Data Handling in Excel

Working Effectively and Confidently with Data

Sandro Leidi, Roger Stern and Myles Kirk - Gushowaty

Statistical Services Centre
The University of Reading, UK

September 2007
Version for the SADC Training Pack
# TABLE OF CONTENTS

Getting started

1.1 Introduction .......................................................... 4
1.2 Using Windows ....................................................... 5
1.3 Spreadsheets .......................................................... 8
1.4 A sample workbook ................................................ 10
1.5 Basic workbook operations ....................................... 12
1.6 Summary .............................................................. 16

Chapter 2: Data manipulation ........................................... 17

2.1 Introduction .......................................................... 17
2.2 Structure of your data .............................................. 17
2.3 Master copies and working copies .............................. 19
2.4 Sorting ................................................................. 20
2.5 Selecting subsets of your data (Filters) ....................... 23
2.6 Referring to data cells using names: named ranges ....... 25
2.7 In summary ............................................................ 27

Chapter 3: Calculations .................................................... 28

3.1 Introduction .......................................................... 28
3.2 Formula Bar .......................................................... 28
3.3 Comparison operators ............................................. 29
3.4 Using column names instead of cell addresses .......... 29
3.5 Functions ............................................................. 30
3.6 In Summary ............................................................ 33

Chapter 4: Further Calculations ......................................... 34

4.1 Introduction .......................................................... 34
4.2 Weighted means ..................................................... 34
4.3 Using functions for coding ....................................... 35
4.4 Attaching labels to numeric codes ............................ 37
4.5 In summary ............................................................ 39

Chapter 5: Multi-level data ................................................ 40

5.1 Introduction .......................................................... 40
5.2 Setting up a unique identifier for households .......... 41
5.3 From one to many: replicating information ............... 42
5.4 From many to one: aggregating information ............ 43
5.5 In summary ............................................................ 45

Chapter 6: Data Entry and Validation .................................. 46

6.1 Introduction .......................................................... 46
6.2 Entering data ......................................................... 46
6.3 Validating data ....................................................... 46
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.4</td>
<td>Missing values</td>
<td>48</td>
</tr>
<tr>
<td>6.5</td>
<td>In summary</td>
<td>49</td>
</tr>
</tbody>
</table>
Getting started

1.1 Introduction

This introductory guide covers the basics of Excel that are needed for entering and manipulating your data. It is in eight chapters.

Chapter 1 reviews the basics of Windows and Excel that will be assumed. This is provided as pre-course reading or may be used during preparatory computer training.

Data handling is covered in Chapter two. The importance of keeping data in a “list” is explained, and data auditing, filtering, and sorting, are introduced among other topics. The data used for illustration are taken from the 1997-98 Kenya National Household Survey. A CD from this survey provides the questionnaires, raw data, data dictionaries and reports 1.

Chapter three is dedicated to performing calculations, and understanding the formula bar. It also covers how to use column names instead of cell address, and explores Excel’s function wizard.

Chapter four looks at how to use functions for coding, how to construct a weighted mean and how to attach labels to numeric codes. These processes are more advanced, but with some diligent work they can be mastered.

Chapter five looks at linking and merging data collected at different levels, which are naturally stored in separate worksheets. This task is often performed by database managers and requires some rigour in the use of spreadsheet functions.

Data entry and validation are examined briefly in chapter 6. Here we describe how to correctly enter and validate your data using Excel, and how to deal with missing values.

Chapter seven describes data handling facilities in SSC-Stat, a statistical add-on program, that simplify some data manipulation tasks that Excel is commonly used for. These are to “stack” data to put it into list format and to simplify the use of Excel’s powerful VLOOKUP function, which can be used to group data or to attach labels to numeric codes.

This guide is used as supporting material for sessions of the SADC training. Other sessions look at the production of tables and graphs in Excel.

1 This is an excellent resource for the training and we are grateful to CBS (Central Bureau of Statistics) of Kenya for providing the information and for granting permission to use the example in this guide.
1.2 Using Windows

There are many different versions of Windows. These notes apply for all versions from Windows 95. Fig. 1.1 shows a typical Windows desktop. This is the base from which to open application programs, such as Excel.

You must be comfortable with the use of your computer mouse to use Windows effectively. The table below shows the four basic mouse operations.
Mouse operations

<table>
<thead>
<tr>
<th>Operation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clicking</td>
<td>Move the mouse so the tip of the pointer touches the element you want to use. Then press and release the left mouse button.</td>
</tr>
<tr>
<td>Right-clicking</td>
<td>Same as the above, but you press and release the right button.</td>
</tr>
<tr>
<td>Double-clicking</td>
<td>Press and release the left button twice in rapid succession</td>
</tr>
<tr>
<td>Dragging</td>
<td>Press and hold down the left button. Then, with the button still pressed, move the mouse across the screen. Release the button when you are at your destination.</td>
</tr>
</tbody>
</table>

If you need practice, then here is a short exercise.

1. Double-click\(^2\) on the **My Computer icon** on the Windows desktop. This will open the My Computer window, as shown in Fig. 1.2.

2 With some versions of Windows you only have to single-click to open the window.
2. Click the **Minimise** button. This reduces the **My Computer** window to a button on the taskbar.

3. Click on the **My Computer** icon on the **taskbar**, see Fig. 1.2. This restores the **My Computer** window to the desktop.

4. Click the **Maximise** button. The **My Computer** window now fills the whole screen. The **Maximize** button changes to a **Restore Down** button.

5. Click the **Restore Down** button. The **My Computer** Window is now restored to its original size.

6. Move the mouse pointer, so it is on the title bar of the **My Computer** window, see Fig. 1.2. Drag the window to the bottom right-hand corner of the screen, and then release the mouse button. Then drag the Window back to roughly its original position.

7. Point to the lower left-hand corner of the **My Computer** Window. The pointer turns into a double-headed arrow. Drag the window corner down, then release the button. This has enlarged the Window. Return it to roughly its previous size.³

8. Repeat this operation, but start with the pointer at an edge of the Window, rather than a corner. This changes the shape of the window.

9. Click on the Close Button. This closes the **My Computer** window.

If you are a relative beginner, then take time to practice these operations, both in the type of exercise above, and when you start using an application, like Excel.

---

³ When you are working with more than one window, you will often wish to arrange them conveniently on the desktop. This action is known as ‘tiling’.
If you need more information about Windows, then take time also to access the On-line Help system. This provides tutorials and other information to help you to use Windows effectively. We show an example of the Windows Help in Fig. 1.3.

Beginners to computing often look for books they can read to gain experience. If you are in this situation we urge you mainly to practice with on-line help and information instead. The best way to gain experience in computing is simply to use a computer!

One of the key features of Windows is that the way you use all applications is basically the same. So look for the common points, like how to use Help, how to open and save files, and so on. Then with each new application, all you have to do is look for the new features, and they are probably why you chose that application.

Once you gain in experience you will also find that you do not need to wait for a formal training course before starting to use a new application. Ask the presenters of your next course what they do. You will find that they often taught themselves, using books and the on-line information, for many of the applications that they are now trainers for.

1.3 Spreadsheets

Excel is a piece of software. It is an application to evaluate and present information in a spreadsheet format. Spreadsheets were originally developed for simple business uses, such as a financial report or an inventory management.

Excel is now so flexible that it is used for many other applications, including data analysis. You can use Excel to enter data and for simple analyses, such as the construction of appropriate tables and graphs. The results can then be transferred to other applications, such as Word for including in a report, or PowerPoint for including in a presentation.

To start Excel you probably have an icon on the desktop that you can click, just as you did to open the My Computer window. Otherwise click the Start button, then Programs and then look for Excel. The screen will look roughly as shown in Fig. 1.4.
The Excel window shown in Fig. 1.4 is where you will analyse your data. If you are not experienced in Excel, take time to note the elements shown in Fig. 1.4.

You instruct Excel by using the menus or the toolbars. For example, to open an Excel file, click on the File menu, and then on Open.

Excel also offers toolbars, and they provide a one-click access to many of the Excel facilities. For example, in Fig. 1.4, clicking on the icon that looks like an open file is the same as using the menu item File, Open.

Excel documents are called workbooks and each workbook is made up of individual worksheets. In Excel 2003, each worksheet can have up to 256 columns (labeled A, B, C and so on) and 65,536 rows\(^4\) and a single workbook can have up to 255 worksheets. In Excel 2007 each worksheet can have up to 16,384 columns and 1,048,576 rows and the

---

\(^4\) If these dimensions are limiting, you should probably be using other software anyway. Most statistics packages do not have such limits.
number of worksheets in a single workbook is limited by the memory available on your own PC.

1.4 A sample workbook

For illustration an Excel workbook that contains data from the 1997-98 national household survey conducted by the Central Bureau of Statistics (CBS) Kenya, is used. We have just used some of the data from a single district called Murang’a in the Central Province.

This section reviews how you can open a workbook and navigate round it.

We assume that you are in Excel\textsuperscript{5}. If so then:

1. Click the Open button on the Excel toolbar. The Open dialogue box appears, looking something like Fig. 1.5.

2. Click on the drop-down list to locate the folder with the Excel file to open, called AllData.xls.

3. Either double-click on the workbook to open it, or click its name and then click on the Open button. The selected workbook, see Fig. 1.6, will open into Excel.

\textsuperscript{5} If Excel is not already open, then an alternative is to use Windows Explorer. Find the data file and double-click on it, to open Excel with the file ready for use.
There are three ways to move from sheet to sheet in the workbook. Practice each of these ways.

a. Click on the tab at the bottom of the sheet, to move to the one you want.

b. Click on the worksheet navigation buttons.

c. Right-click on any of the worksheet navigation buttons, and select from the drop-down list of sheet names that will appear.

1.4.1 Freezing panes

Within any worksheet you can use the vertical and horizontal scrollbars (see Fig 1.6) to look at all the data. However, notice that as you move down the rows the top row showing the column names disappears, so you no longer know which data each column contains. To keep the top row always visible do the following: position the mouse cursor in cell A2 and from the menu select Windows > Freeze Panes. You may want to do this in all worksheets.

Now check that you can answer the following questions:

d. From the worksheet labeled “Corehead”, how many households were included in this survey?
e. How many rows of data are there in the three sheets called “Corehead”, “Expenditure” and “Agriculture”. If they are not the same number, can you think why this should be?

f. Excel names the columns of data as A, B, C, etc. What is the name of the last column with data in the “CorePerson” sheet? So how many columns of data are there in this sheet?

g. How many rows of data are there in the sheet called “CorePerson”? How could you use this information (together with your answer in a.) to give an idea of the average number of people per household in the district?

1.5 Basic workbook operations

We continue to use the workbook that was introduced in the previous section. This section reviews how to add and delete sheets, and how data can be copied and moved.

In this section and beyond it is important that you understand the logic of the task. Do not simply try to remember the instructions. In Windows there are often different ways of doing the same thing and these will soon become automatic for the common tasks.

The first task is to add a new sheet to the workbook. This is so you can then copy some data to this new sheet for your working.

1. Use Insert > Worksheet to insert a blank worksheet.

2. In the new worksheet, go to the tab at the bottom of the screen and right-click Sheet 1. You should get a popup menu as shown in Fig. 1.7.

Fig. 1.7 Popup menu gives basic operations on the current sheet
3. Click on the option to **Rename** and call the sheet “Temp” to signify you will use it for your temporary work.

4. Now practice copying and pasting. Return to the sheet called CoreHead. Make the top corner cell, \(<A1>\) the active cell by clicking on it and then drag the mouse down to the cell \(<D14>\). If things go wrong, then use the \(<\text{Undo}>\) button and try again.

5. Press the CTRL key. With CTRL still pressed, put the cursor on the cell \(<F1>\) and drag the mouse down to the cell \(<G14>\), see Fig. 1.8. Then use \(<\text{Edit}>\)<Copy>\(^6\) to copy the selected parts of the sheet to the clipboard.

![Fig. 1.8 Selecting parts of a worksheet](image)

6. Return to the new sheet that you called “Temp” and use **Edit > Paste**\(^7\). This inserts the values of the copied cells into the new sheet.

7. Now we show a different way of pasting. Start by using the **Undo** button to remove the contents you have just pasted. Then use **Edit > Paste Special**, rather than just the simple Paste. Click on the Paste Link as shown in Fig. 1.9.

8. The new worksheet contains pointers back to the original data. This means that the data are not duplicated, and any correction in the original data will automatically change the working copy.

---

\(^6\) Or use the Copy button on the toolbar, or press \(<\text{Ctrl}>\)<C> or \(<\text{Shift}>\)<Insert>; pressing a combination of keys is known as a shortcut.

\(^7\) Or use the Paste button on the toolbar, or use \(<\text{Ctrl}>\)<V>.
We now show one way of moving a range of cells.

1. In the sheet called **Temp**, select the data in the first four columns.
2. Move the mouse to a border of the area that is selected. The pointer will change from a + to an arrow.
3. Drag the selected area down 3 rows and across one column, as shown in Fig. 1.11.
4. Click on any cell to deselect the cell range.
5. When using Windows there are often different ways of doing the same thing. Can you think of another way of moving a range of cells\(^8\)? Try to do this by pressing the Undo button, and then repeating the move.

You should regularly save your work when you make changes to a workbook. Excel and other Windows applications offer you two options for saving. You can save using the Save command, or by pressing <Ctrl> <S>, which saves the changes back into the original file. Or you can use the Save As command, which allows you to save your new work into another file with a different name.

1. Click on **File > Save As** to open the corresponding dialogue box.
2. Check which folder you are saving your file in, see Fig. 1.12. If necessary change the folder to the one you wish to use.
3. Give the file a new name, and click on the **Save** button.

*Fig. 1.12 The Save As dialogue*

Before finishing this section, delete the temporary worksheet that you created for your working.

1. Make sure that you are on the sheet called **Temp**.

\(^8\) For example you could try cutting the selected range, and then pasting them to their new destination.
2. Then use the **Edit** menu and then **Delete Sheet**. Excel will ask you if you really want to continue. Say that you do.

3. Note that the **Undo** button on the toolbar is now inactive. This is an operation that can not be undone.

4. Can you think of another way you could have done this task?9

### 1.6 Summary

This section has introduced some of the key aspects that apply whatever field of application you use for Excel.

If you have used Excel before, then most of the material in this chapter should have been familiar. However, Excel is a large package and existing users may have found some points that were new, and will help them to use the software more effectively.

One feature of Windows is that there is usually more than one way of accomplishing the general tasks that have been introduced in this chapter. It is much more important to understand the logic of the task, than to remember the steps associated with one way of doing the work. Always use logic and not memory.

A key feature of Windows is that once you have mastered any application, then others have some similar features. So, once you are confident in your use of Excel, this should also help with your use of other software.

---

9 You can right-click on the tab at the bottom of the sheet. This gives a popup menu and one option is to delete the sheet.
Chapter 2: Data manipulation

2.1 Introduction

This chapter looks at features of Excel that are particularly needed for handling data prior to analysis. Excel is an “all-purpose” tool and to use it effectively for statistical work, you must apply the same discipline that would be automatic when using a statistics package. We explain here what we mean by this, using the data set introduced in chapter 1 of this guide.

One element of “good practice” is to avoid making any worksheet too complicated. It is better to keep worksheets simple and give them names that show clearly what they contain. This makes it easier to find a particular set of data, or a table of results. We now show the basics of Excel, including:

- How to keep your data in what Excel calls a “List”.
- How data can be entered and checked in Excel.
- How names are used in Excel and the value of adding comments to a cell.
- How to use special features that only work when the data are in the form of a list, including filtering, and sorting.

2.2 Structure of your data

For effective statistical work you need an efficient structure for holding, managing and analysing the data. This is usually obtained through organising the data in rectangular arrays of rows and columns.

- Rows contain all the information collected about a single ‘individual’ or ‘unit’ being studied: the ‘Cases’
- Columns contain different measurements or information about each Case: the ‘Variables’

The first row of the list contains a label for each column. Fig. 2.1 shows part of the data from the 1997-98 household survey stored in the “Corehead” worksheet. Seven columns are shown, and just the first 15 of the 330 household heads in the district are shown here.

---

10 Even more important for survey work is that Excel’s powerful facility for tabulation (Pivot Tables) requires data to be in the form of a list. This is in a separate guide.
Fig. 2.1 Data from the CBS survey is in list form.

These rectangular structures are called ‘Lists’ in Excel.

- **Lists** make it easy to exchange data with other software such as databases and statistical packages.

- **Lists** give structure to the way data are handled. In particular, they encourage the user to store information of the same kind within columns. So a column with numbers should not contain text, and a text column should not include numbers that will need summarising.

When you set up a list:

- Where possible, put each list into its own worksheet and give the worksheet an appropriate name. This makes it easier to find your data.

- If you need to put other data into the worksheet, leave at least one blank row and one blank column between the list and other data.

- Do not have blank columns or blank rows within the list. In particular, do not use a blank row to separate the column labels from the first row of data. You may find it helpful to distinguish the column labels from the data by using bold face or a different background colour.

- Use column labels that make it easy to understand the meaning of the column.

- Make the top row visible at all times by **Freezing Panes** (see section 1.4.1).
2.3 Master copies and working copies

To preserve the integrity of the data, it is good practice to create a ‘mastercopy’ from which you can retrieve the original data should you make a mistake which overwrites the original data records.

- Return to the worksheet named Corehead and select Tools > Protection > Protect Sheet. Do not type a password, but un-tick all tick boxes (except the very top one) as shown in Fig 2.2, to prevent users selecting any cell. Then press OK.

Fig. 2.2 the Protect sheet dialogue box

Now protect the worksheet named CorePerson. You are now unable to make any changes to the protected worksheets.

So how can you do further data manipulation on these lists? The solution is to make a working copy of the locked worksheets.

- Make the Corehead worksheet active by clicking on the name tab:
- Right-click on the “Corehead” sheet name tab, select Move or Copy, tick Create a copy and select (move to end). The dialogue box is shown in Fig 2.3

2.3.1 This action inserts a new worksheet named Corehead (2) into the current workbook.

- Repeat this action with the worksheet named CorePerson to create a working copy of this worksheet, which is automatically labelled CorePerson (2).
Finally, unprotect both working copies (Corehead (2) and Coreperson (2)) using Tools > Protection > Unprotect sheet.

2.4 Sorting

Sometimes datasets are easier to check if they are sorted. In this section we explain how to sort datasets in ascending and descending order, as well as how to add a column of data that is a regular sequence of integer values.

2.4.1 Entering regular sequences of values

1. On the Corehead(2) sheet insert a new column before column A: position the mouse cursor in column A and use Insert > Column. Give cell A1 the name Index.

2. Type the number 1 in cell A2 and the number 2 in cell A3.

3. Select cells A2 and A3. Notice the small black box at the lower right hand corner of the selected cells; this is called the fill handle.

4. Move the mouse over the fill handle, so the pointer changes to a small +. Double-click on the fill handle: this fills the rest of the column with a regular sequence of consecutive integers up to the number 330. This process is called auto-fill.

5. Now replace the number 2 in cell A3 with the number 3. Select the range A2:A3 and double click on the fill handle to complete the auto-fill action. The result is shown in Fig. 2.4 Notice that the subsequent cells are filled with the numbers 5, 7, 9 etc. In other words, they differ by 2, which is the same as the difference between the two numbers that we selected to start the sequence.

Fig. 2.4 Entering a regular sequence of integers
6. Now select the cell range A2:A4. Press and hold the Ctrl key (notice there are now two black crosses) and drag the fill handle to the end of the column. The result is shown in Fig. 2.5. The subsequent cells now repeat the pattern of the selected cells (1, 3, 5, 1, 3, 5).

Fig. 2.5 Entering repeats of a regular sequence

![Excel screenshot showing repeated sequence in cells A2 to A16]

Many other options for creating regular sequences are available using the menu item Edit, Fill. You may like to experiment with them.

7. Finally, overwrite whatever sequence of numbers you last entered in column A with integers from 1 to 330 in steps of 1.

2.4.2 Sorting columns

Sometimes features of the dataset are easier to check if the dataset is sorted, say, on the age of the household head. To do so:

1. In the sheet named “Corehead(2)” place the mouse cursor inside column L which indicates the age of the head of the household. Then click the Sort ascending icon on the toolbar menu. How old is the youngest household head? How many household heads are of that age?

2. Next click on the Sort descending icon on the toolbar: how old is the oldest household head? How many household heads are of that age?

So far we have used only one column for sorting, but Excel offers sorting by up to 3 columns at a time. Should you need a more powerful sorting feature, Excel 2007 offers sorting by 64 columns at a time.
3. Use **Data > Sort** and complete the dialogue box as shown in Fig. 2.6 to sort by Cluster, day, and Household number in this order. The sorted dataset is shown in Fig 2.7. Look at column F, labeled ‘day’: when did enumeration start and end in the first cluster? (dates are listed in day/month/year format, e.g. 29497 is the 29\textsuperscript{th} of April, 1997).
Now put the data back into their original order. Fortunately you constructed an index column (column A in Fig. 2.7) which used to be in ascending order. So you can sort on this column to get back to the original order. Alternatively, as you have just done the sorting, you could use \textbf{Edit > Undo Sort} or \texttt{<Ctrl><Z>} or click on the undo arrow.

\textbf{2.5 Selecting subsets of your data (Filters)}

\textbf{2.5.1 Autofilter}

The Autofilter facility in Excel is a useful tool for examining your data on the fly and for selecting subsets of rows ("Cases"). Autofilter only works on data that is in list format.

To activate the Autofilter in your data, select the "Corehead(2)" sheet:

1. Click any cell in the list.

2. Select the menu item \textbf{Data > Filter > AutoFilter}. This inserts a button with a downward arrow in each of the column headers as shown in fig 2.8

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2_8.png}
\caption{The active Autofilter showing a list of unique values of age}
\end{figure}

3. Click the arrow at the top of the column L, labeled Q1.4, that is AGE of the respondent. You will get a drop-down list as shown in Fig. 2.8, giving several options:

- \textbf{(All)} – Displays all cases. At present this will have no effect as we are currently displaying all of them. However, this option is used to remove any filter that has been applied to this column.
• (Top 10…) – This allows you to display only the cases with the 10 highest numerical values. In fact you can choose to display a different number of values or a percentage, or the lowest few values (e.g. highest 5, lowest 3, highest 10 percent). So this option is more flexible than the name “Top 10” implies. Note that this option does nothing if all the values in the column are non-numeric.

• (Custom…) – This allows you to perform more complex filtering. We give an example in section 2.6.2

• The custom filter will allow you to filter a particular value in the column – if you click on one of these, it will display only those cases that have that value.

• If your data contains missing values (blanks), you will be given options to display cases with blank values, or those with non-blank values.

4. Select the value 25 in the drop-down list. You should get 6 records of household heads that are 25 years old. Notice that the colour of both the downward arrow of the active Auto filter and the number of the visible row has changed from black to navy blue.

Now let’s see the combined effect of more than one active filter.

5. Click on the downward arrow of column J, labeled Q1.1, that is SEX of the respondent and click on number 1 to show only males household heads, of age 25 years. There should be 4 of them.

You can use Edit > Copy or <Ctrl><C> to copy the selected cases and paste them into another worksheet for further calculation. This is important because, unfortunately, most other Excel facilities to compute summary statistics ignore any filtering that you have done, and act on all the cases in the list whether filtered or not.

2.5.2 Custom filters

Figure 2.9 shows an example of the Custom filter.

6. Select (All) in both the AGE (column L) and SEX (column J) columns to return to displaying all cases in the list.

7. In column AGE, click on the downward arrow and select (Custom…) to get the Custom Autofilter dialogue box. Complete this dialogue as shown in Fig. 2.9 to select only those household heads of age between 20 and 25 years. There should be 18 such cases.
2.6 Referring to data cells using names: named ranges

Using Excel to perform calculations (as described in the next section), you usually need to specify which data cells are to be used. There are various ways in which this can be done.

Each cell in a worksheet can be referred to by its column letter and row number. For example, G14 refers to the cell in the 14th row of column G. G14 is called the address of that cell. You can also refer to a rectangular block of cells by giving the addresses of the top left and bottom right cells in the block, separated by a colon. For example, C4:D5 refers to the block of cells C4, C5, D4 and D5.

An alternative and more convenient way of referring to a cell or block of cells is to give it a name. Names are easier to remember than cell addresses, and it is a good idea to choose names that describe the data in the cell(s). For example, you might name a cell “Total” which contains the sum of a number of values. You can then refer to that cell as ‘Total’ without needing to worry where it is in the worksheet or even in which worksheet, because names are available throughout the entire workbook.

Define a name that refers to the entire list with the information about household heads:

- Make sure that the worksheet “Corehead (2)” is displayed.
- Select the entire list, that is, the cell range A1:BT331.* When selecting cell ranges make sure that you define only the region that contains data, as you may need the other blank cells for future calculations. If, for example, you define an entire column as your cell range and then later perform calculations in that column, your results or answers may become integrated into previous calculations, ultimately miscalculating the data.
• Use the dialogue box from the menu selection **Insert > Name > Define.** Give this area the name **corehead_data** (spaces are not allowed in names) and press the **Add** button: the dialogue box should look like the one shown in Fig 2.11

![Fig.2.11  Define Name dialogue box](image)

- **Fig.2.11  Define Name dialogue box**

- Now move to a different worksheet; then press the downward arrow in the Name box (the little box on the top left hand corner of the sheet, just below the toolbar Save icon). This shows a list of named ranges that have been defined in the workbook; select **corehead_data** to highlight the cell range from A1 to BT331. Notice that the small box to the top left of the worksheet now displays the name **corehead_data.** It will do that every time you select that data area.

  It would also be useful to name the data in each column with the column name in this worksheet. You could do it by using **Insert > Name > Define** repeatedly, but this can be very tedious and it is easy to make a mistake.

  A quicker way is as follows:

  - While the **corehead_data** named range is highlighted, use **Insert > Name > Create.** You will see a small dialogue asking where the names are. Excel has guessed correctly that they are in the top row of the data. Click OK.

  - Excel has now created a name for each data column (starting in row 2) corresponding to the column header in row 1.

  - Click on the downward arrow next to the Name box. This gives you a list of all the names that you have created. You can click on any of these names to select the relevant column. Try it!

Thus, some named ranges correspond to a block of cells, others indicate a single column. Naming single columns is a powerful shortcut to speed up computation, because it avoids having to manually select the cell range to be used in calculations.
2.7 In summary

You should now be able navigate through basic formatting in Excel, (how to keep your
data in lists, how to sort, filter, save, copy, etc…). This knowledge is integral to
understanding and dealing with Excel for statistical work, and will help when learning the
more advanced chapters in this guide.
Chapter 3: Calculations

3.1 Introduction

This chapter introduces some of the calculation functions of Excel, including:

- how to perform simple arithmetic calculations using Excel’s formula bar,
- how to use Excel as a column calculator,
- how to use column names instead of cell address to perform calculations

3.2 Formula Bar

3.2.1 Arithmetic operators

Excel's formula bar works like a calculator, and can also carry out numerous operations using hundreds of different formulae. To perform basic calculations, enter = into the formula bar, followed by the calculation. e.g. =8+2. Excel will calculate the figures, and return the answer of 10 into the cell that you have highlighted. Any expression starting with = is a formula. You may also use the forward slash, /, for division, as well as, +, -, for plus and minus, * for multiplication and ^ for exponentiation. Excel has more inbuilt functions to perform complex operations, with a dedicated facility called the Function wizard that allows you to choose from a long catalogue of functions subdivided by category.

To try out the Function wizard, highlight cell A333, then in the formula bar (on the right hand side of the Name box), press the Insert function button labelled fx. From the category list, select Math & Trig and browse through the functions that Excel offers until you reach SQRT, which stands for Square Root, then click ok. A dialog box will appear that will calculate the square root of any number for you. Try it out by entering 49, excel will return the answer of 7 to cell A333. If you look in the formula bar, it will read: =SQRT(A333) informing you that it has calculated a square root and returned the answer to cell A333, now click undo.

The formula bar can calculate straight figures, but it can also perform calculations using cell addresses. If, for example, you had the number 10 in cell A2 and the number 15 in cell C2, you could calculate the sum of the cells by entering, =A2+C2. Excel will return the answer 25.

3.2.2 Arithmetic operators using cell addresses

Calculate the average number of people per room in the main dwelling:

- In the Corehead (2) sheet, locate the column with entries for Q4.2, labelled “number of people sleeping in the main dwelling”. This should be in column AS. The adjacent column AT should have entries for Q4.3, labelled “number of rooms in the main dwelling”.

• Insert a blank column after Q4.3 by right-clicking on the column to its right (AU) and selecting insert, and label it “people per room”. In the cell immediately below it type \[ \frac{AS2}{AT2} \] then press Enter. Format the cell to show two decimal places by right clicking on the cell and selecting **Format cells**, then in the number category set your decimal places to two.

• Next double click on the fill handle to auto-fill the rest of the column (refer to section 2.4 to refresh your memory about this). The result is shown in Fig. 3.1

![Fig. 3.1 explicit division using relative cell reference](image1)

Fig. 3.1 explicit division using relative cell reference

![Fig. 3.2 using a comparison operator](image2)

Fig. 3.2 using a comparison operator

3.3 **Comparison operators**

Insert a blank column on the right of column AU and label it “crowded”. In the next cell below type \[ AU2 > 3 \] and press Enter, then double-click to auto-fill the rest of the column.

Using the ‘greater than’ sign yields a logical result which is FALSE if the ‘average number of people per room’ is not larger than 3, and TRUE otherwise, as shown in Fig 3.2. Are there any crowded dwellings according to this arbitrary definition?

3.4 **Using column names instead of cell addresses**

Now clear the cells in the column labelled ‘crowded’ without deleting the column: click on the AV letters and press <Delete>.

Now select the `corehead_data` range from the Name box list and use the menu selection **Insert > Name > Create** to create names for each individual column in the range. In the Name box drop-down list check that there is a named range called ‘people_per_room’
Label the blank column to the right of column ‘people per room’ ‘crowded’. In the first blank cell now type \(\text{people per room} > 3\) and press Enter. Finally, auto-fill the rest of the column.

This will yield the same result as in the previous section\(^\text{12}\) but we have now used column labels instead of cell references, as introduced in section 2.7.

Fig. 3.3a shows the results from the computations on this page and Fig. 3.3b shows the actual content of the cells that have been auto-filled.

![Fig 3.3a showing the results of formulae](image1)

![Fig 3.3b showing the formulae in cells](image2)

### 3.5 Functions

- Still in the Corehead (2) sheet, locate the columns labelled Q3.11 to Q3.18 which store the “time taken to reach the nearest …” service, from food market and public transport to schools and doctors. It is of interest to find the average time taken to reach any of these eight services.

- To do this, insert a blank column on the right of column Q3.18 and label it ‘average time’.

You could now enter a formula that spells out the mean by typing \(\frac{\text{AJ2}+\ldots+\text{AQ2}}{8}\) but as this is such a frequent operation, Excel has inbuilt routines, called functions, that will do it.

\(^{12}\) If the cells display #NAME? as the result of this formula, from the main menu use Tools > Options. Go to the Calculations tab and tick the box to Accept labels in formulas. Then re-type the initial formula and auto-fill the rest of the column.
First use the SUM function: type \[ \text{SUM}(AJ2:AQ2) / 8 \] and press Enter. Format the cell to show 1 decimal place.

You can bypass the SUM function by using the AVERAGE function for computing the mean. Replace what you typed earlier with \[ \text{AVERAGE}(AJ2:AQ2) \] press Enter and finally autofill the rest of the column. You get the same result as in the previous paragraph. This is shown in Fig. 3.4.

![Fig. 3.4 computing the mean using a statistical function](image)

You can also use the Function Wizard to locate such operations as SUM and AVERAGE, by Clicking the fx button and browsing through the functions until you find them.

We will now take a closer look at the Function Wizard. Click on the fx button to activate the Insert Function dialog box.

- Select a Category, choose the Statistical category and press the letter M on the keyboard to cycle through all those statistical functions beginning with ‘M’, as you scroll over each function a description of it will appear below. The Function wizard box is shown in Fig. 3.5.

- Find the functions that will give you the minimum and the maximum average time from column AR. What are their names?

- A colleague asks you to compute the median time: can you find the relevant function? Where would you click for help on this function?

There is a help button in the bottom left hand corner of the dialogue box, which offers succinct help on the currently selected function.
3.5.1 Using names in formulae and functions

Return to the column labelled ‘people per room’ and insert a blank column on the right: label it ‘tophalf’

- Move to the cell below the label in row 2 and type:

  =people per room>median(people per room). Now auto-fill the rest of the column. In this way we have identified with TRUE those households with more than the median number of people per room of the dwelling, and with FALSE those households with a number of people per room in the dwelling less than or equal to the median.

But what is the median number of people per room for the 330 households in this dataset?

- Move to the second blank row (row 333) at the bottom of the column labelled ‘people per room’ and type =median(people per room). The result is 1.23 (see Fig. 3.6a), or about one and a quarter of a person per room.

And how does the median compare with the mean?

- Move to the row 334 at the bottom of the column labelled ‘people per room’ and type =average(people per room). The result is 1.35 (see Fig. 3.6a), i.e. one and a third of a person per room.
3.5.2 Combining functions

The ability to combine functions is a powerful feature of Excel. Let’s start with a simple example: what is the percentage of households with more than two people per room?

Move to row 335 at the bottom of the column labelled ‘people per room’. We will first use the COUNTIF function, this counts the number of cases that conform to your specifications. To count out how many households have more than 2 people per room type: `=COUNTIF(people per room,">2")`.

Next combine this with the calculations needed to obtain a percentage count, the final formula will look like this: `=100*COUNTIF(people per room,">2")/COUNT(people per room)`.

The answer is 10.6% (see Fig. 3.6a).

Fig 3.6a  Fig 3.6b

![Image of spreadsheets showing formulas]

Fig. 3.6b shows what has been typed in each cell to obtain these summaries of columns. You can get this view of your worksheet by pressing <Ctrl> and the ¬ or ~ key.

3.6 In Summary

In this chapter we have explored how to use Excel for various calculations. You should now be able to use Excel for simple arithmetic calculations as well as more advanced column calculations. Keep exploring the program, as there are other valuable formulae in Excel that will help you to organise, format and understand your data with clarity.
Chapter 4: Further Calculations

4.1 Introduction

This chapter introduces some of the more advanced operations that can be achieved, including:

- Calculating a weighted mean
- Using functions for coding
- Attaching labels to numeric codes

4.2 Weighted means

The overall mean found in section 3.5.1 is unweighted; that is to say it has been calculated using the same identical weight of 1 for the mean of each individual household. Using identical weights for each mean is called the unweighted mean.

Alternatively, to compute a weighted mean with weight proportional to, say, the number of people in each dwelling, we pre-multiply each household mean by the number of people in the household, or weight, and divide by the sum of the weights.

This can be done in two steps as follows:

First select the data area named corehead_data in the name box, and then create names from all labels in the top row as shown in section 2.7

Now move to the blank row 336 in the column ‘people per room’ and type:

=SUMPRODUCT(Q4.2, people_per_room)/SUM(Q4.2) (Note the underscores)

The answer is 1.69 (see Fig. 4.1a), i.e. one and two thirds of a person per room, giving one third more of a person per room than the unweighted mean.

Fig. 4.1b shows what has been typed in each cell to obtain these summaries of columns. You can get this view of your worksheet by pressing <Ctrl> and the ¬ or ~ key.
4.3 Using functions for coding

Not only can functions be combined, as seen in section 3.5.2, but functions can also be used as arguments within other functions. In this section we present some examples that are useful in the context of coding.

4.3.1 Coding numerical ranges with IF

Return to the newly computed column labelled ‘average time’ (see Fig. 4.2). Use AutoFilters to check the range of this column, which is between 1 and 7, and assume it is of interest to code numerical values in two groups with a cut-off point at 4.

Insert a new column on the right hand side of column AR and call it ‘average time coded’. In the first blank cell type =IF(AR2>4,”above 4”,”up to 4”) and autofill the column. The first argument of IF performs a logical test on the condition, the second argument is returned if the test is true, the third argument is returned if the test is false.

Nesting functions within functions: nested IFs

We can also arbitrarily set two cut-off points at 3 and 5 to make three groups. This can be done using an IF function as argument within IF, also known as ‘nested IFs’. Undo the previous function that you established above, and replace it with:

=IF(AR2>5,”above 5”,IF(AR2>3,”>3 to 5”,”up to 3”))

and then autofill the column.

We can also nest several IFs: to make 4 groups with cut-off points at 3, 4 and 5. Undo your previous work and input:

=IF(AR2>5,”above 5",IF(AR2>4,”>4 to 5",IF(AR2>3,”>3 to 4","up to 3")))

Now autofill the column. The result is shown in column AS in Fig. 4.2.
4.3.2 Coding numerical ranges with AND

It is also possible to attach a label that identifies values within a predetermined numerical range. Say we want to single out numerical values between 3 and 5. Insert a new blank column and label it “coded range” and in the first blank cell below type

\[ =IF(AND(AR2>=3,AR2<=5),"3 to 5","") \]

and then autofill the column. The result is shown in column AT in Fig. 4.2. Note that we could also explicitly type a label for values <3 or >5 instead of typing double quotes to return a blank.

4.3.3 Coding text with OR

We can even recode cells with text, a handy feature for when we have no access to the actual numbers but only to the derived value labels. If we want to recode column AS, ‘average time coded’ into three groups with a cut-off at 3 and 5, we can set multiple criteria as the logical test using the function OR within IF as follows.

Insert a new column labelled ‘re-coded labels’ to the right of the column named, ‘Coded range) and in the first blank cell type

\[ =IF(OR(AS2=">3 to 4",AS2=">4 to 5"),">3 to 5",AS2) \]

and then autofill the column. Note that the last argument to this function copies the original value labels. This is shown in column AU in Fig. 4.2

4.3.4 Coding using text functions

We can quickly recode the two middle groups in column ‘average time coded’ into one group using a single condition, because both labels begin with the same character sign. Insert a blank column labelled ‘re-coded text’ to the right of ‘re-coded labels’, and in the first blank cell below type \[ =IF(LEFT(AS2,1)=">",">3 to 5",AS2) \] and then autofill the column. The result is shown in column AV of Fig. 4.2.

Text functions are especially useful when we do not have access to the actual numbers but just to the derived value labels. Create a new column to the right of ‘re-coded text’ and name it Cut-off 3. To make two groups with a cut-off at 3, input:

\[ =IF(RIGHT(AS2,1)="3",AS2,"above 3") \]

and then autofill the rest of the column. The result is shown in column AW in figure 4.2

4.3.5 Coding using a dynamically computed cut-off

Finally we show how to code numerical values using a cut-off dynamically computed on the same column. Insert a blank column to the right of column, ‘Cut-off 3’ and label it, ‘below or above average’. In the first blank cell type

\[ =IF(AR2>AVERAGE(average time),"above","below") \]

and then autofill the rest of the column. The result is shown in column AX in Fig. 4.2
4.4 Attaching labels to numeric codes

Now move to the worksheet data_dictionary. A data dictionary generally stores two strings of text- one called ‘variable label’, which gives an intelligible explanation of the information stored in a column, whose full name is usually abbreviated for convenience, and another called ‘value label’, which defines the meaning of the numerical values stored in a column, e.g. 1 for male and 2 for female.

For instance, what is stored in column Q0.7? We can see in Fig. 4.3 that the variable Q0.7 is labelled “is the respondent the head of the household?” and the value label for 1 is ‘Head’ and the value label for 2 is ‘Other member’.
Can you see what is stored in Q1.1 and what is represented by the values 1 and 2? These variable and value labels are very useful for obtaining output ready for presentation like graphics and tables.

Statistics packages have special features to assign value labels to numerical codes in the same data column. In Excel we must duplicate the information in the list by inserting a new column that shows (ideally side-by-side) the value label that corresponds to the integer stored in the corresponding row.

4.4.1 A simple example

In the worksheet Corehead (2), locate the data column labelled RURBAN. Use Data > Filter > AutoFilter to find out how many unique numerical codes there are in this data column. How do we know what these numerical values correspond to? Refer to Fig. 4.3 to see that RURBAN is labelled as ‘Urban or rural cluster’ and the value 1 indicates a rural cluster and the value 2 an urban cluster, so now we can assign text to the values.

Insert a name that defines the cell range A22:B23 in the worksheet data_dictionary and call it ‘rurban_label’.

Next return to the Corehead (2) worksheet, insert a blank column on the right of RURBAN and label it ‘RURBAN: urban or rural cluster’. In the first cell below type \(=\)VLOOKUP(G2,rurban_label,2) and autofill the rest of the column. Use the AutoFilter to make sure that there are urban clusters.

Notice we gave 3 arguments to this VLOOKUP function:
- lookup_value (here cell G2),
- table_array (here the rurban_label name)
- and col_index_num (here the number 2, since the value labels are stored in the second column of the named table).

### 4.5 In summary

Make sure that before continuing on you thoroughly understand the calculations we have performed. You should able to: calculate a weighted mean, use functions for coding, and attach labels to numeric codes.
Chapter 5: Multi-level data

5.1 Introduction

Survey data are often recorded at different levels of resolution. For example, in a household survey some information is recorded for the household, (e.g. type of dwelling) while other information is recorded for each individual living in the household, (e.g. age of individual). This is known as having data at multiple levels or multi-level data.

In the current workbook we have an example of multi-level data, because the Corehead worksheet stores information at the household level, while the CorePerson worksheet stores information at the individual member (of the household) level, as shown in Fig. 5.1.

Fig 5.1 showing data collected at multiple levels

<table>
<thead>
<tr>
<th>Household level</th>
<th>Household member level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

With multi-level data it is often of interest to transfer information between levels, for example, replicate household data at the individuals’ level or summarise individuals’ data at the household level. It is also often used to merge certain results.

This task is known as match-merging, because the merging operation requires new data rows to be matched to existing data rows that are identified by a unique number. Match-merging is handled by mainstream statistical packages via their data management features once unique identifiers are set up. Excel requires more discipline, but much can be done by using spreadsheet functions efficiently.
5.2 Setting up a unique identifier for households

Our datasets have no unique household identifier, because the household number is unique only within a cluster. So we must create it explicitly by concatenating the cluster number with the household number in both the household and person level datasets.

Move to the Corehead (2) worksheet and insert a new column in place of the current column A and label it ‘HH_ID_corehead’. In cell A2 type \[=\text{VALUE}(E2\&F2)\] and press Enter, then autofill the rest of the column. This signs & concatenates two numbers as text instead of adding them up and the VALUE function converts text to its numerical value. Finally sort the newly added column in ascending order. This is necessary because the VLOOKUP function that we use in section 4 requires the leftmost column of its lookup ‘table’ to be numeric and sorted.

Now do the same in the CorePerson (2) worksheet and label the new column ‘HH_ID_coreperson’. In cell A2 type \[=\text{VALUE}(D2\&E2)\] Also sort this newly added column in ascending order.

The resulting identifier columns are shown in Fig 5.2 a and 5.2b.

Insert the name ‘HH_ID_coreperson’ that defines the cell range $A2:$A$1783 in the CorePerson (2) worksheet. Also insert the name, age_coreperson that defines the range $N2:$N$1783 in the CorePerson (2) sheet, currently storing the age of each individual.

Refer to section 2.7 if you to review the process of creating names.
5.3 From one to many: replicating information

Wanting to attach the household level information to each of the members in the household, we need to replicate this information because there are several members in a single household. We have already seen an example in section 4.4, when we attached value labels to numerical codes.

As the data at the household level is at a higher level of hierarchy, we have a single record per household that needs replicating for as many rows as there are individuals belonging to that specific household. To replicate the same information for each member belonging to the same household we use the household dataset as the ‘lookup table’ argument to the VLOOKUP function.

In section 2.7 we defined the name corehead_data. Use the name box to locate it and make sure that the leftmost column called HH_ID_corehead is sorted in ascending order.

Move to the CorePerson (2) worksheet, insert a blank column in place of column B and label it ‘HH_head_age’. This is the column in which we want to replicate the age of the household head, which is stored in column N in the Corehead (2) worksheet. Move to row 2 and type \(=\text{VLOOKUP}(A2,\text{corehead_data},14,\text{FALSE})\) then autofill the rest of the column.

The result is shown in Fig. 5.3. Check that the first few non missing values replicate the age of the household head correctly, by cross-reference to the Corehead (2) worksheet.

![Fig 5.3 showing data from the household level replicated at the person level](image)

What do you notice about the first household? #N/A shows as the result of using the argument FALSE to force VLOOKUP to find an exact match of the household identifier. This means that the household in question is missing from the household level data.
Try inserting another blank column indicating the gender of the head of the household. You could copy paste the VLOOKUP formula and change the first argument back to A2. Then amend the third argument as needed by finding the column number in the corehead_data name that corresponds to the label for gender, (column L).

5.4 From many to one: aggregating information

It is often useful to aggregate data from a lower to a higher level, and then merge with the data collected at a higher level. Here we summarise data collected at the person level, stored in the CorePerson (2) worksheet, and then merge the results with the data collected at the household level, stored in the Corehead (2) worksheet.

5.4.1 Conditional counting

While in the Corehead (2) worksheet, insert a new blank column on the right hand side of HH_ID coreperson and label it ‘count of HH members’. In cell B2 type =COUNTIF(HH_ID_coreperson,A2) then autofill the rest of the column. The result is shown in Fig. 5.4.

Fig 5.4 showing data from the person level aggregated at the household level

This conditional formula has counted how many members have been enumerated per household surveyed. Check that the COUNTIF function has returned the right result by summing the frequencies of the first few households in the worksheet Coreperson (2).
5.4.2 Conditional summing

In the Corehead (2) worksheet, insert another new blank column on the right hand side of ‘count of HH members’ and label it ‘mean age of HH’. In cell C2 type

=SUMIF(HH_ID_coreperson,A2,age_coreperson)/B2

and autofill the rest of the column. The result is shown in Fig. 5.5. This conditional formula has summed up the ages of members in individual households and then averaged it\(^{13}\).

---

**Fig 5.5 showing the mean age from the person level merged with the household level data**

<table>
<thead>
<tr>
<th></th>
<th>HH_ID_corehead</th>
<th>count of HH members</th>
<th>mean age of HH</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>633</td>
<td>4</td>
<td>23.3</td>
</tr>
<tr>
<td>3</td>
<td>641</td>
<td>4</td>
<td>19.8</td>
</tr>
<tr>
<td>4</td>
<td>653</td>
<td>7</td>
<td>22.3</td>
</tr>
<tr>
<td>5</td>
<td>673</td>
<td>2</td>
<td>25.5</td>
</tr>
<tr>
<td>6</td>
<td>679</td>
<td>8</td>
<td>13.8</td>
</tr>
<tr>
<td>7</td>
<td>685</td>
<td>3</td>
<td>24.7</td>
</tr>
<tr>
<td>8</td>
<td>691</td>
<td>1</td>
<td>39.0</td>
</tr>
<tr>
<td>9</td>
<td>709</td>
<td>3</td>
<td>8.0</td>
</tr>
<tr>
<td>10</td>
<td>736</td>
<td>4</td>
<td>14.0</td>
</tr>
<tr>
<td>11</td>
<td>741</td>
<td>5</td>
<td>24.0</td>
</tr>
</tbody>
</table>

---

5.4.3 Array functions for other summaries

Suppose we need to compute the median age per household? No dedicated MEDIANIF function exists, so Excel’s array functions must be used. These are typed just like other functions but we hold down the Ctrl+Shift keys before pressing Enter (also known as array-entering).

In the Corehead (2) worksheet insert another blank column and label it ‘median age of HH’. In the first blank cell in row 2, type

=MEDIAN(IF(HH_ID_coreperson=A2,age_coreperson))

then hold down <Ctrl>+<Shift> and press Enter. You should notice that Excel has enclosed the array-function in curly brackets (you do not type the curly brackets). Finally, autofill the rest of the column; the result is shown in Fig. 5.6.

Check that this array-function has returned the right answer by finding the median age value for a household with an odd number of members in the Coreperson (2) worksheet.

---

\(^{13}\) Excel 2007 has a new function called AVERAGEIF that eliminates the need to combine SUMIF with an explicit division.
Fig 5.6 showing the median age from the person level merged with the household level data

5.5 In summary

Sometimes you may want to show the actual values, perhaps sorted, and with summary results for different groups, as separate rows on the sheet. In this case you should make a copy of the data sheet, and the master copy, and then mix the data and results on the master copy(2) sheet. In this way you maintain original versions of your data for referencing and future work. You should now be able to: setup unique identifiers, and aggregate information, including conditional counting, summing, and array functions.
Chapter 6: Data Entry and Validation

6.1 Introduction
Excel is often used for small scale data studies, (larger projects are generally handled more effectively using data entry and validation software). We will now briefly go over the basic functions of Excel, which can be used for entering and validating data.

These Include:

- Entering data
- Validating data
- Auditing Data
- Missing values

6.2 Entering data
Excel is commonly used for entering data from small studies. For large surveys it is generally better to use software that is specially designed for data entry and verification. Hence data entry is not discussed here.\footnote{A guide called “Disciplined Use of Spreadsheets for Data Entry” describes the facilities available in Excel. This can be downloaded from \url{www.ssc.reading.ac.uk}}

6.3 Validating data
Excel offers facilities for restricting values that can be typed into a column, a process that Excel calls validation. To show how validation is used, assume that the ages of the household head are restricted to over 35 years of age. To set up a validation rule do the following:

- Make another copy of the Corehead worksheet, (as learned in section 2.4,) select the whole of the column S, labelled Q1.4, by clicking on the letter S.
- Use Data > Validation and complete the dialogue as shown in Fig. 6.1. Click OK. If you want to allow blank cells, untick the Ignore blank tickbox prior to clicking OK.
Now move to the bottom of column K and try entering ages below 36 years to make sure you cannot type such values.

Checks like these can be used to restrict values to data that have not been entered yet; However, these checks do not work with data that are not typed into the cells, like pasted, filled in by a macro, or as it is the case here, entered pre-validation. So another feature is needed.

6.3.1 Auditing: post-entry checks

Excel also offers facilities for checking values that have already been typed into a column, a process that Excel calls auditing. To activate the auditing features proceed as follows:

- Use Tools > Auditing > Show Auditing Toolbar and click on the button to Circle Invalid data. The result is shown in Fig. 6.2, and makes it easy to pick out the incorrect data values.

- You can remove the circles by clicking the Clear Validation Circles button. Note that a maximum of 256 circles will be shown at one time.
• The column label will always be circled unless we remove the validation, i.e. allow any value, on the whole of row 1.

• There is no need to highlight a specific column for auditing: provided the dataset is organized in a list then all columns that have validation rules will be checked for data entry consistency.

6.4 Missing values

Missing values often occur when data are collected. In general they can indicate that information was not available, something was impossible to measure, or the recorded information was lost or became impossible to read. In surveys they might occur because the respondent:

• Refused to reply

• Did not know

• The question was not applicable

Statistical packages allow you to specify that an item of data is missing by using a special character (often a star * or a dot .). Some packages allow you to specify more than one type of missing value, for example to distinguish between data that have been lost and data that were not applicable to that particular subject.

Unlike statistics packages, Excel does not have specific facilities to handle missing values. We recommend that you leave empty (blank) any cells that represent missing values. In general Excel will ignore these cells when performing calculations. DO NOT set missing values to zero, as it makes calculations incorrect. NEVER use the Excel
function =NA or the error code #N/A (“Not Applicable”), which cause all calculations involving that cell to return an error.

We also recommend adding a comment in the blank cell to indicate the reason for the missing value. You can add comments to any cell by using the menu item **Insert > Comment**, and typing a suitable message into the box. A small red triangle appears in the top right corner of the cell, and the message appears whenever you position the mouse over that cell.

### 6.4.1 Informative missing values

It is advisable to use special entry codes such as negative integers to represent missing values we can explain. We can then easily attach labels to these (see section 2.7) to indicate that the respondent, for example, ‘declined to reply’ or ‘could not recall’.

It is then easy to exclude such negative values from all subsequent calculations of summary statistics by specifying that the values to be included are positive. We discuss this in more details in a document on Excel’s Pivot Tables.

### 6.5 In summary

We can now see that with some diligent work, Excel can be used to perform data entry and validation processes on a small amount of data. You should now be able to: enter data, validate data, audit data, and deal with missing values.