Computational Statistics and Data Analysis 52 (2008) 4602–4606

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Teaching statistics with Excel 2007 and other spreadsheets

John C. Nash*

Telfer School of Management, University of Ottawa, Ottawa, Ontario K1N 9B5, Canada

ARTICLE INFO

Article history:
Available online 12 March 2008

ABSTRACT

This article considers which activities in teaching statistics may be suitable candidates for the application of spreadsheets, and whether spreadsheets in general and Excel 2007 in particular are suitable for these tasks.

© 2008 Elsevier B.V. All rights reserved.

1. A menu

This article considers the tools we use to help teach statistics, especially spreadsheets. Among spreadsheet processors, the most popular is Microsoft Excel, of which a new flavor has recently been released, which we will call Excel 2007.

In order to organize the discussion below, I will try to address the following subjects:

- What activities are involved in teaching where a spreadsheet might be used?
- Are spreadsheets in general appropriate or inappropriate to these tasks, or could they be made so?
- Where a spreadsheet might be appropriate, is Excel 2007 a good choice?

In approaching these questions, I will try to separate those matters I consider fundamental from those that are preferences, and attempt to provide good justifications for such distinctions. However, it is worth noting that style and tradition often trump substance and evidence, so the former may ultimately be the main driver of adoption of a tool, even the wrong tool (Bacon, 2006; Lyon, 2002). Bloodletting endured as a medical treatment for a very long time. Unfortunately, the debate on whether spreadsheets are suitable for use in teaching statistics has also endured, and this paper will in many places repeat arguments made before (see, for example, the content and bibliography of Dell'Omodarme and Valle (2006), as well as material referenced in the next section).

2. Context

I have been interested in spreadsheets since they were introduced in the early 1980s and have taught statistics in various venues since 1969. Nash (2006) provides references to most of my earlier work on spreadsheets and statistics, including the proof-of-concept TellTable project (www.telltable.com). I have done some work to provide test spreadsheets for Gnumeric (Nash and Goldberg, 2005). Being involved in studying open source knowledge generation models, and favoring platform independent tools, I do not regularly use Microsoft Office.

Since this paper is one of a set (see McCullough (2008), McCullough and Heiser (2008), Su (2008) and Yalta (2008)) some topics are given a very light treatment here.

3. What activities are involved in teaching where a spreadsheet might be used?

Let us first eliminate the obvious negatives, and conclude that spreadsheets are not suitable for mathematical statistics. Tools such as Maple or Mathematica are much more relevant in that area. However, the vast bulk of students and the major

* Tel.: +1 613 236 6108.
E-mail address: nashjc@uottawa.ca.

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.03.008

burden of teaching concerns introductory one- or two- term courses of study in other programs. This implies that we are going to need to support data management, computations and graphics for fairly simple data sets, and spreadsheets have often been chosen as the tools.

Spreadsheets get used widely in the management of teaching tasks, but as that is not specific to the teaching of statistics, I will not include such uses here, even though it has been the major application of spreadsheets in my own teaching.

4. Are spreadsheets appropriate to the tasks?

4.1. General issues

The spreadsheet paradigm, that is, results change when data is changed, is an extremely useful one for quickly testing ideas and verifying possibilities. However, most spreadsheet processors do not maintain this mode of operation across all their features. A number of computational dialogs have been introduced in different spreadsheet processors (with variation by version of the processor) where the output remains static even when the input data is changed. In Excel 2007, the Data Analysis/Regression dialog is one such tool. Using some data gathered in a teaching exercise for length and width of students' middle fingers I ran a simple regression of length versus width, then changed one length from 95 mm to 495 mm. The regression output cells did not change. However, when I drew an *XY* graph of the data and added a trend line, then made the same change, the trend line did change. (Thanks to a referee for pointing this out.) Here we have a mixed paradigm for the same calculation. This is like having a brake lever that sometimes acts as an accelerator—pure danger. This is not a criticism of spreadsheets in general, but of unfortunate choices made in their implementation. So readers may try tests like this, my data is available at http://macnash.admin.uottawa.ca/~nashjc/allfingerx.xls.

Data retrieved from databases may raise even more such problems due to timing issues, since active databases do get updated regularly. Here we begin to need a concurrent version control system like CVS (http://www.nongnu.org/cvs/) or Subversion (http://subversion.tigris.org/). These tools allow multiple users to simultaneously edit files, with appropriate features and safeguards to maintain the integrity of the content, allowing for reconciliation of conflicts and recovery of different versions.

4.2. Classroom teaching of introductory statistics

With variation in detail and depth, most introductory applied statistics courses cover the following topics:

- data
- univariate, bivariate and occasionally three dimensional statistical graphics
- descriptive statistics
- probability
- discrete and continuous probability distributions
- sampling distributions
- hypothesis testing for one or two samples
- interval estimation
- count data (goodness of fit)
- analysis of variance (at least one-way, sometimes 2-way)
- regression, at least simple, often multiple

We will look very briefly at the suitability of spreadsheets for each of these.

4.2.1. Data

Modern spreadsheet processors are capable of handling tables of data that are well beyond the size needed for teaching. For the most part the user simply "enters the data" and the spreadsheet processor algorithmically decides how to handle and format it. This can lead to some unfortunate errors if the data should be treated as text but is handled as numbers, or if it is meant to record a date or time. Date handling has been the subject of much discussion over the years (e.g., O'Beirne (2000)), but are a technical matter that could be resolved with appropriate standardization activity. I personally recommend dates be recorded as numbers using a format such as YYYYMMDD to ensure proper sort order, that is, later dates sort as higher numbers. This avoids relying on software-defined date handling. While really an admission of defeat, it bypasses concerns of "locale" where we need to worry about the month/day/year versus day/month/year forms and so forth. Furthermore, there is no need to explain mathematical and software details that are beyond the scope of most undergraduates. Unfortunately, software is sometimes very fascist in "helping you" by automatically formatting cells. For example typing the characters 2007/11/17 in my Excel 2007 results in a cell with the entry 17/11/2007. If you have your regional setting with English (USA) rather than English (Canada), you will get 11/17/2007. This is why it is advisable to leave out the slashes.

A statistical issue, which again could be resolved, is that spreadsheets do not handle missing values very well (Kirschenbaum, 2006). Worse, some users hide data by various tricks such as tiny fonts or setting the same foreground and background colors for cells. This can be used, for example, to avoid displaying intermediate or verification calculations to leave a more esthetically pleasing screen. Even without trickery, users may put spaces or other whitespace characters into cells that can then be acted upon in ways that are not equivalent to an undefined cell, or may hide data accidentally when entering colors or font sizes.

4.2.2. Univariate, bivariate and occasionally three dimensional statistical graphics

I have heard my students saying that they like "Excel" (usually synonymous with "spreadsheet" to them—an idea that suggests a student survey assignment whose results could interest trademark lawyers) for the graphics. However, I would contend that spreadsheets do not do a particularly good job compared to statistical software. Consider the ease of creating a histogram in either Minitab (www.minitab.com) or Rcmdr (http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr) with Excel or Gnumeric. Boxplots are highly useful as summaries of distributions and particularly for comparisons, but are not available in Excel 2007, though they are a feature in Gnumeric. In Gnumeric, I found that I seem to need to group the data myself in order to get multiple boxplots; this is not very helpful when I want to do a quick analysis.

4.2.3. Descriptive statistics

Spreadsheets seem to carry out descriptive statistics calculations satisfactorily. There are also functions listed for correlation and covariance. We may want to ask what they do with missing values (see above). When there are missing values and more than two variables, there are sensible choices to only include observations that are complete or to compute pairwise correlations using all observations that have both variables recorded. I hope to build some test spreadsheets to uncover the behavior of different spreadsheet processors, and welcome assistance in this and similar efforts.

4.2.4. Probability

Spreadsheets are useful tools for dealing with most of the elementary probability topics taught in a typical statistics course.

4.2.5. Discrete and continuous probability distributions

For teaching purposes, the probability functions of spreadsheets are one of their most useful features. Moreover, unlike a calculator, we can review and edit the expressions, alter them, copy them, and so forth. There are two fringe issues. First, the spreadsheet processor may lack particular functions, though this ought to disappear in the next few years. For example, Excel lacks the cumulative hypergeometric probability; Gnumeric has a boolean switch to offer individual or cumulative hypergeometric probabilities. Second, and for the present review secondarily, the computed values may be incorrect in some way. For example, the normal distribution is computed as a body rather than tail area, so applications requiring tail areas may be problematic. The body area suffices for teaching in almost all cases, but may be insufficient for some specialized work. Try computing $(\text{normsinv}(\text{normdist}(z)) - z)$ for values of z from 8 to 8.3. While "better" functions may not always be necessary, we do need better documentation and didactic material to inform our use of software such as a spreadsheet processor.

4.2.6. Sampling distributions

Spreadsheets do quite well in illustrating sampling distributions. For Excel 2003, McCullough and Wilson (2005) had serious concerns regarding the quality of random number generators, but these are unlikely to be paramount in a teaching situation, though there is a danger in introducing students to a less-than-adequate tool for real-life work. Moreover, most statistical packages are more suitable here.

4.2.7. Hypothesis testing for one or two samples

Spreadsheets offer some tools for hypothesis testing, but in my view they are so obviously "patched" onto the application as to be less than ideal for teaching. I much prefer the statistical packages and how they present the possibilities. My principal dissatisfaction is that the user must construct the whole test – the hypothesis, the computation of the test statistic and the setup of the call to the distribution function – and then call such a function to get the equivalent of the "table value" one used to look up in the back of the textbook. Contrast this with the interface and output provided by Minitab or Rcmdr. For example, using only a few mouse clicks, the latter produces

```
>t.test (Davis$ weight, alternative= ‘greater’,mu=63, conf.level=.95)
One Sample t-test
data: Davis$ weight
t=2.6232, df=199, p-value=0.004692
alternative hypothesis: true mean is greater than 63
95 percent confidence interval:
64.03611 Inf
sample estimates:
mean of x
65.8
```

Note that the interface generates the command from a graphical user interface, so the user can later on reuse this command in a script to save effort. Also observe that we can talk of variables and observations so that the test does not need us to specify a set of cells for the test, but instead allows the user to focus on the ensemble of data we call a 'variable'; while spreadsheets allow named ranges, we think of variables in teaching statistics.

4.2.8. *Interval estimation*

Here the spreadsheets are, at present, weak. They offer a "CONFIDENCE" function that gives the margin of error for a $100^*(1 - \text{alpha})\%$ interval (with default 95%) for a single-sample mean based on the Gaussian distribution.

4.2.9. *Count data (goodness of fit)*

Spreadsheets supply functions for the Chi-square distribution, but as in hypothesis tests do not seem to allow for the appropriate preparation of the count data so that goodness of fit and contingency tables can be quickly tested. Given that the spreadsheet is naturally a tabular form, this should make us recognize that the interests of those who build the spreadsheet processors are not aligned with those of us who teach and practice statistics.

4.2.10. *Analysis of variance*

ANOVA has been part of spreadsheets for quite a while, and the tools have been evolving. Single factor, two factor, and two factor with replication ANOVA is supported in Gnumeric and Excel to an extent that is suitable for course teaching. However, the interface may not suit how we wish to present material in courses. If the partitioning of the sum of squares is important in a course, then (and I am grateful to a referee for noting this), one may wish the computational details, as shown in http://faculty.vassar.edu/lowry/ch13pt1.html. If data is changed, users may want to ensure that the output from the dialog has been updated.

4.2.11. *Regression*

Regression is supported by most spreadsheets to an extent suitable for teaching. Personally, I do not find the interface convenient. As mentioned above, changing the input data does not change the regression output.

5. Is Excel 2007 a good choice as a spreadsheet processor?

Let me state up front that I do not like the 2007 incarnation of Excel. Violating the "If it ain't broke, don't fix it" and the KISS principles, Excel 2007 seems little more than the overstuffed toy box of a spoiled child. More than 40 icons at the top of a screen are confusing. Worse, the clutter makes it very difficult to find the few things one does need.

Do we need this clutter? Most of us use only a few features and we likely work most efficiently if we can effectively place, size and alter them. Surely this is the essence of personalization. As the menus are stored in XML, there is a possibility of altering them, but this does not seem to be easy as yet. My suspicion is that a lively third-party market for menu templates and tools will develop, even to return Microsoft Office menus to their classic appearance (e.g., for Access, there is http://software.techrepublic.com.com/download.aspx?docid=301725). When I searched "Help" using "customize menu" I was offered a party planner, a weekly meal planner, and a dinner party journal. This is worse than unhelpful.

Turning to statistics teaching, among the plethora of eye-candy graphs, I do not find histograms or boxplots, though histograms are in the Data Analysis add-in. Among the "new" graphical tools, I was told by Aaron Erickson of the possibility of colored bars scaled to the sizes of the numbers in the cells. These are activated by "Conditional formatting" in Excel 2007. Unfortunately, there is an example from Computerworld where these seem to be scaled improperly (http://www.computerworld.com/html/office_2007b/p11.html). In a sense, such bars provide a sort of histogram, so if properly implemented this could be a useful idea.

Most of the statistical tools in Excel are performed in the Data Analysis Add-in. Under Excel 2007, one needs to find this, yet the Add-ins tab at the top of the screen only shows the default add-ins. The Data Analysis Add-in is installed to the "Data" tab. Also one needs to go to the unnamed "Office" button and choose "Excel options" to activate this add-in. This is three places to look.

Turning to the add-in itself, it is worth mentioning that "histogram" is a very strange function for statisticians in that it is essentially a 'tally' or frequency of observations in ranges defined by the user as bin boundaries. There is no graph, and the binning is manual. Hardly a great statistical tool.

When one does finally find a tool, there is the concern about whether it does its job properly. McCullough (2004) argues that the sluggish response of Microsoft in fixing errors compromises the software. As McCullough shows, the Gnumeric programmers (who are but a few part-timers) were able to fix in several weeks the same errors that Microsoft was unable and/or unwilling to fix in several years. Excel's statistical capabilities have been the subject of many complaints about errors and inaccuracies over many years. Finding and properly documenting such errors or weaknesses in scientific and statistical software is a difficult and tedious job, and reporting on them takes many pages of highly technical detail. Our personal and professional gratitude is owed to those who carry out such studies, such as those reported by McCullough (2004), McCullough and Wilson (2005) and McCullough (2008) and the apparently ongoing and large study by David Heiser (http://www.daheiser.info/). From the point of view of accuracy, my opinion is that Excel 2007 provides sufficient accuracy for most of the tasks in elementary statistics courses, but that it is very poor pedagogy to teach students to use a tool that is inadequate for "real-life" use.

6. Conclusion

Spreadsheets are powerful tools for end-user computing (see, for instance, the program of the EuSpRIG 2004 conference at http://eusprig.org/eusprig-2004-conference-programme.pdf). However, the developers and vendors have vastly different agendas from statistics teachers, so we should not expect them to be well suited to our needs in the classroom. Most of us need tools that work well and offer clean, unambiguous interfaces. This means that for most spreadsheet applications I will use Gnumeric. However, for statistics, I used to use Minitab. With more students using Macintosh and Linux where a bundled textbook/Minitab package has not been available to us, I now would use the open-source R, either natively or via the Rcmdr interface.

References

Dell'Omodarme, M., Valle, G., 2006. Teaching statistics with Excel and R, ArXiv Physics e-prints, ref. physics/0601083. http://arxiv.org/pdf/physics/0601083.pdf.

Bacon, Jono, 2006. Context vs. content. http://www.oreillynet.com/pub/wlg/9236, Feb. 23, 2006.

Kirschenbaum, Leif, 2006. Missing value representation in Excel. http://tolstoy.newcastle.edu.au/R/help/06/01/18871.html, 11 Jan. 2006.

Lyon, Jack M., 2002. Content vs. presentation. http://www.planetpublish.com/mainpage.asp?webpageid=231, June 7, 2002.

McCullough, B.D., 2004. Fixing statistical errors in spreadsheet software: The cases of Gnumeric and Excel, CSDA Statistical Software Newsletter. www.csdassn.org/software_reports.cfm.

McCullough, B.D., Heiser, David A., 2008. On the accuracy of statistical procedures in Microsoft Excel 2007. Computational Statistics and Data Analysis 52 (10), 4570–4578.

McCullough, B.D., 2008. Microsoft's 'Not the Wichmann-Hill' random number generator. Computational Statistics and Data Analysis 52 (10), 4587–4593.

McCullough, B.D., Wilson, Berry, 2005. On the accuracy of statistical procedures in Microsoft Excel 2003. Computational Statistics and Data Analysis 49 (4), 1244–1252.

Nash, J.C., 2006. Spreadsheets in statistical practice—another look. The American Statistician 60 (3), 287–289.

Nash, J.C., Goldberg, Jody, 2005. Why, how and when spreadsheet tests should be used. In: Ward, David (Ed.), Proceedings of the EuSPrIG 2005 Conference: Managing Spreadsheets in the light of Sarbanes-Oxley. EuSpRIG, Greenwich UK, pp. 155–160.

O'Beirne, Patrick, 2000. Spreadsheet year 2000 issues; Tips, traps and answers to Frequently Asked Questions (FAQ). http://www.sysmod.com/y2ksprds.htm.

Su, Yu-Sung, 2008. It's easy to produce Chartjunk using Microsoft Excel 2007 but hard to make good graphs. Computational Statistics and Data Analysis 52 (10), 4594–4601.

Yalta, A.Talha, 2008. The reliability of statistical distributions in Microsoft Excel 2007. Computational Statistics and Data Analysis 52 (10), 4579–4586.