
Review of “Applied Statistics with Microsoft Excel” by Gerald Keller

Neville Hunt
Coventry University

n.hunt@coventry.ac.uk

This book provides a general introduction to statistics for first year undergraduates in any user-discipline. It is written for “statistics practitioners” rather than statisticians – according to the author the latter are those who work with the mathematics of statistics. Running to 670 pages it covers more topics than many books at this level, providing a thorough treatment of analysis of variance (including multiple comparisons and interactions), multiple regression (including indicator variables and autocorrelation) and non-parametric tests (including Kruskal-Wallis and Friedman tests). However, it does not cover decision analysis or time series decomposition, which may reduce its appeal in the economics and business field.

The book is well written with an emphasis on applications and minimal mathematics. Key concepts are highlighted and the layout is attractive and easy to follow. Throughout the book the author follows an explicit three-step approach: identify the technique, compute the statistics, and interpret the results. For those who like such things, there are flow charts to help the student to decide what type of analysis to undertake. The author aims to keep manual calculation to a minimum so as to spend more time on understanding and interpreting the results – an admirable philosophy.

Exercises

Each chapter contains a substantial number of exercises based on real (or at least realistic) scenarios, although most are written with an American readership in mind. The exercises are conveniently annotated to indicate their field of application – for example, Politics or Science. Brief answers to even-numbered exercises are supplied at the back of the book. There is also an Instructor’s Solutions Manual with worked solutions to all exercises and a Student Solutions Manual with worked solutions to even-numbered exercises. Additional exercises will eventually be provided at the companion web-site, although at the time of writing this review the site was not in place. Hopefully it is not easy for an enterprising student to purchase the Instructor’s Solutions Manual!

Most exercises have an associated data set provided on the accompanying CD in Excel, ASCII, JMP and Minitab formats. The data sets are helpfully organised into folders by chapter and each file name indicates the exercise or example in the text to which it refers. In all there are more than 600 data sets, some as large as 22, 000 observations, which is an excellent resource. Unfortunately it is not always clear whether the data are real. For example, one data set purports to contain the salaries of 480 (presumably American) university professors, but I hesitate to indulge in any comparative analysis with the UK without some guarantee that the data are real – I would also need to know the year to which the data refer. The lack of source detail on these data sets does detract somewhat from their usefulness. Later chapters conclude with several interesting longer case studies and these do generally have clear source details.

Excel

In many ways I would prefer to end my review here on a largely positive note. However I am bound to address the main feature of the book, which is its link with Microsoft Excel 2000. At the outset I must confess to being an Excel enthusiast. I believe it is a wonderful tool for aiding the teaching and learning of statistics at an elementary level. In my experience students learn a great deal by building a spreadsheet themselves to carry out a simple analysis. In so doing they discover how each formula works and can experiment with the values of parameters and observe the effect on the results. Dynamic graphics and simulations can be used very effectively to illustrate fundamental concepts such as the central limit theorem and least squares. Excel also has an array of splendid tools for charting, tabulating, validating, editing and recoding raw data. Unlike most statistical software, Excel is not a “black box” but allows the student to observe its inner workings. Again, unlike most statistical software, Excel is almost universally available to students either at college or at home.

With all this in mind I was disappointed to find that this book makes surprisingly little use of the spreadsheet functionality of Excel. Some use is made of the probability distribution functions, although I was amused to find nine pages devoted to traditional Normal probability tables and just one page to Excel’s NORMDIST and NORMINV functions! No attempt is made to plot Binomial and Poisson distributions, which seems an opportunity missed. Some helpful simulation experiments are suggested to reinforce the ideas of sampling distributions, confidence intervals and estimation, although sadly these are described as “optional”.

Add-ins

The book relies mainly on the Data Analysis ToolPak add-in that is supplied with Excel. This is supplemented by a further add-in called Data Analysis Plus supplied on the book’s companion CD, which performs additional analyses not included in Excel’s resident ToolPak – for example, non-parametric tests. Data Analysis Plus is very easy to install (and uninstall), appears as an additional menu item on the Tools menu, and has been designed to look and feel just like the Data Analysis add-ins. In almost all cases in the book where computation is required, readers are given instructions on how to use the relevant macro from either Data Analysis or Data Analysis Plus (see Fig 1) and sample output is displayed (see Fig 2).

- 1 Type the frequencies into adjacent columns.
- 2 Click **Tools, Data Analysis Plus, and Contingency Table**.
- 3 Specify the **Input Range: A1:D5**
- 4 Click **Labels**, if the first row and first column of the input range contain the names of the categories.
- 5 Specify the value of a **(Alpha)**, and click **OK: 0.05**

Fig 1 Instructions for Contingency Table analysis

These instructions make Data Analysis Plus very easy to use. The output is not always as complete as one might like. For example, in Fig 2 - which is fairly typical of all the macros - the output does not include a table of expected frequencies, or a table showing the contribution of each cell to the chi-squared statistic. Admittedly both the p-value and critical value are helpfully provided, which means that either approach to testing may be used. However, the significance level of 0.05 used on input is not included in the output - we are required to remember that. Moreover the stated p-value of 0 here gives us no indication as to how small the p-value actually is, which is rather sloppy.

Contingency Table

	<i>Dem</i>	<i>Rep</i>	<i>Ind</i>	<i>TOTAL</i>	
Reduce		118	193	45	356
Pay	83	132	41	256	
Social	109	88	31	228	
Health		101	31	28	160
TOTAL		411	444	145	1000

chi-squared Stat 68.9037
 df 6
 p-value 0
 chi-squared Critical 12.5916

Fig 2 Output from Contingency Table Analysis

Use of Excel’s Data Analysis ToolPak is unwise on two counts:

- The output from these macros is “dead” text, that is, it is not dynamically linked to the data. This means that if I spot a mistake in my data and correct it the output does not automatically recalculate and I must run the macro again. This is totally contrary to the whole idea of a spreadsheet environment.

- The ToolPak is known to be unreliable with some macros giving daft results under certain circumstances. Worse still, these circumstances are not well documented and Microsoft shows no inclination to either acknowledge the errors or correct them.

I regret to report that Data Analysis Plus shares both of these failings. When presented with two perfectly correlated columns of data the correlation macro crashes with a division by zero error – I don't think we can blame Microsoft for that. The boxplot macro refuses to accept more than one set of data – yet surely the main purpose of a boxplot is to compare distributions among different groups? Hopefully no student will be so unkind as to perform a z-test with a significance level of 0.0000001 – the Data Analysis Plus critical-value turns out to be the rather arbitrary 5,000,000, which confirms that the author is using Excel's own functions in these macros rather than implementing his own algorithms. To be fair these are all features that can easily be corrected in a later version, but they do not suggest that Data Analysis Plus is any more robust than the original ToolPak.

The author does recognise the problems with Excel's statistical capability with the following caution at the start of the book:

The current versions of Excel use algorithms that are not robust. This means that under certain circumstances, Excel's calculations may be incorrect. In some extreme cases, the errors can be very large. Until Microsoft improves this aspect of its products, students are cautioned when using Excel on sets of data other than the ones included with this book. On these datasets Excel works perfectly.

I am sure that many will feel very uneasy about this. We have only reached page 23 of the book when we encounter Excel's histogram - with the upper class limits plotted at the midpoints of the columns. This is not highlighted as an error nor is there any mention of frequency density or the fact that Excel will draw each column the same width even when the intervals are of unequal width! By page 74 we have used Data Analysis Plus to draw a boxplot and it is pointed out that the quartiles on the boxplot do not agree with those from the ToolPak's Descriptive Statistics macro. The differences are not discussed – and maybe they don't really matter? Certainly as the book progresses the Excel content becomes much less significant and the book could be used perfectly well in conjunction with some other statistical software.

Conclusion

This is basically a very sound textbook with many good features. Its deliberate use of Excel is well motivated with the students' best interests in mind. However, the particular way in which it uses Excel is highly questionable. It is a pity that the author did not take the opportunity to rewrite Excel's ToolPak, correcting the blunders, and offer Data Analysis Plus as a replacement rather than a supplement. It is also perfectly possible to program a macro to present the output as "live" formulae, which would constitute a great improvement. My own approach to using Excel would have been to discard the macros altogether and either create, or instruct students how to create, spreadsheet templates to carry out all the different analyses. Once analyses become too complex for this to be done easily, then it is time to move on to a reputable statistical package.