THE UNRELIABLITY OF EXCEL'S STATISTICAL PROCEDURES

by Bruce D. McCullough

Bruce McCullough is on the faculty of Decision Sciences at Drexel University in Philadelphia and is Software Editor for the International Journal of Forecasting. He has written extensively on the accuracy of statistical and econometric software, and his paper "Is it safe to assume that software is accurate?" won the Best Paper 2000-2001 award in the International Journal of Forecasting.

Introduction

In the small world where computer science overlaps with statistics, it was well known that Microsoft Excel was riddled with statistical errors. It was so well known that no one bothered to write about it. In the larger world, however, it remained Microsoft's dark secret. Professional statisticians wrote textbooks with titles like "Statistics with Excel," and a generation of students learned to do statistics with Excel. "Surely," the student reasoned, "it is safe to use Excel for statistics. If it weren't, my professor would have chosen a different software package." So these students went on to use Excel in the business world. It is quite conceivable that more statistical calculations are performed in Excel than in any statistical software package.

Testing the Accuracy of Statistical Software

Several years ago, I developed a methodology for testing the accuracy of statistical software (McCullough, 1998 and 1999), and I applied this method to some major statistical packages, including SAS, SPSS, and S-Plus. I found a few errors in each of them (McCullough, 1999). A coauthor and I applied the same methodology to Excel 97 (McCullough and Wilson, 1999), and we found numerous errors. So egregious were these errors that we advised people who conduct statistical analyses of data not to use Excel.

The scope of these errors is not minor. My methodology analyzes three areas: random number generation, estimation (which has four components: univariate, ANOVA, linear regression, and nonlinear regression), and statistical distributions (for example, tabulating the normal distribution or calculating *p*-values). Excel failed in all three areas.

In the estimation area, we found Excel wanting in all four components. When we applied Excel Solver to 27 problems in the nonlinear least squares regression suite, Solver gave incorrect answers 21 times. In fact, it missed completely 21 times. For example, it returned a coefficient of 454.12 when the correct answer is 238.94. Rick Hesse and others have found errors in specific functions that I did not examine, such as the LINEST, TREND, LOGEST, and GROWTH worksheet functions.

Microsoft's Track Record

It's not as if Microsoft would have to develop new algorithms to solve these problems. For most of the inaccuracies, good algorithms have already been developed and are well known in the statistical community. Microsoft simply used bad algorithms to begin with, and it never bothered to replace them with good algorithms. Revision after revision, in Excel 4.0, Excel 5.0, Excel 95 through Excel 97 and beyond, Microsoft has allowed the errors to persist—unbeknownst to its legions of users.

So unbelievable was Microsoft's cavalier attitude toward accuracy that I came to believe (McCullough, 2002) the company might be catering to a demand for inaccurate statistical software. There is simply no other way to explain Microsoft's lack of response. Contrast Microsoft's behavior with that of a responsible software company such as SAS. When SAS becomes aware of an error, it publishes the error on its Web site, often with a workaround, so that users can avoid the problem. SAS fixes the problem quickly, often by the next minor release, and almost always by the next major release. And SAS fixes problems correctly.

In its Excel XP release, Microsoft attempted to fix some statistical problems, but it did not do a good job

(McCullough and Wilson, 2002). This failure presaged Microsoft's attempt at a major overhaul with Excel 2003. While it fixed many functions, it failed to fix many others.

Perhaps most embarrassing was Microsoft's attempt to install a new random number generator (RNG). In its natural state, the RNG should produce numbers between zero and one. Microsoft chose a very well-known RNG (called the Wichmann-Hill RNG), but could not make it work right: Excel would occasionally spit out negative numbers. What makes this so embarrassing is that the source code for this algorithm is very easy to obtain. Hence it is fair to say that Microsoft did not correctly implement an algorithm for which source code is widely available. Nor did it do adequate testing before releasing the product. In our analysis of Excel 2003, we wrote that "Excel 2003 is an improvement over previous versions, but not enough has been done that its use for statistical purposes can be recommended" (McCullough and Wilson, 2005, p. 1244). Assuming that Microsoft will make another attempt to fix Excel, given Microsoft's track record, it will not be enough for the company to say that it has "fixed" errors. Microsoft will have to prove that it has fixed them correctly.

Warnings, Faults, and Workarounds

Professional statisticians continue to write books with titles like "Statistics with Excel," but they now warn students not to bet their jobs on Excel's accuracy. They advise students to use a real statistical package when they need to do statistics.

If Dante had to conjure a new circle for the 21st century, it would contain persons condemned to do statistics with Excel. What are these poor, unfortunate souls to do? To their succor has come a retired engineer who, in a tour de force, has catalogued Excel's statistical errors and offered many workarounds. These can be found at David A. Heiser's Web site entitled "Microsoft Excel 2000 and 2003: Faults, Problems, Workarounds, and Fixes," which is located at

http://www.daheiser.info/excel/frontpage.html

In this issue of *Foresight*, Rick Hesse provides another example of Microsoft's decision to use a bad algorithm and its refusal to fix this problem over the years. Fortunately for those who have to use Excel, Professor Hesse also provides a workaround. Note that while Professor Hesse does use Excel Solver, he has verified the results using SAS.

References

McCullough, B. D. (2002). *Proceedings of the 2001 Joint Statistical Meeting* [CD-ROM]: *Does Microsoft fix errors in Excel?* Alexandria, VA: American Statistical Association.

McCullough, B. D. (1999). Assessing the reliability of statistical software: Part II. *The American Statistician, 53*(2), 149-159.

McCullough, B. D. (1998). Assessing the reliability of statistical software: Part I. *The American Statistician, 52*(4), 358-366.

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

McCullough, B. D. & Wilson, B. (2002). On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. *Computational Statistics and Data Analysis, 40*(4), 713-721.

McCullough, B. D. & Wilson, B. (1999). On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics and Data Analysis, 31*(1), 27-37.

Contact Info:
Bruce D. McCullough
Department of Decision Sciences
and Department of Economics
Drexel University
bdmccullough@drexel.edu

No statistical procedure in Excel should be used until Microsoft documents that the procedure is correct; it is not safe to assume that Microsoft Excelâ€™s statistical procedures give the correct answer. Persons who wish to conduct statistical analyses should use some other package. Â© 2008 Elsevier B.V. All rights reserved. Real Statistics Using Excel. Everything you need to perform real statistical analysis using Excel .. â€¦ â€¦ .. Â© Real Statistics 2020. Skip to content.Â Excel Capabilities. Matrices and Iterative Procedures. Linear Algebra and Advanced Matrix Topics. Other Mathematical Topics. But properties of statistical procedures for network structures identication may depend on distribution of X âˆˆ K. Problem: construct distribution free statistical procedure for network structure identication. Denition: statistical procedure Î´ is distribution free in class K, if risk function Risk(S, Î¸, Î´) does not depend from distribution of vector X from class K for any S. Petr Koldanov (NRU HSE). S. Peterburg, Russia, April 14, 2017 12.