ARTICLE IN PRESS
Available online at www.sciencedirect.com
SCIENCE DIRECT
Computational Statistics & Data Analysis ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Short communication

On the accuracy of statistical distributions in Microsoft Excel 2003

Leo Knüsel*

Department of Statistics, University of Munich, Munich D-80539, Germany

Received 25 January 2004; accepted 19 February 2004

Abstract

Some of previously indicated errors in Microsoft Excel 97 and Excel XP have been eliminated in Excel 2003. But some others have not been corrected in Excel 2003 and new ones have been found as is shown by numerical examples.

Keywords: Accuracy of statistical distributions; Microsoft Excel 2003

1. Introduction and summary

What accuracy is required of a program computing statistical distributions? Is it enough to give the wanted probabilities with just four decimal places and to output 0.0000 if the correct probability is smaller than 0.00005, or is it necessary to demand an accuracy standard of four to six correct significant digits even for probabilities as small as 10^{-100}, e.g. 0.3654×10^{-87}, where the four digits in the mantissa have to be correct? We do not plead for the latter high standard although it is often easier to compute small tail probabilities of a distribution than to compute correct values in the central probability range of 0.001 to 0.999 due to the faster convergence of many algorithms in the tails of a distribution. But we think that the user of a program can expect as a minimum requirement that the answer given by the program is correct and reliable as it is printed out. If the computed probability is given as 0.0000 and the correct value is smaller than 0.5×10^{-4}, then the answer given by the computer is correct and reliable in our sense. But if the answer given by the program is 4.820×10^{-15} and the correct

* Tel.: +49-8921803195; fax: +49-892180993196.
E-mail address: knuesel@stat.uni-muenchen.de (L. Knüsel).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2004.02.008

value is 4.073×10^{-15} then the answer given by the computer makes the user believe that the result is correct to four significant digits where in fact it is wrong by 18%, as $|app - exa|/exa = 0.18$ where exa = exact value and app = approximate value.

In Knüsel (1998) it was shown that in Excel 97 the computation of some discrete distributions (Binomial, Poisson, Hypergeometric) and of some continuous distributions (Normal, Chi-square, F, t) is unsatisfactory. In Excel 2003 most of the bugs pointed out in that paper have been corrected, but some have been replaced by new ones (Poisson, Binomial), and tail probabilities of the hypergeometric distribution are still missing. We also show that the gamma and beta distribution are not always computed correctly in Excel 2003. Furthermore I reported in Knüsel (2002) that Excel XP can compute *negative variances* and that the random number generator in the analysis toolpack (add-in package) is unsatisfactory as there exist only $2^{15} - 1 = 32767$ distinct real random numbers (uniform or normal); the first bug is no more to be found in Excel 2003 but the random number generator in the analysis toolpack has not changed. So my opinion about Excel 97 and Excel XP remains unchanged with the new version; also Excel 2003 can be recommended only for applications in which the accuracy of statistical distributions does not matter.

2. Numerical accuracy of some statistical distributions in Excel 2003

2.1. Poisson distribution (lower tail probability)

If X denotes a random variable with a Poisson distribution with parameter λ (=mean), then the Excel function POISSON computes for given k (=number of cases) and λ the tail probability $\Pr\{X \leqslant k\}$. Table 1 shows, that Excel 2003 now computes the correct results in the central part of the distribution, but extreme lower tail probabilities were computed correctly with Excel 97 while Excel 2003 rounds these probabilities down

Table 1
Poisson distribution

k	λ	$\Pr\{X \leqslant k\}$		
		Exact	Excel 97	Excel 2003
0	200	1.3839 E-87	Exact	0
10	200	4.1096 E-71	Exact	0
50	200	6.8158 E-37	Exact	0
100	200	3.7236 E-15	Exact	0
103	200	2.8916 E-14	Exact	0
104	200	5.6170 E-14	Exact	2.7254 E-14
110	200	2.4813 E-12	Exact	2.4524 E-12
133	200	2.9439 E-7	Exact	Exact
134	200	4.4562 E-7	No result	Exact
200	200	0.51879	No result	Exact
250	200	0.99972	No result	Exact

Table 2
Binomial distribution

k	n	ϑ	$\Pr\{X \leqslant k\}$		
			Exact	Excel 97	Excel 2003
10	1030	0.5	3.111 E-287	Exact	0
50	1030	0.5	3.941 E-225	Exact	0
100	1030	0.5	1.394 E-169	Exact	0
200	1030	0.5	5.4578 E-92	Exact	0
390	1030	0.5	3.1820 E-15	Exact	0
391	1030	0.5	5.2410 E-15	Exact	2.0590 E-15
400	1030	0.5	3.8974 E-13	Exact	3.8655 E-13
499	1030	0.5	0.16704	Exact	Exact
500	1030	0.5	0.18311	No result	Exact
515	1030	0.5	0.51243	No result	Exact
550	1030	0.5	0.98655	No result	Exact
575	1030	0.5	0.99992	No result	Exact

to zero. So a bad algorithm in Excel 97 has been replaced by another bad algorithm in Excel 2003.

2.2. *Binomial distribution (lower tail probability)*

If X denotes a random variable with a binomial distribution with parameters n (=number of trials) and ϑ (=probability for a success), then the Excel function BINOMDIST computes for given k (=number of successes), n and ϑ the tail probability $\Pr\{X \leqslant k\}$. Table 2 shows the same behaviour as observed with the Poisson distribution; Excel 2003 now computes the correct results in the central part of the distribution, but extreme lower tail probabilities were computed correctly with Excel 97 while Excel 2003 rounds these probabilities down to zero. Again a bad algorithm in Excel 97 has been replaced by another bad algorithm in Excel 2003.

2.3. *Hypergeometric distribution (only point probabilities)*

If X denotes a random variable with a hypergeometric distribution with parameters N (=total number of balls in the urn), M (=number of white balls in the urn), and n (=sample size), then the Excel function HYPGEOMDIST computes for given k (=number of white balls in the sample), N, M, and n the point probability $\Pr\{X = k\}$. These probabilities are now computed correctly also for $N \geqslant 1030$. But there is still no option in Excel 2003 to compute the tail probabilities $\Pr\{X \leqslant k\}$ for the hypergeometric distribution! In my view this is a weak point as the hypergeometric distribution plays a very important role in many statistical fields (random sampling without replacement).

2.4. *Gamma distribution*

If X denotes a random variable with a Gamma distribution with shape parameters α, then the Excel function GAMMADIST computes for given x and α the probability

Table 3
Gamma distribution

x	α	$\Pr\{X < x\}$	
		Exact	Excel 2003
0.3	0.2	0.816 527	Exact
0.2	0.2	0.764 435	Exact
0.1	0.2	0.676 043	Exact
0.2	0.1	0.879 420	Exact
0.1	0.1	0.827 552	#NUM!
0.05	0.1	0.775 539	Exact
0.2	0.05	0.939 209	Exact
0.1	0.05	0.911 258	#NUM!
0.05	0.05	0.882 244	#NUM!
0.01	0.05	0.815 560	Exact
0.001	0.05	0.727 179	Exact

Table 4
Inverse beta distribution

p	α	β	x such that $\Pr\{X < x\} = p$	
			Exact	Excel 2003
0.001	5	2	0.181 386	0.181 396
1E-6	5	2	0.044 427	0.042 969
0.001	10	100	0.027 946	Exact
1E-6	10	100	0.012 149	0.011 719

$\Pr\{X < x\}$. (We only consider the standard case with scale parameter $\beta = 1$ in Excel notation). As can be seen from Table 3 this function can have numerical problems just in the most important central part of the distribution.

2.5. *Inverse beta distribution*

If X denotes a random variable with a Beta distribution with shape parameters α and β, then the Excel function BETAINV computes for given probability p the value of x such that $\Pr\{X < x\} = p$. (We only consider the standard case, where the range of the distribution is the interval $[0, 1]$, i.e. $A = 0$ and $B = 1$ in Excel notation). As can be seen from Table 4 the results are not correct for small values of p.

Acknowledgements

The author would like to thank the referees for helpful comments that led to improvements of the paper.

References

Knüsel, L., 1998. On the accuracy of the statistical distributions in Microsoft Excel 97. Comput. Statist. Data Anal. (CSDA) 26, 375–377.

Knüsel, L., 2002. On the reliability of Microsoft Excel XP for statistical purposes. Comput. Statist. Data Anal. (CSDA) 39, 109–110 (With Corrigendum in CSDA, 40, 835).