



Contents lists available at ScienceDirect

# Computational Statistics and Data Analysis

journal homepage: [www.elsevier.com/locate/csda](http://www.elsevier.com/locate/csda)

## The accuracy of statistical distributions in Microsoft® Excel 2007

A. Talha Yalta

TOBB University of Economics and Technology, Sogutozu Caddesi No: 43, Sogutozu, 06560, Ankara, Turkey

### ARTICLE INFO

#### Article history:

Available online 12 March 2008

### ABSTRACT

We provide an assessment of the statistical distributions in Microsoft® Excel versions 97 through 2007 along with two competing spreadsheet programs, namely Gnumeric 1.7.11 and OpenOffice.org Calc 2.3.0. We find that the accuracy of various statistical functions in Excel 2007 range from unacceptably bad to acceptable but significantly inferior in comparison to alternative implementations. In particular, for the binomial, Poisson, inverse standard normal, inverse beta, inverse student's  $t$ , and inverse  $F$  distributions, it is possible to obtain results with zero accurate digits as shown with numerical examples.

© 2008 Elsevier B.V. All rights reserved.

### 1. Introduction

Considered as a major revision, the new Excel 12, released along with the Microsoft Office 2007 suite, promises a large number of new features as well as improved functionality. The Excel 2007 product guide<sup>1</sup> mentions, among other things, an overhauled user interface, better data integration, faster calculation performance, and “dramatic visual effects” without giving a word’s consideration to numerical accuracy. Given the fact that Excel is the ubiquitous application for managing and analyzing data in the world of education, business, and the government, it is highly probable that more statistical calculations are done using Excel than all statistical software combined. Therefore, the main objective of this study is to provide an assessment of the statistical distributions in Excel versions 97 through 2007 and discuss Microsoft’s performance on correcting reported flaws effectively and in a timely fashion. Software testing is important as it reflects the ongoing concern of the academic community regarding the reliability of commonly used programs providing statistical functionality.

In the following section, we report on the earlier studies by various authors regarding errors in Excel’s statistical distributions. In section three, we examine the reliability of various statistical distributions in Excel versions 97, 2000, 2002, 2003, and 2007 along with two competing spreadsheet programs, namely Gnumeric 1.7.11 and OpenOffice.org Calc 2.3.0. Our comparison is merely illustrative and the specific distributions chosen are those highlighted by earlier studies regarding Excel’s statistical functions. Section four offers the conclusions.

### 2. Fixing errors in Excel’s statistical distributions

Keeling and Pavur (2007) and Yalta (2007) show that, in general, software vendors do respond to software testing by correcting documented flaws in their programs, resulting in improved software reliability. One exception to this is Microsoft Corporation, which, to this day, largely remained indifferent to numerous studies pointing out the unacceptable accuracy errors of the statistical procedures in their Excel product. Microsoft’s cavalier attitude toward accuracy is so shocking that it prompted McCullough (2001) to ask the rhetorical question, “Does Microsoft Fix Errors in Excel?” His conclusion was that it could be a profit maximizing strategy on behalf of Microsoft to satisfy a large class of users’ demand for software which is flashy and easy-to-use, if not very accurate and reliable.

E-mail addresses: [yalta@etu.edu.tr](mailto:yalta@etu.edu.tr), [talhayalta@gmail.com](mailto:talhayalta@gmail.com).

<sup>1</sup> Available for download at <http://office.microsoft.com/en-us/excel/HA101680001033.aspx> (retrieved July 4, 2007).

**Table 1**  
Microsoft's performance on correcting errors in Excel's statistical distributions

Distribution	Excel 97	Excel 2000	Excel 2002	Excel 2003	Excel 2007
Binomial	Flaws reported	Not fixed	Not fixed	Poor fix	Not fixed
Hypergeometric	Flaws reported	Not fixed	Not fixed	Poor fix	Not fixed
Poisson	Flaws reported	Not fixed	Not fixed	Poor fix	Not fixed
Normal	Flaws reported	Not fixed	Not fixed	Fixed	
Inv. normal	Flaws reported	Not fixed	Poor fix	Poor fix	Not fixed
Inv. chi-square	Flaws reported	Not fixed	Not fixed	Poor fix	Not fixed
Inv. <i>t</i>	Flaws reported	Not fixed	Not fixed	Poor fix	Not fixed
Inv. <i>F</i>	Flaws reported	Not fixed	Not fixed	Poor fix	Not fixed
Gamma				Flaws reported	Not fixed
Inv. beta				Flaws reported	Not fixed

The earliest study regarding Excel's statistical distributions is by Knüsel (1998), who used SAS for Windows 6.12 and his own ELV utility as a benchmark to assess the reliability of the elementary statistical distributions in Excel 97. Reporting a long list of errors in discrete distributions namely Poisson, binomial, and hypergeometric as well as continuous distributions such as standard normal, inverse standard normal, inverse chi-square, inverse *F*, and inverse *t*, Knüsel concluded, "So one has to warn statisticians against using Excel functions for scientific purposes." In a follow-up study, McCullough (2001) noted that all the errors cataloged previously remained uncorrected in Excel 2000. McCullough also found that the only observed change in Excel 2002 (included in Microsoft Office XP) was to the inverse normal distribution, which, even after being "fixed," could not return more than three digits of accuracy in the tail of the distribution. Shortly thereafter, Knüsel (2002) affirmed these results and warned once again that Excel should not be used for statistical purposes.

Wondering whether there was something inherently difficult about correcting errors in spreadsheet software, McCullough (2004) investigated how an alternative package, Gnumeric (v1.0.4), coped with the flaws surprisingly similar to those found in Excel (97/2000/2002). He found that the few part-time volunteers who maintain and develop Gnumeric were able to fix all the problems in statistical procedures within several weeks after the errors have been reported.

With the release of Office 2003, Microsoft announced extensive improvements<sup>2</sup> in Excel's statistical functions – including functions for discrete distributions, continuous distribution inverse functions, as well as the normal distribution and its related functions – that "stretch such functions to their limits." Attempting to verify Microsoft's claims, Knüsel (2005) found that some of the inaccurate results cataloged previously have been corrected while some have stayed, and some have been replaced by new errors. He also showed that the gamma and beta distributions were not always computed correctly in Excel 2003. Knüsel's results were shortly confirmed by McCullough and Wilson (2005), who advised persons desiring to conduct statistical analysis of data against using Excel 2003.

Table 1 shows Microsoft's performance on correcting published errors in Excel's statistical functions over the last five major revisions. As it will be shown in the next section, many reliability issues previously reported for different versions of Excel remain unresolved in Excel 2007. Barring the standard normal distribution, which now *appears* to be accurate, the accuracy of these functions range from unacceptably bad to acceptable but significantly inferior in comparison to alternative implementations.

### 3. Numerical accuracy of various statistical functions in Excel 2007

Knüsel (1995) asks the question "What accuracy is required of a program computing statistical distributions?" His answer is that if the computed probability is given as 0.0000 and the exact value is smaller than  $0.5 \times 10^{-4}$ , then the answer given by the computer is correct and reliable. On the other hand, if the program returns  $4.820 \times 10^{-15}$  and the correct value is  $4.073 \times 10^{-15}$ , then this answer is unacceptable since it makes the user believe that the result is accurate to four significant digits while in fact only one of these digits is correct. Accordingly, if the answer cannot be calculated accurately, the program should note this rather than misleading the user by returning an incorrect result.

We certainly agree with Knüsel that the answer given by a program should always be correct as it is printed out. In addition, we second McCullough and Vinod (1999) that software should be judged inadequate if a poor or a defective algorithm is used when there is a known alternative that can provide an accurate answer to a particular problem. We should like to note that today there exists well-known, freely available, and easy to find algorithms that can provide extremely accurate answers for the computation of most statistical distributions. In fact, according to today's computing standards, a good program offering statistical functionality is expected to provide six or seven digits of accuracy on commodity hardware for probabilities as small as  $10^{-300}$  and even smaller. But how do the statistical procedures in Excel 2007 fare against today's computing standards?

Below, we provide an assessment of the accuracy of various statistical distributions in Excel versions 97 through 2007. The specific functions chosen are those previously highlighted in different studies including McCullough (2001, 2004), McCullough and Wilson (1999, 2002, 2005) and Knüsel (1998, 2002, 2005). The parameter combinations presented in the

<sup>2</sup> See Microsoft Knowledgebase Article #828888.

**Table 2**Binomial distribution with parameters ( $k, n = 1030, p = 0.5, \sigma = 1$ )

$k$	EXACT	ELV Ed.2	EXCEL 97/2K/XP	EXCEL 2003/2007	CALC 2.3.0	GNUMERIC 1.7.11
1	8.96114E–308	0	8.95245E–308	0	Exact	Exact
2	4.61499E–305	0	Exact	0	Exact	Exact
100	1.39413E–169	0	Exact	0	Exact	Exact
200	5.45781E–92	Exact	Exact	0	Exact	Exact
300	2.91621E–42	Exact	Exact	0	Exact	Exact
390	3.18196E–15	Exact	Exact	0	Exact	Exact
391	5.24099E–15	Exact	Exact	2.05902E–15	Exact	Exact
400	3.89735E–13	Exact	Exact	3.86553E–13	Exact	Exact
410	3.19438E–11	Exact	Exact	3.19406E–11	Exact	Exact
420	1.76037E–09	Exact	Exact	Exact	Exact	Exact
500	1.83106E–01	Exact	#NUM!	Exact	Exact	Exact
550	9.86550E–01	Exact	#NUM!	Exact	Exact	Exact
575	9.99920E–01	Exact	#NUM!	Exact	Exact	Exact
589	9.99998E–01	Exact	#NUM!	Exact	Exact	Exact

tables are for illustrative purposes and largely follow those used by the aforementioned studies. The “exact” numbers were calculated using Mathematica 5.2, which can use symbolic methods to perform computations with arbitrary precision.<sup>3</sup> These values were also checked with the program ELV (Knüsel, 2003), which can provide exact results with six digits of accuracy for upper and lower tail probabilities as well as upper and lower quantiles of nine elementary statistical distributions for probabilities as small as  $10^{-100}$ .

Two additional programs included in this study are Gnumeric 1.7.11 and OpenOffice.org Calc 2.3.0, which, although being open source software, employ dissimilar subroutines for the calculation of various statistical distributions. More specifically, released under the GNU General Public License (GPL), Gnumeric is allowed to use the math/stats library of the GNU-R statistical environment, which is also a GPL program. On the other hand, because OpenOffice.org is released under the GNU Lesser General Public License (LGPL), it cannot share the libraries that are ordinary GPL licensed and therefore has to rely on its custom developed procedures for statistical distributions.<sup>4</sup> Both of these alternative spreadsheet applications provide a range of features roughly equivalent to those offered by Excel.

All computations were done using an Intel Centrino Duo 2.16 GHz notebook computer with 2 GB memory and running under the GNU/Linux platform (kernel 2.6.18.8). The Windows XP operating system used for running different versions of Excel as well as the ELV utility was virtualized using VirtualBox 1.4.0.

### 3.1. Binomial distribution

If a random variable  $X$  has a binomial  $\text{Bi}(n, p)$  distribution with the parameters  $n$  for the number of trials and  $p$  for the probability for a success, then the Excel function  $\text{BINOMDIST}$ , with the logical argument  $\text{cumulative} = \text{TRUE}$  ( $\Sigma = 1$ ), computes for a given  $k$ , the probability  $\text{Pr}\{X \leq k\}$ . As Table 2 shows, Excel 97, 2000, and XP can correctly compute probabilities as small as  $4.6 \times 10^{305}$  while giving no result for central probabilities.<sup>5</sup> We also see that this bad algorithm was later replaced by an even worse algorithm, so that Excel 2003 and 2007 can mislead the users by providing results with zero accurate digits. Gnumeric and Calc, on the other hand, can provide exact results even for extremely small probability values.

### 3.2. Hypergeometric distribution

If a random variable  $X$  has a hypergeometric  $H(N, M, n)$  distribution with the parameters  $n$  for the number of draws without replacement from a finite population of size  $N$  with  $M$  distinctive members, then the Excel function  $\text{HYPGEOMDIST}$  computes, for a given  $k$  number of successes, the point probability  $\text{Pr}\{X = k\}$ . For the sake of comparison with Knüsel (1998), here we consider the case where exactly  $k$  white balls are selected after making 500 draws from an urn of 1030 balls and the number of white balls in the urn is given as 515. As Table 3 shows, Excel 97, 2000, and XP cannot compute these probabilities for  $N \geq 1030$  while more recent versions of Excel can return an accurate answer only for the range  $187 < k < 313$ . Gnumeric and Calc, on the other hand, can provide exact results for all possible values of  $k \in \{0, 1, \dots, 500\}$ . In addition, Gnumeric also offers the Boolean option  $\text{cumulative}$  for computing the tail probabilities  $\text{Pr}\{X \leq k\}$ , lack of which in Excel was criticized twice by Knüsel (1998, 2005) and is still missing in Excel 2007.

<sup>3</sup> See McCullough (2000) for a detailed assessment of this program in terms of numerical accuracy.

<sup>4</sup> See de Laat (2005) for a discussion of the significance of different property regimes for software development.

<sup>5</sup> For more information, see Microsoft Knowledgebase Article #828888, which also provides links to pseudo-code for the algorithms used to avoid the overflow problems affecting various discrete probability distributions including the binomial, hypergeometric, and Poisson distributions in Excel 2002 and earlier.

**Table 3**Hypergeometric distribution with parameters ( $k, N = 1030, M = 515, n = 500$ )

$k$	EXACT	ELV Ed.2	EXCEL 97/2K/XP	EXCEL 2003/2007	CALC 2.3.0	GNUMERIC 1.7.11
0	1.60137E-280	0	#NUM!	0	Exact	Exact
100	7.46483E-83	Exact	#NUM!	0	Exact	Exact
187	1.53541E-15	Exact	#NUM!	0	Exact	Exact
188	4.13038E-15	Exact	#NUM!	Exact	Exact	Exact
200	1.65570E-10	Exact	#NUM!	Exact	Exact	Exact
300	1.65570E-10	Exact	#NUM!	Exact	Exact	Exact
312	4.13038E-15	Exact	#NUM!	Exact	Exact	Exact
313	1.53541E-15	Exact	#NUM!	0	Exact	Exact
400	7.46483E-83	Exact	#NUM!	0	Exact	Exact
500	1.60137E-280	0	#NUM!	0	Exact	Exact

**Table 4**Poisson distribution with parameters ( $k, \lambda, \Sigma$ )

$k$	$\lambda$	$\Sigma$	EXACT	ELV Ed.2	EXCEL 97/2K/XP	EXCEL 2003/2007	CALC 2.3.0	GNUMERIC 1.7.11
0	200	0	1.38390E-87	Exact	Exact	0	Exact	Exact
103	200	0	1.41720E-14	Exact	Exact	0	Exact	Exact
104	200	0	2.72538E-14	Exact	Exact	Exact	Exact	Exact
133	200	0	1.01322E-07	Exact	Exact	Exact	Exact	Exact
134	200	0	1.51227E-07	Exact	#NUM!	Exact	Exact	Exact
200	200	0	2.81977E-02	Exact	#NUM!	Exact	Exact	Exact
314	200	0	2.23568E-14	Exact	#NUM!	Exact	Exact	Exact
315	200	0	1.41948E-14	Exact	#NUM!	0	Exact	Exact
400	200	0	5.58069E-36	Exact	#NUM!	0	Exact	Exact
900	200	0	1.73230E-286	0	#NUM!	0	Exact	Exact
1E+03	1E+03	1	0.508409	Exact	#NUM!	Exact	#VALUE!	Exact
1E+05	1E+05	1	0.500841	Exact	#NUM!	0.679499	#VALUE!	Exact
1E+07	1E+07	1	0.500084	Exact	#NUM!	0.952000	#VALUE!	Exact
1E+09	1E+09	1	0.500008	Exact	#NUM!	0.994979	#VALUE!	Exact

### 3.3. Poisson distribution

If a random variable  $X$  has a Poisson  $Po(\lambda)$  distribution with the mean parameter  $\lambda$ , then the Excel function POISSON, with the logical argument *cumulative* = FALSE ( $\Sigma = 0$ ), computes for a given  $k$ , the probability  $\Pr\{X = k\}$ . As can be seen from Table 4, Excel 97, 2000, and XP can correctly compute the extreme lower tail probabilities while giving no result for central and upper tail probabilities. Microsoft “fixed” this problem by replacing a poor algorithm with another poor algorithm, so that Excel 2003 and 2007 can compute the correct results in the central part of this distribution while rounding extreme tail probabilities down to zero. This performance is significantly inferior in comparison to those of Gnumeric and Calc, which can provide exact results even for probabilities as small as  $1.73 \times 10^{-286}$ . In addition, Poisson’s cumulative distribution function is known to converge toward 0.5 for large values of  $k = \lambda$ , however, results computed with Excel 2003 and 2007 seem to converge toward 1, resulting in answers with zero accurate digits. In this case, Calc also fails to calculate the correct answer, however, by returning an error message it avoids misleading the user. Gnumeric can provide the exact result even for  $\lambda = 1 \times 10^9$ .

### 3.4. Gamma distribution

If a random variable  $X$  has a gamma  $\Gamma(\alpha, \beta)$  distribution with the shape parameter  $\alpha$  and the scale parameter  $\beta$ , then the Excel function GAMMADIST, with the logical argument *cumulative* = TRUE ( $\Sigma = 1$ ), computes, for a given  $x$ , the probability  $\Pr\{X \leq x\}$ . Excel versions 97 through 2007 print out these probabilities with 6 or more significant digits by default while only 5 of these digits may be correct as can be seen from Table 5. Also, for  $x \leq 0.11$ , Excel may return the error #NUM!, which is a fault known since Excel 2000,<sup>6</sup> and still not fixed in Excel 2007. Our tests show that Gnumeric and Calc seem to provide exact results with at least six accurate digits in this distribution.

### 3.5. Inverse standard normal distribution

If a random variable  $X$  has a standard normal  $N(\mu, \sigma^2)$  distribution with the mean  $\mu = 0$  and variance  $\sigma^2 = 1$ , then the Excel function NORMSINV computes, for a given probability  $p$ , the value of  $k$  such that  $\Pr\{X \leq k\} = p$ . As Table 6 shows,

<sup>6</sup> See Microsoft Knowledgebase Article #215214.

**Table 5**Gamma distribution with parameters ( $x, \alpha, \beta = 1, \Sigma = 1$ )

$x$	$\alpha$	EXACT	ELV Ed.2	EXCEL 97/2K/XP/2003/2007	CALC 2.3.0	GNUMERIC 1.7.11
0.1	0.1	0.827552	Exact	#NUM!	Exact	Exact
0.2	0.1	0.879420	Exact	0.879419	Exact	Exact
0.2	0.2	0.764435	Exact	0.764434	Exact	Exact
0.3	0.2	0.816527	Exact	Exact	Exact	Exact
0.3	0.3	0.726957	Exact	Exact	Exact	Exact
0.4	0.3	0.776381	Exact	0.776380	Exact	Exact
0.4	0.4	0.701441	Exact	Exact	Exact	Exact
0.5	0.4	0.748019	Exact	0.748018	Exact	Exact
0.5	0.5	0.682689	Exact	Exact	Exact	Exact
0.6	0.5	0.726678	Exact	Exact	Exact	Exact

**Table 6**Inverse standard normal distribution with parameter ( $p$ )

$p$	EXACT	ELV Ed.2	EXCEL 97/2K	EXCEL XP	EXCEL 2003/2007	CALC 2.3.0	GNUMERIC 1.7.11
5E-1	0	Exact	Exact	5.47142E-10	-1.39214E-16	Exact	Exact
1E-1	-1.28155	Exact	Exact	Exact	Exact	Exact	Exact
1E-2	-2.32635	Exact	-2.32634	Exact	Exact	Exact	Exact
1E-3	-3.09023	Exact	-3.09024	-3.09025	Exact	Exact	Exact
1E-4	-3.71902	Exact	-3.71947	-3.71909	Exact	Exact	Exact
1E-5	-4.26489	Exact	-4.26546	-4.26504	Exact	Exact	Exact
1E-6	-4.75342	Exact	-4.76837	-4.75367	Exact	Exact	Exact
1E-7	-5.19934	Exact	-5000000	-5.19969	Exact	Exact	Exact
1E-15	-7.94135	Exact	-5000000	-7.93597	Exact	Exact	Exact
1E-16	-8.22208	Exact	-5000000	-8.29366	Exact	Exact	Exact
1E-100	-21.2735	Exact	-5000000	-8.29366	Exact	Exact	Exact
1E-197	-29.9763	No solution	-5000000	-8.29366	Exact	Exact	Exact
1E-198	-30.0529	No solution	-5000000	-8.29366	-30	Exact	Exact
1E-300	-37.0471	No solution	-5000000	-8.29366	-30	Exact	Exact

this function in Excel 97 and 2000 is unacceptably bad and can return results with no accurate digits. Microsoft “fixed” this problem *twice*<sup>7</sup> by first tightening the convergence criteria for the iterative inversion algorithm in Excel XP and then by improving the accuracy of the normal distribution in Excel 2003. Except for the boundary case  $p = 0.5$ , this Excel function can now be judged acceptable by the standards of Knüsel for probabilities larger than  $1 \times 10^{-200}$ . On the other hand, as Knüsel (2005) mentions, it is easier to compute small probabilities due to the faster convergence in the tails of a distribution and there is no reason why a good algorithm, such as the one used in Calc or Gnumeric, cannot return exact results with six accurate digits for  $p > 1 \times 10^{-300}$  for this procedure. Finally, the reason ELV cannot return a solution for the last three examples is a design decision by Knüsel (2003) to not support probabilities less than  $1 \times 10^{-100}$ .

### 3.6. Inverse chi-square distribution

If a random variable  $X$  has a chi-square  $\chi^2(n)$  distribution with  $n$  degrees of freedom, then the Excel function CHIINV computes, for a given probability  $p$ , the value of  $k$  such that  $\Pr\{X > k\} = p$ . Excel 97, 2000, and XP report these quantiles with 9 or 10 significant digits, however, as can be seen from Table 7, only one or two of these digits may be correct if  $p$  is small. The algorithm used in Excel 2003 and 2007 seems to be acceptable by the standards of Knüsel (1995), however, it also shows unstability when  $p$  is small or  $n$  is large. This function in Calc is unacceptably bad and can return misleading output with zero accurate digits. Our tests show that it is very difficult to fault Gnumeric in this distribution as well.

### 3.7. Inverse beta distribution

If a random variable  $X$  has a beta  $Be(\alpha, \beta)$  distribution with the non-negative shape parameters  $\alpha$  and  $\beta$ , then the Excel function BETAINV computes, for a given probability  $p$ , the value of  $k$  such that  $\Pr\{X \leq k\} = p$ . As Table 8 shows, this function in Excel versions 97 through 2007 is very unreliable and can return incorrect results for not-so-small probability values. The corresponding BETAINV function in Calc also produces inaccurate output for small  $p$  values and needs to be fixed immediately. Gnumeric, on the other hand, can return exact results even when  $p$  is extremely small.

<sup>7</sup> See Microsoft Knowledgebase Article #826772.

**Table 7**Inverse chi-square distribution with parameters ( $p, n$ )

$p$	$n$	EXACT	ELV Ed.2	EXCEL 97/2K/XP	EXCEL 2003/2007	CALC 2.3.0	GNUMERIC 1.7.11
2E-1	1	1.64237	Exact	1.64238	1.64238	Exact	Exact
2E-1	5	7.28928	Exact	7.28927	Exact	Exact	Exact
1E-1	1	2.70554	Exact	2.70554	Exact	Exact	Exact
1E-1	5	9.23636	Exact	9.23635	Exact	Exact	Exact
1E-5	1	19.5114	Exact	19.5037	Exact	Exact	Exact
1E-5	5	30.8562	Exact	30.7987	Exact	Exact	Exact
1E-6	1	23.9281	Exact	24.3664	Exact	23.9293	Exact
1E-6	5	35.8882	Exact	35.6115	Exact	35.8896	Exact
1E-7	1	28.3740	Exact	#NUM!	#NUM!	28.5474	Exact
1E-7	5	40.8630	Exact	#NUM!	#NUM!	40.9382	Exact
1E-12	1	50.8441	Exact	#NUM!	#NUM!	48	Exact
1E-12	5	65.2386	Exact	#NUM!	#NUM!	320	Exact
0.48	778	779.312	Exact	#NUM!	#NUM!	Exact	Exact
0.50	780	779.333	Exact	Exact	Exact	Exact	Exact
0.52	782	779.353	Exact	#NUM!	#NUM!	Exact	Exact

**Table 8**Inverse beta distribution with parameters ( $p, \alpha = 5, \beta = 2$ )

$p$	EXACT	ELV Ed.2	EXCEL 97/2K/XP/2003/2007	CALC 2.3.0	GNUMERIC 1.7.11
1E-1	4.89684E-01	Exact	Exact	Exact	Exact
1E-2	2.94314E-01	Exact	2.94315E-01	Exact	Exact
1E-3	1.81386E-01	Exact	1.81396E-01	Exact	Exact
1E-4	1.12969E-01	Exact	1.13037E-01	Exact	Exact
1E-5	7.07371E-02	Exact	7.03125E-02	7.07370E-02	Exact
1E-6	4.44270E-02	Exact	4.29688E-02	4.44268E-02	Exact
1E-7	2.79523E-02	Exact	3.12500E-02	2.78232E-02	Exact
1E-8	1.76057E-02	Exact	3.12500E-02	1.65959E-02	Exact
1E-9	1.10963E-02	Exact	3.12500E-02	2.83309E-03	Exact
1E-10	6.99645E-03	Exact	3.12500E-02	2.83309E-04	Exact
1E-11	4.41255E-03	Exact	3.12500E-02	2.83309E-05	Exact
1E-12	2.78337E-03	Exact	3.12500E-02	2.83309E-06	Exact
1E-13	1.75589E-03	No solution	3.12500E-02	2.83309E-07	Exact
1E-100	6.98827E-21	No solution	3.12500E-02	0.00000E+00	Exact

**Table 9**Inverse t-distribution with parameters ( $p, n = 1$ )

$p$	EXACT	ELV Ed.2	EXCEL 97/2K/XP	EXCEL 2003/2007	CALC 2.3.0	GNUMERIC 1.7.11
2E-1	1.37638E+00	Exact	Exact	Exact	Exact	Exact
1E-1	3.07768E+00	Exact	Exact	Exact	Exact	Exact
1E-2	3.18205E+01	Exact	3.18210E+01	Exact	Exact	Exact
1E-3	3.18309E+02	Exact	3.18289E+02	Exact	Exact	Exact
1E-4	3.18310E+03	Exact	3.18527E+03	Exact	Exact	Exact
1E-5	3.18310E+04	Exact	3.17383E+04	Exact	Err:523	Exact
1E-6	3.18310E+05	Exact	3.12500E+05	Exact	Err:523	Exact
1E-7	3.18310E+06	Exact	5.00000E+06	Exact	Err:523	Exact
1E-8	3.18310E+07	Exact	5.00000E+06	1.00000E+07	Err:523	Exact
1E-11	3.18310E+10	Exact	5.00000E+06	1.00000E+07	Err:523	Exact
1E-12	3.18310E+11	Exact	5.00000E+06	1.00000E+07	Err:523	3.18327E+11
1E-13	3.18310E+12	No solution	5.00000E+06	1.00000E+07	Err:523	3.18338E+12
1E-100	3.18310E+99	No solution	5.00000E+06	1.00000E+07	Err:523	#NUM!

### 3.8. Inverse $t$ distribution

If a random variable  $X$  has a Student's  $t$  ( $t(n)$ ) distribution with  $n$  degrees of freedom, then the Excel function TINV computes, for a given probability  $p$ , the value of  $k$  such that  $\Pr\{|X| > k\} = p$ . As Table 9 shows,<sup>8</sup> Excel 97, 2000, and XP can produce output with only two or three digits if  $p$  is small. The "improved" search method used in Excel 2003 and 2007 performs

<sup>8</sup> While Excel's TINV function requires a 2-tailed probability, following the earlier studies on Excel's statistical distributions, here we report the 1-tailed probabilities computed by using  $2 \times p$ .

**Table 10**  
Inverse  $F$  distribution with parameters ( $p, n_1 = 1, n_2 = 1$ )

$p$	EXACT	ELV Ed.2	EXCEL 97/2K/XP	EXCEL 2003/2007	CALC 2.3.0	GNUMERIC 1.7.11
5E-1	1	Exact	Exact	Exact	Exact	Exact
4E-1	1.89443E+00	Exact	Exact	Exact	Exact	Exact
2E-1	9.47214E+00	Exact	9.47216E+00	Exact	Exact	Exact
1E-1	3.98635E+01	Exact	3.98636E+01	Exact	Exact	Exact
1E-2	4.05218E+03	Exact	Exact	Exact	Exact	Exact
1E-3	4.05284E+05	Exact	4.05312E+05	Exact	Exact	Exact
1E-4	4.05285E+07	Exact	4.05273E+07	Exact	Exact	Exact
1E-5	4.05285E+09	Exact	#N/A	1.00000E+09	Err:523	Exact
1E-6	4.05285E+11	Exact	#N/A	1.00000E+09	Err:523	Exact
1E-12	4.05285E+23	Exact	#N/A	1.00000E+09	Err:523	Exact
1E-13	4.05285E+25	No solution	#N/A	1.00000E+09	Err:523	Exact
1E-100	4.05285E+199	No solution	#N/A	1.00000E+09	Err:523	Exact

better,<sup>9</sup> however, instead of returning the answer “1000000000” it would be preferable if Excel returned “#NUM!” so as not to mislead the users for small  $p$ . The corresponding TINV function in Calc, while acceptable by Knüsel’s standards, also has room for improvement. Here, we see that Gnumeric 1.7.11 also can provide an inaccurate answer for  $p < 1 \times 10^{-11}$ , however, this is already fixed in Gnumeric 1.7.91 which was released a few weeks after we notified the developers about this problem.

### 3.9. Inverse $F$ distribution

If a random variable  $X$  has an  $F(n_1, n_2)$  distribution with  $n_1$  and  $n_2$  degrees of freedom, then the Excel function FINV computes, for a given probability  $p$ , the value of  $k$  such that  $\Pr\{X > k\} = p$ . As Table 10 shows, this function in Excel 97, 2000 and XP is unacceptably bad. The “improved” function in Excel 2003 and 2007 performs better but can become unstable and return “1000000000” for not-so-small probability values. Calc also fails to compute the correct answer for  $p < 1 \times 10^{-4}$ , however, inconvenient it may be, this is acceptable from an accuracy perspective because the user is not misled. Gnumeric, on the other hand, can provide exact results even when  $p = 1 \times 10^{-100}$ .

## 4. Conclusion

It is our understanding that the algorithms for the computation of various statistical distributions in Excel 2007 can be inaccurate and/or unstable, and therefore can be unsafe to use. In particular, for the binomial, Poisson, inverse standard normal, inverse beta, inverse student’s  $t$ , and inverse  $F$  distributions, it is possible to obtain results with zero accurate digits. Our results also show that the alternative Gnumeric and OpenOffice.org Calc programs, which employ dissimilar subroutines for the computation of statistical distributions, provide better accuracy in general in comparison to Excel 2007. In particular, Gnumeric can uniformly return exact values with at least six digits of accuracy for probabilities as small as  $10^{-300}$  in all of our tests except one, and this is already fixed within a few weeks after we contacted the developers about the problem. Calc has important numerical difficulties for the computation of the quantiles of various distributions including the inverse chi-square, inverse beta, inverse  $t$ , and inverse  $F$  distributions. Once notified about the problems, Calc developers expressed their intention to correct these flaws with the upcoming OpenOffice.org version 2.4.

A new trend in computing is Web-based applications, which facilitate the collaborative creation and modification of documents over the Internet in real time. The recently introduced Google Spreadsheet, a part of the Google Docs online service, is one such application competing with the microcomputer software evaluated in this study. Our cursory examination of Google Spreadsheet finds gross errors in the accuracy of the standard normal, binomial, hypergeometric, and Poisson distributions. Consequently, a thorough evaluation of this application is necessary to help researchers and practitioners make the decision whether to move from the PC to the grid.

It is a known fact that Excel is commonly used in a wide range of decision making processes from options trading to research in physical laboratories. Offering statistical functionality in a computer program is a serious matter and it brings important responsibilities to the software vendor. Microsoft has repeatedly shown its lack of interest to this concernment by releasing new versions of Excel without first correcting the problems documented by different authors on various different occasions. Because of Microsoft’s lack of commitment to accuracy, it is now possible to find on the Internet various users’ custom scripts and macros for proper computation of statistical distributions in Excel.<sup>10</sup> It is unclear when, if at all, Microsoft will properly fix Excel’s inaccurate procedures for all of which there are free, well-known, and reliable alternatives. Meanwhile, researchers should continue to avoid using the statistical functions in Excel 2007 for any scientific purpose.

<sup>9</sup> It is worth noting that, for all of our tests, Excel’s distribution functions corresponding to the quartile functions are able to calculate the correct probability, given the inverse. Consequently, the problems noted for various inverse distribution functions are strictly shortcomings of the “improved” numerical inversion routines used in Excel 2003 and 2007 rather than side effects of poor distribution routines.

<sup>10</sup> One such example is “Smith’s Workbook” (available at <http://members.aol.com/iandjmsmith/iansNApage.htm>), which contains a set of VBA procedures for calculating probability related quantities in Excel with significantly improved accuracy.

## Acknowledgements

I would like to express my appreciation to the anonymous referees, whose comments led to significant improvements to this manuscript.

## References

- de Laat, P., 2005. Copyright or copyleft? An analysis of property regimes for software development. *Research Policy* 34, 1511–1532.
- Keeling, K.B., Pavur, R.J., 2007. A comparative study of the reliability of nine statistical software packages. *Computational Statistics and Data Analysis* 51, 3811–3831.
- Knüsel, L., 1995. On the accuracy of the statistical distributions in GAUSS. *Computational Statistics and Data Analysis* 20, 699–702.
- Knüsel, L., 1998. On the accuracy of statistical distributions in Microsoft Excel 97. *Computational Statistics and Data Analysis* 26, 375–377.
- Knüsel, L., 2002. On the reliability of Microsoft Excel XP for statistical purposes. *Computational Statistics and Data Analysis* 39, 109–110.
- Knüsel, L., 2003. Computation of Statistical Distributions - Documentation of Program ELV, <http://www.stat.uni-muenchen.de/~knuesel>, [Online; retrieved July 4, 2007].
- Knüsel, L., 2005. On the accuracy of statistical distributions in Microsoft Excel 2003. *Computational Statistics and Data Analysis* 48, 445–449.
- McCullough, B.D., 2000. The accuracy of Mathematica 4 as a statistical package. *Computational Statistics* 15, 279–299.
- McCullough, B.D., 2001. Does microsoft fix errors in Excel? In: Proceedings of the 2001 Joint Statistical Meetings. American Statistical Association, Alexandria, VA, [CD-ROM].
- McCullough, B.D., 2004. Fixing Statistical Errors in Spreadsheet Software: The Cases of Gnumeric and Excel. [http://www.csdassn.org/software\\_reports/gnumeric.pdf](http://www.csdassn.org/software_reports/gnumeric.pdf), [Online; retrieved July 4, 2007].
- McCullough, B.D., Vinod, H.D., 1999. The numerical reliability of econometric software. *Journal of Economic Literature* 37, 633–655.
- McCullough, B.D., Wilson, B., 1999. On the accuracy of statistical procedures in Microsoft EXCEL 97. *Computational Statistics and Data Analysis* 31, 27–37.
- 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, 713–721.
- McCullough, B.D., Wilson, B., 2005. On the accuracy of statistical procedures in Microsoft Excel 2003. *Computational Statistics and Data Analysis* 49, 1244–1252.
- Yalta, A.T., 2007. The numerical reliability of GAUSS 8.0. *The American Statistician* 61, 262–268.