Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 8 (6): 1122-1124, 2008
ISSN 1812-5654

Optimal Estimator for Sample Size Using Monte-Carlo Method

H. Bevrani, M. Ghorbani and M.K. Sadaghiani
Department of Statistics, Faculty of Mathematical Science,
Research Institute for Fundamental Sciences, University of Tabriz, Tabriz, Iran

Abstract: In this study, we construct the optimal estimator for sample size, which were sufficient for maintenance the demanded accuracy and reliability. The goal of this paper is presenting three estimators such as follow. The first one which is traditional approach and rough enough is based on the Chebyshev's inequality. The second one is based on the central limit theorem, but it doesn't take into account the accuracy of the normal approximation. The third estimator is based on Berry-Esseen's inequality that takes into account the accuracy of the normal approximation and is guaranteed.

Key words: Chebyshev's inequality, Berry-Esseen's inequality, Monte-Carlo method

INTRODUCTION

The Monte Carlo method provides approximate solutions to a variety of mathematical problems (Bauer, 1958). As is well known, the Monte-Carlo method is composed from three composite parts. Firstly, this is a simulation of random variables with the known distributions, secondly, construction of probability models for real processes and at last, problems of theory of the statistical estimation (Rubenstein, 1981). Certainly, the basic ideas of this method are the law of large numbers and the central limit theorem (Ermakov, 1971; Bevrani, 2003; Gentle, 2004). In both cases the sample size is unknown. Frequently there is a question, whether enough the available statistical data that the inference made on their basis, were exact and reliable, in other words, whether available sampling is representative. Also it is rather general problem. Therefore the purpose of the given article is the estimation of sample's value for the Monte-Carlo method.

THE MONTE-CARLO METHOD

Let it is required to calculate approximately model I with the help of a Monte-Carlo method. Then it is necessary to find an random variable U, such, that its mathematical expectation is equal I: EU = I.

Let's consider (n+1) independent identically distributed random variables U_1, U_2,...,U_n with the finite second moments. Then from the central limiting theorem it follows that;

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}U_i - I\right| < 3\sqrt{\frac{DU}{n}}\right\} \approx \Phi(3) \approx 0.998 \quad (1)$$

where, $\Phi(x)$ is a standard normal distribution function.

This relation means, that if we have sufficiently big amount of observations U_1, U_2,..., U_n, the required model can be approximately calculated as follows:

$$I_n^* = \frac{1}{n}\sum_{i=1}^{n}U_i \approx I \quad (2)$$

Thus, with the probability near to 0.998, we mistake on value, not exceeding $3\sqrt{\frac{DU}{n}}$. Easy to see, that $EI_n^* = I$.

ESTIMATION OF SAMPLE SIZE

Let's consider the problem on the accuracy of the approximation $I_n^* \approx I$. Unfortunately, unlike the determined (nonrandom) schemes, analysis of random data requires more then one parameter describing the accuracy, as event $|I_n^* - I| \le \varepsilon$ is random, for any $\varepsilon \in (0, 1)$, that is, for one sampling this event may happen and for any another-may not. Therefore alongside with the parameter ε describing the accuracy, we'll set one more parameter $\gamma \in (0, 1)$-confidence of a statistical inference. We'll require, that the probability of the indicated event was not less then γ, that is,

$$P\left\{|I_n^* - I| \ge \varepsilon\right\} \le 1 - \gamma \quad (3)$$

Corresponding Author: H. Bevrani, Department of Statistics, Faculty of Mathematical Sciences,
Research Institute for Fundamental Sciences, University of Tabriz, Tabriz, Iran
Tel: +98(0411)339 2871

Thus it is clear, that ε should be close to zero and γ should be close to unit, characterizing our confidence of the regularity of the inference. Now we are passing to the estimation of a sample size. We'll start with traditional approaches, using the Chebyshev's inequality and the central limiting theorem. Then we'll consider more accurate estimates which take into account an error of normal approximation. These estimates will be based of the Berry-Esseen's inequality and it's more exact analogue for the case of smooth distributions.

Solution based of the Chebyshev's inequality: On the Chebyshev's inequality:

$$P\left\{\left|I_n^* - I\right| \geq \varepsilon\right\} \leq \frac{DI_n^*}{\varepsilon^2} \tag{4}$$

Hence, condition (3) is satisfied, if $\frac{DI_n^*}{\varepsilon^2} \leq 1-\gamma$. Denote $DU = \sigma^2$, then $DI_n^* = \frac{\sigma^2}{n}$ and a low bound for the number of observations will look like:

$$n \geq \frac{\sigma^2}{\varepsilon^2(1-\gamma)} \tag{5}$$

Solution based on the central limit theorem: As is well known, the Chebyshev's inequality is rather rough, therefore, using the Central Limiting Theorem (CLT) instead of it permits to hope, that estimates for the necessary sample size and appropriate accuracy would be more optimistically. CLT implies, that for the sufficiently big n

$$P\left\{\left|I_n^* - I\right| \geq \varepsilon\right\} = P\left\{\left|\frac{\sum U_i - nI}{\sqrt{nDU}}\right| \geq \frac{\varepsilon\sqrt{n}}{\sqrt{DU}}\right\} \approx 2\left[1-\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{DU}}\right)\right] \tag{6}$$

Taking into account requirements on the confidence of our inference, we obtain, that probability (6) should be no more than $1-\gamma$:

$$2\left[1-\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{DU}}\right)\right] \leq 1-\gamma,$$

when in view of definition α-quantiles z_α of the standard normal law we obtain

$$n \geq \frac{z_{\frac{1+\gamma}{2}}^2 \cdot \sigma^2}{\varepsilon^2} \tag{7}$$

As it is easy to see, estimates (5) and (7) differ only in the factors $(1-\gamma)^{-1}$ and $z_{\frac{1+\gamma}{2}}^2$. For example, let us assign $\gamma = 0.95$, then the condition (5) requires that the relation was $\frac{n\varepsilon^2}{\sigma^2}$ not less than 20, while the (7) one-only 3.85 ($z_{0.975} = 1.96$), that is more, than five times better. Such in the image, the CLT allows to receive more optimistically estimates, however optimism from apparent advantage of the solution based on the CLT, doesn't owe us to weaken. The matter is that the Chebyshev's inequality gives though rough, but absolutely correct, guaranteed estimates for the sample's value and for the accuracy. At the same time, attracting the CLT, we use approximate equality (6), which brings itself an error into the inference. In the following section we'll correct this lack.

Solutions which take into account the accuracy of the normal approximation: The Berry-Esseen inequality as an estimate of the rate of convergence in the CLT is well known in the probability theory. This estimate holds for an arbitrary distribution with the finite third moment.

Assume, that the random variable U has the finite third moment and denote $\beta^3 = M|U-I|^3$. Then, applying the Berry-Esseen's inequality to the accuracy estimation of relation (6), we obtain:

$$\Delta_n = \left|P\left\{\left|I_n^* - I\right| \geq \varepsilon\right\} - 2\left[1-\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{DU}}\right)\right]\right| \leq \frac{L_3}{\sqrt{n}} \tag{8}$$

where, $L_3 = \frac{C_0\beta^3}{\sigma^3}$ and C_0 is an absolute constant with the upper bound $C_0<0.7655$ (Shiganov, 1986; Korolev and Shevtsova, 2006). Thus, more accurate estimate for the sample's value is as follows:

$$n \geq \frac{z_{(\frac{1+\gamma}{2}+\frac{L_3}{\sqrt{n}})}^2 \sigma^2}{\varepsilon^2} \tag{9}$$

Results analysis: Let $\sigma^2 = 1$. Then the required sample's value can be easily computed with the help of relations Eq. 5, 7, 9 and 11. The outcomes of these computations are shown in the Table 1 and 2 (Appendix). The first Table 1 is constructed for $\varepsilon = 0.001$ and the second one-for $\varepsilon = 0.01$. The upper rows contain values of confidence level γ, the second and the third ones-values of the samples sizes, obtained by using the Chebyshev's inequality (Eq. 5) and the CLT (Eq. 7), accordingly. A marginal left column contains the values of L_3 (from 0.7655 till 2.1655 with the step 0.1). We consider so lower bound for L_3, because as it follows from the Lyapunov's inequality $\frac{\beta^3}{\sigma^3} \geq 1$ and therefore $L_3 \geq C_0$. The sample's value

can be found in the intersection of the row with appropriate value L_3 and the column with required confidence level γ.

APPENDIX

Table 1: Estimations for the sample's value when $\varepsilon = 0.001$

γ	0.900	0.925	0.950	0.975	0.990
$n_{cheb.}$	10000000	13333333	20000000	40000000	100000000
n_{CLT}	2705544	3170053	3841458	5023886	6634896
L_3	Sample's value for an arbitrary distribution				
	---	---	---	---	---
0.7655	2720423	3188840	3867783	5071654	6743228
0.8655	2722372	3191303	3871240	5077962	6757754
0.9655	2724322	3193768	3874702	5084286	6772370
1.0655	2726274	3196235	3878169	5090626	6787076
1.1655	2728226	3198704	3881640	5096982	6801874
1.2655	2730180	3201175	3885116	5103355	6816766
1.3655	2732135	3203649	3888595	5109743	6831751
1.4655	2734091	3206124	3892080	5116148	6846832
1.5655	2736048	3208602	3895569	5122570	6862009
1.6655	2738007	3211082	3899062	5129007	6877283
1.7655	2739967	3213564	3902560	5135462	6892657
1.8655	2741928	3216048	3906063	5141933	6908130
1.9655	2743890	3218534	3909570	5148421	6923705
2.0655	2745853	3221022	3913081	5154926	6939382
2.1655	2747818	3223512	3916598	5161447	6955164

Table 2: Estimations for the sample's value when $\varepsilon = 0.01$

γ	0.9	0.925	0.95	0.975	0.99
$n_{cheb.}$	100000	133333	200000	400000	1000000
n_{CLT}	27055	31701	38415	50239	66349
L_3	Sample's value for an arbitrary distribution				
	---	---	---	---	---
0.7655	28576	33637	41170	55482	80313
0.8655	28781	33899	41552	56255	82894
0.9655	28986	34164	41939	57052	85761
1.0655	29193	34432	42332	57874	88974
1.1655	29401	34702	42730	58724	92614
1.2655	29611	34974	43134	59602	96785
1.3655	29822	35249	43545	60511	101626
1.4655	30035	35527	43962	61453	107320
1.5655	30249	35808	44385	62430	114096
1.6655	30465	36091	44815	63444	122226
1.7655	30682	36377	45252	64498	131976
1.8655	30900	36666	45695	65596	143523
1.9655	31121	36958	46147	66739	156865
2.0655	31343	37253	46605	67933	171817
2.1655	31566	37551	47072	69180	188119

ACKNOWLEDGMENT

This research has been supported by the Research Institute for Fundamental Sciences, Tabriz, Iran. The authors would like to thank this support.

REFERENCES

Bauer, W.F., 1958. The Monte Carlo method. J. Soc. Ind. Applied Math., 6 (4): 438-451.

Bevrani, H., 2003. Generalization of a Monte-Carlo method for a solve of definite integrals. Reports of 11th All-Russia Conference Mathematical Methods of Pattern Recognition, Moscow, pp: 208-210.

Ermakov, S.M., 1971. The Monte-Carlo method and related problems. Nauka, Moscow (In Russian).

Gentle, J.E., 2004. Random Number Generation and Monte Carlo Methods. Springer.

Korolev, V.Y. and I.G. Shevtsova, 2006. On the accuracy of normal approximation. Probability Theor. Appl., 50 (2): 298-310.

Rubenstein, R.Y., 1981. Simulation and the Monte Carlo Method, Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York.

Shiganov, I.S., 1986. Refinement of the upper bound on the constant in the central limit theorem. J. Soviet Math., 35: 2545-2551.