Disciplined Use of Spreadsheet Packages for Data Entry

2nd Edition revised for Excel 2007

February 2009

University of Reading
Statistical Services Centre

Biometrics Advisory and Support Service to DFID

Contents

1. Introduction 3

2. An example 4

3. Facilitating the data entry process 6

3.1 Introduction 6

3.2 Freezing or splitting panes 8

3.3 Drop-down lists 8

3.4 Data validation 9

3.5 Adding comments to cells 11

3.6 Formatting columns 11

3.7 Data auditing 12

4. The metadata 14

4.1 Body data 15

4.2 Column information 15

4.3 Page information 16

4.4 Entering the date 16

4.5 Using multiple sheets 16

5. Checking the data after entry 17

5.1 Use of plots to highlight outliers 17

5.2 Data/Filter as a data checking tool 18

5.3 Tabulations 19

6. More complicated data sets 21

7. Conclusions 25

1. Introduction

In our guide called "*Data Management Guidelines for Experimental Projects*",[1] we noted that spreadsheets are commonly used for data entry because they are familiar, in widespread use and very flexible. However, we also stated, as a warning:

Their very flexibility means they can result in poor data entry and management. They should thus be ***used with great care****. Users should apply the* ***same rigour and discipline*** *that is obligatory with more structured data entry software.*

This guide is to explain what we mean by "great care" and "rigour and discipline" and hence to show how a spreadsheet package can be used effectively for data entry. In the illustrations we have used Excel 2007 as an example of a spreadsheet package, but our strategy applies equally to other spreadsheets.

We begin the next section with a simple example that shows a set of data that has been poorly entered, and we discuss the problems that can then arise. In Section 3 we consider features in Excel that facilitate simple and reliable data entry. The principle here is that the simpler the data-entry process, the more reliable will be the data that are entered.

In Section 4 we describe how to organise the data-entry process. This emphasises the need to set up the worksheet to include all the metadata, i.e. all associated background information relating to the data. This background information includes where the data came from, when they were collected, what the data values represent and so on.

Simplifying the data entry process and recognising the importance of the metadata has implications for the task of organising the data entry system within the spreadsheet. The organising phase now takes a little longer, but we believe this extra effort is justified. It is important to separate the task of organising the spreadsheet for data entry from the actual entry of the data.

Finally we cover a number of other issues, including validation checks after data entry and the entry of more complicated data sets.

[1] The Statistical Services Centre has written a series of guidelines for DFID. A list of the current titles is given on the back cover of this booklet.

2. An example

Figure 1 shows a typical data set that has been entered in Excel. This is a simple set of data that can be entered very effectively in a spreadsheet.

Figure 1. A typical data set in spreadsheet style

	A	B	C	D	E	F	G	H	I
1	block	plotwb	plot	species	rcd	height	branch	crown_0	crown_90
2	1	1	101	A.polycantha	12.7,13.3	438	23	673	730
3	1	2	102	A.indica	15.1	415	17	374	354
4	1	3	103	A.nilotica	11.1	350	20	268	375
5	1	4	104	Albizia lebeck	21.1	553	17	700	620
6	1	5	105	control					
7	2	1	201	A.indica	15.1	470	19	420	395
8	2	2	202	Control					
9	2	3	203	Albizia lebeck	12	300	12	394	322
10	2	4	204	A.polycantha	DEAD				
11	2	5	205	A.nilotica	10.1	343	22	420	401
12	2	1	201	A.nilotica	10	330	23	443	402
13	3	2	302	A.polycantha					
14	3	3	303	Control					
15	3	4	304	A.indica	26	410	21	415	440
16	3	5	305	A.polycantha	14.25	635	23	852	880
17	4	1	401	Control					
18	4	2	402	A.nilotica	12.5	373	23	602	500
19	4	3	403	A.polyantha	25.8	630	25	920	750
20	4	4	404	A.indica	18.5	404	22	420	370
21	4	5	405	Albizia lebeck	198	465	10	352	340

The data were entered by a clerk, who – as instructed – typed what was written on the recording sheet in the field. However, this has led to errors (ringed in Figure 1). For example, two of the names under the "species" heading have been typed slightly differently from the names for the same species used elsewhere in the same column. In the column headed "rcd", row 2 has two measurements entered, while in row 10, instead of a numerical value, the cell reports that the plant is dead. Such entries will cause problems when the data are transferred to a statistics package for analysis.

Most of these errors can be avoided if some thought is given to the layout of the data in the spreadsheet before data collection in the field commences. This activity is in fact the responsibility of the researcher, not of the data-entry clerk. This guide attempts to show how to avoid these problems.

There are other deficiencies if these data are all that is to be computerised. For instance, there is no indication as to where the data came from, when the data were collected or what they represent. There is no record of the units of measurements used. Such information is highly relevant and should be included with the data in the spreadsheet. This is especially true if the dataset is to be integrated with datasets from

other studies, or is to be passed to someone else for analysis. The entry of the metadata is discussed in Section 4.

You may think that this is a caricature and that real data would not be entered so poorly. Figure 2 shows some of the data from Figure 1 entered in a different way.

Figure 2 - An alternative way to enter data

	A	B	C	D	E	F
1	Crown 0					
2			Block			
3	Species	1	2	3	4	Mean
4	A.polycantha	673		852	920	815
5	A.indica	374	420	415	420	407
6	A.nilotica	268	420	443	602	433
7	Albizia lebeck	700	394		352	482
8	Species Mean	504	411	570	574	
9						
10	Crown 90					
11			Block			
12	Species	1	2	3	4	Mean
13	A.polycantha	730		880	750	787
14	A.indica	354	395	440	370	390
15	A.nilotica	375	401	402	500	420
16	Albizia lebeck	620	322		340	427
17	Species Mean	520	373	574	490	

The layout in this example is even worse. It confuses data entry and analysis.

The display of the data shown in Figure 2 is sometimes convenient as a prelude to the analysis. In Section 5 we show that with Excel it is easy to enter the data properly and then display the values as shown in Figure 2.

You should not necessarily expect the software to help in recommending good procedures for data entry. For example, if you decided to use Excel for the analysis of variance, then it would expect the data in the form shown in Figure 2. You would then find that Excel's facilities for ANOVA are limited (see the guide "*Excel for Statistics*"). The next step might be to ask about transferring the data to a statistics package but you would not find a good statistics package that expects the data in the tabulated form shown in Figure 2. They all expect it in the "rectangular" shape shown in Figure 1.

Thus, as we stated at the beginning of Section 1, Excel gives you total flexibility to enter your data as you wish, but no guide as to how to enter the data well.

3. Facilitating the data entry process

3.1 Introduction

This section describes how the data could have been entered so as to eliminate the mistakes and discrepancies illustrated in Figure 1. If the guidelines, given in the following sections, had been applied, the data could look like Figure 3.

Figure 3 - Worksheet after data entry guidelines have been used

	A	B	C	D	E	F	G	H	I
1	block	plotwb	plot	species	rcd	height	branch	crown_0	crown_90
2	1	1	101	A.polycantha		438	23	673	730
3	1	2	102	A.indica	15.1	415	17	374	354
4	1	3	103	A.nilotica	11.1	350	20	268	375
5	1	4	104	Albizia lebeck	21.1	553	17	700	620
6	1	5	105	Control					
7	2	1	201	A.indica	15.1	470	19	420	395
8	2	2	202	Control					
9	2	3	203	Albizia lebeck	12	300	12	394	322
10	2	4	204	A.polycantha					
11	2	5	205	A.nilotica	10.1	343	22	420	401
12	3	1	301	A.nilotica	10	330	23	443	402
13	3	2	302	A.polycantha					
14	3	3	303	Control					
15	3	4	304	A.indica	26	410	21	415	440
16	3	5	305	A.polycantha	14.25	635	23	852	880
17	4	1	401	Control					
18	4	2	402	A.nilotica	12.5	373	23	602	500
19	4	3	403	A.polycantha	25.8	630	25	920	750
20	4	4	404	A.indica	18.5	404	22	420	370
21	4	5	405	Albizia lebeck	19.8	465	10	352	340

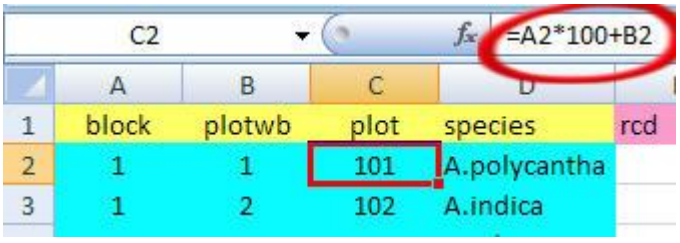
It is sensible to spend a little time thinking about the data, before rushing into using Excel. In this example, the data in the columns A-D would have been determined at the planning stage of the experiment and will be the same for all the measured variables. They could have been entered into a worksheet before any measurements were taken. A paper printout of Figure 4, which has additional information called meta-data (see Section 4) could be used in the field to collect the data.

Figure 4 - Paper data collection sheet

	A	B	C	D	E	F	G	H	I	J
1	Study code	TS695								
2	Study Title	Evaluation of rotational woodlands and fodder for soil fertility								
3	Site									
4	Scientist									
5	Project	4.1								
6	Project title	Systems evaluation and dissemination, developing choices for farmers								
7	Objectives	Assessment of tree growth								
8	Design	Random complete block								
9	Date									
10	Trait-title					Root collar diameter (cm)	Tree height (cm)	No. of branches ()	Crown cover diameter (cm)	
11	Units									
12	Orientation								0 degrees	90 degrees
13	Trait-name	block	plotwb	plot	species	rcd	height	branch	crown_0	crown_90
14		1	1	101	A.polycantha					
15		1	2	102	A.indica					
16		1	3	103	A.nilotica					
17		1	4	104	Albizia lebeck					
18		1	5	105	Control					
19		2	1	201	A.indica					
20		2	2	202	Control					

If you compare Figure 1 with Figure 3, you will notice that there is an extra column, named ***plot***. It is useful to have a column that uniquely defines each row of data. In this example, after the ***block*** and ***plotwb*** (plot within block) columns have been entered, ***plot*** has been calculated as follows: plot=100*block+plotwb, as shown in Figure 5.

Figure 5 - Calculating the *plot* column



When the data have been collected, you may decide to format the columns so that the data are displayed, for example to 2 decimal places for heights or lengths or as integers for data in the form of counts. You may wish to have range checks, so that typing data outside the minimum and maximum values is prohibited. You can add comments to cells, for example, to indicate that an animal has died or that a tree has been destroyed by an elephant!

The next few sub-sections illustrate the ideas suggested above.

3.2 Freezing or splitting panes

When entering data, it is useful to be able to keep the headings of columns always visible as you scroll down the screen. This can be achieved by freezing the panes. For example, assuming the column headers are in row 1 as in Figure 3, then by selecting **View → Window → Freeze Panes → Freeze Top Row,** the headings in row 1 will always be displayed, regardless of how many rows of observations are to be entered. A similar effect can be achieved by choosing **View → Window → Split**. To remove the freezing or the split, select **View → Window → Freeze Panes → Unfreeze Panes** or **View → Window → Split** (this feature acts as a toggle).

3.3 Drop-down lists

There are usually ways to avoid typing a sequence more than once. If the ***species*** column had a repeating list of text strings, these could be typed just once and then the spreadsheet's **Fill** option from the **Editing** section of the **Home** ribbon could be used to fill the remaining cells.

We illustrate the situation shown in Figure 1, where the species names do not form a repeating sequence. It is however more typical because it follows the species randomisation order in the field. The same four species plus the control are each repeated four times within the column. When the same text string is entered many times, typing errors inevitably occur. This is shown in Fig. 1 where the species name, A.Polycantha has been mis-typed as A.Polyantha in block 4.

Figure 6 shows the creation of a drop-down list. The five species names for block 1 are entered into cells D2:D6. These five names can then be used to create the drop-down list. Highlight all the ***species*** data cells, i.e. D2:D21 as in Fig.6. Select **Data → Data Tools → Data Validation → Allow: → List**. For the **Source** of the list, highlight the names already typed for block l, i.e. cells D2:D6 (as shown).

Once the drop-down list has been created, selecting a cell in that column will bring up an arrow (on the right side of the cell). Clicking on this arrow will display the drop-down list (see Figure 7). An appropriate selection can then be made from the drop-down list.

New data entered in these cells must either be a string from the given list, or else the cell can be left blank. Entry into the cells can be forced by unchecking the **Ignore blank** field in the **Data Validation** dialog box shown in Figure 6.

The rest of the ***species*** can now be selected from the drop-down list. For example, Figure 7 shows the selection of A.Indica for plot 201.

Figure 6 - Creating a drop-down list for species

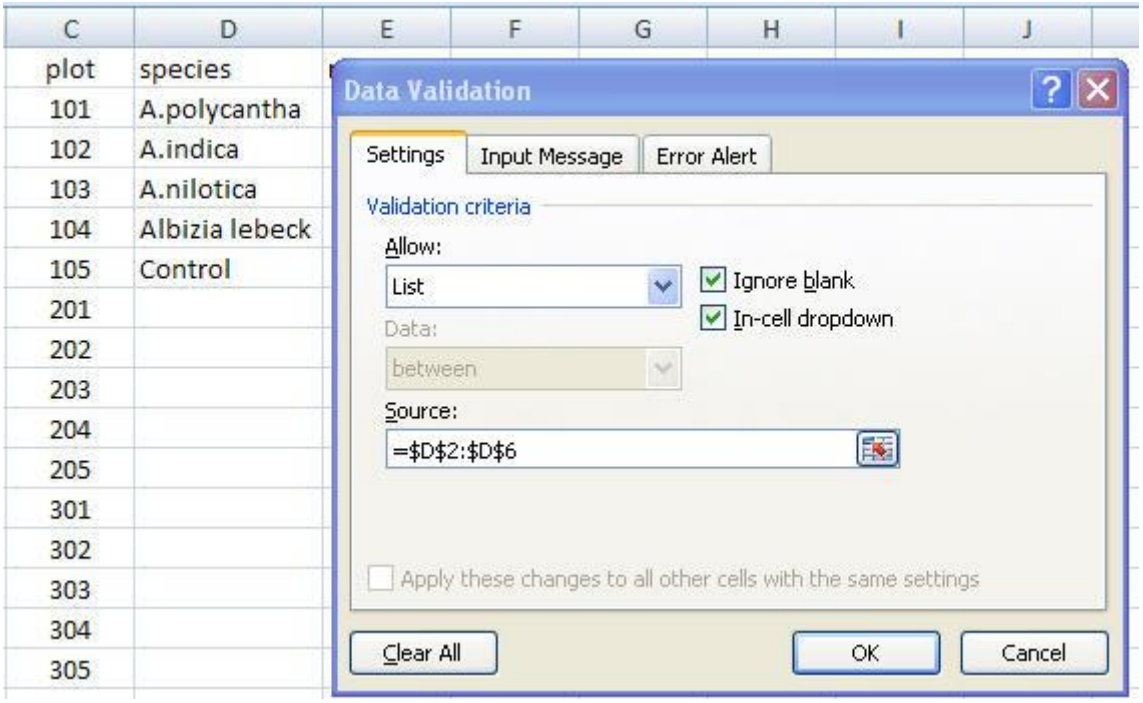
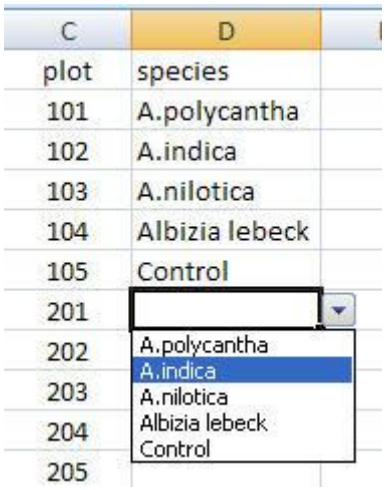


Figure 7 - Selecting a *species* from a drop-down list



Using drop-down lists for data entry helps to ensure that errors in spelling do not occur.

3.4 Data validation

Validation checks can and should be set on ranges of cells within the spreadsheet. A range could be an entire column/row, several columns/rows, or just a single cell. The validation rules apply when new data are entered.

One validation tool available in Excel is the facility to set up range checks for numerical data. For example, the measurements recorded for the variable ***rcd*** are

expected to be in the range from 10 to 26. To set up a range check, **highlight** cells E2 to E21, select **Data → Data Tools → Data Validation → Allow: → Decimal** and set the **Minimum** as **10** and the **Maximum** as **26** as shown in Figure 8. At the same time, an Input Message and an Error Alert can be set up and their effects are shown in Figure 9. The Input Message is displayed when each cell in the column is activated (Figure 9a). This reminds the data entry person of the range of values allowed. The Error Alert Message is displayed when a value outside this range is typed (Figure 9b). Note that only the data cells are highlighted, not the variable name at the top.

Figure 8 - Setting up range checks

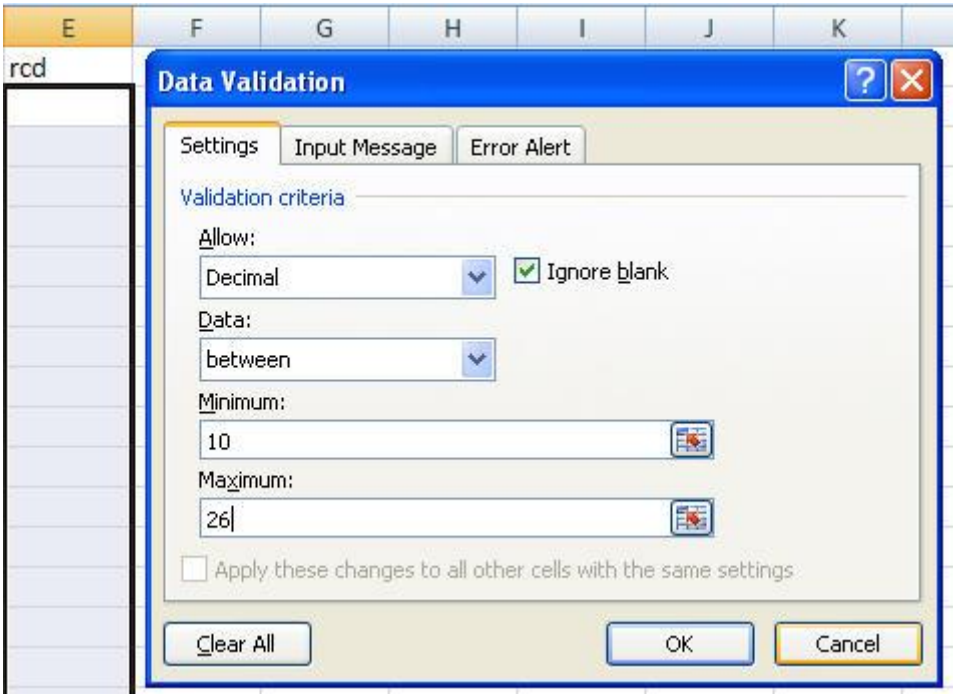


Figure 9 - Validation at data entry

Figure 9a

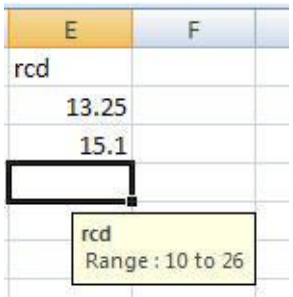
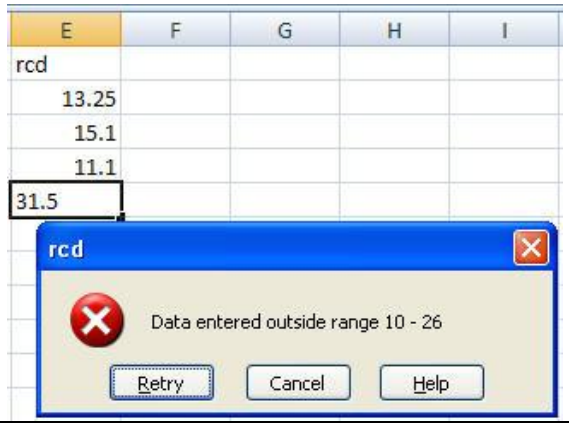


Figure 9b



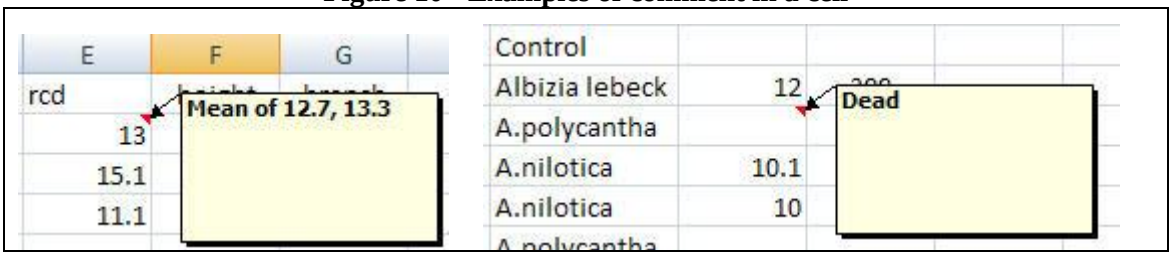
3.5 Adding comments to cells

Excel has a facility for adding comments to a cell. These differ from values within the cell. Comments should be used for any unusual observations or questions concerning a particular data value. When entering the data for ***rcd***, a decision had to made for ***plot*** 101, where 2 values were entered on the data recording sheet. We chose to calculate the mean and add a comment to the cell, as shown in Figure 10.

If there had been several plots with two values recorded, 2 columns of ***rcd*** data could have been entered and a third column could have been used to calculate the mean. With the specific cell highlighted, the sequence **Review → Comments → New Comment** from the ribbon allows comments to be entered.

Comments are also useful in explaining why certain values are missing. For example in Figure 1 the value **Dead** had been entered in cell **D10**. With the exception of the column header, all cells in a column should have the same data type. The data type for column D is numeric. The string "Dead" in cell D10 is therefore inappropriate for the cell, but can be put as a comment to indicate why the value is missing. Figure 10 shows comments that have been added to cells. A comment is set on a single cell but can be copied to a range of other cells.

Figure 10 - Examples of comment in a cell

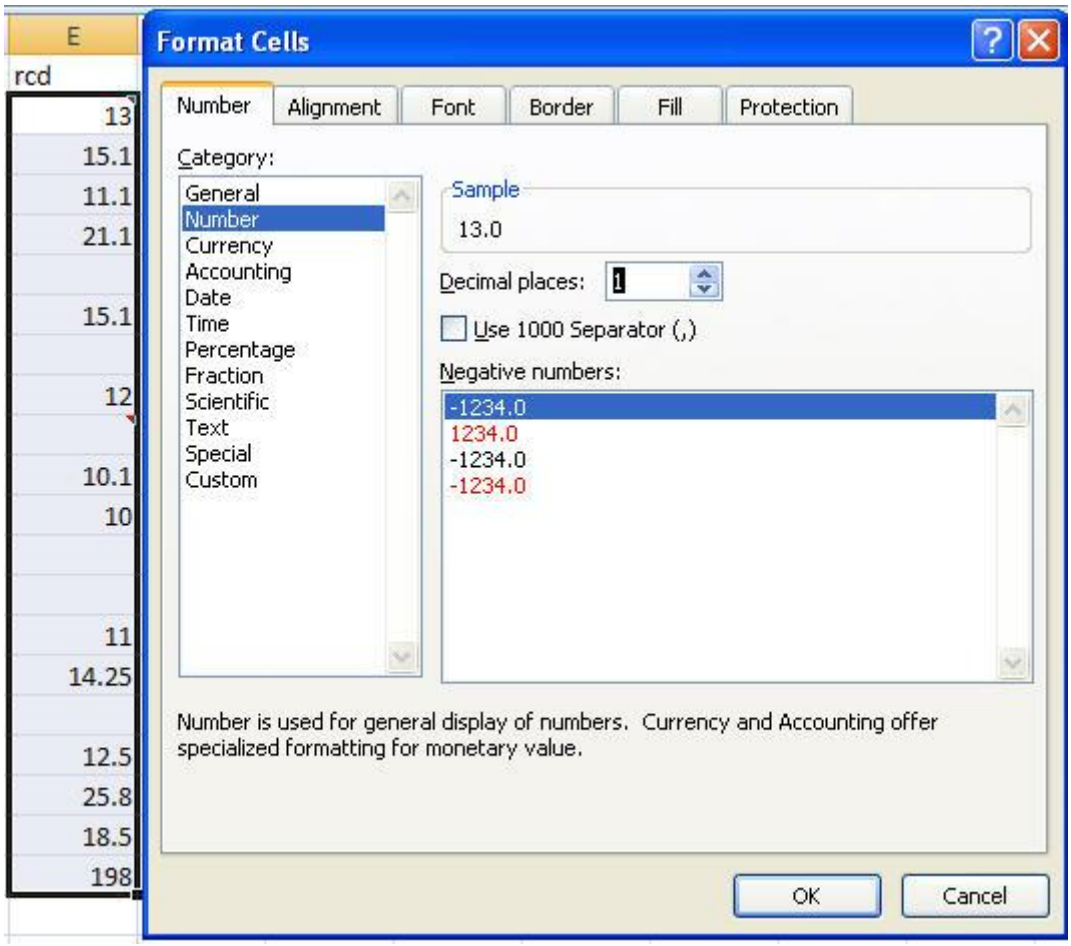


When you wish to remove a comment from a cell, for example when a query has been resolved and the correct value has been entered, highlight the cell, select **Review → Comments → Delete**

3.6 Formatting columns

Excel suppresses trailing zeros by default. For example, in Figure 10 the value **13.0** is displayed as **13**. To display the column of data to 1 decimal place, highlight the data, then use **Home → Cells → Format → Format Cells → Number → Decimal places → 1** as shown in Figure 11. This also shows that there is flexibility in how the data are displayed, e.g. font size, cell borders etc.

Figure 11 - Formatting data

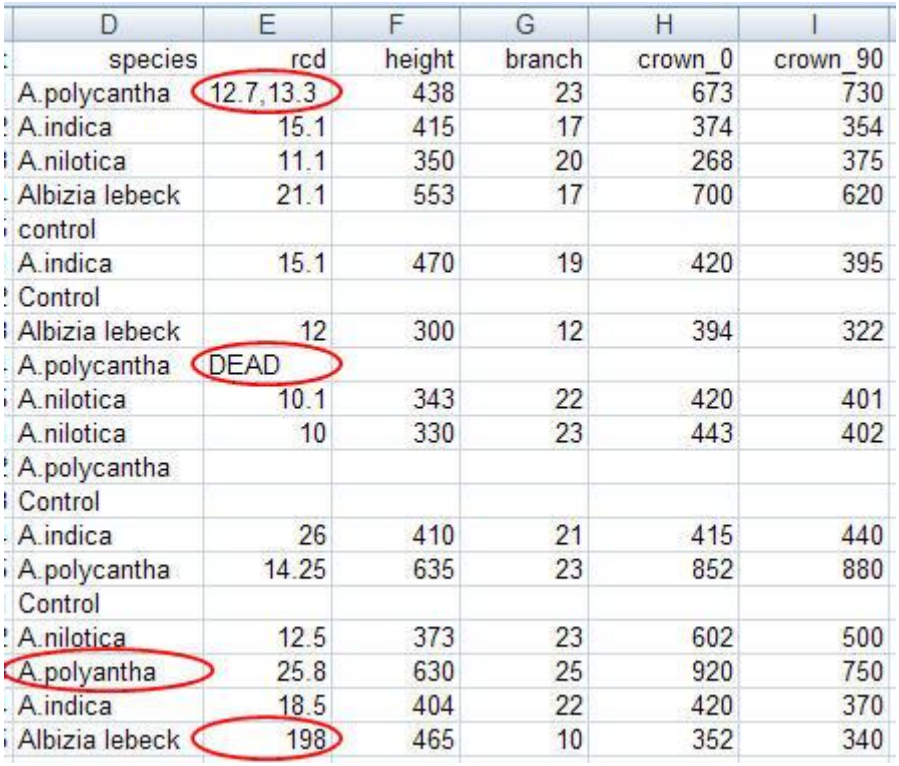


3.7 Checking Existing Data

The suggestions given above are all intended to aid data entry. However, there is also a facility, previously known as auditing, for checking data that have already been entered. It is recommended that you use this facility whenever you add validation rules following the data entry or when you make changes to existing rules. Validation rules are very flexible. The type of data such as whole number, decimal, text, etc., can be specified; data not of the correct type will be rejected on data entry. You can then set further restrictions to accept only those numbers within a given range, or text strings of a particular length, for example.

We illustrate facility on the data in Figure 1, where validation rules have been added for both *species* and *rcd*. Select **Data → Data Tools → Data Validation → Circle Invalid Data**. Figure 12 shows the errors circled in red for the variables ***rcd*** and ***species*.**

Figure 12 - Auditing of existing data



	D	E	F	G	H	I
	species	rcd	height	branch	crown_0	crown_90
	A.polycantha	12.7,13.3	438	23	673	730
	A.indica	15.1	415	17	374	354
	A.nilotica	11.1	350	20	268	375
	Albizia lebeck	21.1	553	17	700	620
	control					
	A.indica	15.1	470	19	420	395
	Control					
	Albizia lebeck	12	300	12	394	322
	A.polycantha	DEAD				
	A.nilotica	10.1	343	22	420	401
	A.nilotica	10	330	23	443	402
	A.polycantha					
	Control					
	A.indica	26	410	21	415	440
	A.polycantha	14.25	635	23	852	880
	Control					
	A.nilotica	12.5	373	23	602	500
	A.polyantha	25.8	630	25	920	750
	A.indica	18.5	404	22	420	370
	Albizia lebeck	198	465	10	352	340

4. The metadata

In the previous section we looked at some methods of validating data both during data entry and afterwards. However the data we started with is still not complete. In this section we advocate adding rows and columns to the spreadsheet before the body of data. These extra rows will store documentation that provides background information about the data, i.e. it comprises the "meta-data".

Figure 13 shows a blank spreadsheet divided into input areas. The number of rows and columns in these areas is not fixed but can increase or decrease depending on the data to be entered.

Figure 13 - Input areas in an Excel Worksheet

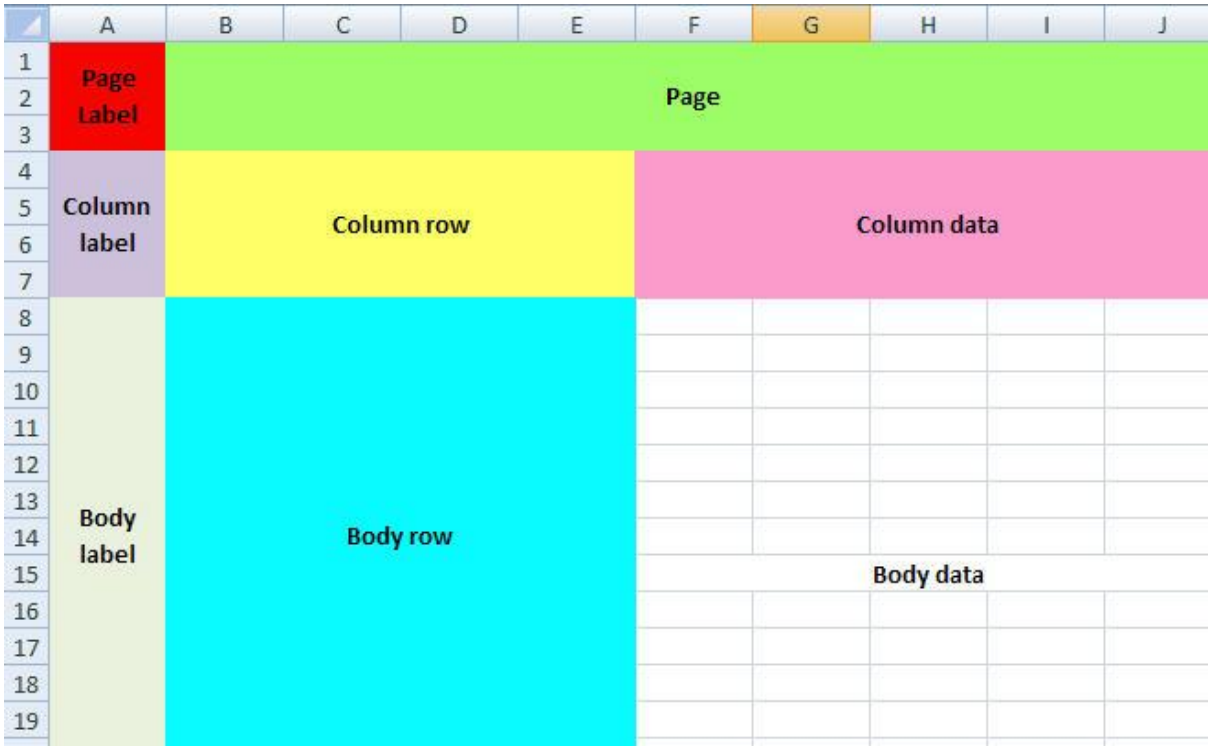


Figure 14 shows the data from Figure 3 with this structure.

It is useful to adhere to a strict code as to what should be entered into each section of the spreadsheet. Keep in mind the following three aims:

1. Encourage completeness;
2. Avoid unnecessary duplication;
3. Minimise errors.

Figure 14 - Metadata

	A	B	C	D	E	F	G	H	I	J
4	Scientist	Riobert Otsyna								
5	Project	4.1								
6	Project title	Systems evaluation and dissemination, developing choices for farmers								
7	Objectives	Assessment of tree growth								
8	Design	Random complete block								
9	Date	29 November 1999								
10	Trait-title					Root collar diameter (cm)	Tree height (cm)	No. of branches ()	Crown cover diameter (cm)	
11	Units								0 degrees	90 degrees
12	Orientation									
13	Trait-name	block	plotwb	plot	species	rcd	height	branch	crown_0	crown_90
14		1	1	101	A.polycantha	13	438	23	673	730
15		1	2	102	A.indica	15.1	415	17	374	354
16		1	3	103	A.nilotica	11.1	350	20	268	375
17		1	4	104	Albizia lebeck	21.1	553	17	700	620
18		1	5	105	Control					
19		2	1	201	A.indica	15.1	470	19	420	395
20		2	2	202	Control					
21		2	3	203	Albizia lebeck	12	300	12	394	322

4.1 Body data

The **Body data** area contains all values that have been observed or measured. With reference to the data in Figure 3, this is the range of cells **E2:I21**. It is the range **F14:J33** in Figure 14. Values in columns **A** to **D** (of Figure 3) are not measurement data but correspond to information relating to the design of the experiment. These, with the exception of the first row, fall into the **Body row** area of Figure 13. The **Body Label** area can be used to add more information about the body data.

4.2 Column information

In the **Column Label**, **Column Row** and **Column Data** areas we add information to describe the measurements. Here a measurement corresponds to a unique identifiable column of data.

The column headers in the first row of Figure 3 go into the column data areas of Figure 13 (i.e. F13:J13). We still need more information. For instance, the data file does not clarify what specific measurements are being made, so it is necessary to include a description of what ***rcd***, ***height***, ***branch***, etc. actually represent. It is also necessary to specify the units of measurement being used for each of these variates. These are shown in Figure 14 in the range of cells F10:J12.

4.3 Page information

All that remains to be entered now is documentation related to the whole dataset. Effective documentation requires that, given a data set, one should be able to retrieve the original protocol plus all other related data. This documentation appears in the **Page** area of the spreadsheet with the **Page Label** area providing labels for each piece of information. The minimum documentation is the study code and its title. It is also useful to include the location where the data was collected and the name of the person responsible for the study. The study objectives, the design and other protocol details should be entered if available.

4.4 Entering the date

All observed raw data should be associated with a date of observation. Before entering the date of measurement ensure that the computer date setting matches the expected date input format. Remember that **9/2/99** could refer either to **9 February 1999** or **2 September 1999** depending on the date settings on your computer. (These can be checked by selecting **Control Panel → Regional and Language Options** from the Windows **Start** menu.) To ensure you have the correct date, use **Home → Cells → Format →Format Cells → Date** and choose a display setting that includes the name of the month, e.g. February 9, 1999

If all observations are taken on the same date, the date should be entered in the **Page** section with an appropriate label in the **Page label** area. This is shown in Figure 14. If this is not the case the dates should be entered into the **Body row** area with an appropriate label in the **Column row**. Where there are a limited number of dates, for example where the values were measured on one of just three or four dates, a drop-down list should be created with these dates and values chosen from the list on data entry.

4.5 Using multiple sheets

Entering the data plus the metadata as described above gives us Figure 14.

An alternative is to put the "Page information" on to a separate sheet in the Excel Workbook. This is often convenient when there is a lot of information at the dataset level.

In such cases you may still have a small "Page" section in each data sheet that describes the type of measurements that are entered in that sheet.

5. Checking the data after entry

We have demonstrated in Section 3 how auditing can be done to check existing data against the validation rules. This is a single operation using the command **Data → Data Tools → Data Validation → Circle Invalid Data**. Here we give some further steps that may be undertaken to validate the data after it has been entered.

5.1 Use of plots to highlight outliers

Scatter plots are useful tools for helping to spot suspicious inputs (i.e. outliers). Many would be trapped at the data entry stage if validation rules had been set up but there may still be some values that differ substantially from the rest. Figure 15 shows a scatter plot of Root Collar Diameter (referred to earlier as **rcd**) where the x-axis corresponds to the order in which the values appear in the data file. This plot shows that all data records lie from about 10 units to about 25 units.

Such a plot is quick and easy to produce and can be done for all observed variables. It is also useful to produce scatter plots of pairs of variables if knowledge about the two variables being plotted are expected to show a definitive pattern. Figure 16 shows tree height plotted against root collar diameter. There is at least one suspicious observation.

Figure 15 - Scatter plot of Root Collar Diameter

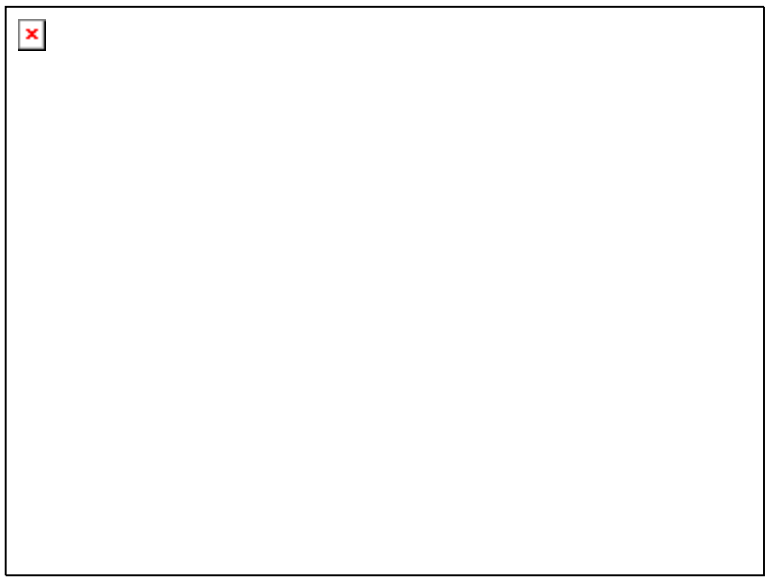
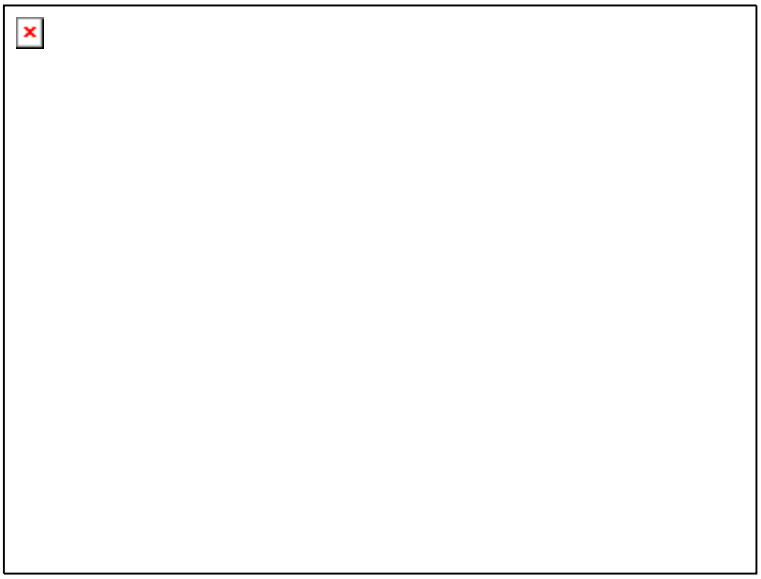


Figure 16 - Tree Height against Root Collar Diameter

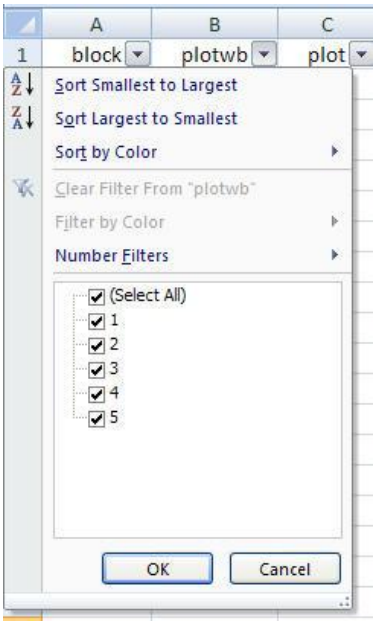


5.2 Data/Filter as a data checking tool

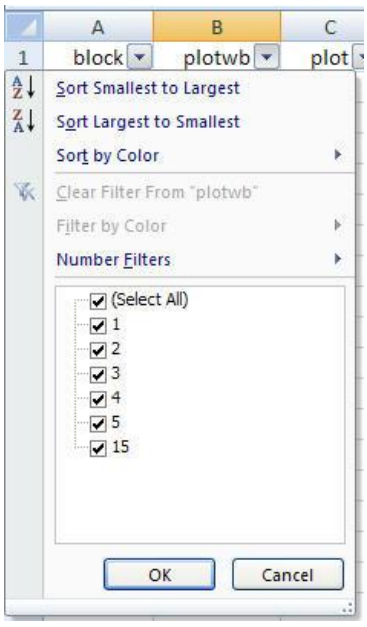
In the example dataset we are using, we know that there are four blocks with five plots in each block. In Figure 17a, the **Data → Filter** option has been used for ***plotwb***. Clicking the autofilter tab for ***plotwb*** gives a list of values in that column. It shows that only values 1 to 5 have been entered in ***plotwb.*** If, for example, the value 15 had been entered instead of 5, the list would also contain the value 15 (Figure 17b) and so the error could easily be spotted.

Figure 17 - AutoFilter to spot unexpected entries

17a



17b



5.3 Tabulations

Frequency tables can be generated using Excel's pivot table facility. Generating tables in Excel is described in Appendix 1 of the booklet "Using Excel for Statistics".

In Figure 18, the count of the number of observations within each of the 4 blocks plus the means of the variables are shown for the data in Figure 3 and Figure 1. Using pivot tables is another way of spotting data entry errors.

Figure 18 - Pivot tables to spot data entry errors

	A	B	C	D	E	F	G
1	Pivot table for Figure 3 dataset						
2		Data					
3	Block	Count of block	Mean rcd	Mean height	Mean branch	Mean Crown_0	Mean Crown_90
4	1	5	15.08	439.00	19.25	503.75	519.75
5	2	5	12.40	371.00	17.67	411.33	372.67
6	3	5	11.75	458.33	22.33	570.00	574.00
7	4	5	19.15	468.00	20.00	573.50	490.00
8	**Grand Total**	**20**	**14.95**	**436.86**	**19.79**	**518.07**	**491.36**
9							
10							
11	Pivot table for Figure 1 dataset						
12		Data					
13	Block	Count of block	Mean rcd	Mean height	Mean branch	Mean Crown_0	Mean Crown_90
14	1	5	15.08	439.00	19.25	503.75	519.75
15	2	6	11.80	360.75	19.00	419.25	380.00
16	3	4	12.63	522.50	22.00	633.50	660.00
17	4	5	19.15	468.00	20.00	573.50	490.00
18	**Grand Total**	**20**	**14.95**	**436.86**	**19.79**	**518.07**	**491.36**

Note that, if you had required your data to be in the layout shown in Figure 2, this could have been accomplished from the recommended format using pivot tables, yielding results as shown in Figure 19.

Figure 19 - Using pivot tables to obtain an alternative layout for the data

	A	B	C	D	E	F
1	**Mean of Crown_0**	**block**				
2	**species**	**1**	**2**	**3**	**4**	**Grand Total**
3	A.indica	374	420	415	420	407
4	A.nilotica	268	420	443	602	433
5	A.polycantha	673		852	920	815
6	Albizia lebeck	700	394		352	482
7	Control					
8	**Grand Total**	**504**	**411**	**570**	**574**	**518**
9						
10	**Mean of Crown_90**	**block**				
11	**species**	**1**	**2**	**3**	**4**	**Grand Total**
12	A.indica	354	395	440	370	390
13	A.nilotica	375	401	402	500	420
14	A.polycantha	730		880	750	787
15	Albizia lebeck	620	322		340	427
16	Control					
17	**Grand Total**	**520**	**373**	**574**	**490**	**491**

6. More complicated data sets

There are various ways that studies can produce data that have a more complicated structure than the example that we have considered here. The most common complication is when measurements are taken at different levels. For example, a survey might have data at the "village", "household" and "person" levels. For illustration, we consider an example of an experiment with some data at the "plot" level and other measurements at the "plant" level.

To emphasise the concepts, we show the system in Excel by a set of figures that should be self-explanatory. Readers wanting more details could look at the notes from the Data-entry Course Notes that were prepared by the Statistical Services Centre for the Institut de Recherche Agronomique de Guinée (IRAG). The notes are available from the SSC web site.

The data are from an experiment to compare nine different varieties of potatoes. The design is a randomised complete block with 3 replicates, giving 27 plots in all. Within each plot, some measurements were made at the plot level while others were made on 20 plants within each plot.

Figure 20 - Measurements at the "plot" level"

	A	B	C	D	E	F	G	H	I
1	Study:			Evaluation of varieties of potatoes in Guinea					
2	Location:			Bareng Station					
3	Experimental Area:			20.16 m^2					
4	Experimental Protocol:			pdt98.docx					
5	Description:			Final Harvest (date)	Weight harvested, by grade per plot (kg)				Dry matter (%)
6	Unit:								
7	Plot	Replicate	Treatment	Hdate	Grade28	Grade35	Grade45	Grade55	Dry_M
8	101	1	Mondial	02/07/1998	3.20	3.00	5.00	3.80	13.00
9	102	1	Anais	24/06/1998	2.00	4.00	7.30	3.40	14.44
10	103	1	70050/96	02/07/1998	1.30	5.00	4.10	3.10	14.44
11	104	1	Desiree	24/06/1998	6.20	6.90	6.90	5.00	18.00
12	105	1	71343	24/06/1998	4.00	7.10	5.30	1.90	17.40
13	106	1	Nicola	02/07/1998	6.60	8.00	5.20	2.80	17.66
14	107	1	87/72/4	02/07/1998	3.60	2.90	5.00	3.00	16.00

The Excel workbook used for the data entry consists of six worksheets, as shown in Figure 20.

The **Yield** worksheet in Figure 20 shows the metadata and some of the measurements made at the "plot" level. The **Entomology** and **Pathology** worksheets also contain data for the 27 plots, and the plot number is repeated in each sheet.

Figure 21 shows the design of the **Tuber** worksheet that was used for entering the "plant" level data, and was a "copy" of the paper data collection sheet. The body of the sheet contains 20 rows (plants) and 27 (plots) * 2 (counts of the number of stems and tubers) columns. This layout makes it easier for the data collection person to record the counts. Figure 22 shows part of the worksheet with some data entered.

Figure 21 – Design of the Tuber worksheet

Plot	**101**		**102**		**103**		**…**		**309**	
Plant	**Stem**	**Tuber**	**Stem**	**Tuber**	**Stem**	**Tuber**			**Stem**	**Tuber**
1										
2										
3										
….										
19										
20										

Figure 22 - Counts at the "plant" level

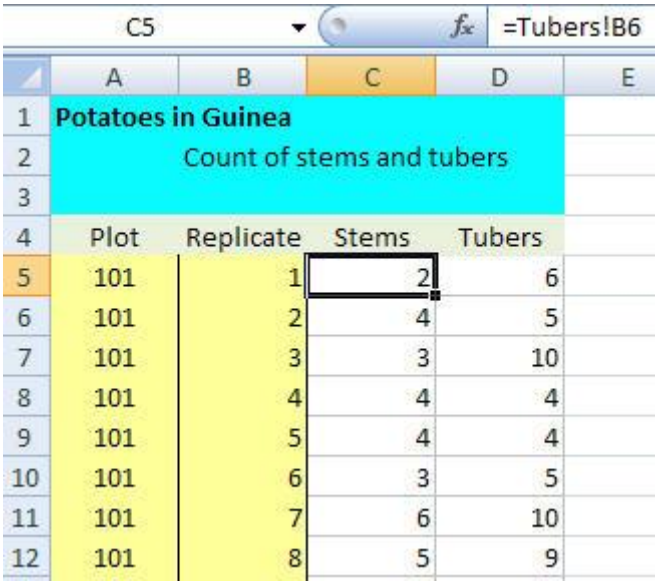
	A	B	C	D	E	F	G	H
1		Potatoes in Guinea						
2		Count of stems and tubers						
3								
4	Plot	101		102		103		
5		Stems	Tubers	Stems	Tubers	Stems	Tubers	Stem
6	1	2	6	3	8	3	5	
7	2	4	5	4	5	4	7	
8	3	3	10	4	8	2	7	
9	4	4	4	4	5	4	6	
10	5	4	4	2	6	3	10	
11	6	3	5	4	4	5	6	
12	7	6	10	2	9	4	8	

The **Plants** worksheet is a copy of the data in the **Tuber** worksheet, with the data reorganised into columns of length 540 (i.e. 27 plots * 20 plants). It is shown in Figure 23, where, for example, cell C5 = cell B6 in the **Tuber** worksheet. Data should not be entered into the **Plants** worksheet, since values will be updated

automatically whenever a change is made to the **Tuber** worksheet. This arrangement of the data is the format needed to generate pivot tables, and is also the format required by most statistics packages. An example is shown in Figure 24.

Be aware that if there is a blank in the cell Tubers!B6, then the cell C5 will contain a zero. The way round this is to use an "if" function, for example, "=IF(Tubers!B6="","",Tubers!B6)".

Figure 23 - Plant data suitable for data analysis

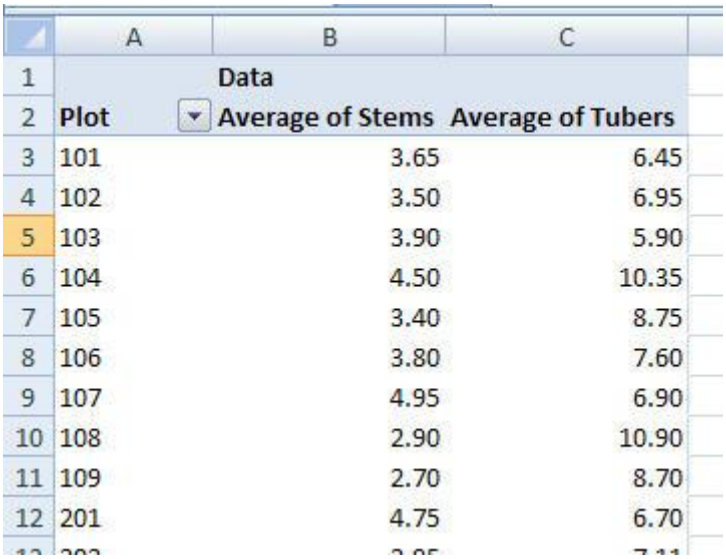


C5 =Tubers!B6

	A	B	C	D	E
1	Potatoes in Guinea				
2		Count of stems and tubers			
3					
4	Plot	Replicate	Stems	Tubers	
5	101	1	2	6	
6	101	2	4	5	
7	101	3	3	10	
8	101	4	4	4	
9	101	5	4	4	
10	101	6	3	5	
11	101	7	6	10	
12	101	8	5	9	

The final worksheet, **Table 1** (shown in Figure 24), was generated using the pivot table facility in Excel. It gives the plot means for each of the 20 plants within a plot.

Figure 24 - An example of a pivot table



	A	B	C
1		Data	
2	Plot	Average of Stems	Average of Tubers
3	101	3.65	6.45
4	102	3.50	6.95
5	103	3.90	5.90
6	104	4.50	10.35
7	105	3.40	8.75
8	106	3.80	7.60
9	107	4.95	6.90
10	108	2.90	10.90
11	109	2.70	8.70
12	201	4.75	6.70

This example of a more complicated dataset illustrates the following:

- the use of multiple worksheets to store data;
- the use of **Plot** in each sheet to link the data in separate sheets;
- a method of linking data in separate sheets so that it is stored only once but can be viewed in different ways;
- a quick way of producing summary tables.

7. Conclusions

Our main conclusion is that a spreadsheet package, such as Excel, can be used for effective data entry, particularly for data sets with a simple structure.

Even for simple data entry, it is important to separate the task of organising the spreadsheet from that of actually entering the data.

Excel provides a range of aids to effective data entry, some of which have been described in Section 3 of this booklet.

It is simple and useful to include the "metadata" within the data entry process. Compare Figure 14 with Figure 1 to review the potential of Excel for complete data entry for simple data sets.

It is also possible to use Excel for the entry of more complex data sets, as was outlined in Section 6 of this booklet. This takes more planning and we recommend that, for such tasks, consideration also be given to the use of a database package, such as Access, or a specialised data entry system, such as EpiInfo, which is distributed by the Centers for Disease Control and Prevention (CDC), Atlanta, U.S.A. More information about EpiInfo can be found from the Internet from www.cdc.gov/epo/epi/epiinfo.htm, from where it can be downloaded free of charge.

In supporting the use of a spreadsheet package for data entry we have been driven, to some extent, by its popular use. It is clear that many people will continue to use a spreadsheet for their data entry and this document suggests ways of making the entry effective. There are limitations. For example there are no easy facilities for skipping fields, conditional on the entry of initial codes. There are no automatic facilities for "double entry". The graphics in Excel, that were illustrated in Section 5, are meant primarily for presentation and there are no boxplots (or other exploratory techniques) that could assist in data scrutiny.

Spreadsheets are intended as "jack-of-all-trades" software. They are certainly not the master of data entry. Hence if the data entry component is large, or complex, we suggest that a spreadsheet should **not** be the only contender for the work.

The Statistical Services Centre is attached to the Department of Applied Statistics at The University of Reading, UK, and undertakes training and consultancy work on a non-profit-making basis for clients outside the University.

These statistical guides were originally written as part of a contract with DFID to give guidance to research and support staff working on DFID Natural Resources projects.

The available titles are listed below.

- *Statistical Guidelines for Natural Resources Projects*
- *On-Farm Trials – Some Biometric Guidelines*
- *Data Management Guidelines for Experimental Projects*
- *Guidelines for Planning Effective Surveys*
- *Project Data Archiving – Lessons from a Case Study*
- *Informative Presentation of Tables, Graphs and Statistics*
- *Concepts Underlying the Design of Experiments*
- *One Animal per Farm?*
- *Disciplined Use of Spreadsheets for Data Entry*
- *The Role of a Database Package for Research Projects*
- *Excel for Statistics: Tips and Warnings*
- *The Statistical Background to ANOVA*
- *Moving on from MSTAT (to Genstat)*
- *Some Basic Ideas of Sampling*
- *Modern Methods of Analysis*
- *Confidence & Significance: Key Concepts of Inferential Statistics*
- *Modern Approaches to the Analysis of Experimental Data*
- *Approaches to the Analysis of Survey Data*
- *Mixed Models and Multilevel Data Structures in Agriculture*

The guides are available in both printed and computer-readable form. For copies or for further information about the SSC, please use the contact details given below.

Statistical Services Centre, The University of Reading
P.O. Box 240, Reading, RG6 6FN United Kingdom

tel: SSC Administration +44 118 931 8025
fax: +44 118 975 3169
e-mail: statistics@reading.ac.uk
web: http://www.reading.ac.uk/ssc/