Proceedings of the 2004 Winter Simulation Conference
R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

NUMERICAL ACCURACY ISSUES IN USING EXCEL FOR SIMULATION STUDIES

Kellie B. Keeling

Business Information Technology
Pamplin 1007
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0235, U.S.A.

Robert J. Pavur

Business Computer Information Systems
PO Box 305249
University of North Texas
Denton, TX 76203-5249, U.S.A.

ABSTRACT

Many researchers use Excel to perform simulations, but with each upgrade to Excel – Excel 97, Excel 2000, Excel XP, and Excel 2003 – numerical accuracy problems have been noted. In the latest version, Excel 2003, some substantial changes have been made to its algorithms as noted on its Web site. This paper discusses generating random numbers in Excel – including Uniform, Normal, and Poisson variates. In addition, the study assesses how Excel's accuracy stacks up to other statistical software by using the NIST Statistical Reference Datasets tests as certified benchmarks of numerical accuracy. This paper will reveal that Excel 2003 still has room for improvement.

1 INTRODUCTION

The use of Excel in performing statistical analyses has dramatically increased over the years. In fact, it has been thought that more basic statistical calculations are performed using Excel than in all other statistical computer packages combined (McCullough and Wilson 2002). Excel has also emerged as a popular tool for creating simulations in a variety of applications including Markov models, discrete-event simulation, capacity planning, and biological systems (Laverty, Miket, and Kelly 2002; Greasley 1998; Yang, Haddad, and Chow 2001; Brown 1999).

While it seems reasonable to expect that new versions of Excel include corrections to previously identified errors that has always not been the case with past upgrades. With the release of Excel 2003, it is only fair to assess its accuracy to determine if Microsoft has made sufficient strides in upgrading the accuracy of the program.

Knusel (2004) has indicated that several previously identified errors in Excel have finally been eliminated. For example, if the mean of the Poisson is 200, the probability of a Poisson random variable being less than 200 is displayed correctly in Excel 2003 but not in Excel 97, Excel 2000, or Excel XP. Thus, a better algorithm is used to compute these values in Excel 2003.

However, Knusel (2004) notes that there are still problems with the accuracy of the statistical distributions. For example, the probability that a Poisson random variable with mean equal to 200 is less than 100 is displayed correctly in Excel 97. It is also displayed correctly in Excel 2000 and XP. But in Excel 2003, this probability is displayed as zero. Thus, very small values are rounded down. This same type of accuracy problem is noted for the binomial distribution. Hence, some probabilities are more accurate in previous versions of Excel than in the 2003 version.

McCullough and Wilson (2002) investigated numerical accuracy of Excel 2000 and Excel XP (also called Excel 2002). These researchers used the "Statistical Reference Datasets" (called StRD) produced by the (American) National Institute of Standards and Technology (NIST). These datasets include certified statistically accurate computations for a suite of analyses for each of the following procedures: univariate summary statistics, one-way ANOVA, linear regression, and nonlinear least squares. Many statistical software reviewers have taken advantage of these datasets to assess the numerical accuracy of computational software. These data sets can be downloaded by visiting the Web site `<http://www.itl.nist.gov/div898/strd/>` and are classified by difficulty as lower, average, and higher.

There have been five major revisions of Excel (Excel 5.0, Excel 95, Excel 97, Excel 2000, and Excel XP) that have not corrected some easily remedied calculation difficulties with the sample variance (McCullough and Wilson 2002). In their paper, Excel XP is compared to a competing statistical software package called Stata (Statacorp 2001). Numerical inaccuracies were noted in Excel for the higher difficulty data sets with respect to standard deviation calculations.

Cryer (2002) illustrated numerous weaknesses in Excel's computing algorithms; for example, inaccuracies in the standard deviation, the first quartile, zero-intercept regression models, and normal probability plots. McCullough (1998, 1999) published a two-part paper in the American Statistician on the reliability of statistical software. He explains rounding, truncation, and algorithm errors often encountered in computational software. The NIST StRD benchmarks and tests for random number generators are discussed. Before these data sets were available, a collection of data sets by Wilkinson (1985) were useful in discovering deficiencies in statistical software programs.

In this paper, the generation of random numbers in Excel is examined. A discussion is presented about using the built-in functions for creating random numbers in an Excel worksheet and in Visual Basic for Applications. These results show several issues in generating Uniform, Poisson, and Normal variates which could affect researchers who need to generate a large number of random variables.

An additional component of this paper is comparing Excel 2003 to several standard statistical software packages (SAS, SPSS, and Minitab) and Statcrunch. These packages are much more comprehensive as statistical packages than Excel 2003. Statcrunch (2002), a free, web-based package, is a recent competitor in the statistical software arena. This software performs many of the same basic statistical functions as Excel. The numerical testing in this paper also includes Excel 2000 and Excel XP for comparison purposes although McCullough and Wilson (2002) have previously reported on the errors produced by these software packages. Our results illustrate that Excel 2003 is finally an improvement over previous versions although additional improvement is still needed.

2 GENERATING RANDOM DATA

2.1 Generation of Random Numbers Using RAND() and RND in Excel 2003

The command to generate uniform random numbers greater than and equal to 0 and less than 1 is RAND(). This function has no argument. Pressing F9 in the formula bar will change the formula to the value of the random number. This is useful if you do not want the numbers to update when the worksheet is recalculated.

The algorithm for the random number generator in Excel 2003 is at `<http://support.microsoft.com/default.aspx?scid=kb;en-us;828795&Product=xl2003>` and the algorithm used in the previous versions of Excel can be found at `<http://support.microsoft.com/default.aspx?scid=kb;en-us;86523>`. The initial release of Excel 2003 contained a bug that allowed values generated by the RAND() command to be negative. The first service pack for Excel 2003 fixed this problem. All available service packs should be installed before generating random numbers in Excel 2003. Updates are at `<http://windowsupdate.microsoft.com>`.

Visual Basic for Applications code can be written to create replications of a simulation. The RND [number] function is used to generate random numbers, where number is an optional argument. This number determines how RND generates a random number. If the number is less than 0, the same number is generated each time using the number as the seed. If the number is greater than 0 or omitted, then the next random number in the sequence is generated. If the number is 0, then the most recently generated number is returned. For a given initial seed, the same number sequence is generated because the next call to RND will use the previous number as the seed for the next number in the sequence.

The RANDOMIZE [number] function uses a number to initialize the RND's random number generator with a new seed value. If the number is omitted, the system timer is used as the new seed. If RANDOMIZE is not used, the RND function (with no argument) will use the same seed as the first time it was called and; thereafter, uses the previously generated number as the seed. The RANDOMIZE statement, without an argument, should be called before a RND in order to initialize the random number generator with a seed based on the system timer. Note, if a sequence of random numbers is to be repeated, the RND function with a negative number argument should be called immediately before using RANDOMIZE with a numeric argument.

Some custom functions using the RND statement may not recalculate when the worksheet is recalculated. Custom functions created in Excel only recalculate cell ranges that are passed as arguments to the function. If the result of the function depends on cells not explicitly called by the function, the function can be made VOLATILE so that it will recalculate correctly. This command should be added to the custom function. It is noted that this may slow down the time it takes the worksheet to recalculate when any changes are made.

Caution should be given when using the RND function because it does return the value of 0. For example, when generating exponential values with the formula "– mean * LN(RND)", an error will result when RND returns a 0. In this case, a function can be used to generate the random number as follows:

```
Function myRand()
    Application.Volatile
    Randomize
    myRand = RND()
    Do While myRand = 0
        myRand = RND()
    Loop
End Function
```

2.2 Random Numbers from the Uniform Distribution in Excel 2003

Knusel (2002) reports that random numbers generated by the uniform distribution option under Tools, Data Analysis, Random Number Generation are generated from integers from the set {0,1,…,32767} in Excel XP. This generation of random variables from a uniform distribution appears to be the same as in Excel 2003. When trying to generate any number of random variables from the normal, uniform, or any of the distributions listed in the drop-down menu, an error message saying “Integer is not valid” is displayed if the number of random variables requested is 32768 or greater.

To understand how random the uniform distribution observations are, 10,000 uniform random numbers over the interval 0 to 1 were generated in columns A-J using the seed 31, then multiplied by 32767, and then sorted. A count was taken of the repeats. For this seed, the number of repeats is 12083 or 36.88%. This result is similar to the birthday problem in which there are 32767 days in a year and there are 10,000 people in a room on a planet with this newly defined year. Scientific studies which may be affected by this many repeats need to exercise caution in using this generator.

2.3 Random Numbers from the Normal Distribution in Excel 2003

Knusel (2002) noticed that Excel XP generated values of standard normal random variables with exceedingly large values in magnitudes, especially on the negative end of the normal distribution. To examine the symmetry of generated normal random variables, 10,000 standard normal random numbers were generated in each of 10 columns using, again, the random seed 31. The proportion of random numbers below -9, -8, -7, -6, -5, -4, -3, -2, and -1 were recorded. The proportion of random numbers greater than 1, 2, 3, 4, 5, and 6 were noted as well.

These results are presented in Table 1. This table illustrates that the proportion of the numbers below the Z values -1, -2, and -3 are roughly in line with the theoretical proportions. However, the number of generated observations less than -4, -5, -6, -7, -8, and -9 are much higher than the theoretical proportions. Clearly, a Z value of -9 or smaller is unacceptable for the number of observations generated in these samples. The proportion of numbers above Z values of 1, 2, and 3 are approximately in line with the theoretical proportions. The proportion of generated observations above 4 and 5 are slightly higher than the theoretical proportions. The number of repeated observations was 13.59%. For a continuous distribution this may be unacceptable for many scientific studies.

2.4 Random Numbers from the Poisson Distribution in Excel 2003

Knusel (2004) notes that Excel 2003’s POISSON function displays zeros for the probability of a Poisson random variable being less than 103 for a mean of 200. The exact probabilities are equal to 2.8916E-14 or smaller. Actually it is not the algorithm that is at fault here but the fact that Excel 2003 displays each digit starting with the 15th decimal as a 0. This is due to Excel using double precision to store numbers which includes only 15 significant digits of precision.

Since the normal random values generated by Excel 2003 resulted in unacceptable numbers on the tails, the probabilities of tail values for Poisson generated values are also considered. To examine the tail probabilities of generated observations from a Poisson distribution, 10,000 Poisson random numbers with a population mean of 175 were generated in each of 10 columns. From Table 2, the proportion of the Poisson values below the small values in the left tail appears to conform to the theoretical proportion. However, for Poisson values in the upper tail, several of the proportions were much smaller than the theoretical proportions. At least for this particular generation of Poisson observations, the right tail may be problematic in scientific studies that assume that that the observations are representative of a population of Poisson distributed random numbers.

3 RESULTS

A benchmark analysis of 4 statistical software packages and 3 versions of Excel are presented in this section. Reference datasets with certified computational results from NIST are used to allow objective evaluation of statistical-software. This analysis is timely since new versions of statistical software have appeared since the McCullough and Wilson (2002) analysis of numerical inaccuracies was re ported. Minitab 14 appeared in November of 2003. SPSS 12.0 is another recent upgrade. This comparative analysis will show how Excel 2003 compares to statistical software packages, as well as previous versions of Excel.

Table 3 displays the certified accurate values of the mean and standard deviation for the listed data sets of

Table 1: Proportion of Generated Random Normal Observations below and above Specified Values

Z Value	Proportion Below	Theoretical Proportion		Z Value	Proportion Above	Theoretical Proportion
-9	0.00005	1.13E-19		1	0.15881	0.1586600000
-8	0.00005	6.22E-16		2	0.02261	0.0227500000
-7	0.00005	1.28E-12		3	0.00127	0.0013499000
-6	0.00005	0.0000000010		4	0.00006	0.0000317000
-5	0.00005	0.0000002870		5	0.00005	0.0000002870
-4	0.00007	0.0000317000		6	0	0.0000000010
-3	0.00141	0.0013499000				
-2	0.02299	0.0227500000				
-1	0.15892	0.1586600000				

Table 2: Proportion of Generated Random Poisson Observations below and above Specified Values

Poisson Value	Proportion Below	Theoretical Proportion		Poisson Value	Proportion Above	Theoretical Proportion
130	0.00027	0.0002233000		225	0.00021	0.0001618540
129	0.00018	0.0001620000		230	0.00007	0.0000404450
127	0.00008	0.0000844300		231	0.00006	0.0000302700
125	0.00003	0.0000425000		233	0.00004	0.0000167520
123	0.00003	0.0000208000		235	0.00003	0.0000091210
122	0.00001	0.0000143000		325	0.00001	< 1E-10

Table 3: Certified Accurate Statistical Computations for StRD Calculations

Data Set	Certified Correct Value	
	Mean	Standard Deviation
PiDigits	4.5348000000000	2.8673390602887
Lottery	518.9587155963300	291.6997274709690
Lew	-177.4350000000000	277.3321680443160
Mavro	2.0018560000000	0.0004291234540
Michelso	299.8524000000000	0.0790105478191
NumAcc1	10000002.0000000000000	1.0000000000000
NumAcc2	1.2000000000000	0.1000000000000
NumAcc3	1000000.2000000000000	0.1000000000000
NumAcc4	10000000.2000000000000	0.1000000000000

NIST's StRD. These data sets are listed from lower to higher difficulty. As seen in Tables 4 and 5, the statistical software packages perform rather well in accurately computing the mean. There is no format option in Minitab or Statcrunch to display more significant digits. So, for these two packages, the recorded results were simply the displayed results. For the other software packages, the output was formatted for 15 significant digits. It may be worth noting that the only two software packages to compute the mean of the NumAcc4 data set correctly are SAS and SPSS although all three versions of Excel come close.

In Tables 6 and 7, several of the software packages show weaknesses in computing the standard deviations. Excel 2003 shows a definite improvement over Excel 2000 and XP. SAS and SPSS do well in computing the standard deviations. Minitab appears to round off too much. StatCrunch does not display a standard deviation for the NumAcc4 data set. At least, there are no negative standard deviations.

On the Microsoft website, there is detailed information in the knowledge base about the improvements made to the statistical functions in the Analysis Toolpak. This can be found at: `<http://support.microsoft.com/default.aspx?scid=kb;en-us;829208&Product=xl2003>`. There are a number of articles in the knowledge base that detail known problems with Excel based on the particular version. Researchers should consult this information to determine if the statistical procedures they are using have been identified as having accuracy issues for their version of Excel.

4 CONCLUSIONS

Simulation researchers who use Excel may make the assumption that software companies extensively test their software for every contingency so only accurate results are reported. Unfortunately, there are limits to all statistical software. Since the generation of random numbers is an essential component of simulation studies, care should be taken when using Excel's random number generation procedures.

Excel 2003 has corrected many problems that have plagued previous versions. For example, the calculation of probabilities with the standard normal cumulative distribution function is now corrected. Someone may counter that the errors in Excel are rare and the corrections make little differences. However, this is not very comforting when the results of a particular analysis are computed in two different software packages and the results don't agree.

Table 4: Actual Mean Computations for SAS, SPSS, and Minitab

Data Sets	SAS Ver 8.02	SPSS 12.0	Minitab 14.0
PiDigits	4.53480000000000	4.53480000000001	4.53480000000000
Lottery	518.95871559633000	518.95871559633000	519.00000000000000
Lew	-177.43500000000000	-177.43500000000000	-177.40000000000000
Mavro	2.00185600000000	2.00185600000000	2.00190000000000
Michelso	299.85239999999900	299.85240000000000	299.85000000000000
NumAcc1	10000002.00000000000000	10000002.00000000000000	10000002.00000000000000
NumAcc2	1.20000000000000	1.20000000000000	1.20000000000000
NumAcc3	1000000.20000000000000	1000000.20000000000000	1000000.00000000000000
NumAcc4	10000000.20000000000000	10000000.20000000000000	10000000.00000000000000

Table 5: Actual Mean Computations for Excel 2000, XP, 2003, and Statcrunch

Data Sets	Excel 2000	Excel XP	Excel 2003	StatCrunch Ver3
PiDigits	4.53480000000000	4.53480000000000	4.53480000000000	4.53480000000000
Lottery	518.95871559633000	518.95871559633000	518.95871559633000	518.95874000000000
Lew	-177.43500000000000	-177.43500000000000	-177.43500000000000	-177.43500000000000
Mavro	2.00185600000000	2.00185600000000	2.00185600000000	2.00185600000000
Michelso	299.85240000000000	299.85240000000000	299.85240000000000	299.85240000000000
NumAcc1	10000002.00000000000000	10000002.00000000000000	10000002.00000000000000	10000002.00000000000000
NumAcc2	1.19999999999999	1.19999999999999	1.19999999999999	1.20000000000000
NumAcc3	1000000.20000000000000	1000000.20000000000000	1000000.20000000000000	1000000.20000000000000
NumAcc4	10000000.20000010000000	10000000.20000010000000	10000000.20000010000000	10000000.00000000000000

Table 6: Actual Standard Deviations for SAS, SPSS, and Minitab

Data Sets	SAS Ver 8.02	SPSS 12.0	Minitab 14.0
PiDigits	2.86733906028870	2.86733906028871	2.86730000000000
Lottery	291.69972747096900	291.69972747096900	291.70000000000000
Lew	277.33216804431600	277.33216804431600	277.30000000000000
Mavro	0.00042912345400	0.00042912345400	0.00042900000000
Michelso	0.07901054781904	0.07901054781908	0.07900000000000
NumAcc1	1.00000000000000	1.00000000000000	1.00000000000000
NumAcc2	0.09999999999999	0.10000000000000	0.10000000000000
NumAcc3	0.10000000003492	0.10000000003465	0.10000000000000
NumAcc4	0.10000000055879	0.10000000056079	0.10000000000000

Table 7: Actual Standard Deviations for Excel 2000 XP, 2003, and Statcrunch

Data Sets	Excel 2000	Excel XP	Excel 2003	StatCrunch Ver3
PiDigits	2.86733906028871	2.86733906028871	2.86733906028871	2.86733910000000
Lottery	291.69972747096900	291.69972747096900	291.69972747096900	291.69974000000000
Lew	277.33216804431600	277.33216804431600	277.33216804431600	277.33215000000000
Mavro	0.00042912345385	0.00042912345385	0.00042912345400	0.00042912347000
Michelso	0.07901054823365	0.07901054823365	0.07901054781905	0.07901054600000
NumAcc1	1.00000000000000	1.00000000000000	1.00000000000000	1.00000000000000
NumAcc2	0.10000000000027	0.10000000000027	0.10000000000027	0.10000000000000
NumAcc3	0.10723805294764	0.10723805294764	0.10000000003493	0.10723805400000
NumAcc4	0.00000000000000	0.00000000000000	0.10000000055884	N/A

The authors of this paper concur with Sawitzki (1994) who emphasized that “The presence of errors does not mean unusability for all purposes; the absence of error reports does not imply a recommendation.” Also, there is that possibility that some combination of features might produce unintended results which are difficult to predict and which are not reported in any review of numerical accuracies of software packages.

REFERENCES

Brown. 1999. A methodology for simulating biological systems using Microsoft Excel. *Computer Methods and Programs in Biomedicine* 58(2): 181-190.

Cryer, J. 2002. Problems with using Microsoft Excel for statistics. In *Proceeding of the 2001 Joint Statistical Meetings, American Statistical Association*. Alexandria, Virginia.

Greasley, A. 1998. An example of a discrete-event simulation on a spreadsheet. *Simulation* 70(3): 148-166.

Knusel, L. 2002. On the reliability of Microsoft Excel XP for statistical purposes. *Computational Statistics and Data Analysis* 39: 109-110.

Knusel, L. 2004. On the accuracy of statistical distributions in Microsoft Excel 2003. *Computational Statistics and Data Analysis*: To Appear.

Laverty, W. H., M. J. Miket, and I. W. Kelly. 2002. Simulation of hidden Markov models with EXCEL. *Journal* *of the Royal Statistical Society Series D-The Statistician.* Part 1 51: 31-40.

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

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

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

Sawitzki, G. 1994. Report on the reliability of data analysis systems. *Computational Statistics and Data Analysis* 18: 289-301.

Statacorp. 2001. Stata Statistical Software: Release 7.0. Stata Corporation, College Station, Texas.

Statcrunch. 2002. Statistical software for data analysis on the Web, Integrated Analytics LLC. Available online via `<http://www.statcrunch.com>` (accessed August 3, 2004)

Wilkinson, L. 1985. *Statistics quiz: problems which reveal deficiencies in statistical programs*. Evanston, Illinois: Systat Inc.

Yang, Y. H., K. Haddad, and C. W. Chow. 2001. Capacity planning using Monte Carlo simulation: an illustrative application of commonly available PC software. *Managerial Finance* 27(5): 33-54.

AUTHOR BIOGRAPHIES

KELLIE B. KEELING, Ph.D. is an Assistant Professor at Virginia Polytechnic Institute and State University. She received her Ph.D. in Management Science from the University of North Texas. Her research interests include data mining, simulation, and the use of technology in the classroom. She has recently published in *Communications in Statistics*, *Communications of the ACM*, and *Multivariate Behavioral Research*. She has received Virginia Tech's university award for excellence in teaching. She is a coauthor of the textbook *Introduction to Business Statistics: A Computer Integrated, Data Analysis Approach* published by South-Western Publishers. Her email address is `<kkeeling@vt.edu>`.

ROBERT J. PAVUR, Ph.D. is a Professor at the University of North Texas. He holds a master and doctorate degree from Texas Tech University. He has papers published in such journals as the *Annals of Operations Research*, *IEEE Transactions on Reliability*, *European Journal of Operational Research*, *Journal of the Operational Research Society*, and *International Journal of Operations and Quantitative Management*. He is a coauthor of the textbook *Introduction to Business Statistics: A Computer Integrated, Data Analysis Approach* published by South-Western Publishers. He has held various officer positions in the southwest regional Decision Science Institute and received the distinguished service award in 2000. His email address is `<pavur@unt.edu>`.