An introduction to using Microsoft Excel for quantitative data analysis

Contents

1 Introduction ........................................................................................................................................... 1
2 Why use Excel? .................................................................................................................................... 2
3 Quantitative data analysis tools in Excel .......................................................................................... 3
4 Entering your data ............................................................................................................................. 6
5 Preparing your data .......................................................................................................................... 7
6 Appendix A: Using Excel statistical functions ............................................................................... 8
7 Appendix B: Loading and using the Data Analysis ToolPak ............................................................ 12

1 Introduction

The aim of this short guide is to provide an introduction to using Microsoft Excel (hereafter: Excel) for quantitative data analysis within the context of a business and management research project. It covers some of the key features of Excel that are particularly useful when doing a research project. Further guides give information on the use of Excel to apply various analysis techniques discussed in Chapter 13.

The guide is not written for a specific version of Excel although it includes screenshots for Excel 2010. Most of the functionality referred to in the guide is also available in earlier and later versions, although the user interface has changed somewhat.

The guide assumes that you are already familiar with the basics of using Excel (e.g. how to create worksheets, enter data, use formulae and functions, create charts (graphs), print and work, etc.). If you have never used Excel, there are many reasonably-priced textbooks to get you started. Alternatively, you may find that Excel training or support material is available in your institution. There are also various websites, including Microsoft’s Office Support area (http://office.microsoft.com/en-001/support/?CTT=97) that offer advice to get you started.
2 Why use Excel?

With so many specialist software packages available, why use Excel for statistical analysis? Convenience and cost are two important reasons: many of us have access to Excel on our own computers and do not need to source and invest in other software. Another benefit, particularly for those new to data analysis, is to remove the need to learn a software program as well as getting to grips with the analysis techniques. Excel also integrates easily into other Microsoft Office software products which can be helpful when preparing reports or presentations.

2.1 What you can do with Excel

As a spreadsheet, Excel can be used for data entry, manipulation and presentation but it also offers a suite of statistical analysis functions and other tools that can be used to run descriptive statistics and to perform several different and useful inferential statistical tests that are widely used in business and management research. In addition, it provides all of the standard spreadsheet functionality, which makes it useful for other analysis and data manipulation tasks, including generating graphical and other presentation formats. Finally, even if using bespoke statistical software, Excel can be helpful when preparing data for analysis in those packages.

2.2 Limitations of Excel

Useful though it is, Excel does have limitations for statistical analysis. It remains first and foremost a spreadsheet package. Inevitably it does not cover many of the more advanced statistical techniques that are used in research. More surprisingly, it lacks some common tools (such as boxplots) that are widely taught in basic statistics. There is also concern amongst some statisticians over the format of specific output in some functions. The extensive range of graph (chart) templates is also criticised for encouraging bad practice in data presentation through inappropriate use of colour, 3-D display, etc. Despite these limitations Excel remains a very valuable tool for quantitative data analysis as you will see.

2.2.1 Making the choice

Many basic analysis projects involving primarily data exploration, descriptive statistics and simple inferential statistics can be successfully completed using standard Excel. More advanced projects, especially those involving multivariate analysis are more challenging in
Excel and in such cases it is worth considering using specialist analysis software such as IBM SPSS.

3 Quantitative data analysis tools in Excel

Excel includes a large number of tools that can be used for general data analysis. Here our primary concern is those that are relevant to the statistical and related analysis techniques introduced in Chapter 13. Four sets of tools are particularly useful:

3.1.1 Statistical functions

Excel offers a broad range of built-in statistical functions. These are used to carry out specific data manipulation tasks, including statistical tests. An example is the AVERAGE function that calculates the arithmetic mean of the cells in a specified range. A list of Excel functions referred to in this and other guides is included in Appendix A along with instructions on how to access them.

3.1.2 Data Analysis ToolPak

The Data Analysis ToolPak is an Excel add-in. It contains more extensive functions, including some useful inferential statistical tests. An example is the Descriptive Statistics routine that will generate a whole range of useful statistics in one go. An introduction to loading and using the ToolPak add-in is included at Appendix B.

The ToolPak is not available in Excel for Mac. See Appendix B for an alternative.

3.1.3 Charts

Excel’s in-built charts (graphs) cover most of the chart types introduced in Chapter 13 and are invaluable in data exploration and presentation. We illustrate their use in Chapter 13 and also in the other guides.

3.1.4 Pivot tables

Pivot tables provide a way of generating summaries of your data and organising data in ways that are more useful for particular tasks. They are extremely useful for creating contingency

---

1 For clarity, Excel functions will be written in upper case letters as they appear in the Excel formula bar.
tables, cross-tabulations and tables of means or other summary statistics. A brief introduction to creating pivot tables is given in the guide *Data exploration in Excel: univariate analysis*.

### 3.2 Preparing Excel for analysis

Before starting, check that your Data Analysis ToolPak has been loaded. Do this by selecting the Data tab; the Data Analysis command should appear in Analysis group on the right-hand side of the ribbon (Figure 1). If it does not do so, follow the procedures shown in Appendix B.

*Figure 1 – Data Analysis ToolPak command*

![Data Analysis ToolPak command](image)

### 3.3 Setting up your data for analysis

Typically there are two options for getting your data into Excel:

1. Import the data in a suitable format from, for example, an online survey tool.
2. Enter the data manually.

If you are going to enter your data manually use a single worksheet to hold all the data in your dataset and set up the worksheet with variables (questions) as the columns and the cases (e.g. respondents) as the rows. An individual cell, therefore, contains a respondent’s answer to a specific question. Figure 2 shows a small dataset entered into Excel.

*Figure 2 – Sample dataset*
3.3.1 Allocate column headers

In the first row, give each column a simple, informative header that will be easy to understand when entering data or reviewing output\(^2\). Avoid just using question numbers (e.g. Q1, Q2, etc.) as these can be confusing if you have a large number of questions. Instead, use a simple naming system. A variable measuring customer satisfaction, for example, could be headed CSat: easy to remember and not likely to be confused during analysis. Ensure each header is unique (this will facilitate subsequent analysis and avoid confusion when interpreting output).

3.3.2 Allocate each case a unique ID

If they do not have one already, allocate each case in the dataset a unique numerical identifier (ID). The easiest way to do this is simply to number them consecutively from 1 through to \(n\) (where \(n\) is the number of cases). For clarity, it is best to put the ID as the first column in the worksheet. Giving each respondent a unique ID aids in sorting and tracking individual responses when (for example) cleaning the data or checking outliers. A simple, consecutive number ID system also makes it easy to reorder the data if needed. If you are transferring data

\(^2\) Some software programs (such as IBM SPSS) have particular requirements or restrictions on variable names (e.g. no spaces). If you are using Excel to prepare data for analysis in another program, it is useful to check those requirements before naming your variables.
from paper copies of a questionnaire, it is useful to write the ID number onto the paper copy to make it easier to check any errors.

4 Entering your data

Once the spreadsheet is set up, simply enter the data into the appropriate cell as required. Numerical data can be entered as numbers, other data, such as Likert scale data, may need to be coded (see Chapter 13). With nominal data you have two options:

- Enter the values as words (e.g. male/female), appropriately abbreviated if required (e.g. m/f). Ensure you are consistent in spelling and format as Excel will treat each variation as a different value.
- Enter the re-coded numerical values (e.g. 0/1 for male/female), ensuring you keep a record in a code book (Chapter 13). A worksheet in the workbook is a useful place to record details of your variables and to store your code book as shown in Figure 3.

Which to do depends on your analysis needs. Some tools in Excel (e.g. pivot tables) work well with text and generate meaningful output but some analysis tasks may require numerically coded data. If you are exporting your data to another software package, check the format required by that package. In some cases, it may be helpful to have both formats. You can do this by creating a copy of the column containing the original data, then selecting the new column and using Home > Find & Select > Replace to replace the original values with the new ones. Ensure you give the new column a unique header.

Figure 3 – Using a worksheet to record details of dataset variables
4.1 Importing data

If you are importing the data from another electronic file, check that the layout is suitable (i.e. respondents as rows, variables as columns), add or modify variable names if required, add respondent ID if needed and check that the data has imported correctly.

4.2 Managing your data

Once you have created your dataset, ensure that you back it up in a secure place, not on your PC or laptop. If you make any changes to your master dataset, record those changes and create a duplicate back-up.

Give files a meaningful name. It is also helpful to date them as this makes it easier to track back if you need to do so. Worksheet tabs can also be named to help you manage your data.

5 Preparing your data

Once your data are entered you can follow the steps in Chapter 13 to prepare your data for analysis. If you need to carry out data transformation, such as recoding variables or calculating summated scores, do so now. (Hint: you can use functions such as SUM and AVERAGE to help you with creating summated scales.) If you are creating new variables during data transformation ensure they are given unique column headers.

In Excel a simple way of carrying out the reverse coding for scales is to set up a new variable and calculate the new values for each cell using the following equation:

\[
\text{new value} = (\text{highest value in scale} + 1) - \text{respondent's original score}
\]

So for a 5-point scale with an original score of 2, the reverse coded value would be:

\[
\text{new value} = (5 + 1) - 2 = 4
\]
6 Appendix A: Using Excel statistical functions

You have almost certainly used the Autosum (∑) SUM function which returns the sum of a range of cells, even if you were not aware that it is only one of a suite of such tools. In fact Excel includes a very large number of functions covering a broad range of applications, including statistics.

6.1 Accessing functions

You can access the statistics functions in Excel via Formulas > More Functions > Statistical which opens up a menu of available functions (see Figure 4). You can then select the function you desire. It will be inserted in the active cell so ensure you choose a destination cell away from your dataset or your data will be overwritten by the function output.

Figure 4 – Excel’s statistical functions menu

You can also access functions via Formulas > Insert Functions which opens up the Insert Functions menu into which you can type the name of the function you want, select a Category from which to choose, or use one of the functions listed in the Select a Function window.
6.2 Example statistical functions

Table 1 shows some of the statistical functions available in Excel that are particularly relevant to the analysis techniques introduced in Chapter 13.

Table 1 – Selected Excel statistical functions (Excel 2010 and 2013)

<table>
<thead>
<tr>
<th>Function name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>Returns the arithmetic mean (average) of the given numbers</td>
</tr>
<tr>
<td>CHISQ.DIST.RT</td>
<td>Returns the right-tailed probability for the chi squared distribution</td>
</tr>
<tr>
<td>CHISQ.TEST</td>
<td>Returns the p-value for the chi-squared test of association</td>
</tr>
<tr>
<td>CONFIDENCE.T</td>
<td>Returns the margin of error for a confidence interval for the mean</td>
</tr>
<tr>
<td>COUNT</td>
<td>Counts the number of cells in a range that contain numbers</td>
</tr>
<tr>
<td>COUNTIF</td>
<td>Counts the number of cells in a range that meet a given condition</td>
</tr>
<tr>
<td>COUNTA</td>
<td>Counts the number of cells in a range that are not empty</td>
</tr>
<tr>
<td>KURT</td>
<td>Returns the kurtosis of a dataset</td>
</tr>
<tr>
<td>MAX</td>
<td>Returns the maximum value of the given numbers</td>
</tr>
<tr>
<td>MEDIAN</td>
<td>Returns the median of the given numbers</td>
</tr>
<tr>
<td>MIN</td>
<td>Returns the minimum value of the given numbers</td>
</tr>
<tr>
<td>MODE.SNGL</td>
<td>Returns the mode of the given numbers</td>
</tr>
<tr>
<td>PEARSON</td>
<td>Returns the Pearson correlation coefficient (r) of two variables</td>
</tr>
<tr>
<td>SKEW</td>
<td>Returns the skewness of a dataset</td>
</tr>
<tr>
<td>STDEV.P</td>
<td>Returns the standard deviation of the given numbers, based on the population</td>
</tr>
<tr>
<td>STDEV.S</td>
<td>Returns the standard deviation of the given numbers, based on a sample</td>
</tr>
<tr>
<td>VAR.P</td>
<td>Returns the variance of the given numbers, based on the population</td>
</tr>
<tr>
<td>VAR.S</td>
<td>Returns the variance of the given numbers, based on a sample</td>
</tr>
</tbody>
</table>

6.3 Using a function

We will introduce specific functions in the other guides but the following example of applying the AVERAGE function to calculate the mean age in the sample dataset in Figure 2 illustrates their use:

- Select the cell in which you wish the calculation to be placed (Hint: if you are using the same worksheet as your dataset, avoid cells that are immediately adjacent to your data).
• Select Formulas > More Functions > Statistical > AVERAGE to open the Function Argument dialogue box (Figure 5).

Figure 5 – AVERAGE Function Argument dialogue box

• Select the range of cells to which the function should be applied (Figure 6).

Figure 6 – Selecting the range of cells

• Select OK. The result shows the arithmetic mean (AVERAGE) of the chosen numbers (Hint: Type a descriptor of the function in an adjacent cell so that you can remember what has been calculated as shown in Figure 7).

Figure 7 – Function result
<table>
<thead>
<tr>
<th>ID</th>
<th>Temp</th>
<th>Distance</th>
<th>Gender</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>32.0</td>
<td>male</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>27.0</td>
<td>female</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
<td>35.5</td>
<td>male</td>
<td>33</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>28.2</td>
<td>female</td>
<td>42</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>32.8</td>
<td>female</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>25</td>
<td>28.8</td>
<td>female</td>
<td>28</td>
</tr>
<tr>
<td>7</td>
<td>26</td>
<td>29.3</td>
<td>female</td>
<td>33</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>29.5</td>
<td>female</td>
<td>41</td>
</tr>
<tr>
<td>9</td>
<td>18</td>
<td>37.2</td>
<td>male</td>
<td>29</td>
</tr>
<tr>
<td>10</td>
<td>26</td>
<td>27.9</td>
<td>female</td>
<td>36</td>
</tr>
<tr>
<td>11</td>
<td>28</td>
<td>26.3</td>
<td>male</td>
<td>35</td>
</tr>
<tr>
<td>12</td>
<td>26</td>
<td>28.1</td>
<td>male</td>
<td>18</td>
</tr>
</tbody>
</table>

Mean age: 30.5
Appendix B: Loading and using the Data Analysis ToolPak

The Data Analysis ToolPak is an add-in. It is installed when you install Excel but needs to be loaded prior to use. To do so:

- Select File > Options > Add-Ins to open up the Add-ins dialogue box (Figure 8).

Figure 8 – Excel Add-Ins dialogue box

- In the box marked Manage, select Excel Add-ins and click OK.
- In the Add-ins available box, select the Analysis ToolPak check box and click OK (Figure 9).

(Note: your available add-ins may differ from those shown in Figure 8 depending on what else has been installed on your machine.)
Once the Data Analysis ToolPak is loaded, it will be available via the Data > Data Analysis (see Figure 1).

### 7.1 Functions available in the Data Analysis ToolPak

Table 2 shows some of the functions available in the Data Analysis ToolPak that are particularly relevant to the analysis techniques introduced in Chapter 13.

**Table 2 – Selected Data Analysis ToolPak functions (Excel 2010 and 2013)**

<table>
<thead>
<tr>
<th>Function name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anova: Single Factor</td>
<td>Performs a one-way analysis of variance (ANOVA)</td>
</tr>
<tr>
<td>Correlation</td>
<td>Creates a correlation matrix showing the Pearson correlation coefficient ($r$) for each pair of variables of $N$ cases selected</td>
</tr>
<tr>
<td>Descriptive statistics</td>
<td>Calculates a range of univariate descriptive statistics, including measures of central tendency, dispersion, skewness and kurtosis for a variable</td>
</tr>
<tr>
<td>Histogram</td>
<td>Generates a histogram for a range of data (this function also generates a table of the data on which the histogram is based and can be used for Pareto analysis)</td>
</tr>
<tr>
<td>$t$-Test: Paired Two-Sample for Means</td>
<td>Performs a $t$-test to compare the means of a paired sample</td>
</tr>
<tr>
<td>$t$-Test: Two-Sample Assuming Equal Variances</td>
<td>Performs a $t$-test to compare the means of two independent samples, assuming equal variances</td>
</tr>
</tbody>
</table>
### Function name | Description
---|---
t-Test: Paired Two-Sample Assuming Unequal Variances | Performs a t-test to compare the means of two independent samples, assuming unequal variances

### 7.2 Using functions in the Data Analysis ToolPak

We will introduce specific ToolPak functions in the other guides but the following example of applying the Descriptive Statistics function to generate a range of descriptive statistics for the variable age in the sample dataset in Figure 2 illustrates their use:

- Select Data > Data Analysis to open the Data Analysis menu dialogue box (Figure 10).

**Figure 10 – Data Analysis menu dialogue box**

- Select the desired function, in this case Descriptive Statistics, which opens the relevant dialogue box (Figure 11).
- In the dialogue box, enter the desired range in the Input Range box. If you have included the column header, select the Labels in First Row box. Confirm where you want the output to go. The default setting is New Worksheet Ply which creates a new worksheet for the output; since most ToolPak outputs are quite large, this is a sensible option.
- Select Summary Statistics to get descriptive statistics for your chosen data; you can also select an appropriate confidence interval for the mean if desired (the default is 95%).

**Figure 11 – Descriptive Statistics dialogue box**
Click OK. The output will be shown in a new worksheet (Figure 12). Note that here the column widths have been adjusted to make it easier to read.

Figure 12 – Descriptive Statistics output for variable Age

Note also that this output is not dynamically linked to the original dataset so changes to the dataset will not automatically be updated in the output. You will need to run a new analysis.

Once created, the output can be cut-and-pasted into word-processing software for further editing.
(Hint: if using the Descriptive Statistics function, you can select multiple adjacent variables and the function will report the output for each one.)

7.2.1 Excel for Mac

If you are using Excel for Mac (2008 or 2011) the data analysis ToolPak add-in is not available within Excel. An alternative is a downloadable add-in called StatPlus:mac LE which is available on the following link (note that this is a free version of a more powerful commercial add-in, so ensure that you download the correct version):