

mathcentre community project

encouraging academics to share maths support resources

All mccp resources are released under a Creative Commons licence

**Getting Started with Data Manipulation in Microsoft Excel – Audio Transcript**

**Title slide**

Welcome to the ‘Getting started with data manipulation in Microsoft Excel’ tutorial.

**Overview**

In this tutorial, you will learn about some of the relatively simple functionality that Excel offers. If you would like to complete the steps in Excel yourself alongside this video, or would like to practice later, please download the Excel file entitled “Getting Started With Excel” that accompanies this video.

The video demonstration that follows, as well as the downloadable file, uses Microsoft Excel 2013. Other versions are very similar and hold the same functionality. However, versions prior to 2007 will have different menu bars.

This tutorial is aimed at complete beginners, so if you have never used Excel before, this is the video for you.

**Navigation**

Moving to Excel, we can see a small dataset that we will be using for all of our examples. It contains data relating to patients who attended appointments at a local hospital. We can see their name, their age, date of appointment and the clinic that they attended.

Data is usually stored in this format in Excel; where each row represents an observation, in this case a person, while each column represents a variable, in this case patient information.

Each sheet within an Excel spreadsheet is made up of cells. We can identify a cell via its column letter and row number. For example, this is cell D5.

The entire area that we can see here is known as a sheet and we can have multiple sheets within our Excel file. We simply tab between them simply by clicking these tabs at the bottom.

If we want to delete a sheet – *Sheet2* and *Sheet3* are empty – we would simply right-click on the sheet name and click *Delete*.

We can also rename our sheets. If I right-click on *Sheet1* here, *Rename*, and I’m going to rename that to *demo*.

We can see that the top row here gives us the names of the variables that we are using within our data set, and this is standard for Excel files, and many other statistical packages.

It is often useful to keep the top row visible. If we had a lot more data in here and scrolled down, we lose the top row. To keep the top row visible, select *View* on the menu bar, and click *Freeze Panes*, and select *Freeze Top Row*. We now see if we scroll down, the top row is kept visible at all times.

If you do not like the font theme, size, colour or formatting, then you can change these using the *Font* options within the *Home* tab. Highlight the row, column or cell you wish to change and select the new options within the tab.

You can also highlight the whole sheet by clicking the top left corner.

**Checking for duplicates**

It is always important to validate your data prior to performing any statistical analysis. One validation measure is checking for duplicates, where an observation in the data set has accidentally been entered more than once. We can clearly see a duplicate in our data set on rows three and four, and we could remove them simply by right-clicking on the row and selecting *Delete*.

However, in a much larger data set it will not be so easy to recognise duplicate entries, so we will let Excel find them for us. You need to click anywhere within the data first, I’m going to select cell B2, then select the *Data* tab on the menu bar, and within *Data Tools*, click *Remove Duplicates*.

We leave the *My data has headers* box ticked, as this is true in our case. We only want to delete duplicate entries where all variables are equivalent between observations, so leave all of the boxes ticked and click on *OK*.

The duplicate entry will be deleted and you will be given some information on the number of duplicates found and removed, as well as the number of unique entries that remain.

You can click *Undo* to get the entries back, or alternatively it is always a good idea to keep a copy of the original data set.

## Sorting and Filtering

## Currently our observations are ordered alphabetically by Forename, but suppose we want them in a different order. We can use the sort functionality within Excel to help us with this. Select anywhere within your data, click the *Data* tab, and then within the *Sort & Filter* menu, click *Sort*.

Again we leave the *My data has headers* box ticked. Then select the variable you wish to sort on. I’m going to choose Surname. I want to sort by the values of Surname, and I want them to be alphabetical, so I’ll leave these two options.

We can now see that the data is sorted by Surname.

We can sort by more than one column; for example, sorting by Surname A to Z, adding a level and sorting by Age as well, we can see that Henry now comes before Anna.

It is important to note that if you add new data, it will not be automatically sorted. You will need to re-sort the data if you add new observations.

You can also use the arrow buttons next to the *Sort* button, but these provide less options and it is easier to make mistakes.

The *Filter* option in the same section allows us to view a subset of our data. Suppose, for example, we just want to view some of the clinics, say A1 and A2. Click anywhere within the data and select the *Filter* button.

Arrows will appear at the top of each column. Select the variable you wish to filter on and untick those that you do not want to see.

We have now kept only clinics A1 and A2. Information is displayed in the bottom left corner of the screen, telling us how many records were found. To remove the Filter, simply click on *Clear*.

The Text Filters within the filter options give additional choices. For example, if we wanted to select all of the clinics beginning with B, we could just select *Begins With…*, and then type B. It is not case sensitive. There is additional information here on wildcards which may come in useful.

There are similar filter options for numbers. With Age for example, we could select everybody under the age of 18, by selecting *Number Filters*, *Less Than…*, typing 18 and clicking *OK*.

I’ll now remove these filters by pressing *Clear*.

To remove the filter options from your data, simply click on *Filter* button again.

## IF function

Functions in Excel are used perform mathematical calculations and manipulate data.

In this section the IF function is introduced. It’s one of over 300 functions available in Excel and can be very useful for data manipulation.

Functions may be entered into the *Formula Bar*, just below the main menu bar and the formula will then relate to the cell that was selected, G5 in this case.

Alternatively, simply click on the cell that you wish your formula to relate to and start typing. The formula will then also appear in the *Formula Bar*.

You may have noticed that I started typing the formula with an equals sign. This is true for all formulae – it is how you indicate to Excel that you are writing a formula.

The *Formulas* tab on the menu bar also provides a way of accessing available formulae. Clicking *Insert Function* allows you to see all of the functions that are available within Excel. If you aren’t sure which function you need, type a description into the box and click *Go*.

Now, back to the IF function. Suppose we want to group our patients by age, where anyone aged under aged 18 is classed as a child and the rest are adults. We’re going to record this in column F.

First, I’ll give it a header – Age Group.

Next, we type the formula. So, for our first patient, we’ll enter this into cell F2. We know we’re using a formula, so we start with the equals sign, and then we follow this with the name of the formula, in this case, IF. We can see some useful information then relating to the formula.

Next, we list the inputs of the function, contained within brackets. As soon as we open the brackets, we are told what the inputs are. First, we need to enter the logical test. This is whether or not the patient is aged under 18. We’re therefore testing if cell C2 is less than 18, so we type in C2 and we’ll see that the text and cell border outline are matched in colour. This helps us identify that what we are doing is correct. Alternatively, we can click on the cell instead, which is often easier, and the cell’s reference within the formula will be populated for us.

Now we need to complete the logical test. We want to know if the value in cell C2 is less than 18, so we simply write this inequality in using standard mathematical notation. We follow this with a comma, as indicated in the list, and we need to say what happens if this is true. If it is true, we want to record that they are a child, so we would write Child next – as it is a text string we need to use double quotation marks. Again, follow this with a comma, and next we write what happens if this is false, in this case, that they are an Adult.

Finally, we close the formula with a bracket and hit *Enter*.

We can see that this returned what we expected. If you need to change the formula, click on the cell and change the text.

Obviously we do not want to type this out for every patient. Instead, we can drag the formula down the column. Click on the cell and hover over the small square that appears in the bottom-right corner, until a black cross appears.

Holding down the left-hand button on the mouse, we then drag this down to the next cell. We can see that the formula automatically updates to look at the next cell down in the column.

To populate the whole column, we can drag all of the way down, like this…, or an easier option is to find the small black cross, then left double click. It will populate the cells as far down as there is content in the adjacent cells.

## Absolute and Relative referencing

Now consider the scenario where we wanted to change between 16 and 18, and suppose also we had a lot more data and there were other columns that were also impacted. It would be rather time-consuming to keep changing them between 16 and 18, or indeed any other number. This is where absolute referencing comes in.

If we write the age limit in another cell, I’m going to use cell K2, then we can refer to it there. I’ll also label this in case we forget what it is.

Now, we can refer to the cell in the formula, instead of writing the number. So, we delete 16 and instead we’re going to write K2, or we could have just simply selected K2.

If we now drag to complete the other formulae, we soon run into problems. Excel has tried to help us by updating the formulae as before, but this time we did not want it to.

We therefore are going to use absolute referencing, which is where we instruct Excel to not update certain parts of the formula when we drag. Instead, we fix them.

Here, we want to fix K2 in the formula. We don’t want it to update for each row. To do this, we use dollar signs. Type a dollar before the K, and a dollar before the 2.

The dollar sign fixes whatever comes next, so here we have fixed the K and the 2. Now, if we drag, we can see that each of the formulae contain the correct reference. Each time, it refers to K2.

A quicker way to enter absolute references is using F4 on the keyboard. Click somewhere within the cell reference, and press F4. The first dollar sign has disappeared; this means that the row, 2 in this case, is now fixed. The column is not fixed. Pressing it again removes all referencing, and we can simply scroll through all of the options by continually pressing F4.

Using relative and absolute referencing effectively can save us a lot of time and effort when manipulating data in Excel.

## Additional help and resources

That concludes this ‘Getting Started With Data Manipulation in Microsoft Excel’ tutorial. Thank you for watching and I hope you have found it useful.

For additional support and resources, see the statstutor website and the Excel help files.