Guide to using Microsoft Excel®

Contents

Introduction

Microsoft Excel® is a widely available, relatively inexpensive, computer package. It can be run on most computers and is available from a wide variety of stores and suppliers. Microsoft Excel® is **not** specifically designed to be a statistical package; it is, in fact, a spreadsheet. Spreadsheets are computer packages that were originally designed to allow data to be easily manipulated and handled. Microsoft Excel® is very versatile and the package comes with basic statistical functions. It can also be used to produce tables and graphs. Microsoft Excel's abilities to perform statistical functions can be upgraded. Microsoft Excel® comes as part of Microsoft's Office suite of packages and thus can be found on many of the world's computers. The statistical utilities the package comes with are limited, and do not cover all the statistics described in the book. In particular very few non-parametric statistical tests are available via Excel®. Some professional statisticians have criticised Excel® for producing inaccurate statistics. If you are thinking of analysing a large more complex set of data then a package like SPSS may be more suitable. Alternatively, you can by software that can extend Excel's capabilities.

Entering data

Advice is given in the book on coding data. In general use a column for each variable and a row for each case. If recording the heights of a number of different subjects variables you could include things like, age, sex, country of birth and of course height. Each variable would have its own column and each case its own row. In

Excel® data is entered on to what is called a worksheet. If you click on the Excel® icon it will normally open with a new worksheet as in Figure 1.

Notice how the columns are marked with letters and the rows with numbers. Each box on the spreadsheet is known as a cell. Enter the data by typing into the cells. The cell where data is being entered is termed the active cell. The outline of the active cell is highlighted in black.

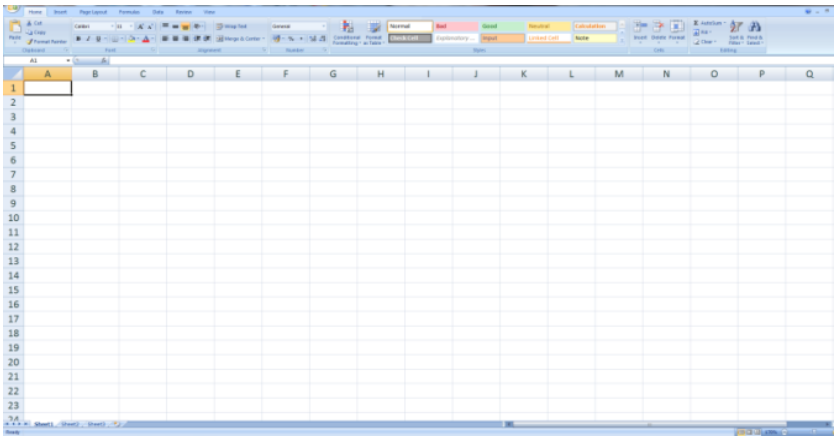


Figure 1: An Excel® workbook.

Co-ordinates

You can identify an individual cell from co-ordinates formed by the letter that denotes its columns and the number that denotes its row. Thus A1 is the first cell of the spreadsheet. The co-ordinates that locate the cells are a very important aspect of spreadsheets.

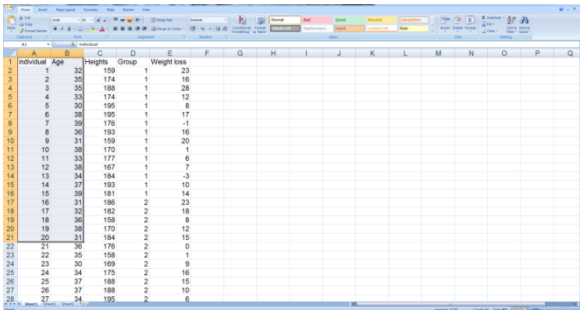


Figure 2: Excel® spreadsheet indicating an array of

These co-ordinates are used to tell the programme which numbers to perform calculations upon. If we gave Excel® the command A1*6 it would indicate that we want to multiply whatever number was in the cell A1 by 6. Similarly, if we gave the command A1*B2 it would indicate that we wanted Excel® to multiply the number in

cell A1 by that in cell B2. We can also ask Excel® to perform functions in relation to groups of cells (numbers). Groups of cells are identified using pairs of co-ordinates; for example the formula A1:B32 would signify all those cells from A1 to B34 inclusive. This group is enclosed in the red box on the picture below (Figure 2) which shows some of the data derived from the experiment involving symphadiol. Notice how the variables have been coded, and that the experimental group is treated as a variable. That is, it is given its own column and each case is ascribed to an experimental group. A number has also been given to identify each individual that took part in the study.

Descriptive Statistics with Excel® (Chapter 6)

Excel® can be used to calculate descriptive statistics, once you have entered the data on the spreadsheet you can access many of the basic functions of the programme by clicking on the function symbol: f_x. This is a small button on the last of the tool bars at the top of the worksheet.

This will bring up a menu that will look like figure 3 below; if you then select **Statistical** from the second look-up menu, you will then see a list of Excel's statistical functions.

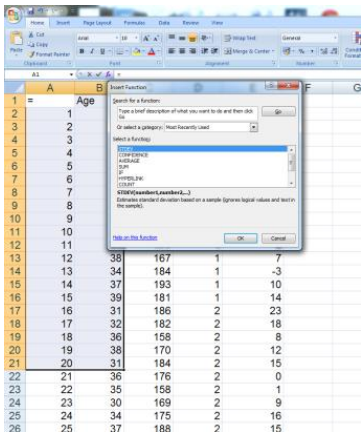


Figure 3: The dialogue box of the function key

If you click-on **AVERAGE**; a new dialogue box will appear asking you to identify the range of numbers for which you want to obtain the average. You must enter the co-ordinates in the first dialogue box. The answer to this calculation (the statistic) is given in the active cell. In figure 4, the average for the cells C2:C63 is being requested.

For most of the descriptive statistical functions within Excel, the process is similar. When working with columns of data, we recommend using an active cell beneath the column of interest and requesting the descriptive required. You will need to select

additional active cells as you call up more statistics. Commonly required descriptive statistics and their function name (Figure 3) in Excel® are given below (Table 1).

Statistic	**Function Name**
Arithmetic Mean	AVERAGE
Standard Deviation	STDEV
Variance	VAR
Sample Size	COUNT
Mode	MODE
Median	MEDIAN
Maximum	MAX
Minimum	MIN
Quartiles	QUARTILE: Need to specify which quartile in a separate cell
Range	No function need to calculate just subtract Minimum from Max (see Chapter 6)

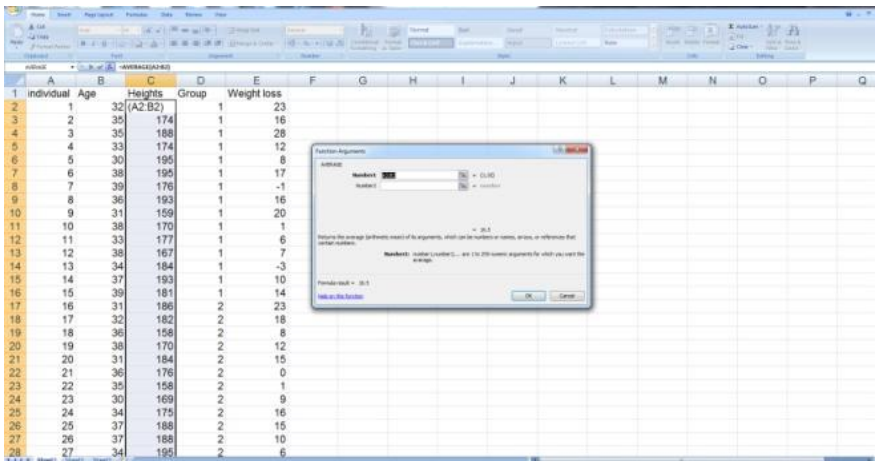


Figure 4: Calculating the average (Mean)

Excel® Equations

You may have noticed that when you use the statistics functions that Excel® writes information into the formula bar. You can find the formula bar just below the toolbars. Simultaneously to writing into the formula bar the same information will be written to the active cell. What Excel® is doing is writing a formula that reflects the statistical procedures that you are asking it to perform. Indeed it is possible not to use the dialogue boxes described above and write your own formulas. For simple calculations you may want to do this as it is sometimes a faster option.

If, for example, you want to add lots of numbers together you can use Excel's Sum facility. To use this facility first highlight the column of numbers that you wish to add together, and press the button marked ∑. You will see that the numbers in the column will be added together and the sum placed in the cell at the bottom of the column that was highlighted.

If you activate the cell that contains the Sum you will see in the formula bar that an equation is written. The form of that equation will be **SUM (Co-ordinates of first cell highlighted: Co-ordinates of the last cell co-ordinated)**. The formula bar informs the operator what procedures are being applied to the active cell. Thus = SUM (A1: A5) informs us that the number in the active cell is the summation of those values between and including cells A1 to A5. Cells can also be summed across rows using the same procedure that has been described for columns above. You can tell that a cell contains a formula because the content of the cell will start with an = sign.
You can write formulas of your own and apply them to new cells using standard computer mathematical notation. Some of the commonly used symbols are listed below.

/ = divide
* = multiplied by
- = minus
+ = plus
** = square of
Log10 (Co-ordinate) = Logarithm to the base 10 of the number at the co-ordinates in brackets.

To write a formula, first click on the cell where you want the answer to your calculation to appear. If, for example, you want to calculate an average (mean), you can simply write = AVERAGE(A2:A33). The mean for the numbers between the cells A2 and A33 will be calculated. If you had written =A2/A33 then Excel® would divide the number in cell A2 by that in cell A33. Note, that even when you use the dialogue boxes Excel® still actually works by using formulas. The dialogue boxes simply make writing the formulas easier (they do it for you).

You can repeat the formula for other columns (or rows) of data. When you want to repeat a calculation procedure across columns or down rows you can use Microsoft Excel's **Copy and Paste** facility.

If you want to calculate the mean for when you have highlighted the correct cells type Ctrl-c on the key board (i.e. press the keyboard key marked Ctrl and at the same time press the key marked c, Use the same action as when using the shift key to

capitalise). The cell will become outlined with a dashed line. Now highlight the cells where you want the formula to be placed, (in this example in cells B12 and B13) and type Ctrl-v. If all has gone well, you should now see that the values for B12 (i.e. the sum of B2 to B11) and B13 (i.e. the mean of cells B2 to B11) have been calculated and will appear in cells B12 and B13.

Inferential Statistics with Excel®

The basic Excel® package gives access to a limited range of inferential statistics, these are described below. If you need to calculate many different statistics with complex data sets you will either need to upgrade your computer package or use a package specifically designed to perform statistical analysis.

Standard Error and confidence limits

You can ask Excel® to calculate a confidence limit for any sample mean. You need to know the standard deviation of the sample, the sample size and the confidence level you are working to. First click on the function key on the menu tool bar; then select **Statistical** from the right hand box and then **CONFIDENCE** from the left hand box. A dialogue box will appear (Figure 5). Next, put your values for standard deviation, sample size and alpha in the appropriate boxes.

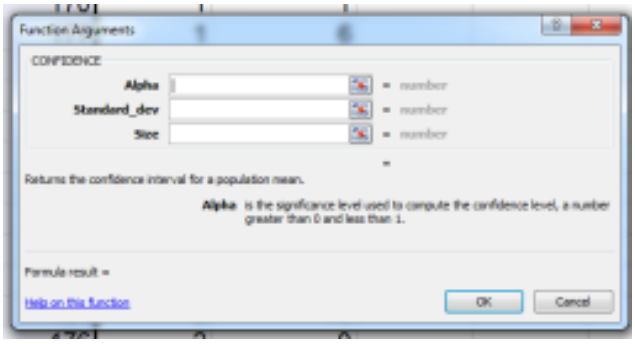


Figure 5: The dialogue box for calculating confidence limits

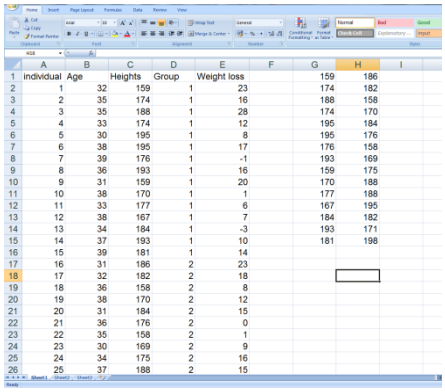
Note that alpha is the significance level so that if you want to calculate 95% confidence limits you will need to use an alpha or significance level of 0.05. The actual confidence limits are formed by the mean ± the confidence interval (see Chapter 11). Note that the formula is based on the normal distribution, for smaller data sets the t-distribution needs to be used.

There seems to be no direct way of calculating a standard error in Excel®, but if you obtain the standard deviation of your sample, the standard error can then be calculated using the formula given in chapter 11 of the book.

Z, T and F tests. (Chapter 12)

An F-test can be obtained from Excel® by using the function key and then selecting **FTEST**. Here some of the issues with regard to using a spreadsheet rather than a programme dedicated just to statistics will emerge. The dialogue boxes you see that ask for two arrays of numbers to be entered – array is another word for list. Each of the arrays is formed from one of the samples that you wish to compare, experiment and control for example. If, however, you have followed our advice and entered the data in columns, you will have just one array for each variable, with a separate column indicating which sample each case belongs to (see Figure 6). To overcome this problem, when using Excel®, you will need to copy and paste each sample into a separate column. This is why we do not recommend the basic version of Excel® or anything other than simple data sets.

Let us say we wanted to examine the data from the syphadiol experiment and decided to perform an F-test to compare the variance of the heights of the sample of participants in group 1 with those of group 2.



	A	B	C	D	E	F	G	H	I
1	individual	Age	Heights	Group	Weight loss		159	186	
2	1	32	159	1	23		174	182	
3	2	35	174	1	16		188	158	
4	3	35	188	1	28		174	170	
5	4	33	174	1	12		195	184	
6	5	30	195	1	8		195	176	
7	6	38	195	1	17		176	158	
8	7	39	176	1	-1		193	169	
9	8	36	193	1	16		159	175	
10	9	31	159	1	20		170	188	
11	10	38	170	1	1		177	188	
12	11	33	177	1	6		167	195	
13	12	38	167	1	7		184	182	
14	13	34	184	1	-3		193	171	
15	14	37	193	1	10		181	198	
16	15	39	181	1	14				
17	16	31	186	2	23				
18	17	32	182	2	18				
19	18	36	158	2	8				
20	19	38	170	2	12				
21	20	31	184	2	15				
22	21	36	176	2	0				
23	22	35	158	2	1				
24	23	30	169	2	9				
25	24	34	175	2	16				
26	25	37	188	2	15				

Figure 6: Excel® workbook showing the heights of the participants in treatment groups 1 and 2 copied into columns G and H.

We would first need to cut and paste the data into separate columns (see Figure 6). It is good practice to give these new columns headings, such that you do not forget which sample they belong to. In figure 6 we have cut out the data for group 1 and group 2 and placed them in to the columns G and H. Next, you call up the **Statistical** dialogue box in the normal way and select **FTEST**. You indicate in the

dialogue box the groups you want to compare; remember you need to give the location of each of the columns (arrays). For group 1 the array is G5:G19 and for Group 2 it is H5:H19 (Figure 7). If you remember that when describing the array the colon stands for “to” they are easier to understand. Thus G5:G19 simply tells the computer to use the numbers from G5 to G19.

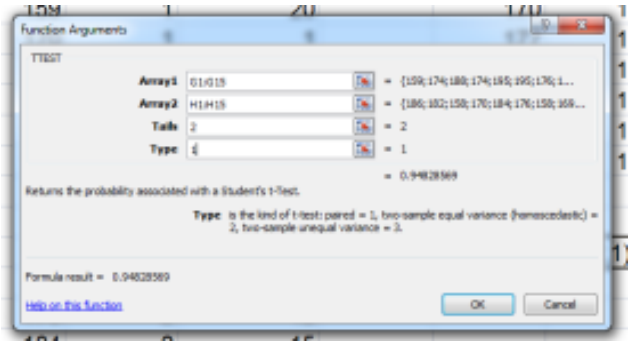


Figure 7: Using a dialogue box (F-Test) to indicate the location of arrays of data.

For the F-test the result will appear at the bottom of the dialogue box where you have just entered the location of the data to be analysed. When you click **OK** the result will also be entered into the active cell. The result you obtain will be the P value for the test. The value for F is not actually given. You cannot just present the F value you will need to give the statistic as well as the degrees of freedom.

The actual value can be calculated by selecting the command **FINV** from the Statistical menu. You will be asked to input the probability (you have just calculated this) and the degrees of freedom from the two samples you are comparing. The degrees of freedom from the smaller sample is entered first.

An alternative to the described route to obtain an F-test is to use the **Data Analysis** sub-set of Excel®. This can be obtained from the Data menu (far end of the tool bar). Unfortunately When the Excel® package is loaded on to computers, sometimes the data analysis omitted. This facility can be loaded at anytime by going to the **File** menu and then selecting **Options** (down the bottom). Next select **Analysis toolpak** click the check box and then **OK**.

We will look at the students-t test using this facility – you will find that using the **Data Analysis** option will give you more output than that from the function facility. We will analysis the same data set as used for the F-test.

Having loaded the Analysis toolpak select **Data Analysis** from the **Data** menu, then select an appropriate t-test, there are three on offer; given that the F-tests indicates

that the variances are not significantly different, you can select the t-test that assumes equal variance (Figure 8).

Note that you are now given several options – the first thing you need to think of is whether or not you will include labels. If you want to use labels (we suggest you do) click on the **Labels** check box. Excel® will assume that the first entry in the column of data is the name of the sample (we have used the labels Group 1 and Group 2). You can now enter the range of the cells for the two samples remembering to include the labels if you have selected this option.

The next question you are asked is, what the hypothesised difference between the means is, the default answer is 0, which is the value that would normally be used (i.e. the hypothesis of no difference-the null hypothesis.

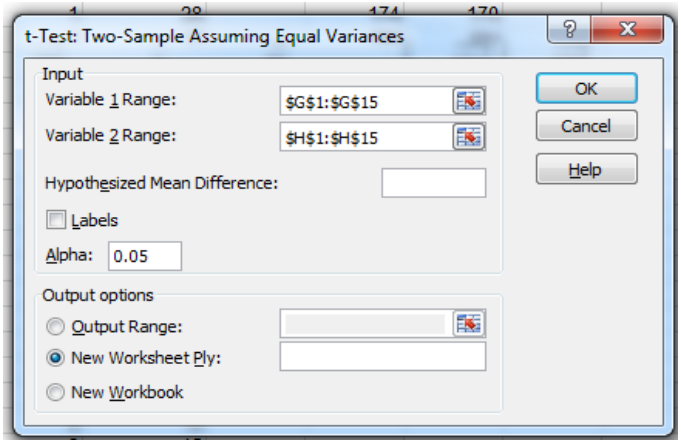
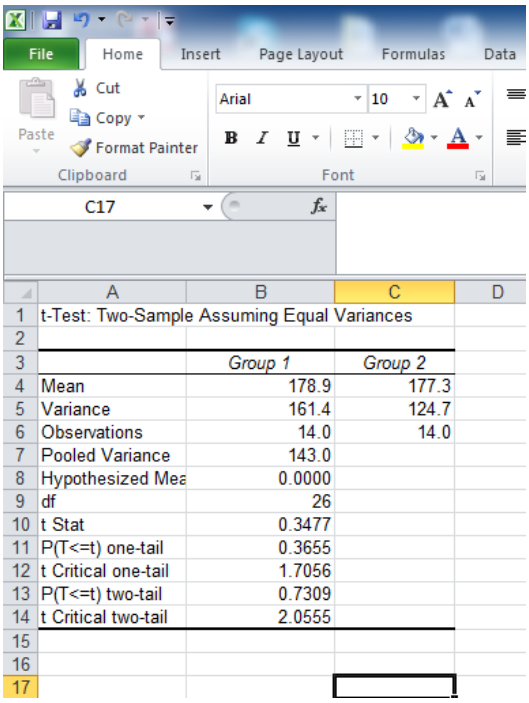


Figure 8: The t-test dialogue box

Finally you must decide where the output (i.e. the statistics generated by the test will be placed. We suggest you accept the new worksheet option; in this option the answers are placed on a new worksheet that can be accessed by clicking on the worksheet tabs (see Figure 6) at the bottom of the current worksheet.

The output will appear in a new worksheet that will automatically appear on top of the current sheet; the output may at first seem a little confusing (Figure 9); but a little careful analysis will dispel any initial panic. With the exception of the pooled variance the first 6 lines of information are statistics that describe the two samples that were compared. The pooled variance is a descriptive statistic of the combined samples (see Chapter 12). Moving down the first column of information you will see t Stat, this is the actual statistic; t for this comparison thus equals 0.07522. Moving down you are given the P value for the one tailed test, along with the critical value for the

one-tailed test, this is the value that, if you were performing a t-test by hand, you would look up from the t-distribution tables. It is the value that the t-statistic is compared against. In the last two rows the same information is given but for a two-tailed test.



	A	B	C	D
1	t-Test: Two-Sample Assuming Equal Variances			
2				
3		*Group 1*	*Group 2*	
4	Mean	178.9	177.3	
5	Variance	161.4	124.7	
6	Observations	14.0	14.0	
7	Pooled Variance	143.0		
8	Hypothesized Mea	0.0000		
9	df	26		
10	t Stat	0.3477		
11	P(T<=t) one-tail	0.3655		
12	t Critical one-tail	1.7056		
13	P(T<=t) two-tail	0.7309		
14	t Critical two-tail	2.0555		
15				
16				
17				

Figure 9: The output from the t-test

Paired-t-test

The paired-t-test follows exactly the same format as the other t-tests. You must, however remember that numbers in the same row as each other will be treated as a pair (a repeat measure) thus, you must set up your data accordingly.

Z-Test

The procedure for conducting a Z test is similar to that for the t-test accept that you will need to know the variance of each sample, these can be calculated using the function key and the VAR command.

ANOVA (Chapter 13).

Calculating an ANOVA using Excel® is similar to calculating a t-test. You must have loaded the Analysis toolpak and then click-on **Data Analysis** from the Data menu, then select an appropriate ANOVA. There are three tests on offer, single factor, two factor with replication and two factor without replication. Single factor is analogous to the one-way ANOVA and two factor with the two-way ANOVA. The Two-way ANOVA without replication is not covered in the textbook.

We will do a single factor analysis comparing the means of the heights of each of the four treatment groups from the symphadiol experiment. As with the other tests you will need to copy and paste each sample (treatment group) of heights into a separate column (Figure 10).

Next from the **Tools** menu select **Data Analysis**, and click on one factor ANOVA. You must now tell Excel® how the data are arranged (columns or rows). In this example we have arranged the data in columns. So columns is selected. You must then, as always, indicate were the data are located. This time you must enter the co-ordinates for the block of data being analysed. Enter the co-ordinates in the Input Range box (Figure 10).

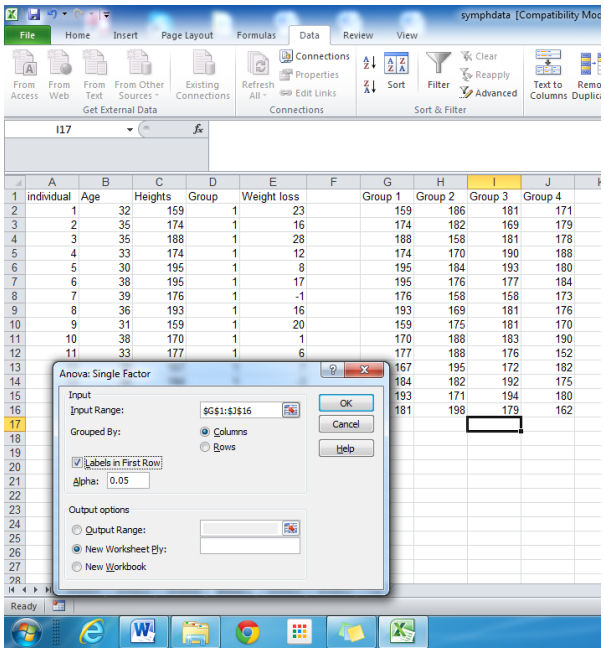


Figure 10: Arranging the data into separate columns and the ANOVA dialogue box.

In the dialogue box you are also asked to enter a value for alpha. This is equivalent to the significance level that you want to use for rejection of the null-hypothesis. It is the probability level that Excel® will use to produce the critical value for F. Finally you have the option of selecting where the output will be placed – we recommend selecting the default and allowing a new worksheet to be produced.

The output is given as two tables – the first provides descriptive statistics and the second the actual ANOVA table. The summary table gives descriptive statistics for each of the treatment groups.

	A	B	C	D	E	F	G	H
1	Anova: Single Factor							
2								
3	SUMMARY							
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*			
5	Group 1	15	2685	179	150.1429			
6	Group 2	15	2680	178.6667	144.381			
7	Group 3	15	2707	180.4667	93.8381			
8	Group 4	15	2640	176	94.85714			
9								
10								
11	ANOVA							
12	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
13	Between Groups	155.8667	3	51.95556	0.430079	0.732251	2.769433	
14	Within Groups	6765.067	56	120.8048				
15								
16	Total	6920.933	59					
17								

Figure 11: The output from the ANOVA test

The ANOVA table follows the format given in the textbook, with the exception that a value F crit is given. This is the value that if doing the calculations by hand you would look up in a book of statistical tables using 56 and 3 degrees of freedom.

Correlation and Regression (Chapter 18)

You can obtain a correlation coefficient using the f_x and selecting Statistical and then **CORREL** from the dialogue box. You will be asked for the co-ordinates when you click on **CORREL**; a new dialogue box will appear asking you to identify the two ranges of numbers for which you want to obtain the correlation co-efficient; enter the co-ordinates in the first dialogue box. The answer to this calculation (the statistic) is given in the active cell. Using this method you will need to look up the significance of the result, using the appropriate sample size, in a book of statistical tables. Note that the array of numbers defined in the uppermost box is treated as the x-axis value, and that in the lower box, the y-axis value. Note also that, as with the paired-t-test, numbers in the same row as each other will be treated as the pair of x and y values being correlated.

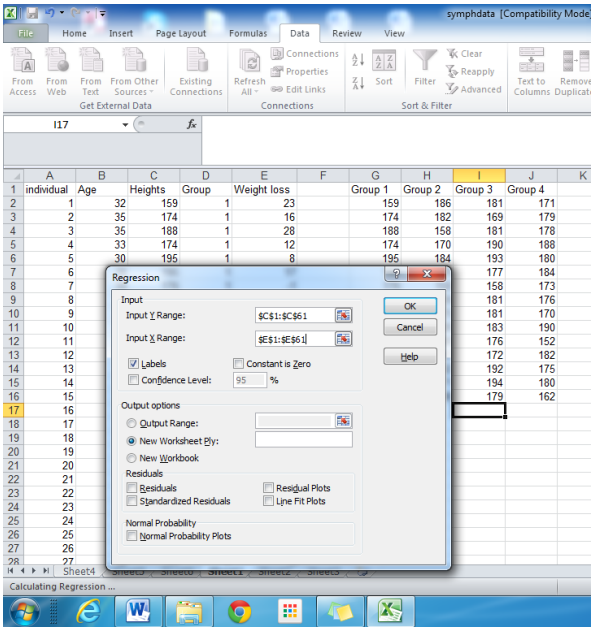


Figure 12: The regression analysis dialogue box.

You can also find correlation and regression by clicking-on **Data Analysis** from the **Data** menu. The output, however for the correlation achieved by using this method is the same as in that described above and we do not recommend this approach. For regression however, the **Data Analysis** command provides a good rapid method to obtain regression statistics (Figure 12). First select **regression** from the Statistical sub-set of the **Data Analysis** menu; you then need define the arrays of data that you wish to compare. Enter the array co-ordinates in to the dialogue box (Figure 12). You then have various options to select.

You can adjust the confidence limits used by clicking in the Confidence limits check box and then altering the value in the adjacent box. You can also select options that will help you analyse residuals. Residuals are beyond the scope of the textbook. Residuals can be used to provide insight into your data and to see if your data meet the assumption behind the regression analysis.

The output from this analysis comes in the form of three tables (Figure 13), one to provide correlation statistics, an ANOVA table and a table of statistics covering the regression equation.

The regression summary, gives the Pearson's correlation coefficient (given as Multiple R) and the coefficient of determination (given as R Square). You can ignore the Adjusted R square value as this should be used only when the analysis has used more than one y variable. The ANOVA table may appear slightly out-of-place, after all you had asked for a regression analysis to be performed. ANOVA and regression are very similar procedures and it is common for statistical packages to use an ANOVA to determine the significance of the correlation between variables. Thus, the value in this table given as the "Significance of F" is the P value for the regression coefficient.

In the final table, separate statistics for the slope term (weight) and for the intercept. The values for intercept and for the slope term are given under the heading coefficients. Thus the equation for this regression is:

$$y = 173.5 + \text{Weight}(0.37).$$

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.273103774					
5	R Square	0.074585671					
6	Adjusted R Square	0.058630252					
7	Standard Error	10.50839638					
8	Observations	60					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	516.2024595	516.2025	4.674629	0.034749179	
13	Residual	58	6404.730874	110.4264			
14	Total	59	6920.933333				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	173.5288394	2.682922922	64.67903	9.18E-56	168.1583897	178.8992891
18	Weight loss	0.369790195	0.171033744	2.162089	0.034749	0.027429216	0.712151174

Figure 13: The output from the regression analysis

The table also gives the standard errors for each of these terms as well as t-test to determine their significance from the value of 0. Note how the significance value for the t statistic is the same as that for the F statistic from the ANOVA table. Finally, the table also provides lower and upper 95% confidence limits.

Chi-Square tests (Chapter 16)

There is no direct method for obtaining a chi-square test using Excel®. The mathematics of the chi-square test however is relatively straightforward and therefore Excel's calculation function can be used to perform such tests. The information you need to supply Excel® is a list the observed values and a list of expected values. The textbook covers how the expected values should be determined, the method of their determination will vary depending on the type of chi-square test.

The observed and the expected frequencies values can be entered as separate columns or as separate contingency tables (see Chapter 16). Once you have done this click f_x and select **Statistical** and choose **CHITEST**, enter the arrays for the observed frequencies (actual_range box) and then the expected frequencies. When you then click OK the result (the P value) will be placed in the active cell. The significance level used is P=0.05). It is not sufficient to give just a P value for a statistical test you must also provide the degrees of freedom used and the value of

the actual statistic. To calculate this select **CHINV**. Input the P value and then the degrees of freedom. Again the result will be placed in the active cell.

Additional software and further reading

You can buy additional add-on software that will enable Excel® to perform more statistical functions. Two packages that are available are WinStat and Statisxl their web sites are http://www.winstat.com/index.htm and http://www.statistixl.com respectively. These packages have both been reviewed favourably.

Further information on using Excel® can be found in various texts, for example:

Harvey, G. *Excel 2013 for Dum*mies, London, John Wiley & Sons Inc.