph into the Chart Editor. A Regression line can be added by clicking on **Elements** then **Fit Line Total** if you have not defined any markers, or **Fit Line Subgroups** if you have defined markers.

This produces the following graph.



<u>Video Tutorial – Scatter Plot with regression line</u>

https://www.youtube.com/watch?v=blfflA-34pQ

STATISTICAL INFERENCE IN SPSS

Introduction

This part will introduce the basic methods of statistical inference available in SPSS. It will assume some familiarity with concepts in statistical inference including hypothesis testing and confidence intervals. If you are unfamiliar with these concepts, you are strongly recommended to read an introductory text in medical statistics such as Campbell and Machin “Medical Statistics A Common Sense Approach”.

The methods will be illustrated by the Foundry data set that was considered in Part I. The purpose of this study was to examine whether dust increased respiratory morbidity. In this study the measure of respiratory morbidity are “Ever had asthma", “Ever had bronchitis”, “Measured FEV” and “Measured FVC”. The variable “Predicted FEV” and “Predicted FVC” are the values that are expected for a person’s demographic characteristics including Age, Height and Sex. Exposure to dust is measured by two variables “Exposed/Un-exposed” and dust levels recorded only for exposed workers. Because smoking is a confounding factor in this study, smoking behaviour has been recorded in terms of current smoking status (smknow), smoking history (smkever), and consumption (cigno) and duration of smoking (cigyrs).

During this part of the practical you may need to refer to the notes from Part I. If you are starting the tutorial at this point rather than continuing from Part I, you will need to open the SPSS data as preciously shown on page 17.

Categorical Variable

In the first part of the study we examined whether there was any relationship between exposure to dust and smoking. Using the cross-tabs procedure we can generate the following table.

Do you smoke now * Exposure Group Crosstabulation

			Exposure Group		
			Unexposed	Exposure to Dust	Total
Do you smoke now	No	Count	43	39	82
		% within Exposure Group	68.3%	53.4%	60.3%
	Yes	Count	20	34	54
		% within Exposure Group	31.7%	46.6%	39.7%
Total		Count	63	73	136
		% within Exposure Group	100.0%	100.0%	100.0%

From the table above it can be seen that the percentage of workers who currently smoke is higher for those exposed to dust than those who are not, 47% as compared to 32%.

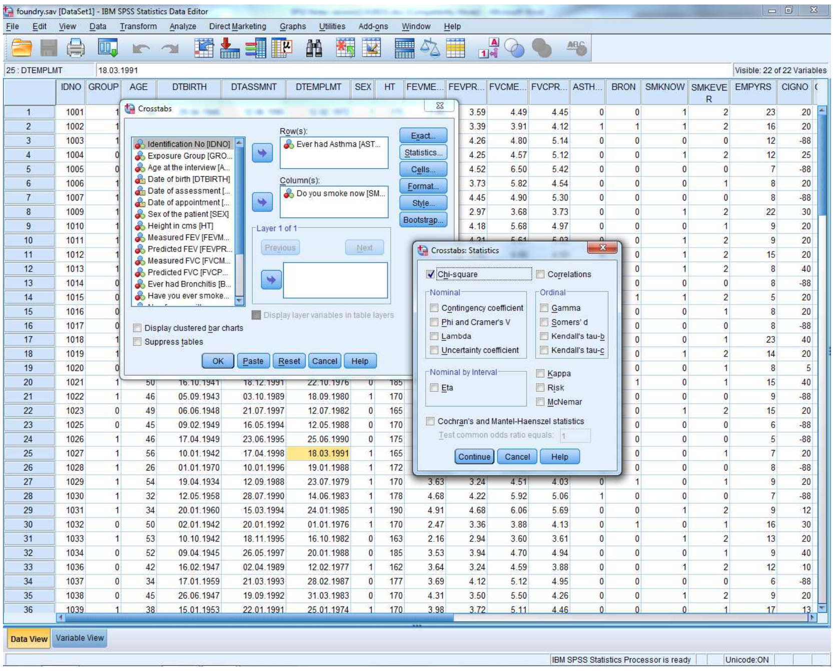
We will now examine whether respiratory symptoms as measured by the variable **asthma** relate to smoking. Using cross-tabs procedure again we obtain the following table.

Ever had Asthma * Do you smoke now Crosstabulation

			Do you smoke now		
			No	Yes	Total
Ever had Asthma	No	Count	77	48	125
		% within Do you smoke now	93.9%	88.9%	91.9%
	Yes	Count	5	6	11
		% within Do you smoke now	6.1%	11.1%	8.1%
Total		Count	82	54	136
		% within Do you smoke now	100.0%	100.0%	100.0%

The Chi-squared test and Fisher's Exact test

Amongst those who currently smoked 11.1% had experienced symptoms of asthma whilst only 6.3% amongst those who did not. Does this suggest that smoking may be related to asthma or might this difference be due to chance - that is explained by sampling variation? One way in which we can examine this is by a chi-squared test. This can be carried out by re-running the cross-tab procedure including the chi-squared statistics option as follows. In the cross-tabs panel (see illustration below) we select Statistics to reveal the second panel that lists possible statistics. In this panel we have selected **chi-squared.**


foundry.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Crosstabs
Exact...
Statistics...
Cells...
Format...
Style...
Bootstrap...
Previous
Next
OK Paste Reset Cancel Help
Crosstabs: Statistics
Continue Cancel Help
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

25 : DTEMPLMT 18.03.1991

Visible: 22 of 22 Variables

Crosstabs

Variables:
- Identification No [IDNO]
- Exposure Group [GRO...
- Age at the interview [A...
- Date of birth [DTBIRTH]
- Date of assessment [...
- Date of appointment [...
- Sex of the patient [SEX]
- Height in cms [HT]
- Measured FEV [FEVM...
- Predicted FEV [FEVPR...
- Measured FVC [FVCM...
- Predicted FVC [FVCP...
- Ever had Bronchitis [B...
- Have you ever smoke...

Row(s): Ever had Asthma [AST...

Column(s): Do you smoke now [SM...

Layer 1 of 1

☐ Display layer variables in table layers

☐ Display clustered bar charts

☐ Suppress tables

Crosstabs: Statistics

☑ 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

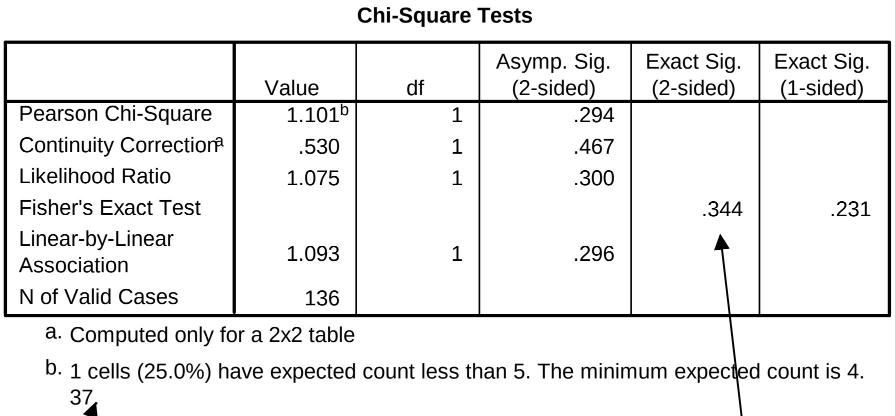
Nominal by Interval
☐ Eta

☐ Kappa
☐ Risk
☐ McNemar

☐ Cochran's and Mantel-Haenszel statistics

Test common odds ratio equals: 1

Then click on **continue** then **OK** to get the analysis below



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.101[b]	1	.294		
Continuity Correction[a]	.530	1	.467		
Likelihood Ratio	1.075	1	.300		
Fisher's Exact Test				.344	.231
Linear-by-Linear Association	1.093	1	.296		
N of Valid Cases	136				

a. Computed only for a 2x2 table

b. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4. 37.

The panel above gives the results of a chi-squared test of no association between asthma and smoking. In interpreting this table we are concerned with the columns headed “Asymp.Sig” and “Exact Sig.”. These columns give the p-values for the significance test. Firstly it is usually recommended that you consider a 2-sided rather than 1-sided test. As one of the cells has an expected count less than or equal to 5, it is recommended that we take the Fisher’s Exact Test value as our result – that is 0.344. Assuming the conventional 0.05 significance level, this result is considered non-significant. In reporting results of statistical tests you are strongly recommended to give the p-value rather than just write “significant” or “non-significant”. In reporting this we might write “there was no evidence of an association between smoking and asthma (Fisher’s Exact p=0.344)." Had the expected count been greater than 5 and the table greater than 2 by 2 it is suggested that you report the straight forward Chi-squared test p-value. If the expected count is greater than 5 but the table is a 2 by 2 then report the continuity correction p-value.

Exercise Using the cross-tabs procedure examine whether there is a relationship between current smoking status and bronchitis symptoms.

Are the expected numbers greater than 5 for all cells?

Fill in the spaces and delete as appropriate in the following statement:

“Amongst those that currently smoked ___% had experienced symptoms of bronchitis whereas ___% of non-smokers experience such symptoms. This was statistically significant/non significant at a 5% level using a two-tailed continuity corrected chi-squared test with p=______ “

Exercise Now use the cross-tabs procedure to examine the relationship between Exposure to dust and symptoms of bronchitis and asthma. Record your conclusions below using either the continuity corrected chi-squared or Fisher’s exact test as appropriate.

We have found no statistically significant relationship between exposure to dust and either asthma or bronchitis symptoms. For bronchitis symptoms you should have obtained the following tables.

Ever had Bronchitis * Exposure Group Crosstabulation

			Exposure Group		
			Unexposed	Exposure to Dust	Total
Ever had Bronchitis	No	Count	59	62	121
		% within Exposure Group	93.7%	84.9%	89.0%
	Yes	Count	4	11	15
		% within Exposure Group	6.3%	15.1%	11.0%
Total		Count	63	73	136
		% within Exposure Group	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.620[b]	1	.106		
Continuity Correction[a]	1.807	1	.179		
Likelihood Ratio	2.735	1	.098		
Fisher's Exact Test				.169	.088
Linear-by-Linear Association	2.601	1	.107		
N of Valid Cases	136				

[a] Computed only for a 2x2 table

[b] 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.95.

Whilst 15% (11/73) of the exposed worker had symptoms of bronchitis and only 6% (4/63) of non-exposed, this difference was not statistically significant at the 5% level (p=0.179). There are several explanations for this. There may be no relationship between the exposure to dust and respiratory disease. Alternatively, the study may have lacked statistical power to detect small differences. It should be noted also that only 11% (15/136) of the sample reported such symptoms.

<u>Video Tutorial – Chi-square test</u>

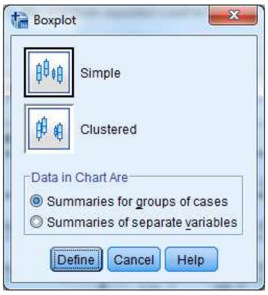
https://www.youtube.com/watch?v=wfIfEWMJY3s

CONTINUOUS OUTCOME MEASURES

We will now consider the lung function measurements. Given that lung function is age and size dependent it is usual to divide measured lung function by the expected lung function. In Part I we constructed such a variable.

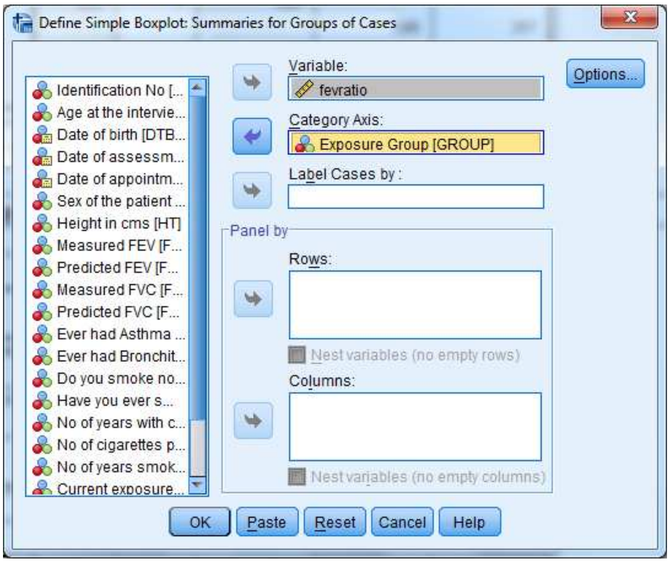
<u>Exercise</u> Using the Compute option in Transform construct new variable **fevratio** and **fvcratio** defined by **fevmeas/fevpred** and **fvcmeas/fvcpred**.

We now want to examine whether workers exposed to dust have reduced lung function. First we might examine this graphically with a box plot. Going to the graph menu, select **boxplot**.



Data in Chart Are
● Summaries for groups of cases
○ Summaries of separate variables

Select simple to get and transfer variable names in the usual way (see below).

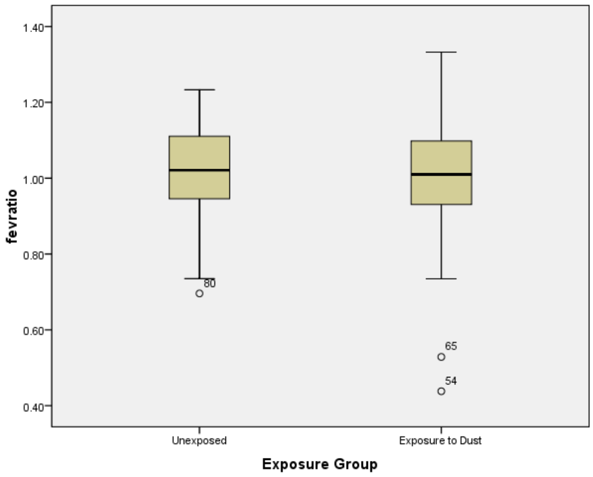


Variable: fevratio

Category Axis: Exposure Group [GROUP]

Label Cases by:

This gives the following plot



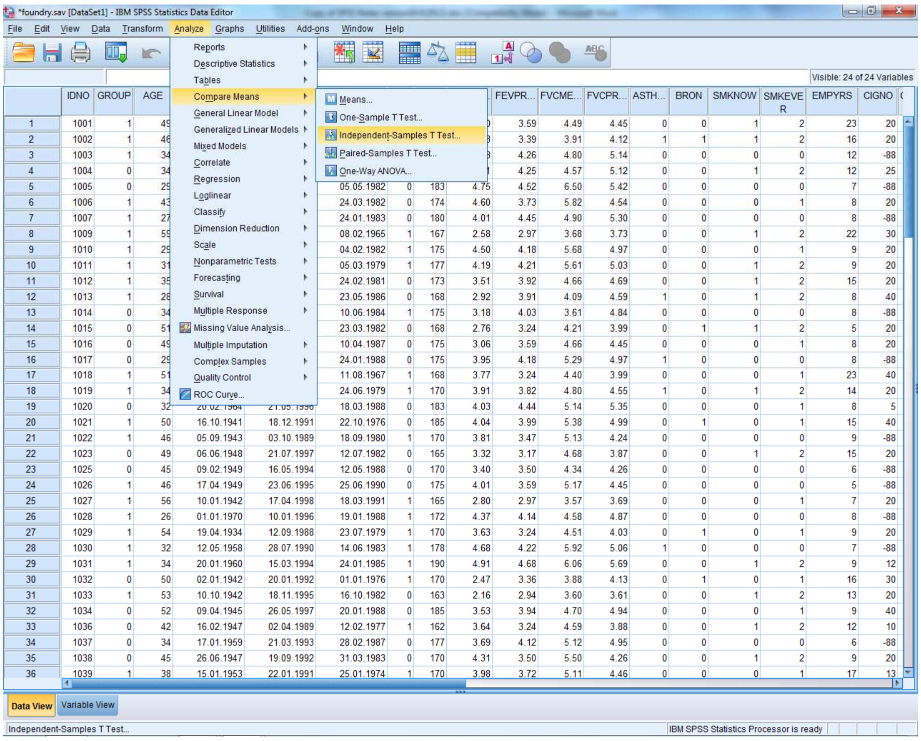
The box represents the inter-quartile range; the whiskers represent the range. The solid line in the middle represents the median. This suggests that there is little difference between the dust exposed and non-exposed workers. Other **Analysis** options we might use to compare the lung function of exposed and non-exposed workers are **Explore** in the **Descriptive** section and the **Means** under **Compare Means**.

Exercise Use **Explore** and **Means** options to compare lung function of exposed with non-exposed workers using fvcratio and fevratio. Record the results below.

	Mean	*Standard Deviation*	*Median*	*Max*	*Min*	*N*
Exposed						
Non Exposed						

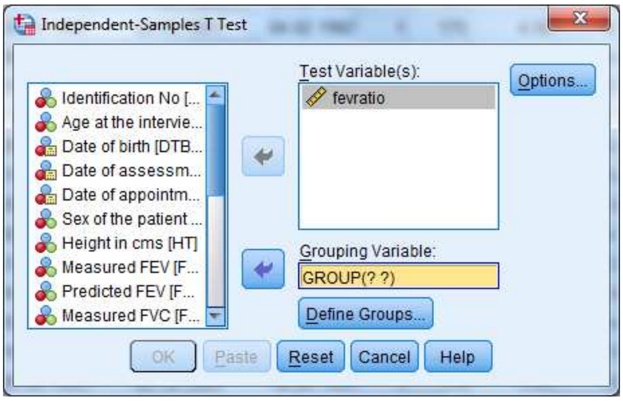
Comparison of Means Using a t-test

The t-test procedure can be used for statistical comparison of the mean **FEV ratio** of the exposed compared to non-exposed workers. It will also give the confidence interval for the difference of the two means. For the test go to **Compare means** then **Independent Sample t-test**



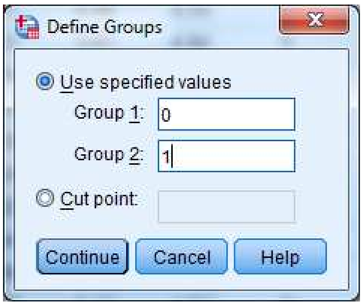
The following panel (below left) then appears into which we have selected **fevrat** as the test variable and group defining the exposure.

Note (? ?) marks beside the variable name **group**. Click on **Define Groups** to add the codes for the codes "0" and "1" for the two groups as shown (in the panel on the right).


Independent-Samples T Test

Test Variable(s): fevratio

Grouping Variable: GROUP(? ?)


Define Groups

● Use specified values

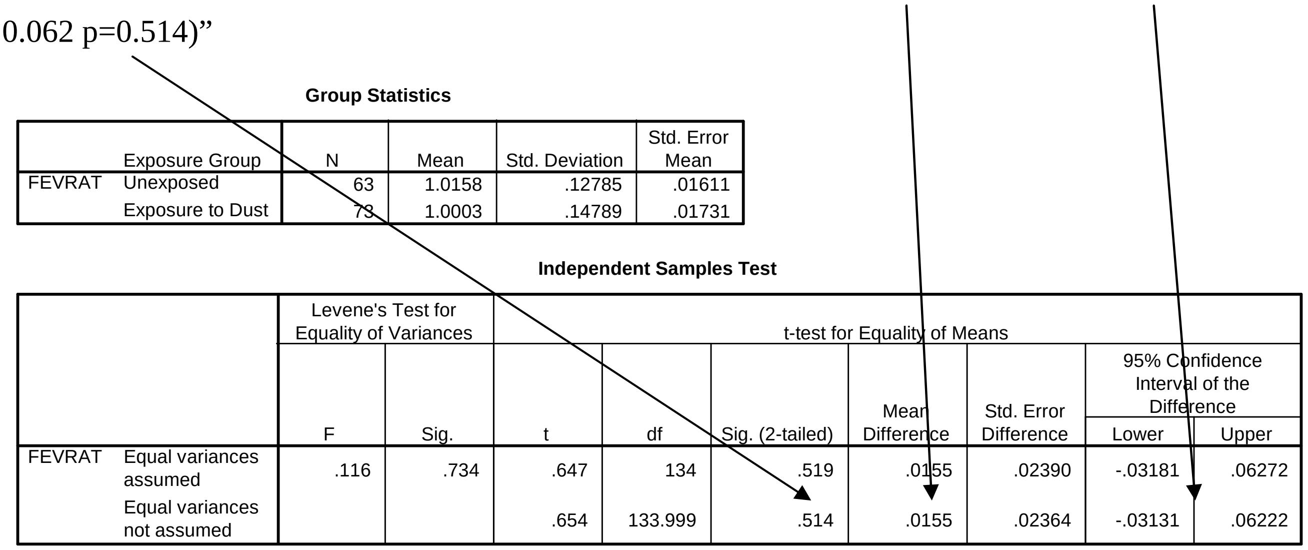
Group 1: 0

Group 2: 1

○ Cut point:

The ability to select groups by choice of codes simplifies things when there are more than two groups in the data set.

Clicking **Continue** then **Ok** gives the results below. The first summarises the data of the two groups. The second presents two analyses. The first two columns of data, the Levene's F-Test of equality of variance – the assumption of a t-test is that the means for each group have the same variance. The remainder summarise a t-test for equal and un-equal variance. Please note, we recommend always using the t-test assuming an unequal variance, unless there is a very strong belief that the two groups have equal variance. Therefore we take the second row of t-test results although in this case it makes little difference. The result can be summarised as "there was no evidence of increased FEV ratio for workers exposed to dust (mean diff=0.0155, 95% c.i -0.031 to 0.062 p=0.514)"



Group Statistics

	Exposure Group	N	Mean	Std. Deviation	Std. Error Mean
FEVRAT	Unexposed	63	1.0158	.12785	.01611
	Exposure to Dust	73	1.0003	.14789	.01731

Independent Samples Test

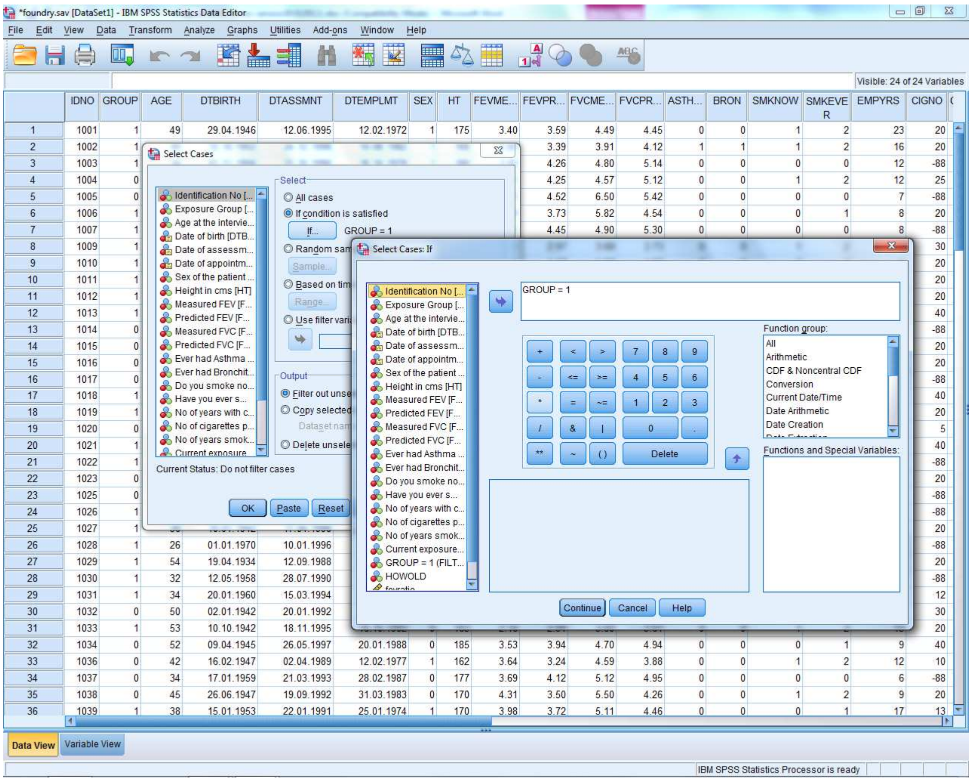
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
FEVRAT	Equal variances assumed	.116	.734	.647	134	.519	.0155	.02390	-.03181	.06272
	Equal variances not assumed			.654	133.999	.514	.0155	.02364	-.03131	.06222

<u>Video Tutorial – Independent groups t-test</u>

https://www.youtube.com/watch?v=8alv3kZt8Ug

<u>Exercise</u> Compare mean FVC ratio for the exposed and non-exposed subjects using a t-test

From the analyses there appears to be no evidence that exposure to dust affects respiratory function. It may be argued nevertheless that being categorised as "exposed" or "not exposed" is a crude assessment for exposure. Dust exposure has been recorded for subjects in the exposed group. We will now carry out some analysis on just the exposed subjects. First we select these from the data. This was shown in Part I of the tutorial. Under **Data** we choose **Select** cases then **If condition is satisfied** as shown below. We add the condition **group=1** subsequent analysis will only be on the dust exposed group.



*foundry.sav [DataSet1] - IBM SPSS Statistics Data Editor

Select Cases

Select: ● If condition is satisfied — GROUP = 1

Output: ● Filter out unselected cases

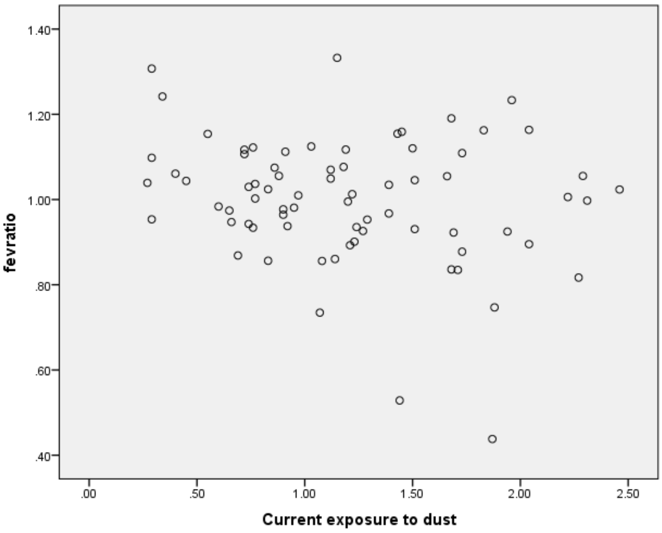
Current Status: Do not filter cases

Select Cases: If

GROUP = 1

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
OK Paste Reset
Continue Cancel Help
Function group:
Functions and Special Variables:
Data View Variable View
IBM SPSS Statistics Processor is ready

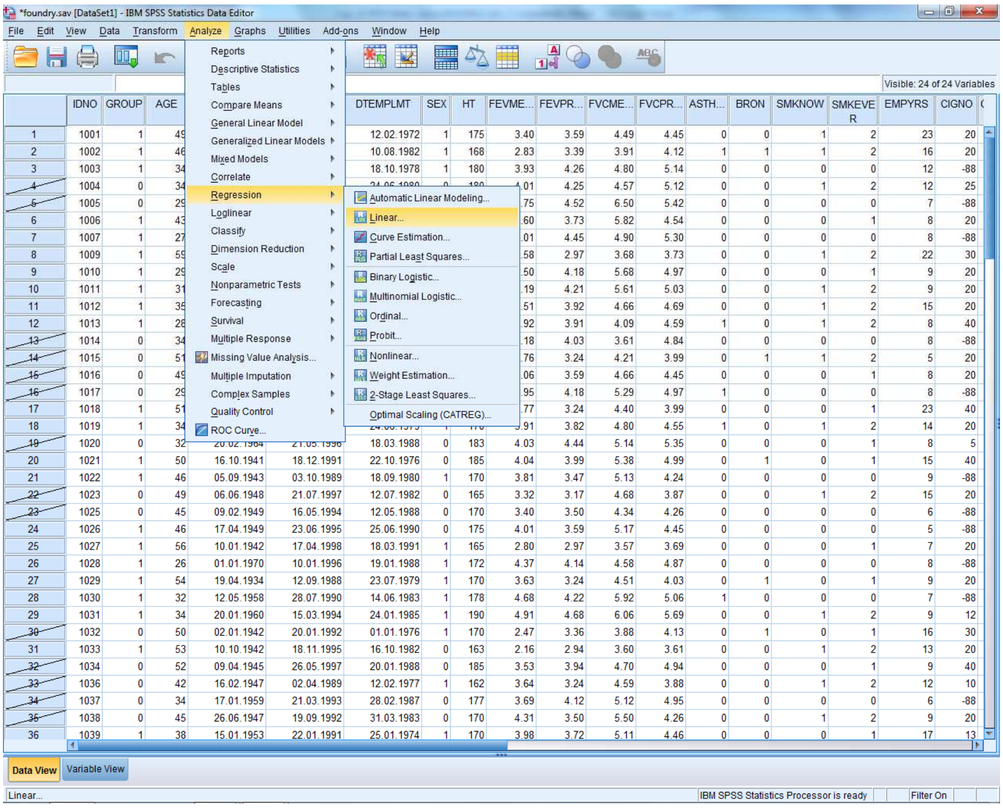
Below displays a scatter plot of FEV ratio compared to dust for subjects for the exposed group.



There is some suggestion from this that respiratory function may be reduced for those with higher exposure.

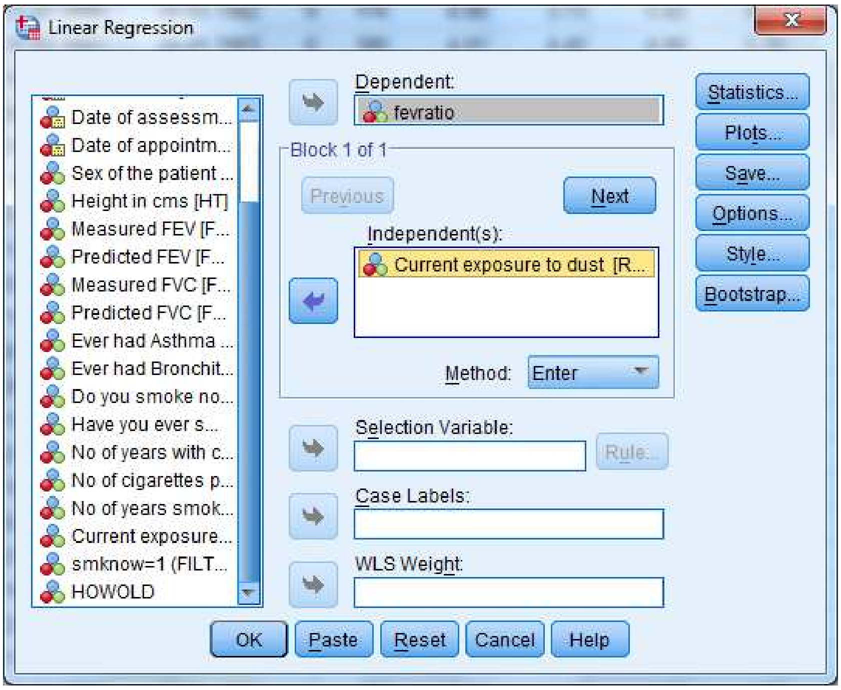
LINEAR REGRESSIONS

To test this we use a linear regression to fit a straight line in the form Y=A + BX. Where, Y is the dependent variable **fevratio** and X is an independent variable **respdust**. If the gradient (B) is negative this would indicate reduced respiratory function with increased dust. To do this in SPSS, whilst keeping select cases as exposure group 1, go to the **Regression** then **Linear** as shown



	IDNO	GROUP	AGE			DTEMPLMT	SEX	HT	FEVME...	FEVPR...	FVCME...	FVCPR...	ASTH...	BRON	SMKNOW	SMKEVER	EMPYRS	CIGNO
1	1001	1	49			12.02.1972	1	175	3.40	3.59	4.49	4.45	0	0	1	2	23	20
2	1002	1	46			10.08.1982	1	168	2.83	3.39	3.91	4.12	1	1	1	2	16	20
3	1003	1	34			18.10.1978	1	180	3.93	4.26	4.80	5.14	0	0	0	0	12	-88
~~4~~	1004	0	34			24.06.1980	0	180	4.01	4.25	4.57	5.12	0	0	1	2	12	25
~~5~~	1005	0	29						.75	4.52	6.50	5.42	0	0	0	0	7	-88
6	1006	1	43						.60	3.73	5.82	4.54	0	0	0	1	8	20
7	1007	1	27						.01	4.45	4.90	5.30	0	0	0	0	8	-88
8	1009	1	59						.58	2.97	3.68	3.73	0	0	1	2	22	30
9	1010	1	29						.50	4.18	5.68	4.97	0	0	0	1	9	20
10	1011	1	31						.19	4.21	5.61	5.03	0	0	1	2	9	20
11	1012	1	35						.51	3.92	4.66	4.69	0	0	1	2	15	20
12	1013	1	28						.92	3.91	4.09	4.59	1	0	1	2	8	40
~~13~~	1014	0	34						.18	4.03	3.61	4.84	0	0	0	0	8	-88
~~14~~	1015	0	51						.76	3.24	4.21	3.99	0	1	1	2	5	20
~~15~~	1016	0	49						.06	3.59	4.66	4.45	0	0	0	1	8	20
~~16~~	1017	0	29						.95	4.18	5.29	4.97	1	0	0	0	8	-88
17	1018	1	51						.77	3.24	4.40	3.99	0	0	0	1	23	40
18	1019	1	34				1	170	.91	3.82	4.80	4.55	1	0	1	2	14	20
~~19~~	1020	0	32	20.02.1964	21.05.1996	18.03.1988	0	183	4.03	4.44	5.14	5.35	0	0	0	1	8	5
20	1021	1	50	16.10.1941	18.12.1991	22.10.1976	0	185	4.04	3.99	5.38	4.99	0	1	0	1	15	40
21	1022	1	46	05.09.1943	03.10.1989	18.09.1980	1	170	3.81	3.47	5.13	4.24	0	0	0	0	9	-88
~~22~~	1023	0	49	06.06.1948	21.07.1997	12.07.1982	0	165	3.32	3.17	4.68	3.87	0	0	1	2	15	20
~~23~~	1025	0	45	09.02.1949	16.05.1994	12.05.1988	0	170	3.40	3.50	4.34	4.26	0	0	0	0	6	-88
24	1026	1	46	17.04.1949	23.06.1995	25.06.1990	0	175	4.01	3.59	5.17	4.45	0	0	0	0	5	-88
25	1027	1	56	10.01.1942	17.04.1998	18.03.1991	1	165	2.80	2.97	3.57	3.69	0	0	0	1	7	20
26	1028	1	26	01.01.1970	10.01.1996	19.01.1988	1	172	4.37	4.14	4.58	4.87	0	0	0	0	8	-88
27	1029	1	54	19.04.1934	12.09.1988	23.07.1979	1	170	3.63	3.24	4.51	4.03	0	1	0	1	9	20
28	1030	1	32	12.05.1958	28.07.1990	14.06.1983	1	178	4.68	4.22	5.92	5.06	1	0	0	0	7	-88
29	1031	1	34	20.01.1960	15.03.1994	24.01.1985	1	190	4.91	4.68	6.06	5.69	0	0	1	2	9	12
~~30~~	1032	0	50	02.01.1942	20.01.1992	01.01.1976	1	170	2.47	3.36	3.88	4.13	0	1	0	1	16	30
31	1033	1	53	10.10.1942	18.11.1995	16.10.1982	0	163	2.16	2.94	3.60	3.61	0	0	1	2	13	20
~~32~~	1034	0	52	09.04.1945	26.05.1997	20.01.1988	0	185	3.53	3.94	4.70	4.94	0	0	0	1	9	40
~~33~~	1036	0	42	16.02.1947	02.04.1989	12.02.1977	1	162	3.64	3.24	4.59	3.88	0	0	1	2	12	10
~~34~~	1037	0	34	17.01.1959	21.03.1993	28.02.1987	0	177	3.69	4.12	5.12	4.95	0	0	0	0	6	-88
~~35~~	1038	0	45	26.06.1947	19.09.1992	31.03.1983	0	170	4.31	3.50	5.50	4.26	0	0	1	2	9	20
36	1039	1	38	15.01.1953	22.01.1991	25.01.1974	1	170	3.98	3.72	5.11	4.46	0	0	0	1	17	13

In the following panel transfer the variables as shown.



Dependent: fevratio

Independent(s): Current exposure to dust [R...

Method: Enter

There are several tables of results generated by the linear regression option. The most useful of these is the table of coefficients shown below.

The coefficients are the values of A and B in the equation of the line **fevratio=A+B.respdust**

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.069	.041		26.019	.000
	Current exposure to dust	-.057	.031	-.212	-1.830	.071

a Dependent Variable: fevratio

The coefficient for respiratory is written -0.057. The column labelled “Sig.” gives the p-value for the statistical test that the regression coefficients differ from zero. This tell us that the constant is significantly different from zero which is not particularly interesting as we do not expect the intercept of the line with the y-axis to be zero. It gives a p-value of 0.071for the test that the gradient differs from zero. There is some suggestion of a negative gradient, but this is not significant at the conventional 5% significance level.

The **Model Summary** table reproduced below tells one how well the line fits that data. The result for R^2 (written “R square”) is 0.045. This is an estimate of the proportion of the variance explained by the model. A line that fits the data perfectly will have an R^2 equal to 1. Where as a line that does not explain anything in the data will have an R^2 of zero. A value of R^2 equal to 0.045 is therefore not at all good as only 4.5% of the variation in the data is being explained by the line.

Model Summary

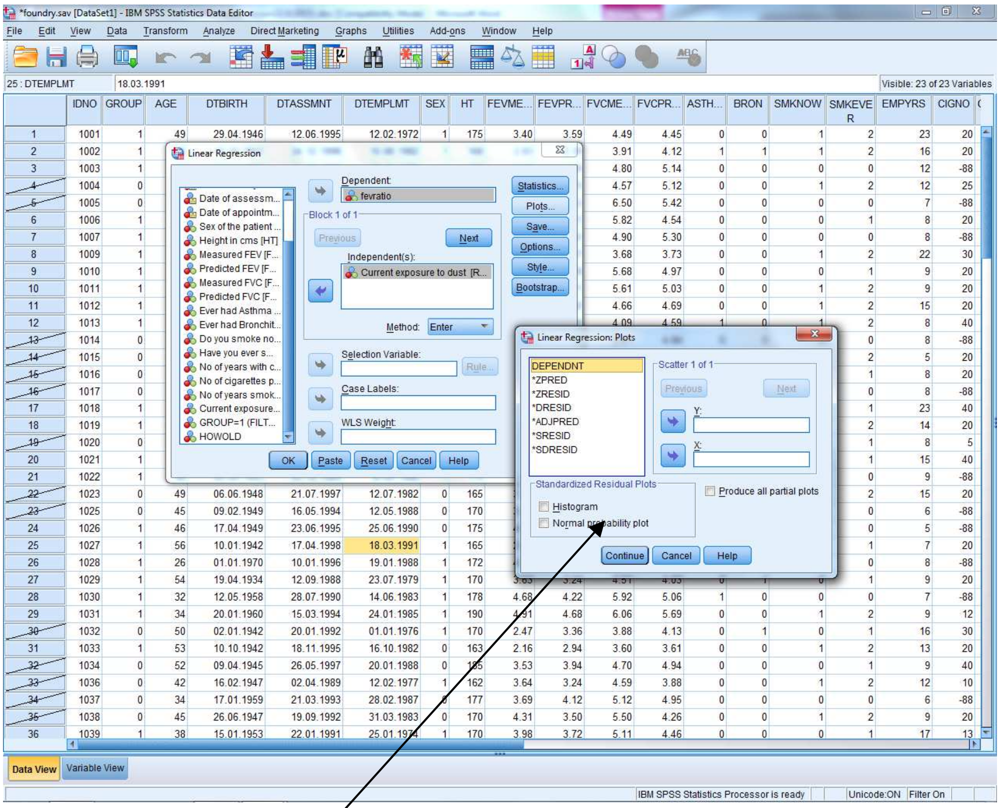
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.212[a]	.045	.032	.14553

a. Predictors: (Constant), Respiratory Dust

The conclusion that can be drawn from this is that whilst there is a slight suggestion of reduced respiratory function with increased dust exposure, the evidence is weak.

Model Checking

The linear regression model described by the coefficients allows one to estimate a predicted value. The difference between the observer value and the predicted value is called a residual. Where a model fits badly the regression line will have large residuals. If we consider the scatter plot above for FEV ratio compared to respiratory dust the residuals will be large. One of the assumptions of a regression model is that the residuals will have a normal distribution. One way to check this graphically is to use **normal probability plot**. This compares the residuals against a normal distribution. Such a plot can be obtained from linear regression in SPSS as shown



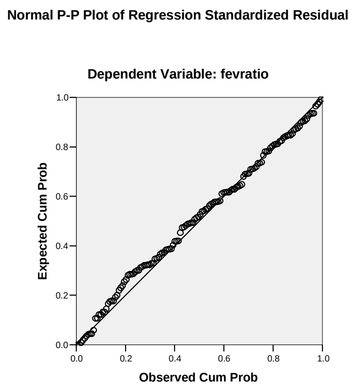
Dependent: fevratio

Block 1 of 1

Independent(s): Current exposure to dust [R...

Method: Enter

Just select the normal probability plot options. Then the plot will be added to the output when it is re-run. If the residuals are normally distributed the plotted points are on the diagonal line. The plot below suggests that the data are approximately normally distributed. If the data were skewed the points would bulge away from the line.



Video Tutorial – Independent groups t-test (Video 2 included model checking)

https://www.youtube.com/watch?v=vnQIW5ts3eM

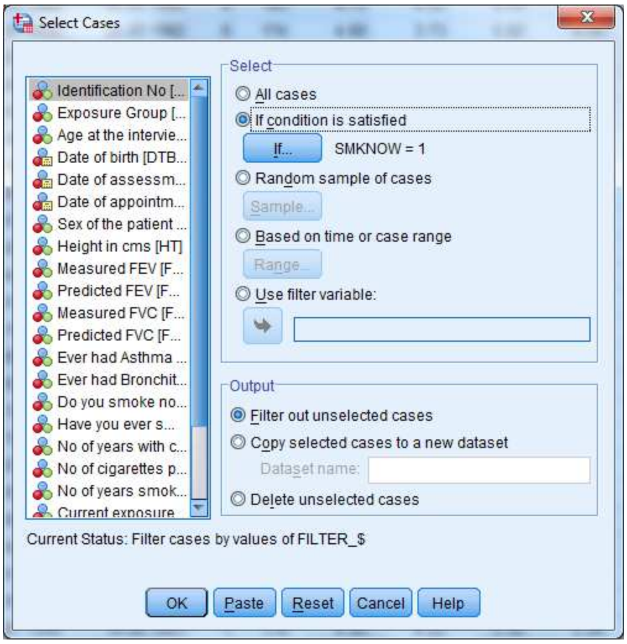
https://www.youtube.com/watch?v=U2p16pCHW3c

Exercise Examine the relationship between FVC ratio and dust levels using the methods above.

NON-PARAMETRIC METHODS

Where data is not normally distributed, statistical analyses that assume a normal distribution may be inappropriate. This is especially a concern where the sample size is small (<50 in total). Variables that are discrete (take only integer values) or have an upper or lower limit are by definition non-normal. Sometimes the distribution of the data is approximately normal so this is not a problem, particularly where the sample size is large, but for some variables it may be unreasonable to treat the data as normally distributed. To illustrate this we will compare the number of cigarettes smoked by "exposed" and "non-exposed" workers who currently smoke.

Before you start this you will need to reselect all cases as follows. To do this go to **Data** then **Select case** and change the if condition to **smknow=1** as shown**.**



Select Cases

Select

○ All cases

● If condition is satisfied

If... SMKNOW = 1

Output

● Filter out unselected cases

Current Status: Filter cases by values of FILTER_$

The frequency table for cigs per day for current smokers is given below.

No of cigarettes per day

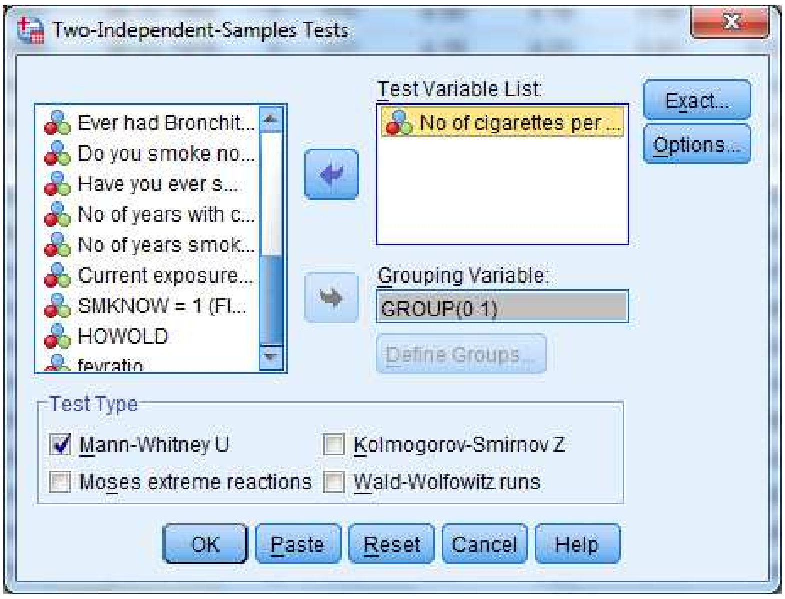
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3	2	3.7	3.7	3.7
	5	1	1.9	1.9	5.6
	6	1	1.9	1.9	7.4
	10	3	5.6	5.6	13.0
	12	2	3.7	3.7	16.7
	15	6	11.1	11.1	27.8
	18	1	1.9	1.9	29.6
	20	23	42.6	42.6	72.2
	25	6	11.1	11.1	83.3
	30	7	13.0	13.0	96.3
	40	2	3.7	3.7	100.0
	Total	54	100.0	100.0	

More than half the sample (30/54) give values of 20 or 30 cigs. per day. The variable is not even approximately normally distributed.

Exercise Use the **Explore** option under **Descriptive statistics** to determine the median and inter-quartile range for **No Cigs** consumed for Exposed and Non-dust exposed workers.

Suppose we wanted to compare the median number of cigarettes smoked per day by smokers according to dust exposure group. The method one uses is the Mann-Whitney U-test, which is called a rank based **non-parametric** method. The analysis is based not on the raw data values but on the ranks of the data. The procedure ranks the values of numbers of cigarettes smoked from smallest to largest.

The Mann-Whitney U-Test is carried out as follows. Under **Analysis** select **Non-parametric – Legacy Dialogs** to give a choice of non-parametric procedure. As we are going to compare two groups the choice in this case is then **2-Independent Groups**. In this panel select, Mann-Whitney U-test, **No cigs** as the test variable and **Group** as the grouping variable as shown.



Test Variable List: No of cigarettes per ...

Grouping Variable: GROUP(0 1)

Test Type

☑ Mann-Whitney U ☐ Kolmogorov-Smirnov Z

☐ Moses extreme reactions ☐ Wald-Wolfowitz runs

This generates the following output

Ranks

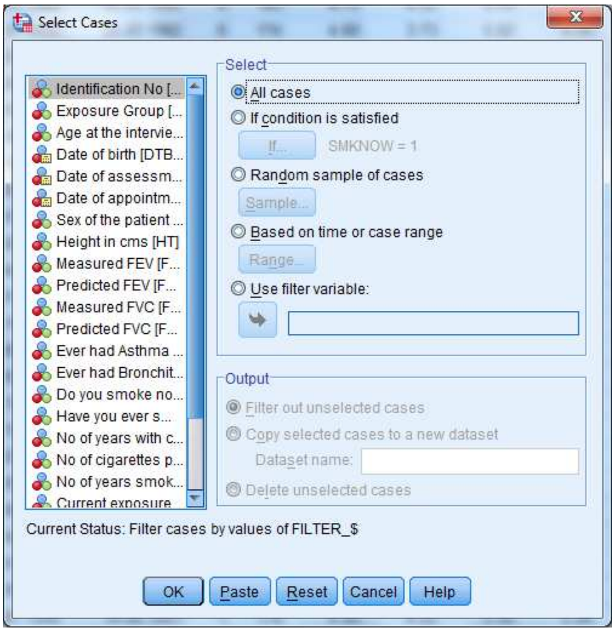
	Exposure Group	N	Mean Rank	Sum of Ranks
No of cigarettes per day	Unexposed	20	25.45	509.00
	Exposure to Dust	34	28.71	976.00
	Total	54		

Test Statistics[a]

	No of cigarettes per day
Mann-Whitney U	299.000
Wilcoxon W	509.000
Z	-.767
Asymp. Sig. (2-tailed)	.443

a. Grouping Variable: Exposure Group

In the tables above note the mean rank for each group and the significance level. The mean rank is slightly lower for the unexposed group but this is not statistically significant at a 5% significance level. Hence, we conclude that there is no difference between the median number of cigarettes smoked by "exposed" and "non-exposed" workers. Before moving on to the next analysis we need to select all subjects from the data menu.



Select
● All cases
○ If condition is satisfied
○ 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
Dataset name:
○ Delete unselected cases

Current Status: Filter cases by values of FILTER_$

Video Tutorial – Mann-Whitney U-test

https://www.youtube.com/watch?v=ALfW6DayQks

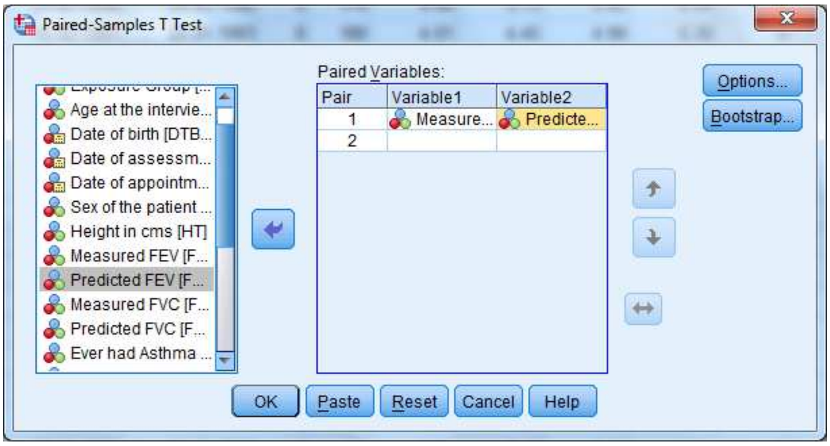
COMPARISONS OF RELATED OR PAIRED VARIABLES

For most of the analysis above we have compare the "exposed" and "non-exposed" groups of workers. In some circumstances we want to compare measures within the same subject. Such comparisons are sometimes referred to as **paired** or **pair-matched.**

Continuous Outcome Measures

One might want to compare the mean of a continuous measure at one time point with the mean of the same measure at a different time point. Whilst this may not be a sensible analysis for this data, we can illustrate this for a continuous variable by comparing FEV measured with FVC measured.

To compare the mean measured FEV with mean predicted FEV we select a **Paired samples T-test** in the **Compare means** submenu. This gives the panel below. Pairs of variables are selected by highlighting the pair of variables in the window to the left then clicking on the select button to transfer to the **Paired Variable** window as shown.



Paired-Samples T Test

Exposure Group [...
Age at the intervie...
Date of birth [DTB...
Date of assessm...
Date of appointm...
Sex of the patient ...
Height in cms [HT]
Measured FEV [F...
Predicted FEV [F...
Measured FVC [F...
Predicted FVC [F...
Ever had Asthma ...

Paired Variables:

Pair	Variable1	Variable2
1	Measure...	Predicte...
2		

Results are given below

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Measured FEV	3.7938	136	.73936	.06340
	Predicted FEV	3.7552	136	.45619	.03912

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Measured FEV & Predicted FEV	136	.739	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Measured FEV - Predicted FEV	.03860	.50632	.04342	-.04726	.12447	.889	135	.376

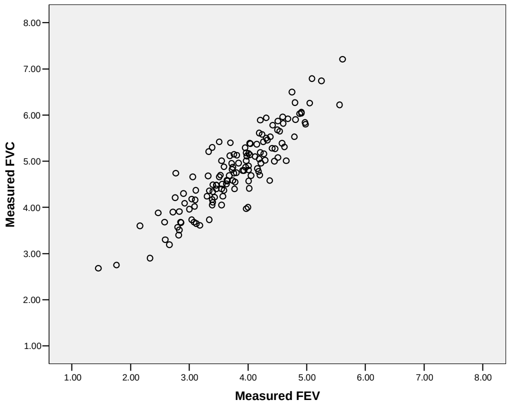
It is readily apparent that mean *measured FEC* is slightly greater than mean predicted *FEV***.** However, we report this as “Measured FVC was not significantly higher than measured FEV as (meandiff=0.038, 95% c.i. -0.0473 to 0.1245, p>0.05)”

Video Tutorial – Paired Samples t-test

https://www.youtube.com/watch?v=MJGk2sg4EZU

Exercise Compare the mean measured FVC with the mean predicted FVC.

The above method of analysis compares the mean value for the two variables. It does not tell one how close individual values are for the same subject. A visual way in which one can do this is with a scatter plot of the two variables as shown below. We get a visual impression that FEV and FVC are quite strongly correlated. By choosing the same numerical range for both axes we can see also that the values for FVC are systematically larger than for FEV.



Analysis of Binary Outcomes that are Related

Suppose we wish to compare the proportion of workers who had bronchitis symptoms with the proportion who had asthma symptoms. One might first construct the cross-tabulation using the cross tabs procedure. Both row and column percentages have been added.

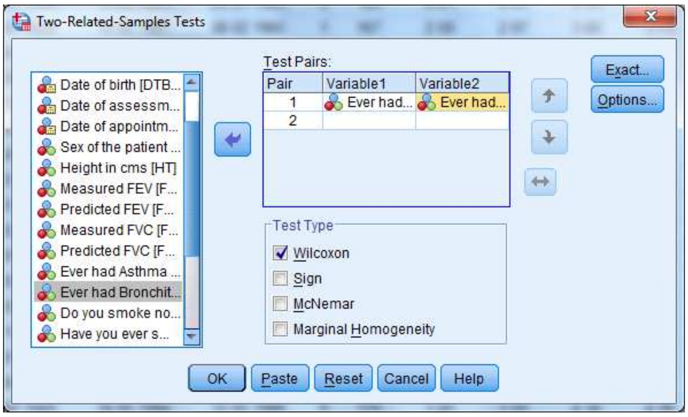
Ever had Bronchitis * Ever had Asthma Crosstabulation

			Ever had Asthma		Total
			No	Yes	
Ever had Bronchitis	No	Count	113	8	121
		% within Ever had Bronchitis	93.4%	6.6%	100.0%
		% within Ever had Asthma	90.4%	72.7%	89.0%
	Yes	Count	12	3	15
		% within Ever had Bronchitis	80.0%	20.0%	100.0%
		% within Ever had Asthma	9.6%	27.3%	11.0%
Total		Count	125	11	136
		% within Ever had Bronchitis	91.9%	8.1%	100.0%
		% within Ever had Asthma	100.0%	100.0%	100.0%

Careful examination of this table reveals that 11% (15/136) of workers reported bronchitis whilst only

8% (11/136) had asthma. These two proportions can be compared using McNemar's test. This is available under **2 Related samples** in the **Non-parametric** sub menu. Select the pair of variables in the same way as for a paired t-test and select the **McNemar**

Option



Test Pairs:

Pair	Variable1	Variable2
1	Ever had...	Ever had...
2		

Test Type

☑ Wilcoxon

□ Sign

□ McNemar

□ Marginal Homogeneity

.

This gives the following results

Test Statistics(b)

	Ever had Asthma & Ever had Bronchitis
N	136
Exact Sig. (2-tailed)	.503(a)

a Binomial distribution used.

b McNemar Test

The p-value for the McNemar test is not significant (p=0.503) so we conclude that symptoms of bronchitis are no more common in this population than symptoms of asthma.

Video Tutorial – Paired Binary McNemars Test

https://www.youtube.com/watch?v=3JNGOtKR28I

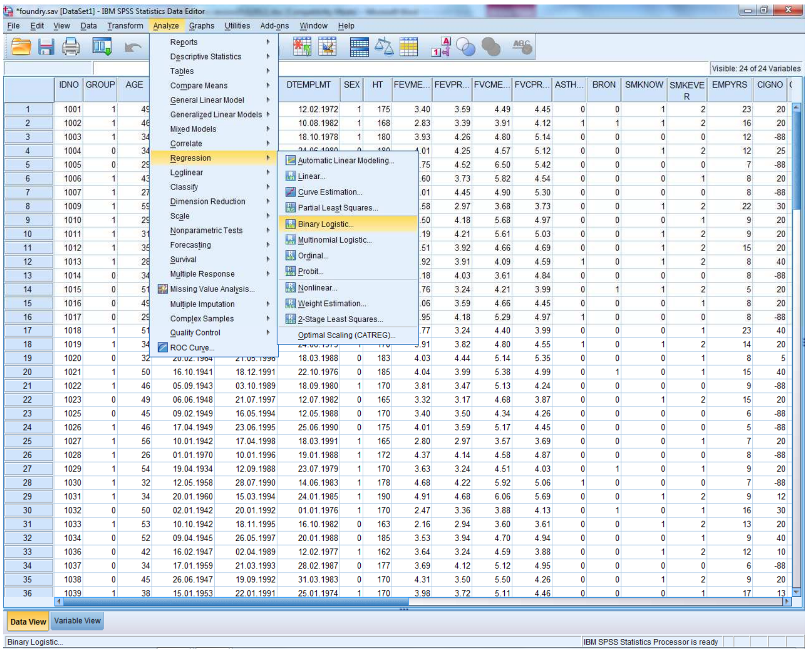
Related Ordinal Data

For ordered categorical or quantitative variables that are not plausibly normal the suggested procedure is to use the **Wilcoxon** procedure. This is selected from the same panel as McNemar Test (see above).

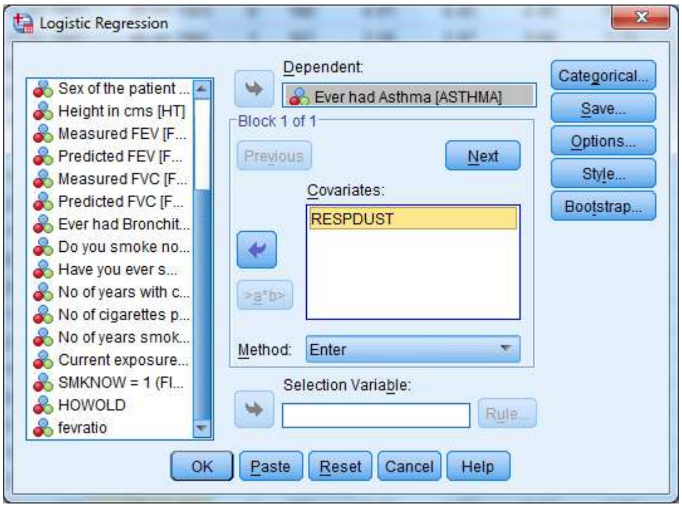
LOGISTIC REGRESSIONS

It is possible to apply regression techniques to a binary outcome e.g. Ever had Asthma Yes or No and test the effect of predictors on this outcome. We use logistic regression to fit a straight line of the form Y=A + BX.

Where Y is a link function called **logit** that converts the dependent variable **Ever had Asthma** from a binary (0=No 1=Yes) variable into a probability of success (i.e. probability of Yes anwer) and X is the standard independent variable **respdust**. Unlike before, the gradient B is a coefficient and can not be interpreted as linear regression, alternatively the coefficient can be altered using the exponential function so as to be considered an odds ratio (we will explain how to interpret this later) To do this in SPSS, go to the **Regression** then **Binary Logistic** as shown



In the following panel transfer the variables as shown.



Dependent: Ever had Asthma [ASTHMA]

Covariates: RESPDUST

Method: Enter

If the variable you wish to assess is a categorical variable then click **Categorical** and transfer the appropriate variable into the appropriate box, note also click the radio button indicating **Reference Category** as the first. Then finally click **ok**. There are several tables of results generated by the linear regression option. The most useful of these is the table of coefficients shown below.

The coefficients are the values of A and B in the equation of the line logit(**asthma**)=**A+B.respdust**

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	respdust	.892	.414	4.649	1	.031	2.439
	Constant	-3.190	.543	34.468	1	.000	.041

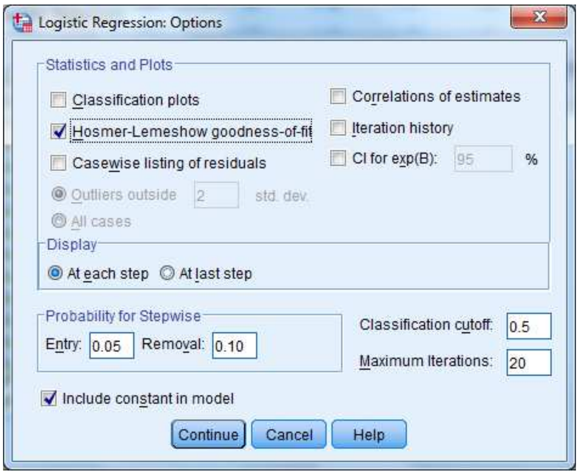
a Variable(s) entered on step 1: respdust.

The coefficient for respiratory is written 0.892 but the odds ratio is given as 2.439. An odds ratio falls into three distinct groups, equal to 1, less than and greater than 1. An odds ratio of 1 indicates no change in the likelihood of having the event (asthma) as the predictor changes (respdust). Odds ratios greater than 1 indicate an increased likelihood of asthma and an odds ratio less than 1 indicates a decreased likelihood of asthma. So in this case we say that as respdust increase by one unit the likelihood of having asthma increases by a multiple of 2.439. If the variable being tested was a categorical variable say gender then one of the categories would be classed as the reference category and the odds ratio would refer to the difference between the two groups. Say the odds ratio was 3.2 and Males are the reference category we would say, Females are approximately 3 times as likely as males to develop Asthma.

The column labelled "Sig." gives the p-value for the statistical test that the odds ratios significantly differ from one. This tells us that with a p-value of 0.031 for the test that the odds ratio for respdust differs from one. There is strong suggestion of an increased likelihood of having asthma, this is significant at the conventional 5% significance level.

Model Checking

In order to assess the Logistic regression models ability to represent the data we use a statistical test called the Hosmer & Lemeshow test. It is based on grouping cases into 10 equally spaced groups of risk and comparing the observed probability with the expected probability within each group. To perform this in SPSS repeat the process as if you where performing the logistic regression, so **Analyse – Regression – Binary Logistic** and click the box **Options** to get



Logistic Regression: Options

Statistics and Plots

☐ Classification plots
☑ Hosmer-Lemeshow goodness-of-fit
☐ Casewise listing of residuals
● Outliers outside 2 std. dev.
○ All cases
☐ Correlations of estimates
☐ Iteration history
☐ CI for exp(B): 95 %

Display

● At each step ○ At last step

Probability for Stepwise

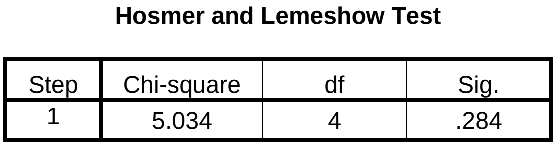
Entry: 0.05 Removal: 0.10

Classification cutoff: 0.5

Maximum Iterations: 20

☑ Include constant in model

Tick the box corresponding to the **Hosmer-Lemeshow goodness of fit**, then click **Continue** followed by **Ok**. In the same output that you received before the following table should appear,



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.034	4	.284

In this case we have only included one variable (respdust) you may require there to be several variables in your model. The above table may therefore have several steps, in that case then the last step will always be the result for the final model. The model is deemed unsuitable if the p-value (in the Sig column) is less than 0.05, therefore in this case as the p-value is 0.284 the model is deemed to be adequate.

Video Tutorial – Paired Binary McNemars Test

https://www.youtube.com/watch?v=iZoaXETWAN4

Exercise Examine the relationship between Bronchitis and dust levels using the methods above.

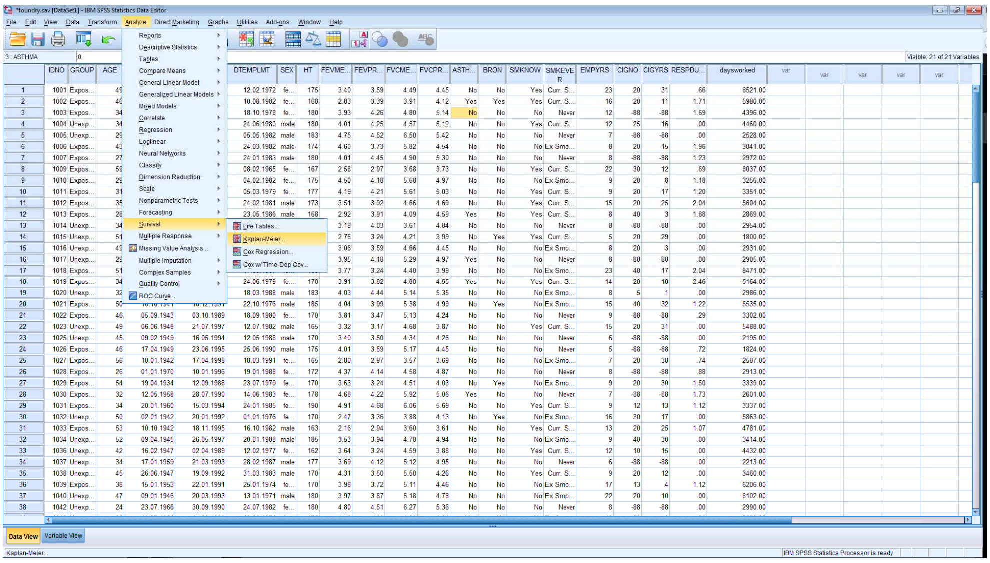
SURVIVAL ANALYSIS

A common form of data in medical research is **survival** or **time to an event** data. This where we are interested in comparing the length of **time** (since observations began) before an **event** occurs (e.g. death, disease occurrence, or a recovery). However, it may be that the event does not occur in all study subjects. Subjects, in whom the event has not been observed to have occurred during the study period, have their time **censored** based on the last date when the subject was observed to be event free. It may then be that you wish to compare the expected time to event in two (or more) groups. For example, does a particular treatment improve the survival time of a patient with a chronic condition. Note, it may be that the subjects enter the study at different time points and so the length of time until censor date may be different for each subject.

We investigate time to event using survival analysis techniques such as Kaplan Meier plots or Log Rank tests. To do this the data will require a **time variable** representing the time (e.g. days or months) between beginning observations (e.g. start of study, or entry into study) and either the event or the censor date (in those where the event was not observed). A second variable (e.g. event present Yes/No) will then identify if the event occurred in the subject or was censored.

The Foundry data provided should contain time until assessment in days (see Exercise on page 24). Assume for now that Asthma was diagnosed at this assessment, the calculated time variable can then be the time from employment until last assessment in the study where if the Asthma present variable is *Yes* then it represents time to the event, and if the Asthma present variable is *No* then it represents the time to censor date.

A Kaplan Meier plot can be used to describe if Asthma was occurring earlier in those exposed to dust vs those not exposed to dust. To generate a Kaplan Meier plot in SPSS, go to the **Analysis** > **Survival > Kaplan-Meier**

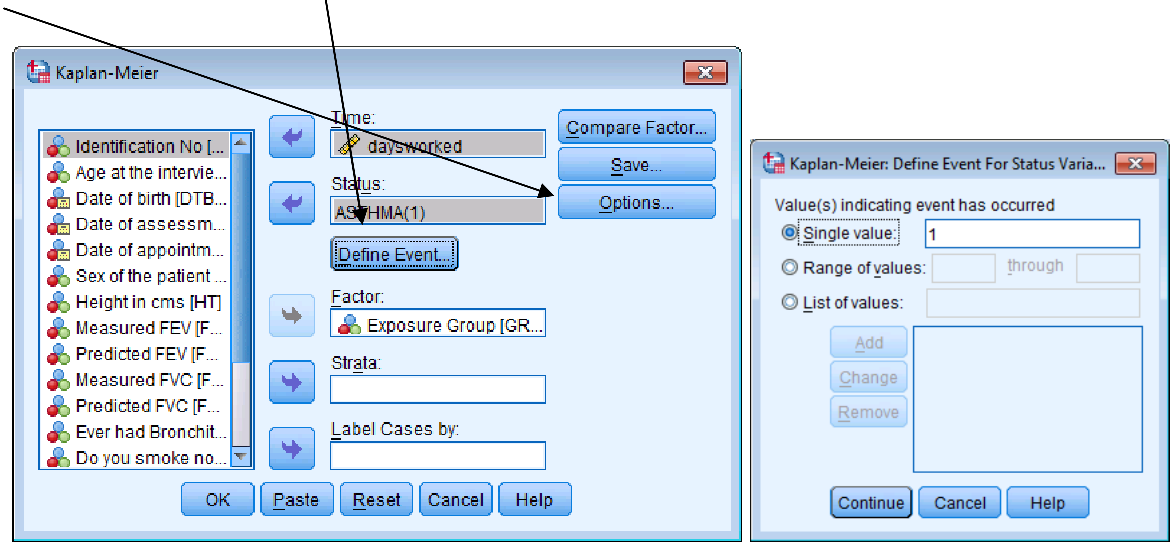


3 : ASTHMA 0

Visible: 21 of 21 Variables

	IDNO	GROUP	AGE			DTEMPLMT	SEX	HT	FEVME...	FEVPR...	FVCME...	FVCPR...	ASTH...	BRON	SMKNOW	SMKEVER	EMPYRS	CIGNO	CIGYRS	RESPDU...	daysworked
1	1001	Expos...	49			12.02.1972	fe...	175	3.40	3.59	4.49	4.45	No	No	Yes	Curr. S...	23	20	31	.66	8521.00
2	1002	Expos...	46			10.08.1982	fe...	168	2.83	3.39	3.91	4.12	Yes	Yes	Yes	Curr. S...	16	20	11	1.71	5980.00
3	1003	Expos...	34			18.10.1978	fe...	180	3.93	4.26	4.80	5.14	No	No	No	Never	12	-88	-88	1.69	4396.00
4	1004	Unexp...	34			24.06.1980	male	180	4.01	4.25	4.57	5.12	No	No	Yes	Curr. S...	12	25	16	.00	4460.00
5	1005	Unexp...	29			05.05.1982	male	183	4.75	4.52	6.50	5.42	No	No	No	Never	7	-88	-88	.00	2528.00
6	1006	Expos...	43			24.03.1982	male	174	4.60	3.73	5.82	4.54	No	No	No	Ex Smo...	8	20	15	1.96	3041.00
7	1007	Expos...	27			24.01.1983	male	180	4.01	4.45	4.90	5.30	No	No	No	Never	8	-88	-88	1.23	2972.00
8	1009	Expos...	59			08.02.1965	fe...	167	2.58	2.97	3.68	3.73	No	No	Yes	Curr. S...	22	30	12	.69	8037.00
9	1010	Expos...	29			04.02.1982	fe...	175	4.50	4.18	5.68	4.97	No	No	No	Ex Smo...	9	20	8	1.18	3256.00
10	1011	Expos...	31			05.03.1979	fe...	177	4.19	4.21	5.61	5.03	No	No	Yes	Curr. S...	9	20	17	1.20	3351.00
11	1012	Expos...	35			24.02.1981	male	173	3.51	3.92	4.66	4.69	No	No	Yes	Curr. S...	15	20	25	2.04	5604.00
12	1013	Expos...	28			23.05.1986	male	168	2.92	3.91	4.09	4.59	Yes	No	Yes	Curr. S...	8	40	3	1.88	2869.00
13	1014	Unexp...	34						3.18	4.03	3.61	4.84	No	No	No	Never	8	-88	-88	.00	2954.00
14	1015	Unexp...	51						2.76	3.24	4.21	3.99	No	Yes	Yes	Curr. S...	5	20	29	.00	1800.00
15	1016	Unexp...	49						3.06	3.59	4.66	4.45	No	No	No	Ex Smo...	8	20	3	.00	2931.00
16	1017	Unexp...	29						3.95	4.18	5.29	4.97	Yes	No	No	Never	8	-88	-88	.00	2905.00
17	1018	Expos...	51						3.77	3.24	4.40	3.99	No	No	No	Ex Smo...	23	40	17	2.04	8471.00
18	1019	Expos...	34			24.06.1979	fe...	170	3.91	3.82	4.80	4.55	Yes	No	Yes	Curr. S...	14	20	18	2.46	5164.00
19	1020	Unexp...	32			18.03.1988	male	183	4.03	4.44	5.14	5.35	No	No	No	Ex Smo...	8	5	1	.00	2986.00
20	1021	Expos...	50			22.10.1976	male	185	4.04	3.99	5.38	4.99	No	Yes	No	Ex Smo...	15	40	32	1.22	5535.00
21	1022	Expos...	46	05.09.1943	03.10.1989	18.09.1980	fe...	170	3.81	3.47	5.13	4.24	No	No	No	Never	9	-88	-88	.29	3302.00
22	1023	Unexp...	49	06.06.1948	21.07.1997	12.07.1982	male	165	3.32	3.17	4.68	3.87	No	No	Yes	Curr. S...	15	20	31	.00	5488.00
23	1025	Unexp...	45	09.02.1949	16.05.1994	12.05.1988	male	170	3.40	3.50	4.34	4.26	No	No	No	Never	6	-88	-88	.00	2195.00
24	1026	Expos...	46	17.04.1949	23.06.1995	25.06.1990	male	175	4.01	3.59	5.17	4.45	No	No	No	Never	5	-88	-88	.72	1824.00
25	1027	Expos...	56	10.01.1942	17.04.1998	18.03.1991	fe...	165	2.80	2.97	3.57	3.69	No	No	No	Ex Smo...	7	20	38	.74	2587.00
26	1028	Expos...	26	01.01.1970	10.01.1996	19.01.1988	fe...	172	4.37	4.14	4.58	4.87	No	No	No	Never	8	-88	-88	.88	2913.00
27	1029	Expos...	54	19.04.1934	12.09.1988	23.07.1979	fe...	170	3.63	3.24	4.51	4.03	No	Yes	No	Ex Smo...	9	20	30	1.50	3339.00
28	1030	Expos...	32	12.05.1958	28.07.1990	14.06.1983	fe...	178	4.68	4.22	5.92	5.06	Yes	No	No	Never	7	-88	-88	1.73	2601.00
29	1031	Expos...	34	20.01.1960	15.03.1994	24.01.1985	fe...	190	4.91	4.68	6.06	5.69	No	No	Yes	Curr. S...	9	12	13	1.12	3337.00
30	1032	Unexp...	50	02.01.1942	20.01.1992	01.01.1976	fe...	170	2.47	3.36	3.88	4.13	No	Yes	No	Ex Smo...	16	30	17	.00	5863.00
31	1033	Expos...	53	10.10.1942	18.11.1995	16.10.1982	male	163	2.16	2.94	3.60	3.61	No	No	Yes	Curr. S...	13	20	25	1.07	4781.00
32	1034	Unexp...	52	09.04.1945	26.05.1997	20.01.1988	male	185	3.53	3.94	4.70	4.94	No	No	No	Ex Smo...	9	40	30	.00	3414.00
33	1036	Unexp...	42	16.02.1947	02.04.1989	12.02.1977	fe...	162	3.64	3.24	4.59	3.88	No	No	Yes	Curr. S...	12	10	15	.00	4432.00
34	1037	Unexp...	34	17.01.1959	21.03.1993	28.02.1987	male	177	3.69	4.12	5.12	4.95	No	No	No	Never	6	-88	-88	.00	2213.00
35	1038	Unexp...	45	26.06.1947	19.09.1992	31.03.1983	male	170	4.31	3.50	5.50	4.26	No	No	Yes	Curr. S...	9	20	12	.00	3460.00
36	1039	Expos...	38	15.01.1953	22.01.1991	25.01.1974	fe...	170	3.98	3.72	5.11	4.46	No	No	No	Ex Smo...	17	13	4	1.12	6206.00
37	1040	Unexp...	47	09.01.1946	20.03.1993	13.01.1971	male	180	3.97	3.87	5.18	4.78	No	No	No	Ex Smo...	22	20	10	.00	8102.00
38	1042	Unexp...	24	23.07.1966	30.09.1990	24.07.1982	fe...	180	4.80	4.51	6.27	5.36	No	No	No	Never	8	-88	-88	.00	2990.00

In the following panel transfer the variables as shown (note the status is the event present vs censored variable) and **Define Event** in this cases Asthma present is equal to 1. Note, in the **Options** button click Kaplan-Meier.



Time: daysworked

Status: ASTHMA(1)

Factor: Exposure Group [GR...

Strata:

Label Cases by:

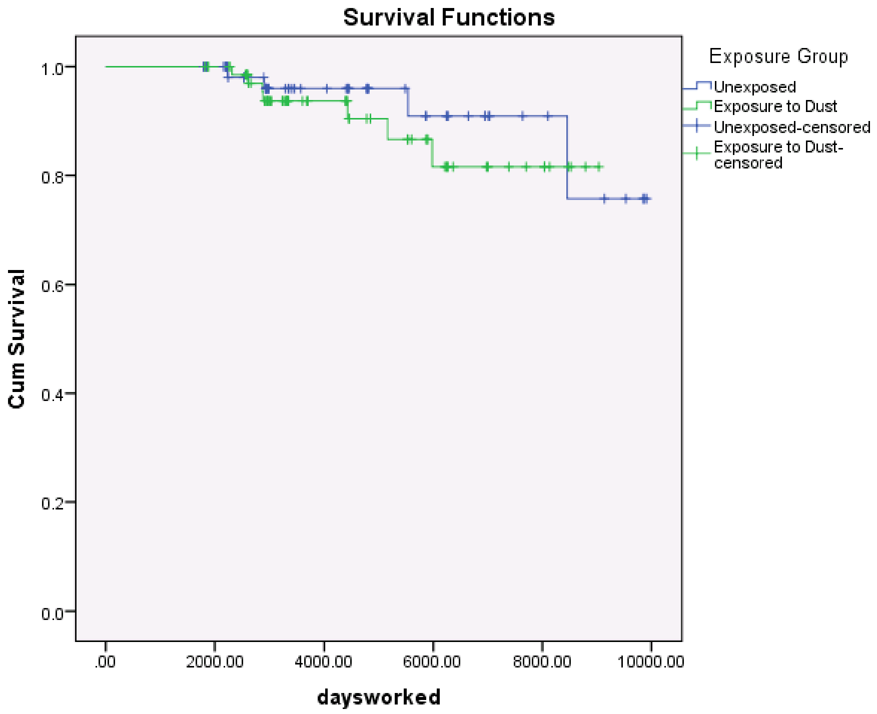
Value(s) indicating event has occurred

● Single value: 1

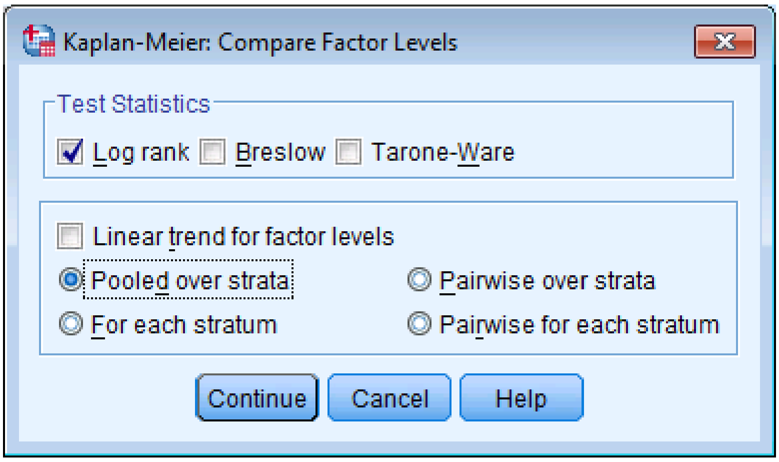
○ Range of values: through

○ List of values:

Resulting in the following Kaplan-Meier plot describing the rate at which Asthma occurs as working days increases in those subjects exposed to dust vs unexposed to dust.



This appears to indicate that those exposed to dust developed Asthma at a quicker rate than those unexposed to dust. To test the hypothesis that a significant difference is present we can perform a log-rank test. Repeat the process outlined for the Kaplan Meier (**Analysis** > **Survival > Kaplan-Meier)** except this time click **Compare Factor..** followed by **Log rank**


Kaplan-Meier: Compare Factor Levels

Test Statistics

☑ Log rank □ Breslow □ Tarone-Ware

□ Linear trend for factor levels

● Pooled over strata ○ Pairwise over strata

○ For each stratum ○ Pairwise for each stratum

Giving….

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	.475	1	.491

Test of equality of survival distributions for the different levels of Exposure Group.

Indicating (sig=0.491 i.e. greater than 0.05) that we fail to find evidence to reject the null hypothesis that there is no significant difference in the rate of asthma diagnosis in the exposed to dust vs unexposed to dust.

Video Tutorial – Kaplan-Meier plot and Log rank test

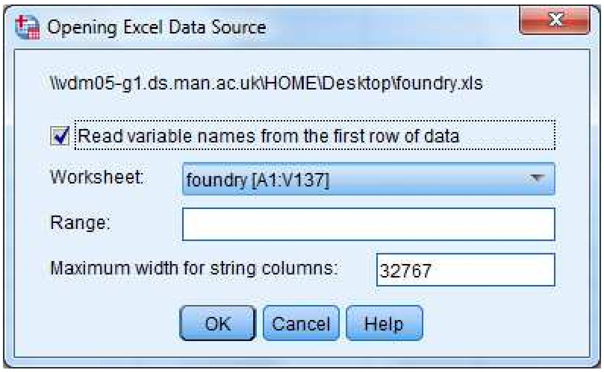
https://www.youtube.com/watch?v=Tw1WVxiXHsk

Exercise Repeat the Kaplan Meier and Log Rank test to determine if exposure to dust increased the rate at which Bronchitis was diagnosed.

READING AN EXCEL FILE INTO SPSS

It may be the case that data may already be stored in another data format. **SPSS** can read many of these. For example you can retrieve an **Excel** file into **SPSS**. If you put the variable names in the first row of your spreadsheet, they can be copied as variable names in **SPSS** file. Unlike Stats Direct, SPSS is only able to read a single work sheet it cannot read a complete work book with several sheets. In order that **SPSS** can read it, the **Excel** file needs to be saved in the version 4 format.

The data from the foundry study is saved in a spreadsheet located on **the shared area**. The names of the variables have been entered in the first row. You may wish to check this by going to **EXCEL**. The procedure for retrieving the data from **EXCEL** is similar to retrieving an **SPSS** data file. Click on the **File** option at the top of the screen, then on the **Open** sub-option followed by the **Data** option so that the screen above appears. At this point change the file type to **Excel** and **Open** the spreadsheet named **foundry**. The following screen should appear.


Opening Excel Data Source

\\wdm05-g1.ds.man.ac.uk\HOME\Desktop\foundry.xls

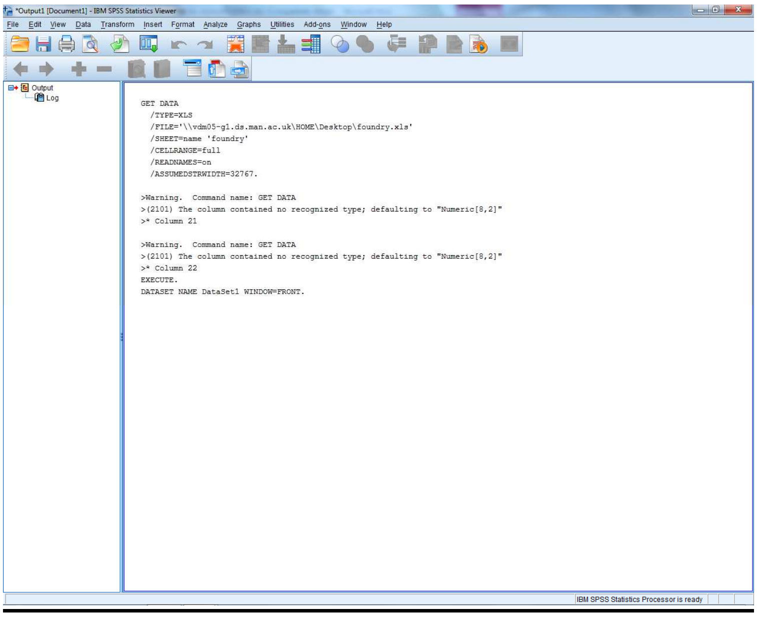
☑ Read variable names from the first row of data

Worksheet: foundry [A1:V137]

Range:

Maximum width for string columns: 32767

Unless there is other data on the spreadsheet that we do not want to read we need not specify a range. As we want to read the variable names, you have to click **Read variable names** button then press **OK**. You may get an output window with a warning or explaining variable names, types and their formats similar to this.


*Output1 [Document1] - IBM SPSS Statistics Viewer

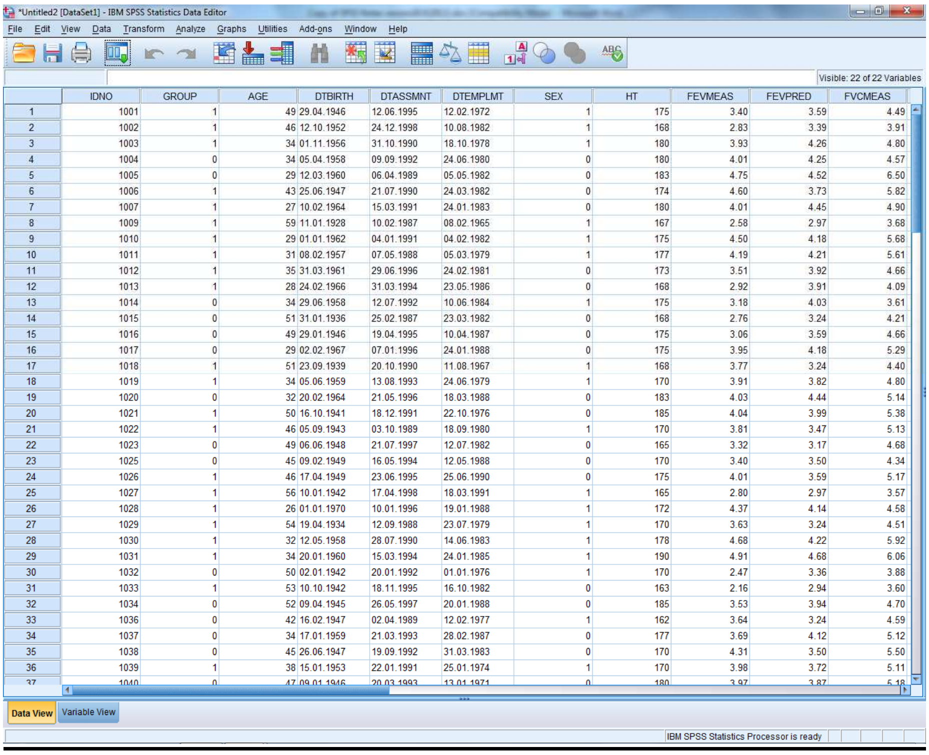
File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help
Output
Log
IBM SPSS Statistics Processor is ready

```
GET DATA
  /TYPE=XLS
  /FILE='\\vdm05-g1.ds.man.ac.uk\HOME\Desktop\foundry.xls'
  /SHEET=name 'foundry'
  /CELLRANGE=full
  /READNAMES=on
  /ASSUMEDSTRWIDTH=32767.

>Warning.  Command name: GET DATA
>(2101) The column contained no recognized type; defaulting to "Numeric[8,2]"
>* Column 21

>Warning.  Command name: GET DATA
>(2101) The column contained no recognized type; defaulting to "Numeric[8,2]"
>* Column 22
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.
```

If you switch over to **Data editor** screen by clicking **Window** option on the menu bar or by using the button on the status bar at the bottom of the screen, you will be able to see the variable names and values in their proper columns. Now all the Foundry data has been read from the spreadsheet. If we want to add variable labels and value labels we would need to go to **variable view**.


*Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 22 of 22 Variables

	IDNO	GROUP	AGE	DTBIRTH	DTASSMNT	DTEMPLMT	SEX	HT	FEVMEAS	FEVPRED	FVCMEAS
1	1001	1	49	29.04.1946	12.06.1995	12.02.1972	1	175	3.40	3.59	4.49
2	1002	1	46	12.10.1952	24.12.1998	10.08.1982	1	168	2.83	3.39	3.91
3	1003	1	34	01.11.1956	31.10.1990	18.10.1978	1	180	3.93	4.26	4.80
4	1004	0	34	05.04.1958	09.09.1992	24.06.1980	0	180	4.01	4.25	4.57
5	1005	0	29	12.03.1960	06.04.1989	05.05.1982	0	183	4.75	4.52	6.50
6	1006	1	43	25.06.1947	21.07.1990	24.03.1982	0	174	4.60	3.73	5.82
7	1007	1	27	10.02.1964	15.03.1991	24.01.1983	0	180	4.01	4.45	4.90
8	1009	1	59	11.01.1928	10.02.1987	08.02.1965	1	167	2.58	2.97	3.68
9	1010	1	29	01.01.1962	04.01.1991	04.02.1982	1	175	4.50	4.18	5.68
10	1011	1	31	08.02.1957	07.05.1988	05.03.1979	1	177	4.19	4.21	5.61
11	1012	1	35	31.03.1961	29.06.1996	24.02.1981	0	173	3.51	3.92	4.66
12	1013	1	28	24.02.1966	31.03.1994	23.05.1986	0	168	2.92	3.91	4.09
13	1014	0	34	29.06.1958	12.07.1992	10.06.1984	1	175	3.18	4.03	3.61
14	1015	0	51	31.01.1936	25.02.1987	23.03.1982	0	168	2.76	3.24	4.21
15	1016	0	49	29.01.1946	19.04.1995	10.04.1987	0	175	3.06	3.59	4.66
16	1017	0	29	02.02.1967	07.01.1996	24.01.1988	0	175	3.95	4.18	5.29
17	1018	1	51	23.09.1939	20.10.1990	11.08.1967	1	168	3.77	3.24	4.40
18	1019	1	34	05.06.1959	13.08.1993	24.06.1979	1	170	3.91	3.82	4.80
19	1020	0	32	20.02.1964	21.05.1996	18.03.1988	0	183	4.03	4.44	5.14
20	1021	1	50	16.10.1941	18.12.1991	22.10.1976	0	185	4.04	3.99	5.38
21	1022	1	46	05.09.1943	03.10.1989	18.09.1980	1	170	3.81	3.47	5.13
22	1023	0	49	06.06.1948	21.07.1997	12.07.1982	0	165	3.32	3.17	4.68
23	1025	0	45	09.02.1949	16.05.1994	12.05.1988	0	170	3.40	3.50	4.34
24	1026	1	46	17.04.1949	23.06.1995	25.06.1990	0	175	4.01	3.59	5.17
25	1027	1	56	10.01.1942	17.04.1998	18.03.1991	1	165	2.80	2.97	3.57
26	1028	1	26	01.01.1970	10.01.1996	19.01.1988	1	172	4.37	4.14	4.58
27	1029	1	54	19.04.1934	12.09.1988	23.07.1979	1	170	3.63	3.24	4.51
28	1030	1	32	12.05.1958	28.07.1990	14.06.1983	1	178	4.68	4.22	5.92
29	1031	1	34	20.01.1960	15.03.1994	24.01.1985	1	190	4.91	4.68	6.06
30	1032	0	50	02.01.1942	20.01.1992	01.01.1976	1	170	2.47	3.36	3.88
31	1033	1	53	10.10.1942	18.11.1995	16.10.1982	0	163	2.16	2.94	3.60
32	1034	0	52	09.04.1945	26.05.1997	20.01.1988	0	185	3.53	3.94	4.70
33	1036	0	42	16.02.1947	02.04.1989	12.02.1977	1	162	3.64	3.24	4.59
34	1037	0	34	17.01.1959	21.03.1993	28.02.1987	0	177	3.69	4.12	5.12
35	1038	0	45	26.06.1947	19.09.1992	31.03.1983	0	170	4.31	3.50	5.50
36	1039	1	38	15.01.1953	22.01.1991	25.01.1974	1	170	3.98	3.72	5.11
37	1040	0	47	09.01.1946	20.03.1993	13.01.1971	0	180	3.97	3.87	5.18

If you don't have variable names in the **Excel** file then when retrieving it into **SPSS** file you should not click **Read variable names** button, just press **OK** button and you get the following screen.

You then have to define the variable names by clicking the **Variable View** as described above.

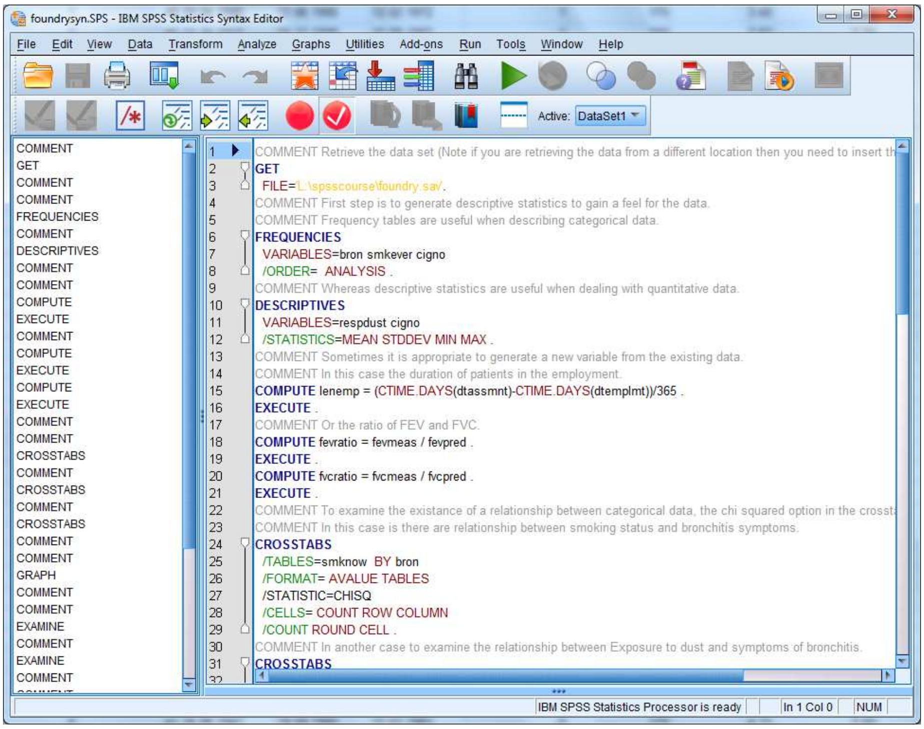
Having read data from an excel spreadsheet it is important to check what has been read in and amend each of the variable properties in the variable view window. For example if a column on the spreadsheet contained a mix of numeric and string data (besides the variable name at the top) either one or the other may be set to missing.

CREATING A SPSS SYNTAX

To date we have used SPSS interactively, an alternative is to create an SPSS syntax file containing the commands. There are two reasons for this: -

- It makes it easier and quicker to rerun an analysis if we make changes to the raw data.
- It documents the analysis that we have performed.

The screen shot below illustrates part of the syntax file for the analysis that we have done.

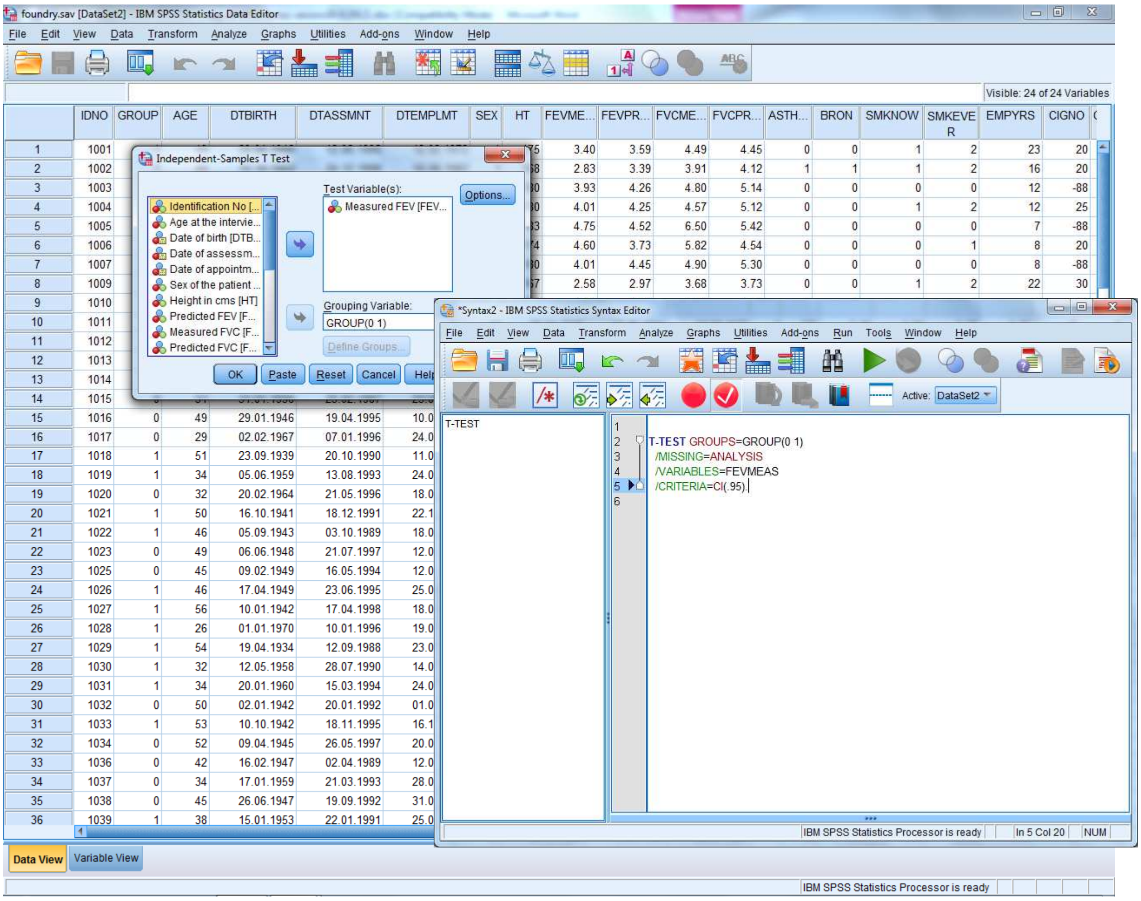


foundrysyn.SPS - IBM SPSS Statistics Syntax Editor

```
COMMENT
GET
COMMENT
COMMENT
FREQUENCIES
COMMENT
DESCRIPTIVES
COMMENT
COMMENT
COMPUTE
EXECUTE
COMMENT
COMPUTE
EXECUTE
COMPUTE
EXECUTE
COMMENT
COMMENT
CROSSTABS
COMMENT
CROSSTABS
COMMENT
CROSSTABS
COMMENT
COMMENT
GRAPH
COMMENT
COMMENT
EXAMINE
COMMENT
EXAMINE
COMMENT
```

```
COMMENT Retrieve the data set (Note if you are retrieving the data from a different location then you need to insert th
GET
  FILE='L:\spsscourse\foundry.sav'.
COMMENT First step is to generate descriptive statistics to gain a feel for the data.
COMMENT Frequency tables are useful when describing categorical data.
FREQUENCIES
  VARIABLES=bron smkever cigno
  /ORDER=  ANALYSIS .
COMMENT Whereas descriptive statistics are useful when dealing with quantitative data.
DESCRIPTIVES
  VARIABLES=respdust cigno
  /STATISTICS=MEAN STDDEV MIN MAX .
COMMENT Sometimes it is appropriate to generate a new variable from the existing data.
COMMENT In this case the duration of patients in the employment.
COMPUTE lenemp = (CTIME.DAYS(dtassmnt)-CTIME.DAYS(dtemplmt))/365 .
EXECUTE .
COMMENT Or the ratio of FEV and FVC.
COMPUTE fevratio = fevmeas / fevpred .
EXECUTE .
COMPUTE fvcratio = fvcmeas / fvcpred .
EXECUTE .
COMMENT To examine the existance of a relationship between categorical data, the chi squared option in the crossta
COMMENT In this case is there are relationship between smoking status and bronchitis symptoms.
CROSSTABS
  /TABLES=smknow  BY bron
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ
  /CELLS= COUNT ROW COLUMN
  /COUNT ROUND CELL .
COMMENT In another case to examine the relationship between Exposure to dust and symptoms of bronchitis.
CROSSTABS
```

This looks complicated but we do not need to learn this because SPSS can do this for us using the menus. You may of notices a button **paste** on the interactive commands. We will illustrate this using the t-test command. If we click **paste** instead of running the command then the syntax is pasted into a new file.



foundry.sav [DataSet2] - IBM SPSS Statistics Data Editor

Visible: 24 of 24 Variables

	IDNO	GROUP	AGE	DTBIRTH	DTASSMNT	DTEMPLMT	SEX	HT	FEVME...	FEVPR...	FVCME...	FVCPR...	ASTH...	BRON	SMKNOW	SMKEVER	EMPYRS	CIGNO
1	1001							75	3.40	3.59	4.49	4.45	0	0	1	2	23	20
2	1002							68	2.83	3.39	3.91	4.12	1	1	1	2	16	20
3	1003							30	3.93	4.26	4.80	5.14	0	0	0	0	12	-88
4	1004							30	4.01	4.25	4.57	5.12	0	0	1	2	12	25
5	1005							33	4.75	4.52	6.50	5.42	0	0	0	0	7	-88
6	1006							74	4.60	3.73	5.82	4.54	0	0	0	1	8	20
7	1007							30	4.01	4.45	4.90	5.30	0	0	0	0	8	-88
8	1009							67	2.58	2.97	3.68	3.73	0	0	1	2	22	30
9	1010																	
10	1011																	
11	1012																	
12	1013																	
13	1014																	
14	1015																	
15	1016	0	49	29.01.1946	19.04.1995	10.0												
16	1017	0	29	02.02.1967	07.01.1996	24.0												
17	1018	1	51	23.09.1939	20.10.1990	11.0												
18	1019	1	34	05.06.1959	13.08.1993	24.0												
19	1020	0	32	20.02.1964	21.05.1996	18.0												
20	1021	1	50	16.10.1941	18.12.1991	22.1												
21	1022	1	46	05.09.1943	03.10.1989	18.0												
22	1023	0	49	06.06.1948	21.07.1997	12.0												
23	1025	0	45	09.02.1949	16.05.1994	12.0												
24	1026	1	46	17.04.1949	23.06.1995	25.0												
25	1027	1	56	10.01.1942	17.04.1998	18.0												
26	1028	1	26	01.01.1970	10.01.1996	19.0												
27	1029	1	54	19.04.1934	12.09.1988	23.0												
28	1030	1	32	12.05.1958	28.07.1990	14.0												
29	1031	1	34	20.01.1960	15.03.1994	24.0												
30	1032	0	50	02.01.1942	20.01.1992	01.0												
31	1033	1	53	10.10.1942	18.11.1995	16.1												
32	1034	0	52	09.04.1945	26.05.1997	20.0												
33	1036	0	42	16.02.1947	02.04.1989	12.0												
34	1037	0	34	17.01.1959	21.03.1993	28.0												
35	1038	0	45	26.06.1947	19.09.1992	31.0												
36	1039	1	38	15.01.1953	22.01.1991	25.0												

Independent-Samples T Test

Test Variable(s): Measured FEV [FEV...

Grouping Variable: GROUP(0 1)

*Syntax2 - IBM SPSS Statistics Syntax Editor

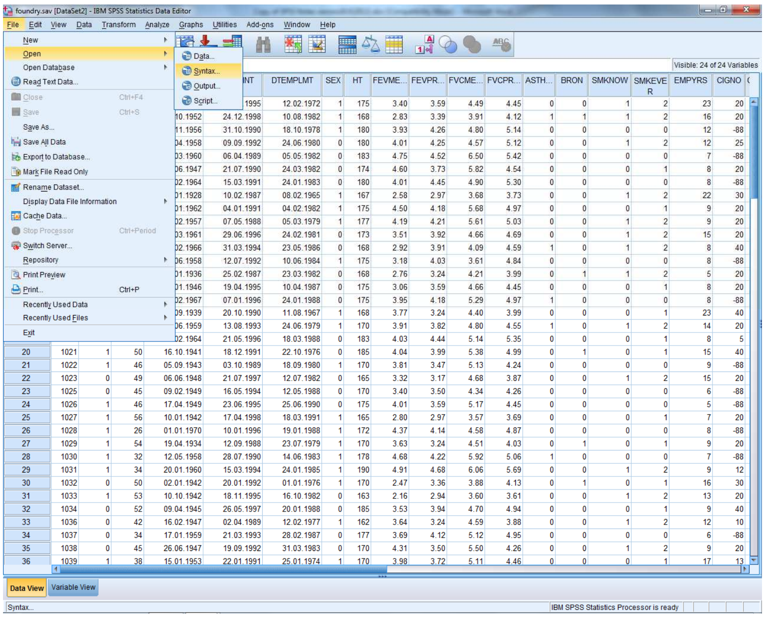
```
T-TEST
```

```
T-TEST GROUPS=GROUP(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=FEVMEAS
  /CRITERIA=CI(.95).
```

The first time in a session that you click paste a new file is created. Using the same method as for the t-test above you can add further commands to the syntax. It is possible to run the entire syntax all at once or alternatively only specific commands.

To run the entire syntax click **run** on the options bar followed by **all**. To run a specific command, highlight the command in the main window and click **run** followed by **selection**. A Syntax can be edited through copy and paste commands or alternatively through more detailed written commands, described in the help file. Because the syntax file is a separate file from SPSS needs to be saved separately at the end of the session, using **File** and **Save**. At the start of a new session, you can reopen an existing syntax file.

Please find located along with the data set on the **Shared Data area** and/or the **website** address a syntax that will give the appropriate output for the SPSS exercises within the notes.



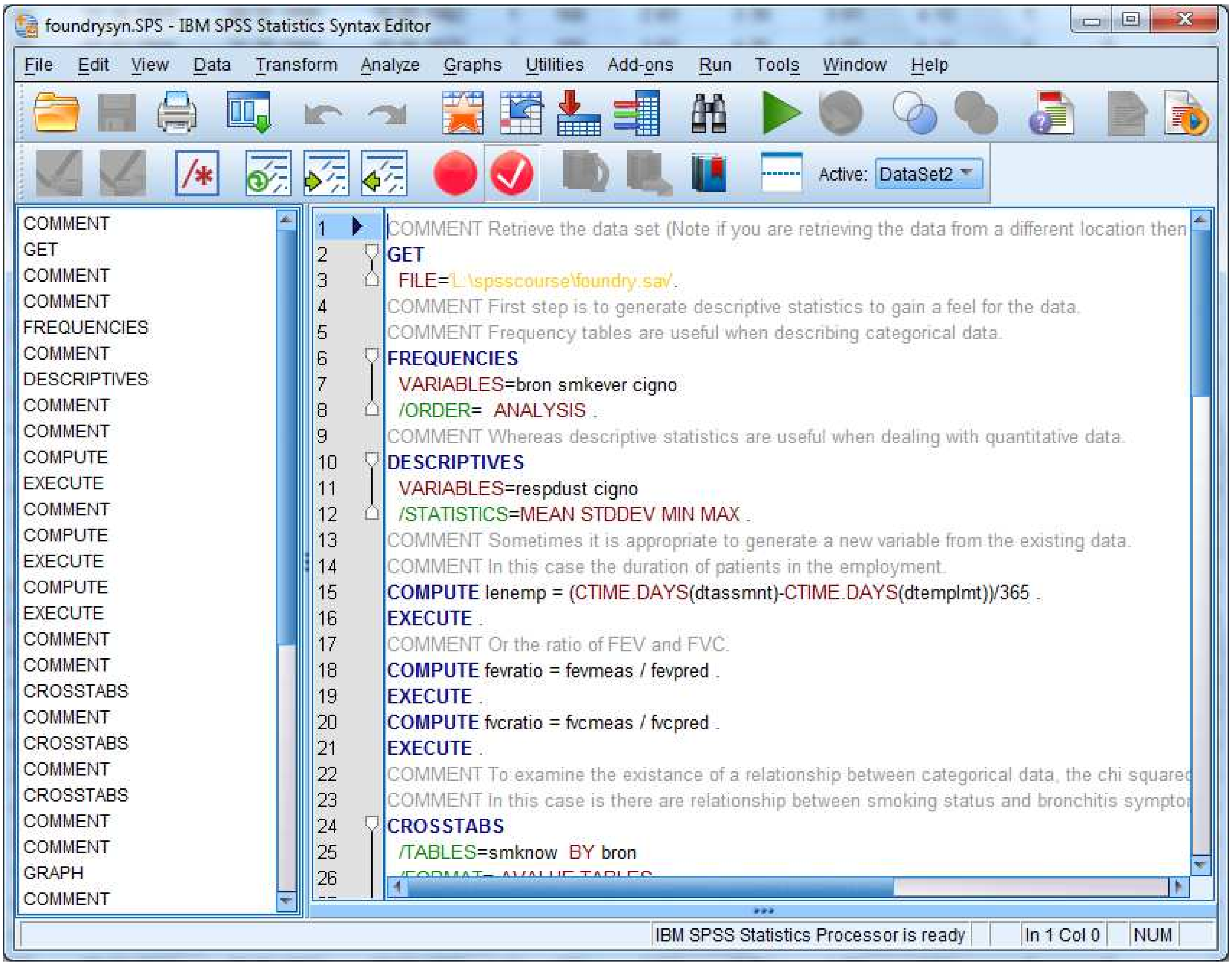
					INT	DTEMPLMT	SEX	HT	FEVME...	FEVPR...	FVCME...	FVCPR...	ASTH...	BRON	SMKNOW	SMKEVER	EMPYRS	CIGNO
					1995	12.02.1972	1	175	3.40	3.59	4.49	4.45	0	0	1	2	23	20
				10.1952	24.12.1998	10.08.1982	1	168	2.83	3.39	3.91	4.12	1	1	1	2	16	20
				11.1956	31.10.1990	18.10.1978	1	180	3.93	4.26	4.80	5.14	0	0	0	0	12	-88
				04.1958	09.09.1992	24.06.1980	0	180	4.01	4.25	4.57	5.12	0	0	1	2	12	25
				03.1960	06.04.1989	05.05.1982	0	183	4.75	4.52	6.50	5.42	0	0	0	0	7	-88
				06.1947	21.07.1990	24.03.1982	0	174	4.60	3.73	5.82	4.54	0	0	0	1	8	20
				02.1964	15.03.1991	24.01.1983	0	180	4.01	4.45	4.90	5.30	0	0	0	0	8	-88
				01.1928	10.02.1987	08.02.1965	1	167	2.58	2.97	3.68	3.73	0	0	1	2	22	30
				01.1962	04.01.1991	04.02.1982	1	175	4.50	4.18	5.68	4.97	0	0	0	1	9	20
				02.1957	07.05.1988	05.03.1979	1	177	4.19	4.21	5.61	5.03	0	0	1	2	9	20
				03.1961	29.06.1996	24.02.1981	0	173	3.51	3.92	4.66	4.69	0	0	1	2	15	20
				02.1966	31.03.1994	23.05.1986	0	168	2.92	3.91	4.09	4.59	1	0	1	2	8	40
				06.1958	12.07.1992	10.06.1984	1	175	3.18	4.03	3.61	4.84	0	0	0	0	8	-88
				01.1936	25.02.1987	23.03.1982	0	168	2.76	3.24	4.21	3.99	0	1	1	2	5	20
				01.1946	19.04.1995	10.04.1987	0	175	3.06	3.59	4.66	4.45	0	0	0	1	8	20
				02.1967	07.01.1996	24.01.1988	0	175	3.95	4.18	5.29	4.97	1	0	0	0	8	-88
				09.1939	20.10.1990	11.08.1967	1	168	3.77	3.24	4.40	3.99	0	0	0	1	23	40
				06.1959	13.08.1993	24.06.1979	1	170	3.91	3.82	4.80	4.55	1	0	1	2	14	20
				02.1964	21.05.1996	18.03.1988	0	183	4.03	4.44	5.14	5.35	0	0	0	1	8	5
20	1021	1	50	16.10.1941	18.12.1991	22.10.1976	0	185	4.04	3.99	5.38	4.99	0	1	0	1	15	40
21	1022	1	46	05.09.1943	03.10.1989	18.09.1980	1	170	3.81	3.47	5.13	4.24	0	0	0	0	9	-88
22	1023	0	49	06.06.1948	21.07.1997	12.07.1982	0	165	3.32	3.17	4.68	3.87	0	0	1	2	15	20
23	1025	0	45	09.02.1949	16.05.1994	12.05.1988	0	170	3.40	3.50	4.34	4.26	0	0	0	0	6	-88
24	1026	1	46	17.04.1949	23.06.1995	25.06.1990	0	175	4.01	3.59	5.17	4.45	0	0	0	0	5	-88
25	1027	1	56	10.01.1942	17.04.1998	18.03.1991	1	165	2.80	2.97	3.57	3.69	0	0	0	1	7	20
26	1028	1	26	01.01.1970	10.01.1996	19.01.1988	1	172	4.37	4.14	4.58	4.87	0	0	0	0	8	-88
27	1029	1	54	19.04.1934	12.09.1988	23.07.1979	1	170	3.63	3.24	4.51	4.03	0	1	0	1	9	20
28	1030	1	32	12.05.1958	28.07.1990	14.06.1983	1	178	4.68	4.22	5.92	5.06	1	0	0	0	7	-88
29	1031	1	34	20.01.1960	15.03.1994	24.01.1985	1	190	4.91	4.68	6.06	5.69	0	0	1	2	9	12
30	1032	0	50	02.01.1942	20.01.1992	01.01.1976	1	170	2.47	3.36	3.88	4.13	0	1	0	1	16	30
31	1033	1	53	10.10.1942	18.11.1995	16.10.1982	0	163	2.16	2.94	3.60	3.61	0	0	1	2	13	20
32	1034	0	52	09.04.1945	26.05.1997	20.01.1988	0	185	3.53	3.94	4.70	4.94	0	0	0	1	9	40
33	1036	0	42	16.02.1947	02.04.1989	12.02.1977	1	162	3.64	3.24	4.59	3.88	0	0	1	2	12	10
34	1037	0	34	17.01.1959	21.03.1993	28.02.1987	0	177	3.69	4.12	5.12	4.95	0	0	0	0	6	-88
35	1038	0	45	26.06.1947	19.09.1992	31.03.1983	0	170	4.31	3.50	5.50	4.26	0	0	1	2	9	20
36	1039	1	38	15.01.1953	22.01.1991	25.01.1974	1	170	3.98	3.72	5.11	4.46	0	0	0	1	17	13

By clicking **File**, **Open** and **Syntax** as shown above the following screen will appear.

Locate the appropriate file named **foundrysyn.sps** and click open.

The file can then be run in the same manner described on the previous page.

The syntax contains a set of codes and appropriate descriptions to produce the appropriate output for each exercise (page 18 onwards) through out the notes. The descriptions of the tests performed can be found in the syntax script after COMMENT.



```
COMMENT
GET
COMMENT
COMMENT
FREQUENCIES
COMMENT
DESCRIPTIVES
COMMENT
COMMENT
COMPUTE
EXECUTE
COMMENT
COMPUTE
EXECUTE
COMPUTE
EXECUTE
COMMENT
COMMENT
CROSSTABS
COMMENT
CROSSTABS
COMMENT
CROSSTABS
COMMENT
COMMENT
GRAPH
COMMENT
```

```
COMMENT Retrieve the data set (Note if you are retrieving the data from a different location then
GET
  FILE='L:\spsscourse\foundry.sav'.
COMMENT First step is to generate descriptive statistics to gain a feel for the data.
COMMENT Frequency tables are useful when describing categorical data.
FREQUENCIES
  VARIABLES=bron smkever cigno
  /ORDER=  ANALYSIS .
COMMENT Whereas descriptive statistics are useful when dealing with quantitative data.
DESCRIPTIVES
  VARIABLES=respdust cigno
  /STATISTICS=MEAN STDDEV MIN MAX .
COMMENT Sometimes it is appropriate to generate a new variable from the existing data.
COMMENT In this case the duration of patients in the employment.
COMPUTE lenemp = (CTIME.DAYS(dtassmnt)-CTIME.DAYS(dtemplmt))/365 .
EXECUTE .
COMMENT Or the ratio of FEV and FVC.
COMPUTE fevratio = fevmeas / fevpred .
EXECUTE .
COMPUTE fvcratio = fvcmeas / fvcpred .
EXECUTE .
COMMENT To examine the existance of a relationship between categorical data, the chi squared
COMMENT In this case is there are relationship between smoking status and bronchitis symptor
CROSSTABS
  /TABLES=smknow  BY bron
  /FORMAT= AVALUE TABLES
```

IBM SPSS Statistics Processor is ready | In 1 Col 0 | NUM

Please note the first comment regarding the retrieval of the dataset.

CHOOSING THE APPROPRIATE STATISTICAL PROCEDURE

In this tutorial we have illustrated some of the basic statistical procedures available in SPSS. These are summarised in the table below.

	Plausibly Continuous and Normal	**Ordinal or Ordered Categorical**	**Binary and Unordered Categories**
Comparison of Independent Two Groups	Box-plot Independent groups t-test	Box-plot or Cross-tabulation of ordered categories Mann-Whitney U-test	Cross-tabulation Chi-squared test Fisher's exact test
Comparison of more than Two groups	Analysis of variance (ANOVA)	*Kruskal Wallis analysis of Variance**	Cross-tabulation Chi-squared test
Comparison of two related outcomes	Paired samples t-test	Wilcoxon Matched Pairs	McNemar's Test
Relationship between a dependent variable and one or more independent variables	Scatter plot Regression *Pearson's correlation coefficient*	*Spearman correlation or Kendall's correlation coefficient*	*Phi coefficient* *Logistic Regression*

* Not illustrated

For a more comprehensive chart for selecting methods see www.graphpad.com/www/book/choose.htm. We conclude by noting that SPSS has some serious weaknesses for analysis of medical data. For example many of the methods give only p-values and no confidence interval. For example the Mann-Whitney U-Test is a comparison of two medians but it does not give the confidence interval for the difference of the two medians as recommended in many guidelines for medical research publication. In this aspect the program **StatsDirect**, also available in the Micro-labs, is much better as the corresponding procedure gives a confidence interval of the difference between medians.