Basic Statistics in MS Excel

Excel 97 version

OCTOBER 1998

Acknowledgements

These notes were originally developed for Excel 5.0 by Catherine Davies and were based on similar ones produced for the Minitab package by Linda Williams.

The Excel 97 version has been prepared by Helen Talbot of Edinburgh University Computing Service and Tim Jackson of BioSS.

CONTENTS

1 Introduction

1.1 Aims and Expectations

Microsoft Excel 97 runs on any PC with a Microsoft Windows 95, 98 or NT interface. Some familiarity with both PCs and the Windows environment is useful, but having none will not exclude you from being able to use the package. For ease of use a mouse or compatible pointing device is strongly recommended although this is by no means a necessity.

By the end of the course you should be able to use Excel to explore your data. Microsoft Excel is an extremely powerful package, with a wide range of applications, which extend far beyond those dealt with in this course. The simplicity of both entering and manipulating data makes this a useful package for performing preliminary statistical exploration of data.

1.2 A Brief Summary of the Course Content

The course will look at how to explore and examine data by

- summary statistics: measures of average (mean, median) and spread (standard deviation)
- histograms
- scatter plots and other diagrams
- tables
- general data manipulation
- t-tests
- analysis of variance
- regression
- filtering

The pointing finger ☞ indicates when you are expected to use the computer to practice the concepts explained earlier.

1.3 Types of data

Before you do any type of analysis with your data, you should consider what type of data you have. Data fall into two broad categories: continuous and discrete.

Continuous data are usually measurements or observations of some sort, e.g. weight, height. They can usually take any value in some number range (e.g. 2.1Kg, -8.9964 °C, 4.1415 m) and will have associated units (m, kg, °C etc)

Discrete data can only take integer (whole number) values, e.g. 7, 24, 907. For example, the number of animals in an experiment is discrete. A subset of discrete data is categorical data, where the numbers denote a category, rather than a numerical value, e.g. 1=male, 2=female

1.4 Miscellaneous

1.4.1 Organising your data within workbooks

A **workbook** is like a file, into which one can place as many **worksheets**, **dialog sheets** (for the creation of customised dialog boxes), and **module sheets** (for editing and recording macros) as memory allows. Usually data are entered, manipulated and printed from worksheets and it is these that we will concentrate on.

1.4.2 Links between Excel and other applications

One of the most powerful features of Excel is its ability to establish links not only between objects in different workbooks (enabling the worksheets of one book to be updated automatically when those of the other are updated by the user), but also **remote links** with other applications. These are used to incorporate data from a file created in another application. To establish these kinds of links Excel uses object linking and embedding (OLE) and dynamic data exchange (DDE). The ability to form such links has numerous advantages, including the automatic update of any object embedded in a document of another application when the original Excel object is updated in any way. There are situations when it is better not to use OLE and DDE and these include using a spreadsheet in a number of different ways in a single document.

1.4.3 Output

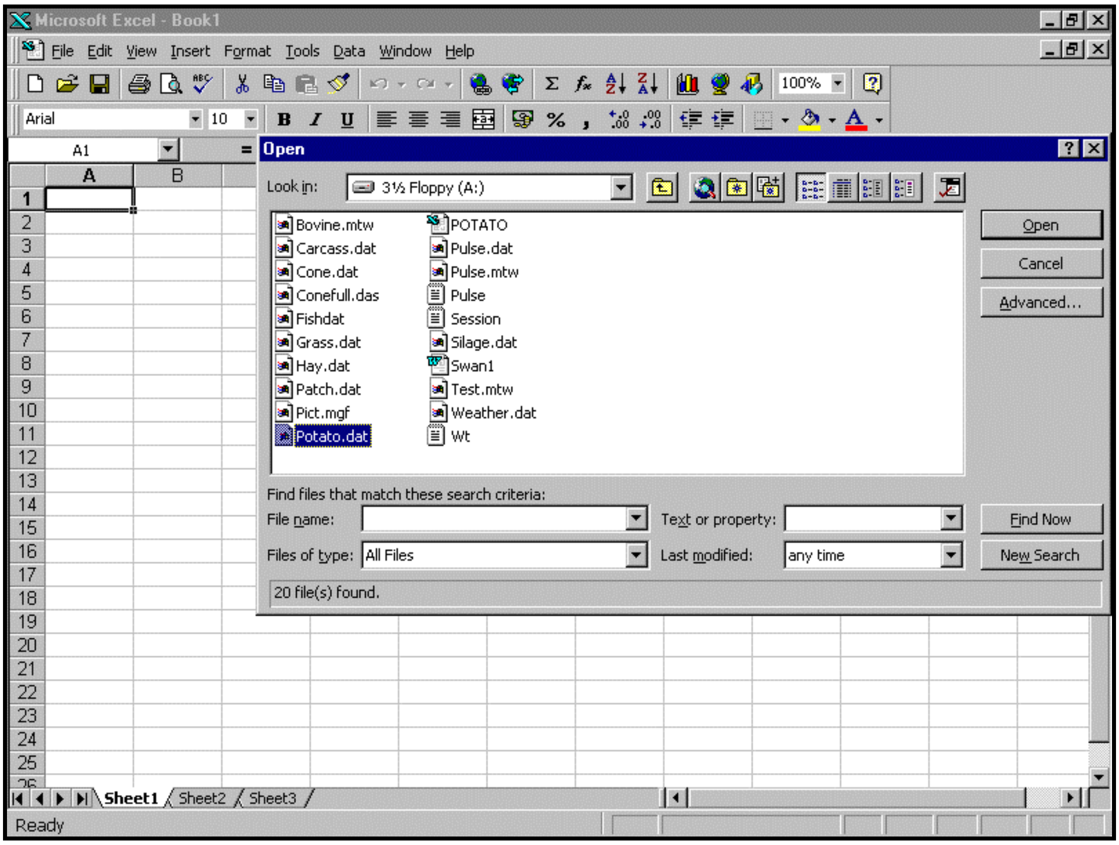
When an analysis tool is applied to a dataset the user is given the option of saving the output table to the same sheet as the input, to a separate sheet in the same workbook, or to a new workbook. Once output, the information can easily be moved to a new location using the cut and paste commands, and formatted as desired.

1.5 Entering and storing data

1.5.1 From ascii files

If you have your data in the form of an ascii file which you want to import into Excel, simply click on the **File** menu and select **Open** from the drop-down menu. Next select the directory and file from the dialog box which appears and click on the OK button. This is illustrated in Figure 1.5.1

Note if no file list appears make sure **Files of type**: contains **All Files**. On identifying the file to be opened the **TextWizard** will be activated . This displays a series of dialog boxes guiding you through the steps required to specify how you want text distributed across columns (see Figure 1.5.2). Excel automatically recognises column breaks in the document and displays these in a separate dialog box. The user is permitted to modify these, adding or deleting line breaks in the process.



Look in: 3½ Floppy (A:)

Bovine.mtw
Carcass.dat
Cone.dat
Conefull.das
Fishdat
Grass.dat
Hay.dat
Patch.dat
Pict.mgf
Potato.dat
POTATO
Pulse.dat
Pulse.mtw
Pulse
Session
Silage.dat
Swan1
Test.mtw
Weather.dat
Wt

Find files that match these search criteria:

File name:

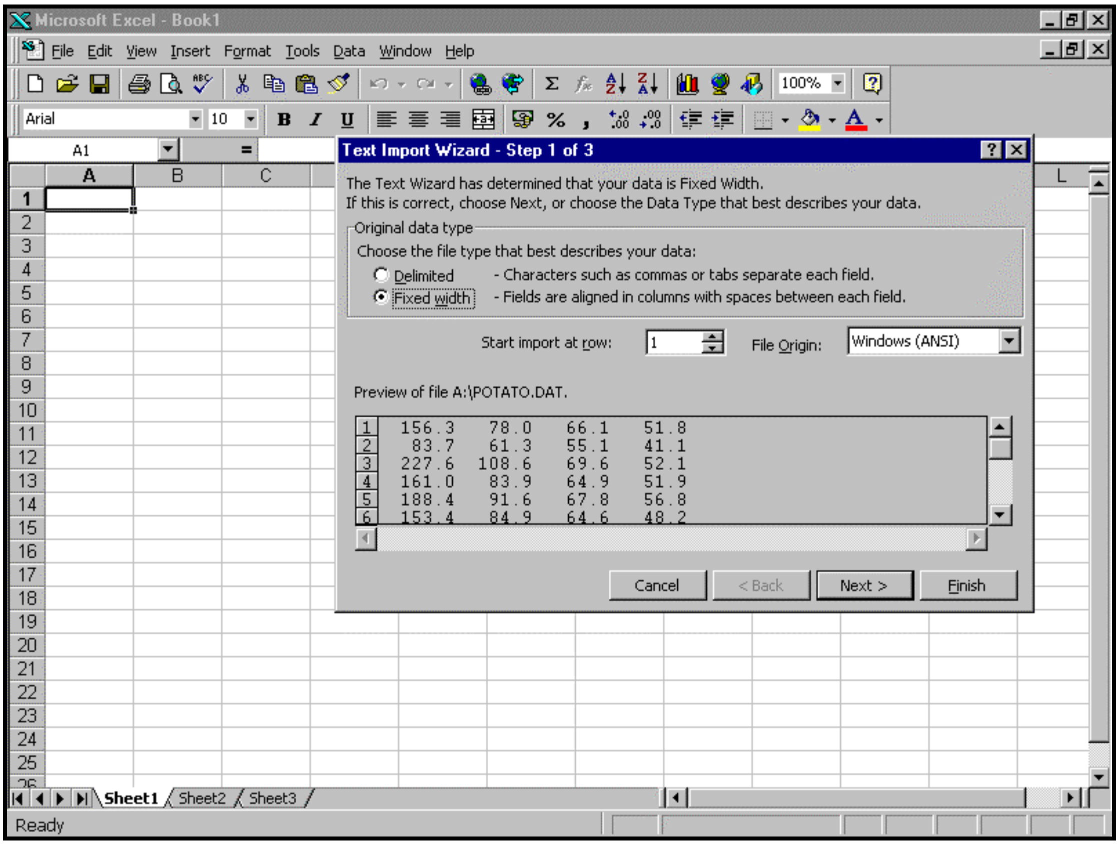
Text or property:

Files of type: All Files

Last modified: any time

20 file(s) found.

Figure 1.5.1


Microsoft Excel - Book1
File Edit View Insert Format Tools Data Window Help
Arial
10
100%
A1
Text Import Wizard - Step 1 of 3
Cancel
< Back
Next >
Finish
Sheet1
Sheet2
Sheet3
Ready

The Text Wizard has determined that your data is Fixed Width.
If this is correct, choose Next, or choose the Data Type that best describes your data.

Original data type

Choose the file type that best describes your data:

○ Delimited - Characters such as commas or tabs separate each field.

● Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: 1

File Origin: Windows (ANSI)

Preview of file A:\POTATO.DAT.

```
1   156.3    78.0    66.1    51.8
2    83.7    61.3    55.1    41.1
3   227.6   108.6    69.6    52.1
4   161.0    83.9    64.9    51.9
5   188.4    91.6    67.8    56.8
6   153.4    84.9    64.6    48.2
```

Figure 1.5.2

From Minitab, Saving as an Excel Worksheet
A Minitab data worksheet can be saved directly as an Excel worksheet by clicking on **File**, selecting **Save Worksheet As**, and selecting **Excel** in the **Save as Type** box. The worksheet is given the extension .XLS and can be opened in Excel in the same way as any other Excel worksheet.

Cut-and-Paste
In Minitab, select the block of data you wish to import. Click on **Edit**, then on **Copy**. Then switch to Excel, and activate a worksheet. Highlight the position of the first cell of the first column that you want the data to occupy, then click on the **Paste** command located in the **Edit** menu. This method will work in any Windows supported package.

1.5.2 Inserting data within a worksheet

In order to perform any statistical analysis on data, the data to be analysed must first be entered and organised into the columns and rows of a worksheet. **Text labels** can be added at the beginning of the columns or rows containing the data, and the worksheet may be formatted, either before or after analysis has taken place. If text labels are included these should also be included in the **input range** of the data. Excel will then use these in any output table it generates. Omitting labels will result in Excel providing default labels such as variable1 and variable2.

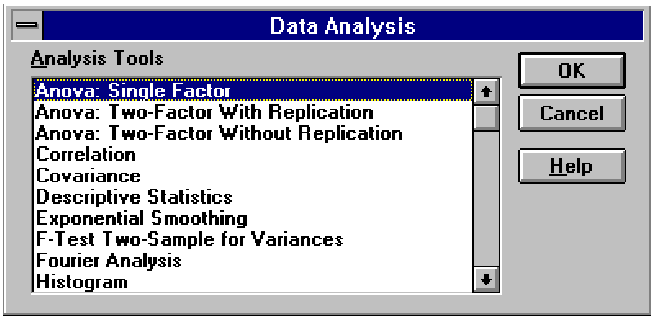
1.5.3 Saving and retrieving your work

Saving and retrieving an Excel workbook
To save a workbook simply select **Save** from the **File** menu, choose the directory and drive where you want to save the workbook, and specify a name, which you are advised to make 8 characters or less in length. Excel will automatically add the extension .XLS to the end of the filename unless you specify that you want to save the file as anything other than an Excel Workbook. Click on OK when you are satisfied. Retrieving a workbook is very similar to saving, except that you select **Open** from the **File** menu.

Saving a worksheet as a text file
The data in a worksheet can be saved as a text file by clicking on **File**, selecting **Save As**, and selecting either **Text (tab delimited)** or **Formatted text (space delimited)** from the **Save As Type** box. This saves the data in a file with extension .TXT or .PRN respectively. The saved file can then be read by both word processing packages and most statistical packages.

1.6 Tools for data analysis

In Microsoft Excel there is a set of special analysis tools known as the **Analysis ToolPak**. This provides a number of basic statistical analyses which can be applied to a wide range of data. The Analysis ToolPak may be activated by choosing **Data Analysis** from the **Tools** menu. To perform an analysis select a tool from those listed and enter the appropriate input range, output range and options in the dialog box which appears.



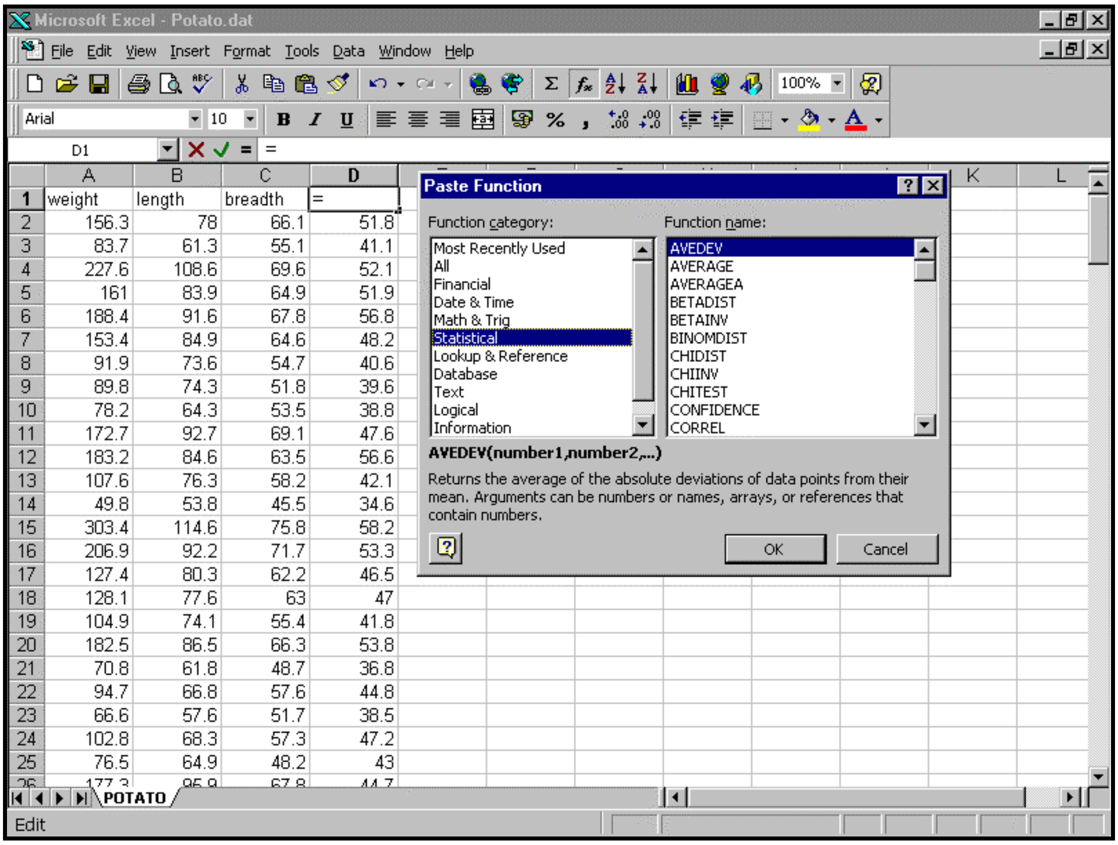
N.B If the **Data Analysis** command does not appear in the **Tools** menu select **Add-Ins** from the **Tools** menu and check the **Analysis ToolPak** option. If this is not listed you will need to run the Setup program to install the Analysis ToolPak.

The analyses offered in the ToolPak include:

- Anova
- Correlation
- Covariance
- Descriptive statistics
- Exponential smoothing
- F-test : Two sample for variances
- Fourier analysis
- Histogram
- Moving average
- Random number generation
- Rank and percentile
- Regression
- Sampling
- t-test
- z-test

1.7 Functions

In addition to the tools provided in the Analysis ToolPak there are also a large number of **worksheet statistical functions** available. Many of these are used by the analysis tools above, but there are others which are unique to the **Function Wizard**. It looks like this on the top tool bar fx , and only appears when a cell is selected by the user.


Microsoft Excel - Potato.dat
File Edit View Insert Format Tools Data Window Help
100%
Arial
10
B I U
D1
=
POTATO
Edit

	A	B	C	D
1	weight	length	breadth	=
2	156.3	78	66.1	51.8
3	83.7	61.3	55.1	41.1
4	227.6	108.6	69.6	52.1
5	161	83.9	64.9	51.9
6	188.4	91.6	67.8	56.8
7	153.4	84.9	64.6	48.2
8	91.9	73.6	54.7	40.6
9	89.8	74.3	51.8	39.6
10	78.2	64.3	53.5	38.8
11	172.7	92.7	69.1	47.6
12	183.2	84.6	63.5	56.6
13	107.6	76.3	58.2	42.1
14	49.8	53.8	45.5	34.6
15	303.4	114.6	75.8	58.2
16	206.9	92.2	71.7	53.3
17	127.4	80.3	62.2	46.5
18	128.1	77.6	63	47
19	104.9	74.1	55.4	41.8
20	182.5	86.5	66.3	53.8
21	70.8	61.8	48.7	36.8
22	94.7	66.8	57.6	44.8
23	66.6	57.6	51.7	38.5
24	102.8	68.3	57.3	47.2
25	76.5	64.9	48.2	43
26	177.3	95.9	67.8	44.7

Paste Function

Function category:

- Most Recently Used
- All
- Financial
- Date & Time
- Math & Trig
- Statistical
- Lookup & Reference
- Database
- Text
- Logical
- Information

Function name:

- AVEDEV
- AVERAGE
- AVERAGEA
- BETADIST
- BETAINV
- BINOMDIST
- CHIDIST
- CHIINV
- CHITEST
- CONFIDENCE
- CORREL

AVEDEV(number1,number2,...)

Returns the average of the absolute deviations of data points from their mean. Arguments can be numbers or names, arrays, or references that contain numbers.

Figure 1.7.1

The functions can be viewed by activating the **Function Wizard** button on the standard toolbar or by using

Insert > Function...

and then select the required worksheet function from the Function Wizard. Figure 1.7.1 shows the dialogue box produced by the Function Wizard. Note the large number of functions and how they are categorised in the left-hand box to make them easier to find.

1.8 Formatting worksheets

In order to help you perform the task below, and others given in this course, it may be useful to familiarise yourself with a few basic procedures for formatting worksheets. If you are already an experienced Excel user skip to the next section, otherwise read on.

Shortcut menu:
Many of the commands used whilst editing/formatting a worksheet are accessible via the shortcut menu. This can be activated by highlighting an object (e.g. a worksheet cell) and clicking on the right mouse button.

Inserting rows:
Highlight the row where you want to insert an extra row and activate the shortcut menu.

Select **Insert...**. Alternatively select **Insert > Row** from the standard menubar.
The inserted row will then precede the row you highlighted.

Changing the row/column width:
Highlight the row(s) or column(s) to be formatted, activate the shortcut menu and select **Row height/Column width**. Alternatively select **Format > Row height/Column width** from the standard menubar.

Formatting cells:
Select the cells to be formatted and activate the shortcut menu. Click on **Format cells** and make the necessary changes to the dialog box which appears. Alternatively select **Format > Cells...** from the standard menubar.

Referencing cells
Throughout this document you will need to reference various sections of the worksheet. To do this use for example $A$2:$A$10 to denote all items of column A between row 2 and row 10. Similarly you can use $A$2:$A$10,$B$2:$B$10 to denote using columns between A and B between rows 2 and 10.

☞**Exercise 1.1:** Import the data from the file POTATO.DAT and name the columns ‘weight’, ‘length’, ‘breadth’, and ‘depth’. Save the workbook as an Excel workbook named POTATO.XLS.

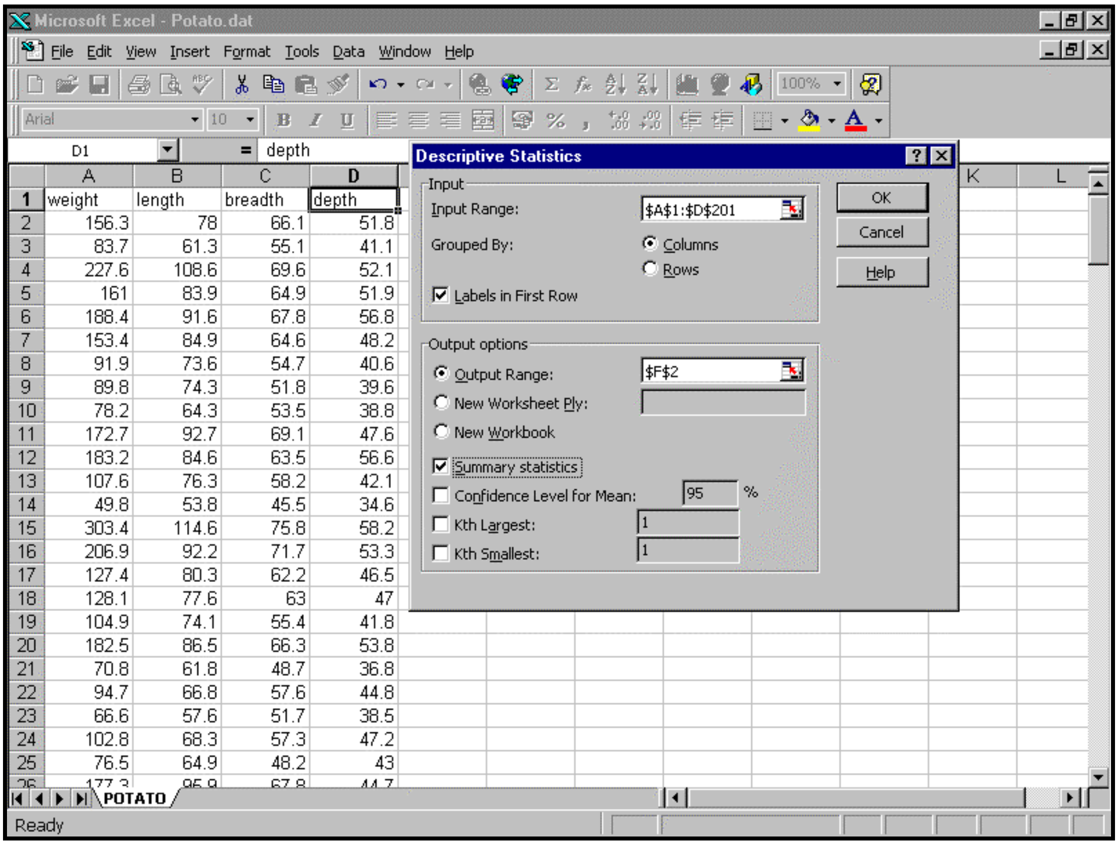
2 Exploratory data analysis for one variable

2.1 Introduction to exploratory methods

Exploratory data analysis methods are used to examine the data, prior to using more formal methods (e.g. t-test, ANOVA etc.). They are useful for looking for patterns, identifying outliers, and examining possible deviations from assumptions (e.g. Normality).

2.2 The DESCRIPTIVE STATISTICS tool

By now, you should have the POTATO data stored in your worksheet, and you should also have named the columns. The next step is to examine some basic statistics for the data. This we do using the **Descriptive Statistics** option, found under the **Tools** menu in the **Data Analysis Menu**.


Microsoft Excel - Potato.dat
File Edit View Insert Format Tools Data Window Help
Arial 10 100%
D1 = depth
POTATO
Ready
OK
Cancel
Help

	A	B	C	D
1	weight	length	breadth	depth
2	156.3	78	66.1	51.8
3	83.7	61.3	55.1	41.1
4	227.6	108.6	69.6	52.1
5	161	83.9	64.9	51.9
6	188.4	91.6	67.8	56.8
7	153.4	84.9	64.6	48.2
8	91.9	73.6	54.7	40.6
9	89.8	74.3	51.8	39.6
10	78.2	64.3	53.5	38.8
11	172.7	92.7	69.1	47.6
12	183.2	84.6	63.5	56.6
13	107.6	76.3	58.2	42.1
14	49.8	53.8	45.5	34.6
15	303.4	114.6	75.8	58.2
16	206.9	92.2	71.7	53.3
17	127.4	80.3	62.2	46.5
18	128.1	77.6	63	47
19	104.9	74.1	55.4	41.8
20	182.5	86.5	66.3	53.8
21	70.8	61.8	48.7	36.8
22	94.7	66.8	57.6	44.8
23	66.6	57.6	51.7	38.5
24	102.8	68.3	57.3	47.2
25	76.5	64.9	48.2	43
26	177.3	95.9	67.8	44.7

Descriptive Statistics

Input
Input Range: $A$1:$D$201
Grouped By: ● Columns ○ Rows
☑ Labels in First Row

Output options
● Output Range: $F$2
○ New Worksheet Ply:
○ New Workbook
☑ Summary statistics
□ Confidence Level for Mean: 95 %
□ Kth Largest: 1
□ Kth Smallest: 1

Figure 2.2.1

The order in which we call up the Descriptive Statistics tool is:

1. **Select the Tools menu at the top of the screen.**
2. **Select Data Analysis... from the drop down menu.**
3. **Select Descriptive Statistics from the list of data analysis options.**

A box headed **Descriptive Statistics** should now appear on your screen as shown in Figure 2.2.1. It allows you to choose the columns you want to describe, the output option and the statistics you want.

The above method of describing procedures to choose from the menus is fairly time consuming and in future shall be abbreviated using arrows. For example: **Tools > Data Analysis... > Descriptive Statistics** represents the above three step procedure.

Once the Descriptive Statistics tool box is displayed you can select the variables that you wish to summarise by activating the input range box and then highlighting the input range on the worksheet. Remember to include any labels representing the data sets if these exist. Next check or uncheck the output options listed depending on the statistics you wish to display. The kth largest option will display the kth largest value in each of the data sets and the kth smallest, the kth smallest value. A 95% mean confidence interval is also available as well as summary statistics, including such statistics as the mean, median, standard deviation, and skewness of the data sets.

If you have selected all four variables, your output should look like this:

Weight		*Length*		*Breadth*		*Depth*	
Mean	135.394	Mean	76.673	Mean	60.483	Mean	47.113
Standard Error	4.427358	Standard Error	1.01841	Standard Error	0.674183	Standard Error	0.514953
Median	120.3	Median	75.95	Median	59.45	Median	47
Mode	100.3	Mode	80.3	Mode	64.6	Mode	51.9
Standard Deviation	62.6123	Standard Deviation	14.4025	Standard Deviation	9.534382	Standard Deviation	7.282534
Sample Variance	3920.3	Sample Variance	207.4319	Sample Variance	90.90443	Sample Variance	53.03531
Kurtosis	2.901618	Kurtosis	0.288492	Kurtosis	-0.04986	Kurtosis	0.109112
Skewness	1.157945	Skewness	0.384037	Skewness	0.102262	Skewness	0.190064
Range	435.9	Range	85.6	Range	56.9	Range	42.7
Minimum	23.7	Minimum	39.6	Minimum	34.7	Minimum	26.9
Maximum	459.6	Maximum	125.2	Maximum	91.6	Maximum	69.6
Sum	27078.8	Sum	15334.6	Sum	12096.6	Sum	9422.6
Count	200	Count	200	Count	200	Count	200
Confidence Level(95.0%)	8.730561	Confidence Level(95.0%)	2.008262	Confidence Level(95.0%)	1.329459	Confidence Level(95.0%)	1.015465

The following will give a brief summary of some of the statistics found in the above output, explaining the ideas and meanings behind them, rather than the mathematics of each.

1 Mean

This is the most common measure of average. It is the sum of the values, divided by the number of values.

2 Median

The median is defined so that half the values are smaller and half the values are larger. If you sort the values into increasing order, the median is the value in the middle (if you have an odd number of values) or half way between the two middle values (if even)

3 Standard deviation

This is a measure of spread of the values around the mean. It is the square root of the mean squared deviation from the mean.

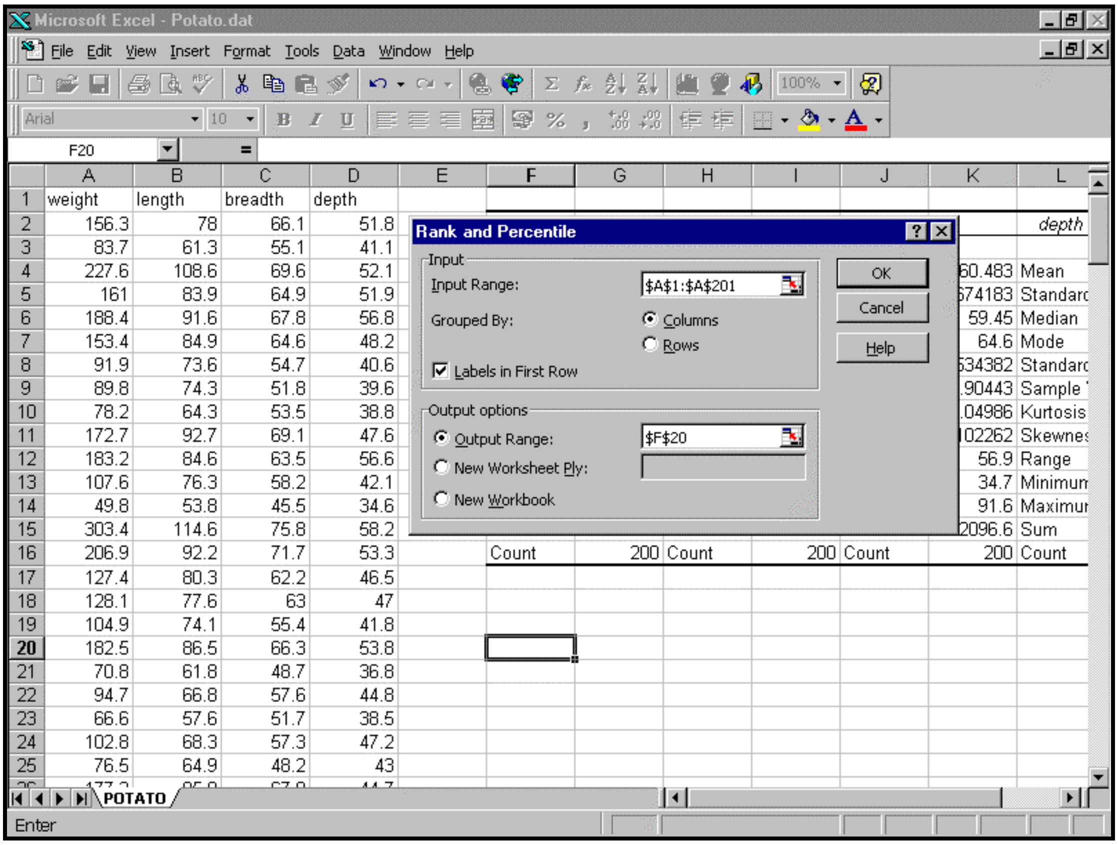
4 Standard error

This measures the sampling variability in some quantity (e.g. a mean) which is estimated from a sample.

5 Mode
The value which occurs most frequently

2.3 The RANK AND PERCENTILE tool

This tool, like the Descriptive Statistics tool can be found in the Analysis ToolPak and provides further summary information for data sets. To use it simply select **Rank and Percentile tool** from the **Tools > Data Analysis** menu and insert the input and output ranges for the data.


Microsoft Excel - Potato.dat
File Edit View Insert Format Tools Data Window Help
Arial
10
100%
F20
=
POTATO
Enter

Rank and Percentile

Input
Input Range: $A$1:$A$201
Grouped By: ● Columns ○ Rows
☑ Labels in First Row

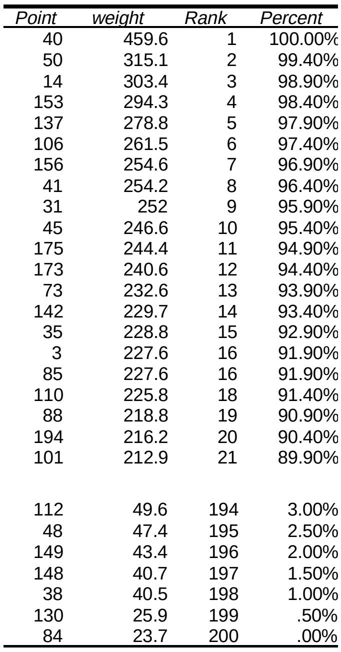
Output options
● Output Range: $F$20
○ New Worksheet Ply:
○ New Workbook

OK
Cancel
Help

	A	B	C	D
1	weight	length	breadth	depth
2	156.3	78	66.1	51.8
3	83.7	61.3	55.1	41.1
4	227.6	108.6	69.6	52.1
5	161	83.9	64.9	51.9
6	188.4	91.6	67.8	56.8
7	153.4	84.9	64.6	48.2
8	91.9	73.6	54.7	40.6
9	89.8	74.3	51.8	39.6
10	78.2	64.3	53.5	38.8
11	172.7	92.7	69.1	47.6
12	183.2	84.6	63.5	56.6
13	107.6	76.3	58.2	42.1
14	49.8	53.8	45.5	34.6
15	303.4	114.6	75.8	58.2
16	206.9	92.2	71.7	53.3
17	127.4	80.3	62.2	46.5
18	128.1	77.6	63	47
19	104.9	74.1	55.4	41.8
20	182.5	86.5	66.3	53.8
21	70.8	61.8	48.7	36.8
22	94.7	66.8	57.6	44.8
23	66.6	57.6	51.7	38.5
24	102.8	68.3	57.3	47.2
25	76.5	64.9	48.2	43

	F	G	H	I	J	K	L
2							depth
4						60.483	Mean
5						674183	Standarc
6						59.45	Median
7						64.6	Mode
8						634382	Standarc
9						.90443	Sample '
10						.04986	Kurtosis
11						102262	Skewnes
12						56.9	Range
13						34.7	Minimun
14						91.6	Maximur
15						2096.6	Sum
16	Count	200	Count	200	Count	200	Count

Figure 2.3.1



Point	weight	Rank	Percent
40	459.6	1	100.00%
50	315.1	2	99.40%
14	303.4	3	98.90%
153	294.3	4	98.40%
137	278.8	5	97.90%
106	261.5	6	97.40%
156	254.6	7	96.90%
41	254.2	8	96.40%
31	252	9	95.90%
45	246.6	10	95.40%
175	244.4	11	94.90%
173	240.6	12	94.40%
73	232.6	13	93.90%
142	229.7	14	93.40%
35	228.8	15	92.90%
3	227.6	16	91.90%
85	227.6	16	91.90%
110	225.8	18	91.40%
88	218.8	19	90.90%
194	216.2	20	90.40%
101	212.9	21	89.90%
112	49.6	194	3.00%
48	47.4	195	2.50%
149	43.4	196	2.00%
148	40.7	197	1.50%
38	40.5	198	1.00%
130	25.9	199	.50%
84	23.7	200	.00%

The table above was generated by applying the Rank and Percentile tool to the weight variable.

2.4 Graphical representation

2.4.1 An Introduction to Excel 97 graphics

Charts in Excel 97 are easy to produce, and simple to understand. The **ChartWizard** has an icon like this: 📊 It guides you through the process of creating a chart step-by-step, enabling you to verify your data selection, select a chart type, and decide whether or not to add such items as axes labels, titles and a legend. A sample of the chart is also displayed, and it is possible to go back to a particular step and make changes to this before leaving the ChartWizard and outputting the chart to a worksheet.

Excel offers many different **chart types** for you to choose from, and each one has one or more **subtypes**, or variations, which you can choose to illustrate your data somewhat differently. Subtypes and chart types can also be used in combination to create a variety of looks. Once a chart has been created many of the chart items can be moved, re-sized or reformatted.

A variety of **autoformats** are available for each chart type, and it is possible for you to add to these, creating **custom autoformats**. Excel also enables the user to combine multiple chart types in a single chart. Hence for example, a chart could be created displaying two data series, one represented by columns and the other by a single line.

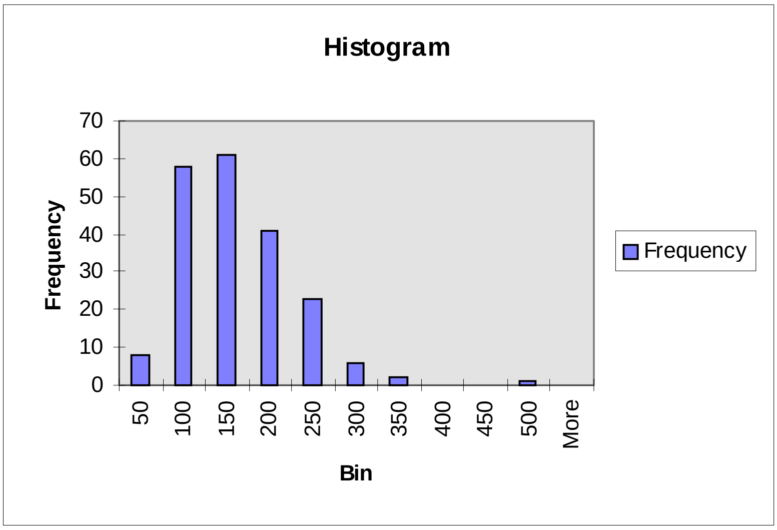
2.4.2 Simple visual aids

1) Histogram

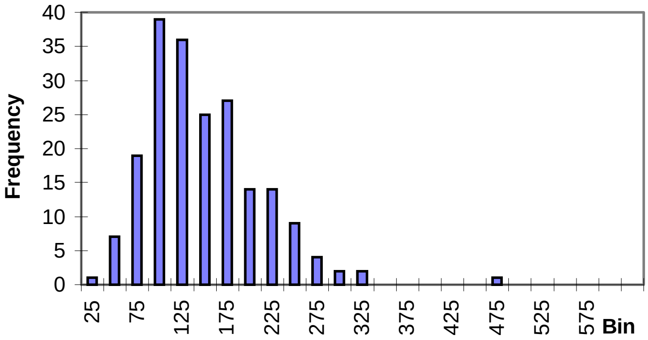
Histograms of data series can be created using the Analysis ToolPak's **Histogram tool**. The data are grouped into a series of intervals (known as bins) and the number of observations that fall into each are calculated and displayed both in a table and graphically, as a histogram.

Tools > Data Analysis... > Histogram...

On selecting the above options from the menu bar a dialog box will be displayed titled 'Histogram'. Insert the input range of the data series, either entering the reference by keyboard or highlighting the corresponding input range on the worksheet. Check the box marked **label** if one is included in the input range. Set up the ranges for your histogram divisions somewhere on the sheet. It could be a column of numbers starting at 50 and going up intervals of 50 as shown under the column headed bin below. Next enter the reference for the bin range; any range can be used. The range specified is of cells, where say, E2=50,E3=100, … ,E11=500, then the reference given is $E$2:$E$11. If this box is left empty Excel will generate a default range. Define the cell at the top, left corner of where you want the output to appear and then check the chart output box to display the histogram. Click on the OK button when you are satisfied with your selection. The histogram for the variable **weight** could look like this after you have stretched it vertically:



Below is the output generated by the Histogram tool for the **weight** data using a step-size of 25 as opposed to 50 in the previous graph.



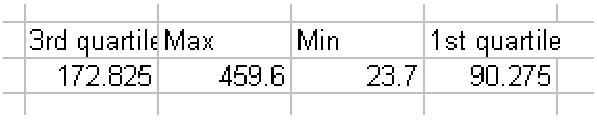
Bin	*Frequency*
25	1
50	7
75	19
100	39
125	36
150	25
175	27
200	14
225	14
250	9
275	4
300	2
325	2
350	0
375	0
400	0
425	0
450	0
475	1
500	0
525	0
550	0
575	0
600	0
More	0

From both of the histograms, you can see that the data are positively skewed, i.e. more than half the area of the histogram is to the right of the mode.

☞ **Exercise 2.1:** Explore the other data in the worksheet (breadth, length, depth) with histograms. Describe the skewness of the data.

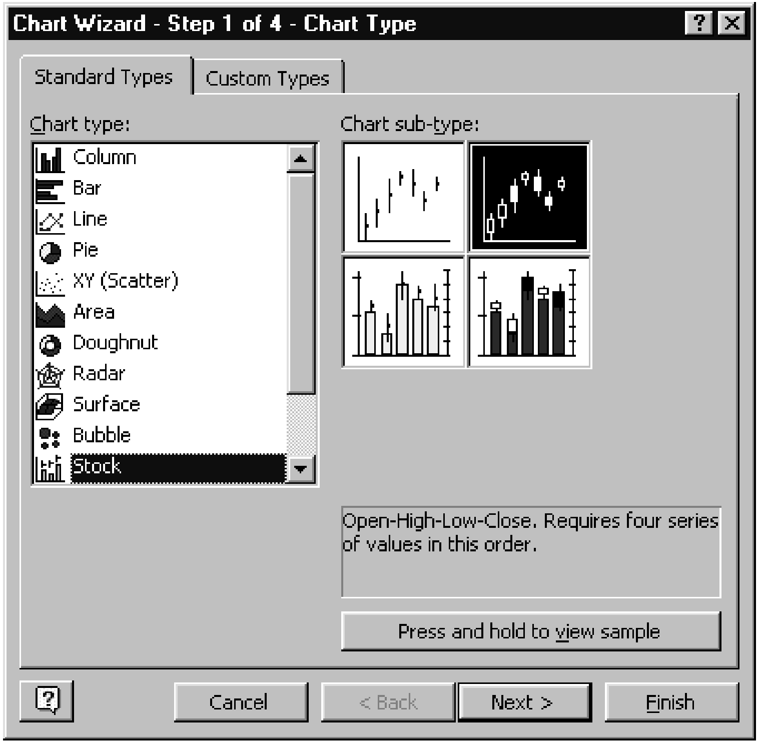
2) Boxplot

The boxplot is a visual representation of three of the summary statistics; the median, and the lower (Q1) and upper (Q3) quartile. In Excel there is no standard tool for creating a boxplot. However, by using the **Open-High-Low-Close** subtype of the **Stock Chart type** (used for stock prices) it is possible to create a chart that looks very similar. To do this you first need to create a table containing the maximum, minimum and lower and upper quartiles:



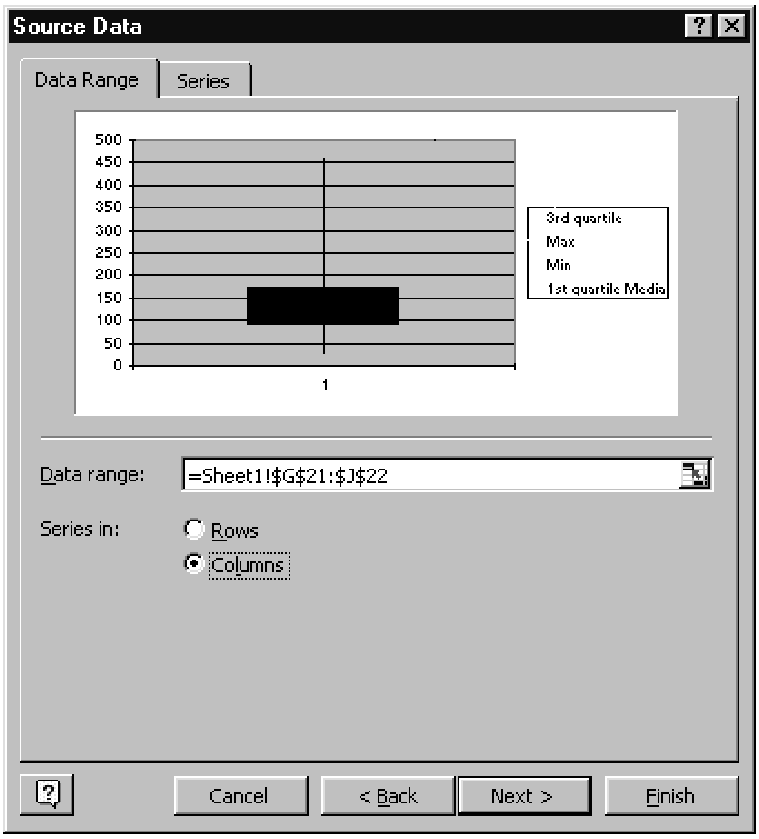
3rd quartile	Max	Min	1st quartile
172.825	459.6	23.7	90.275

Most of these statistics can be extracted from the output produced by the Descriptive statistics tool. The **Function Wizard** can also be employed to calculate these and any other statistics that are absent. Next click the Chart Wizard button on the standard toolbar. Enter the input range for the statistics table and select the **Open-High-Low-Close subtype** from the **Stock Chart** option.



Open-High-Low-Close. Requires four series of values in this order.

Ensure that the data series option is set to columns in the second dialog box. A preview of the chart will appear in the box window.



Data range: =Sheet1!$G$21:$J$22

Series in: ○ Rows ● Columns

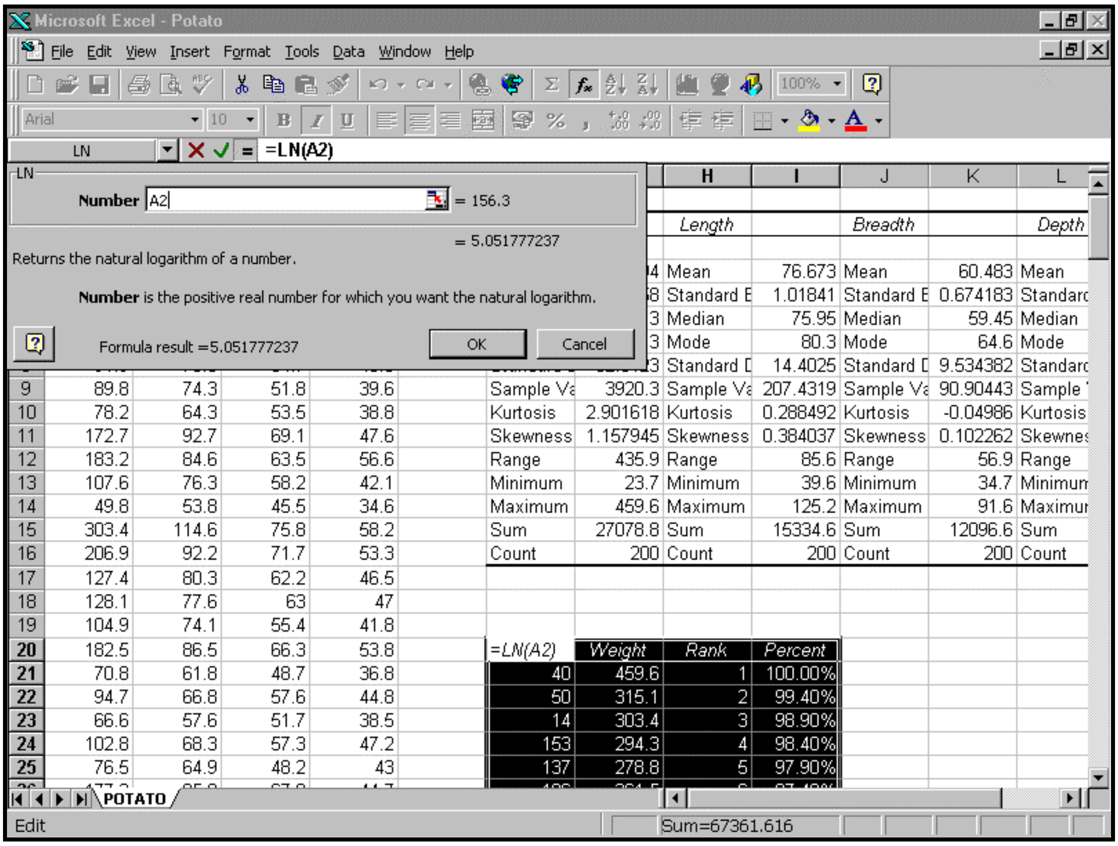
Note that there is no line representing the median point. This must be added in by hand using the drawing tools provided by Excel.

2.4.3 Transformations

Transformations change the shape of the distribution, and are useful in many analyses where the Normal is ideal. The Normal distribution has a symmetrical bell shape. There are usually good reasons to want the data to be as Normal as possible. The square root, log (to the base e, or to the base 10), and -1/y are three of the more common transformations and progressively pull in the right tail and push out the left. There are three ways of transforming datasets in Excel. The first and most straight forward is to enter the required formulae directly into the cells allocated to the transformed dataset, using the Autofill method to copy a formula from one cell to another. The second is to use the Function Wizard to apply a built-in function followed once again by the Autofill method. The third is to write a Visual Basic program, referencing the raw data from the worksheet, and outputting the transformed data to a specified range on the worksheet. In this course we shall consider only the first two options.

Transforming a single data value

Select the cell into which the transformed value is to be placed and click the **Function Wizard** button on the standard toolbar. Select Math and trig functions and then find LN in the group on the right hand side of the display box. Enter the input range and parameter values into the box provided. Remember, the input range can be entered automatically by simply dragging the cursor over the relevant cells. Figure 2.2.1 shows the dialogue box for the log function LN



LN =LN(A2)

LN

Number A2 = 156.3

= 5.051777237

Returns the natural logarithm of a number.

Number is the positive real number for which you want the natural logarithm.

Formula result =5.051777237

								H	I	J	K	L
								Length		*Breadth*		*Depth*
							4	Mean	76.673	Mean	60.483	Mean
							8	Standard E	1.01841	Standard E	0.674183	Standard
							3	Median	75.95	Median	59.45	Median
							3	Mode	80.3	Mode	64.6	Mode
							3	Standard D	14.4025	Standard D	9.534382	Standard
9	89.8	74.3	51.8	39.6		Sample Va	3920.3	Sample Va	207.4319	Sample Va	90.90443	Sample
10	78.2	64.3	53.5	38.8		Kurtosis	2.901618	Kurtosis	0.288492	Kurtosis	-0.04986	Kurtosis
11	172.7	92.7	69.1	47.6		Skewness	1.157945	Skewness	0.384037	Skewness	0.102262	Skewnes
12	183.2	84.6	63.5	56.6		Range	435.9	Range	85.6	Range	56.9	Range
13	107.6	76.3	58.2	42.1		Minimum	23.7	Minimum	39.6	Minimum	34.7	Minimum
14	49.8	53.8	45.5	34.6		Maximum	459.6	Maximum	125.2	Maximum	91.6	Maximur
15	303.4	114.6	75.8	58.2		Sum	27078.8	Sum	15334.6	Sum	12096.6	Sum
16	206.9	92.2	71.7	53.3		Count	200	Count	200	Count	200	Count
17	127.4	80.3	62.2	46.5								
18	128.1	77.6	63	47								
19	104.9	74.1	55.4	41.8								
20	182.5	86.5	66.3	53.8		*=LN(A2)*	*Weight*	*Rank*	*Percent*			
21	70.8	61.8	48.7	36.8		40	459.6	1	100.00%			
22	94.7	66.8	57.6	44.8		50	315.1	2	99.40%			
23	66.6	57.6	51.7	38.5		14	303.4	3	98.90%			
24	102.8	68.3	57.3	47.2		153	294.3	4	98.40%			
25	76.5	64.9	48.2	43		137	278.8	5	97.90%			

Edit Sum=67361.616

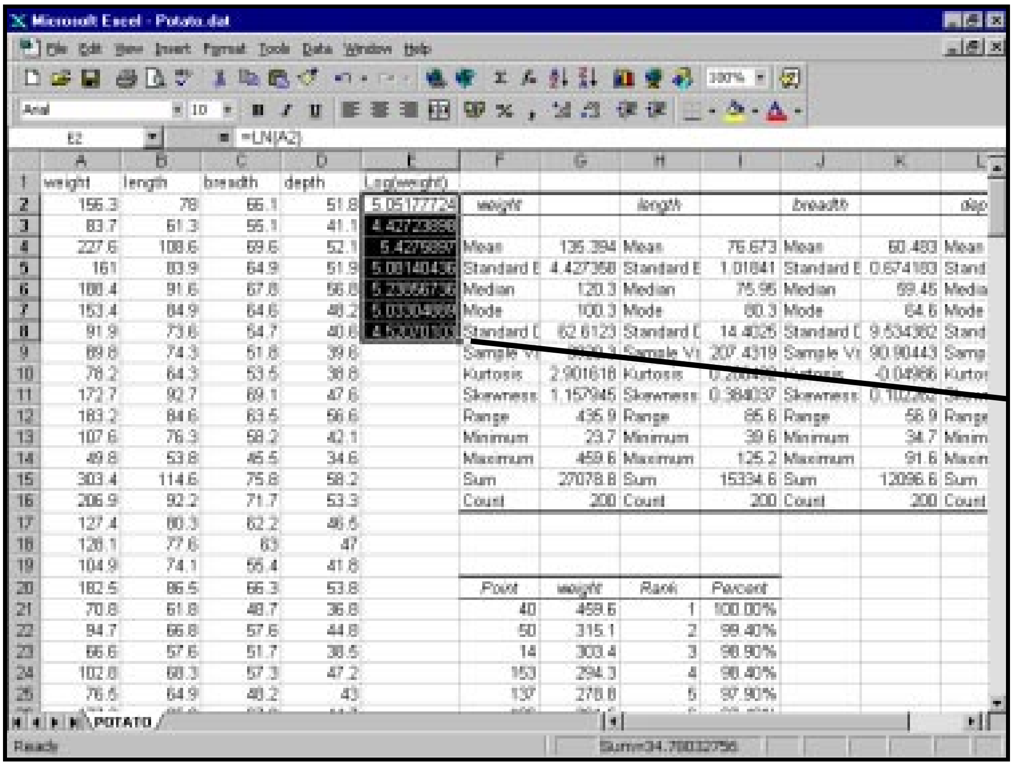
Figure 2.4.1

Both the log and the square root functions can be applied in this way, but the formula -1/y will need to be entered manually. This brings us to the second option for transforming datasets, which we shall illustrate using the -1/y transformation. To do this simply select the output cell, type ‘ = ‘ and enter the **formula**, eg **= -1/A20**, where A20 accesses the data cell containing the value for y. Click on the enter key or green tick on the formula bar. The **formula bar**, situated at the top of the worksheet, shows the changes you make as you build your formula.

Copying formulae and references

When cells are copied, Excel automatically adjusts relative references and the relative parts of mixed references in the area where the copied cells are pasted. Relative references are of the form C$4 and $E5. C$4 means keep the row constant and use the column whereas $E5 means keep the column constant and use the row. Note you can also use the worksheet as part of the formula so for instance POTATO!B16 references cell B16 in worksheet POTATO

To copy a formula select the cell and use the fill handle to drag the formula into the adjacent cells. See Figure 2.4.2 Alternatively copy the cell and select the cells in which you want the formula pasted. Click on **Edit > Paste Special**, select **Paste Formula** from the options and click on the **OK** button.


Microsoft Excel - Potato.dat
File Edit View Insert Format Tools Data Window Help
Arial
10
100%
POTATO
Ready

E2 =LN(A2)

	A	B	C	D	E	F	G	H	I	J	K	L
1	weight	length	breadth	depth	Log(weight)							
2	156.3	78	66.1	51.8	5.05177724	weight		length		breadth		dep
3	83.7	61.3	55.1	41.1	4.42723898							
4	227.6	108.6	69.6	52.1	5.4275897	Mean	135.394	Mean	76.673	Mean	60.483	Mean
5	161	83.9	64.9	51.9	5.08140436	Standard E	4.427358	Standard E	1.01841	Standard E	0.674183	Stand
6	188.4	91.6	67.8	56.8	5.23856736	Median	120.3	Median	75.95	Median	59.45	Media
7	153.4	84.9	64.6	48.2	5.03304869	Mode	100.3	Mode	80.3	Mode	64.6	Mode
8	91.9	73.6	54.7	40.6	[illegible]	Standard D	62.6123	Standard D	14.4025	Standard D	9.534382	Stand
9	89.8	74.3	51.8	39.6		Sample Vi	[illegible]	Sample Vi	207.4319	Sample Vi	90.90443	Samp
10	78.2	64.3	53.5	38.8		Kurtosis	2.901618	Kurtosis	[illegible]	Kurtosis	-0.04986	Kurtos
11	172.7	92.7	69.1	47.6		Skewness	1.157945	Skewness	0.384037	Skewness	[illegible]	[illegible]
12	183.2	84.6	63.5	56.6		Range	435.9	Range	85.6	Range	56.9	Range
13	107.6	76.3	58.2	42.1		Minimum	23.7	Minimum	39.6	Minimum	34.7	Minim
14	49.8	53.8	45.5	34.6		Maximum	459.6	Maximum	125.2	Maximum	91.6	Maxim
15	303.4	114.6	75.8	58.2		Sum	27078.8	Sum	15334.6	Sum	12096.6	Sum
16	206.9	92.2	71.7	53.3		Count	200	Count	200	Count	200	Count
17	127.4	80.3	62.2	46.5								
18	128.1	77.6	63	47								
19	104.9	74.1	55.4	41.8								
20	182.5	86.5	66.3	53.8		Point	weight	Rank	Percent			
21	70.8	61.8	48.7	36.8		40	459.6	1	100.00%			
22	94.7	66.8	57.6	44.8		50	315.1	2	99.40%			
23	66.6	57.6	51.7	38.5		14	303.4	3	98.90%			
24	102.8	68.3	57.3	47.2		153	294.3	4	98.40%			
25	76.5	64.9	40.2	43		137	278.8	5	97.90%			

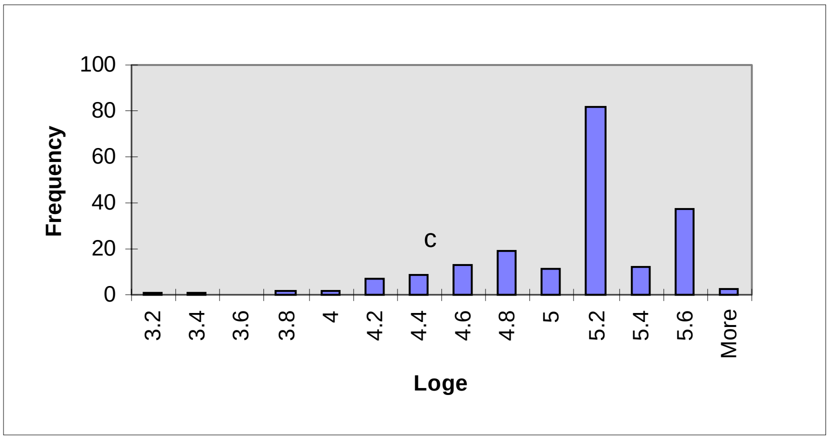
Sum=34.78032756

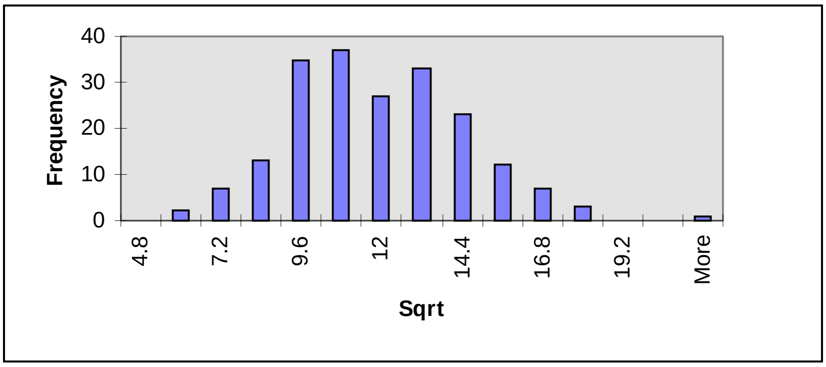
Fill handle

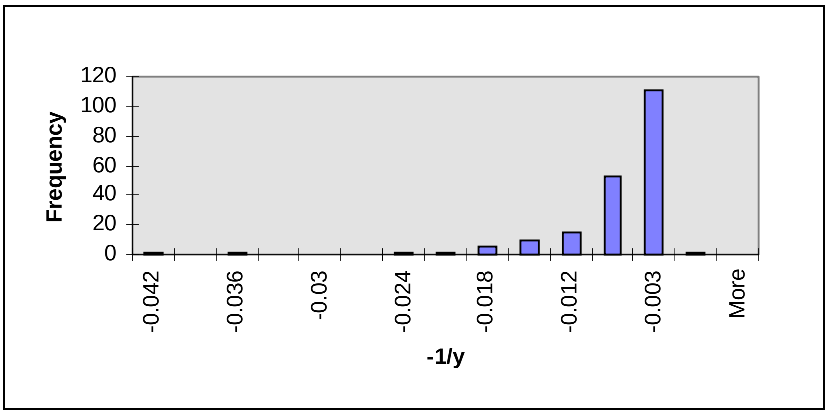
Figure 2.4.2

To chart the transformed data click on : **Tools > Data Analysis > Histogram**

☞ **Exercise 2.2:** Produce the following histograms taking the loge, sqrt and -1/y of the weight variable**.**







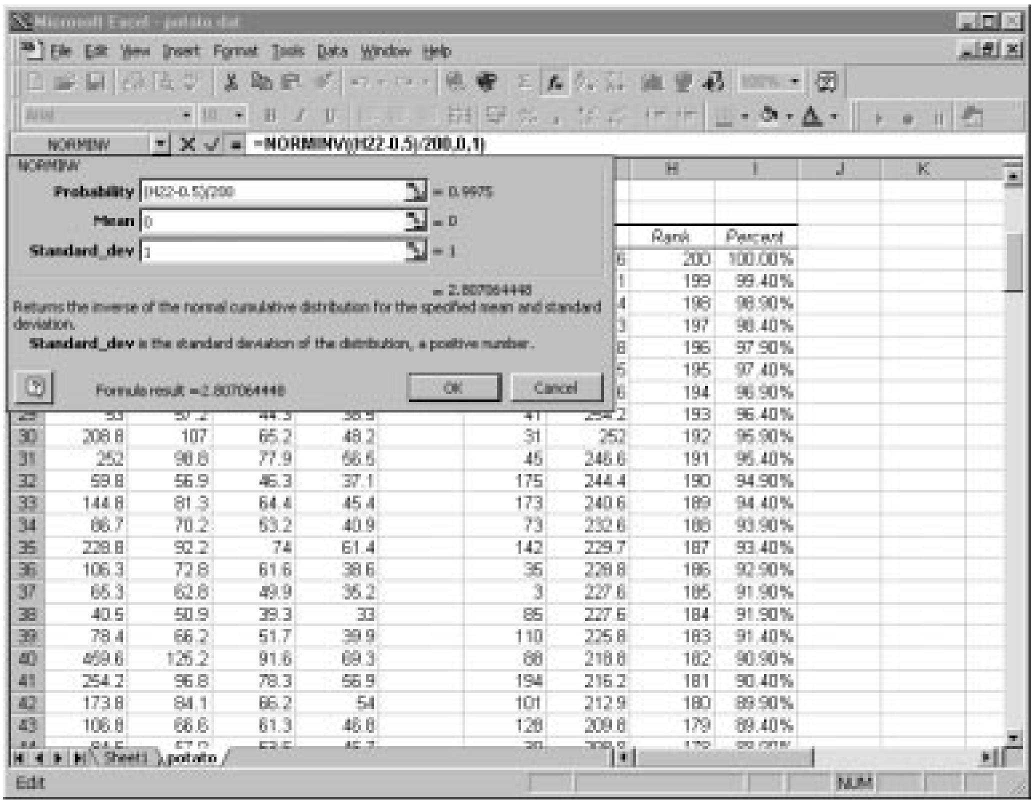
2.4.4 Normal scores

Nscores is an abbreviation for normal scores. If data are plotted against their normal scores, and if the result is a reasonably straight line, we accept that the distribution is Normal.

To create the nscores first rank the weight data from 1-200 using either the **Rank and Percentile tool** or the **Function Wizard**'s **Rank** statistical function. In another column store the nscores. These are derived using the formula: $nscore(X_{(j)}) = NORMINV(\frac{j-0.5}{200},0,1)$

where j is the rank of data point Xj , and NORMINV is a built-in function which returns the inverse of the normal cumulative distribution.

Note in Figure 2.4.3 the probability is calculated as (H22-0.5)/200 since the rank values are stored in column H starting at row 22



Microsoft Excel - potato.dat
File Edit View Insert Format Tools Data Window Help
NORMINV
OK
Cancel
Sheet1
potato
Edit
NUM

=NORMINV((H22-0.5)/200,0,1)

NORMINV

Probability (H22-0.5)/200 = 0.9975

Mean 0 = 0

Standard_dev 1 = 1

= 2.807064448

Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation.

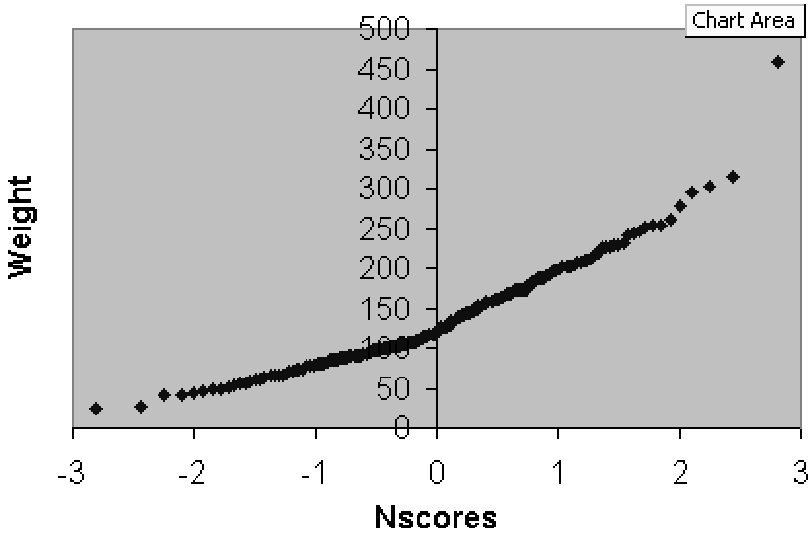
Standard_dev is the standard deviation of the distribution, a positive number.

Formula result =2.807064448

								H	I
								Rank	Percent
							6	200	100.00%
							1	199	99.40%
							4	198	98.90%
							3	197	98.40%
							8	196	97.90%
							5	195	97.40%
							6	194	96.90%
29	53	57.2	44.3	38.5		41	254.2	193	96.40%
30	208.8	107	65.2	48.2		31	252	192	95.90%
31	252	98.8	77.9	66.5		45	246.6	191	95.40%
32	59.8	56.9	46.3	37.1		175	244.4	190	94.90%
33	144.8	81.3	64.4	45.4		173	240.6	189	94.40%
34	86.7	70.2	63.2	40.9		73	232.6	188	93.90%
35	228.8	92.2	74	61.4		142	229.7	187	93.40%
36	106.3	72.8	61.6	38.6		35	228.8	186	92.90%
37	65.3	62.8	49.9	35.2		3	227.6	185	91.90%
38	40.5	50.9	39.3	33		85	227.6	184	91.90%
39	78.4	66.2	51.7	39.9		110	225.8	183	91.40%
40	459.6	126.2	91.6	69.3		88	218.8	182	90.90%
41	254.2	96.8	78.3	56.9		194	216.2	181	90.40%
42	173.8	84.1	66.2	54		101	212.9	180	89.90%
43	106.8	66.6	61.3	46.8		128	209.8	179	89.40%

Figure 2.4.3

☞ **Exercise 2.3:** Calculate the nscores for the weight data and plot these against the ranked weight value which was calculated by the percentile. Your output should look something like this:



Can you locate the outlier on the graph?

Since this is not straight, we can conclude that the weights are not Normally distributed.
Which of the transformations do you think gives the closest approximation to a Normal distribution?

3. Exploratory data analysis for several variables

When you have several variables recorded for each unit, you might want to examine the relationships between the variables as well as studying each variable individually. To illustrate the techniques described in this chapter import the ascii file **weather.dat** from the floppy disk and name the columns **day, hour, cloud, rad, temp, humidity, rain, windsp, winddir**. These variables were recorded hourly for seven days in May 1982 in Edinburgh. You will notice that some of the values in the winddir column are denoted by *; this is because there are missing values. Many of the analyses in the Data Analysis Toolpak have not been designed to handle missing values and therefore changes may need to be made to a dataset before a tool is applied. The inability of Excel to identify and respond to missing values is one of the drawbacks of the package for statistical analysis.

3.1 Simple graphs

3.1.1 Plotting against time order

It is a good idea to examine each variable separately before any sort of analysis between variables and examining the ordered data is a good starting point. To create a plot of **cloud** against order you simply need to plot **cloud** in an x-y plot and Excel will assume it is ordered. So highlight the **cloud** column, click on the chart wizard and choose **xy scatter**.

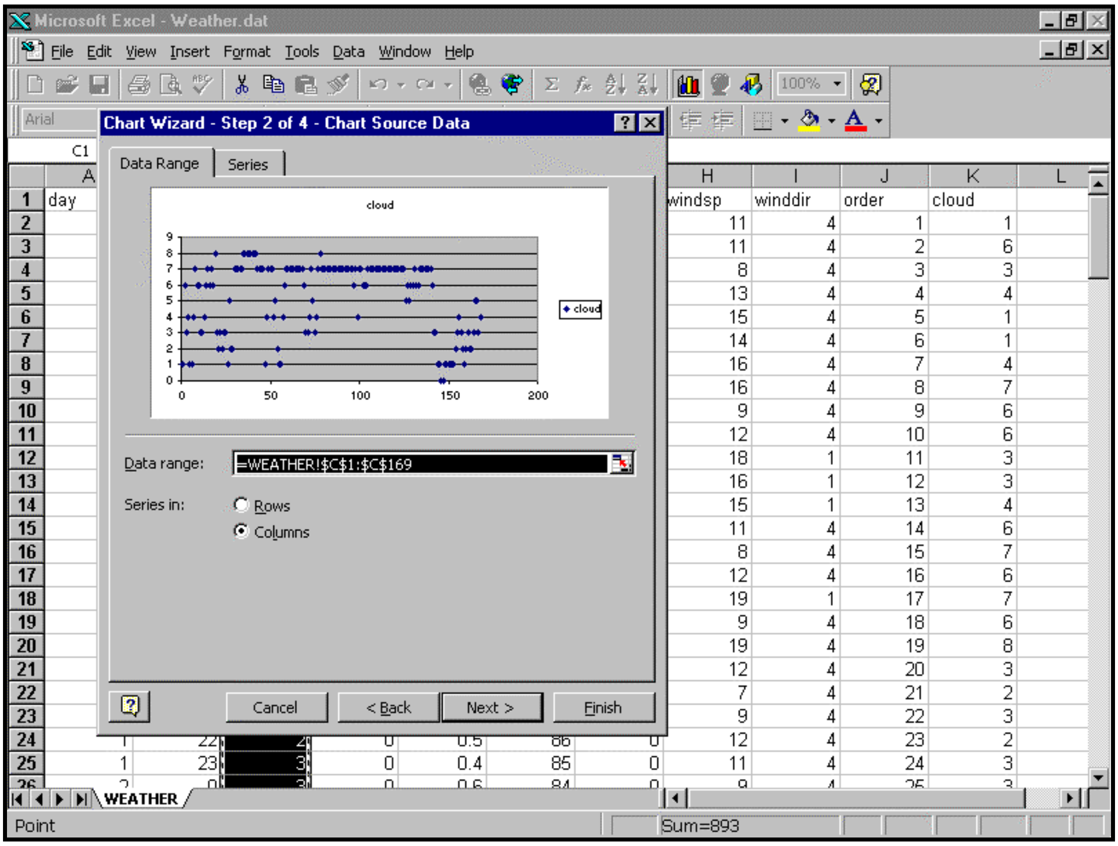

Microsoft Excel - Weather.dat
File Edit View Insert Format Tools Data Window Help
100%
Arial
C1
cloud
0
1
2
3
4
5
6
7
8
9
0
50
100
150
200
cloud
WEATHER
Point

Chart Wizard - Step 2 of 4 - Chart Source Data

Data Range | Series

Data range: =WEATHER!$C$1:$C$169

Series in: ○ Rows ● Columns

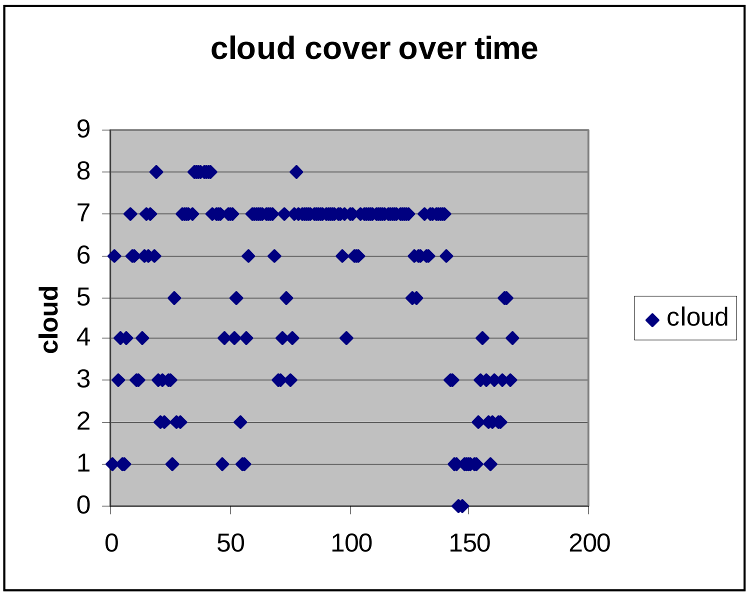
Cancel | < Back | Next > | Finish

	A							H	I	J	K	L
1	day							windsp	winddir	order	cloud	
2								11	4	1	1	
3								11	4	2	6	
4								8	4	3	3	
5								13	4	4	4	
6								15	4	5	1	
7								14	4	6	1	
8								16	4	7	4	
9								16	4	8	7	
10								9	4	9	6	
11								12	4	10	6	
12								18	1	11	3	
13								16	1	12	3	
14								15	1	13	4	
15								11	4	14	6	
16								8	4	15	7	
17								12	4	16	6	
18								19	1	17	7	
19								9	4	18	6	
20								19	4	19	8	
21								12	4	20	3	
22								7	4	21	2	
23								9	4	22	3	
24		1	22	2	0	0.5	86	0	12	4	23	2
25		1	23	3	0	0.4	85	0	11	4	24	3

Sum=893

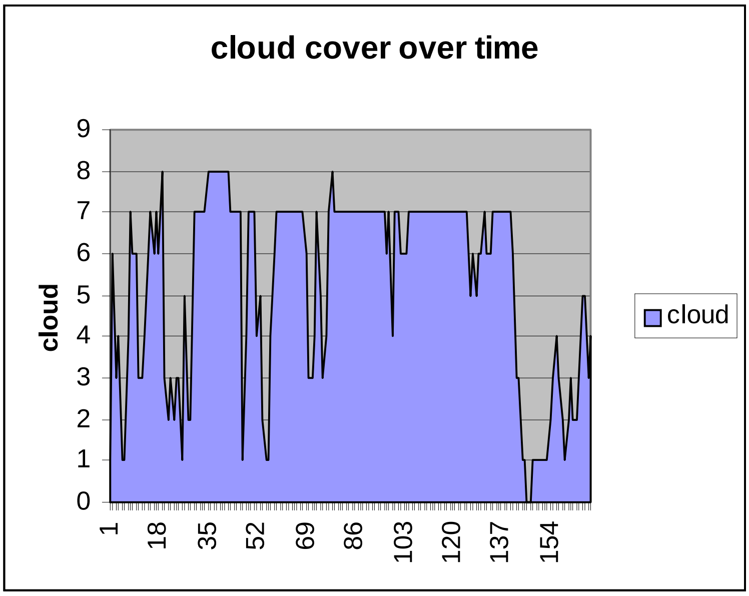
Figure 3.1.1

Step two of the chart wizard is illustrated in Figure 3.1.1 This will result in the following plot.



You can see from this plot that there are 8 discrete values for cloud. In fact it is measured in eighths of cover. The value 1 say indicating that coloud cover is one eight of the total.

It is difficult to see what is going on from day to day so use a different type of subplot format which will join up each point and shade in the area below the joined points. See next diagram for this. It is obtained by choosing **Area** in place of **xy scatter**.



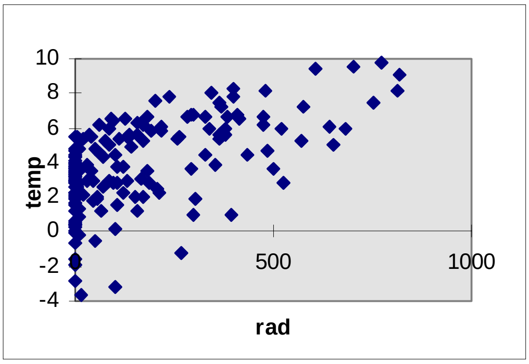
Some graphs are not improved by this type of treatment including those for categorical and discrete data as demonstrated for **cloud** above.

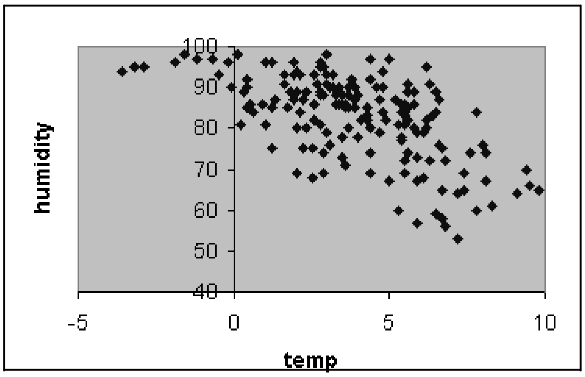
Exercise 3.1: Look at the **rain** data and explore whether or not the data should be joined by a line.

Exercise 3.2: Now look at **Humidity, Rad and Temp**

3.1.2 Plotting two variables.

Look at the following plots





These show simple relationships between variables and are only useful if at least one of the variables is continuous. The result is more useful if both variables are continuous as shown above. The second plot has been improved by formatting the y-axis so that the lower limit is 40.

The correlation between pairs of variables can be calculated by :

Tools > Data Analysis…> Correlation.

Selecting **temp**, **rad** and **humidity** as the input ranges creates the following correlation table.

	rad	***temp***	***humidity***
rad	1		
temp	0.590231	1	
humidity	-0.6189	-0.61039	1

Notice that the correlation coefficient is positive for the first pair, and negative for the second.

Correlation measures the strength of the relationship between two variables, and always lies between -1 and +1. It can be interpreted as follows:

Close to -1 ➔ strong negative relationship (the plot should have a reasonably straight line heading from top left to bottom right)

Between -1 and 0 ➔ negative relationship

About 0 ➔ little relationship

Between 0 and 1 ➔ positive relationship

Close to 1 ➔ strong positive relationship (the plot should have a reasonably straight line heading from bottom left to top right)

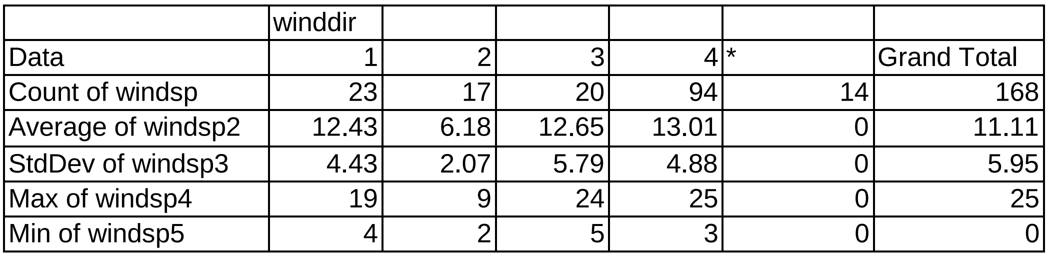
The order in which you specify the variables is not important. Care must be taken with the correlation coefficient, since it only measures the strength of straight line relationships, and can give misleading answers with more complicated relationships. Also, correlation does not imply cousation.

3.2 Tabulating data

Tabulation is one way of examining two categorical variables by producing a table showing the number of units in each combination of categories.

3.2.1 Pivot tables.

A pivot table is an interactive worksheet table which provides an easy way for you to display and analyse summary information about data stored in an Excel worksheet. The pivot table is simple to create and easy to modify. Here is an example of a pivot table:



	winddir					
Data	1	2	3	4	*	Grand Total
Count of windsp	23	17	20	94	14	168
Average of windsp2	12.43	6.18	12.65	13.01	0	11.11
StdDev of windsp3	4.43	2.07	5.79	4.88	0	5.95
Max of windsp4	19	9	24	25	0	25
Min of windsp5	4	2	5	3	0	0

This pivot table is a summary of **windsp** categorised by **winddir**. The data on **windsp** is presented as split into 4 categories plus the missing value category denoted by *

3.2.2 Creating a simple pivot table.

Before commencing this section make sure your work sheet values are not overwritten by plots or other information. You may prefer to read the data in again.

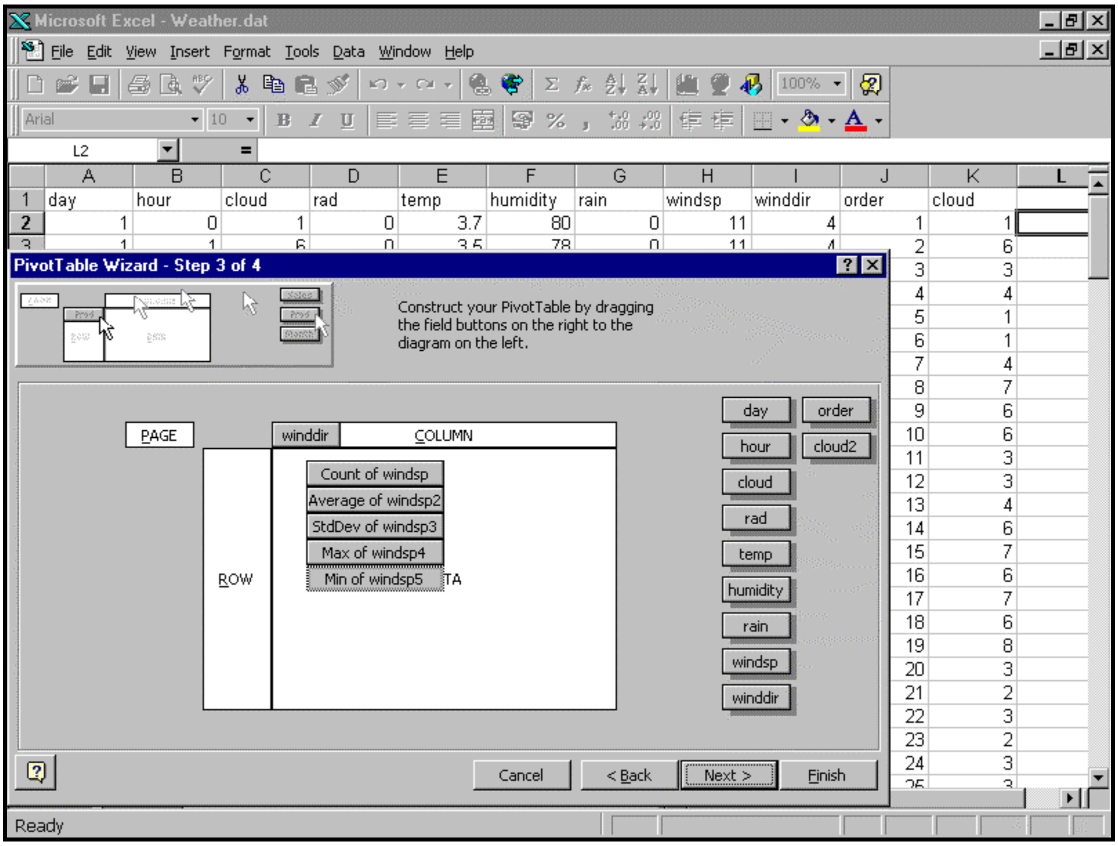
1) Create the pivot table by using:

Data> Pivot table report…

2) To specify the layout, follow the wizard and use Figure 3.2.1 to indicate what you enter into the boxes. You should note that if you highlight variables using the mouse before you start the pivot table report then only those variables will appear in the pivot table selection. The list of variables in Figure 3.2.1 is the result of no selection before starting the table.

3) To create the table, drag the boxes containing the variable names in the columns on the right hand side of the Wizard screen to the area marked column, row or data. In the example in Figure 3.2.1 a box containing **winddir** was moved into the area marked COLUMN. **windsp** was then dragged five times to the area in the centre of the table marked DATA. The boxes in the DATA area were then double clicked on successively and the appropriate statistic chosen (Count, Average, StdDev, Max and Min).

Complete the table by clicking Next. Enter the location for the output of the table and finally click Finish. The final stage is shown in Figure 3.2.2. The result of this pivot table is shown in section 3.2.1.



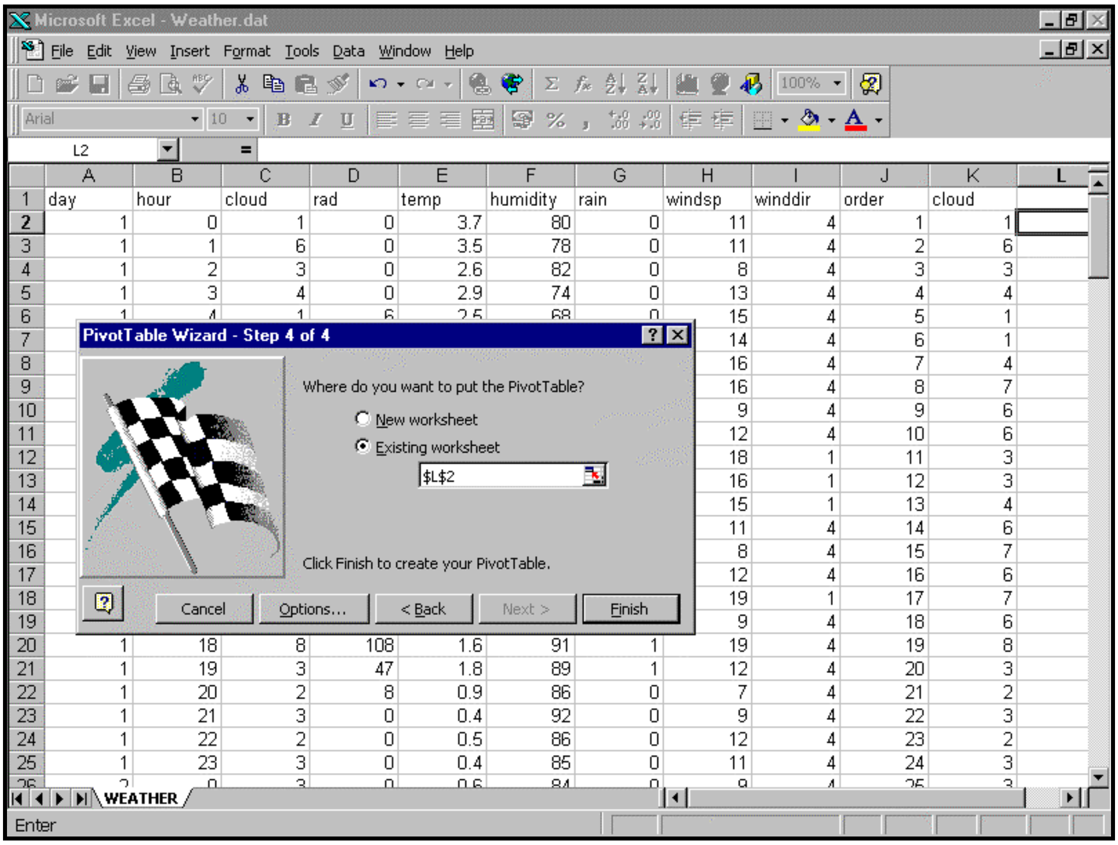
Construct your PivotTable by dragging the field buttons on the right to the diagram on the left.

Column: winddir

Data: Count of windsp, Average of windsp2, StdDev of windsp3, Max of windsp4, Min of windsp5

	A	B	C	D	E	F	G	H	I	J	K
1	day	hour	cloud	rad	temp	humidity	rain	windsp	winddir	order	cloud
2	1	0	1	0	3.7	80	0	11	4	1	1
										2	6
										3	3
										4	4
										5	1
										6	1
										7	4
										8	7
										9	6
										10	6
										11	3
										12	3
										13	4
										14	6
										15	7
										16	6
										17	7
										18	6
										19	8
										20	3
										21	2
										22	3
										23	2
										24	3

Figure 3.2.1


Microsoft Excel - Weather.dat
File Edit View Insert Format Tools Data Window Help
Arial
10
100%
L2
=
PivotTable Wizard - Step 4 of 4
Cancel
Options...
< Back
Next >
Finish
WEATHER
Enter

Where do you want to put the PivotTable?

○ New worksheet

● Existing worksheet

$L$2

Click Finish to create your PivotTable.

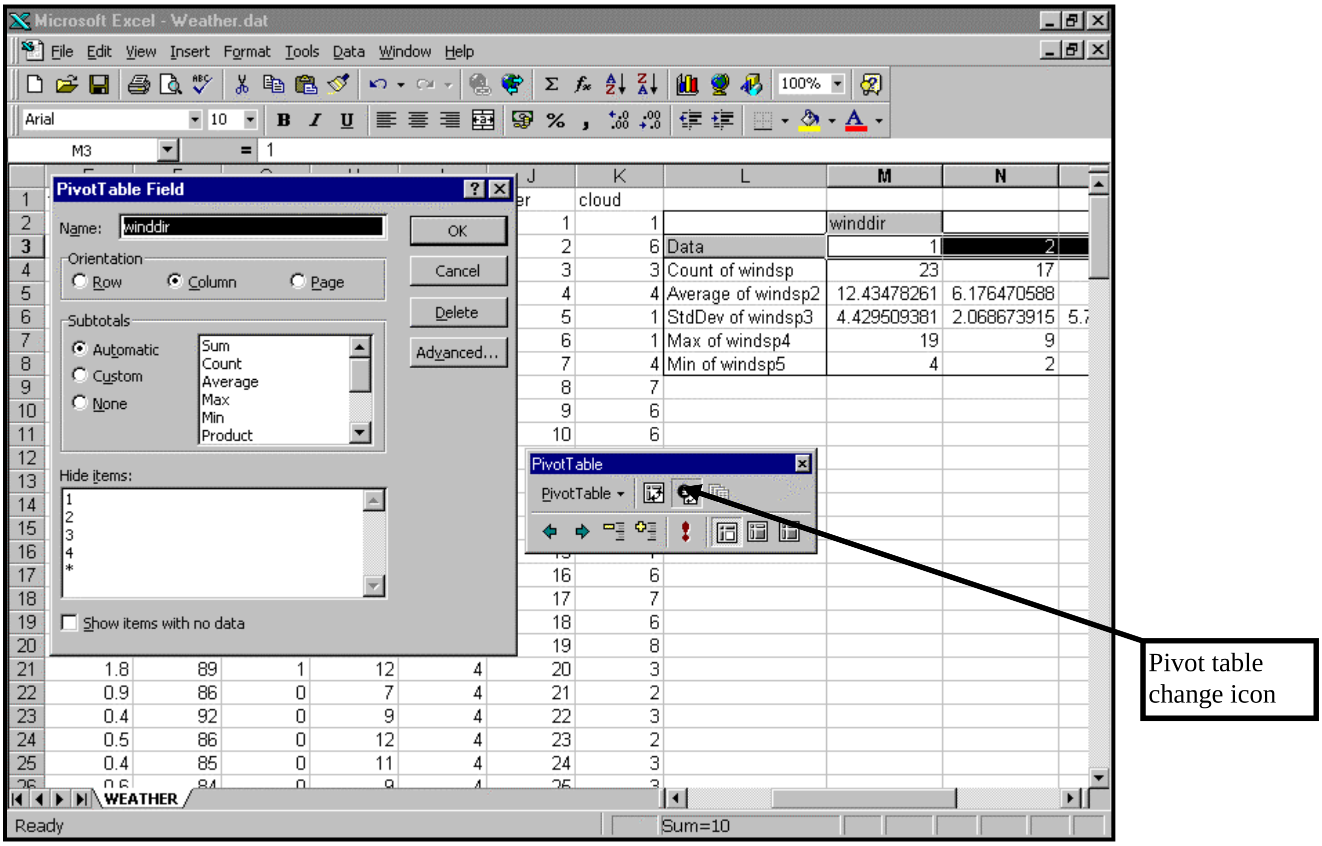
	A	B	C	D	E	F	G	H	I	J	K
1	day	hour	cloud	rad	temp	humidity	rain	windsp	winddir	order	cloud
2	1	0	1	0	3.7	80	0	11	4	1	1
3	1	1	6	0	3.5	78	0	11	4	2	6
4	1	2	3	0	2.6	82	0	8	4	3	3
5	1	3	4	0	2.9	74	0	13	4	4	4
6	1	4	1	6	2.5	68	0	15	4	5	1
7								14	4	6	1
8								16	4	7	4
9								16	4	8	7
10								9	4	9	6
11								12	4	10	6
12								18	1	11	3
13								16	1	12	3
14								15	1	13	4
15								11	4	14	6
16								8	4	15	7
17								12	4	16	6
18								19	1	17	7
19								9	4	18	6
20	1	18	8	108	1.6	91	1	19	4	19	8
21	1	19	3	47	1.8	89	1	12	4	20	3
22	1	20	2	8	0.9	86	0	7	4	21	2
23	1	21	3	0	0.4	92	0	9	4	22	3
24	1	22	2	0	0.5	86	0	12	4	23	2
25	1	23	3	0	0.4	85	0	11	4	24	3

Figure 3.2.2

3.2.3 Making changes to a pivot table.

To make changes to a pivot table once it has been created, select any data cell in the table and click on the mouse shortcut menu using the right hand mouse button. The short cut menu will appear. Choose the one option you want. You can for instance add other variables to the table or delete them. You can change say, stdev of **windsp** to variance of **windsp** using **Field...** option off the right hand mouse button.

Alternatively a range of options are available by highlighting the pivot table you want to change. Now by looking at Figure 3.2.3 click on the Pivot Table Icon change button and you will get the result in Figure 3.2.3. You should look for this icon not only on the Excel worksheet but it is part of a movable menu which could be in amongst the buttons at the top of the screen. The dialogue box in Figure 3.2.3 was activated by selecting the **winddir** button on the worksheet pivot table and clicking the **Pivot Table** > **Field…** option of the Pivot table Menu bar. As you can see it is also possible to hide items (categories) from a field, such as the missing values from the **winddir** field. To do this simply highlight the item and click OK. So if you want to hide the missing values, click on the * in the bottom box then click OK.



Microsoft Excel - Weather.dat
File Edit View Insert Format Tools Data Window Help
Arial
10
100%
M3
PivotTable
PivotTable ▾
WEATHER
Ready
Sum=10

M3 = 1

PivotTable Field

Name: winddir

Orientation: ○ Row ● Column ○ Page

Subtotals: ● Automatic ○ Custom ○ None

Sum
Count
Average
Max
Min
Product

Hide items: 1, 2, 3, 4, *

□ Show items with no data

OK | Cancel | Delete | Advanced...

J	K	L	M	N
er	cloud			
1	1		winddir	
2	6	Data	1	2
3	3	Count of windsp	23	17
4	4	Average of windsp2	12.43478261	6.176470588
5	1	StdDev of windsp3	4.429509381	2.068673915
6	1	Max of windsp4	19	9
7	4	Min of windsp5	4	2
8	7			
9	6			
10	6			

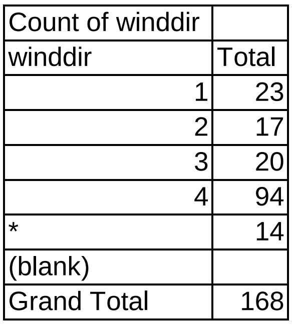
Row	E	F	G	H	I	J	K
17						16	6
18						17	7
19						18	6
20						19	8
21	1.8	89	1	12	4	20	3
22	0.9	86	0	7	4	21	2
23	0.4	92	0	9	4	22	3
24	0.5	86	0	12	4	23	2
25	0.4	85	0	11	4	24	3

Pivot table change icon

Figure 3.2.3

☞ **Exercise 3.3:** Now try to produce these pivot tables:

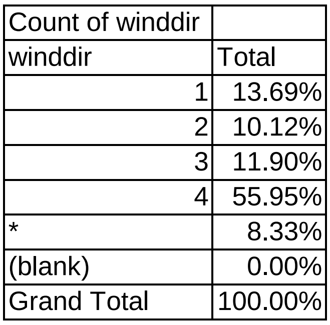
Count of rain	
rain	Total
0	137
1	31
Grand Total	168



Count of winddir	
winddir	Total
1	23
2	17
3	20
4	94
*	14
(blank)	
Grand Total	168

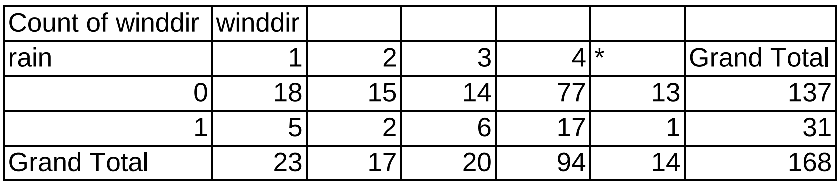
Note if you obtain a horizontal table then get Pivot Table Field (see previous tip) and click Row in the orientation panel.

You can choose various different options for output. For example, try **% of column** from the options at step 3 of 4 of the Pivot Table Wizard to produce something like this:

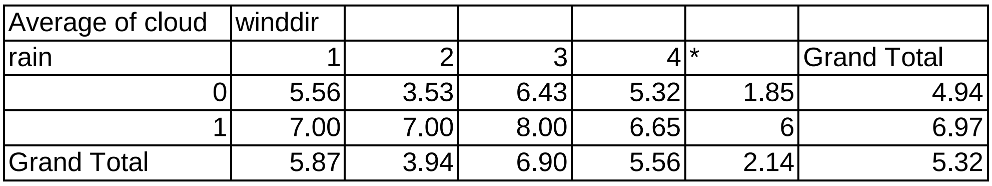


Count of winddir	
winddir	Total
1	13.69%
2	10.12%
3	11.90%
4	55.95%
*	8.33%
(blank)	0.00%
Grand Total	100.00%

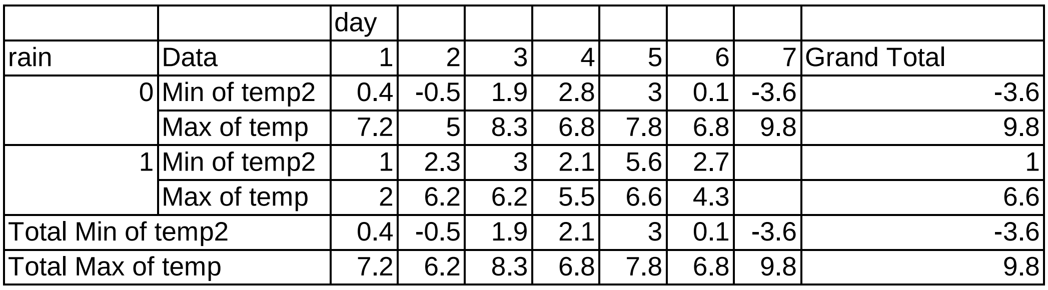
☞ **Exercise 3.4:** Now produce the following three tables using the **pivot table wizard**.



Count of winddir	winddir					
rain	1	2	3	4	*	Grand Total
0	18	15	14	77	13	137
1	5	2	6	17	1	31
Grand Total	23	17	20	94	14	168



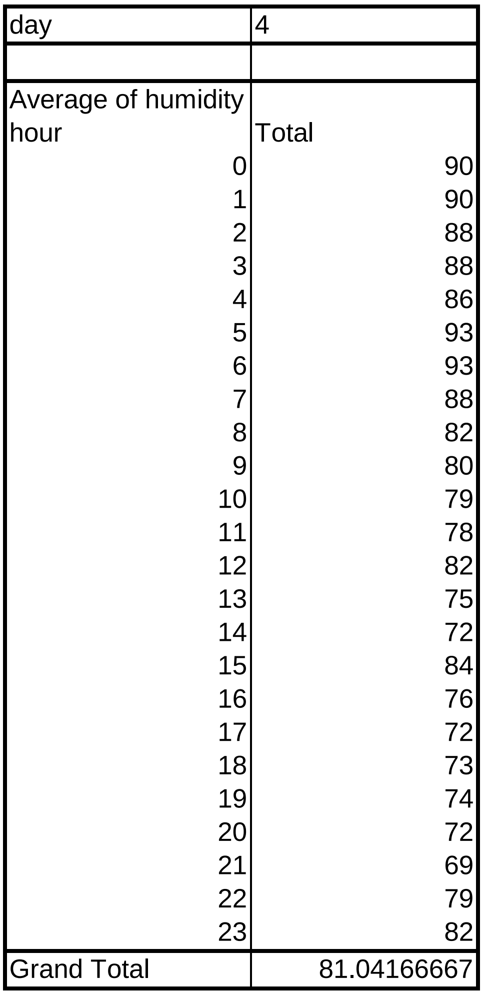
Average of cloud	winddir					
rain	1	2	3	4	*	Grand Total
0	5.56	3.53	6.43	5.32	1.85	4.94
1	7.00	7.00	8.00	6.65	6	6.97
Grand Total	5.87	3.94	6.90	5.56	2.14	5.32



		day							
rain	Data	1	2	3	4	5	6	7	Grand Total
0	Min of temp2	0.4	-0.5	1.9	2.8	3	0.1	-3.6	-3.6
	Max of temp	7.2	5	8.3	6.8	7.8	6.8	9.8	9.8
1	Min of temp2	1	2.3	3	2.1	5.6	2.7		1
	Max of temp	2	6.2	6.2	5.5	6.6	4.3		6.6
Total Min of temp2		0.4	-0.5	1.9	2.1	3	0.1	-3.6	-3.6
Total Max of temp		7.2	6.2	8.3	6.8	7.8	6.8	9.8	9.8

3.2.4 Filtering the data in a pivot table

To filter the data in a pivot table use the page field. The page field is shown in step 3 of the pivot table wizard. See Figure 3.2.1. This breaks the data up into separate pages so that summary statistics can be viewed for one particular value of a field at a time. This option is particularly valuable when charting data whose factors possess many levels; any chart plotted from tabulated data will change when a new page field item is selected. Below is a pivot table of the average **humidity** categorised by **day** and **hour**. **Day** has been inserted as the page field area of the pivot table and **hour** as a row field. When first produced this table shows All values for day. The top line has the word "day" in one box and the word "All" in the box on the right. Beside All is an downward pointing arrow. If you click on this arrow with the mouse it will allow you to choose a particular value of day. The following table is day 4.



day	4
Average of humidity	
hour	Total
0	90
1	90
2	88
3	88
4	86
5	93
6	93
7	88
8	82
9	80
10	79
11	78
12	82
13	75
14	72
15	84
16	76
17	72
18	73
19	74
20	72
21	69
22	79
23	82
Grand Total	81.04166667

3.3 Graphs and categories

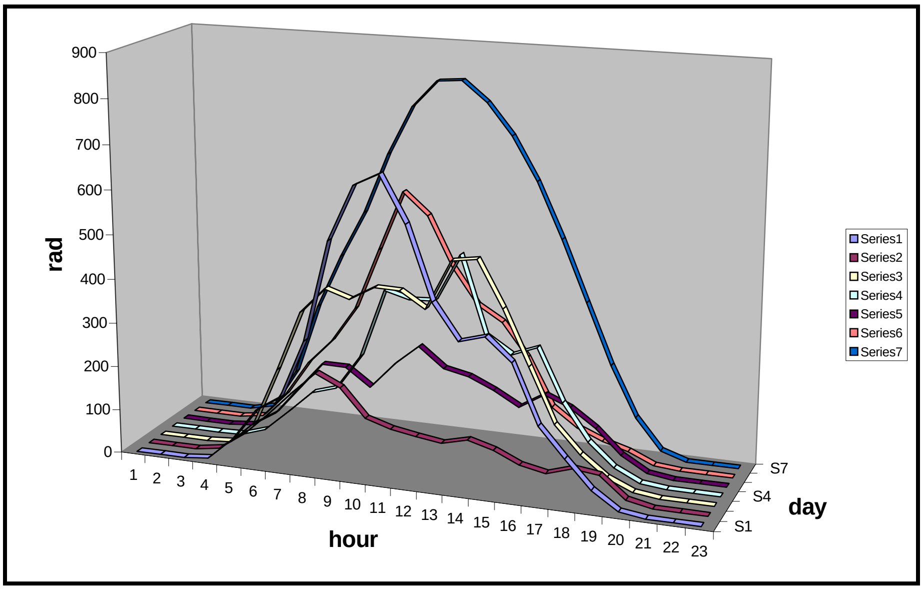
3.3.1 Plotting 2 variables categorised by qualitative variable

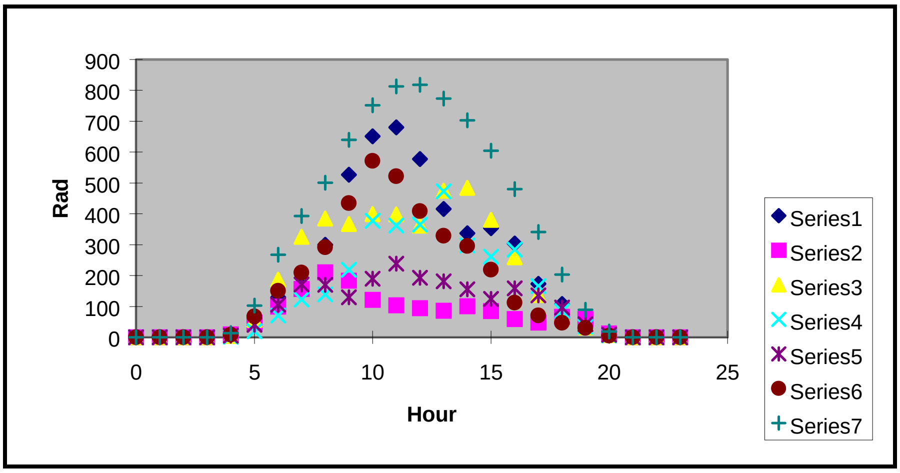
By using both the pivot table wizard and the chart wizard it is possible to obtain a graphical representation of two continuous variable categorised by some qualitative variable.

First create a pivot table, using **hour** as a row field, **day** as a column field and **rad** as a data field. Since there is only one value of **rad** in each combination of fields, the function sum simply returns this value. Once generated the table was highlighted (excluding the grand total row and column) and the **chart wizard** activated. The first graph was created using the line type and the second using the xy scatter plot

The next page shows the generated table of **rad** split by both **hour** and **day**. The graphs both show **rad** against **hour** split by **day**

Sum of rad	day							
hour	1	2	3	4	5	6	7	Grand Total
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	6	9	4	2	5	8	13	47
5	56	48	56	20	40	67	103	390
6	130	100	187	70	106	150	268	1011
7	166	157	326	122	171	209	393	1544
8	300	210	385	139	170	291	501	1996
9	527	184	367	219	130	434	640	2501
10	651	121	399	377	190	571	752	3061
11	680	104	398	362	238	522	813	3117
12	578	94	362	365	193	409	818	2819
13	416	86	475	473	181	329	773	2733
14	337	101	484	298	156	295	703	2374
15	354	85	381	261	124	219	605	2029
16	304	59	259	284	159	112	480	1657
17	173	48	137	166	136	71	341	1072
18	108	66	76	85	96	47	203	681
19	47	59	32	34	42	30	89	333
20	8	12	7	6	8	5	19	65
21	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0
Grand Total	4841	1543	4335	3283	2145	3769	7514	27430





The format of the two graphs is the default and can be changed in all respects by the user. For example the symbols on the latter graph can be customised by double clicking on each data series in turn and selecting various options from the dialog box which appears. You can also design your own symbols copying and pasting them onto the selected data series. The orientation of the 3-D line graph can also be changed by the user clicking on the graph and dragging any of the six nodes. This way the graph can be viewed from all perspectives.

3.3.1 Filtering

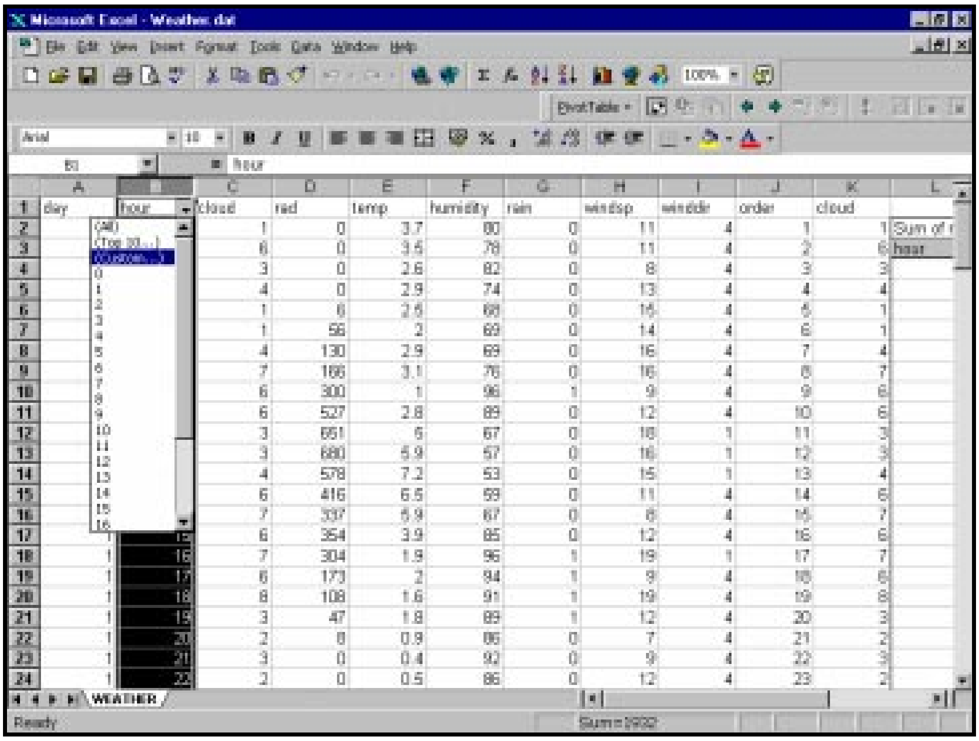
Three way relationships can often be explored by looking at the relationship between two of them for different ranges of values of the third. A convenient way of doing this in Excel is to use the Autofiller feature. Consider the relationships between cloud and temperature for different times of the day.

First categorise the time periods into “day” and “night”. Do this by considering day to be 7am to 6pm and night to be 7pm to 6am. Since the times are recorded as a 24 hour clock this translates to 0-6 (night) 7-18 (day) 19-23 (night)

So highlight the weather data and use

Data > Filter… > Autofilter

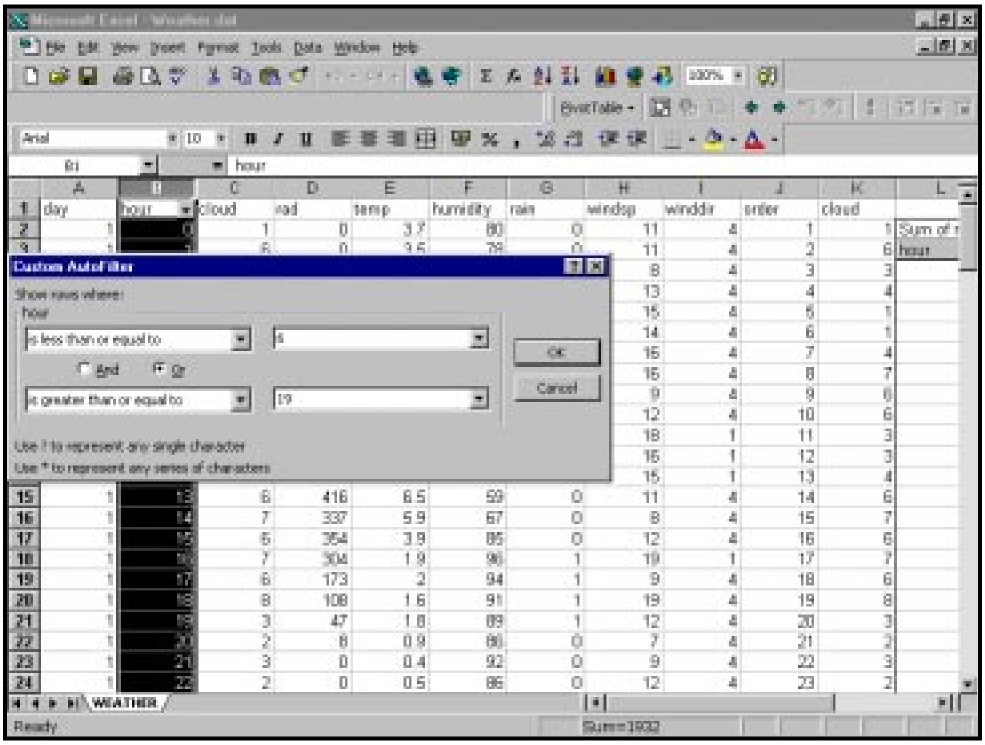
You will get drop down arrows directly on the column labels of the list. Figure 3.3.1 shows the drop down arrow for column B **(hour)**



	A	B	C	D	E	F	G	H	I	J	K	L
1	day	hour	cloud	rad	temp	humidity	rain	windsp	winddir	order	cloud	
2			1	0	3.7	80	0	11	4	1	1	Sum of r
3			6	0	3.5	78	0	11	4	2	6	hour
4			3	0	2.6	82	0	8	4	3	3	
5			4	0	2.9	74	0	13	4	4	4	
6			1	6	2.5	68	0	15	4	5	1	
7			1	56	2	69	0	14	4	6	1	
8			4	130	2.9	69	0	16	4	7	4	
9			7	186	3.1	76	0	16	4	8	7	
10			6	300	1	96	1	9	4	9	6	
11			6	527	2.8	89	0	12	4	10	6	
12			3	651	5	67	0	18	1	11	3	
13			3	680	5.9	57	0	16	1	12	3	
14			4	578	7.2	53	0	15	1	13	4	
15			6	416	6.5	59	0	11	4	14	6	
16			7	337	5.9	67	0	8	4	15	7	
17	1	15	6	354	3.9	85	0	12	4	16	6	
18	1	16	7	304	1.9	96	1	19	1	17	7	
19	1	17	6	173	2	94	1	9	4	18	6	
20	1	18	8	108	1.6	91	1	19	4	19	8	
21	1	19	3	47	1.8	89	1	12	4	20	3	
22	1	20	2	8	0.9	86	0	7	4	21	2	
23	1	21	3	0	0.4	92	0	9	4	22	3	
24	1	22	2	0	0.5	86	0	12	4	23	2	

Figure 3.3.1

By clicking on the arrow beside hour, a list of the distinct items in that column will appear. On selection of any one of them only those rows matching the constraint will appear. By choosing (Custom…) the dialogue box shown in figure 3.3.2 appears. The dialogue shown has been adjusted to select the night time values. Use the down arrows to change from "equals" to "is less than or equal to" and similarly for "is greater than or equal to". Remember to change And to Or between the two conditions.


Microsoft Excel - Weather.dat
File Edit View Insert Format Tools Data Window Help
100%
PivotTable
Arial
10
B1
hour
OK
Cancel
WEATHER
Ready
Sum=1932

Custom AutoFilter

Show rows where:

hour

is less than or equal to | 6

○ And ● Or

is greater than or equal to | 19

Use ? to represent any single character

Use * to represent any series of characters

	A	B	C	D	E	F	G	H	I	J	K	L
1	day	hour	cloud	rad	temp	humidity	rain	windsp	winddir	order	cloud	
2	1	0	1	0	3.7	80	0	11	4	1	1	Sum of r
3	1	1	6	0	3.5	78	0	11	4	2	6	hour
4								8	4	3	3	
5								13	4	4	4	
6								15	4	5	1	
7								14	4	6	1	
8								16	4	7	4	
9								16	4	8	7	
10								9	4	9	6	
11								12	4	10	6	
12								18	1	11	3	
13								16	1	12	3	
14								15	1	13	4	
15	1	13	6	416	6.5	59	0	11	4	14	6	
16	1	14	7	337	5.9	67	0	8	4	15	7	
17	1	15	6	354	3.9	85	0	12	4	16	6	
18	1	16	7	304	1.9	96	1	19	1	17	7	
19	1	17	6	173	2	94	1	9	4	18	6	
20	1	18	8	108	1.6	91	1	19	4	19	8	
21	1	19	3	47	1.8	89	1	12	4	20	3	
22	1	20	2	8	0.9	86	0	7	4	21	2	
23	1	21	3	0	0.4	92	0	9	4	22	3	
24	1	22	2	0	0.5	86	0	12	4	23	2	

Figure 3.3.2

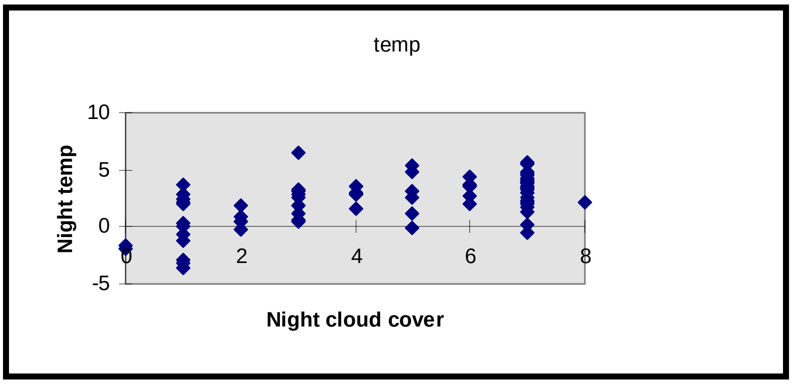
Complete the criteria as shown to use only the “night” data. Note you need to click the circle beside **Or** For day you need to click the circle beside **And** to change the conditions**.** Note if you change OR to AND then you will get no data resulting from the filter.

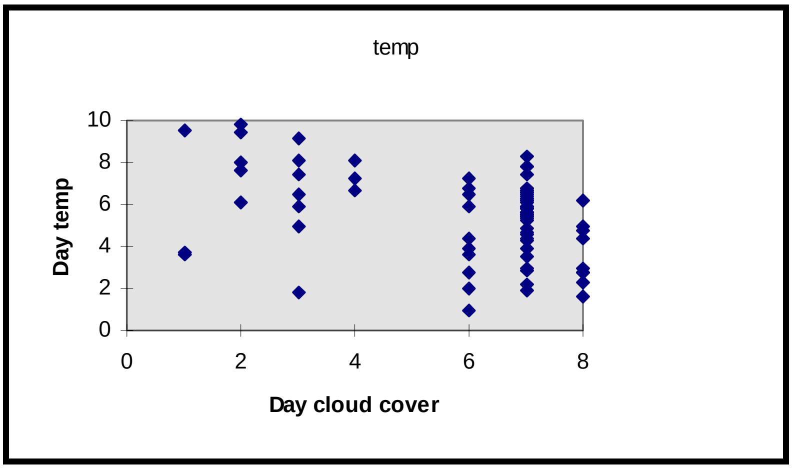
☞ **Exercise 3.5:** Use the above filtering technique to split the 24 hour data into day and night.

☞ **Exercise 3.6:** Now use the filtered data to plot night **cloud** cover against night time **temp** and day time **cloud** against day time **temp**.

☞ **Exercise 3.7:** Try plotting **humidity** split by night and day against **temperature**.

For the first of these tasks you should get plots like these:





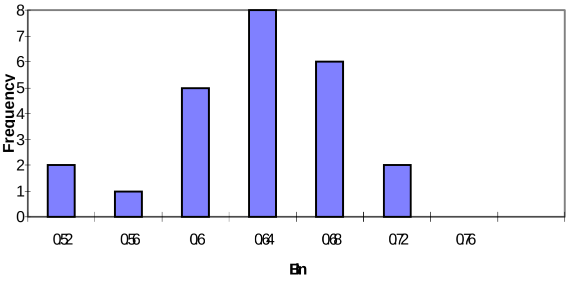
4 T-tests and analysis of variance

4.1 *Hypothesis testing*

Sometimes the exploratory methods described in the earlier chapters are sufficient for understanding data because the patterns they reveal are unambiguous and easy to interpret. On other occasions it is difficult to judge whether apparent differences or effects are real. We may wish to check formally whether such patterns could have arisen by chance.

Hypothesis testing is a statistical method to assess the extent to which the observed data are consistent with a specified hypothesis. We will illustrate some simple examples using data from a laboratory grass drying experiment. Read in the data from the file **grass.dat** into two columns, and label them **shape** and **drytime**. Shape describes the shape of the drying curves obtained from each of the 24 samples. We shall examine the shape data, using the **Descriptive Statistics** tool (see section 2.2) and the **Histogram tool** (see section 2.4.2), both of which are accessible from the **Analysis ToolPak**.

shape	
Mean	0.6152
Standard Error	0.0113
Median	0.6195
Mode	
Standard Deviation	0.0556
Sample Variance	0.0031
Kurtosis	0.4830
Skewness	-0.5819
Range	0.2340
Minimum	0.4840
Maximum	0.7180
Sum	14.7642
Count	24.0000
Confidence Level(95%)	0.0222



A model for the drying process predicts an average value of 0.58. Are the data consistent with this?

4.2 *The t-test*

The t-test is used to test the hypothesis that a sample is consistent with a specified mean, or to test for the equality of the means of two sample data sets.

4.2.1 One sample t-test

The one sample t-test tests a set of data against a particular value. In the case of our example it assesses whether the difference between 0.58 and the mean of the data could have happened by chance.

No formal one-sample t-test exists in Excel. However we can find a range of values that would be plausable for the true mean. This is called a **confidence interval**, and is a range of values that would not be rejected by a t-test at a particular **level of significance** (e.g. 5%).

In Excel a **confidence interval** can be obtained either by using the **Descriptive Statistics** tool in the **Analysis ToolPak,** or by selecting **CONFIDENCE** from the **FunctionWizard's** list of statistical functions. Both methods are fairly straightforward and hence we shall only look at one of them.

Select **Tools > Data Analysis > Descriptive Statistics** from the menubar, enter the input range of the sample, and tick the confidence interval option. The conventional approach is to accept the **hypothesised value** if it lies within the **95% confidence interval**. Thus ‘95’ would be inserted as the confidence level for the mean. Click **OK**. In the case of our example the following output appears:

shape	
Confidence Level(95.000%)	0.022240656

The **sample mean** is 0.6152, thus the corresponding 95% **confidence interval** for the mean is **(0.5929,0.6374)**. Since 0.58 doesn’t lie in this interval, the **null hypothesis** is rejected.

The higher the confidence level, the wider the interval. For example, with a 99% confidence interval:

shape	
Confidence Level(99.000%)	0.02922928

Thus a 99% confidence interval for the mean would be (0.5859,0.6444).

The theory behind the t-test and confidence intervals is explained in the appendix.

4.2.2 Two sample t-test

The second variable in the data set represents the time taken for each grass sample to reach a certain moisture content. The first six are based on samples from a grass cutting machine, and the subsequent eighteen are in batches of six from each of the three machines which condition the grass to accelerate the drying process.

Here, we do not want to compare the data with a predetermined value, as with the **shape** example, but to compare the results of one treatment with another. We shall first investigate whether there is any difference between the means of the **conditioned** and **unconditioned** grass, i.e. whether the conditioning has any effect.

The test for the difference is called the **two-sample t-test.** In Excel there are three two-sample t-tests, one for samples with unequal variance, one for samples with equal variance and a third for paired samples. The samples are not paired and we shall assume for the moment that the samples have **unequal variances**.

Select **Tools > Data Analysis > t-test: Two sample assuming unequal variances** from the menubar. In the dialog box insert the input range for each of the samples. If the samples have been labeled check the label option. Once the output range has been selected, click OK.

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	*Variable 2*
Mean	8.03915	4.225689
Variance	2.780167295	1.82203
Observations	6	18
Hypothesized Mean Difference	0	
df	7	
t Stat	5.075216959	
P(T<=t) one-tail	0.000719346	
t Critical one-tail	1.894577508	
P(T<=t) two-tail	0.001438691	
t Critical two-tail	2.36462256	

In the above example no labels were specified, and hence the two samples were issued with default labels: variable1 and variable2. **Variable1** corresponds to the **unconditioned grass** sample, and **variable2** to the **conditioned grass** sample.

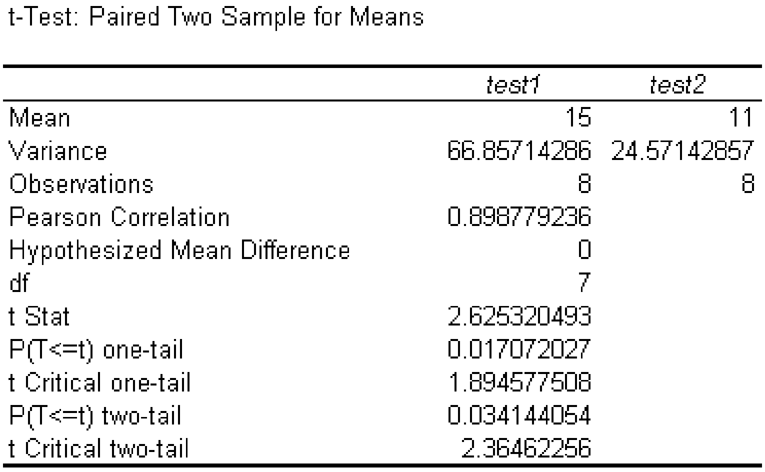
A slightly more powerful test can be performed if we can assume that the standard deviations are the same under both treatments. This is done by **`pooling'** the **standard deviations**, i.e. taking the weighted average. In Excel, this is done by activating the option '**t-test: Two sample assuming equal variance'** in the **Analysis ToolPak**. This will give a similar output, but the results are slightly different:

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	*Variable 2*
Mean	8.03915	4.225688889
Variance	2.78016729	1.822030065
Observations	6	18
Pooled Variance	2.03978853	
Hypothesized Mean	0	
df	22	
t Stat	5.66412731	
P(T<=t) one-tail	5.3643E-06	
t Critical one-tail	1.71714419	
P(T<=t) two-tail	1.0729E-05	
t Critical two-tail	2.07387529	

This, more powerful, test produces a smaller confidence interval for the difference in means, this being reflected by the p-value i.e. P(T<=t) two-tail.

The **Paired Two-Sample for Means** t-test is used when there is a natural pairing of observations in the samples, for example, when a sample group is tested twice-- before and after an experiment. Read in data from the file **tobacco.dat**. Test1 and Test2 represent the results from 2 different tests, carried out on the two halves of the same leaf, for one leaf from each of 8 tobacco plants. Selecting the option '**t-test: Paired Two-Sample for Means**' from the **Data Analysis** menu gives the following output:

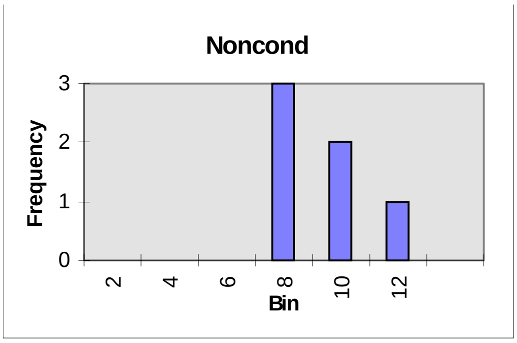


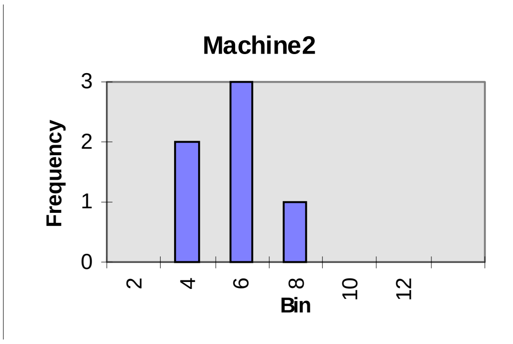
t-Test: Paired Two Sample for Means

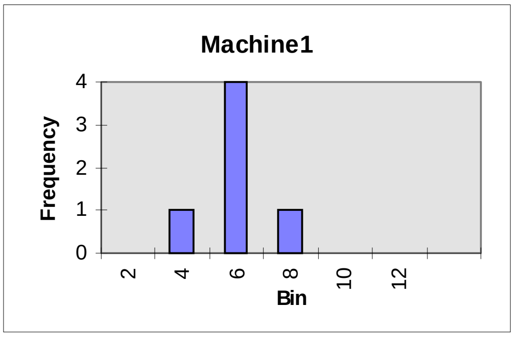
	test1	*test2*
Mean	15	11
Variance	66.85714286	24.57142857
Observations	8	8
Pearson Correlation	0.898779236	
Hypothesized Mean Difference	0	
df	7	
t Stat	2.625320493	
P(T<=t) one-tail	0.017072027	
t Critical one-tail	1.894577508	
P(T<=t) two-tail	0.034144054	
t Critical two-tail	2.36462256	

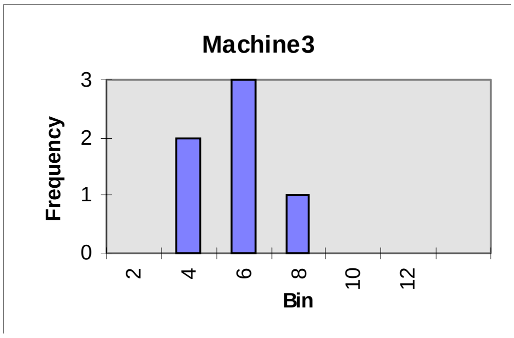
Next, using **grass.dat** again, we shall compare the different conditioned treatments. We can compare any pair of treatments using a t-test, as above.

Set up four columns labeled : **noncond, machine1,machine2,machine3.** By copying and pasting transfer the drytime data to each of these columns, the first 6 values being placed under noncond, the next six under machine1, and so on. Using the **Histogram tool** in the **Analysis ToolPak** plot a histogram for each of the four samples.

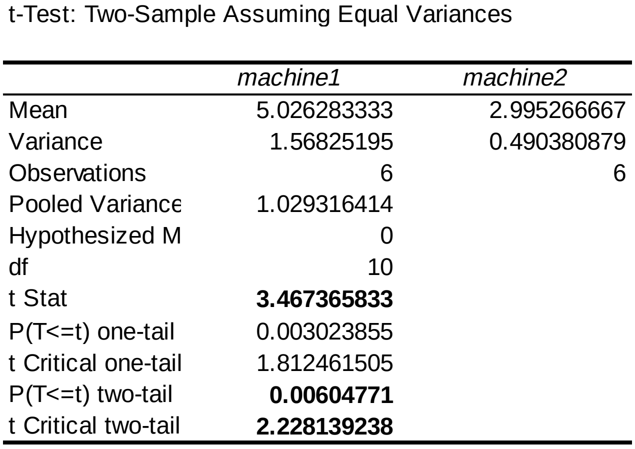






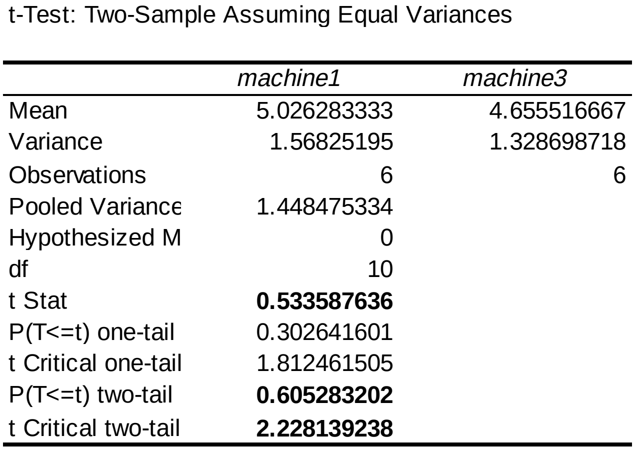


Considering the relatively small sample size it is reasonable to assume a Normal distribution. You can test the pairs of treatments as before using the **'t-test: two sample assuming equal variance'** option in the **Analysis ToolPak.** For example:



t-Test: Two-Sample Assuming Equal Variances

	machine1	*machine2*
Mean	5.026283333	2.995266667
Variance	1.56825195	0.490380879
Observations	6	6
Pooled Variance	1.029316414	
Hypothesized M	0	
df	10	
t Stat	**3.467365833**	
P(T<=t) one-tail	0.003023855	
t Critical one-tail	1.812461505	
P(T<=t) two-tail	**0.00604771**	
t Critical two-tail	**2.228139238**	



t-Test: Two-Sample Assuming Equal Variances

	machine1	*machine3*
Mean	5.026283333	4.655516667
Variance	1.56825195	1.328698718
Observations	6	6
Pooled Variance	1.448475334	
Hypothesized M	0	
df	10	
t Stat	**0.533587636**	
P(T<=t) one-tail	0.302641601	
t Critical one-tail	1.812461505	
P(T<=t) two-tail	**0.605283202**	
t Critical two-tail	**2.228139238**	

and so on.

4.3 Analysis of Variance

Alternatively, we can use a single test to look for treatment differences, using an **F-test.** This is called an **analysis of variance.** It allows us to test the hypothesis that the means are the same under all treatments.

In Microsoft Excel the **Analysis of Variance tools** can be found in the **Analysis ToolPak.**

Three options exist which enable the user to carry out either a **singe factor ANOVA**, a **two factor ANOVA with replication**, or a **two factor ANOVA without replication**.

There are a number of limitations associated with both the ANOVA tools and the Regression tool. The first is the inability to include **missing values** in the data set (Excel views these as text values), and the second is the inability to perform more complex analyses on the data within Excel once the standard ANOVA or linear regression techniques have been applied. However it must be emphasized that for a basic statistical analysis these tests are sufficient. As well as being easy to apply to a data set they offer a number of additional features such as residual plots, and Normal probability plots. The later are a necessity in a regression analysis as they enable the user to identify whether the assumptions of the analysis (e.g. a Normal residual variance) have been met.

4.3.1 One-way analysis of variance

Arrange the treatment groups in adjacent columns and select:

Tools> Data Analysis> Anova: Single factor

Insert the input range, including the treatment labels if these are available. After selecting an output range, click on the OK button. Carry out a singe factor analysis of variance on the four conditioning grass treatments: **noncond, machine1,machine2**, and **machine3**. Your output should look like this:

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
noncond	6	48.2349	8.03915	2.780167
machine1	6	30.1577	5.026283	1.568252
machine2	6	17.9716	2.995267	0.490381
machine3	6	27.9331	4.655517	1.328699

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	79.47904	3	26.49301	17.18234	9.31E-06	3.098393
Within Groups	30.83749	20	1.541875			
Total	110.3165	23				

When there are only two treatments, the **F-statistic** is simply the **pooled t-statistic squared**, as can be seen by testing the effect of conditioning:

Replace the three machine treatment groups with a single treatment group labelled **cond** and select:

Tools > Data Analysis > ANOVA: Single factor

The following output should be obtained:

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
uncond	6	48.2349	8.03915	2.780167
cond	18	76.0624	4.225689	1.82203

ANOVA

Source of Variatior	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	65.44119	1	65.44119	32.08234	1.07E-05	4.300944
Within Groups	44.87535	22	2.039789			
Total	110.3165	23				

From the earlier example, the **pooled t-statistic** was found to be 5.66, which is the square root of 32.08, the F-statistic.

4.3.2 Two-way analysis of variance

The value of an experiment can usually be increased greatly by careful design, prior to beginning the experiment. Suppose a set of treatments are being compared in an experiment involving several **replicates**. If the replicates can be put into groups of similar type, where the variation is less than over the entire experiment, then comparisons between the treatments can be made with greater accuracy. These groups are known as **blocks**. It is also important to randomise the allocation of treatments to units within the blocks, in order to guard against the possible effects of patterns within blocks, or in the way they are managed (e.g. one area of a field may be muddy, lambs may have the same mother etc.).

We shall illustrate this using data from a field experiment on the drying rate of grass. Import the data set **silage.dat** and store the data in 5 columns: **initmois, dryrate, treatmt, block** and **index** respectively. You should always examine the data using graphs before any analysis is performed, so we shall examine the pattern of the data by looking at a **scatterplot**, categorised by the treatment types.

To perform the ANOVA analysis the **dryrate** data must first be unstacked into **7 columns**, representing **treatment**, and **3 rows** representing the **replicate**. This can be done either by copying and pasting into new worksheet cells, or by using the **PivotTable Wizard** to create the table below.

Sum of dryrate	treatmt							
block	1	2	3	4	5	6	7	Grand Total
1	2.08	2.24	2.54	2.84	2.42	2.68	2.78	17.58
2	2.19	2.14	2.19	3.37	2.61	3.01	3.02	18.53
3	2.25	2.07	2.02	2.77	2.52	1.88	2.34	15.85
Grand Total	6.52	6.45	6.75	8.98	7.55	7.57	8.14	51.96

If you prefer, you can reformat the data by copy-and-paste.

Once the data has been reformatted in this way, make the following selection:

Tools > Data Analysis > ANOVA: two-way without replication

NB Excel views replication in a different context. In the case of our example 'two-way ANOVA without replication' must be selected even though replication is used in the experimental model. For further information see the Help documentation for the two-way ANOVA.

Enter the input range, including the row and column labels but not the Grand Totals and the output range. Click OK. Your output should look like this:

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
1	7	17.58	2.511429	0.079581
2	7	18.53	2.647143	0.244957
3	7	15.85	2.264286	0.095362
1	3	6.52	2.173333	0.007433
2	3	6.45	2.15	0.0073
3	3	6.75	2.25	0.0703
4	3	8.98	2.993333	0.107633
5	3	7.55	2.516667	0.009033
6	3	7.57	2.523333	0.337633
7	3	8.14	2.713333	0.118933

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	0.527514	2	0.263757	4.011419	**0.046338**	3.88529
Columns	1.730381	6	0.288397	4.386158	**0.014178**	2.996117
Error	0.789019	12	**0.065752**			
Total	3.046914	20				

It is clear that we are confident about the existence of a treatment effect. Our estimates of the treatment means will also be more accurate, than if we had not used a block structure.

Once we've decided that treatment differences exist, we may want to identify what the differences are by comparing particular individual treatments. We do this by constructing **t-tests**, using the mean square error in the ANOVA as the estimate of the variance.

The **standard error** of the **differences** between pairs of treatments is calculated by :

$$sed = \sqrt{MSErrorx(\frac{1}{n_1} + \frac{1}{n_2})}$$

where n_1 and n_2 are the number in each treatment. You can calculate this value by selecting an empty cell on a worksheet and typing **= SQRT(MSError*((1/n_1) + (1/n_2)))** where **MSError** is the cell reference of the mean square error, and use this value to test for differences between pairs of treatments by :
You can calculate a **p-value** from t using the following steps:

$$t = \frac{(mean1 - mean2)}{sed}$$

1. Activate the **FunctionWizard** fx on the toolbar.
2. Select **TDIST** from the list of statistical functions.
3. Enter the **absolute t-value** as x, the number of **degrees of freedom** of the MSE as **df**, and **2** as the **number of tails**.

The answer given (call it p, say) is the probability that t could be greater than the absolute value of the number you fed in. This is known as the **p-value**.

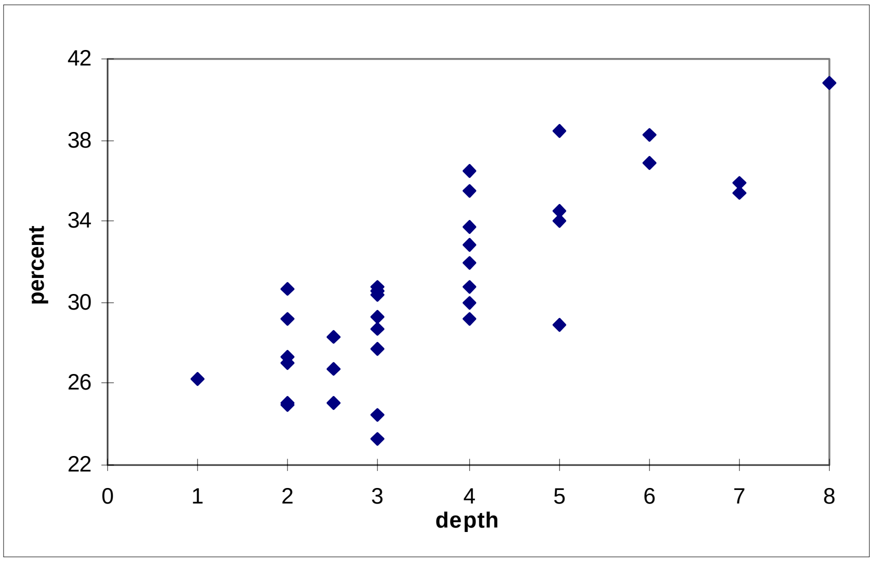
For example, for a comparison between treatments 1 and 2,
s.e.d = 0.209367, df =12, t = 0.11145, and p = TDIST(0.11145,12,2) =0.913104.
i.e. There is no evidence of a difference between treatments 1 and 2.

5 Regression

Regression is a method for describing, estimating and testing the relationship between a **response variable** (e.g. live weight of a sheep) and some **explanatory** (also called the **predictor**) **variable(s)** (e.g. age, feeding etc.). The simplest situation is a **linear relationship**, with one response variable (y) and one explanatory variable (x). The **y** variable is a **random** variable**,** while the **x** variable may be another r**andom** variable or it may be a **controlled** variable (e.g. taking measurements at specific times).

5.1 Simple linear regression

Read the file **carcass.dat** into three columns, labelled **grade**, **percent** and **depth**, and examine the plot of percent against depth.



This plot shows a reasonably linear relationship, so we will try **linear regression.**
Tools > Data analysis... > Regression
Enter the percent cell range in the Y range box, and the depth cell range in the X range box. You will need to have removed the row with the missing depth before doing the regression or you will get an error message about non-numeric data. Ensure that the labels box is checked if necessary. Click on OK. The output should look like this:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8152
R Square	0.6646
Adjusted R Square	0.6547
Standard Error	2.5972
Observations	36.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.0000	454.4441	454.4441	67.3697	0.0000
Residual	34.0000	229.3479	6.7455		
Total	35.0000	683.7920			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99%*	*Upper 99%*
Intercept	22.5755	1.0950	20.6162	0.0000	20.3501	24.8009	19.5878	25.5632
depth	2.2263	0.2712	8.2079	0.0000	1.6751	2.7775	1.4863	2.9664

The **standard error** is calculated as the square root of the **Error Mean Square (MS). R** squared indicates how much of the total variation in the dependent variable (y) is explained by the regression. The closer R squared is to 100, the better the fit of the line to the data. R squared is the square of the coefficient of correlation (when the variables are random).
R-square (adj.) is the R-sq. adjusted for degrees of freedom.

	depth	*percent*
depth	1	
percent	0.815226	1

Next comes a familiar Analysis of Variance table, with the p-value from the hypothesis that the variables are independent (i.e. the variable percent does not change as depth changes).

The final table gives detailed information about the **regression equation**. The first column contains the regression **coefficients**. These describe the best fitting straight line:

percent = 22.6 + 2.23 depth.

The latter columns give further information about these coefficients, including their **standard deviation**. These values are useful for obtaining confidence intervals of the intercept and slope. The **t-ratio** and the **p-value** of a coefficient are under the hypothesis that the coefficient is 0. The last columns specify **confidence intervals** for the parameter estimates at both the 5% level and at the level you specified in the initial regression dialog box.

5.2 Regression options

1) Residuals

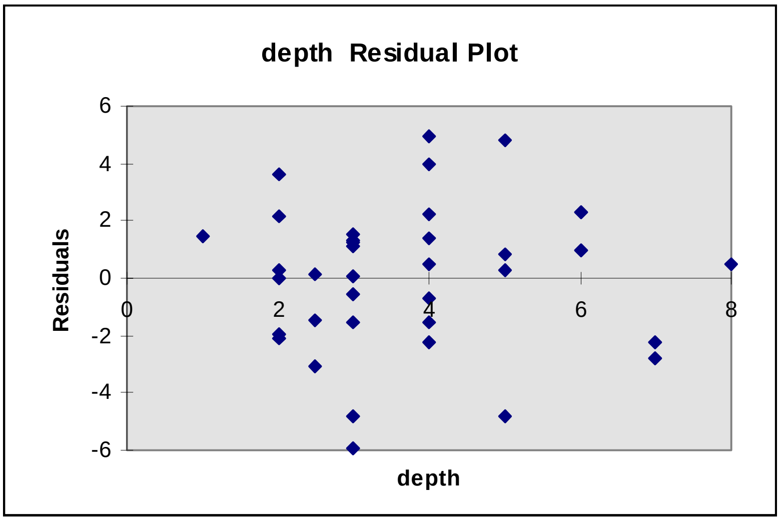
It is always a good idea to check your residuals, to make sure that there is no pattern (e.g. to check that the residuals don't become more variable as x increases, etc.) and to check for **outliers** (i.e. data points with a standardised residual much greater than +/-2). To do this call up the regression dialog box as before and check the options for residuals, standardised residuals and residual plot. The following will appear:

RESIDUAL OUTPUT

Observation	*Predicted percent*	*Residuals*	*Standard Residuals*
1	31.4807	3.9793	1.5321
2	33.7070	0.8330	0.3207
3	27.0281	2.1519	0.8285
4	31.4807	0.4593	0.1768
5	33.7070	-4.8170	-1.8547
6	29.2544	1.0956	0.4218
7	28.1413	-3.0813	-1.1864
34	38.1597	-2.7797	-1.0702
35	35.9333	0.9667	0.3722
36	31.4807	-0.6707	-0.2582

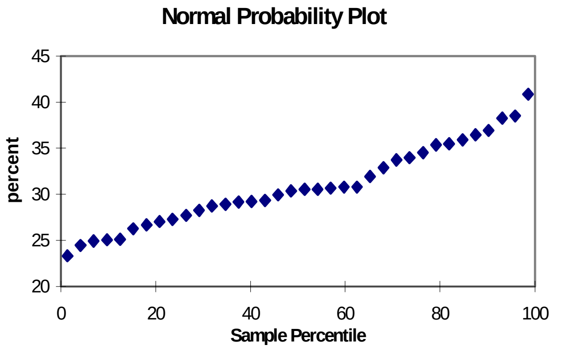
PROBABILITY OUTPUT

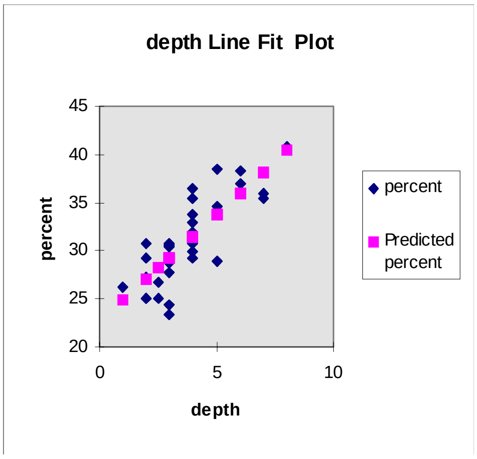
Percentile	*percent*
1.3889	23.3
4.1667	24.44
6.9444	24.96
9.7222	25.06
12.5000	25.1
15.2778	26.27
18.0556	26.71
93.0556	38.23
95.8333	38.49
98.6111	40.85



There does not appear to be any trend. If there was, then a transformation of at least one of the variables could be tried.

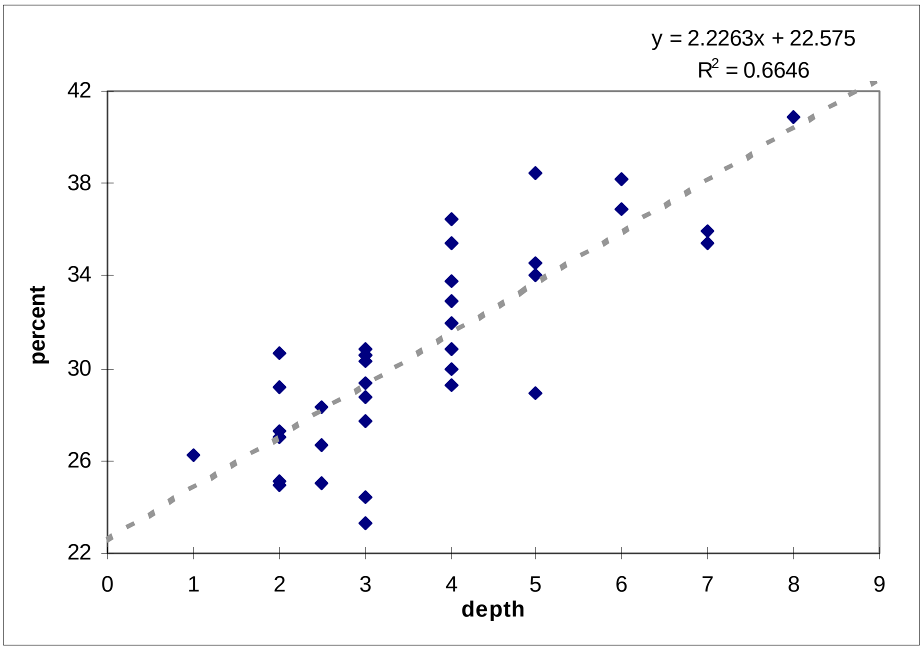
Two other options available are for a **line fit plot** and a **normal probability plot.**





2) Adding a line through your data.

In addition to the line fit plot option above it is also possible and more useful to fit a straight line through a data series by clicking on the series, selecting **Chart > Add Trendline** and selecting the **linear** option on the **Type** tab. Options also exist to display the regression equation and R^2 value on the chart.


y = 2.2263x + 22.575
R^2 = 0.6646
percent
42
38
34
30
26
22
0
1
2
3
4
5
6
7
8
9
depth

6 Working with databases

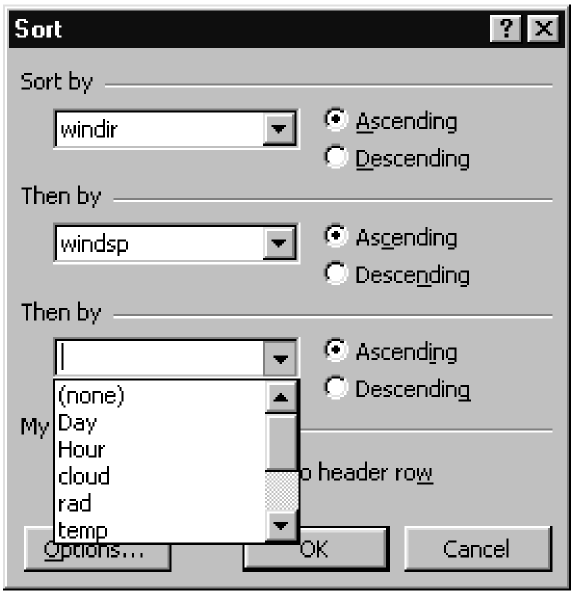
Excel 97 provides powerful functionality for managing and analysing data stored in databases. The application distinguishes between two types of database : internal and external databases. Internal databases reside in worksheets and are referred to as lists. The fields are positioned in columns and the records in rows. Most tables of data on a worksheet can be treated as internal databases. External databases include dBase, Paradox, Oracle, and FoxPro. Excel is able to access information from these via another Microsoft application, MS Query. Through this, data satisfying a set criterion can be imported into an Excel workbook.

It is wise to gain some familiarity with the database features offered in Excel, even if you are not intending to work with databases in the future ; the database features available can be used to great effect when analysing any type of data set, as shall be demonstrated in the following pages.

Read the file weather.dat into nine columns and label them **day, hour, cloud, rad, temp, humidity, rain, windspd,** and **winddir** respectively. If you have covered section 3 of this course then you will already have imported this data into Excel. The dataset consists of nine fields (i.e. variables) and 168 records (i.e. observations). When performing most database tasks - such as finding, sorting, or subtotaling data, Microsoft Excel will automatically recognise this sort of list as a database.

6.1 Sorting data

Databases can be sorted by any number of fields in Excel. To perform a sort select any cell in your database (list) and click on **Data > Sort**. Provided there are no blank rows or columns in your dataset Excel will automatically calculate the database range. A sort can be performed by three fields at a time. Sorting by more than three fields will require a sequence of independent sorts, sorting by the least significant fields first.



Sort

Sort by: windir ● Ascending ○ Descending

Then by: windsp ● Ascending ○ Descending

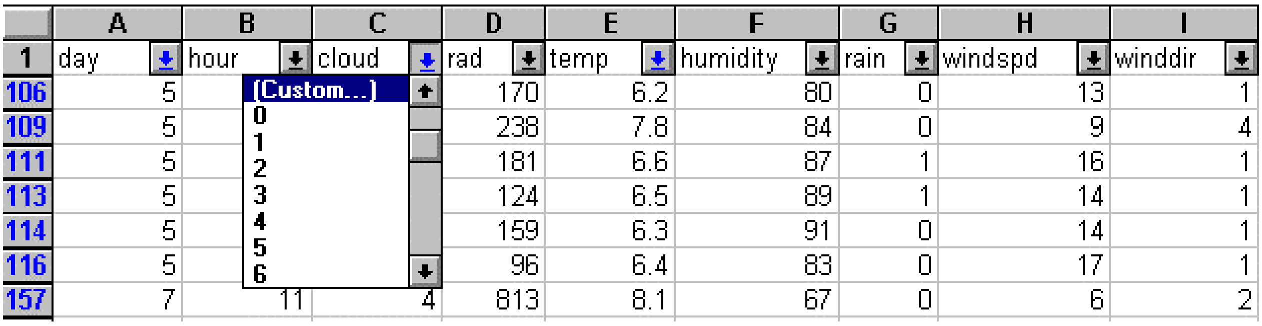
Then by: | ● Ascending ○ Descending

6.2 *Filtering data*

Filtering a database enables you to display only those records that meet a certain criteria. There are two methods of filtering data in Excel. One uses autofilters and the other advanced autofilters.

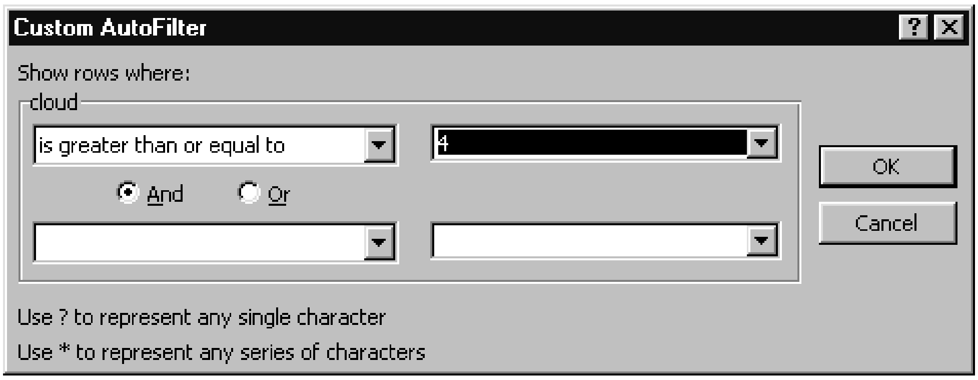
6.2.1 Autofilters

Select any cell in your data range and click on **Data > Filter > AutoFilter**. Drop down controls will be placed on top of the field names, listing all the categories present in each field.



	A	B	C	D	E	F	G	H	I
1	day	hour	cloud	rad	temp	humidity	rain	windspd	winddir
106	5			170	6.2	80	0	13	1
109	5			238	7.8	84	0	9	4
111	5			181	6.6	87	1	16	1
113	5			124	6.5	89	1	14	1
114	5			159	6.3	91	0	14	1
116	5			96	6.4	83	0	17	1
157	7	11	4	813	8.1	67	0	6	2

Data can be filtered by any number of categories in any number of fields. The custom option in each of the drop-down lists allows relationships other than equal to be specified for the filter.



Custom AutoFilter

Show rows where:

cloud

is greater than or equal to 4

● And ○ Or

Use ? to represent any single character

Use * to represent any series of characters

Filtering data in this way has many advantages when performing a preliminary statistical analysis. Not only does it facilitate the exploratory data analysis of several variables, but it also saves time when creating charts and pivot tables, as these will be updated automatically every time the criteria range is adjusted.

6.2.2 Advanced autofilters

This method differs from autofilters in that criteria is specified for the filter using a criteria range. This is best illustrated by example.

Rain	cloud	temp
=1	<4	
=1		>7

This is known as a criteria range and is placed on a worksheet. By selecting **Data > Filter > Advanced filter...** and specifying the range of the table above as the criteria, all observations with rain =1 and cloud cover <4 or rain=1 and temperature > 7 will be displayed.

The overriding benefits of using advanced autofilters as opposed to the standard autofilters are:

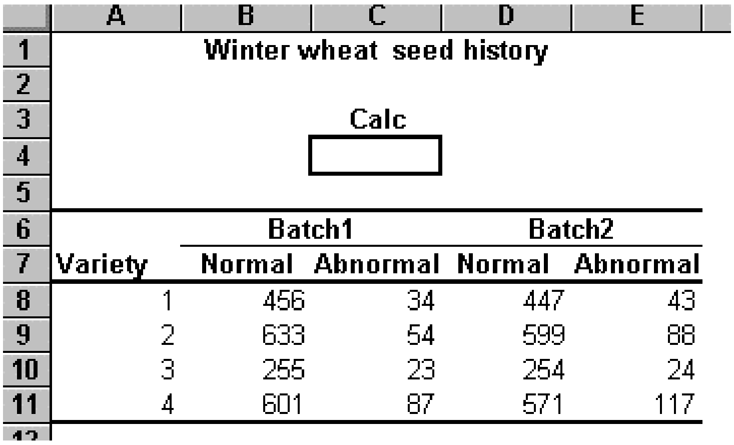
1) Multiple criteria is possible

2) An unlimited no. of criteria per field can be specified.

3) Computed criteria is allowed

4) Variables can be used as part of the criteria.

A computed criteria is simply a criteria which contains a formula. The ability to specify a formula makes filtering in this way extremely powerful.



	A	B	C	D	E
1		Winter wheat seed history			
2					
3			Calc		
4					
5					
6		Batch1		Batch2	
7	Variety	Normal	Abnormal	Normal	Abnormal
8	1	456	34	447	43
9	2	633	54	599	88
10	3	255	23	254	24
11	4	601	87	571	117

Consider the table above. This lists the number of normal and abnormal seeds found in two batches of germinating seeds from each of four varieties. Suppose we wish to display only those varieties where the proportion of normal seeds is greater in batch 1 than in batch 2. To do this we place the following formula in the calculate box and filter the data range A7:E11 specifying C4 as the criteria range.

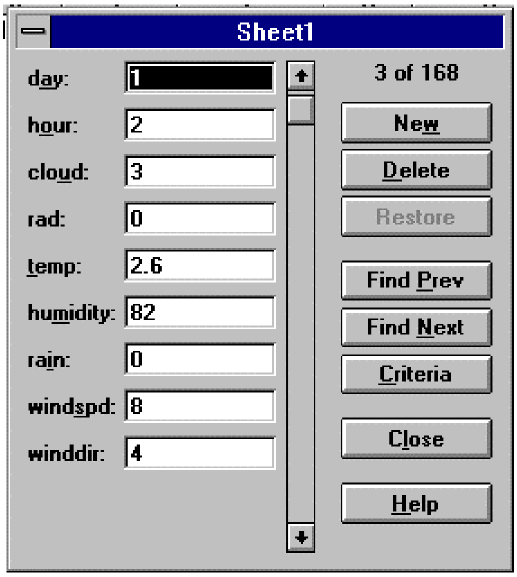
Criteria range: '=B8/(B8+C8)>D8/(D8+E8)'

NB Labels must be unique when performing any kind of database operation on a list. Thus the labels in range B7: E7 should be renamed before attempting the filter outlined above eg Norm1, Abnorm1, Norm2, Abnorm2 respectively.

6.3 Data forms

Not only can databases be edited on a spreadsheet, but they can also be edited and filtered using a data form. To display data in a data form select any cell in the data range and click on **Data > Form...** To set a criteria range click on the criteria option and enter your criteria into the relevant field boxes. Both expressions such as ">300" and computed criteria can be matched.

To display those records matching the set criteria use the Find Next and Find Previous buttons. To regain access to the original list, choose the criteria button and then click on the Clear button.



day:	1
hour:	2
cloud:	3
rad:	0
temp:	2.6
humidity:	82
rain:	0
windspd:	8
winddir:	4

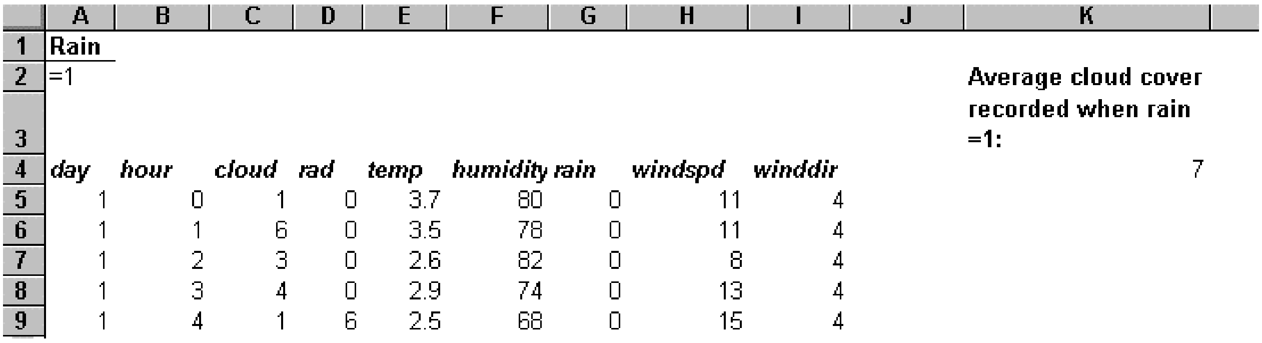
☞ **Exercise 6.1:** By experimenting with both autofilters and advanced autofilters display all those observations made on days five and seven with cloud cover greater than or equal to 4 and temperature greater than 6. Perform the same operation using the **data form** criteria facility.

6.4 D functions

These functions have been created to facilitate database calculations. They are used to summarise only those values that meet a complex criteria, and are of the following form:

DFUN(Database range, Field, Criteria range)

In order to use a database function a criteria range must be specified (see Section 6.2.2).



	A	B	C	D	E	F	G	H	I	J	K
1	Rain										
2	=1										Average cloud cover recorded when rain =1:
3											
4	*day*	*hour*	*cloud*	*rad*	*temp*	*humidity*	*rain*	*windspd*	*winddir*		7
5	1	0	1	0	3.7	80	0	11	4		
6	1	1	6	0	3.5	78	0	11	4		
7	1	2	3	0	2.6	82	0	8	4		
8	1	3	4	0	2.9	74	0	13	4		
9	1	4	1	6	2.5	68	0	15	4		

The value of K4 in the example above was calculated using the DAVERAGE function (ie “=DAVERAGE(A4:I172,”cloud”,A1:A2)”). What is the average cloud cover when rain=0?

See Section 9 for a complete list of all database functions.

7 Bits and Pieces

This chapter will cover some miscellaneous topics not always discussed with the experimental scientist to any great degree. It may not be covered in the two days of teaching, but is important none the less.

7.1 Count Data

The majority of the data used in the examples so far have been either continuous and/or categorical. However, count data often forms a large part of an experimenter's results, and should not always be analysed in the same way as continuous data. To illustrate this, open the worksheet **bovine.xls**. The data concern four types of bovine disease recorded over two years, 1991 and 1992.

7.1.1 Chi-squared tests

An experiment in which each observation may be classified by two types of category can be summarised in a table, as shown in section 3.3. This type of table is also called a **two-way contingency table** and can be analysed using a chi-squared test. The Null Hypothesis for a chi-squared test is that the two category types are independent, i.e. for the data in the example, the proportion of each disease is the same for each year. The **Chi-squared test** in Excel calculates the **p-value** of the **Chi-squared statistic**, given both the actual data values and the expected data values (i.e. the number of animals we would expect in each category, if they were independent). Each category combination is called a **cell** (e.g. the cell for BVDV in 1997 contains 597 animals). In order to perform the test the data must be arranged in the following format:

Actual

	BVDV	IBR	PI3	RSV	Total
1	597	259	72	91	1019
2	649	231	106	96	1082
Total	1246	490	178	187	2101

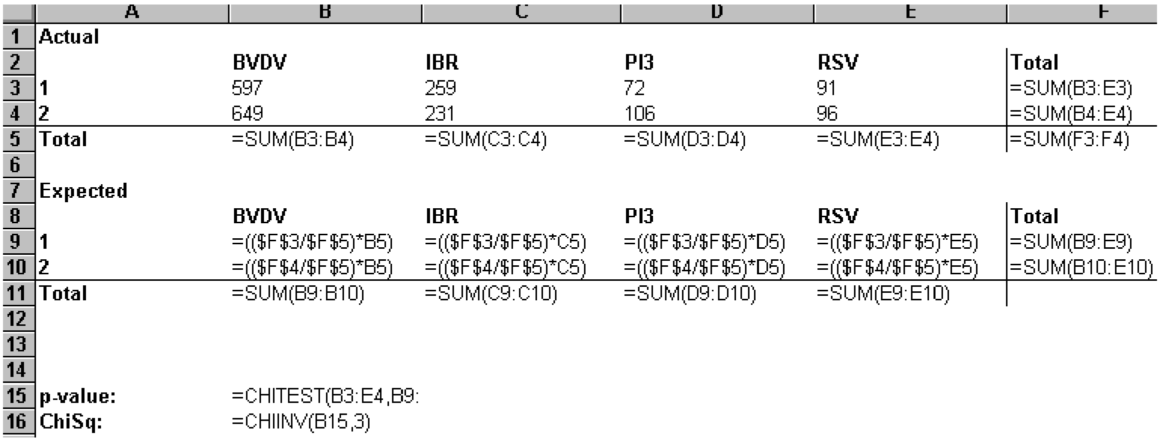
Expected

	BVDV	IBR	PI3	RSV	Total
1	604.32	237.65	86.33	90.70	1019
2	641.68	252.35	91.67	96.30	1082
Total	1246	490	178	187	

The expected values for each cell can be calculated using formulae and the autofill facility. They are given by:

$$\mathbf{E_{ij} = (row\ total)\ x\ (column\ total)\ /\ (overall\ total)}$$

Below is a table of formulae for the example above. Note that the reference $A$2 is unchanged by the autofill facility whereas A2 is adjusted accordingly for each cell.



	A	B	C	D	E	F
1	**Actual**					
2		**BVDV**	**IBR**	**PI3**	**RSV**	**Total**
3	1	597	259	72	91	=SUM(B3:E3)
4	2	649	231	106	96	=SUM(B4:E4)
5	**Total**	=SUM(B3:B4)	=SUM(C3:C4)	=SUM(D3:D4)	=SUM(E3:E4)	=SUM(F3:F4)
6						
7	**Expected**					
8		**BVDV**	**IBR**	**PI3**	**RSV**	**Total**
9	1	=(($F$3/$F$5)*B5)	=(($F$3/$F$5)*C5)	=(($F$3/$F$5)*D5)	=(($F$3/$F$5)*E5)	=SUM(B9:E9)
10	2	=(($F$4/$F$5)*B5)	=(($F$4/$F$5)*C5)	=(($F$4/$F$5)*D5)	=(($F$4/$F$5)*E5)	=SUM(B10:E10)
11	**Total**	=SUM(B9:B10)	=SUM(C9:C10)	=SUM(D9:D10)	=SUM(E9:E10)	
12						
13						
14						
15	**p-value:**	=CHITEST(B3:E4,B9:				
16	**ChiSq:**	=CHIINV(B15,3)				

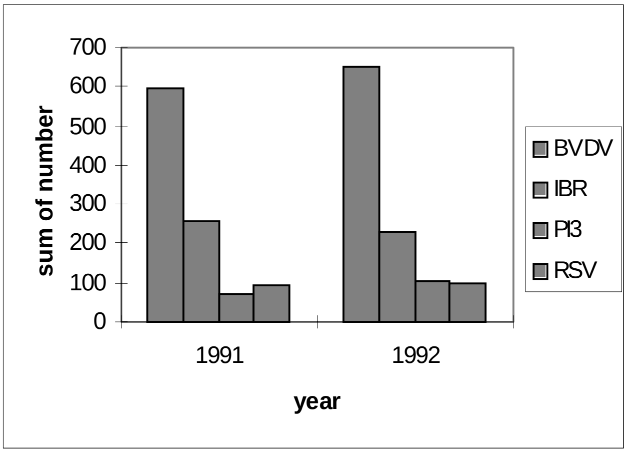
To perform the test click on the **FunctionWizard** and select **CHITEST** from the list of statistical functions. Enter the range for the table of observations (excluding total values and labels) and the range for the table of expected values (again excluding total values and labels) . Click **OK.** The **p-value** of the **Chi-squared statistic** will appear in the active cell.

The value of the Chi-squared statistic (i.e. 8.517) can be obtained by selecting **CHIINV** from the **FunctionWizard** and entering the **p-value** as '**probability**' and **3** as the '**degrees of freedom**'. The degrees of freedom are given by **DF=(r-1)(c-1)** where r is the no. of rows and c the no. of columns.

Note: If more than 20% of cells have expected values of less than 5, the chi-squared test should not be used as it stands. One way around this problem is to combine cells.
The Chi-squared test should not be used with very small numbers, as it is only an approximation. If you have very small values, Fisher's Exact Test is more appropriate.

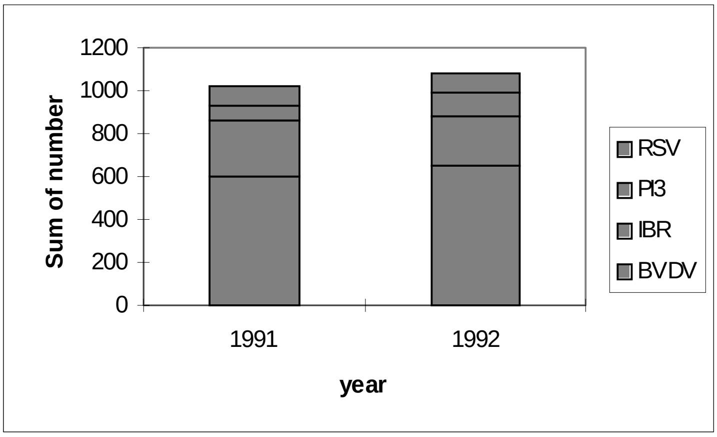
7.1.2 Graphs

Histograms and dotplots are useful for comparing a few graphs, but if you have, say, three treatments on two types of animal (control and vaccinated), then presenting these graphically with a histogram is difficult. In the example above, comparing four disease types over two years would be difficult, and impossible to do on just one histogram. A more suitable type of graph is the **chart**, which can produce, for the above data set, the following graph:

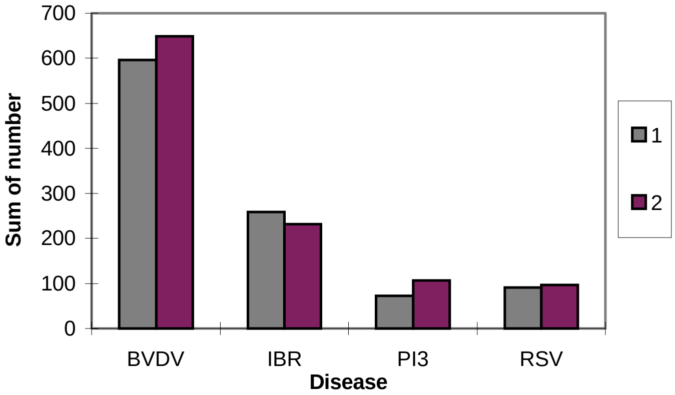


To produce this graph choose the **Column-Format1** chart type option from the **Chart Wizard.**

You might prefer to have your columns on top of each other, rather than side-by-side. To do this select the **Column-Format3** chart type from the **Chart Wizard.** This will produce a chart similar to the one below.



You may wish to compare the disease between years. To do this simply change the orientation of the data series in the ChartWizard i.e. change the option for data series to rows if the rows represent the year, or to columns if the columns represent the year.



You may also wish to change the **colour** and **shading** of your charts. To do this simply double click on the data series you want to reformat. A dialog box will appear containing a number of attributes associated with the data series which the user can customise. Make your selection and click on the OK button. The chart will be updated accordingly. Similarly changes can be made to **legends** and **axes** simply by double clicking on them. To add titles, legends or gridlines to a chart activate the chart by double clicking its border and select the appropriate option from the menu.

7.2 Random allocation

Randomization of treatments to units (e.g. sheep) must be used to make the comparison between treatments ‘fair' and unbiased. Since we don't know which units are going to respond better to the treatments we must make sure that each unit has an equal chance of being allocated a particular treatment. Randomising the allocation ensures that the results aren’t biased, and that we can perform valid statistical analyses.

7.2.1 Simple Randomization

This is the most straightforward type of randomization. We shall illustrate it using a simple example. Suppose we have a flock of 12 sheep that we wish to randomly allocate to two groups: control and treatment. To do this we would arbitrarily number the sheep 1 to 12, in any order. and pick 6 for one group and 6 for the other group at random.
The **Sample tool** in the **Analysis ToolPak** generates a random sample from a given population. However values for the sample are drawn with replacement (i.e. a population value can be selected more than once) , and thus is inappropriate for our needs.
An alternative method follows:

1. Insert the values 1 to 12 in a worksheet column
2. In an adjacent column generate random numbers between 0 and 1 using the function **RAND** (i.e. type **'=RAND()'** into each cell). Here the **Autofill** facility can be employed to great effect.
3. Copy the **values** of the two columns and paste them to the same location using the **Copy** and **Paste special > values** commands. This is essential for the next step; copying only the values disassociates them from the randomization formulae.
4. Select the values and click on **Data > Sort** . Sort by the column containing the randomised values.

The 12 sheep have now been randomised and can be allocated to one of the two treatment groups. For example, on a single run the sheep 1...12 were randomised (sorted) as follows: 6,4,9,12,1,8,5,11,10,7,3,2. Hence the control group contained the sheep 1,4,6,8,9,12 and the treatment group contained the sheep 2,3,5,7,10,11.

7.2.2 Equality of groups (or blocking)

It is often a good idea to make your groups as similar as possible, given known characteristics of the units (e.g. the live weight of the sheep). One possible solution is to pair similar sheep together (for two treatments), and number one in each pair 1 and the other 2. Now, we have six pairs of similar sheep. We still need to assign treatments randomly, so we select a sheep at random from each pair. To do this enter 3 ones and three twos into a column. Click **Tools > Data Analysis > Sampling** and insert the range for these data values into the input range box. Select the random option, the no. of samples (i.e. 6) and define an output range. Click OK. The result from this run was 1,2,1,1,2,1. The appropriate sheep were then allocated to the control group and the remainder to the treatment group. This is a form of blocking.

7.3 Repeated measures

Repeated measures are measurements of the same characteristic on the same unit on more than one occasion, such as measuring the weights of a group of sheep over a period of weeks. The analysis of repeated measures data is not a simple technique, and it is often easier to circumvent, by selecting a particular feature and analysing that.

7.3.1 The basal level

Variation between individual units is often the cause of failing to find a significant result. With repeated measures, it may be possible to remove the between unit variation by letting the unit serve as its own control. For example, measuring the weight of each sheep before any treatment can serve as the basal level for the sheep. Removing this basal level, by subtraction from each subsequent measurement, leaves just the change over time, and removes the between sheep variation. It is often better to record more than one basal value per unit, and use the mean of these as the basal level.

7.3.2 Response feature analysis

In practice, there is often a particular feature of interest to the experimenter that can be specified prior to the analysis. For example, the height and timing of a peak in the measurements, length of time required to return to a basal level etc. If such features can be identified, this reduces the analysis to a simple ANOVA checking for differences between the groups. Alternatively, you may be interested in whether there is any noticeable difference between the groups at a particular time. Again, this reduces the analysis to a standard ANOVA.

☞ **Exercise 7.1 :** An example of this type of analysis may be done on the data set **TEMP.XLS**. It concerns the temperatures recorded during the progression of a disease. The sheep are in three groups: sheep 1-5 are given treatment 1, sheep 6-10 are given treatment 2, and sheep 11 and 12 are the control. The sheep have been randomised and the sheep numbers are just for ease in identifying them in this example. Temperatures were recorded once a day for 20 days (-2 to 17), with the challenge given on day 0. The features of interest in this case are the maximum temperatures recorded and the day on which this maximum occurred.

7.3.3 Linear/near-linear response

This is really a particular case of the response feature analysis, with the overall trend or increase being the feature of interest. This is sometimes referred to as `last minus first', where the initial value for each unit is subtracted from the first, and the results analysed with a one-way ANOVA. This is a simplified form of orthogonal contrast. Other orthogonal contrasts can be used when there is no prior feature of interest. Alternatively, regression with respect to time can be used to estimate the trend, which can be used in ANOVA. This makes efficient use of all the data recorded.

8 The on-line help system

8.1 On-line help

If you need any further help on the topics covered in this course, or need some tips on how to perform a particular operation in Excel this is where you should begin.

The on-line help system provides step-by-step instructions and reference information for all tasks and features including functions, commands, toolbar buttons and Visual Basic.

Office Assistant

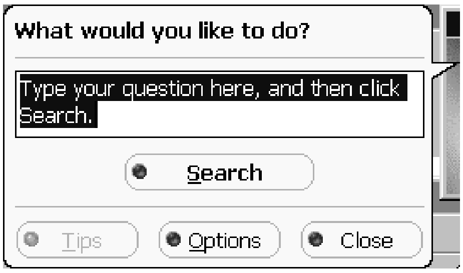
Help!

You can open help either using the **Help** menu on the standard menubar or by double clicking the **Office Assistant** button on the toolbar. The latter will display the **Search** dialog box (and a paperclip face or whatever **Office Assistant** you have chosen in Options).

Help > Contents
This gives you a hypertext menu, broken down into sections. To view a topic simply click on the title. A list of related topics will then be displayed.

Help > Index
This activates an alphabetical index of tasks and features.

Search
When you click on Office Assistant the following is displayed:

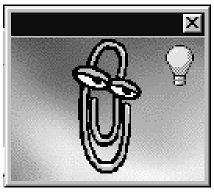


What would you like to do?

Type your question here, and then click Search.

Select the Search option to search Help for the exact information you need. For example typing Data Analysis and selecting **Search** will activate a list of all topics associated with data analysis. To view a particular topic simply click on the icon next to it.

TipWizard
If there is a quicker or more efficient way of performing the action you've just performed the TipWizard will inform you of it. Whenever there is a new tip the light bulb will light up.



Click on the Office Assistant and then click on **Tips** to display the TipWizard.
For previous tips click on the **back** button.

Help button
This provides detailed instructions and reference information if you are not sure what something does. Click on the **What's this?** button from the help menu on the standard toolbar and then choose a command or click a screen element to get detailed information about it.

Status bar
Positioned at the bottom of the screen, this gives a description of the currently selected command.

8.2 Reference Manuals

- ♦ *Microsoft Excel User's guide*
- ♦ *Microsoft Excel Visual Basic User's guide*
- ♦ *Microsoft Excel Visual Basic Reference*

The reference manuals and on-line help that accompany the package are designed to work together to help you get your job completed quickly and efficiently. Most of the help you will ever need is available on the on-line help system. For many of the analysis tools this includes the mathematics behind the workings. The reference manuals are not quite as thorough in this respect but do give useful examples of many of the commands and features available in the package. Easy to read and not particularly technical, they are a good way to see what's available, and browse. All the manuals have a comprehensive index at the back.

There are also several web pages on the Internet which give information on statistics with Excel (although most deal with earlier versions of Excel) :

- FAQ's on Excel and Statistics: R Gerard, City Univ., London
  http://www.city.ac.uk/~sc397/excelfaq/faq.html

- Introductory Course: M Pelosi & T Sandifer
  http://kraken.mvnet.wnec.edu/~jletkows/DOSTATS/update.html

- Introduction to Excel in Statistics: J Currall, Univ. of Glasgow
  http://www.stats.qla.ac.uk/cti/activities/reviews/96_11/excel/intro.html

- Teaching Statistics with Excel: N.Hunt, Univ. of Coventry
  http://www.stats.qla.ac.uk/cti/activities/reviews/96_05/excel.html

- Probability & Statistics: L Wolf, Lancaster Day School
  http://www.lcds.pvt.k12.pa.us/acaddept/math/probstat.htm

9 Summary of commands

Commands are given in alphabetical order, with a description of their function, pages of importance and the pathway to them.

- **Add-Ins** - displays a list of installed system add-ins. Page 5.
  Tools > Add-Ins...

- **ANOVA** - performs an analysis of variance with or without replication, testing whether a variable has the same mean in different groups. Page 39.

Tools > Data Analysis... > Anova: Single-Factor
> Anova: Two-Factor with Replication
> Anova: Two-Factor without Replication

- **AutoFilter** - to quickly find and work with a subset of your data without moving or sorting it. Page 31,48.

Data > Filter > Autofilter...

- **Boxplot** - draws a diagram showing the median, upper and lower quartiles, and upper limit (= max {upper range, Q3 + 1.5*interquartile range}) and lower limit (= min {lower range, Q1 + 1.5*interquartile range}). Page 14.

- **Chi-square test**- performs a test, and gives the value and the degrees of freedom. Page 51

fx **Insert > Function...> Statistical > CHITEST**

- **Chart** - creates a chart from data in a worksheet. This can either be embedded in a worksheet or displayed on a separate chart page. Page 12-17,21-23,29-31,52-53

📊 **Insert > Chart**

- **Confidence Interval** - a range of values for an estimate. It is the range of values that would not be rejected by a test based on the data. Page 35-36,44.

fx **Insert > Function ...> Statistical > Confidence**
OR
Tools > Data Analysis... > Descriptive Statistics

- **Correlation** - measures the strength of the linear relationship between two variables. Page 23.

Tools > Data Analysis... > Correlation

- **Cumulative distribution function** - the cdf for a value x is the probability that (for a specified distribution) a variable has a value less than or equal to x. CDF (x) = Prob(X<=x). Used in these notes to give p-values. Page 42, 52.

  fx **Insert > Function... > Statistical >** ***Cumulative distribution function*** (eg. TDIST)

- **Descriptive Statistics** - gives a table of summary statistics for the columns indicated, including the mean, median, standard deviation, and standard error of the mean. The option Confidence Interval also calculates a confidence interval for the mean. Page 9-11

  Tools > Data Analysis... > Descriptive Statistics...

- **Error bars -** Indicate the degree of uncertainty (the plus or minus range) for the data.

  Format > Selected Data Series > Error bars

- **Filter -** used to filter columns of data satisfying a specified criterion. Page 31,48-49

Data > Filter > Autofilter
> Advanced filter

- **Fitted values** - the values the data would have if a regression model fitted the data exactly. In Excel these are called the predicted values and are given by checking the residual option in the dialog box. Page 44.

  Tools > Data Analysis... > Regression... > ***Residuals***

- **Cell format -** enables borders to be added to cells, and the font and number format customised. Page 7.

  Format > Cells...

- **Chart format** - Once the chart has been activated the chart type, format and data series can be customised. Page 52-53.

  Format > Chart type...
  Format > Source data...
  Format > Chart options...

- **FunctionWizard** - a series of dialog boxes which guide you through the process of selecting a built-in Excel function or customised function and inserting it into a worksheet formula. Page 5-6,14,18,35,42,52.

  fx

- **Histogram** - draws a diagram, showing the frequency of observations in a number of intervals (bins). Page 12-14

  Tools > Data Analysis... > Histogram...

- **Import ascii data** - reads data from an ascii file. Page 2.
  File > Open > *...*

- **Percentile** - Calculates the percentage rank of data in a column. See Rank. Page 11.

  Tools > Data Analysis... > Rank and percentile....
  fx **Insert > Function > Statistical > PERCENTRANK**

- **Trendline** - shows the trend or direction of data in a series. Select the series on the chart to add a trendline to before calling the trendline procedure. Useful for adding a fitted line to a scatter plot. Page 45.

  Chart > Add Trendline...

- **Normal scores** - calculates what the relative values of observations would be if the data were Normally distributed. Page 18.

  fx **Insert > Function > Statistical > NORMINV...**

- **One-way ANOVA** - tests whether a variable has the same mean in two or more groups. Does not analyse blocking structures. Page 39-40.

  Tools > Data Analysis... > ANOVA: Single factor...

- **Open** - opens a workbook, previously saved as a workbook, or a file from another application. Page 2.

File > Open

- **Rank** - numbers the data in a column of length n in ascending order. i.e. smallest = 1,largest = n and so on. Page 11.

Tools > Data Analysis... > Rank and percentile....
fx **Insert > Function > Statistical > RANK**

- **Regression** - models a linear relationship between the dependant variable and one or more independent variable(s). Page 43.

Tools > Data Analysis...> Regression

- **Residuals** - the difference between the data and their fitted values. Check the residual option in the Residual dialog box. Page 44.

Tools > Data Analysis... > Regression > *Residuals*

- **Save -** saves the active workbook in the previously specified format if the file has been saved before, or in the newly selected format if the file is being saved for the first time. In the case of text files, only the active worksheet will be saved. To save under a new name, or in a different format, use the ***Save As* option rather than *Save*.** Page 4.
  File > Save
  File > Save As

- **Scatterplot** - Plots two variables against each other, categorised by a third if specified. Page 21-22.

Insert > Chart > XY(Scatter)

- **Autofill** - fills a column of data with a pattern of numbers (e.g. 1,2, 3, 4). The first few values in the series must be specified and the fill range including these values highlighted. Page 16, 51.

Edit > Fill > Series

- **Pivot table** - an interactive worksheet table that summarises large amounts of data using the format and calculation method you choose. Page 24-28.

Data > Pivot table...

- **Sort** - Sorts columns of data according to a predetermined order. (Default: Ascending or descending order). Page 47.

Data > Sort

- **Tally chart** - a table of counts and percentages of individual values. The columns specified must contain categorical data. Page 27.

Data > Pivot table...

- **TextWizard** - A series of dialog boxes which allows you to specify how you want the data of a text file to be distributed across columns. Page 2.

Data > Text to columns...

- **t-test** - performs a t-test (see Definitions of Statistics). Page 35.

Tools > Data Analysis > t-test: Two sample assuming equal variance.
> t-test: Two sample assuming unequal variance

- **Two-way ANOVA** - tests the effect of two classifying factors (e.g. treatments and blocks) on a variable. Page 40-42.

  Tools > Data Analysis.> Anova: Two-factor without Replication

10 Built-in Statistical functions

The following is a complete list of the built-in Excel statistical functions. They can be accessed from the **Function Wizard** by using the **Insert > Function** command.

AVEDEV	Returns the average of the absolute deviations of data points from their mean
AVERAGE	Returns the average of its arguments
AVERAGEA	Returns the average of its arguments. Evaluates text and FALSE as 0 and TRUE as 1.
BETADIST	Returns the cumulative beta probability density function
BETAINV	Returns the inverse of the cumulative beta probability density function
BINOMDIST	Returns the individual term binomial distribution probability
CHIDIST	Returns the one-tailed probability of the chi-squared distribution
CHIINV	Returns the inverse of the chi-squared distribution
CHITEST	Returns the test for independence
CONFIDENCE	Returns a confidence interval for a population
CORREL	Returns the correlation coefficient between two data sets
COUNT	Counts how many numbers are in the list of arguments
COUNTA	Counts the number of non-empty cells and the values in list of arguments
COUNTBLANK	Counts empty cells in a specified range of cells.
COUNTIF	Counts the number of cells within a range that meet the given condition
COVAR	Returns covariance, the average of the products of paired deviations
CRITBINOM	Returns the smallest value for which the cumulative binomial distribution is less than or equal to criterion value
DEVSQ	Returns sum of squares of deviations of data points from their sample mean.
EXPONDIST	Returns the exponential distribution
FDIST	Returns the F probability distribution
FINV	Returns the inverse of the F probability distribution
FISHER	Returns the Fisher transformation
FISHERINV	Returns the inverse of the Fisher transformation
FORECAST	Calculates/predicts a future value using existing values
FREQUENCY	Returns a frequency distribution as a vertical array
FTEST	Returns the result of an F-test
GAMMADIST	Returns the gamma distribution
GAMMAINV	Returns the inverse of the gamma cumulative distribution
GAMMALN	Returns the natural logarithm of the gamma function, G(x)
GEOMEAN	Returns the geometric mean
GROWTH	Returns values along an exponential trend
HARMEAN	Returns the harmonic mean
HYPGEOMDIST	Returns the hypergeometric distribution
INTERCEPT	Returns the intercept of the linear regression line
KURT	Returns the kurtosis of a data set
LARGE	Returns the k-th largest value in a data set
LINEST	Returns the parameters of a linear trend
LOGEST	Returns the parameters of an exponential trend
LOGINV	Returns the inverse of the lognormal distribution
LOGNORMDIST	Returns the cumulative lognormal distribution
MAX	Returns the maximum value in a list of arguments. Ignores text.
MAXA	Returns the maximum value in a list of arguments. Does not ignore logical values and text.
MEDIAN	Returns the median of the given numbers
MIN	Returns the minimum value in a list of arguments

MINA	Returns the minimum value in a list but does not ignore logical values and text
MODE	Returns the most common value in a data set
NEGBINOMDIST	Returns the negative binomial distribution
NORMDIST	Returns the normal cumulative distribution
NORMINV	Returns the inverse of the normal cumulative distribution
NORMSDIST	Returns the standard normal cumulative distribution
NORMSINV	Returns the inverse of the standard normal cumulative distribution
PEARSON	Returns the Pearson product moment correlation coefficient
PERCENTILE	Returns the k-th percentile of values in a range
PERCENTRANK	Returns the percentage rank of a value in a data set
PERMUT	Returns the number of permutations for a given number of objects
POISSON	Returns the Poisson distribution
PROB	Returns the probability that values in a range are between two limits
QUARTILE	Returns the quartile of a data set
RAND	Returns a random number between 0 and 1
RANK	Returns the rank of a number in a list of numbers
RSQ	Returns the square of the Pearson product moment correlation coefficient through the given data points.
SKEW	Returns the skewness of a distribution
SLOPE	Returns the slope of the linear regression line
SMALL	Returns the k-th smallest value in a data set
STANDARDIZE	Returns a normalized value
STDEV	Estimates standard deviation based on a sample
STDEVA	Estimates standard deviation based on a sample but includes text and logical values
STDEVP	Calculates standard deviation based on the entire population
STDEVPA	Calculates standard deviation based on the whole population but includes text and logical values
STEYX	Returns the standard error of the predicted y-value for each x in the regression
TDIST	Returns the Student's t-distribution
TINV	Returns the inverse of the Student's t-distribution
TREND	Returns values along a linear trend
TRIMMEAN	Returns the mean of the interior of a data set
TTEST	Returns the probability associated with a Student's t-Test
VAR	Estimates variance based on a sample
VARA	Estimates variance based on a sample including logical values and text
VARP	Calculates variance based on the entire population
VARPA	Calculates variance based on the entire population (including text)
WEIBULL	Returns the Weibull distribution
ZTEST	Returns the two-tailed P-value of a z-test

11 Definitions of Statistics

ANOVA

Analysis of Variance: calculates how much of the variation is due to the various sources.

One-way

For the example in 4.3.1, with t = no. of treatments (4), and n = total no. of units in the experiment (24).

Total Sum of Squares (SS) = $\sum x^2 - \frac{(\sum x)^2}{n}$, where the x's are all the sample values and n is the number of values.

Factor (or treatment) SS = $\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4} - \frac{(\sum x)^2}{n}$

where x_i are all the values in group i, and n_i is the number of values in group i.

Error SS = total SS - factor SS.

The degrees of freedom are:

Total D.f. = n - 1

Factor D.f. = t - 1

Error D.f. = n - t

The Mean Square (MS) values are the SS divided by the degrees of freedom from the same source and the F-statistic is calculated by : F = $\frac{FactorMS}{ErrorMS}$, and is distributed as

$F_{t-1,\, n-t}$

Two-way

The example in 4.3.2 is an example of a blocked design. t = no. of treatments (7), b = no. of blocks (3) and n = no. of units in the experiment (21).

Total SS is calculated as for the one-way Total SS.

Block SS = $\frac{B_1^2 + B_2^2 + B_3^2}{t} - \frac{(\sum x)^2}{n}$, where B_i = sum of values in block i

Treatment SS = $\frac{T_1^2 + T_2^2 + T_3^2 + T_4^2 + T_5^2 + T_6^2 + T_7^2}{t} - \frac{(\sum x)^2}{n}$, where T_i = sum of values in treatment i.

Error SS = Total SS - (Block SS + Treatment SS)

The degrees of freedom for a blocked design are:

Total D.f. = n - 1 (as above)

Block D.f. = b - 1

Treatment D.f. = t - 1

Error D.f. = (b - 1)(t - 1)

MS and the individual F statistics are calculated as above.

Chi-Squared Tests

The chi-squared $\left(\chi^2\right)$ statistic tests for independence between variables. The total proportion that occurs of a category is expected to be reflected within the subdivisions. So, the expected value within a cell is the overall proportion of the category multiplied by the total number possible. In other words, (row total x column total) / grand total.

To illustrate:

n = n1 + n2 + n3 + n4

	Yes	No	
1	n1	n2	n1+n2
2	n3	n4	n3+n4
	n1+n3	n2+n4	n

For example, to calculate the expected value of ‘Yes:1', the total proportion of ‘1' is (n1+n2) /n, and the total available ‘Yes' is n1 + n3. Hence the expected value for ‘Yes:1' is:

(n1 + n2) * (n1 + n3) /n, and so on. The χ^2 statistic is calculated by: $\sum \frac{(observed - \exp ected)^2}{\exp ected}$

and compared to a χ^2 distribution with (no. of rows - 1)*(no. of columns - 1) degrees of freedom.

Confidence Interval

Given a sample mean of $\bar{x}$, the population mean μ lies in the interval

$$\mu = \bar{x} \pm t_{df} \times se$$

where t_{df} is the appropriate t value for the degrees of freedom (n_1+n_2-2) for standard t-test,

Error DF for an ANOVA t-test).

Correlation

The correlation between two variables, x and y, is calculated by :

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

where r denotes the correlation coefficient

Hypothesis Test

Any of a number of statistical tests to assess the extent to which observed data are consistent with a specified hypothesis.

Mean

This is one of the simplest statistics:

$$\bar{x} = \frac{\sum x}{n}$$

Median

For a sample of n values, the median is given by:

i) if n is odd, the median is the $\left(\frac{n+1}{2}\right)^{th}$ of the ordered values.

ii) if n is even, the median is the mean of the $\left(\frac{n}{2}\right)^{th}$ and the $\left(\frac{n}{2}+1\right)^{th}$ ordered values.

Quartiles

Quartiles mark the data points that have 25% of the data below (lower quartile) the mark and 25% of the data above the mark (upper quartile). The difference between the upper and lower quartiles is called the inter-quartile range.

Regression

For the equation of a linear regression line y = a + bx, a and b are calculated by:

$$\hat{b} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

and

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Significance level

A value of the p-value at which you would reject the hypothesis under consideration. Commonly used significance levels are 5%, 1% and 0.1%.

Standard deviation

The standard deviation is the square root of the variance, σ^2..

$$\sigma^2 = \frac{1}{n-1}\sum(x-\bar{x})^2$$

Standard error (of the difference)

Used as the standard error when calculating t-tests based on an ANOVA.

$$sed = \sqrt{MSError \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Standard error (of the mean)

The standard error indicates the variability in sample means. It is the standard deviation divided by the square root of the sample size, n.

$$sem = \frac{\sigma}{\sqrt{n}}$$

t-tests

1) Standard t-test (not ANOVA)

one sample t-test

Testing a sample mean against an expected mean, μ .

$$t = \left|\frac{\bar{x}-\mu}{sem}\right|$$

t should be compared with the t-distribution with n - 1 degrees of freedom.

two sample t-test

Testing whether there is a difference between two sample means.

$$t = \left|\frac{\bar{x}-\bar{y}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right|$$

where S_p is the pooled standard deviation, and given by:

$$S_p^2 = \frac{\sum(x-\bar{x})^2 + \sum(y-\bar{y})^2}{n_1 + n_2 - 2}$$

t should be compared with the t-distribution with (n_1+ n_2 - 2) degrees of freedom.

2) From an ANOVA

$$t = \frac{\bar{x}-\bar{y}}{sed}$$

t should be compared with the t-distribution with the Error degrees of freedom.

Appendix 2 : Data sets for use in exercises

POTATO

The data consists of the weights and three principal dimensions of 200 potato tubers. The length, measured in mm, is the longest dimension approximately at right angles to length. Depth is measured at right angles to both length and breadth. Values for the first three and final tubers are:

Weight(g)	**Length(mm)**	**Breadth(mm)**	**Depth(mm)**
158.3	78.0	66.1	51.8
83.7	61.3	55.1	41.1
227.6	108.6	69.6	52.1
-	-	-	-
79.7	64.4	51.0	39.2

PATCH

Durations of time were recorded for which a singe ion channel in a patch clamp was open or closed. In total there were 185 openings :

Open time	***Closed time***
26	4
186	10361
256	4912
-	-
14	2501

WEATHER

Hourly weather data in Edinburgh are given for the first seven days of May, 1982. There are 168 observations on each variable, except that the wind direction is missing when the wind speed is zero :

Day	***Hour***	***Cloud (octal)***	***Rad (W/m*m)***	***Temp (Celcius)***	***Relative humidity***	***Rain (m/s)***	***Wind spd (0 or 1)***	***Wind drn (1:N 2:E .)***
1	*0*	*1*	*0*	*3.7*	*80*	*0*	*11*	*4*
1	*1*	*6*	*0*	*3.5*	*78*	*0*	*11*	*4*
1	*2*	*3*	*0*	*2.6*	*82*	*0*	*8*	*4*
-	-	-	-	-	-	-	-	-
7	*23*	*4*	*0*	*2.8*	*95*	*0*	*0*	*

CONE

Data are the resistances (kg) of soil to the penetration of a cone, measured at 15 depths spaced at 3cm intervals. In all, 10 penetrations were made in each of the four plots of ground, which had been treated in different ways. There are 40 rows of data, the first 10 rows are for treatment 1, and so on :

Depth 3 6 9 12 15 18 21 24 17 30 33 36 39 42 45

4	7	8	6	12	13	14	15	15	14	12	30	22	25	24
2	6	6	7	8	6	6	10	10	11	12	23	25	30	29
5	8	6	10	11	8	9	8	9	10	26	24	27	27	29
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	3	8	8	7	7	7	7	6	13	20	21	25	25	22

GRASS

Grass, which had been cut by one of four different machines, was dried in thin layers in a laboratory. The data are two summary statistics obtained from each of 24 grass samples: the first describes the shape of the drying curve, and the second represents the time taken to reach a certain moisture content. The first six observations are based on samples from a grass-cutting machine, and the subsequent 18 are in batches of six from each of three machines which condition the grass in order to accelerate the drying process:

Shape parameter	***Drying time parameter***
0.5034	6.5601
0.5797	0.5487
0.5726	8.6279
-	-
0.6199	5.7259

SILAGE

Grass in swaths was dried in 21 plots in a field. Plots were blocked in groups of size seven, one of which had received each of the seven treatments. The data consist of two summary statistics derived from each plot : a measure of the initial moisture content of the grass, and of the rate of drying over the subsequent three days, together with indicator variables:

Init. moisture	***Drying***	***Treatment***	***Block no.***	***Index no.***
1.586	2.08	1	1	1
1.519	2.19	1	2	2
1.446	2.25	1	3	3
-	-	-	-	-
1.491	2.34	7	3	21

HAY

Grass in swaths was dried in 18 plots in a field. Plots were blocked in groups of size six, one of which had received each of the six treatments. Drying rates and indicator variables

are given :

Drying rate	**Treatment number**	**Block number**	**Index number**
163	1	1	1
174	1	2	2
157	1	3	3
-	-	-	-
95	6	3	18

CARCASS

Each of 37 sheep carcasses was visually graded for fatness, and backfat depth was measured by ultrasound. They were then dissected, the actual fat content was determined. In the data, two grades and one depth are missing :

fatness grade	% fat	Backfat depth
4	35.46	4.0
3	34.54	5.0
2	29.18	2.0
-	-	-
*	30.81	4.0

TOBACCO

Each of 8 tobacco plants had their second leaf tested for effects of a virus. Half of each leaf had the first preparation of virus extract rubbed on and the other half had the second preparation. The number of leisons appearing as small dark rings were counted.

Test1	Test2
31	18
20	17
-	-
7	6

BOVINE

The data concern four types of bovine disease (BVDV, IBR, P13, RSV) recorded over two years, 1991 and 1992.

BVDV	IBR	PI3	RSV
597	259	72	91
649	231	106	96