

Statistically Sound Distribution Plots in Excel

Gaj Vidmar¹

Abstract

Excel is the most widespread and the most powerful general-purpose spreadsheet software, but it is not popular with statisticians. Nevertheless, as a natural means for organising, displaying and analysing large amounts of data, spreadsheets keep gaining importance in statistical education and practice. Aiming at improving such practice rather than fruitlessly and indiscriminately condemning it, the paper provides general considerations on the topic, pointers to the huge body of relevant literature and software, and several concrete examples of data visualisation in Excel in the sense of univariate, bivariate and multivariate distribution plotting. Original and improved Excel solutions for producing dot-density plots, dot plots, stem-and-leaf plots, windowgrams, coplots and parallel coordinates plots are presented, as well as for performing the Box-Cox transformation. Additionally, further possibilities opening with the forthcoming Excel 2007 version, use of various commercial and freeware add-ins, and integration of Excel with statistical software are discussed.

1 Background

Microsoft[®] Excel (henceforth referred to as Excel for brevity) is by far the most widespread and arguably the most powerful spreadsheet, but it is not a very popular tool with statisticians. This is partly justified (McCullough and Wilson, 1999, 2002), but to a large extent also based on prejudice and ignorance, as demonstrated by the publications, software and examples referenced, discussed or introduced in this paper.

As a natural means for organising, displaying and analysing large amounts of data, spreadsheets have gradually but relentlessly found their way into the world of mathematics and statistics. Their potential for interactive teaching was already realised at the onset of widespread Internet use (Hunt and Tyrrell, 1995), and nowadays high-profile introductory statistics textbooks are being published that

¹ University of Ljubljana, Faculty of Medicine, Institute of Biomedical Informatics; gaj.vidmar@mf.uni-lj.si

are either entirely based on Excel (e.g., Anderson, Sweeney and Williams, 2005) or demonstrate its use in parallel with a major statistical software package (e.g., Hawkes and Marsh, 2005). Though its limitations and shortcomings, mainly in terms of numerical algorithms and missing data handling, can take considerable effort and expertise to overcome (Heiser, 2005), Excel has proven to be a particularly valid tool for combining mathematics, statistics and engineering education (de Levie, 2004; Liengme, 2002; Neuwirth and Arganbright, 2004). In the form of electronic or printed monographs with accompanying publicly available add-ins, Excel is also becoming a prominent tool for combining statistical teaching and statistical practice (Steppan, Werner and Yeater, 2001; Myerson, 2005). Last but not least, the market niche of general-purpose or specialised statistical add-ins for Excel as low-cost alternatives to stand-alone statistical packages would not be flourishing as it is² if the producers were only exploiting Excel's deficiencies and omissions, rather than capitalising on Excel's qualities and potential for storing, manipulating, analysing and presenting statistical data.

The field of applying Excel in statistics that is subject to particularly harsh criticism, but also witnessing particularly intense development, is data visualisation. The large public knowledge base consisting of the usenet newsgroup and websites of experts³ is an enormous source of ready-made solutions (e.g., for bivariate density charts⁴), instructions for working around the limitations of Excel's charting facilities (e.g., to create boxplots⁵), and support for task automation via macros and add-ins. At the same time, professional products are covering the range from large-scale categorical data visualisation⁶ through dynamic multivariate exploration⁷ to heat-maps for gene microarray data analysis⁸.

² successful products include (tentatively ordered by decreasing scope) XLSTAT-Pro (<http://www.kovcomp.co.uk/xlstat>), statistiXL (<http://www.statistixl.com>), WinSTAT (<http://www.winstat.com>), Simetar[®] (<http://simetar.com>), SPC XL (<http://www.sigmazone.com>), SigmaXL[®] (<http://www.sigmaxl.com>) and MegaStat[®] (<http://blue.butler.edu/~orris/megastat>)

³ especially recipients of the Microsoft[®] Most Valuable Professional title, such as (in alphabetical order) F. Cinquegrani (<http://www.prodomosua.eu/ppage02.html>), T. Mehta (<http://www.tushar-mehta.com/>), J. Peltier (<http://peltiertech.com>) and A. Pope (<http://www.andypope.info>)

⁴ <http://www.prodomosua.eu/zips/density.exe> or <http://www.j-walk.com/ss/excel/files/gradcontour.htm>

⁵ <http://www.mis.coventry.ac.uk/~nhunt/boxplot.htm> or <http://peltiertech.com/Excel/Charts/BoxWhisker.html>

⁶ e.g., Treemap freeware implementations (<http://research.microsoft.com/community/treemapper>) and commercial extensions (<http://www.panopticon.com>)

⁷ e.g., commercial Miner3D[™] (<http://www.miner3d.com>) and freeware VisuLab (<http://www.inf.ethz.ch/personal/hinterbe/Visulab>)

⁸ e.g., BRB ArrayTools (<http://linus.nci.nih.gov/BRB-ArrayTools.html>)

2 About the paper and the presented solutions

With the presented solutions, the author filled the main gaps left in the collection of basic distribution plots available to an Excel user aware of the above-named resources. The examples are primarily intended for teaching and demonstrational purposes, but they are also useful for producing publication-quality graphics. They range from univariate through bivariate to multivariate visualisations and related procedures.

The remainder of the paper is organised as follows. First, the univariate dot-density plot is addressed together with simple and multi-way dot plots, discussing the existing resources and providing implementation of the two possible solutions, i.e., character-based cell-charts and modified scatter-plots (Section 3). Next, an implementation of the stem-and-leaf plot with a macro is presented (Section 4). Automated histogram plotting and the use of Excel's built-in Solver add-in for optimisation are combined in the presented implementation of the modified Box-Cox power transformation towards normal distribution (Section 5). Kernel density estimation is introduced through windowgrams, which are implemented only with conditional summing formulas (Section 6). In combination with coplot demonstration (scatter-plots conditioned on a third variable) based on array formulas (Section 7), the aim is to make advanced concepts and techniques understandable and accessible to non-mathematical audience. Parallel coordinates plot is implemented with a macro for quick production of presentation graphics as well as for exploratory purposes (Section 8). Additionally, further possibilities for data visualisation with the forthcoming Excel 2007 version and the integration of Excel with academically oriented statistical packages are discussed (Section 9). Section 10 offers some concluding thoughts.

To emphasise the actual user experience and the instructional aspects of the solutions, worksheet screenshots are presented in the figures rather than just the resulting plots. The aim of the text in the worksheets is to make them sufficiently self-explanatory while still stimulating the user's self-discovery. Because of the instructional purpose, some compromise had to be made with respect to the principles of good data visualisation, but elimination of visual clutter (Tufte, 1983) and avoidance of unnecessary use of colour (Wainer and Thissen, 1981) has been pursued relatively strictly. Since random data is generated solely for demonstrational purposes, Excel's built-in random number generation function (RAND) is used despite its known deficiencies (particularly in older versions; see Heiser, 2005, for details).

All the presented solutions were developed on the Microsoft® Windows® platform and tested with Excel versions from 97 upwards. Those not using macros should be completely compatible with the corresponding Apple® Macintosh® versions of Microsoft® Office, while those including VBA code could work with Microsoft® Office for Mac® versions prior to the 2004 edition. The workbooks

can be downloaded from the webpage of the Biostatistical Centre of the Institute of Biomedical Informatics, Faculty of Medicine, University of Ljubljana (<http://www.mf.uni-lj.si/ibmi-english/biostat-center>, follow the link to Software). Some other statistical Excel workbooks can also be found there.

The workbooks described in this paper are designed to be readily available for use upon download. The combination of instructions, comments, plot and axis titles, column headings and formatting elements should provide all the necessary information for the interested instructor, student or practitioner. The scope of four workbooks is primarily instructional: the solutions for dot-density plots, dot plots, windowgrams and coplots require a small further effort from the users who want to transfer them into daily statistical practice. Three workbooks have dual scope: the solutions for stem-and-leaf plots, modified Box-Cox transformation and parallel coordinates plots can be used not only as instructional aids, but also as "proper" applications.

3 Dot-density plots and dot plots

Dot plots (Cleveland, 1985) have recently attracted notable attention in business setting and in the context of management dashboards (Kyd, 2006; Robbins, 2006). As spreadsheets are particularly dominant in such setting, it is not surprising that detailed step-by-step instructions for constructing them were aptly published by Excel experts (O'Day, 2006; Peltier, 2006).

Conceptually, dot plots can be considered as developed from dot-density plots, and the two methods are also technically related in terms of possible Excel implementation. Dot-density plots can be simply and efficiently implemented as cell-charts with the REPT function. As explained and automated by Peltier (2006), the same principle can be used for constructing dot plots, including additional possibilities offered by conditional formatting (e.g., different symbols corresponding to the value of another variable, or highlighting outliers). The other main option for constructing dot plots in Excel involves workarounds for changing the appearance of scatter-plots.

Below, each plot type is presented in a separate subsection, whereby for instructional purposes, dot-density plots are implemented as cell-charts, while the scatter-plot based technique is used to construct univariate and multi-way dot plots.

3.1 Dot-density plots

Figure 1 shows the extremely simple procedure for creating a dot-density plot as a character-based cell-chart. The worksheet also instructs the user how to copy to

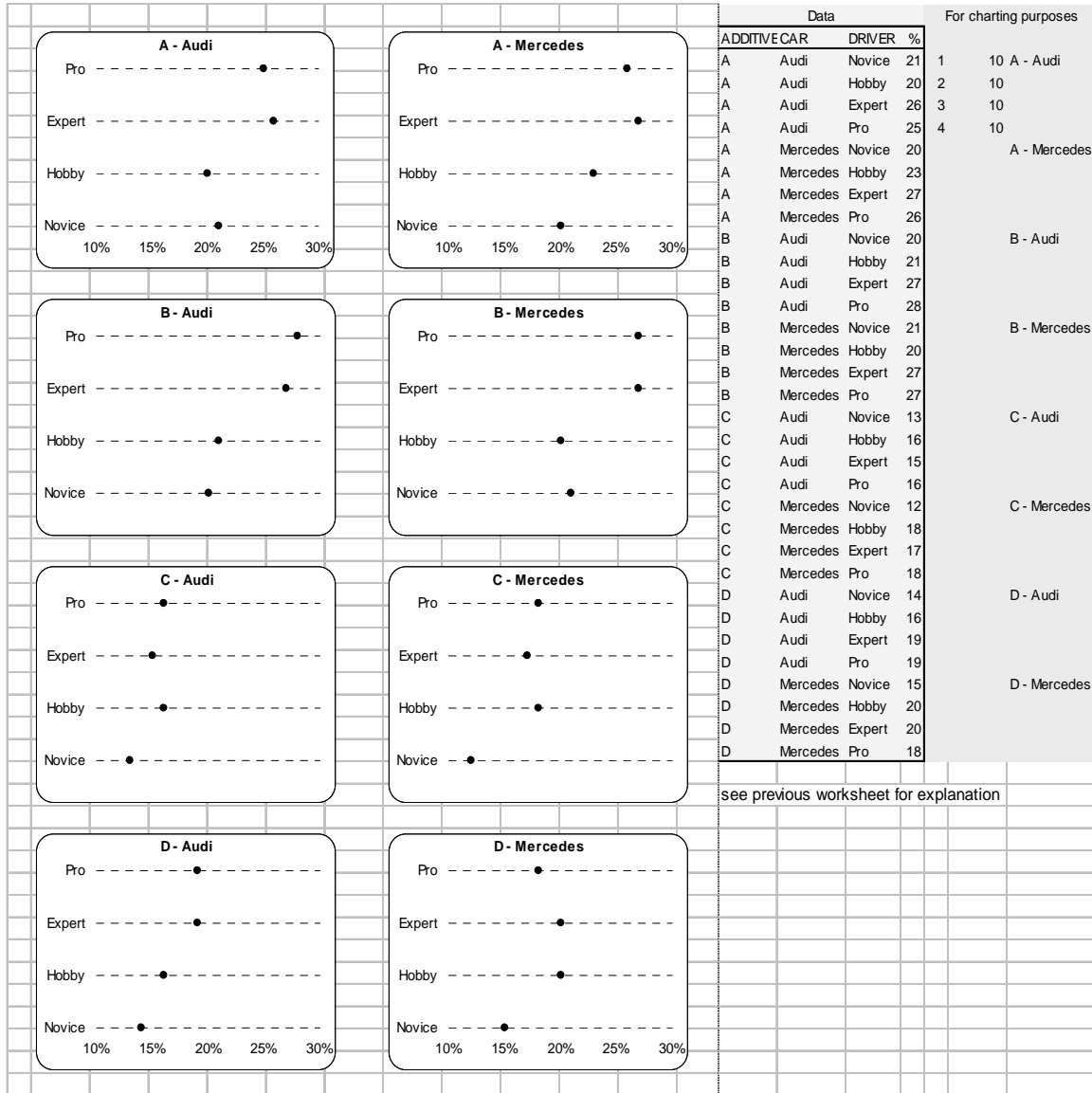


Figure 3: Excel worksheet demonstrating multi-way dot plots with designed-experiment data.

3.2 Dot plots

For dissemination purposes, the presented solutions for a single dot plot and multi-way dot plots are stored together in one workbook with two worksheets (named "Single" and "Multi-way", respectively).

Construction of dot plot in Excel from a scatter-plot by hiding axes and major gridlines and introducing a series with data labels to replace the categorical axis is displayed in Figure 2. It requires the installation of an add-in for linking data

labels to cell contents, which is in any case a mandatory tool for serious data visualisation with Excel.

In Figure 3, multi-way dot plots are illustrated with a realistically constructed dataset from the field of experimental design. The visible mechanics of the solution does not harm the presentation-quality display, which is obtained by setting the Print Area (indicated by the dashed line) to include just the plots and setting the Header (e.g., to "Complete block design experiment with single observation per cell") and the Footer (e.g., to "Effect of car, driver experience, and fuel additive on emission reduction") in the Page Setup dialog to the appropriate explanation.

4 Stem-and-leaf plot

Although stem-and-leaf plot epitomises the best of the pre-computer era of statistical graphics (Tukey, 1977), it should neither be under-estimated nor excluded from introductory statistics curricula. In addition to serving teaching purposes, it can replace tables of raw data (not limited to whole numbers) in publications. Like dot-density plot, it is a perfect candidate for construction with proper formatting of spreadsheet cells.

Count	Stem	Leaves	Number of cases each digit represents: 1			
5	0	23399				
2	1	56				
1	2	9				
	3					
2	4	15				
6	5	035788				
3	6	048				
3	7	355				
6	8	013559				
1	9	6				
1	10	4				

Figure 4: Sample output of the steam-and-leaf macro formatted for presentation.

The basic algorithm is straight-forward enough for ad-hoc manual implementation, whereby the amount of help from computer's calculation would depend on the user's proficiency with the spreadsheet application. For routine use, especially with larger amounts of data and larger numbers, macro automation is required. It was programmed by Maxwell (2006), but some debugging was needed and workbook design had to be added to make it a self-explanatory instructional tool that can also be used for producing publication-quality plots. Instructions for the user to run the macro either via menu or using keyboard shortcut, and to specify which column contains the data, are placed in the first worksheet. Output is produced in a separate worksheet, which also provides instructions for

formatting the plot (Figure 4 presents the result for the sample data) and copying it to the clipboard as a picture.

5 Box-cox transformation

The implementation of the modified Box-Cox transformation towards normal distribution is shown in Figure 5. The modification (subtraction of λ instead of 1; equation 5.1) of the originally proposed transformation (Box and Cox, 1964) is clearly displayed in the worksheet. The reference for the Excel implementation (Swanson, Tayman and Barr, 2000) is also provided, together with the correction of the formula for the maximum-likelihood estimation of the parameter λ .

$$Tx = \begin{cases} (x^\lambda - \lambda)/\lambda & ; \lambda \neq 0 \\ \ln(x) & ; \lambda = 0 \end{cases} \quad (5.1)$$

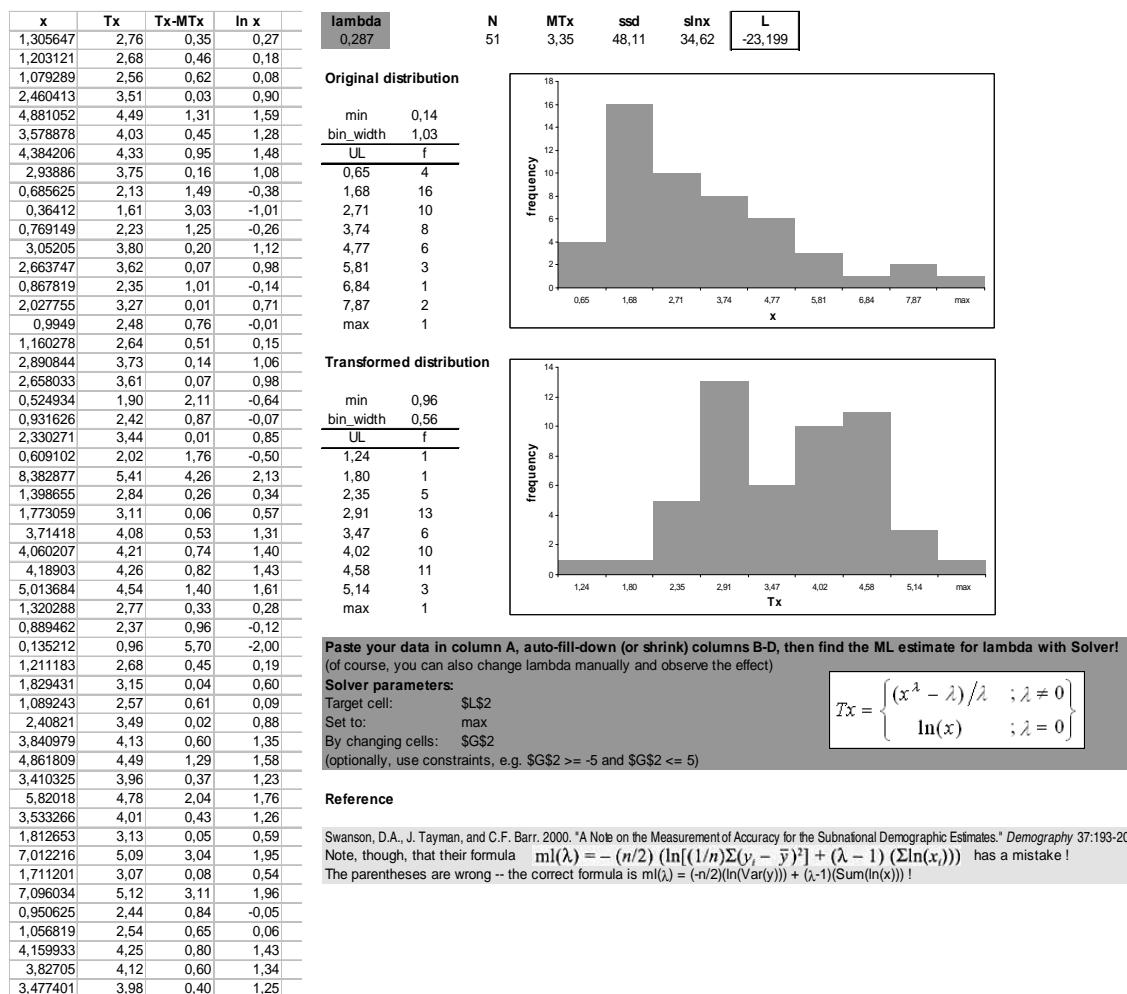


Figure 5: Worksheet for modified Box-Cox transformation.

Automated histogram binning (based on the FREQUENCY array function) and updating is particularly instructive if the user inputs various values of λ manually and observes the effect on the transformed distribution. At the same time, the implementation can be used to introduce the user to maximum-likelihood estimation and the Solver add-in. Although the Solver is far from an ideal and universal optimisation tool, it is very useful for a wide array of applications in statistics, probability, operation research, econometrics and related fields, such as teaching generalised linear models (Graham, 2000) or performing robust regression (Barreto and Maharry, 2006).

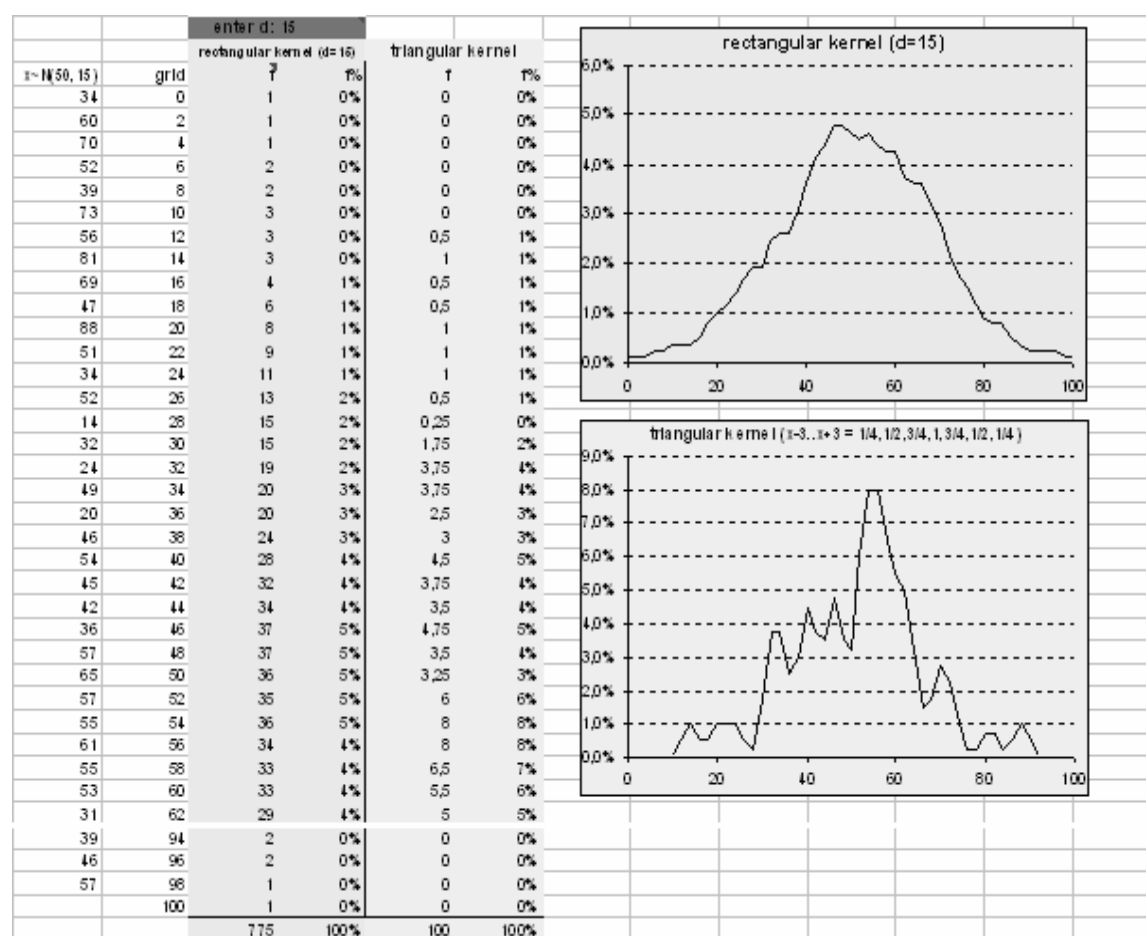


Figure 6: Worksheet for demonstrating windowgrams using rectangular and triangular kernel (15 rows of data cut out).

The essential feature of the presented solution that makes it useful for data analysis practice is that the value of N is based on counting the cells with the x data, and that all the vector quantities involved in the calculations are addressed

via named ranges that refer to the OFFSET function based on N . Hence, input data of practically any length can be entered, imported or pasted into the x column.

6 Windowgram

Smoothing is one of the simplest examples of a graphical method that can substantially facilitate understanding of a given problem, or even enable insight that is far superior to what students can infer from parametric statistical models they are familiar with (Weldon, 2005). Windowgrams provide the most gentle and the least mathematically oriented introduction to kernel density estimation; the latter has long been an essential part of statistics, yet it is too seldom taught in introductory statistics courses (Weldon, 2004).

The instructional spreadsheet, which features rectangular kernel (with adjustable window width) and triangular kernel (with 7-point window, as shown in the plot title) is presented in Figure 6. Both windowgrams are implemented only with conditional summing formulas. In-cell comments provide further guidance and explanation: pointing the cursor to the cell for entering the kernel width displays the message " ≥ 3 (larger value produces more smoothing)", while the comment of the cell with the heading f for rectangular kernel gives an excerpt from the Excel's help on some commonly used array formulas using the SUMIF, COUNTIF, SUM and IF functions. As shown in the worksheet, the data are 50 values sampled from a normal distribution with $\mu = 50$ and $\sigma = 15$. As an exercise, the user can find the smallest d that produces a completely flat density estimate with rectangular kernel.

7 Coplot

Coplots – short for conditional plots – are another name for what is usually called panel plots, trellis display (Becker, Cleveland and Shyu, 1996) or lattice graphics. Like windowgrams, they are arguably under-used in introductory statistics courses (Weldon, 2004). Similarly, the interactive Excel solution (Figure 7) demonstrates that they can be implemented in a spreadsheet without programming and hence more likely understood by non-mathematicians.

The comments in the headers of the columns a and b explain how the data is generated (7.1). Excel's Data Validation feature is used to guide break-point entry upon selecting the input cells: since c ranges from 1 to 50, the lower value is limited to the interval $[1, 46]$ and thus the valid range for the upper value is between the lower value + 1 and 48. To emphasise that non-linearity should regularly be considered when studying relations between a pair of quantities, quadratic trend is fitted to the points in each panel (using Excel's convenient

routine built into the scatter-plot). The scatter-plot of the total sample below the panels allows clear visual comparison with the conditional plots.

$$\begin{aligned}
 &a \sim U(0,1) \\
 &b \sim U(0,1) + \left\{ \begin{array}{l} 0 \quad ; c \leq \text{lower} \\ 2a^2 \quad ; \text{lower} < c \leq \text{upper} \\ 4a^2 \quad ; c > \text{upper} \end{array} \right\} \quad (7.1)
 \end{aligned}$$

As with multi-way dot plots, print area is limited to the plots of interest so that a display for presentation is readily available for printing or export (the later is easily achieved by printing to a file with appropriately set-up printer driver). Conditional cell formatting is used to highlight the data points displayed on the first and the second pair of adjacent panels.

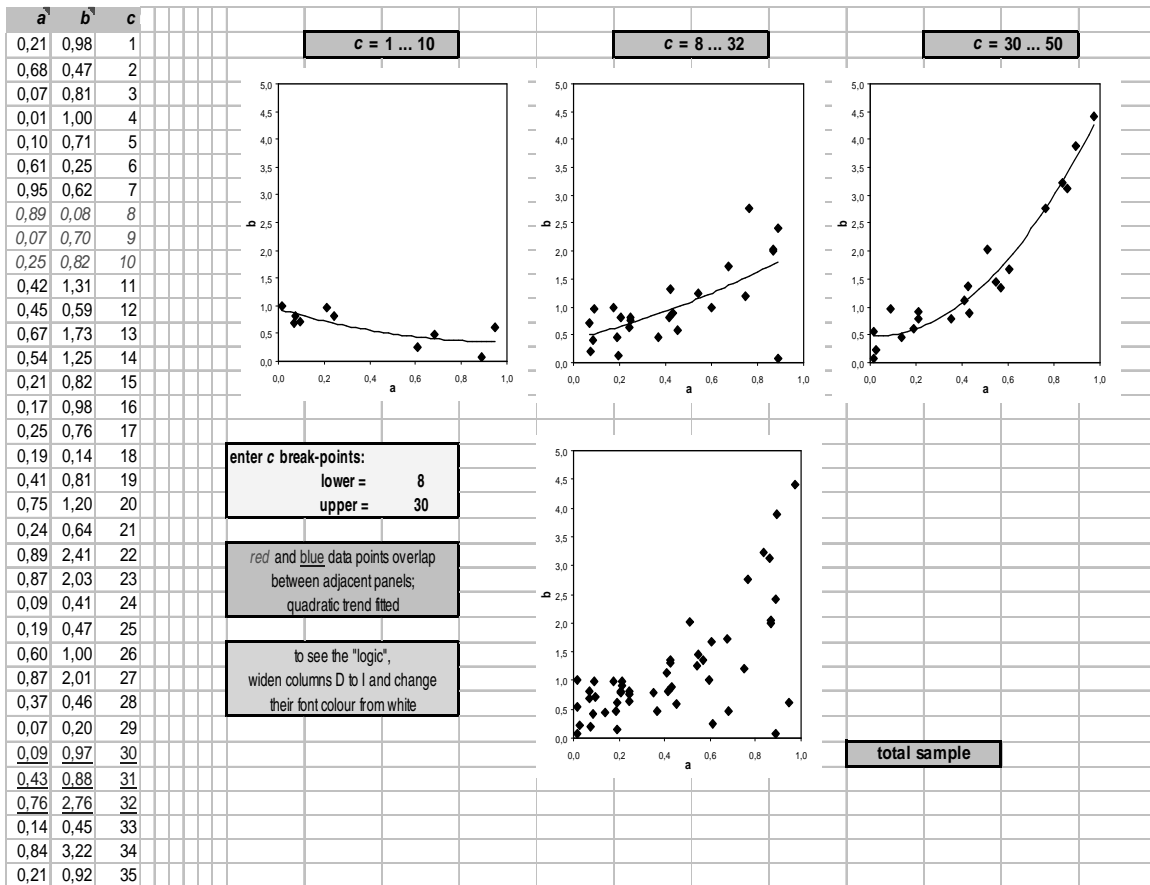


Figure 7: Worksheet for coplot demonstration (bottom 15 rows of data cropped).

x1	x2	x3	etc	
0,872562	0,377655	0,724187	0,099585	Select the data (two or more columns) and run the macro (from the menu, or press CTRL+SHIFT+P)!
0,160465	0,731002	0,874199	0,488534	Empty cells (i.e., missing data) are allowed, but make sure that there are no completely empty rows!
0,140685	0,01553	0,315819	0,18952	The plot is placed on the same worksheet as the data. The width of the plot is proportional to the number of variables (i.e., columns of the selection). The plot is drawn in black-and white with care for high data-ink ratio.
0,87716	0,750209	0,883424	0,751245	
0,798468	0,186594	0,052813	0,046079	
0,964605	0,643516	0,708466	0,389557	If you adjust some properties of the vertical axis, make sure you apply the same changes to the right-hand side of the plot (i.e., to the secondary value axis, which applies only to the first data-point, i.e., to the first series). The easiest way for that is Edit→Repeat (make the change to the primary axis, select the secondary axis and press CTRL+Y). If you add a title to the vertical axis, there is, of course, no need to do that.
0,597506	0,068571	0,067457	0,706973	
0,315575	0,420207	0,445867	0,834996	
0,474247	0,050423	0,514941	0,702131	
0,79045	0,68973	0,423104	0,458557	
0,221567	0,695294	0,13626	0,315879	
0,216896	0,907696	0,968733	0,538041	
0,343077	0,440352	0,008427	0,366289	
0,570123	0,022933	0,714774	0,919678	
0,393155	0,983995	0,220599	0,351666	
0,390496	0,42753	0,000103	0,167513	
0,275702	0,869384	0,741858	0,716521	
0,879767	0,506125	0,630438	0,972763	
0,402041	0,380847	0,729398	0,236713	
0,658159	0,954299	0,104217	0,394362	
0,992494	0,861495	0,419193	0,864391	
0,496892	0,748462	0,707383	0,5271	
0,184875	0,448685	0,232206	0,62227	

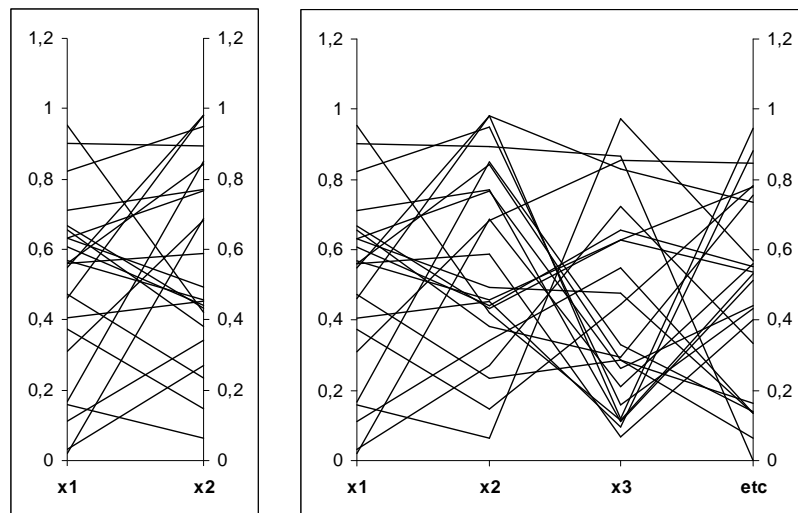


Figure 8: Instructions and sample data for producing parallel coordinates plots (top), and the resulting plot of two (bottom left) and four variables (bottom right).

8 Parallel coordinates plot

Parallel coordinates plot (Inselberg, 1985), sometimes also classified as a type of profile plot (Harris, 2000), has become very popular for dynamic visual exploration or large datasets with the developments in data mining, information visualisation and related fields. Its traditional version is an essential tool for analysis of repeated measurements, which are frequent in biomedical research.

The presented implementation (see Figure 8 for instructions to the user and two examples of produced plots) is static and aimed at quickly producing presentation graphics. Considering how tedious it is to manually produce such a plot in Excel, the macro gives Excel users the possibility to routinely accompany

paired-samples comparisons, method-agreement studies or similar analyses with useful graphics. Given the ease of data reordering and transformations within a spreadsheet, the macro also provides possibilities for parallel-coordinates exploration of larger datasets with more variables (Wegman, 1990).

9 Excel 2007 and beyond

The characteristics of the new, 2007 edition of Microsoft[®] Office, which includes the new Excel 2007, have been revealed in detail in the extensive official blog (Gainer, 2006). The improvements to the basic limitations are drastic and constitute a major leap of Excel into the present-day world of vast amounts of rapidly accumulating information. It is also a leap far beyond the capabilities of any other competing software. First and foremost, worksheet size has been increased from 65536 rows by 256 columns (i.e., $2^{16} \times 2^8 = 2^{24}$ cells) to 1048576 rows by 16384 columns ($2^{20} \times 2^{14} = 2^{34}$ cells, i.e., 1024-fold). Furthermore, the maximum number of worksheets in a workbook is limited only by available memory, and total usable memory has been increased to 2GB (i.e., doubled from the 2003 version). Excel's capabilities regarding storage capacity of a single cell, various aspects of formulas, conditional formatting, printing, searching, filtering and Pivot Tables (which now have the same size limit as the new worksheets) have also been vastly increased (see Gainer, 2006, for details).

On the other hand, the changes to the data visualisation facilities can be labelled as mainly cosmetic. Although the charting engine has been completely rewritten to improve its organisation, consistency and appearance, no really new or statistically sounder chart types have been introduced, despite plenty of qualified requests and suggestions over more than a decade of Excel's popularity. The only segment where a real break-through has been made is visualisation associated with conditional cell formatting. Versatile in-cell bars, gradient-coloured backgrounds, ordered icons sets, top-down rules and cell highlighting rules have been introduced, and in conjunction with the abovementioned capability increases these features can be used for visualisation of large categorical datasets, large-scale business reporting, or a variety of hierarchical or pixel-oriented information visualisation techniques (cf. Keim, 2000).

As far as statistical profession is concerned, the ideal expansion of Excel is its integration with statistical software packages. Probably the most advanced two, i.e., R (R Development Core Team, 2004) and XploRe (Härdle, Klinke, and Müller, 1999) have been successfully connected with Excel using the Component Object Model (COM) technology (an overview is given by Aydınli, Härdle and Neuwirth, 2003). Most notably, the R (D)COM Server and the RExcel add-in (Baier and Neuwirth, 2006), which can be further combined with the R Commander (Fox, 2005) package to allow interfacing with R via menus and dialogs, are opening up real possibilities for making the learning curve for R less

steep. Focusing on visualisation, this connection makes a variety of R's plots readily available to the Excel user. It can also be used to combine Excel's user interface (including controls, such as sliders) and charting capabilities with R's computational power to create dynamic visualisations and illustrated simulations.

10 Conclusion

The starting point of this paper has been pinpointed by one of the most prominent statisticians of today: "*Let's not kid ourselves: the most widely used piece of software for statistics is Excel*" (Ripley, 2002). And it is enough to open just about any newspaper to realise that the observation also applies to graphical presentation of statistical data.

Since Excel's default settings and ready-made procedures for charting compare even less favourably to statistical software than do its data analysis capabilities, this is undoubtedly a problem. However, the stance of the author of this paper is that like any other problem, it is actually an opportunity for finding a solution. In this case, there are various valid solutions at various levels, and it is reasonable to consider the presented Excel techniques among such solutions at the level of basic statistics education, as well as at the level of applying statistics in research and business practice.

References

- [1] Anderson, D.R., Sweeney, D.J., and Williams, T.A. (2005): *Modern Business Statistics* (with CD-ROM and InfoTrac). Mason, OH: South-Western.
- [2] Aydınlı, G., Härdle, W., and Neuwirth, E. (2003): Computational Statistics with Spreadsheets: Towards Efficiency, Reproducibility and Security. *Humboldt University Discussion Papers 373: Quantification and Simulation of Economic Processes*, (26).
<http://edoc.hu-berlin.de/oa/articles/reFDoa9WpdlyU/PDF/25diYZQdlKPLg.pdf>.
- [3] Baier, T. and Neuwirth, E. (2006): *R COM Connectivity*.
<http://sunsite.univie.ac.at/rcom/>.
- [4] Barreto, H. and Maharry, D. (2006): Least Median of squares and regression through the origin. *Computational Statistics & Data Analysis*, **50**, 1391-1397.
- [5] Becker, R.A., Cleveland, W.S., and Shyu, M.-J. (1996): The visual design and control of trellis display. *Journal of Computational and Graphical Statistics*, **5**, 123-155.
- [6] Box, G.E.P. and Cox, D.R. (1964): An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26**, 211-246.

- [7] Cleveland, W.S. (1985): *The Elements of Graphing Data*. Monterey, CA: Wadsworth.
- [8] Fox, J. (2005): The R Commander: A basic-statistics graphical user interface to R. *Journal of Statistical Software*, **14**
<http://www.jstatsoft.org/v14/i09/v14i09.pdf>
- [9] Gainer, D. (2006): *Microsoft Excel 2007 (nee Excel 12): A Discussion of What's New in Microsoft Excel 2007*. <http://blogs.msdn.com/excel/default.aspx>.
- [10] Graham, J. (2000): Regression using Excel's Solver. In Electronic Proceedings of the Thirteenth Annual International Conference on Technology in Collegiate Mathematics, Atlanta, 2000.
<http://archives.math.utk.edu/ICTCM/EP-13/C13/html/paper.html>.
- [11] Härdle, W., Klinke, S., and Müller, M. (1999): *XploRe Learning Guide*. Heidelberg: Springer.
- [12] Harris, R.L. (2000): *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.
- [13] Hawkes, J., Marsh, W. (2005): *Discovering Statistics* (2nd ed.). Charlotte, NC: Hawkes Learning Systems and Quant Systems.
- [14] Heiser, D.A. (2005): *Microsoft Excel 2000 and 2003 Faults, Problems, Workarounds and Fixes*. <http://www.daheiser.info/excel/frontpage.html>.
- [15] Hunt, N. and Tyrrell, S. (1995): *DISCUS – Discovering Important Statistical Concepts Using Spreadsheets*.
http://www.mis.coventry.ac.uk/research/discus/discus_home.html.
- [16] Inselberg, A. (1985): Plane with parallel coordinates. *Visual Computer*, **1**, 69-97.
- [17] Keim, D.A. (2000): Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, **6** (1), 59-78.
- [18] Kyd, C. (2006): *An Excel Tutorial: Compare Metrics by Category Using Excel Dot Plot Charts*. <http://exceluser.com/dash/dotplot.htm>.
- [19] de Levie, R. (2004): *Advanced Excel for Scientific Data Analysis*. New York: Oxford University Press.
- [20] Liengme, B. (2002): *Guide to Microsoft Excel 2002 for Scientists and Engineers* (3rd ed.). Oxford: Butterworth Heinemann.
- [21] Maxwell, N. (2006): *Data Matters with Excel*. Emeryville, CA: Key College Publishing. <http://www.keycollege.com/DM/activities/excel/index.html>.
- [22] McCullough, B.D. and Wilson, B. (1999): On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics and Data Analysis*, **31**, 27-37.

- [23] McCullough, B.D. and Wilson, B. (2002): On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. *Computational Statistics and Data Analysis*, **40**, 713-721.
- [24] Myerson, R.B. (2005): *Probability Models for Economic Decisions* (with CD-ROM, 1st ed.). Belmont, CA: Thomson Brooks/Cole.
- [25] Neuwirth, E. and Arganbright, D. (2004): *The Active Modeler – Mathematical Modeling with Microsoft Excel*. Belmont, CA: Brooks/Cole.
- [26] O'Day, D.K. (2006): *Excel Dot Plots*.
http://processtrends.com/pg_charts_dot_plots.htm.
- [27] Peltier, J. (2006): *Dot Plots*. <http://peltiertech.com/Excel/Charts/DotPlot.html>.
- [28] R Development Core Team (2004): *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
<http://www.R-project.org>.
- [29] Ripley, B.D. (2002): *Statistical Methods Need Software: A View of Statistical Computing* (RSS2002 Opening lecture).
<http://www.stats.ox.ac.uk/~ripley/RSS2002.pdf>.
- [30] Robbins, N.B. (2006): Dot Plots: A Useful Alternative to Bar Charts. *Business Intelligence Network Newsletter*.
<http://www.b-eye-network.com/newsletters/ben/2468>.
- [31] Steppan, D., Werner, J., and Yeater, B. (2001): *Essential Regression and Experimental Design* (Release 2.219, Excel 95/97/2000/2002/2003 Add-In and Electronic Book Package).
<http://www.geocities.com/SiliconValley/Network/1032/>.
- [32] Swanson, D.A., Tayman, J., and Barr, C.F. (2000): A note on the measurement of accuracy for the subnational demographic estimates. *Demography*, **37**, 193-201.
- [33] Tufte, E.R. (1983): *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
- [34] Tukey, J.W. (1977): *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.
- [35] Wainer, H. and Thissen, D. (1981): Graphical data analysis. *Annual Review of Psychology*, **32**, 191-241.
- [36] Wegman, E. (1990): Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, **85**, 664-675.
- [37] Weldon, K.L. (2004): *Some Under-Used, But Simple and Useful, Data Analysis Techniques*. <http://www.math.sfu.ca/~weldon/papers/32.simple.pdf>.
- [38] Weldon, K.L. (2005): From data to graphs to words - but where are the models? In Proceedings of the ISI/IASE Satellite Conference, Sydney, Australia, 2005. <http://www.math.sfu.ca/~weldon/papers/35.words.pdf>.

This tutorial will walk you through plotting a histogram with Excel and then overlaying normal distribution bell-curve and showing average and standard-deviation lines. To produce my random normal samples I used VBA function RandNormalDist by Mike Alexander. I created samples with a mean of 100 and standard deviation of 25, function RandNormalDist(100, 0.25). The actual mean and standard deviation was 100.84 and 27.49 respectively. Normally distribution. The samples can be checked to confirm normally distributed by comparing the mean, median and mode which should all be equal. Mean. Note that Excel wants the range address to use absolute references " hence, the dollar signs. In the Output Options area of the Descriptive Statistics dialog box, describe where and how Excel should produce the statistics. To indicate where the descriptive statistics that Excel calculates should be placed: Choose from the three radio buttons here " Output Range, New Worksheet Ply, and New Workbook.