Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 49 (2005) 1244–1252
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

On the accuracy of statistical procedures in Microsoft Excel 2003

B.D. McCullough[a,*], Berry Wilson[b]

[a] *Department of Decision Sciences, LeBow College of Business, Drexel University, Philadelphia, PA 19104, USA*
[b] *Finance and Graduate Economics Department, Lubin School of Business, Pace University, New York, NY 10038, USA*

Received 21 January 2004; accepted 21 June 2004
Available online 13 August 2004

Abstract

Some of the problems that rendered Excel 97, Excel 2000 and Excel 2002 unfit for use as a statistical package have been fixed in Excel 2003, though some have not. Additionally, in fixing some errors, Microsoft introduced other errors. Excel's new and improved random number generator, at default, is supposed to produce uniform numbers on the interval (0,1); but it also produces negative numbers. Excel 2003 is an improvement over previous versions, but not enough has been done that its use for statistical purposes can be recommended.

Keywords: Benchmarks; Software; Strd; TestU01

1. Introduction

McCullough (2002) posed the rhetorical question, "Does Microsoft Fix Errors in Excel?" and it appears that the answer is, "Yes, but not very well." McCullough (1998) proposed a methodology for assessing the reliability of statistical software in three areas: statistical distributions, estimation, and random number generation. Statistical distributions (e.g., for calculating *p*-values) are assessed using Knüsel's (1989) ELV program. Estimation is assessed using the "Statistical Reference Datasets" produced by the (American) National

* Corresponding author. Tel.: 2158952134; fax: 2158952907.
E-mail addresses: bdmccullough@drexel.edu (B.D. McCullough), bwilson@pace.edu (B. Wilson).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.06.016

Table 1
Inverse normal distribution, $P(X < x) = p$

x	Exact	Excel		
		97/2000	*XP*	2003
0.001	−3.09023	−3.09024	−3.09025	Exact
0.0001	−3.71902	−3.71947	−3.71909	Exact
1E−5	−4.26489	−4.26546	−4.26504	Exact
1E−6	−4.75342	−4.76837	−4.75367	Exact
3E−7	−4.99122	−7.15256	−4.99152	Exact
2E−7	−5.06896	−5000000	−5.06928	Exact

Institute of Standards and Technology,[1] which has four suites of tests: univariate summary statistics, one-way ANOVA, linear regression, and nonlinear least squares. The random number generator (RNG) is subjected to empirical tests of randomness.

McCullough and Wilson (1999) applied this methodology to Excel 97, observed that Excel 97 was deficient in all three areas, and concluded that Excel should not be used for statistical analysis of data. McCullough and Wilson (2002) examined both Excel 2000 and Excel XP, and found no reason to change the conclusion. They did find, however, that Microsoft exhibited a tendency to "fix" these errors in less than acceptable ways, e.g., the inverse normal function and the normal RNG in the analysis toolpak (ATP). Since algorithms that will produce acceptable answers for these procedures are well known, replacing one defective algorithm with another is not evidence that the software developers are familiar with customary practices in the field for which they are writing software.

As can be seen in Table 1 , the inverse normal function in Excel 97/2000 was quite weak. Microsoft attempted to fix this in Excel XP, but did not do a very good job. The standard applied here, as described in Knüsel (1995), is that the program, at default, should display no inaccurate digits. If the program automatically displays six digits, then all six digits should be correct. If the algorithm is only accurate to two digits, then only two digits should be displayed so as not to mislead the user. As Knüsel (2004) shows, Microsoft finally fixed the problem correctly in Excel 2003.

Generating random normal variates in Excel 97 and Excel 2000, by either a function call (i.e., NORMSINV(RAND)) or by using the ATP, regularly produced a value of −50 00 000, though this was "fixed" in Excel XP: the −50 00 000 was changed to −9. Even −9 is far too large for a random normal, and should not be seen in a lifetime of generating random normals. See McCullough and Wilson (2002) for discussion of this problem. In Excel 2003, this problem has been fixed for the function call, but not for the ATP.

In this paper, we assess the reliability of Excel 2003, using the same methodology that we have employed previously. Section 2 discusses statistical distributions, Section 3 discusses the StRD, Section 4 discusses the RNG, and Section 5 presents the conclusions. We note that other researchers (Carlson, 2002; Cryer, 2002; Cook et al., 2000) have reported difficulties that we do not check, and these need to be investigated, too.

[1] http://www.itl.nist.gov/div898/strd.

Table 2
Poisson distribution with $\lambda = 200$, $P(X \leqslant k)$

k	Exact	Excel		Gnumeric	
		97/2000/*XP*	2003	v0.67	v1.1.2
0	1.3839E−87	Exact	0	No result	Exact
10	4.1096E−71	Exact	0	0	Exact
50	6.8158E−37	Exact	0	0	Exact
100	3.723 64 E−15	Exact	0	3.77476E−15	Exact
103	2.8916 E−14	Exact	0	2.86658E−14	Exact
104	5.6170 E−14	Exact	2.7254 E−14	5.61773E−14	Exact
110	2.4813 E−12	Exact	2.4524 E−12	2.48124E−12	Exact
133	2.943 90 E−07	Exact	Exact	Exact	Exact
134	4.456 17 E−07	No result	Exact	Exact	Exact
200	0.518 795	No result	Exact	5.18794E−01	Exact
250	0.999 715	No result	Exact	Exact	Exact

2. Statistical distributions

Knüsel (2004) analyzed the statistical distributions for Excel 2003. He found that although Microsoft had fixed several bugs, at least four distributions were unacceptable: Poisson, Binomial, Gamma and inverse Beta. Microsoft did not correctly fix bugs in the Poisson and Binomial distributions, and the Gamma and inverse Beta functions are not always computed correctly. The inverse Beta function is particularly troubling because, though Microsoft claims to have enhanced its accuracy, in Excel 2003 it still exhibits the same unacceptable behavior as in previous versions of Excel.

It is interesting to observe that the open source Excel-clone called "Gnumeric" (http://www.gnome.org/projects/gnumeric/) was such a good clone that it even had errors similar to Excel. However, the developers of Gnumeric, who are part-time volunteers with no R&D budget, chose to deal with these errors in a different way: by implementing correct fixes. See McCullough (2004a) for a discussion.

As can be seen in Table 2 , Excel's Poisson distribution returned no result for values in the central region near the mean of the distribution in old versions. In Excel 2003, Microsoft obtained an accurate answer in the central region of the distribution in exchange for inaccurate results in the tail. This is not a good "fix" and it is not an isolated incident; Microsoft traded accuracy in the central region for inaccuracy in the tail for the Binomial distribution, also. A good fix is demonstrated by Gnumeric, where an algorithm that was inaccurate in both the central region and the tail was exchanged for an algorithm that provides exact results in both areas. See McCullough (2004a) for a comparison of Gnumeric and Excel.

The performance of Excel in this area, statistical distributions, is still inadequate.

3. StRD

Each of the four suites of StRD tests contains several problems of varying degree of difficulty: low (l), average (a), and high (h). For each problem, NIST computed the correct

Table 3
StRD results for univariate summary statistics. This table shows the number of accurate digits for $\bar{x}$, s and ρ (the mean, standard deviation, and correlation coefficient)

Data set	Excel 97/00/02			Excel 2003		
	$\lambda_{\bar{x}}$	λ_s	λ_ρ	$\lambda_{\bar{x}}$	λ_s	λ_ρ
Pidigits (l)	15	15	15	15	15	13.6
Lottery (l)	15	15	15	15	15	15
Lew (l)	15	15	14.8	15	15	14.8
Mavro (l)	15	9.4	8.1	15	13.1	13.6
Michelso (l)	15	8.3	7.7	15	13.8	13.4
Numacc1 (l)	15	15	15	15	15	15
Numacc2 (a)	14.0	11.6	11.1	14.0	11.6	14.6
Numacc3 (a)	15.0	1.1	0	15	9.5	12.2
Numacc4 (h)	14.0	0	2.1	15	8.3	11.0

answer, say 'c', to several digits (15 digits for linear problems, 11 digits for nonlinear problems). For an answer produced by a statistical package, say, 'x', the number of correct digits can be calculated via the *log relative error* as

$$\lambda = \mathrm{LRE}(x) = -\log_{10}\left(\frac{|x-c|)}{|c|}\right).$$

For example, if $c = 6.54321$ and $x = 6.54300$, then $\mathrm{LRE}(x) = 4.5$ correct digits. Values of LRE less than unity should be set to zero.

3.1. Univariate summary statistics

This suite of tests has benchmark values for the mean ($\bar{x}$), the sample standard deviation (s) , and the correlation coefficient (ρ). The Excel commands for computing these quantities are: 'average', 'stdev' and both 'Correl' and 'Pearson'. Previously Microsoft had used two different algorithms for Correl and Pearson, though they compute the same quantity. The Pearson algorithm was weak, and it has been changed to the same algorithm that is used for Correl. Here we use the Pearson command.[2] Results are presented in Table 3 .

Excel had used an unstable algorithm for calculation of the sample variance and the correlation coefficient. Both these problems have been corrected. Excel's performance on this suite of tests is acceptable.

3.2. Analysis of variance

Since ANOVA produces many numerical results, here only the LRE for the final F-statistic is presented. Results are presented in Table 4 . Previously Excel had employed an

[2] Previously we had missapplied this correlation test to Excel, producing results that overstated accuracy for some datasets and understated it for others. Here we have corrected the error by using *Mathematica* (Wolfram, 1999) to compute benchmarks for the Pearson/Correl statistic. *Mathematica* can be used to compute benchmarks because it returns a perfect score on all four suites of the StRD (McCullough (2000b)).

Table 4
StRD results for ANOVA. This table shows the number of accurate digits in the final F-statistic

Data set	Excel 97/00/02	Excel 2003	Data set	Excel 97/00/02	Excel 2003
SiResist (l)	8.5	12.8	Simon5 (a)	1.1	10.2
Simon1 (l)	14.3	15	Simon6 (a)	0[a]	10.2
Simon2 (l)	12.5	13.9	Simon7 (h)	0[b]	4.2
Simon3 (l)	12.6	13.0	Simon8 (h)	0[a]	1.8
Simon4 (l)	1.7	10.4	Simon9 (h)	0[a]	0
AgWt (a)	1.8	10.2			

[a]Negative within group sum of squares.
[b]Negative between group sum of squares.

Table 5
StRD linear regression results. This table shows the number of accurate digits for the least accurate coefficient ($\hat{\beta}$) and the least accurate standard error thereof ($\hat{\sigma}$)

Data set	Old Excel		Excel 2003	
	$\lambda_{\hat{\beta}}$	$\lambda_{\hat{\sigma}}$	$\lambda_{\hat{\beta}}$	$\lambda_{\hat{\sigma}}$
Norris (l)	12.1	13.8	12.0	14.4
Pontius (l)	11.2	14.3	12.0	12.8
Origin1 (a)	14.7	15	14.7	15
Origin2 (a)	15	15	15	14.8
Filip (h)	0	0	7.2	7.2
Longley (h)	7.4	8.6	13.3	14.7
Wampler1 (h)	6.6	7.2	9.9	10.4
Wampler2 (h)	9.7	11.8	13.4	15
Wampler3 (h)	6.6	11.2	10.1	11.4
Wampler4 (h)	6.6	11.2	8.1	11.8
Wampler5 (h)	6.6	11.2	6.1	12.0

unstable algorithm, but this has been corrected. The zero digits of accuracy for the strenuous Simon9 test is no cause for concern, as this occurs when reliable algorithms are employed (see McCullough, 2000a for a discussion of this point). Excel's performance on this suite of tests is acceptable.

3.3. Linear regression

Since linear regression produces many numerical results, here only the lowest LRE for all the estimated coefficients ($\hat{\beta}$) and the lowest LRE for the standard errors of the coefficients ($\hat{\sigma}$) are presented. Results are presented in Table 5.

Previous versions of Excel either did not check for near-singularity of the design matrix, or did a bad job of checking, and so could return results that were so contaminated by rounding error that they contained not a single correct digit. See the result for the Filip problem. This has been corrected in Excel 2003. Excel's performance on this suite of tests is acceptable.

3.4. Nonlinear regression

This suite has 27 test problems. Excel 97, Excel 2000 and Excel XP did not perform well on this suite. At default, the Excel Solver returns solutions that have zero accurate digits 21 times. Tuned for better performance (automatic scaling invoked, and convergence tolerance set at 1E-7), Excel still produces solutions that have zero accurate digits for 14 of the problems. There have been no changes to this part of Excel, so Excel's performance on this suite of tests is still unacceptable. McCullough (2004b) gives a detailed example of how Solver stops at a point that is not a solution and nonetheless reports that it has found a solution.

3.5. Overall performance on the StRD

While the improvements made to the univariate, ANOVA and linear regression functions are most welcome, Excel still fails the nonlinear suite and, as such, cannot be said to perform well on the StRD. Consequently, *the performance of Excel in this area, the StRD, is still inadequate.*

4. Random number generator

Excel offers two RNGs, one in the ATP and another via a function call, RAND. In versions prior to Excel 2003, both RNGs were unacceptably bad and Microsoft made no changes to the ATP RNG for Excel 2003,[3] so we focus attention on the RNG for RAND. Microsoft claims[4] to have implemented the Wichmann–Hill RNG (Wichmann and Hill, 1982). However, the Wichmann–Hill RNG does not produce negative numbers. It has been reported in some newsgroups and in some press venues (e.g., *PC Magazine*, April 6, 2004, p. 71) that, at default, when RAND should produce numbers on the interval (0, 1), it sometimes produces negative numbers[5] and each of us has independently confirmed this phenomenon. However, even if Microsoft had correctly implemented the Wichmann–Hill RNG, it would still be unacceptable.

It is important that the RNG passes empirical tests; for a discussion see Gentle (2003, Section 2.3). The first standard battery of tests was produced by Knuth (1981), which was supplanted by Marsaglia (1996) DIEHARD tests. These, in turn, have been supplanted by L'Ecuyer and Simard's (2003) TESTU01 program. Gentle (2003, p. 80) recommends TESTU01, which has three batteries: Small Crush, Crush and Big Crush. By way of comparison, DIEHARD takes about 15 seconds to execute on a 1.7 GHz Pentium, while Crush takes

[3] See Microsoft Knowledgebase Article #829208.

[4] See Microsoft Knowledgebase Article #828795.

[5] See Microsoft Knowledgebase Article #834520. To produce negative numbers using RAND, create a worksheet with 2000 rows and 20 columns. Fill all the cells with RAND. Create a row that has the minimum of each column (this is much faster than taking the minimum of all the cells, and makes it easy to find the negative numbers). Hit F9 (recalculate) repeatedly. Once the negative numbers appear, they come in waves.

1 to 2 hours depending on the complexity of the RNG being tested. See McCullough (2004b) for a review of TESTU01.

Just as there was a shakeout amongst RNGs when DIEHARD replaced the Knuth tests, so there will be a shakeout now, as TESTU01 replaces DIEHARD. RNGs previously thought to be acceptable will now be adjudged as unacceptable. One such RNG is the Wichmann–Hill RNG, which passes all the DIEHARD tests but fails "birthday spacings" and "close pairs" tests in the TESTU01 Crush battery. The output from running the Crush test on the Wichmann–Hill RNG with seeds 26656, 2092, 3794 is given below:

```
Generator:            Wichmann–Hill
ulcg_CreateLCG, unif01_CreateCombAdd3
Number of tests:      60
Total CPU time:       02:56:36.79
The following tests gave p-values outside [0.01, 0.99]:
(eps means a value < 1.0e-15)
```

	Test	p-value
9	Multinomial Bits Over	0.9958
10	BirthdaySpacings ($t = 2$)	eps
11	BirthdaySpacings ($t = 4$)	eps
13	BirthdaySpacings ($t = 13$)	2.4e-3
14	ClosePairs ($t = 2$)	eps
15	ClosePairs ($t = 4$)	3.6e-14

```
All other tests were passed
```

As can be seen, four tests produce *p*-values smaller than 1E–10, and so can be considered catastrophic failures: the Wichmann–Hill, when properly implemented, fails empirical tests.

Another important consideration is period length, i.e., the number of calls that can be made to the RNG before it begins to repeat. Microsoft claims[6] that the period is 10E13, and cites the original Wichmann and Hill (1982) article. However, Microsoft neglected to do a complete investigation of the random number generator that it selected, as the erratum to the original article (Wichmann and Hill, 1984) states that the period of the Wichmann–Hill generator is only 6.95E12 ($\approx 2^{43}$). More importantly, even as long ago as 1994, 10E13 was considered a short period length: L'Ecuyer (1994) recommended that the period length of an RNG be at least 2^{60}. There is no discernable reason that Microsoft should have chosen a 20 year old random number generator, and ignored the literature on period length. Essentially, the only packages that use Wichmann–Hill are packages that have been using it for many years. Most developers who add a new RNG add a modern one. For example, the developers of Gnumeric recently upgraded their RNG, and chose the Mersenne-Twister (Matsumoto and Nishimura, 1998), which passes all the Crush tests and has a period of $2^{19\,937} - 1$.

Excel's performance in this area; *random number generation, is still inadequate.*

[6] See Microsoft Knowledgebase article #828795.

5. Conclusions

Microsoft's attempts to fix the many errors in its statistical procedures are welcome, but the results of Microsoft's efforts hardly inspire confidence. Microsoft has implemented many fixes for Excel 2003, and we have no idea which of these fixes were implemented correctly. Until all of Microsoft's claims have been independently verified—and the methodology that we have employed only begins this process of independent verification—persons who wish to use Excel for statistical purposes should exercise extreme caution. We note that persons who use the spreadsheet Gnumeric need not exercise such caution.

One could argue that it is acceptable to use Excel for summary statistics, one-way ANOVA, linear regression, and some of the statistical distributions, but we are extremely concerned about Microsoft's cavalier attitude toward accuracy. For example, the "inverse normal" function is important to Six Sigma methods for quality control. Excel's implementation of this function was known to be quite inaccurate; see our Table 1. Defending the use of an inaccurate algorithm, Microsoft says,[7] "When the functions were originally added to Excel, nobody could anticipate future uses. For example, Six Sigma techniques were not in widespread use." This is perfectly true, and it is perfectly irrelevant: it does not explain why Microsoft failed to upgrade its algorithms year-after-year and version-after-version, even after the Six Sigma became popular. As McCullough (2002, Section 5) observed, "If Microsoft intends for Excel to offer reliable statistical functionality, it should act in the same fashion as the purveyors of reliable statistical software." Specifically, upon becoming aware of a bug, Microsoft should make the existence of the bug known to users, so that they can avoid it or work around it until it is fixed. Microsoft should also make known its plans for fixing the bug—in the next minor upgrade or (no later than) the next major upgrade. Finally, when the bug is fixed (correctly, one would hope), complete details of the fix should be in the release notes of the software.

The cavalier attitude towards accuracy is underscored by the fact that Microsoft could not properly implement an RNG for which source code is readily available, and the quality (or lack thereof) of its "fixes" for its statistical distributions. These bad fixes raise the question of whether Microsoft is using its unwitting customers to debug its product, or whether Microsoft needs to adopt a quality control program. Either way, failure to provide adequate fixes for the problems with the statistical distributions and the random number generator, as well as continued failure on the StRD nonlinear suite, prevent us from revising our conclusion; Persons desiring to conduct statistical analyses of data are advised not to use Excel 2003.

Acknowledgements

We thank Connie Borror, John Nash, and an Associate Editor for useful comments.

[7] See Microsoft Knowledgebase article #828888.

References

Carlson, W., 2002. The use of Microsoft Excel for statistical purposes: a textbook author perspective. Proceedings of the 2001 Joint Statistical Meetings [CD ROM], American Statistical Association, Alexandria, VA.

Cook, H.R., Cox, M.G., Dainton, M.P., Harris, P.M., 2000. Testing the intrinsic functions of excel. NPL Report CISE 27/00, www.npl.co.uk/ssfm/download.

Cryer, J., 2002. Problems with using Microsoft Excel for statistics. Proceedings of the 2001 Joint Statistical Meetings [CD-ROM], American Statistical Association, Alexandria, VA.

Gentle, J.E., 2003. Random Number Generation and Monte Carlo Methods 2*e*. Springer, New York.

Knüsel, L., 1989. Computergestützte Berechnung Statistischer Verteilungen. Oldenburg, München-Wien (An English version of the program can be obtained at http://www.stat.uni-muenchen.de/∼knuesel/elv).

Knüsel, L., 1995. On the accuracy of statistical distributions in Gauss. Comput. Statist. Data Anal. 20, 699–702.

Knüsel, L., 2004. On the accuracy of statistical distributions in Microsoft Excel 2003. Manuscript.

Knuth, D.E., 1981. The Art of Computer Programming, vol. 2, Semminumerical Algorithms. Addison–Wesley, Reading, MA.

L'Ecuyer, P., 1994. Uniform random number generation. Ann. Oper. Res. 53, 77–120.

L'Ecuyer, P., Simard, R., 2003. TESTU01: a software library in ANSIC for empirical testing of random number generators. Manuscript, Department d'Informatique et de Recherche Operationnelle, University of Montreal.

Marsaglia, G., 1996. DIEHARD: a battery of tests of randomness. http://stat.fsu.edu/∼geo.

Matsumoto, M., Nishimura, T., 1998. Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. ACM Trans. Model. Comput. Simul. 8 (1), 3–30.

McCullough, B.D., 1998. Assessing the reliability of statistical software: part I. Amer. Statist. 52, 358–366.

McCullough, B.D., 2000a. Experience with the StRD: application and interpretation. Statist. Comput. 31, 16–21 (Proceedings of the Interface Conference).

McCullough, B.D., 2000b. The accuracy of mathematica 4 as a statistical package. Comput. Statist. 15, 279–299.

McCullough, B.D., 2002. Does Microsoft fix errors in excel? Proceedings of the 2001 Joint Statistical Meetings [CD-ROM], American Statistical Association, Alexandria, VA.

McCullough, B.D., 2004a. Fixing statistical errors in spreadsheet software: the cases of Gnumeric and Excel. Statist. Software Newslett. http://www.ssncsda.org/reports.

McCullough, B.D., 2004b. Some details of nonlinear estimation. In: Altman, M., Gill, J., McDonald, M.P. (Eds.), Numerical Methods in Statistical Computing for the Social Sciences. Wiley, New York, pp. 199–218 (Chapter 8).

McCullough, B.D., Wilson, B., 1999. On the accuracy of statistical procedures in Microsoft Excel 97. Comput. Statist. Data Anal. 31, 27–37.

McCullough, B.D., Wilson, B., 2002. On the accuracy of statistical procedures in Microsoft Excel 2000 and XP. Comput. Statist. Data Anal. 40 (4), 27–37.

Wichmann, B.A., Hill, I.D., 1982. Algorithm AS 183: an efficient and portable pseudo-random number generator. Appl. Statist. 31, 188–190.

Wichmann, B.A., Hill, I.D., 1984. Correction: algorithm AS 183 : an efficient and portable pseudo-random number generator. Appl. Statist. 33, 123.

Wolfram, S., 1999. The Mathematica Book 4*e*. Cambridge University Press, New York.

PDF | All previous versions of Microsoft Excel until Excel 2007 have been criticized by statisticians for several reasons, including the accuracy of statistical functions, the properties of the random number generator, the quality of statistical add-ins, the weakness of the Solver...Â (will be inserted by the editor). On the accuracy of statistical procedures in. Microsoft Excel 2010. Guy MÂ´elard. Received: date / Accepted: date. Abstract All previous versions of Microsoft Excel until Excel 2007 have been. criticized by statisticians for several reasons, including the accuracy of statis-. tical functions, the properties of the random number generator, the quality.