Simulation Study to Compare the Random Data Generation from Bernoulli Distribution in Popular Statistical Packages

Manash Pratim Kashyap
Department of Business Administration
Assam University, Silchar, India
kashayap.manashaus@gmail.com

Nadeem Shafique Butt
PIQC Institute of Quality
Lahore, Pakistan
nadeemshafique@piqc.com.pk

Dibyojyoti Bhattacharjee
Department of Business Administration
Assam University, Silchar, India
dibyojyoti.bhattacharjee@gmail.com

Abstract

In study of the statistical packages, simulation from probability distributions is one of the important aspects. This paper is based on simulation study from Bernoulli distribution conducted by various popular statistical packages like R, SAS, Minitab, MS Excel and PASW. The accuracy of generated random data is tested through Chi-Square goodness of fit test. This simulation study based on 8685000 random numbers and 27000 tests of significance shows that ability to simulate random data from Bernoulli distribution is best in SAS and is closely followed by R Language, while Minitab showed the worst performance among compared packages.

Keywords: Bernoulli distribution, Goodness of Fit, Minitab, MS Excel, PASW, SAS, Simulation, R language

1. Introduction

Use of Statistical software is increasing day by day in scientific research, market surveys and educational research. It is therefore necessary to compare the ability and accuracy of different statistical softwares. This study focuses a comparison of random data generation from Bernoulli distribution among five softwares R, SAS, Minitab, MS Excel and PASW (Formerly SPSS). The statistical packages are selected for comparison on the basis of their popularity and wide usage. Simulation is the statistical method to recreate situation, often repeatedly, so that likelihood of various outcomes can be more accurately estimated. Simulation forms a central part, because of the relative ease with which samples can often be generated from a probability distribution, even when the density function cannot be explicitly integrated Sharma (2006).

McCullough (1998) suggested that random numbers generation should be one of the characteristics of software comparison. Das, Roy, & Bhattacharjee (2009) considered MS Excel and R for comparison, using mean square errors (MSE) of maximum likelihood estimates of the parameter *p* of Bernoulli distribution:

In this paper R 2.11.1, SAS 9.1.3, Minitab 15, MS Excel 2007 and PASW 18 are explored in term of their accuracy of generating random data from Bernoulli distribution. Following steps are used for comparison

(i) Random samples are generated for different sizes (n=30, 50, 100, 250, 500 and 1000), for different values of the parameter *p* in range of 0.1(0.1)0.9 form the Bernoulli distribution using above mentioned softwares.
(ii) The maximum likelihood estimates of the parameter are obtained for a fixed sample size and fixed value of the parameter mentioned in step (i).
(iii) The procedure is replicated one hundred times.
(iv) Chi-Square goodness of fit test is conducted for a given sample size and a given value of *p* for each package, and number of poor fits are recorded in hundred replications.

2. Literature Review

A number of reviews of literature concerning statistical software for microcomputers has been provided by the researcher and offered very useful comments to both and users and vendors. Searle (1989) discussed the review concerning statistical software comparison. On the recommendation of "*The American Statistical Association"* Francis, Heiberger, & Velleman (1975) focused a comprehensive study of the performance of statistical packages. This study subsequently modified and published a monograph. Study by Francis (1981) was the first systematic attempt to evaluate the performance of the software used in academics and industry for critical statistical applications. Dallal (1992) compared different computing packages viz. SAS and SPSS. They analyze unbalance data from fixed model with nested factors. Dallal (1992) found differences between SAS and SPSS results beside some error of calculations of sums of squares in SPSS output. There are certain literature review regarding statistical software comparison Wilkinson & Dallal (1977), Anscombe (1981), Hayes (1982), Wilkinson (1985), Simon James & Stephen (1988, (1989), L'Ecuyer (1992), Wilkinson (1994), Knüsel (1998), Rogers, et al. (1998) etc.

Okunade, Chang, & Evans (1993) compared the output of summary statistics of regression analysis in commonly used statistical and econometric packages such as SAS, SPSS, SHAZM, TSP, and BMDP. Oster (1998) reviewed five statistical software packages (EPI INFO, EPICURE, EPILOG PLUS, STATA, and TRUE EPISTAT) according to criteria that are of most interest to epidemiologists, biostatisticians, and others involved in clinical research. McCullough (1998) proposed testing the accuracy of statistical software packages using Wilkinson's Statistics Quiz in three areas: linear and nonlinear estimation, random number generation, and statistical distributions. Again, McCullough (1999) used his methodology using SAS, SPSS, and S-Plus. McCullough (1999) showed that the reliability of statistical software cannot be taken for granted because he found some weak points in all random number generators; e.g.; S-plus correlation procedures, and the one-way ANOVA and nonlinear least squares routines of SAS and SPSS. Zhou, Perkins, & Hui (1999) reviewed five software packages

viz. MLN, MLWIN, SAS Proc Mixed, HLM, and VARC that can fit a generalized linear mixed model for data with more than two-level structure and multiple number of independent variables. Bergmann, Ludbrook, & Spooren (2000) Compared 11 statistical packages on real dataset. The study was based upon SigmaStat 2.03, SYSTAT 9, JMP 3.2.5, S-Plus 2000, STATISTICA 5.5, UNISTAT 4.53b, SPSS 8, Arcus Quickstat 1.2, Stata 6, SAS 6.12, and StatXact 4. They found that different packages could give very different outcomes for the Wilcoxon-Mann-Whitney test.

3. About the Software Packages

In this section we have given a brief description of selected packages

3.1 R Language

R is an integrated software facility for data manipulation, calculation and graphical display. It has a suite of operators for calculations on arrays; a large coherent and integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and a well developed, simple and effective programming language which includes conditionals, loops, and user defined recursive functions and input and other facilities. R has been widely accepted in the scientific world in general and statistical community in particular.

3.2 SAS

Statistical Analysis System (SAS) is based on SAS programs that define a sequence of operations to be performed on data stored as tables. Although graphical user interfaces to SAS exist (such as the SAS Enterprise Guide), most of the time these GUIs are just a front-end to automate or facilitate generation of SAS programs. SAS components expose their functionalities via application programming interfaces, in the form of statements and procedures.

3.3 Minitab

Minitab is a statistics package which was developed at the Pennsylvania State University by Barbara F. Ryan, Thomas A. Ryan, Jr., and Brian L. Joiner in 1972. Minitab is distributed by Minitab Inc, a privately owned company headquartered in State College, Pennsylvania, with subsidiaries in Coventry, England (Minitab Ltd.) Paris, France (Minitab SARL) and Sydney, Australia (Minitab Pty.).

3.4 MS Excel

Microsoft Excel is an integral part of Microsoft Office package by Microsoft Corporation. MS excel is a window-based spreadsheet. It provides the user with several facilities like data analysis and data handling. The Excel worksheet gives a list of function like Financial, Date and Time, Math, Trigonometrical, Statistical, Logical, Database etc. In addition to this Excel has an add-in called as the Data Analysis Tool Pak that can be used for different types of statistical analysis including simulation from distributions.

3.5 PASW

Predictive Analytical Software (PASW) formerly known as SPSS was first time released in its first version in 1968 after being founded by Norman Nie and C. Hadlai Hull. Nie was then a political science postgraduate at Stanford University, and now Research Professor in the Department of Political Science at Stanford and Professor Emeritus of Political Science at the University of Chicago. PASW is among the most widely used programs for statistical analysis in social science. It is used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations and others. The original SPSS manual by Nie & Dale (1970) has been described as 'Sociology's most influential book'. In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary is stored with the data) are features of the base software.

4. Methodology of Comparison

James Bernoulli (1654-1705) was discovered Bernoulli distribution of random variable X which takes only two values 0 and 1 with probabilities p and q respectively, where $p+q=1$. In other words, if X is a random variable such that $P(X=1)=p$ and $P(X=0)=q$, $q=p-1$ than random variable X is called Bernoulli variate and said to have a Bernoulli distribution. The probability mass function is given by

$$p(X = x) = p^x q^{1-x} \; ; \; x=0 \text{ and } 1; \; p+q=1$$

where, p is the parameter of the Bernoulli distribution.

If $X=(X_1, X_2, X_3, \ldots, X_n)$ is a random sample from the Bernoulli distribution than the maximum likelihood estimator (MLE) of the parameter of the Bernoulli distribution is given by $p = \frac{x}{n}$ for any sequence of n Bernoulli trials resulting in x 'successes'.

The Bernoulli distribution has central importance in the theory of probability and statistics. When there is situation of success or failure than the Bernoulli distribution provides the best result. The Bernoulli distribution has no direct utility, but binomial distribution which can be obtained from Bernoulli distribution has greater importance in describing enormous variety of real life example.

4.1 Simulation and Random Data Generation

In this section we discuss Bernoulli random data generation from softwares used in the study. Following commands can be used for random data generation from Bernoulli distribution

R Language 2.11.1: *rbinom(n, size, prob);* where "*n*" is number of random numbers to be drawn, "*k*" is the number of independent trails and "*prob*" is the parameter of the Bernoulli distribution.

SAS 9.1.3: *RANBIN(seed,n,p,x);* where "*seed"* is the random number seed value, Range of *seed* < 2^31 - 1, "*n*" is an integer number of independent Bernoulli trials, "*p*" is a numeric probability of success parameter and "*x*" is a numeric SAS variable. A new value for the random variate x is returned each time CALL RANBIN is executed.

Minitab 15: *Random n Var; Bernoulli P*.; where "*Random"* is Minitab command to generate random numbers, "*n*" is number of random numbers to be drawn, "*Bernoulli*" keyword specifies the Bernoulli distribution and "*p*" is a numeric probability of success parameter.

MS Excel 2007: The simulation can be done by using Analysis Toolpack add-in.

PASW 18: *RV.BERNOULLI(P);* where "*RV.BERNOULLI"* is PASW keyword and "*P*" is the probability of success Parameter

5. Calculation and Results

Based on the methodology discussed in section 1, simulations from specific software and relevant calculations are performed and numbers of poor fits are recorded in hundred replications. Numbers of poor fits are given in Table 5.1 at each combination of sample size, Probability of success event and packages used.

Table 5.1: No. of poor fits in 100 replications under various probabilities by various packages

Sample Size	Package	Probability								
		0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
30	R	2	2	3	3	2	2	6	4	0
	SAS	4	2	7	3	3	3	7	2	4
	Minitab	4	4	10	3	3	3	7	2	4
	Excel	2	2	2	3	5	6	3	7	3
	SPSS	3	2	4	3	4	4	2	4	5
50	R	1	4	1	4	7	6	7	6	1
	SAS	2	3	2	8	5	8	2	3	2
	Minitab	8	8	2	8	5	8	2	3	2
	Excel	1	0	4	4	9	6	6	4	1
	SPSS	1	4	3	7	5	4	3	5	3
100	R	7	4	4	9	3	5	5	1	5
	SAS	10	4	1	5	2	5	1	4	10
	Minitab	10	9	1	5	2	5	1	4	10
	Excel	3	3	10	7	7	9	6	7	9
	SPSS	6	6	3	5	5	7	4	4	5

Sample Size	Package	Probability								
		0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
250	R	1	4	1	2	4	2	5	4	7
	SAS	7	5	3	4	7	4	3	5	7
	Minitab	7	5	3	4	4	4	3	9	7
	Excel	6	2	6	8	2	5	2	3	1
	SPSS	3	3	1	8	6	3	2	11	4
500	R	2	4	4	4	4	4	6	6	5
	SAS	2	3	3	1	3	1	3	3	2
	Minitab	8	5	4	1	3	3	3	6	2
	Excel	5	5	4	3	4	3	10	3	4
	SPSS	4	6	4	6	2	5	8	5	7
1000	R	5	4	1	4	3	4	6	6	6
	SAS	5	2	1	5	6	5	1	2	5
	Minitab	4	6	10	5	6	11	3	6	5
	Excel	9	6	5	3	2	3	7	8	4
	SPSS	5	3	3	3	5	7	3	3	10

* each cell of the table represent the number of poor fits out of hundred replication at 5% level of significance.

To decide which package is performing best the results of simulation study are pooled over various sample sizes and presented in Table 5.2.

Table 5.2: No. of poor fits (Rank) in 600 replications under various probabilities by various packages

Packages	*p*									Total	Rank
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**		
R	**18(1)**	22(3)	**14(1)**	**26(1)**	**23(1)**	**23(1)**	35(5)	27(2)	24(2)	212	2
SAS	30(4)	19(2)	17(2)	**26(1)**	26(3)	26(2)	**17(1)**	**19(1)**	30(3)	210	1
Minitab	41(5)	37(5)	30(4)	**26(1)**	**23(1)**	34(5)	19(2)	30(3)	30(3)	270	5
Excel	26(3)	**18(1)**	31(5)	28(4)	29(5)	32(4)	34(4)	32(4)	**22(1)**	252	4
SPSS	22(2)	24(4)	18(3)	32(5)	27(4)	30(3)	22(3)	32(4)	34(5)	241	3

It can be clearly seen from Table 5.2 that overall SAS is performing best as out of 5400 replication only 210 poor fits are observed at 5% level of significance, and is followed by R Language with 212 poor fits out of 5400 tests. Based on number of poor fits each package is ranked and Minitab is ranked 5 indicating that it is worst software included in this study in terms of Bernoulli random data generation.

The bold face items in Table 5.2 shows best performance of each package at various success probabilities. Performance of R is best at (P=0.1, 0.3 and 0.6), SAS shows its best performance at (P= 0.7 and 0.8) and Excel shows best performance against (P=0.2 and 0.9). While around middle success probabilities

R, SAS and Minitab is performing equally well. Overall PASW is at number 3 in ranking but it has no best performance at various success probabilities.

Further simulation study is also explored to see the effect of sample size on ability to generate Bernoulli random data by various packages. The results are described in Table 5.3.

Table 5.3: No. of poor fits (Rank) in 900 replications under various sample sizes by various packages

Packages	Sample Size					
	30	**50**	**100**	**250**	**500**	**1000**
R	**24(1)**	37(4)	43(2)	**30(1)**	39(3)	39(2)
SAS	35(4)	**35(1)**	**42(1)**	45(4)	**21(1)**	**32(1)**
Minitab	40(5)	46(5)	47(4)	46(5)	35(2)	56(5)
Excel	33(3)	**35(1)**	61(5)	35(2)	41(4)	47(4)
SPSS	31(2)	**35(1)**	45(3)	41(3)	47(5)	42(3)

This can be seen from Table 5.3 that R is performing best at small sample size only, while SAS is performing consistently over various sample sizes under study. A very slow random number generation was observed for large sample sizes in MS Excel 2007.

It can be concluded from this simulation study that SAS and R Language are performing close to each other. Recommendation can be given in favor of R Language over SAS as R Language is available free of cost.

6. Future Directions

With increase in the number of repetitions things may change but very marginally. However it is essential to take up such simulation study for other probability distributions and hence a conclusion should be reached on the overall simulation capability packages under study.

Acknowledgments

We are thankful for the referees of suggesting useful points, which improve the quality of the paper. We are also thankful to Dr. Muhammad Qaiser Shahbaz who supported us a lot throughout this project.

References

1. Anscombe, F. (1981). *Computing in statistical science through APL*: SPRINGER-VERLAG INC, 175 FIFTH AVE, NEW YORK, NY.
2. Bergmann, R., Ludbrook, J., & Spooren, W. (2000). Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages. *American Statistician, 54*(1), 72-77.
3. Dallal, G. (1992). The computer analysis of factorial experiments with nested factors. *The American Statistician, 46*, 240.

4. Das, K. K., Roy, T. D., & Bhattacharjee, D. (2009). Comparing the Ability of MS Excel and R While Simulating from Poisson Distribution. *Assam University Journal of Science and Technology 4*(2), 1-6.
5. Francis, I. (1981). *Statistical software: a comparative review*: Elsevier North-Holland, Inc. New York, NY, USA.
6. Francis, I., Heiberger, R., & Velleman, P. (1975). Criteria and considerations in the evaluation of statistical program packages. *American Statistician*, 52-56.
7. Hayes, A. (1982). Statistical Software: A survey and Critique of its Development. *Office of Naval Research, Arlington, VA*.
8. Knüsel, L. (1998). On the accuracy of statistical distributions in Microsoft Excel 97. *Computational Statistics & Data Analysis, 26*(3), 375-377.
9. L'Ecuyer, P. (1992). *Testing random number generators*.
10. McCullough, B. (1998). Assessing the reliability of statistical software: Part I. *The American Statistician, 52*(4), 358-366.
11. McCullough, B. (1999). Assessing the reliability of statistical software: Part II. *The American Statistician, 53*(2).
12. Nie, N., & Dale, H. (1970). Bent, and C. Hadlai Hull. 1970. Statistical Package for the Social Sciences: New York: McGraw-Hill.
13. Okunade, A., Chang, C., & Evans, R. (1993). Comparative Analysis of Regression Output Summary Statistics in Common Statistical Packages. *The American Statistician, 47*(4).
14. Oster, R. (1998). An examination of five statistical software packages for epidemiology. *The American Statistician, 52*(3).
15. Rogers, J., Filliben, J., Gill, L., Guthrie, W., Lagergren, E., & Vangel, M. (1998). *Strd: Statistical reference datasets for testing the numerical accuracy of statistical software*: Technical report, National Institute of Standards and Technology, Washington, DC.
16. Searle, S. (1989). Statistical computing packages: Some words of caution. *The American Statistician, 43*(4), 189-190.
17. Sharma, J. (2006). Operations Research: Theory and Applications. *Mac Millan India Ltd*.
18. Simon James, P., & Stephen, D. (1988). Benchmarking numerical accuracy of statistical algorithms. *Computational Statistics & Data Analysis, 7*(2), 197-209.
19. Simon James, P., & Stephen, D. (1989). Assessing the accuracy of ANOVA calculations in statistical software. *Computational Statistics & Data Analysis, 8*(3), 325-332.
20. Wilkinson, L. (1985). Statistics quiz. *IL. SYSTAT, Evanston*.
21. Wilkinson, L. (1994). Practical guidelines for testing statistical software. *Computational Statistics. Physica-Verlag, Berlin*, 111–124.
22. Wilkinson, L., & Dallal, G. (1977). Accuracy of sample moments calculations among widely used statistical programs. *American Statistician, 31*(3), 128-131.
23. Zhou, X., Perkins, A., & Hui, S. (1999). Comparisons of software packages for generalized linear multilevel models. *The American Statistician, 53*(3).