

# INTRODUCTION TO STATSDIRECT

PART 1 .....	2
INTRODUCTION .....	2
Why Use StatsDirect .....	2
ACCESSING STATSDIRECT FOR WINDOWS XP .....	4
DATA ENTRY .....	5
Missing Data .....	6
Opening an Excel Workbook .....	6
Moving around Data Editor .....	8
INTRODUCING ANALYSIS COMMANDS .....	8
Descriptive Statistics .....	8
Analysis By Column Or By Identifier .....	8
Frequency Tables .....	9
Descriptives .....	11
Cross-tabulation, Chi-squared Test and Fisher's Exact Test .....	13
FILE HANDLING .....	17
Saving The Work .....	17
Backup The Work .....	18
Retrieving a StatsDirect File .....	18
MODIFYING THE DATA .....	19
Modifying a new variable by Search and Replace .....	19
PART II .....	22
CONTINUOUS OUTCOME MEASURES .....	22
Histogram .....	22
Comparison of Means Using a t-test .....	25
REARRANGING DATA .....	26
Sorting Data .....	26
Group Split .....	27
EXAMINING THE RELATIONSHIP BETWEEN TWO CONTINUOUS VARIABLES .....	29
Scatter Plots .....	29
Linear Regressions .....	30
Model Checking .....	32
NON-PARAMETRIC METHODS .....	34
COMPARISONS OF RELATED OR PAIRED VARIABLES .....	36
Continuous Outcome Measures .....	36
Analysis of Binary Outcomes that are Related .....	37
Related Ordinal Data .....	38
SUMMARY STATISTIC METHODS .....	39
t-test Using summary Data .....	39
Comparison of Proportions .....	41
CHOOSING THE APPROPRIATE STATISTICAL PROCEDURE .....	42

# PART 1

## INTRODUCTION

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

### Why Use StatsDirect

**StatsDirect** is an easy-to-use package designed for medical researchers. It uses an interface similar to Micro-soft Excel and is able to read **EXCEL** Workbooks. It has some procedures specifically designed for medical researchers. **StatsDirect** has an extensive **Help** Facility that can be accessed by clicking on the local Help Tile. The help provided depends on what part of the system is being used.

StatsDirect is not suitable for analysis of large data sets with many variables and missing data. If your dataset is large and has missing data and many variables, as you might have with a survey or data base, you may find it easier to use SPSS or STATA. For example in SPSS it is easier to select subsets of data for a particular analysis. In SPSS variables can be labelled and text labels can be added to numerical codes making statistical output much more readable. This may be important when analysing large data sets.

This tutorial will assume some familiarity with concepts in statistical inference including hypothesis testing and confidence intervals. This was covered in years 1 and 2 of the medical undergraduate curriculum as part of the Information stream of the course. If you are unfamiliar with these concepts, it is suggested that you read an introductory text in medical statistics such as Campbell and Machin “Medical Statistics A Common Sense Approach”. Some examples are given at the Medical Statistics support web site at

<http://research.bmh.manchester.ac.uk/biostatistics/teaching/statisticalsupport>

StatsDirect has different type of analysis command

- Commands are based on raw data of individual subjects. Data needs to be entered into a spread-sheet with each row representing each subject. This might be into an Excel spreadsheet that is opened by StatsDirect or it may be entered directly into the StatsDirect

spreadsheet. We advise the former as you can enter your data, even when you have not got access to StatsDirect.

- Commands based on summary statistics. For this summary statistics data such as frequencies, mean and standard deviation are entered into a separate panel associated with that analysis. For example suppose one is comparing mean outcome for two groups of patients using a t-test. If one already has the mean standard deviation and number of patients for each group of patients, just these six pieces of information can be entered. You can also use StatsDirect to calculate confidence intervals using proportions/ percentages using frequency data. This is very useful where you do not have access to the original data for example where we are extracting data from published research.
- StatsDirect also has commands for meta-analysis, which is the statistical method used to combine summary data from several studies to obtain an overall result.

This tutorial will focus on analysis based on the first type of data. Nevertheless you may wish to use the second two forms of analysis particularly if you want to use data from published research in your study. A brief introduction is given for summary statistic methods at the end of the tutorial.

No statistical package is completely comprehensive. For example you may decide to use SPSS for the analysis of your raw data, but you might use the summary statistics methods for comparison of your results with other.

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

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 level 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, smoking history, and consumption and duration of smoking.

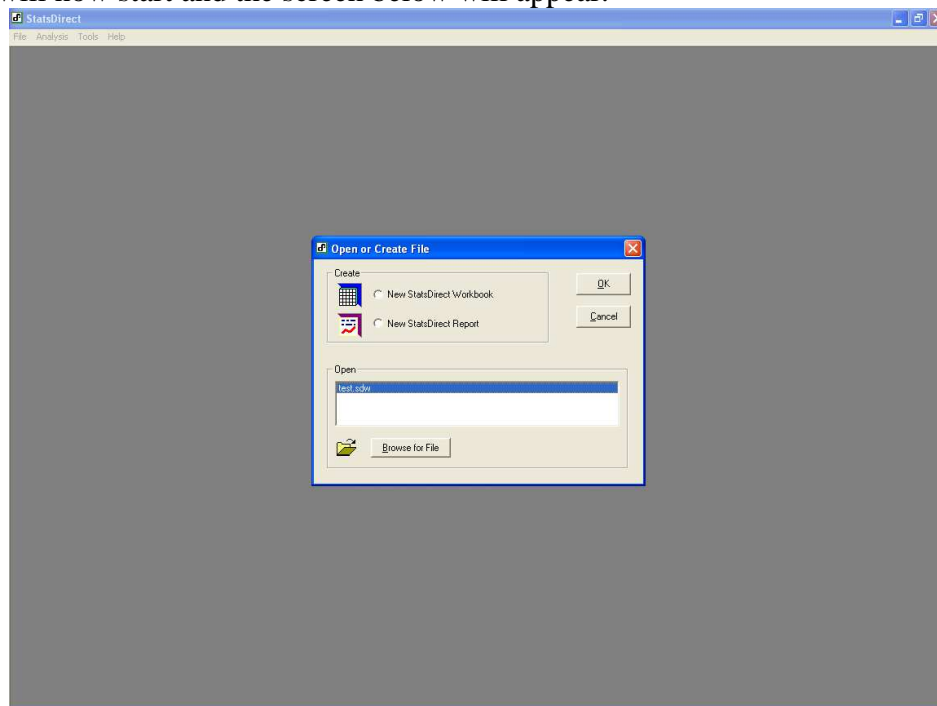
To make the most of this tutorial, work through generating the screens shown at each stage. There are several short exercises in the text to allow you to reinforce your skills.

## ACCESSING STATSDIRECT FOR WINDOWS XP

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

**2009 → MHS → Cluster → StatsDirect.**

**StatsDirect** will now start and the screen below will appear.



There are two sections:

**Create** which has two radio buttons:

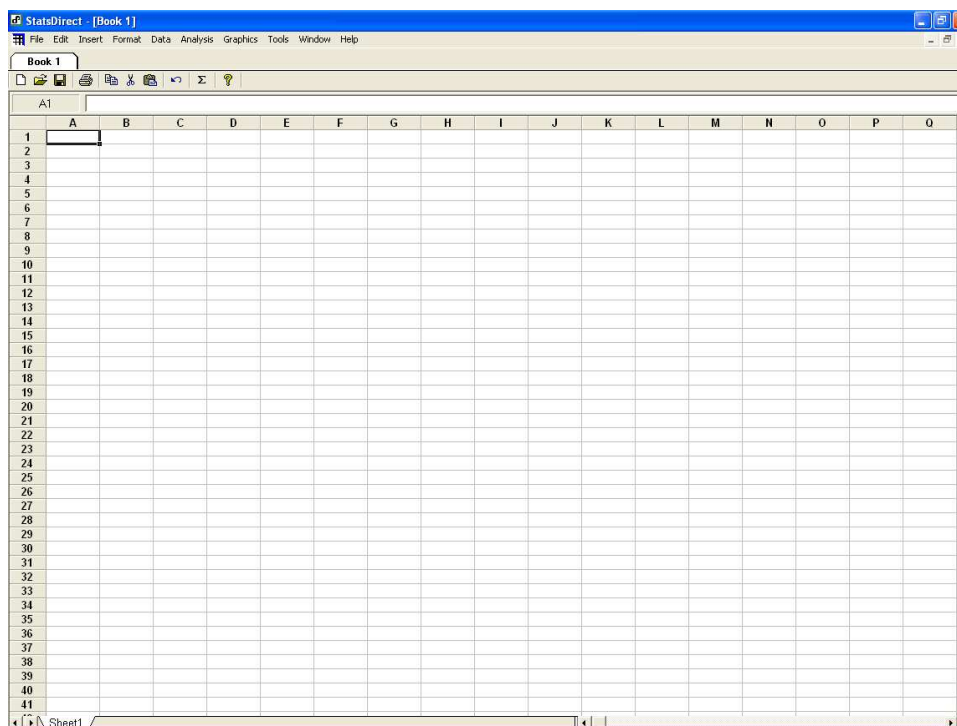
- **New StatsDirect Workbook**
- **New StatsDirect Report**

**Open** which displays data files recently accessed by **StatsDirect** from which you may select a file.

Once you have made your choice clicking the **OK** will complete start-up process. The default will open the file highlighted in the Open window.

The file **test.sdw** which should be visible in the window is a work book that contains example data prepared by the developers of **StatsDirect** to illustrate the analysis commands of the program.

For the purpose of this tutorial you should select **New StatsDirect Workbook** button and click **OK** button. A spreadsheet (**Book 1** highlighted) will be displayed for data entry as shown below



This has the following features at the top:

**Title bar** Displays **StatsDirect** and the name of the active document

**Menu bar** Contains a list of menus

**Document Tabs** Contains a list of worksheets and reports available with current one highlighted. If you have more than one, then you can move between them by clicking on the tab.

**Tool bar** Provides quick access with the mouse to frequently used tools and commands for formatting text and numbers.

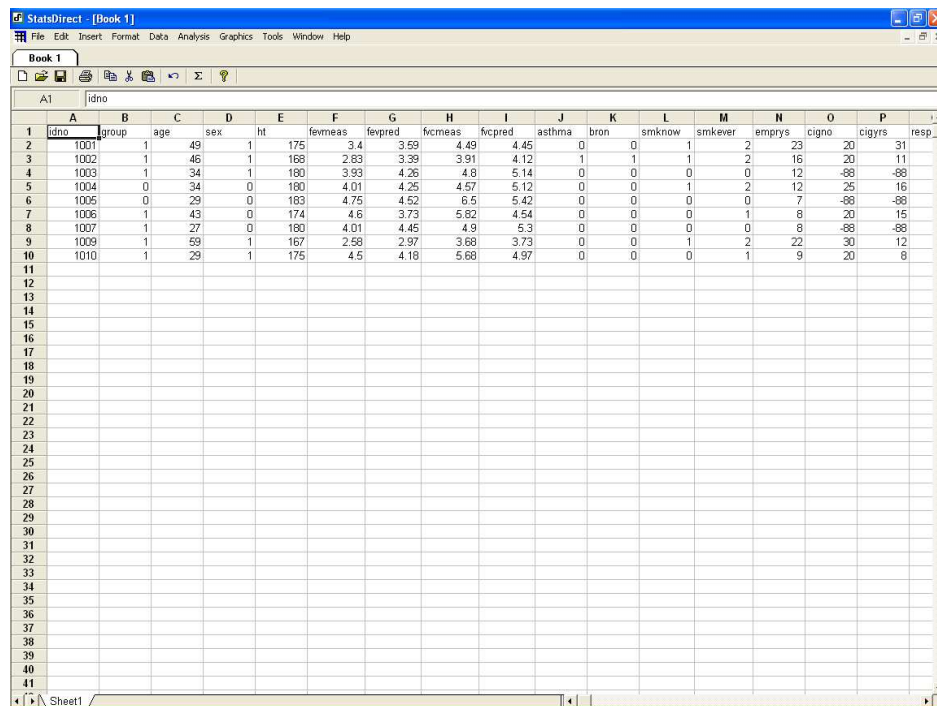
**Formula bar** Contains a cell reference and value

Below the spreadsheet grid at the bottom there is a tab to select the work sheet within the work book.

## DATA ENTRY

Data can be entered directly into **StatsDirect**. Alternatively you can enter the data into an EXCEL spreadsheet and then read it into StatsDirect. Enter the data for each subject as a row of the spreadsheet and enter the data for each variable (question or measurement). Add a label for each variable the first row of each column. In case you should need to transfer the data to another statistical package such as SPSS, it is suggested that the label naming each variable should be alpha-numeric without punctuation as illustrated below. The data should be anonymous, but there should be a unique identifying number so that you can check your data against paper records.

After inputting the first 10 subjects from the foundry workers study the screen might look like that given below.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
	idno	group	age	sex	ht	fevmeas	fevpred	ficmeas	ficpred	asthma	bron	smknow	smkever	emprys	cigno	cigyr	resp
1	1001	1	49	1	175	3.4	3.59	4.49	4.45	0	0	1	2	23	20	31	
2	1002	1	46	1	168	2.83	3.39	3.91	4.12	1	1	1	2	16	20	11	
3	1003	1	34	1	180	3.93	4.26	4.8	5.14	0	0	0	0	12	-88	-88	
4	1004	0	34	0	180	4.01	4.25	4.57	5.12	0	0	1	2	12	25	16	
5	1005	0	29	0	183	4.75	4.52	6.5	5.42	0	0	0	0	7	-88	-88	
6	1006	1	43	0	174	4.6	3.73	5.62	4.54	0	0	0	1	8	20	15	
7	1007	1	27	0	180	4.01	4.45	4.9	5.3	0	0	0	0	8	-88	-88	
8	1009	1	59	1	167	2.58	2.97	3.68	3.73	0	0	1	2	22	30	12	
9	1010	1	29	1	175	4.5	4.18	5.68	4.97	0	0	0	1	9	20	8	

The first row of the spreadsheet contains the entire variable names. Please note that the shaded column on the left-hand side of **StatsDirect** spreadsheet with a number from 1 onwards is the line number, it can not be copied or changed. Do not depend on the grey row numbers on the left as an identifier, as the relationship between data and row number will change if the data is sorted by row.

## Missing Data

It is strongly recommended that you use the character \* to represent a missing value. This is different to other statistical packages such as SPSS where it is advisable to specify a numerical or STATA where . (period) is used.

## Opening an Excel Workbook

You can retrieve an EXCEL workbook containing the data from the study of foundry workers and dust. This is stored on the Shared Data Area located on the desk top. Double click on the **Shared Data icon** which can be found on the left hand side of the desktop, follow **Health Methodology Course Data** and save the Excel file to your Work Space. Alternatively the File can be found by going to the following internet address

<http://research.bmh.manchester.ac.uk/biostatistics/teaching/statisticalsupport>

click the link foundry.sav for the dataset followed by **save** and place it in your workspace. Under the File menu select Open File and then find the file. Once open the screen will look like this

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	IDNO	GROUP	AGE	DTBIRTH	DTASSHNT	DTEMPRT	SEX	HT	FEVMEAS	FEVPRED	FVCMEAS	FVCPRED	ASTHMA	BRON	SMKNO
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	
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	
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	
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	
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	
6	1006	1	49	25.06.1947	21.07.1990	24.03.1982	0	174	4.60	3.73	5.82	4.54	0	0	
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	
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	
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	
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	
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	
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	
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	
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	
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	
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	
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	
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	
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	
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	
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	
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	
23	1024	0	45	09.02.1949	16.05.1994	12.05.1988	0	170	3.40	3.50	4.34	4.26	0	0	
24	1025	0	46	17.04.1949	23.06.1995	25.06.1990	0	175	4.01	3.59	5.17	4.45	0	0	
25	1026	1	56	10.01.1942	17.04.1998	18.03.1991	1	165	2.80	2.97	3.57	3.69	0	0	
26	1027	1	54	19.04.1934	12.09.1988	33.07.1979	1	170	3.63	3.24	4.51	4.03	0	1	
27	1028	1	32	12.05.1958	28.07.1990	14.06.1983	1	178	4.68	4.22	5.92	5.06	1	0	
28	1029	1	34	20.01.1960	15.03.1994	24.01.1985	1	190	4.91	4.68	6.06	5.69	0	0	
29	1030	0	50	02.01.1942	20.01.1992	01.01.1976	0	170	4.47	3.36	3.88	4.13	0	1	
30	1031	1	53	10.10.1942	18.11.1995	16.10.1982	0	163	2.16	2.94	3.60	3.61	0	0	
31	1032	0	52	09.04.1945	26.05.1997	20.01.1988	0	185	3.53	3.94	4.70	4.94	0	0	
32	1033	0	42	16.02.1947	02.04.1989	12.02.1977	1	162	3.64	3.24	4.59	3.88	0	0	
33	1034	0	34	17.01.1959	21.03.1993	28.02.1987	0	177	3.69	4.12	5.12	4.95	0	0	
34	1035	0	45	26.06.1947	19.09.1992	31.03.1983	0	170	4.31	3.50	5.50	4.26	0	0	
35	1036	1	38	15.01.1953	22.01.1991	25.01.1974	1	170	3.98	3.72	5.11	4.46	0	0	
36	1037	0	47	09.01.1946	20.03.1993	13.01.1971	0	180	3.97	3.87	5.18	4.78	0	0	
37	1038	0	24	23.07.1966	30.09.1990	24.07.1982	1	180	4.80	4.51	6.27	5.36	0	0	
38	1039	0	35	11.07.1954	11.08.1989	12.06.1974	1	175	4.16	4.00	4.84	4.81	0	0	
39	1040	1	51	02.02.1944	13.01.1995	19.04.1970	0	170	2.72	3.33	3.90	4.11	0	0	
40	1041	1	48	19.06.1949	27.09.1997	20.04.1980	1	175	3.39	3.63	4.17	4.47	0	0	
41	1042	1	38	01.01.1950	12.03.1988	31.01.1981	1	190	4.81	4.56	5.90	5.59	0	0	
42	1043	1	47	21.08.1945	10.06.1992	24.03.1986	1	175	4.59	3.66	4.88	4.49	0	0	
43	1044	1	39	07.07.1955	04.09.1994	25.06.1982	1	190	5.25	4.53	6.74	5.57	0	0	
44	1045	0	62	16.03.1929	11.05.1991	13.03.1983	0	175	3.04	2.74	3.73	3.49	0	0	
45	1046	0	36	17.07.1961	23.09.1997	02.08.1991	1	173	3.68	3.89	4.69	4.67	0	0	
46	1047	0	35	20.03.1955	17.04.1990	31.03.1982	0	170	3.97	3.80	3.97	3.80	0	0	
47	1048	1	40	06.09.1948	20.11.1988	28.10.1979	1	190	5.05	4.50	6.26	5.54	0	0	
48	1049	0	34	31.12.1959	11.12.1993	29.10.1976	0	182	4.54	4.33	5.65	5.24	0	0	

Using the cursor keys you can quickly examine the data set. Notice the IDNO in column 1. There are 136 subjects in this study (so there are 137 rows in the work sheet).

Details of the numerical codes and the coding are given below.

Variable Name	Variable Description	Value Labels for each code
IDNO	Identification No	
GROUP	Exposure Group	1 = Exposed to dust, 0 = Unexposed
SEX		1= male, 0 = female
HT	Height in cms	
FEVMEAS	Measured FEV	
FEVPRED	Predicted FEV	
FVCMEAS	Measured FVC	
FVCPRED	Predicted FVC	
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	-88if never smoked
CIGYRS	No of years smoked	-88 if never smoked
EMPYRS	No of Years with company	
RESPDUST	Current exposure to dust	

## Moving around Data Editor

You may wish to try the following commands that will assist you in moving around the Data Editor.

<i>Key</i>	<i>Description</i>
Home	First Variable of the same case
End	Last Variable of the same case
Ctrl and (Up Arrow)	First Case of the same variable
Ctrl and (Down Arrow)	Last Case of the same variable
Ctrl and Home	First Value
Ctrl and End	Last Value

## INTRODUCING ANALYSIS COMMANDS

### Descriptive Statistics

The first step in data analysis is to generate descriptive statistics. This will help you check the data. This will also give you a feel for the data and help you identify any inconsistencies that there may be in the data. For example it is useful to look at the maximum and minimum values to check that they are in the range of acceptable values. This is sometimes called data cleaning. Techniques that are commonly used to do this include:

- Frequency Tables
- Descriptive Statistics
- Cross-tabulations
- Plots such as bar charts, histograms and scattergrams

The first three are in the analysis menu. Plots are obtained from the graphics menu.

### Analysis By Column Or By Identifier

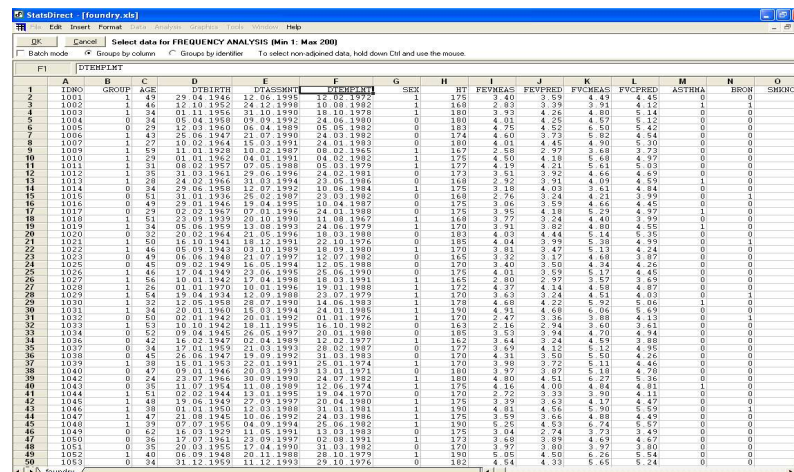
When you select a command that uses raw data it will then ask you to select columns of data in your spreadsheet. Instructions as to which columns to select appear at the top of the screen. Several commands, including some of the descriptive statistics commands, give you the option to **Group by column** or **Group by identifier**. If you select **Group by identifier** you will be asked to select a column that contains a variable that identifies groups. The data in columns that are subsequently selected will be divided according to the values in the identifier. How this option works will be illustrated by the **Frequencies** and **Descriptive** commands.



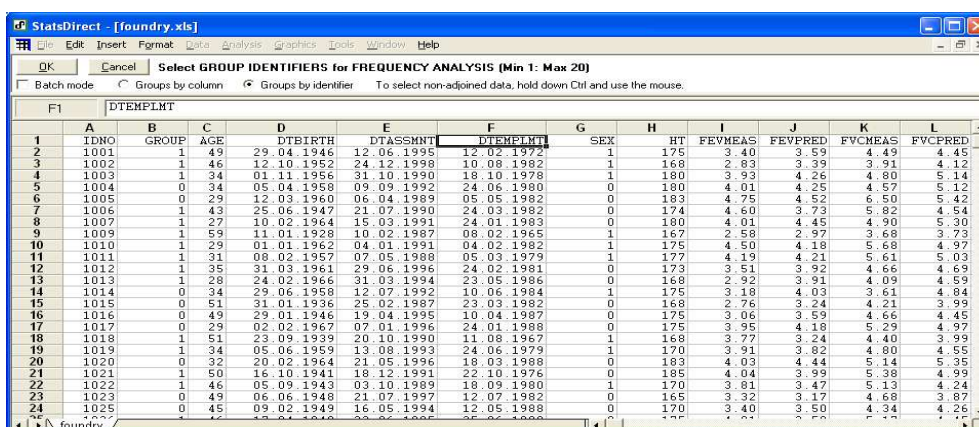
## Frequency Tables

For variables that are categorical or take integer values it is useful to construct a frequency table. To do this in **StatsDirect**, first select the variables for which you want to generate frequencies. If the variables are consecutive you can block them by holding down the left mouse button and drag the mouse across the cells required. If they are not consecutive then select one by one by pressing down the **ctrl** key. Now choose **Analysis** from the Menu bar. From the drop down menu you select **Frequencies**.

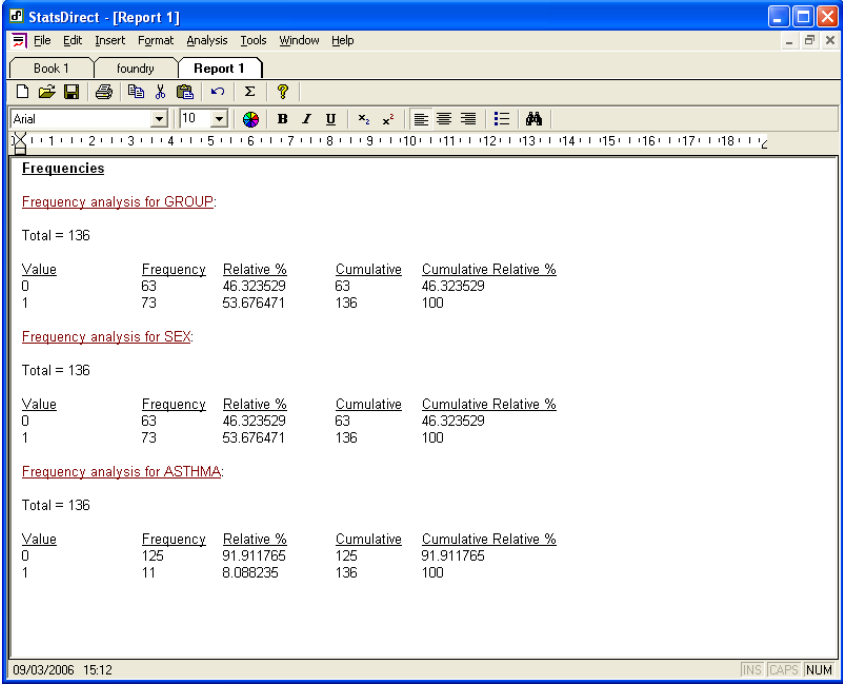
Alternatively you can choose **Analysis** (without choosing the variables on the first) from the Menu bar. From the drop down menu you select **Frequencies**. Since the variables are not chosen previously, at this point a message will appear near the top of the screen (just below the menu bar) saying **Select data for FREQUENCY ANALYSIS [Min1: Max 200]**. Shown below.



If you select the **Group by column** option using the radio button, descriptive statistics will be produced for all data for each variable selected. You select the data the same way as explained above and then click **OK** button. If instead you select the radio button **Group by identifiers** the screen below appears, you could have selected a categorical variable. Separate frequency tables would then be produced for each category. For example if you chose **GROUP** as the identifier variable separate frequency table would be produced for exposed and non-exposed workers. Output for this will be shown later.



You will then be asked where to put the result. You can either select existing output **report1** or **New report**. If you choose **Report1** your output will be appended with the existing output. **New Report** will open a separate page for this output. For example, suppose the three variables **group**, **sex** and **asthma** are chosen to run the frequencies, the output window will look like below



**StatsDirect - [Report 1]**

File Edit Insert Format Analysis Tools Window Help

Book 1 foundry **Report 1**

**Frequencies**

Frequency analysis for GROUP:

Total = 136

Value	Frequency	Relative %	Cumulative	Cumulative Relative %
0	63	46.323529	63	46.323529
1	73	53.676471	136	100

Frequency analysis for SEX:

Total = 136

Value	Frequency	Relative %	Cumulative	Cumulative Relative %
0	63	46.323529	63	46.323529
1	73	53.676471	136	100

Frequency analysis for ASTHMA:

Total = 136

Value	Frequency	Relative %	Cumulative	Cumulative Relative %
0	125	91.911765	125	91.911765
1	11	8.088235	136	100

09/03/2006 15:12 JINS | CAPS | NUM

If instead you had chosen the Group by Identifier option, can only select one variable for the frequency table. Suppose you wanted separate frequencies for smoking status by gender, one would select **sex** as the group identifier variable and **SMKEVER** as the data getting the results below in the output window.

### **Frequencies**

#### Frequency analysis for SMKEVER SEX 1:

Total = 73

<u>Value</u>	<u>Frequency</u>	<u>Relative %</u>	<u>Cumulative</u>	<u>Cumulative Relative %</u>
0	24	32.8767	24	32.8767
1	24	32.8767	48	65.7534
2	25	34.2466	73	100

#### Frequency analysis for SMKEVER SEX 0:

Total = 63

<u>Value</u>	<u>Frequency</u>	<u>Relative %</u>	<u>Cumulative</u>	<u>Cumulative Relative %</u>
0	20	31.746	20	31.746
1	14	22.2222	34	53.9683
2	29	46.0317	63	100

**Exercise 1** 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 20 cigarettes per day?

## Descriptives

The **descriptives** menu in StatsDirect is useful for summarizing quantitative data. It gives the most important descriptive statistics. The output also tells you the number of cases in the analysis and the numbers of missing values. To use this click on the **Analyse** tile choose the **Descriptives** option.

Book 1    **foundry**    **Reg**    **Exact Tests on Counts**  
 File Edit Insert Format Data    **Analysis** Graphics Tools Window Help  
 Chi-Square Tests    **Proportions**  
 Rates    **Distributions**  
 Sample Size    **PLM2**    **SEX**    **H**    **I**    **F**    **K**    **L**    **M**    **N**    **O**    **P**    **G**  
 Randomization    **PLM1**    **1**    **175**    **FEVHES**    **FEVFPD**    **FEVFEAS**    **FEVFCPE**    **ASTHMA**    **BRON**    **SHQNOV**    **SHKEVER**    **EMP**  
 1    **AGE**    **DOBIRTH**  
 2    49    29    04    1946    1972    1    175    3    40    3    59    4    49    4    45    0    0    1    2  
 3    46    12    10    1956    1982    1    169    3    40    3    59    4    49    4    45    0    0    1    2  
 4    34    01    01    1956    1982    1    169    3    40    3    59    4    49    4    45    0    0    1    2  
 5    34    05    04    1956    1982    1    169    3    40    3    59    4    49    4    45    0    0    1    2  
 6    12    05    10    1960    1982    1    169    3    40    3    59    4    49    4    45    0    0    1    2  
 7    43    25    06    1948    1982    1    175    3    40    3    59    4    49    4    45    0    0    1    2  
 8    11    01    1954    1982    1    175    3    40    3    59    4    49    4    45    0    0    1    2  
 9    59    11    01    1928    1982    1    167    2    88    2    97    3    68    3    73    0    0    1    2  
 10    23    05    01    1962    1982    1    159    2    88    2    97    3    68    3    73    0    0    1    2  
 11    08    02    1957    1982    1    159    2    88    2    97    3    68    3    73    0    0    1    2  
 12    35    31    03    1961    1982    1    173    3    51    3    92    4    66    4    69    0    0    1    2  
 13    24    02    1956    1982    1    175    3    40    3    59    4    49    4    45    0    0    1    2  
 14    34    29    06    1958    1982    1    175    3    40    3    59    4    49    4    45    0    0    1    2  
 15    31    01    1936    1982    1    182    3    51    3    92    4    66    4    69    0    0    1    2  
 16    49    29    01    1946    1982    1    175    3    40    3    59    4    49    4    45    0    0    1    2  
 17    29    02    02    1958    1982    1    175    3    40    3    59    4    49    4    45    0    0    1    2  
 18    23    09    1939    1982    1    167    2    88    2    97    3    68    3    73    0    0    1    2  
 19    34    05    06    1959    1982    1    170    3    91    3    82    4    80    4    55    1    0    1    2  
 20    02    01    1964    1982    1    183    4    01    3    82    4    80    4    55    1    0    1    2  
 21    50    16    10    1941    18    12    1991    22    10    1976    0    185    4    04    3    99    5    38    4    99    0    1    0  
 22    05    09    1943    21    07    1997    12    07    1982    0    168    4    04    3    99    5    38    4    99    0    1    0  
 23    49    06    06    1948    21    07    1997    12    07    1982    0    185    4    04    3    99    5    38    4    99    0    1    0  
 24    45    09    02    1949    16    05    1994    12    05    1988    0    170    4    30    4    35    4    34    4    26    0    0    0  
 25    46    17    04    1949    16    05    1994    12    05    1988    0    175    4    30    4    35    4    34    4    26    0    0    0  
 26    56    10    01    1942    17    04    1998    18    03    1991    1    165    2    80    2    97    3    57    3    69    0    0    1  
 27    01    01    1970    21    07    1997    12    07    1982    0    172    4    37    4    44    4    50    4    87    0    0    0  
 28    54    19    04    1934    12    09    1988    27    07    1979    1    178    4    37    4    44    4    50    4    87    0    0    0  
 29    32    12    05    1958    28    07    1998    1    170    4    38    4    43    4    52    4    86    1    0    1  
 30    34    01    1960    28    07    1998    1    190    4    31    4    68    6    06    5    59    0    0    1  
 31    50    02    01    1942    20    01    1992    01    01    1976    1    170    4    37    4    44    4    50    4    87    0    0    1  
 32    10    10    1942    20    01    1992    01    01    1976    1    183    4    01    3    82    4    80    4    55    1    0    1  
 33    52    09    04    1945    26    05    1997    20    01    1988    0    185    3    54    3    94    4    70    4    94    0    0    1  
 34    16    02    1947    26    05    1997    20    01    1988    0    182    3    54    3    94    4    70    4    94    0    0    1  
 35    34    17    01    1959    21    03    1993    28    02    1987    0    177    3    69    4    12    5    12    4    95    0    0    0  
 36    45    26    06    1947    19    09    1992    31    03    1983    0    170    3    31    3    50    5    50    4    26    0    0    2

If you select **Descriptives Report** the screen below will appear

If you select the Group by column option using the radio button, descriptive statistics will be produced for all data for each variable selected. For example if you choose the variables FEVMEAS and FVCMEAS, you get the following in the report window.

**Descriptive statistics**

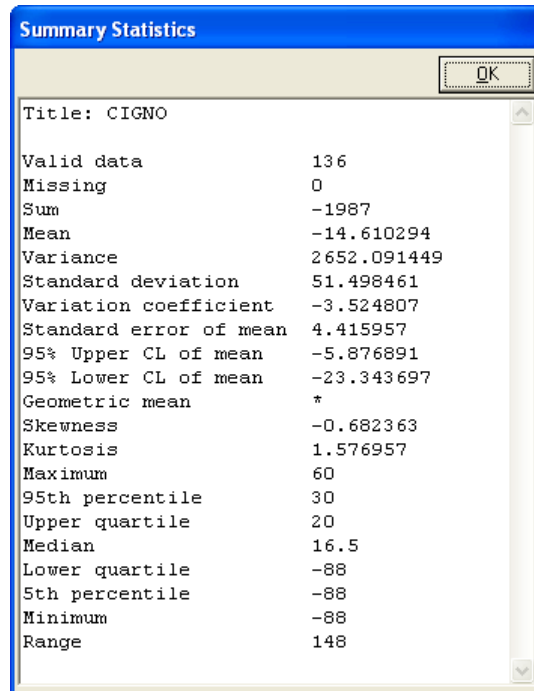
Variables	FEVMEAS	FVCMEAS
Valid Data	136	136
Missing Data	0	0
Mean	3.793824	4.813456
Variance	0.546659	0.705058
SD	0.739364	0.839678
SEM	0.0634	0.072002
Lower 95% CL	3.668438	4.671059
Upper 95% CL	3.919209	4.955853
Geometric Mean	3.714979	4.738105
Skewness	-0.241417	0.041158
Kurtosis	3.228908	3.037622
Maximum	5.61	7.21
Upper Quartile	4.285	5.3775
Median	3.79	4.805
Lower Quartile	3.3525	4.24
Minimum	1.45	2.68
Range	4.16	4.53
Variance coeff.	0.194886	0.174444
Sum	515.96	654.63
Centile 5	2.5885	3.4935

If instead you had selected group by identifier, you could have selected a categorical variable SEX as the identifier. Separate descriptive statistics would then be produced for men and women as shown below. It is only possible to obtain descriptive statistics for one variable at a time. Having selected sex as the identifier and then FEVMEAS as the data you get the result below. The first column give descriptive statistics where SEX = 1, that's men, and the second for SEX= 0 which is women.

**Descriptive statistics**

Variables	FEVMEAS_SEX_1	FEVMEAS_SEX_0
Valid Data	73	63
Missing Data	0	0
Mean	3.8033	3.7829
Variance	0.6703	0.4117
SD	0.8187	0.6416
SEM	0.0958	0.0808
Lower 95% CL	3.6123	3.6213
Upper 95% CL	3.9943	3.9444
Geometric Mean	3.7043	3.7274
Skewness	-0.3677	0.0477
Kurtosis	2.9996	3.3013
Maximum	5.56	5.61
Upper Quartile	4.435	4.21
Median	3.81	3.78
Lower Quartile	3.24	3.4
Minimum	1.45	2.16
Range	4.11	3.45
Variance coeff.	0.2153	0.1696
Sum	277.64	238.32
Centile 5	2.428	2.728

If you select **Quick Univariate Summary**, you cannot divide the data by an identifier. **StatsDirect** will then ask you to select a variable and prepare a summary in a separate window as shown.



The image shows a 'Summary Statistics' dialog box with the title 'CIGNO'. It contains a list of statistical measures and their corresponding values for the variable CIGNO.

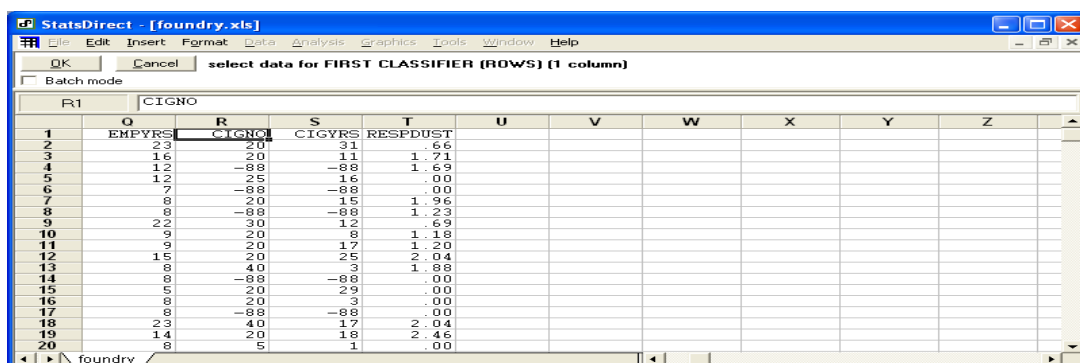
Statistic	Value
Valid data	136
Missing	0
Sum	-1987
Mean	-14.610294
Variance	2652.091449
Standard deviation	51.498461
Variation coefficient	-3.524807
Standard error of mean	4.415957
95% Upper CL of mean	-5.876891
95% Lower CL of mean	-23.343697
Geometric mean	*
Skewness	-0.682363
Kurtosis	1.576957
Maximum	60
95th percentile	30
Upper quartile	20
Median	16.5
Lower quartile	-88
5th percentile	-88
Minimum	-88
Range	148

Note that the mean numbers of cigarettes given per day is  $-14.6102$ . This is because data for non-smokers is recorded as  $-88$  to represent not applicable. Dealing with missing values will be considered below.

**Exercise:** Use the **descriptive** procedure to determine the mean median and range of the ages of the workers. Using the group by identifier option obtain the mean median and range of ages for exposed and non-exposed workers.

### Cross-tabulation, Chi-squared Test and Fisher's Exact Test

To examine the relationship between two categorical variables, a two way Frequency Table can be used called a cross-tabulation. Rows of the data are categories for one variable and columns are categories for the second. Click on **Analyze** then **Crosstabs**. The screen below will appear.

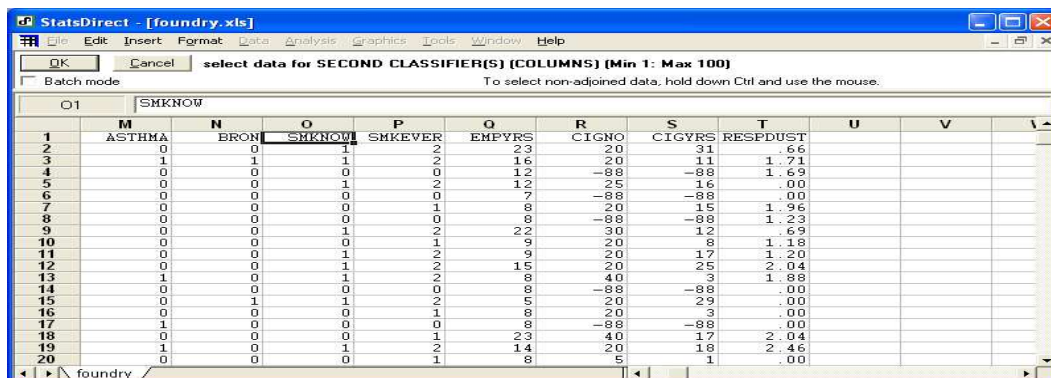


The image shows the StatsDirect software interface with a data table loaded from 'foundry.xls'. The table has columns labeled R1 through Z. The data is organized into rows, with the first row being the header. The table content is as follows:

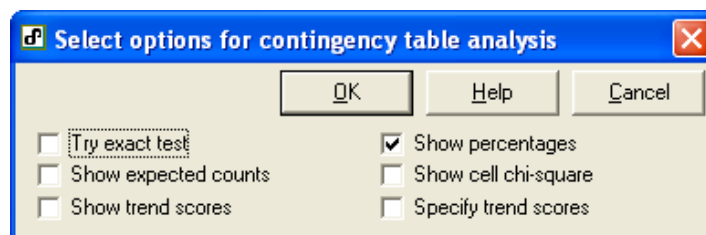
R1	Q	R	S	T	U	V	W	X	Y	Z
1	EMPYRS	CIGNO	CIGYRS	RESPDUST						
2	23	20	31	66						
3	16	20	11	71						
4	12	-88	-88	69						
5	12	25	16	00						
6	7	-88	-88	00						
7	8	20	15	96						
8	8	-88	-88	23						
9	22	30	12	69						
10	9	20	8	18						
11	9	20	17	20						
12	15	20	25	04						
13	8	40	3	88						
14	8	-88	-88	00						
15	5	20	29	00						
16	8	20	3	00						
17	8	-88	-88	00						
18	23	40	17	04						
19	14	20	18	46						
20	8	5	1	00						

Select the column **SMKNOW** then click **OK**. In the next screen select the column **GROUP**.





A third screen appears, which allows you to select a third variable for a three way table. Since we do not wish to do this, click Cancel. The follow panel appears.



Click OK to get the results part of which are shown below

#### Crosstabs

Row variable (first classifier):

**SMKNOW**

Column variable (second classifier):

**GROUP**

	<b>0</b>	<b>1</b>
<b>0</b>	43	39
<b>1</b>	20	34

#### Contingency table analysis

Observed	43	39	<b>82</b>
% of row	52.44%	47.56%	
% of col	68.25%	53.42%	60.29%
Observed	20	34	<b>54</b>
% of row	37.04%	62.96%	
% of col	31.75%	46.58%	39.71%
<b>Total</b>	<b>63</b>	<b>73</b>	<b>136</b>
% of n	46.32%	53.68%	

TOTAL number of cells = 4

#### NOMINAL INDEPENDENCE

Chi-square = 3.106252

DF = 1

**P = 0.078**

G-square = 3.130666

DF = 1

**P = 0.0768**

Fisher-Freeman-Halton exact

not calculated

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% (34/73) as compared to 32% (20/63). This is not statistically significant with a 5% level test. It is recommended that you report the p-value rather than just say significant or not significant. Since the p-value is only slightly larger than 5%, you might write "There was a slight suggestion that workers exposed to dust were more likely to smoke (p=0.078)".

If the cross-tabulation has more than 2 rows or 2 columns, a different set of statistics are produced. For example suppose one compared the two dust exposed and not dust exposed in terms of the smoking status variable SMKEVER, using SMKEVER as the row variable and GROUP as the column variable one gets the following table and output.

**Contingency table analysis**

Observed	24	20	44
% of row	54.55%	45.45%	
% of col	38.1%	27.4%	32.35%
Observed	19	19	38
% of row	50%	50%	
% of col	30.16%	26.03%	27.94%
Observed	20	34	54
% of row	37.04%	62.96%	
% of col	31.75%	46.58%	39.71%
<b>Total</b>	<b>63</b>	<b>73</b>	<b>136</b>
% of n	46.32%	53.68%	

TOTAL number of cells = 6

**NOMINAL INDEPENDENCE**  
 Chi-square = 3.275682 DF = 2 P = 0.1944  
 G-square = 3.299606 DF = 2 P = 0.1921  
 Fisher-Freeman-Halton exact not calculated

**ANOVA**  
 Chi-square for equality of mean column scores = 3.058527  
 DF = 1 P = 0.0803

**LINEAR TREND**  
 Sample correlation (r) = 0.150518  
 Chi-square for linear trend (M²) = 3.058527  
 DF = 1 P = 0.0803

**NOMINAL ASSOCIATION**  
 Phi = 0.155196  
 Pearson's contingency = 0.15336  
 Cramer's V = 0.155196

**ORDINAL**  
 Goodman-Kruskal gamma = 0.146264  
 Approximate test of gamma = 0 SE = 0.133721 P = 0.0655 95% CI = -0.015824 to 0.508351  
 Approximate test of independence: SE = 0.137166 P = 0.0726 95% CI = -0.022577 to 0.515105

If one considers the categories, of smoking to have an order None, Ex, Current, the appropriate statistic is **Chi<sup>2</sup> with Linear Trend**. If instead one did not consider them to be ordered, the appropriate test is **Total Chi<sup>2</sup>** might be appropriate. From this table there is a slight suggestion (p=0.0803) that those exposed to dust (the column Failures) are more likely to smoke and less likely to be non-smokers or ex-smokers.

**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.

---

You should find no statistically significant relationship between exposure to dust and either asthma or bronchitis symptoms. 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 ( $p=0.169$ ). 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.



# FILE HANDLING

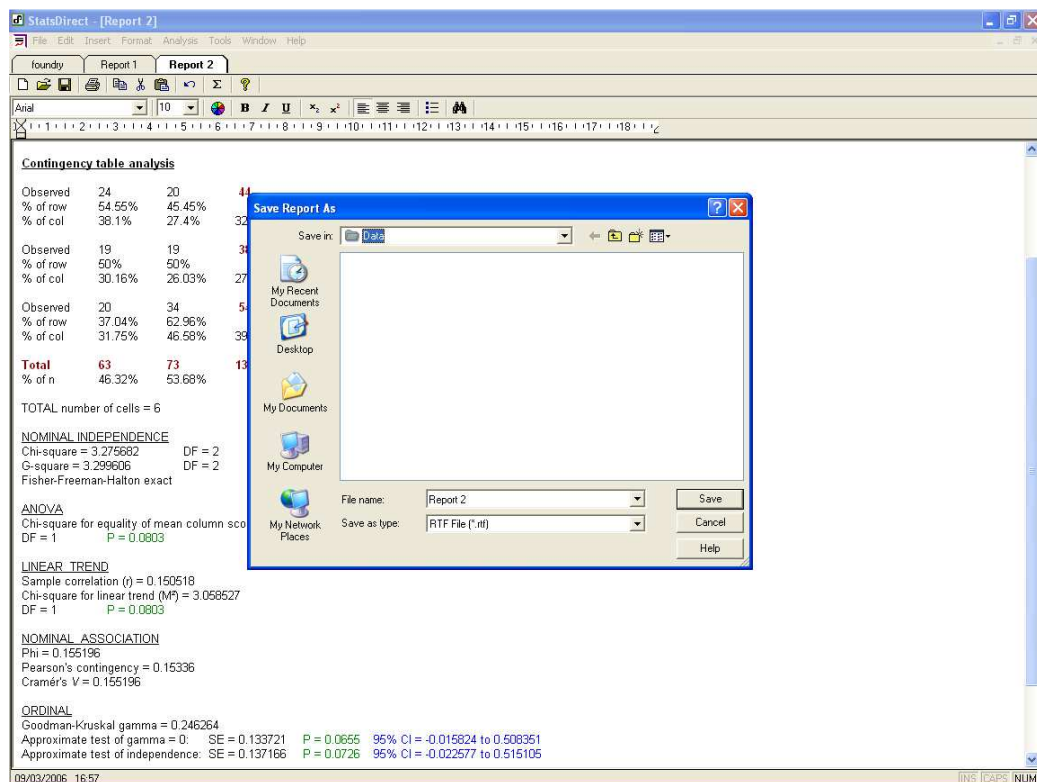
## Saving The Work

**Worksheets** and **Reports** are all saved separately. To save the report that is created here:

**Either:**

a) Select **Save Report As** from the file menu,

**Or:** b) Click on the **Report 1** tab with the right mouse button and select **Save As** from the menu of commonly used options which appears below



The standard **Save As** dialogue box will be displayed as above. You have to choose a drive and a suitable name of the file. For Report the file extension is **.rtf** (which stands for rich text format. Files in that format can be opened by **Word** to edit). Then click the **Save** button. Say, you saved the Report with the name called **survey1** in the pen drive.

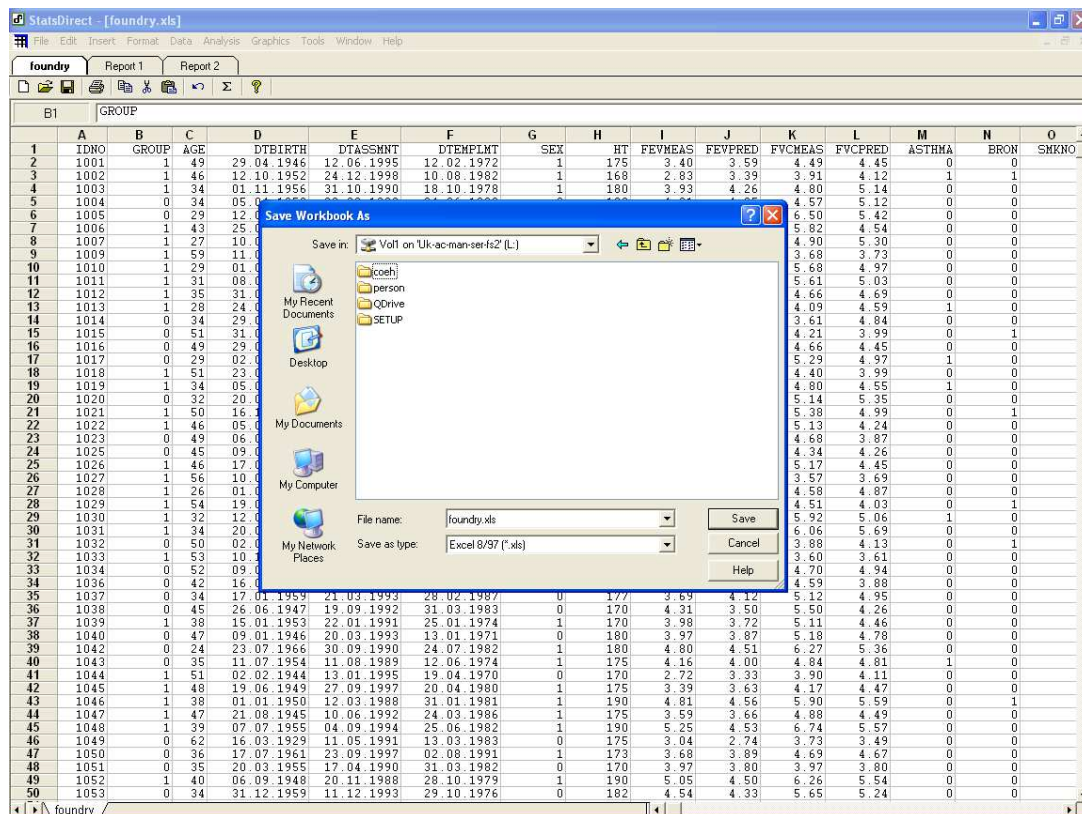
Similarly, to save the Workbook (**Book 1**):

**Either:**

a) Select **Save Workbook As** from the file menu,

**Or:**

b) Click on the **Book 1** tab with the right mouse button and select **Save As** from the menu of commonly used options which appears below



The standard **Save As** dialogue box will be displayed as above. You have to choose a drive and a suitable name. For Workbook file the file extension is **.sdw**. Then click the **Save** button. Say, you saved the workbook with the name called **survey1** in the pen drive.

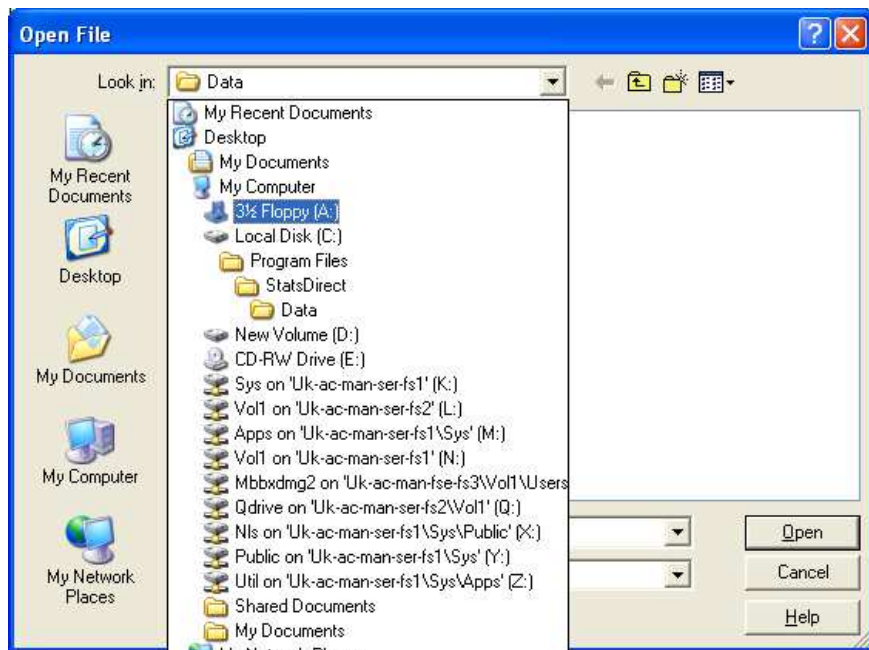
## Backup The Work

When saving Workbook or Report, make sure that you save onto at least two USB pen drives just in case one of your USB pen drives fail. To make a backup copy of your work repeat the **Save As** procedure.

## Retrieving a StatsDirect File

Having returned to the computer, access the system as before and click on the **File** option from the menu bar, then on the **Open File** sub-option.

To open a copy of the **StatsDirect** file on your pen drive, under **Drives:** click on  $\Sigma$  in the **Look in** window to generate a list of the drives. Select the appropriate **Pen Drive** where you saved your file, a list of file names will be displayed. Then choose **survey1.sdw** and click **Open** button (shown below)



## MODIFYING THE DATA

StatsDirect contains the same functions for inserting and deleting rows and columns as a standard spreadsheet such as EXCEL. These are found under the Edit and Insert menus. It is possible to move columns and rows using cut and paste. StatsDirect also contains some additional functions to calculate new variables from old variables.

### Modifying a new variable by Search and Replace

The Search and Replace option in the data menu allows more complex editing than that available in the standard editing menu. It should be noted that this is not as powerful as the recode procedures in SPSS or STATA but it still allows you to do some simple recoding.

In the exercise above we noted that number of cigarettes per day is recorded in the data set with -88 for missing. You might want to change this to the missing value code \*, if you want to calculate the mean numbers of cigarettes smoked by smokers. Alternatively you might change this to 0 if you wanted to calculate the mean number of cigarettes smoked by each worker regardless of whether they were a smoker. For this part of the tutorial we will change the “-88” in the variables **signo** to \*. Select the column and search and replace.

StatsDirect - [foundry.xls]

File Edit Insert Format Data Analysis Graphics Tools Window Help

Book 1 foundry

Transformations  
Sort  
Random Numbers  
Pairwise  
Grouping  
Apply Function  
Clear Missing Data  
Date & Time Intervals  
Dummy Variables  
Fill Series  
Normal Scores  
Rank  
Rotate Data Block  
**Search & Replace**  
Standardization  
Text to Numbers

	A	B		F	G	H	I	J	K	L	M	N	O	P	RES
1	1070			HT	FEVMEAS	FEVPRD	FVCMEAS	FVCPRD	ASTHMA	BRO	SHKNO	SHKEVER	CIGLO	CIGYRS	EMPYRS
2	1091			177	5.61	4.21	7.21	5.03	0	1	1	2	10	5	12
3	1134			160	4.51	3.45	5.08	4.02	0	0	0	0	-88	-88	5
4	1006			165	4.62	3.72	5.31	4.36	0	0	1	2	15	4	7
5	1086			174	4.60	3.73	5.82	4.54	0	0	0	1	20	15	8
6	1018			185	5.56	4.67	6.22	5.59	0	0	0	0	-88	-88	9
7	1018			168	3.77	3.24	4.40	3.99	0	0	0	1	40	17	23
8	1133			175	4.79	4.12	5.53	4.91	0	0	0	1	10	6	9
9	1048			190	5.25	4.53	6.74	5.57	0	0	1	2	3	9	12
10	1061			175	4.26	3.69	5.42	4.52	1	0	1	2	20	5	8
11	1099			170	4.27	3.70	5.16	4.44	0	0	1	2	15	15	12
12	1135			170	3.97	3.53	5.01	4.29	0	0	1	2	30	25	19
13	1052			190	5.05	4.50	6.26	5.54	0	0	0	0	-88	-88	9
14	1029	1	54	170	3.63	3.24	4.51	4.03	0	1	0	1	20	30	9
15	1140	1	56	168	3.43	3.07	4.22	3.82	0	0	0	0	-88	-88	10
16	1026	1	46	175	4.01	3.59	5.17	4.45	0	0	0	0	-88	-88	5
17	1095	1	32	173	4.46	4.01	5.27	4.77	0	0	0	0	-88	-88	6
18	1030	1	32	178	4.68	4.22	5.92	5.06	1	0	0	0	-88	-88	7
19	1087	1	39	165	3.84	3.47	4.96	4.15	0	0	0	1	20	8	17
20	1022	1	46	170	3.81	3.47	5.13	4.24	0	0	0	0	-88	-88	9
21	1010	1	29	175	4.50	4.18	5.68	4.97	0	0	0	1	20	8	9
22	1097	1	56	177	3.74	3.48	4.58	4.38	0	0	0	1	40	15	19
23	1039	1	38	170	3.98	3.72	5.11	4.46	0	0	0	1	13	4	17
24	1065	1	37	175	4.19	3.95	5.05	4.76	0	0	1	2	20	15	7
25	1028	1	26	172	4.37	4.14	4.58	4.87	0	0	0	0	-88	-88	8
26	1075	1	31	180	4.58	4.34	5.39	5.20	0	0	1	2	30	10	8
27	1046	1	38	190	4.81	4.56	5.90	5.59	0	1	1	2	25	20	7
28	1031	1	34	190	4.91	4.68	6.06	5.69	0	0	1	2	12	13	9
29	1085	1	32	182	4.59	4.39	5.96	5.29	0	0	0	1	20	10	8
30	1090	1	29	168	4.05	3.88	4.69	4.56	0	0	1	2	15	4	9
31	1126	1	28	178	4.51	4.34	5.87	5.16	0	0	1	2	3	8	9
32	1148	1	31	168	3.96	3.82	4.87	4.51	0	0	0	1	20	11	5
33	1078	1	59	173	3.30	3.19	4.24	4.09	0	0	0	0	-88	-88	7
34	1059	1	50	170	3.46	3.36	4.40	4.13	0	0	1	2	25	32	22
35	1124	1	45	175	3.80	3.71	4.75	4.55	0	0	0	1	10	11	21
36	1019	1	34	170	3.91	3.82	4.80	4.55	1	0	1	2	20	18	14
37	1021	1	50	185	4.04	3.99	5.38	4.99	0	1	0	1	40	32	15
38	1094	1	50	163	3.08	3.05	3.68	3.73	0	0	0	0	-88	-88	10
39	1054	1	49	170	3.40	3.38	4.11	4.16	0	0	0	1	20	16	16
40	1129	1	28	175	4.22	4.21	4.96	4.99	0	0	0	0	-88	-88	9
41	1114	1	39	178	4.01	4.02	4.81	4.88	0	0	0	0	-88	-88	12
42	1011	1	31	177	4.19	4.21	5.61	5.03	0	0	1	2	20	17	9
43	1153	1	33	180	4.21	4.28	5.89	5.15	0	1	1	2	25	20	17
44	1047	1	47	175	3.59	3.66	4.88	4.49	0	0	0	0	-88	-88	6
45	1131	1	49	172	3.39	3.47	4.05	4.27	0	0	0	1	8	24	24
46	1074	1	53	175	3.39	3.48	5.30	4.34	0	0	1	2	30	37	7
47	1132	1	25	175	4.15	4.29	5.37	5.07	0	1	1	2	20	10	8
48	1098	1	32	171	3.78	3.92	4.55	4.66	0	0	1	2	20	12	16
49	1123	1	37	170	3.70	3.88	5.40	4.59	0	1	1	2	40	26	8
50	1113	1	45	163	3.04	3.19	4.18	3.86	0	0	0	0	-88	-88	22

When the panel below appears insert the values as shown. It is worth clicking on **Count**, to check how many values satisfy the condition. Then click **Replace** to complete the change.

**Search**

Count Replace Delete Help Exit

Search for entries

☒ equal to  
☐ greater than  
☐ less than  
☐ greater than or equal to  
☐ less than or equal to  
☐ not equal to  
☐ match expression

Search for: -88  
 Replace with: 1

Search mode  
☐ Numerical  
☒ Text

You can also code a numeric value to a string value and vice versa by this **Search and Apply** method. Suppose you want to code the variable **sex**, 0 as male and 1 as female. The screen below is an example.

**Search**

Count Replace Delete Help Exit

Search for entries

☒ equal to  
☐ greater than  
☐ less than  
☐ greater than or equal to  
☐ less than or equal to  
☐ not equal to  
☐ match expression

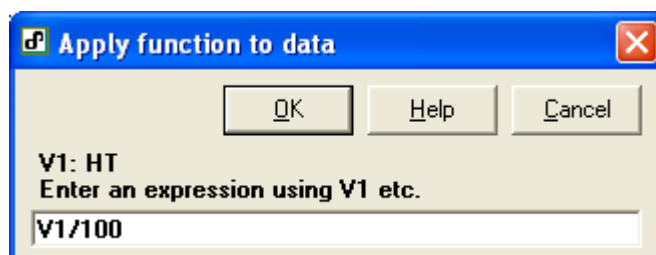
Search for: 0  
 Replace with: male

Search mode  
☐ Numerical  
☒ Text

This recoding is very useful for some analyses and graphs, such as crosstabs boxplots etc.

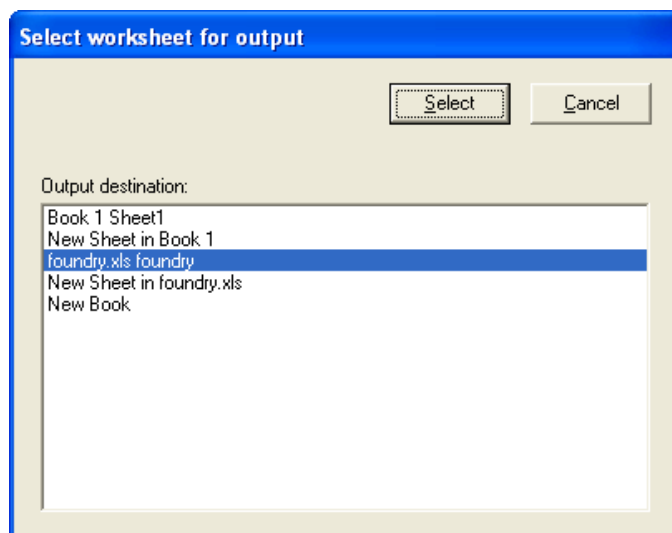
## Creating a new variable by calculation

It is recommended that you store your data to disk just in case you make a mistake. Suppose you want to convert the variable HT in meters. To do that you highlight the variable HT, then click **Data** then **Apply Function**. The following screen will be displayed.



You can see a little window called "Enter an expression using V1 etc." You have to write the mathematical expression there. In our case it should be **V1/100** (**V1** represents **HT**). Then press **OK** button.

Then you have options where to put the variable (shown below).



It is always advisable to put the new variable in the existing data sheet, so choose **foundry.xls**, then press **Select** button. The new variable name will be **HT/100** and it will be placed right at the end of the variable list. You can replace the variable called **HT/100** by any suitable name (say, **HTmeter**).

**Exercise 1:** A variable FEV ratio is need for a later part of this tutorial. This is defined as FEV Measured/ FEV Predicted. Block these two columns on the spreadsheet and then use **Data** then **Apply Function** to construct this new variable.

**Exercise 2:** Use the integer divide operator \ to create a new variable **agegp** grouping ages by decade

[Hint The integer divide operator is \. Select **AGE** then use Apply function in the Data menu with the formula **v1\10**. Compare the **AGE** and **AGEGP** to check your result]

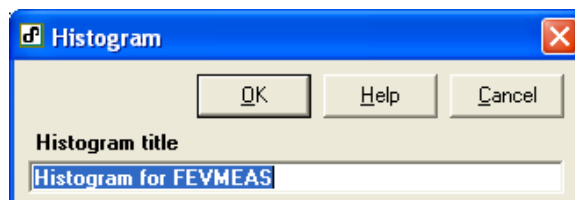
# PART II

## CONTINUOUS OUTCOME MEASURES

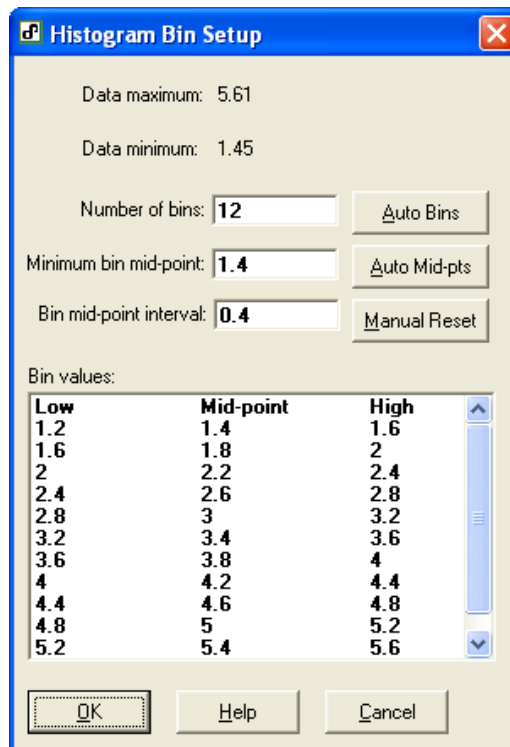
Perhaps the first step in analysing a continuous variable is to examine its distribution. This can be done graphically using a histogram. Histogram can also be used to detect outlier, that is values very different from the other data points.

### Histogram

Histograms are produced for interval variables e.g. age, height etc. Here we will produce a histogram of measured FEV. To do that, click **Graphs** on the menu bar and select **Histogram** from drop down menu. Then select the data and click **OK**. You will get a little window asking for the title for the histogram.

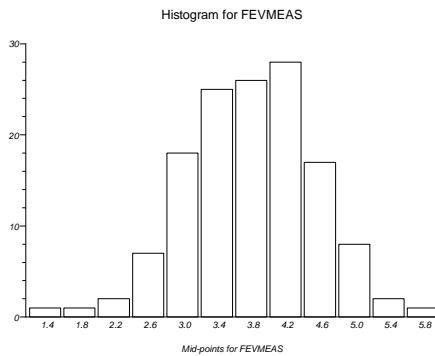


Choose the title and click **OK** button. Then you will get another window like that below.

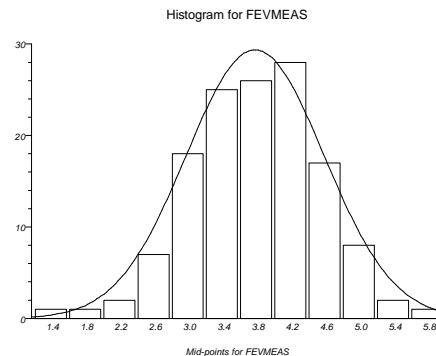


Click **OK** then you will be asked whether you want to overlay the normal curve. If you say “yes” it will overlay the normal curve, otherwise it will produce only the histogram. Then select the title of the axis and the output destination and you will get the histogram (shown below)





Without Normal Curve



With Normal Curve

**Exercise:** Draw a histogram of AGE or of HT

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.

We now want to examine whether workers exposed to dust have reduced lung function. First you might examine this graphically with a box plot also called a Box-and-Whisker plot. To examine whether dust exposure affects lung function you want a separate Box-and-Whisker for exposed and not exposed separately. Make sure no variables are highlighted. Go to the graph menu, select **Box & Whisker**. Then the following screen will appear. Whether a subject is exposed is given by the identifying variable **Group**. Use **Groups by Identifier** rather than **Groups by Column**. Select the radio button for **Groups by Identifier** if necessary. You will then be asked for the column for the group identifier which in this case is the variable **Group**. Then select the variable for FEV Ratio.

StatsDirect - [foundry.xls]

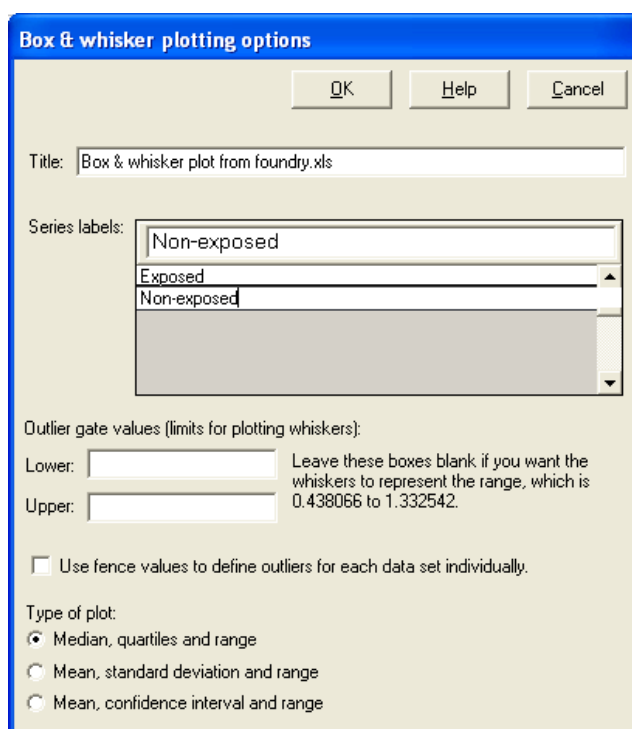
File Edit Insert Format Data Analysis Graphics Tools Window Help

OK Cancel Select GROUP IDENTIFIERS (Min 1: Max 20)

☐ Batch mode ☐ Groups by column ☒ Groups by identifier To select non-adjointed data, hold down Ctrl and use the mouse.

B1	GROUP																
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	RES
	IDNO	GROUP	AGE	SEX	HT	FEVMEAS	FEVPRED	FVCMEAS	FVCPRED	ASTHMA	BROM	SHKNOV	SHKEVER	CIGNO	CIGYRS	EMPYRS	
1	1090	1	31	0	177	5.61	4.21	7.21	5.03	0	1	1	2	10	5	12	
2	1070	1	31	0	177	5.61	4.21	7.21	5.03	0	1	1	2	10	5	12	
3	1091	1	32	1	160	4.51	3.45	5.08	4.02	0	0	0	0	*	-88	5	
4	1134	1	30	0	165	4.62	3.72	5.31	4.36	0	0	1	2	15	4	7	
5	1006	1	43	0	174	4.60	3.73	5.82	4.54	0	0	0	1	20	15	8	
6	1085	1	27	1	185	5.56	4.67	6.22	5.59	0	0	0	0	*	-88	9	
7	1018	1	51	1	168	3.77	3.24	4.40	3.99	0	0	0	1	40	17	23	
8	1133	1	31	1	175	4.79	4.12	5.53	4.91	0	0	0	1	10	6	9	
9	1048	1	39	1	190	5.25	4.53	6.74	5.57	0	0	1	2	3	9	12	
10	1061	1	46	1	175	4.26	3.69	5.42	4.52	1	0	1	2	20	5	8	
11	1099	1	38	1	170	4.27	3.70	5.16	4.44	0	0	1	2	15	15	12	
12	1135	1	44	0	170	3.97	3.53	5.01	4.29	0	0	1	2	30	25	19	
13	1052	1	40	1	190	5.05	4.50	6.26	5.54	0	0	0	0	*	-88	9	
14	1029	1	54	1	170	3.63	3.24	4.51	4.03	0	1	0	1	20	30	9	
15	1140	1	56	0	168	3.43	3.07	4.22	3.82	0	0	0	0	*	-88	10	
16	1026	1	46	0	175	4.01	3.59	5.17	4.45	0	0	0	0	*	-88	5	
17	1095	1	32	1	173	4.46	4.01	5.27	4.77	0	0	0	0	*	-88	6	
18	1030	1	32	1	178	4.68	4.22	5.92	5.06	1	0	0	0	*	-88	7	
19	1087	1	39	1	165	3.84	3.47	4.96	4.15	0	0	0	1	20	8	17	
20	1022	1	46	1	170	3.81	3.47	5.13	4.24	0	0	0	0	*	-88	9	
21	1010	1	29	1	175	4.50	4.18	5.68	4.97	0	0	0	1	20	8	9	
22	1097	1	56	0	177	3.74	3.48	4.58	4.38	0	0	0	1	60	15	19	
23	1039	1	38	1	170	3.98	3.72	5.11	4.46	0	0	0	1	13	4	17	
24	1065	1	37	0	175	4.19	3.95	5.05	4.76	0	0	1	2	20	15	7	
25	1028	1	26	1	172	4.37	4.14	4.58	4.87	0	0	0	0	*	-88	8	
26	1075	1	31	1	180	4.58	4.34	5.39	5.20	0	0	1	2	30	10	8	
27	1046	1	38	1	190	4.81	4.56	5.90	5.59	0	1	1	2	25	20	7	
28	1031	1	34	0	190	4.91	4.68	6.06	5.69	0	0	1	2	12	13	9	
29	1085	1	32	1	182	4.59	4.39	5.96	5.29	0	0	0	1	20	10	8	
30	1090	1	29	0	168	4.05	3.88	4.69	4.56	0	0	1	2	15	4	9	
31	1126	1	48	0	178	4.51	4.34	5.87	5.16	0	0	1	2	3	6	9	
32	1148	1	31	0	168	3.96	3.82	4.87	4.51	0	0	0	1	20	11	5	
33	1078	1	59	1	173	3.30	3.19	4.24	4.09	0	0	0	0	*	-88	7	
34	1059	1	50	1	170	3.46	3.36	4.40	4.13	0	0	1	2	25	32	22	
35	1124	1	45	0	175	5.00	3.71	4.75	4.55	0	0	1	2	10	11	21	
36	1019	1	34	1	170	3.91	3.82	4.80	4.55	1	0	1	2	20	18	14	
37	1021	1	50	0	185	4.04	3.99	5.38	4.99	0	1	0	1	40	32	15	
38	1094	1	50	1	163	3.08	3.05	3.68	3.73	0	0	0	0	*	-88	10	
39	1054	1	49	0	170	3.40	3.38	4.11	4.16	0	0	1	2	20	16	16	
40	1129	1	28	0	175	4.22	4.21	4.96	4.99	0	0	0	0	*	-88	9	
41	1114	1	39	0	178	4.01	4.02	4.81	4.88	0	0	0	0	*	-88	12	
42	1011	1	31	1	177	4.19	4.21	5.61	5.03	0	0	1	2	20	17	9	
43	1153	1	33	0	180	4.21	4.28	5.89	5.15	0	1	1	2	25	20	17	
44	1047	1	47	1	175	3.59	3.66	4.88	4.49	0	0	0	0	*	-88	6	
45	1131	1	49	1	172	3.39	3.47	4.05	4.27	0	0	0	1	8	24	24	
46	1074	1	53	0	175	3.39	3.49	5.30	4.34	0	0	1	2	30	27	7	
47	1132	1	25	1	175	4.15	4.29	5.37	5.07	0	1	1	2	20	10	8	
48	1098	1	32	0	171	3.78	3.92	4.55	4.66	0	0	1	2	20	12	16	
49	1123	1	37	1	170	3.70	3.88	5.40	4.59	0	1	1	2	40	26	8	
50	1113	1	45	1	163	3.04	3.19	4.18	3.86	0	0	0	0	*	-88	22	

Next you will be asked for the statistics used to construct the boxplot. The usual choice is **median and quartile**. Select this then provide titles as requested for the graph and the groups as requested.



**Box & whisker plotting options**

OK Help Cancel

Title: Box & whisker plot from foundry.xls

Series labels:

- Non-exposed
- Exposed
- Non-exposed

Outlier gate values (limits for plotting whiskers):

Lower: Upper: Leave these boxes blank if you want the whiskers to represent the range, which is 0.438066 to 1.332542.

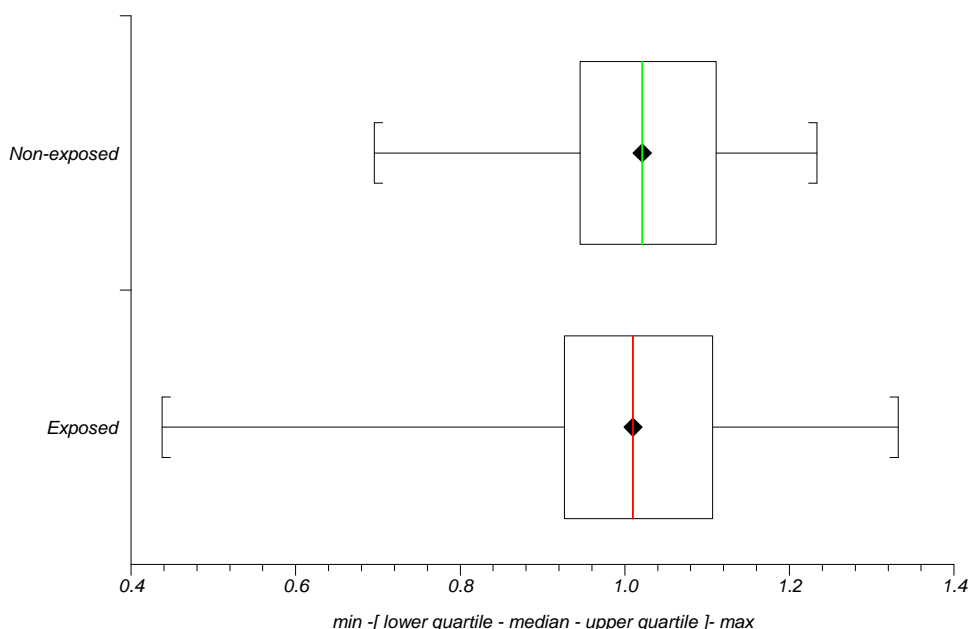
☐ Use fence values to define outliers for each data set individually.

Type of plot:

- ☒ Median, quartiles and range
- ☐ Mean, standard deviation and range
- ☐ Mean, confidence interval and range

The boxplot is displayed below.

FEV Ratio Exposed and Non-Exposed Workers



From this one can see that the groups have similar median and quartiles although the range is much larger for the exposed subjects. Descriptive statistics for each exposure group can be obtained using the **Descriptive** option in the **Analysis** menu. Before you do this make sure no variables are highlighted in order that you can select groups by an indicator variable.



**Exercise:** Use the **Descriptives** 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. Under **Analysis** select **Parametric** then **Unpaired t**

The screenshot shows the StatsDirect software interface. The 'Analysis' menu is open, and 'Parametric' is selected, leading to the 'Unpaired t' option. The background data table has the following columns: T1, A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, RES. The data rows show various numerical values for these variables.

To compare mean FEV ratio for exposed and unexposed select **GROUP** as the identifier variable then **FEVRATIO**. This gives the following in the report window.

### Unpaired t test

Mean of FEVRAT\_GROUP\_1 = 1.0003 (n = 73)  
Mean of FEVRAT\_GROUP\_0 = 1.0158 (n = 63)

#### Assuming equal variances

Combined standard error = 0.0239  
df = 134  
t = 0.6467  
One sided P = 0.2595  
Two sided P = 0.519

95% confidence interval for difference between means = -0.0627 to 0.0318

#### Assuming unequal variances

Combined standard error = 0.0236  
df = 133.999  
t(d) = 0.6536  
One sided P = 0.2572  
Two sided P = 0.5145

95% confidence interval for difference between means = -0.0622 to 0.0313

#### Comparison of variances

Two sided F test is not significant  
No need to assume unequal variances

Two analyses are given. The first assumes variances are equal and the second are unequal. It also gives the results of statistical test comparing the variance. Unfortunately it just summarizes this as significant or non-significant without specifying the values of the magnitude of the variances, standard deviation or the significance level. This is an omission of the software. Above we saw that the standard deviation was 0.1479 for exposed and 0.1278 for non-exposed. As the F-test suggests that they are not significantly different, we take the unequal variance analysis as the t-test results although in this case it makes little difference. The difference of the two means is -0.0155. 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 - 0.0627 to 0.0318 p=0.519)”

**Exercise:** Compare mean FVC ratio for the exposed and non-exposed subjects using a t-test

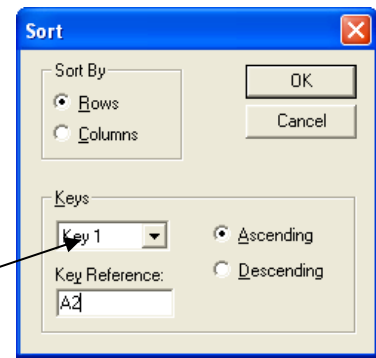
## REARRANGING DATA

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. You will now carry out some analysis on just the exposed subjects, but to do this you need to select just those subjects.

### Sorting Data

There are several ways you can do this. You could sort the data by group then delete the non-exposed subjects. Before you do this on your own data, make sure it is saved to disk in case anything goes wrong. Data in the spreadsheet can be sorted using the **Sort** option under **Data**. Check the help to find details of this procedure as this is a little complicated.

Highlight all the data involved excluding the column labels. Then select **Data** then **Sort** then **Manipulate Worksheet** the panel (right) will appear. Change the cell reference under Key Reference to select the column by which the data is to be sorted. To sort the data by exposure group insert put B1. If a nested sort is required, i.e. sort by exposure group then by sex within each exposure group, alter Key 1 to Key 2 and enter the appropriate Key reference. In this case enter G1.



### Exercise:

- (1) Sort the rows of data by group.
- (2) By using two Keys sort the data first by group and then by age within group.

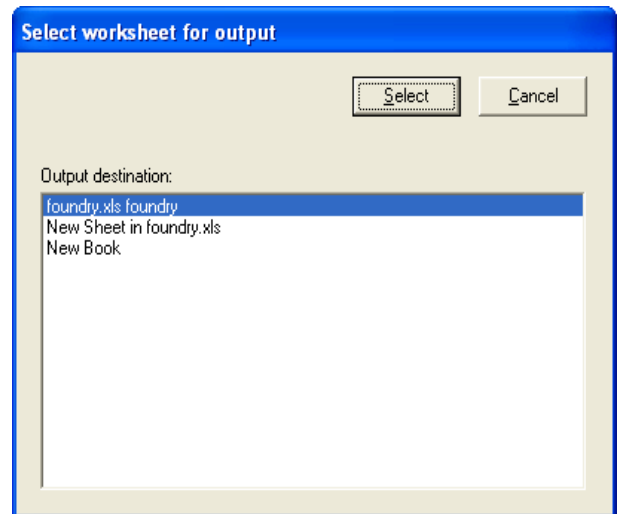
### Group Split

Alternatively a new spreadsheet can be created in which there are separate columns for exposed and non-exposed workers for each variable. Select **Group split** on the **Data** menu. Then select the column that identifies the groups into which you want to split the data, in this case the variable **group**. It then asks us to choose the variables to which we want to apply **Group split**. For this exercise select all the columns RESPDUST and FEVRAT. If you need to select columns that are not adjacent you need to use the **Ctrl** key.

**Group split** then confirms the number of groups in the panel below.



Click **OK** to give the panel to the right. To save confusion it is suggested that you put the split data into a new worksheet in the same work book. A new worksheet is now created as shown below



The screenshot shows the StatsDirect software interface with a worksheet named 'foundry'. The worksheet contains a table with 50 rows and 12 columns (A through L). The data is organized into two main groups based on the 'RESPDUST\*GROUP' variable. The first group (GROUP=1) is in columns A and B, and the second group (GROUP=0) is in columns C and D. The remaining columns (E through L) are empty.

	A	B	C	D	E	F	G	H	I	J	K	L
1	RESPDUST*GROUP=1	FEVRAT*GROUP=1	RESPDUST*GROUP=0	FEVRAT*GROUP=0								
2	1.15	1.332541568	0	1.23325062								
3	0.29	1.307246377	0	1.23255814								
4	2.34	1.241935484	0	1.231428571								
5	1.96	1.233243968	0	1.211538462								
6	1.68	1.190578158	0	1.182897862								
7	2.04	1.163580247	0	1.180974478								
8	1.83	1.162621359	0	1.177464789								
9	1.45	1.158940397	0	1.168341709								
10	2.43	1.154471545	0	1.164983165								
11	0.55	1.154054054	0	1.163507109								
12	1.03	1.124645892	0	1.158501441								
13	0.76	1.122222222	0	1.152631579								
14	1.5	1.12037037	0	1.145539906								
15	1.19	1.117263844	0	1.130573248								
16	0.72	1.116991643	0	1.12345679								
17	0.91	1.112219451	0	1.111398964								
18	1.73	1.109004739	0	1.109725686								
19	2.72	1.106628242	0	1.109489051								
20	0.29	1.097982709	0	1.099502488								
21	1.18	1.076555024	0	1.073979592								
22	0.86	1.074712644	0	1.06870229								
23	1.12	1.068992473	0	1.067092652								
24	0.4	1.06075494	0	1.064301552								
25	1.88	1.055555556	0	1.050884956								
26	2.29	1.055299539	0	1.050295858								
27	1.66	1.054824561	0	1.048498845								
28	1.12	1.049145299	0	1.047318612								
29	1.51	1.045558087	0	1.044736842								
30	1.45	1.043814433	0	1.04								
31	1.27	1.039170507	0	1.028277635								
32	0.77	1.036649215	0	1.025839793								
33	1.39	1.034482759	0	1.021226415								
34	1.74	1.029761905	0	1.017191977								
35	0.83	1.02425876	0	1.014336918								
36	0.46	1.023560209	0	0.9950617284								
37	1.22	1.012531328	0	0.9896193772								
38	0.97	1.009836066	0	0.9842105263								
39	2.22	1.00591716	0	0.9836065574								
40	1.77	1.002375297	0	0.9785522788								
41	3.31	0.9975124378	0	0.9714285714								
42	1.2	0.9952494062	0	0.9702702703								
43	0.6	0.9836448598	0	0.9690721649								
44	1.95	0.9808743169	0	0.9626168224								
45	1.9	0.976945245	0	0.9625360231								
46	1.65	0.974137931	0	0.9595687332								
47	1.39	0.9673659674	0	0.9510703364								
48	0.9	0.9642857143	0	0.9460154242								
49	0.29	0.9536082474	0	0.9451754386								
50	1.29	0.9529780564	0	0.9449760766								

Either the split work-sheet or the original can be selected by the tab at the bottom of the screen.

# EXAMINING THE RELATIONSHIP BETWEEN TWO CONTINUOUS VARIABLES

## Scatter Plots

The first step in examining the relationship between two continuous variables such as respiratory dust and lung function is to construct a scatter plot. This can be obtained from the graphics menu and is shown below.

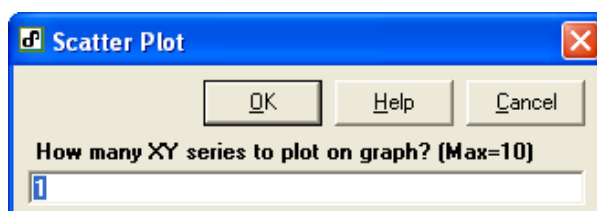
Scatter plots have two axes:

the value of the dependent or response variable on the y axis

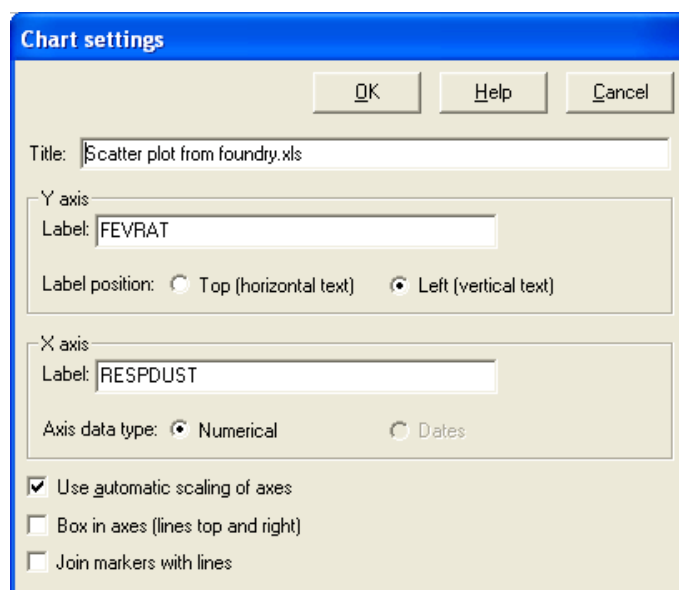
the value of the independent variable on the horizontal axis.

Since we expect respiratory dust to affect FEV rather than the other way round you should choose FEVRATIO as the y-axis and RESPDUST as the x-axis. If you have not already created the variable FEVRATIO, return to on page 721 (Creating new variables) to create it. You will also need to sort the data so that you can select just the expose subjects (Group 1) as described on page 28.

To run the scatter plot click **Graphs** from the menu bar and **Scatter** from the drop down menu. Then a little window will appear like below

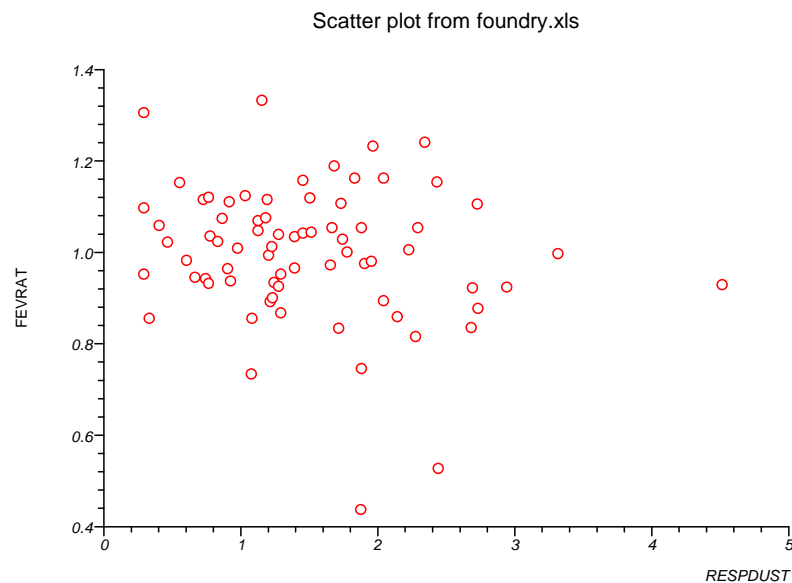


Type the number “how many xy series to plot on graph” and click **OK**. (In our case it will be 1) followed by selecting the Y and X axis data (FEVratio and RESPDUST respectively). Then you see another window with chart setting options. When you amend your options then click **OK**.





The scatter plot of FEV ratio compared to dust for subjects for the exposed group is displayed below.



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

## Linear Regressions

To test this we will use linear regression to fit a straight line of the form  $Y = A + BX$ . Where Y is the dependent variable **fevratio** and X is independent variable **respdust**. If the gradient B is negative, this would indicate reduced respiratory function with increased dust. To do this in StatsDirect go to Regression & Correlation in the Analysis menu then select simple linear regression. And the following screen should appear.

StatsDirect [foundry.xls]

FileEditInsertFormatData

AnalysisGraphicsToolsWindowHelp

foundryReport 1

E11170

<

When requested select FEVRATIO as the Outcome or Y variable. Select RESPDUST as the Predictor or X variable. The following output should appear. This gives an equation for the line.

The screenshot shows a software window titled "Regression and Correlation". On the left, under "Simple linear regression", the "Interpolate X to Y" option is selected. To the right, there is a text input field and a "Calculate" button. Below the input field, the text "Interpolate X to Y" and "Type value for X and press Enter to calculate" is displayed. The main results area, titled "Simple linear regression", contains the following information:

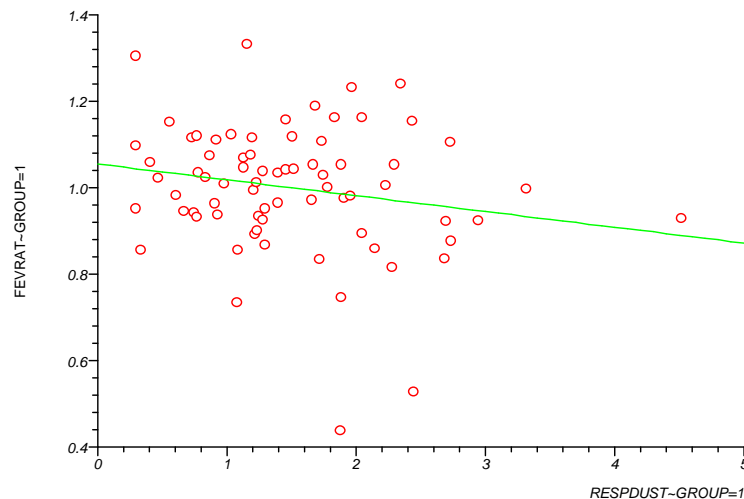
- Equation:** FEVRAT~GROUP=1 = -0.036595 RESPDUST~GROUP=1 + 1.054884
- Standard Error of slope = 0.022004
- 95% CI for population value of slope = -0.08047 to 0.007279
- Correlation coefficient ( $r$ ) = -0.19364 ( $r^2$  = 0.037496)
- 95% CI for  $r$  (Fisher's  $z$  transformed) = -0.405635 to 0.038126
- $t$  with 71 DF = -1.663119
- Two sided  $P$  = 0.1007
- Power (for 5% significance) = 37.09%
- Correlation coefficient is not significantly different from zero

It also gives the correlation coefficient  $r$  and statistical test of whether the gradient of the line is significantly different from zero. It gives a  $p$ -value of 0.1007 for 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. Note that the output states "Correlation coefficient is not significantly different from zero". This assumes that one is using a two-tailed test with a 5% significance level.

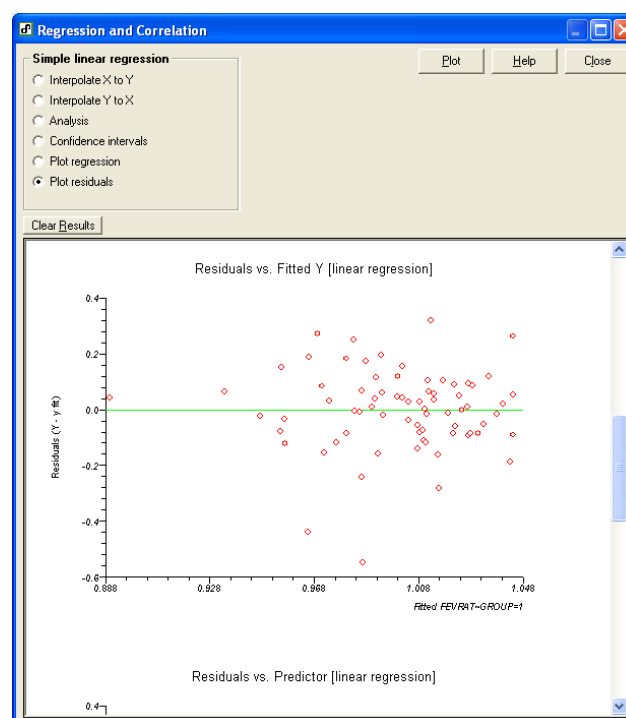
The result for  $r^2$  is also useful. This is an estimate of the proportion of the variance explained by the model. A line that fits the data perfectly will have an  $r^2$  equal to 1. Whereas a line that does not explain anything in the data will have an  $r^2$  of zero. The value of  $r^2$  equal to 0.037 is therefore not at all good – only 3.7% of the variation in the data is being explained by the regression line

## Model Checking

The linear regression model described by the coefficients allows one to estimate a predicted value. The difference between the observed value and the predicted value is called a residual. Where a model fits badly the regression line will have large residuals. If you consider the scatter plot above for FEV ratio compared to respiratory dust the residuals will be large. It is possible to get a plot of the regression line with the data by selecting **Plot regression** shown below.



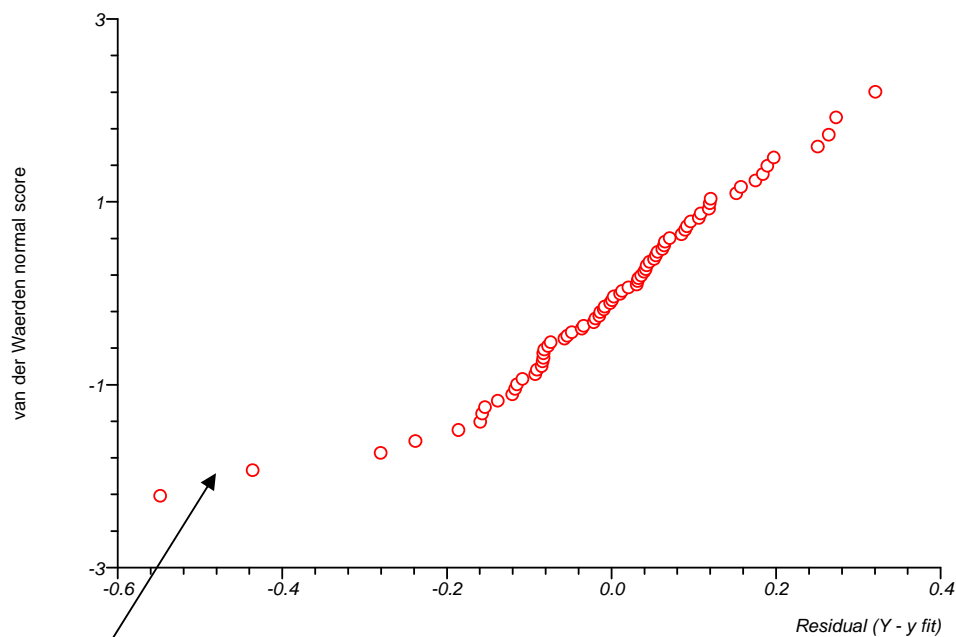
From the plot above two points appear to have large negative residuals suggesting that they are outliers. One of the assumptions of a regression model is that the residuals will have a normal distribution. In StatsDirect this may be carried out by plotting the residuals. Just select plot residual options.





This gives three plots, the first two are of the residuals against the fitted values and residuals against predicted values. From this it can be seen that there are two observations with large negative residuals. The second plot is a normal probability plot shown below. If the residuals are normally distributed the plotted points are in a straight line. Unlike SPSS StatsDirect does not show a line through the normal probability plot so that it is difficult to assess this plot. If one covers up the two points with large negative residuals in the plot below, it suggests that the data are approximately normally distributed. If the data were skewed the points would bulge away from the line.

Normal Plot for Residuals [linear regression]



Large negative residuals.

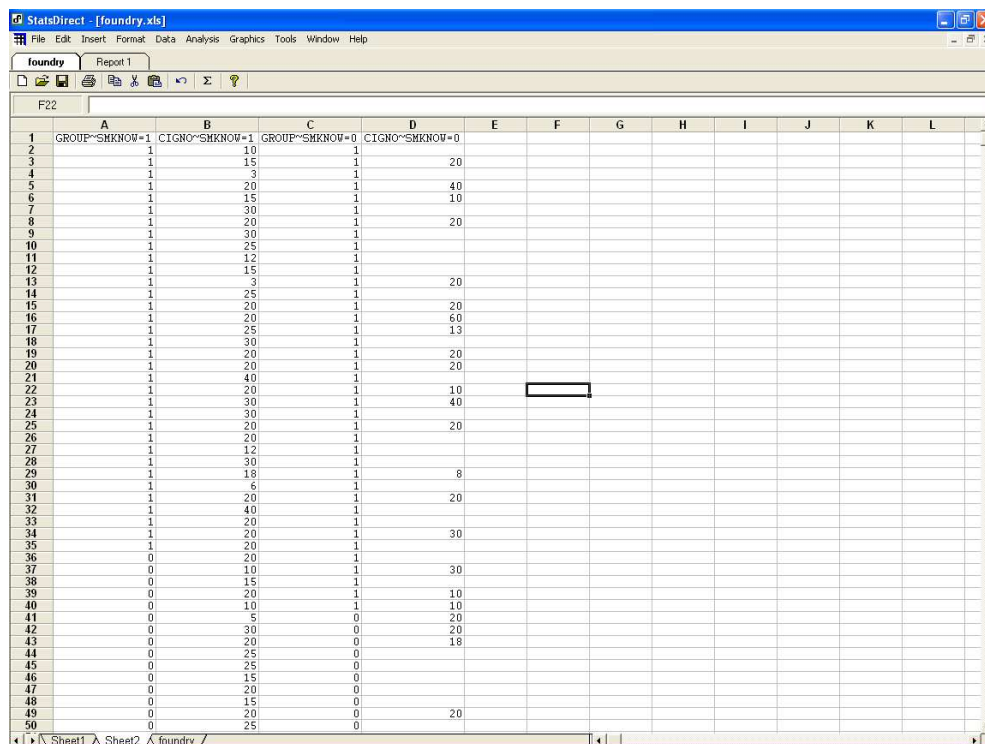
To check the robustness of the analysis it could be rerun excluding the two points with large residuals. The multiple regression procedure (rather than the simple regression) is required to calculate residuals and carry-out more complex statistical analyses.

**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 you can compare the number of cigarettes smoked by "exposed" and "non-exposed" workers who currently smoke.

Either sort the data in the main work sheet so that only current smokers can be selected or use **Group split** to create a new work sheet. Return to the Foundry work sheet. In the Group split option under Data select SMKNOW as the variable to split the data then select the variables GROUP and CIGNO to create separate columns for current smokers and current non-smokers as shown below.



	A	B	C	D	E	F	G	H	I	J	K	L
1	GROUP~SMKNOW=1	CIGNO~SMKNOW=1	GROUP~SMKNOW=0	CIGNO~SMKNOW=0								
2	1	10	1									
3	1	15	1	20								
4	1	3	1									
5	1	20	1	40								
6	1	15	1	10								
7	1	30	1									
8	1	20	1	20								
9	1	30	1									
10	1	25	1									
11	1	12	1									
12	1	15	1									
13	1	3	1	20								
14	1	25	1									
15	1	20	1	20								
16	1	20	1	60								
17	1	25	1	13								
18	1	30	1									
19	1	20	1	20								
20	1	20	1	20								
21	1	40	1									
22	1	20	1	10								
23	1	30	1	40								
24	1	30	1									
25	1	20	1	20								
26	1	20	1									
27	1	12	1									
28	1	30	1									
29	1	18	1	8								
30	1	6	1									
31	1	20	1	20								
32	1	40	1									
33	1	20	1									
34	1	20	1	30								
35	1	20	1									
36	0	20	1									
37	0	10	1	30								
38	0	15	1									
39	0	20	1	10								
40	0	10	1	10								
41	0	5	0	20								
42	0	30	0	20								
43	0	20	0	18								
44	0	25	0									
45	0	25	0									
46	0	15	0									
47	0	20	0									
48	0	15	0	20								
49	0	20	0									
50	0	25	0									

The two left-hand columns are related to the current smokers. You may wish to replace the column headings to prevent confusion.

**Exercise:** Use **Descriptive** and **Frequencies** to generate summary statistics for the numbers of cigarettes smoked by current smokers for exposed and un-exposed workers separately.

From the **Descriptive** command you should obtain the following f tables

### **Descriptive statistics**

Variables	CIGNO (Group=1)	CIGNO (Group=0)
Valid Data	34	20
Mean	20.7059	18.75
Variance	76.9412	33.8816
SD	8.7716	5.8208
SEM	1.5043	1.3016
Lower 95% CL	17.6453	16.0258
Upper 95% CL	23.7664	21.4742
Geometric Mean	18.2073	17.5964
Skewness	0.0916	-0.5262
Kurtosis	3.1335	3.3146
Maximum	40	30
Upper Quartile	26.25	20
Median	20	20
Lower Quartile	15	15
Minimum	3	5
Range	37	25
Variance coeff.	0.4236	0.3104
Sum	704	375
Centile 5	3	5.25

Examining this data one might wish to test if smokers exposed to dust were more likely to smoke more as the mean is larger and the upper quartile also ( although the medians are the same). To test these use the Mann-Whitney U-test. Select Non-parametric from the Analysis menu then select Mann-Whitney. Choose GROUP as the group identifier and then select CIGNO. This should generate the following output.

### **Mann-Whitney U test**

Observations (x) in CIGNO~SMKNOW=1\_GROUP~SMKNOW=1\_1 = 34 median = 20 rank sum = 976  
 Observations (y) in CIGNO~SMKNOW=1\_GROUP~SMKNOW=1\_0 = 20 median = 20  
 U = 381 U' = 299

Exact probability (adjusted for ties):

Lower side P = 0.225 ( $H_1$ : x tends to be less than y)

Upper side P = 0.775 ( $H_1$ : x tends to be greater than y)

Two sided P = 0.4499 ( $H_1$ : x tends to be distributed differently to y)

95.1% confidence interval for difference between medians or means:

K = 231 median difference = 0

CI = 0 to 5

In the tables above note the median for each group and the significance level. Hence, one can conclude that there is no difference between the median number of cigarettes smoked by "exposed" and "non-exposed" workers. The output also gives difference in the medians (0) and the confidence interval for the difference of the median (0 to 5). Note also that the coverage is given as 95.1% rather than 95%.

# COMPARISONS OF RELATED OR PAIRED VARIABLES

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

## 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 measured FVC select a **Paired T** in the **Parametric** submenu. Then chose Group by Column and select the columns FEVMEAS and FEVPRED as shown below.

The screenshot shows the StatsDirect software window with a data table. The table has columns labeled A through P. The 'FEVMEAS' column is highlighted in the 'Select 2 matched columns' dialog box. The 'FEVPRED' column is also highlighted. The dialog box also shows options for 'Batch mode', 'Groups by column', and 'Groups by identifier'.

Results are given below

### Paired t test

For differences between FEVMEAS and FVCMEAS:

Mean of differences = -1.0196 (n = 136)

Standard deviation = 0.373

Standard error = 0.032

95% CI = -1.0829 to -0.9564

df = 135

t = -31.8789

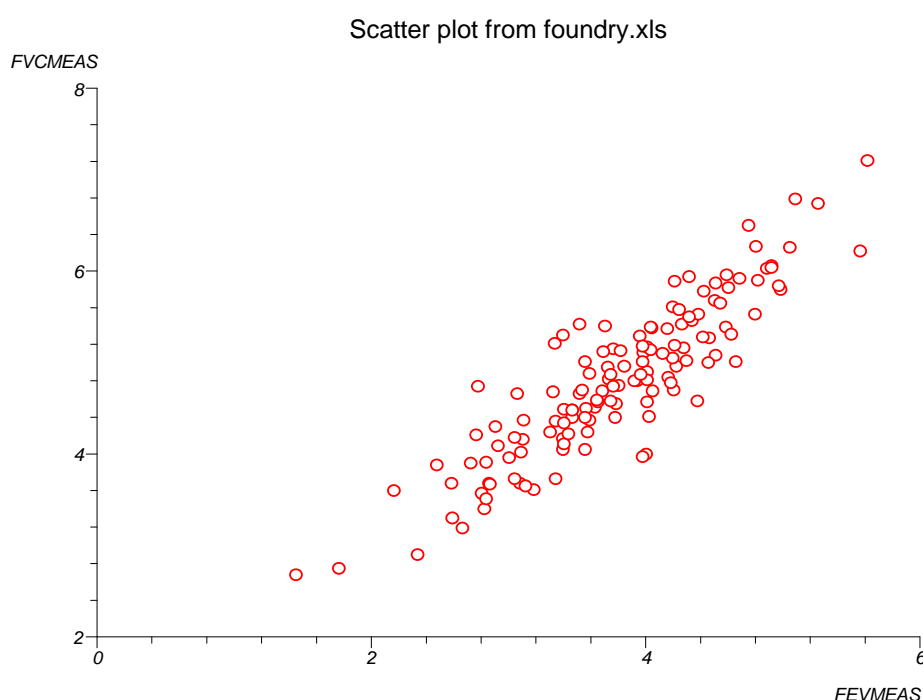
One sided P < 0.0001

Two sided P < 0.0001

It is readily apparent that mean *measured FVC* is greater than mean *measured FEV*. You could report this as “Measured FVC was significantly higher than measured FEV (diff=1.02, 95% c.i. 0.96 to 1.08,  $p<0.0001$ )”

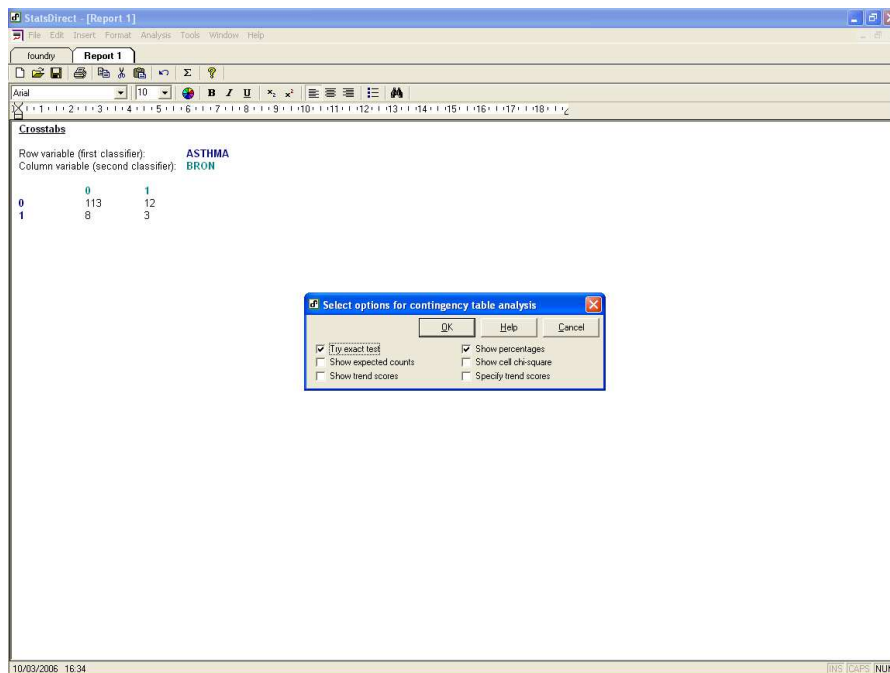
**Exercise:** Compare the mean FEV ratio with the mean FVC ratio.

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. One gets 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.

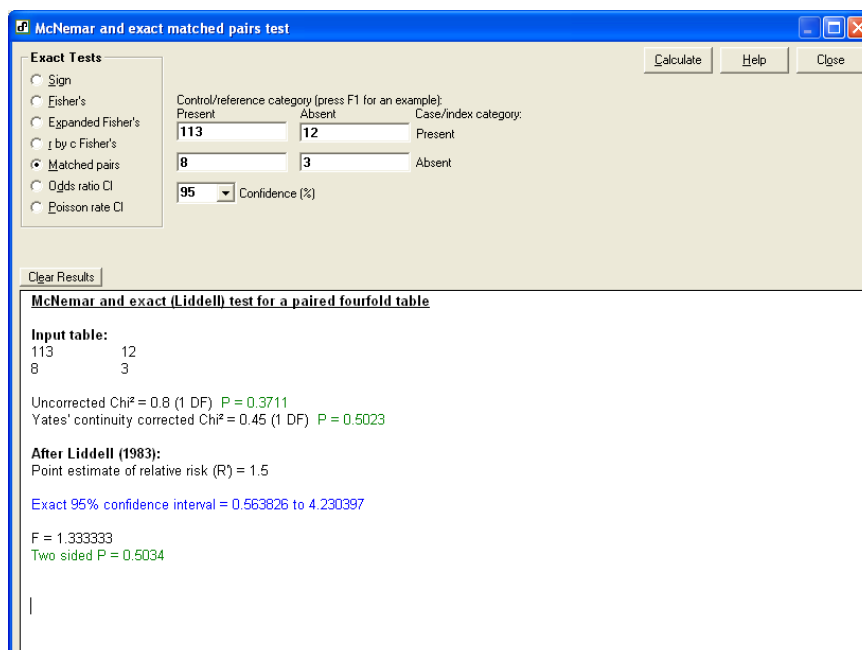


## Analysis of Binary Outcomes that are Related

Suppose we wish to compare the proportion of workers who had bronchitis symptoms with the proportion who had asthma symptoms. Data are paired and the appropriate method statistical inference is McNemar's test. First one should construct the **Crosstabs** procedure in **Analysis menu**. Select ASTHMA as the **First Classifier Row** and BRON as the **Second Classifier row** to get a 2x2 cross-classification. We obtain the following screen from which we can see that only 3 workers had symptoms of both and 8 had symptoms of just Asthma or 12 just Bronchitis. From this table we can calculate 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, which takes account of the pairing of the variable, by selecting **Yes**.



By clicking **Analysis**, **Exact Test on Counts** then **Matched Pairs** the panel below will appear. In this select Matched pair's radio button then click **Calculate** to obtain the results for McNemar's test comparing the proportion of workers with Asthma symptoms as compared to Bronchitis symptoms.



The p-value for the McNemar test with Yates continuity correction is 0.5023. This is not significant at a 5% level so we conclude that symptoms of bronchitis are no more common in this population than symptoms of asthma.

## Related Ordinal Data

For ordered categorical or quantitative variables that are not plausibly normal the suggested procedure is to use the **Wilcoxon** procedure.

## SUMMARY STATISTIC METHODS

StatsDirect has a set of procedures that are based on summary statistics. These methods are based on statistics obtained from previous analysis. For example 0.11 or 11% (15/136) workers had symptoms of Bronchitis. A confidence interval for that proportion using StatsDirect follows. In the **Analysis** menu, choose **Proportions** then **Single** to obtain the following screen in to which the numerator and denominator of the proportion have been entered.

**Tests on proportions**

**Proportion Tests**

☒ Single  
☐ Paired  
☐ Two independent

Total number of observations: 136  
Number responding: 15  
Probability of success on each trial: .5  
Confidence (%): 95

**Single proportion**

Total = 136, response = 15  
Proportion = 0.110294  
Exact (Clopper-Pearson) 95% confidence interval = 0.06306 to 0.175383  
Using null hypothesis that the population proportion equals 0.5  
Binomial one sided P < 0.0001  
Binomial two sided P < 0.0001  
Approximate (Wilson) 95% mid-P confidence interval = 0.067988 to 0.174011  
Binomial one sided mid-P < 0.0001  
Binomial two sided mid-P < 0.0001

The 95% confidence interval is 0.0631 to 0.1754. One might report this as “the proportion of workers who have experienced symptoms of Bronchitis was 11% (15/136, 95% c.i. 6% to 17%)”

### t-test Using summary Data

Summary methods are particularly useful where we want to compare data from different studies where the raw data is not available. Suppose we wanted to compare the data in this study with that reported in say that from workers exposed to paper dust. Let us assume that the summary statistics reported for the FEV Ratio for workers in a study of the effects of paper dust are: mean=0.97, standard deviation=0.14, sample size n=115. From the **Descriptives** option the corresponding values for workers exposed to dust are: mean=1.0007, standard deviation=0.1479, sample size n=73). To carry out a t-test comparing mean FEV Ratio for the foundry workers with the paper workers on selects the option **Parametric** then **Summary Data t** then **Un-paired**.



The screenshot shows the StatsDirect software interface with a data table and a menu path for an unpaired t-test.

**Menu Path:** Analysis > Parametric > Unpaired t > Unpaired

**Data Table:**

	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	FEVMEAS	FEVREFD													
2	3.40	3.59													
3	2.83	3.39													
4	3.93	4.26													
5	4.01	4.25													
6	4.75	4.52													
7	4.60	3.73													
8	4.01	4.45													
9	2.58	2.97													
10	4.50	4.18													
11	4.19	4.21													
12	3.51	3.92													
13	2.92	3.91													
14	3.18	4.03													
15	2.76	3.24													
16	3.06	3.59													
17	3.95	4.18													
18	3.77	3.24													
19	3.91	3.82													
20	4.03	4.44													
21	4.04	3.99													
22	3.81	3.47													
23	3.32	3.17													
24	3.40	3.50													
25	4.01	3.59													
26	2.80	2.97													
27	4.37	4.14													
28	3.63	3.24													
29	4.68	4.22													
30	4.91	4.68													
31	2.47	3.36													
32	2.16	2.94													
33	3.53	3.94													
34	3.64	3.24													
35	3.69	4.12													
36	4.31	3.50													
37	3.98	3.72													
38	3.97	3.87													
39	4.80	4.51													
40	4.16	4.00													
41	2.72	3.33													
42	3.39	3.63													
43	4.81	4.56													
44	3.59	3.66													
45	5.25	4.53													
46	3.04	3.24													
47	3.68	3.89													
48	3.97	3.80													
49	5.05	4.50													
50	4.54	4.33													

This gives the following panel into which the summary statistics can be entered as shown.

Clicking on OK gives the output

The dialog box 'Student's t Tests from Summary Data' contains the following input fields:

- Sample 1 size: 115
- Sample 1 mean: 0.97
- Sample 1 standard deviation: 0.14
- Sample 2 size: 73
- Sample 2 mean: 1.007
- Sample 2 standard deviation: 0.1479

Buttons: OK, Cancel

### Unpaired t test

Mean of \* sample 1 from summary = 1.007 (n = 115)

Mean of \* sample 2 from summary = 0.97 (n = 73)

### Assuming equal variances

Combined standard error = 0.0214

df = 186

t = 1.7277

One sided P = 0.0429

Two sided P = 0.0857

95% confidence interval for difference between means = -0.0792 to 0.0052

### Assuming unequal variances

Combined standard error = 0.0217

df = 147.1343

t(d) = 1.7065

One sided P = 0.045

Two sided P = 0.09

95% confidence interval for difference between means = -0.0798 to 0.0058

### Comparison of variances

Two sided F test is not significant

No need to assume unequal variances



As the standard deviations are very similar (0.14 and 0.1479) and this is confirmed by the F-test, one considers the results under Assuming Equal Variance. Whilst it is tempting to present the One-sided P because it is statistically significant at a 5% level rather than the Two-sided P, usual practice in Medical research is to report the latter unless strong arguments can be made for a One-sided test a priori. Based on these results one might report “the difference in mean FEV ratio for paper workers exposed to dust as compared to foundry workers exposed to dust was -0.04 (95% c.i. – 0.079 to 0.005,  $p=0.086$ ).

## Comparison of Proportions

Suppose now that for the paper workers it had been reported symptoms of Bronchitis was 22% (25/115). For dust exposed workers in the foundry the percentage is 15% (11/73). A z-test comparing these two proportions/percentages can be carried out using the summary data. Going to **Proportions** then **Two independent** on the **Analysis** menu one gets the following panel into which the data is entered.

**Tests on proportions**

**Proportion Tests**

☐ Single

☐ Paired

☒ Two independent

25 Total observations in SAMPLE 1

115 Number responding in SAMPLE 1

11 Total observations in SAMPLE 2

73 Number responding in SAMPLE 2

95 Confidence (%)

Calculate Help Close

Clear Results

**Two independent proportions**

Total 1 = 115, response 1 = 25

Proportion 1 = 0.217391

Total 2 = 73, response 2 = 11

Proportion 2 = 0.150685

Proportion difference = 0.066706

Approximate (Miettinen) 95% confidence interval = -0.052548 to 0.175028

Exact two sided (mid) P = 0.2622

Standard error of proportion difference = 0.058882

Standard normal deviate (z) = 1.132879

Approximate two sided P = 0.2573

Approximate one sided P = 0.1286

The difference in proportions is 0.0667 with confidence interval –0.0525 to 0.175. This might be reported as “The percentage of foundry workers with symptoms of Bronchitis was 15% (11/73) compared to paper workers in study X of 22% (25/115). This difference was not statistically significant at the 5% level (Diff.=7%, Exact two-sided  $p=0.26$ , 95% c.i. –5% to 18%) . “In making comparison with published data one should of course considered whether the method of data collection were comparable. For example were symptoms of Bronchitis ascertained in the same way.

## CHOOSING THE APPROPRIATE STATISTICAL PROCEDURE

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

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

\* Not illustrated

+ Method that can be used with summary statistics

For a more comprehensive chart for selecting methods see

<http://www.graphpad.com/www/book/choose.htm>