Computational Statistics and Data Analysis 52 (2008) 4594–4601

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

It's easy to produce chartjunk using Microsoft®Excel 2007 but hard to make good graphs

Yu-Sung Su

Department of Political Science, The Graduate Center, The City University of New York, 365 Fifth Avenue, New York, NY 10016, USA

ARTICLE INFO

Article history:
Available online 13 March 2008

ABSTRACT

The purpose of default settings in a graphic tool is to make it easy to produce good graphics that accord with the principles of statistical graphics, e.g., [Tufte, E.R., 1990. Envisioning Information. Graphics Press, Cheshire, Conn, Tufte, E.R., 1997. Visual Explanations: Images and Quantities, Evidence and Narrative, 2nd Edition. Graphics Press, Cheshire, Conn, Cleveland, W.S., 1993. Visualizing Data. Hobart Press, N.J Cleveland, W.S., 1994. The Elements of Graphing Data, rev. edition. AT&T Bell Laboratories, Murray Hill, N.J, Wainer, H., 1997. Visual revelations: Graphical tales of fate and deception from Napoleon to Ross Perot. Copernicus, New York, Spence, R., 2001. Information Visualization. ACM Press & AddisonWesley, New York, and Few, S., 2004. Show Me the Numbers. Analytic Press, Hillsdale, NJ]. If the defaults do not embody these principles, then the only way to produce good graphics is to be sufficiently familiar with the principles of statistical graphics. This paper shows that Excel graphics defaults do not embody the appropriate principles. Users who want to use Excel are advised to know the principles of good graphics well enough so that they can choose the appropriate options to override the defaults. Microsoft® should overhaul the Excel graphics engine so that its defaults embody the principles of statistical graphics and make it easy for non-experts to produce good graphs.

Published by Elsevier B.V.

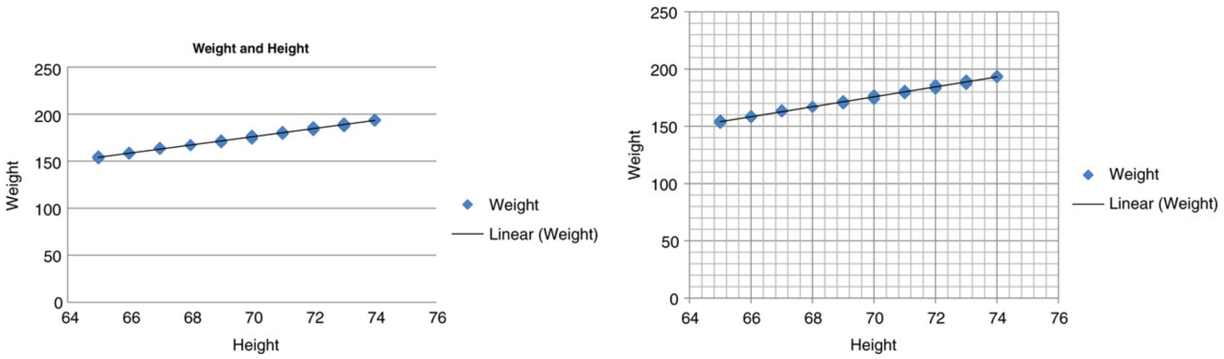
> "Data graphics should draw the viewer's attention to the sense and substance of the data, not to something else. The data graphical form should present the quantitative contents. Occasionally artfulness of design makes a graphic worthy of the Museum of Modern Art, but essentially statistical graphics are instruments to help people reason about quantitative information" (Tufte, 1997, p. 91).

1. Introduction

A well-constructed chart can bring the data to life, revealing the substantive information. A well-designed graphic tool makes it easy for users to achieve this goal. Microsoft®Excel performs poorly in this respect. On the one hand, statisticians seldom use Excel because it is not a sophisticated statistical package, unlike other alternatives (e.g., SAS®, SPSS®, S-PLUS®, STATA®, and R), it does not have programmable procedures to do statistical analysis and make graphs. On the other hand, common users who just want to glimpse and/or pre-process data or make some simple graphs are attracted by the WYSIWYG (the acronym of What You See Is What You Get) nature of Excel. Users can easily modify the design of a chart by clicking buttons. They do not have to type or remember codes or commands for making graphs. Nonetheless, convenience should not supersede the accuracy and clarity of demonstrating the data through graphs. In particular, the default chart types in Excel often distort graphics with redundant symbols, fill-ins and other extraneous graphical elements. Edward Tufte describes these kinds of redundant "interior decoration of graphics" as *chartjunk* (1997, p. 83) because they do not tell viewers anything new about the data.

E-mail address: ys463@columbia.edu.

0167-9473/$ – see front matter Published by Elsevier B.V.
doi:10.1016/j.csda.2008.03.007



(a) Layout 1

(b) Layout 3

Fig. 1. Scatter chart.

Microsoft® has just released Excel 2007 and claims that the new graphics engine has enhanced Excel's graphical capabilities. Various reviews of Excel 2007 have shown that, the novel graphical interface notwithstanding, Excel 2007 continues to provide users with the clutter and fluff in its default chart types (Few, 2006; Walkenbach, 2007). This paper evaluates Excel's graphical capabilities on its default chart types and, since there is no obvious improvement, the following evaluation is applicable to the previous Excel products (Excel 95, Excel 2000, Excel XP, and Excel 2003).[1] This paper is not a manual about how to make good statistical graphics in Excel (See O'Day (2007), Peltier (2007) and Vidmar (2007) and various blog entries in Juice Analytic, http://www.juiceanalytics.com), nevertheless, it employs the philosophy of what counts as good graphics as discussed in works by Tufte (1990, 1997), Cleveland (1993, 1994) and Wainer (1997) when assessing various Excel charts.

2. Defaults matter!

Defaults matter because most users rely heavily on the default settings of the software they use. Nowadays, computing power and various graphical software have advanced to a level where everyone can be an artist, creating various graphics by him/herself. However, greater power comes with greater responsibility. Excel should guide its users toward good graphical design and not toward making junk charts. Microsoft should be more careful about the default settings on Excel charts.

Fig. 1 plots the linear relationship between height and weight of 40 people using the default scatter chart with two different default Excel layouts. These plots remind us how crude an Excel chart can look. Here, and also in the plots of the following sections, I try to adjust the Excel default settings as little as possible so that I can convey the idea of how nonsensical the default graphical outputs look. Some compromises and adjustments had to made though, but *chartjunk* remain.

This type of chart is useful when we want to display a relationship between two variables. Fig. 1 shows that weight and height are positively related. The overlying regression line depicts this linearity. These charts are poorly designed for several reasons:

- The over-stretched axes hinder viewers' ability to read the data. The exaggerated y-axis simply distorts the data revelation. Sometimes we do want the scale to go all the way to the zero point because zero has its meaning in certain cases. This is not the case for this example. No one is expected to have zero weight, thus it is legitimate to start the y-axis from the minimum weight. Rescaling the y-axis in this way will help viewers identify more data points from the chart. In the current design, users can only observe 10 points where in fact there are 40.
- The redundant grid lines and the excess axis labels create nothing but clutter that messes up the graphical area. The purpose of the grid lines is supposed to help viewers better identify the absolute position of the data points, especially when they are dispersed over the plotting rectangle. Because this is not the case here, the grid lines are distracting. In particular, Fig. 1(b) overly uses grid lines, which obscure the crux of this chart—the data. Likewise, both the y-axis and x-axis have too many labels. This is messy.
- The legend and labels are not only superfluous; they are also wrong. The dots here certainly do not only mean weight. Rather, they represent 40 different people with the relative coordination of their weight and height. Also, the line indicator here is redundant. A simple graph like this does not need a legend and labels. Even when labeling is necessary, a better strategy is to use direct labeling as opposed to a legend.
- The filled diamond data points fail to reveal the overlain data points. A better choice would be to use open circles.

[1] In contrast to previous Excel products, Excel 2007 has reduced the number of its default chart types from fourteen to seven; the new categories are: Column, Line, Pie, Bar, Area, Scatter, and other charts.

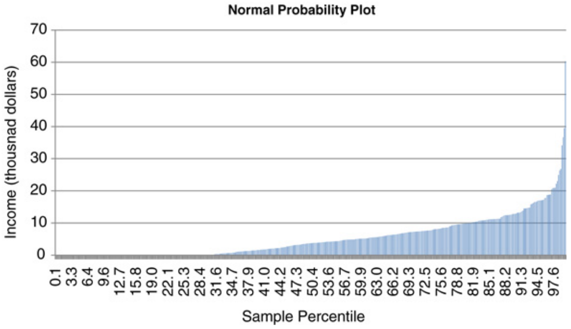


Fig. 2. Normal probability plot.

- The colour scheme can be too light for viewers to read and is also not suitable for publication. In particular, two axis lines are light grey in colour. This implies that the two axes are graphical elements equal in importance to the major grid lines, which is a wrong premise. Similarly, using light blue colour for the data dots is not illuminating enough in print.
- The most untenable default option is that there are no tick marks in Fig. 1(a)! In some default layouts, Excel removes tick marks from the chart. Viewers are going to have a hard time reading Excel graphics without the help of tick marks because Excel tends to mark too many labels on axes. Use of both tick marks and axis labels is a better way to limit the excess use of axis labels.[2]

Fig. 2 is another poorly designed chart in Excel 2007. This type of chart is called a normal probability plot (Chambers et al., 1983), which is a graphical technique for visually evaluating whether or not a data set is normally distributed. The vertical axis is the ordered response value whereas the horizontal axis is normal order statistic median. Thus, if the data set is normally distributed, we would expect to see a straight line.[3] Excel 2007 keeps the option of making this plot in the "data analysis tool pack" add-in. The plot was a line chart in the previous versions of Excel. Excel 2007 makes it into a column chart. This is a wrong choice because the length of each bar does not mean anything and it might mislead the readers to think there is an accumulative pattern for the data. Nevertheless, the excess axis label is an untenable design.

In short, Excel's default chart types are problematic in that they embed superfluous graphical elements and design poorly. Unfortunately, Excel has more than just these defects. The following sections show the way in which Excel misleads users who will end up wasting their time creating fancy designs which do not help to make their data stand out. Instead, they create junk charts.

3. Create junk charts in Excel: Just by clicks of buttons!

In this section, I pretend to be a typical user who tries to make charts using Excel 2007. I select five charts to show the way in which the default charts in Excel can violate the principle of statistical graphics.[4] One common mistake of these charts is the use of three-dimensional (3-D) plotting. 3-D objects often overwhelm the comprehensibility of human eye. Ironically, they can attract users' as well as readers' attention easily. Nonetheless, this does not justify the use of 3-D plotting. In addition, 3-D plotting in Excel often conveys null information and always distorts the data. As shown in Fig. 3, when users scroll down a chart menu, the options immediately next to the 2-D chart types are a variety of 3-D chart types. Excel leads users to produce bad graphical designs just by clicks of buttons.

3.1. Line chart

A line chart is useful to spot trends of continuous data. Sometimes, a line chart is a good choice when you want to compare multiple data series on the same baseline. Fig. 4 shows the monthly expense of six different items from January to June, plotted as a 3-D line chart. This type of chart obscures the data display. Viewers cannot make the comparison of multiple data series because of the overlapping lines. The additional axis, which labels the data series, fails to identify the six data series clearly. Nevertheless, the width of each line is *chartjunk*, which shows null information.

[2] See Cleveland (1994) for the general principles of using tick marks and axis labels, and Wilkinson (2005) for getting The optimal number of axis labels and tick marks by rescaling the data.

[3] The normal probability plot in Excel is incorrectly programmed: it shows a straight line for uniform data, not for normal data (See McCullough and Heiser, 2008, Section 5).

[4] I use R (R Development Core Team, 2007) to create fake data to make the following charts.

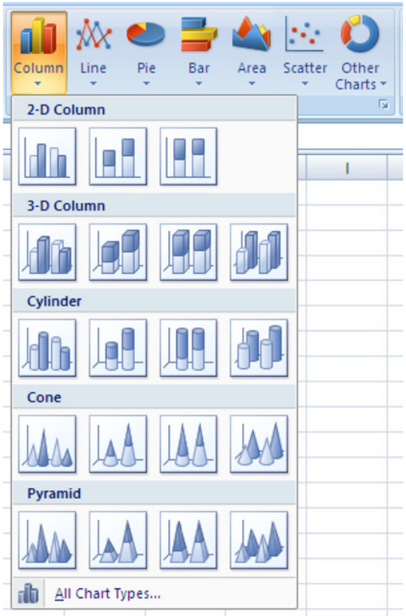


Fig. 3. The chart menu in Excel 2007.

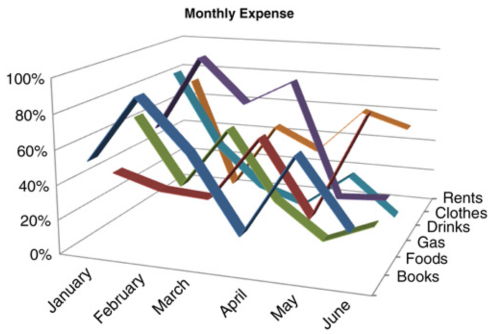


Fig. 4. Line chart.

3.2. Pie chart

A pie chart is useful when users want to show relative proportions of a whole. However, a pie chart is almost never effective.[5] Tufte (1997) wrote, "the only worse design than a pie chart is several of them" (p. 178). The fact that a pie chart can only convey a single series of the data means that a dot plot is an excellent alternative because it can show data positions along a common scale rather than rely on pie chart angles. Fig. 5 compares the proportion of aid money among five major donor countries, plotted as a 3-D pie chart. The 3-D effect and the angle of this pie chart distort the data. Viewers would misidentify Japan (22%) over the US (33%) as the largest aid money donor. Another reason not to use a pie chart is that it wastes data ink. A pie chart uses color, size, angle, 3-D effect, and fill-in to convey only one single message from the data—the relative proportion to a whole. In this case, a table can do a better job than a chart even though graphs are usually better than tables in summarizing numerical information of the data. Don't use graphics to "merely decorate a few numbers" (Tufte, 1997, p. 83).

3.3. Area chart

An area chart is a special case of a line chart in which the area below the line has been coloured. An area chart is not an example of an effective chart, either. The comparison between data series is often misleading and obscured due to the size

[5] There is a lively debate about the cons and pros of the pie chart as a means to display data (Cleveland and McGill, 1984; Simkin and Hastie, 1987; Spence, 1990; Carswell et al., 1991; Rangecroft, 2003). Some argue that the pie chart is still useful in showing a good qualitative view of the data, particularly when there are not too many small pieces. For more recent discussion with this regard, please refer to Few (2007) and a blog entry by Zach Gemignani, "The problem with pie charts" and the comments posted to this entry, http://www.juiceanalytics.com/writing/2006/12/the-problem-with-pie-charts.

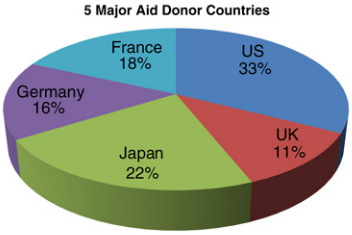


Fig. 5. Pie chart.

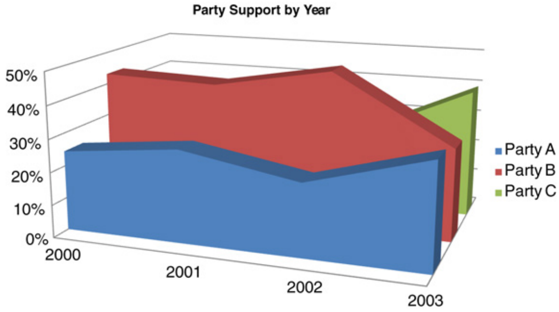


Fig. 6. Area chart.

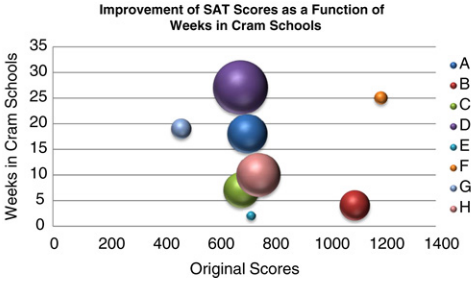


Fig. 7. Bubble chart.

of the areas. Fig. 6 shows the change of support rates for three parties from 2000 to 2003, plotted as a 3-D area chart. The use of the third dimension helps viewers to better identify three different parties, though at the expense of obscuring the data. In particular, the area of Party C is veiled by the area of Party B. Excel offers the solutions such as rotating the charts or using transparency. But these solutions are not going to work in print.

3.4. *Bubble chart*

A bubble chart is a special case of scatter chart. It displays an additional data series by altering the size of the bubbles. However, a bubble chart sometimes can be misleading because people perceive the mapping from value to the size of the bubbles in a way that is not easy to quantify. Fig. 7 shows the improvement in SAT scores for eight students as the function of weeks cramming in school, plotted as a 3-D bubble chart. The bubble chart shows the relative improvement in SAT scores effectively by the size of the bubbles. However, an open circle is a better choice than a colour-filled bubble because data can be masked due to overlaps. The legend on the right of the chart is redundant. A better legend would indicate what the size of the bubbles means. Nevertheless, the colour fill-in and the 3-D effect create a false impression in which these extra dimensions convey no information.

3.5. *Bar chart*

A bar chart displays each data point as a horizontal bar, the length of which corresponds to the value. A 90 degrees counterclockwise bar chart is a column chart. A bar chart is a superior choice over a column chart when you have a lengthy category label. Fig. 8 displays the change of monthly support rate for three candidates, plotted as a 3-D stacked pyramid chart, a subtype of the bar chart. The stacked pyramids fail to clearly show the relative proportion of the three candidates.

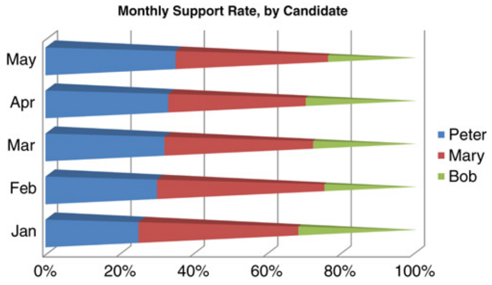


Fig. 8. Pyramid chart.

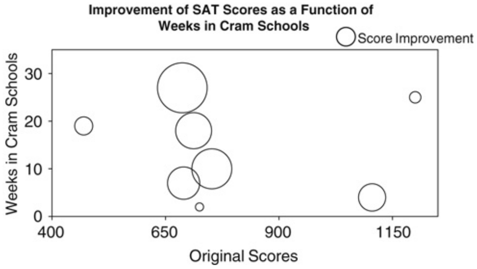


Fig. 9. Improved bubble chart.

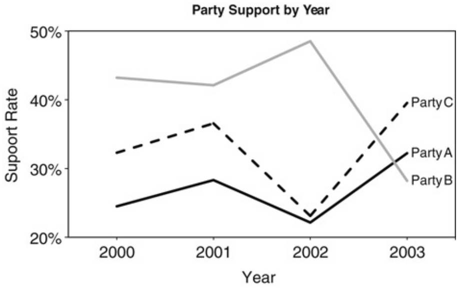


Fig. 10. Improved area chart using line chart.

In addition, the pyramid shape suggests that there are differences between the candidates in some perspective other than support rates. In fact, shape here is just *chartjunk*. It contains no information but distorts the data.

4. Effective charting in Excel is laborious!

It is disappointing that Excel 2007 did not provide any new chart types or make any significant improvement. In order to make charts that effectively display the data, users have to expend great effort to clean up the *chartjunk* in Excel defaults. Here are some steps to do this:

- No fill-in or unnecessary colouring and shadow.
- 3-D effect is seldom a good option to use.
- No busy grid lines and over-crowded labels on axes. Adjust scale for both axes and avoid overstretching axes.
- Tick marks are still necessary, especially when you have data that is dispersed over the plotting rectangle. Tick marks help viewers to better locate data in the relative coordinates.
- Direct labeling is a better alternative to identify data objects than a legend.

Fig. 9 is a revised version of Fig. 7 after cleaning up *chartjunk*. The chart effectively reveals the overlapped data points with open circle. The improved legend shows that the circle size means the improvement in SAT scores. The succinct labeling on both axes enables viewers to concentrate on reading the data information.

Fig. 10 is an improved version of Fig. 6. A line chart is more effective than an area chart in comparing different trend lines. Fig. 10 clearly demonstrates that 2002 and 2003 are the two years that mark distinctive changes among the three trend lines. The pattern is veiled in Fig. 6 because of *chartjunk*.

Fig. 11 is another example of effective charting that summarizes the relationships between an outcome variable and predictors. The plot shows the regression coefficients of various predictors on income, plotted as a stock chart.

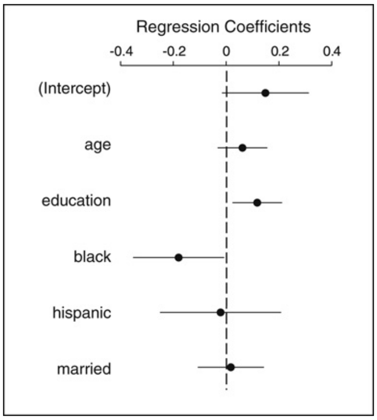


Fig. 11. Stock chart.

Dots represent regression estimates. Bars show the 95% confidence intervals of the estimates (or ±1.96 standard errors). If the intervals of an estimate cross the zero reference line, the estimate is statistically insignificant. Viewers do not have to count stars in a regression table. This figure demonstrates the way in which an effective chart can help viewers to digest numerical information.

To make junk charts in Excel, users just need a few clicks of the buttons. To clean up *chartjunk*, they require more than a few. Some Excel experts have written and developed Excel macro add-ins to do the clean-up jobs (See O'Day (2007), Peltier (2007) and Vidmar (2007)). However, users could save time and energy by searching for and installing these macros. They could have focused more on their works had they had an excellent chart tool at the beginning. A good graphic design certainly helps effectively bring data to life. Excel just does not offer a good guidance, instead it gives users many other chart styles that fail to communicate.

5. Conclusions

A properly-designed chart can help people get the most information out of data and an excellent graphic tool helps people to achieve this task without much labour. The purpose of default settings in a graphic tool is to make it easy to produce good graphics that accord with the principles of statistical graphics. If the defaults do not embody these principles, then the only way to produce good graphics is to be sufficiently familiar with the principles of statistical graphics (Tufte, 1990, 1997; Cleveland, 1993, 1994; Wainer, 1997; Spence, 2001; Few, 2004).

This paper has shown that the default chart types in Excel 2007 do not embody these appropriate principles. Instead, these charts create *chartjunk* that hinder peoples' ability to comprehend the data. Some users have developed add-ins and instructions to redress the malfunctions of the Excel default chart settings and thereby enable users to produce better graphics (O'Day, 2007; Peltier, 2007; Vidmar, 2007). Nevertheless, those who want to use Excel are advised to get to know the principles of good graphing well enough so that they know how to choose the appropriate options to override the defaults. Microsoft® should overhaul its graphics engine so that it embodies the principles of statistical graphics and makes it easy for non-experts to produce good graphs.

Acknowledgements

I would like to thank Dr. Andrew Gelman for his advices on the earlier draft and the suggestion for the title. I would also like to thank Stephen Few for his comments and Masanao Yajima for his inputs on this paper. Finally, I would like to thank three anonymous referees for their valuable comments and suggestions on how to improve this paper.

References

Carswell, C.M., Frankenberger, S., Bernhard, D., 1991. Graphing in depth: Perspectives on the use of three-dimensional graphs to represent lower-dimensional data. Behaviour & Information Technology 10 (6), 459–474.

Chambers, J.M., Cleveland, W.S., Kleiner, B., Tukey, P.A., 1983. Graphical Methods for Data Analysis. PWS Publishers: Duxbury Press, Boston.

Cleveland, W.S., 1993. Visualizing Data. Hobart Press, N.J.

Cleveland, W.S., 1994. The Elements of Graphing Data, rev. edition. AT&T Bell Laboratories, Murray Hill, N.J.

Cleveland, W.S., McGill, R., 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of American Statistical Association 287, 531–554.

Few, S., 2004. Show Me the Numbers. Analytic Press, Hillsdale, NJ.

Few, S., May 2006. Excel's new charting engine: preview of an opportunity missed. http://www.perceptualedge.com/articles/b-eye/excels_new_charting_engine.pdf.

Few, S., August 2007. Perceptual Edge, Save the pies for dessert. http://www.uperceptualedge.com/articles/visual_business_intelligence/save_the_pies_for_dessert.pdf.

McCullough, B.D., Heiser, D.A., 2008. On the accuracy of statistical procedures in microsoft excel 2007. Computational Statistics & Data Analysis 52 (10), 4570–4578.
O'Day, K.D., 2007. Data analysis and visualization with excel tools and charts. http://processtrends.com/.
Peltier, J., 2007. Jon's excel and charting pages. http://peltiertech.com/.
R Development Core Team,, 2007. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Rangecroft, M., 2003. As easy as pie. Behaviour & Information Technology 22 (6), 421–426.
Simkin, D., Hastie, R., 1987. An information-processing analysis of graph perception. Journal of the American Statistical Association 82 (398), 454–465.
Spence, I., 1990. Visual psychophysics of simple graphical elements. Journal of Experimental Psychology: Human Perception and Performance 16, 683–692.
Spence, R., 2001. Information Visualization. ACM Press & Addison Wesley, New York.
Tufte, E.R., 1990. Envisioning Information. Graphics Press, Cheshire, Conn.
Tufte, E.R., 1997. Visual Explanations: Images and Quantities, Evidence and Narrative, 2nd edition. Graphics Press, Cheshire, Conn.
Vidmar, G., 2007. Statistically sound distribution plots in excel. Metodološki Zvezki 83–98.
Wainer, H., 1997. Visual revelations: Graphical tales of fate and deception from Napoleon to Ross Perot. Copernicus, New York.
Walkenbach, J., 2007. Excel 2007 Bible. Wiley, Indianapolis, IN.
Wilkinson, L., 2005. The Grammar of Graphics, 2nd Edition. Springer, New York.