APPLICATION OF THE SUFFICIENT EMPIRICAL AVERAGING METHOD IN ONE STATISTICAL PROBLEM SOLVING

Catherine Zhukovskaya
Faculty of Transport and Mechanical Engineering
Riga Technical University
Kalku 1, LV-1658, Riga, Latvia
E-mail: kat_zuk@hotmail.com

KEYWORDS

estimation, complete sufficient statistic, inventory control, optimization

ABSTRACT

In this article we discuss the new approach to preparing the input data in system simulation, the so called Sufficient Empirical Averaging (SEA) method. It assumes the existence of the complete sufficient statistics for unknown parameters of input variable distributions. The application of this method allows getting unbiased estimates with minimum variance.

INTRODUCTION

We know that the preparation of the primary data is a very important stage in modeling. In particular, it includes the estimation of the parameters of the random variables used in simulation. In the traditional approach, so called the "*Plug-In*" method, the observations of input variables are used for estimation of probability distributions of these variables. According to these distributions the pseudo-random numbers are produced in the process of simulation, and then they replace the latter variables. Here the estimation of probability distributions leads to making mistakes in choosing the form of a distribution and estimating its parameters. However, the primary statistical samples, on the base of which the estimation of the parameters of distribution is executed, is not sufficient, it leads to the bias of estimated parameters and results of simulation.

The alternative to the "Plug-In" method is the "*Resampling Approach*" method (Andronov and Merkuryev 2002). According to this method well known method the random variables aren't worked out by generators of the random numbers but derived from initial statistical samples according to this or that random mechanism. This method doesn't use preliminary information about the type of the distributions of the random variables, i. e. in fact it is nonparametrical. It is a big advantage of such approach. Nevertheless, if such information is available, it should be also used, and in this case the method of *Sufficient Empirical Averaging method* (*SEA*-method) will be effective (Chepurin 1994, 1995, 1999).

The SEA-method is used when the supposed distributions have the sufficient statistics. It is based on the fact, that conditional distributions of the random variables, calculated with fixed values of sufficient statistics, don't depend on the unknown parameters of distributions. Consequently, the necessary random variables could be produced according to their conditional distributions. The received results of the imitative simulation will be unbiased. Moreover, if the applied sufficient statistics are complete (see Lehman 1983) then they will have the least variance.

The aim of this article is to illustrate the application of the SEA-method for one statistical problem of the inventory control. The SEA-method is described in the following section. An inventory problem setting and solving will be given in the third and forth sections. The results of the problem solution are given in the fifth section. The peculiarities of the discussed approach application are covered in the last section.

SUFFICIENT EMPIRICAL AVERAGING METHOD

Let us suppose that our aim is to estimate the mathematical expectation θ of function f of independent random variables $X_1, X_2, ..., X_n$: $\theta = Ef(X_1, X_2, ..., X_n)$. This function describes realiability or performance characteristics of considered system. Distributions of random variables $X_1, X_2, ..., X_n$ are unknown, but sample populations $\{H_i\}$ are available for each $\{X_i\}$.

The traditional approach to solving this problem supposes three stages: 1) hypothesis about distributions of random variables are formed; 2) unknown parameters of these distributions are estimated by using samples $\{H_i\}$; 3) the estimated distributions are used to estimate the mathemtical expectation θ by generation values of variables $\{X_i\}$ and simulating the function f on this base.

An alternate approach is the SEA method (Chepurin 1994, 195, 1999; Andronov, Zhukovskaya and Chepurin 2005). It is based on the concept of the sufficient

statistics. Let $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ as a sample, which has the corresponding distribution with unknown parameter θ. Then $\boldsymbol{S}$ *statistics* is any function of given sample $\boldsymbol{S} = \boldsymbol{S}(X_1, X_2, ..., X_n) = \boldsymbol{S}(\boldsymbol{X})$. The *statistics* is called *sufficient*, if the conditional distribution of sample $\boldsymbol{X} = (X_1, X_2, ..., X_n)$, which has been calculated on condition that the statistics $\boldsymbol{S}$ has a fixed value $\boldsymbol{s}$, which doesn't depend on the parameter θ (Lehman 1959, Cox and Hinkley 1974).

Let us denote $\boldsymbol{X}(\boldsymbol{s}) = \boldsymbol{X} \mid \{\boldsymbol{S} = \boldsymbol{s}\}$ the conditional sequence, that has been calculated on condition $\{\boldsymbol{S} = \boldsymbol{s}\}$. Note that $\boldsymbol{X}(\boldsymbol{s})$ "... is statistically equivalent to the data ($\boldsymbol{X}$), i.e. containing the same amount of information ..." (Chepurin 1999, p.182).

Let us for example estimate the unknown parameter λ of the exponential distribution. The probability density function of this distribution (p.d.f.):

$$f_X(x,\lambda) = \lambda e^{-\lambda x},\ x \geq 0;\ \lambda \in \Omega = (0,\infty) \qquad (1)$$

there $\Omega = (0, \infty)$ – is parametrical space, which consist all possible values of λ.

The sufficient statistics for the parameter λ on the sample $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ is the sum $S_n = X_1 + X_2 + \ldots + X_n$. It has Erlang distributions with parameters λ and n:

$$f_{S_n}(s,\lambda) = \lambda \frac{(\lambda s)^{n-1}}{(n-1)!} e^{-\lambda s}, \quad s \geq 0. \qquad (2)$$

The conditional p.d.f. of the sample $\boldsymbol{X}(\boldsymbol{s})$ at the point $x = (x_1, x_2, \ldots, x_n)$ is calculated by the formula:

$$f_X(\boldsymbol{x},\lambda \mid S_n = s) = \frac{\prod_{i=1}^{n-1}\left(\lambda e^{-\lambda x_i}\right)\lambda e^{-\lambda\left(s-\sum_{i=1}^{n-1} x_i\right)}}{f_{S_n}(x,\lambda)} = \frac{(n-1)!}{s^{n-1}}, \qquad (3)$$

$$\forall x_i \geq 0,\ x_1 + x_2 + \ldots + x_n \leq s. \qquad (4)$$

We see that one doesn't depend on the unknown parameterλ.

Now we suppose that we have generated a random value X. Firstly, we must calculate the sufficient statistics $\boldsymbol{S}$, secondly – generate the random value X on condition $\boldsymbol{S} = \boldsymbol{s}$. By this we use the conditional distribution X for the given $\boldsymbol{s}$.

Figure 1 allows us to compare traditional approach and SEA-method.

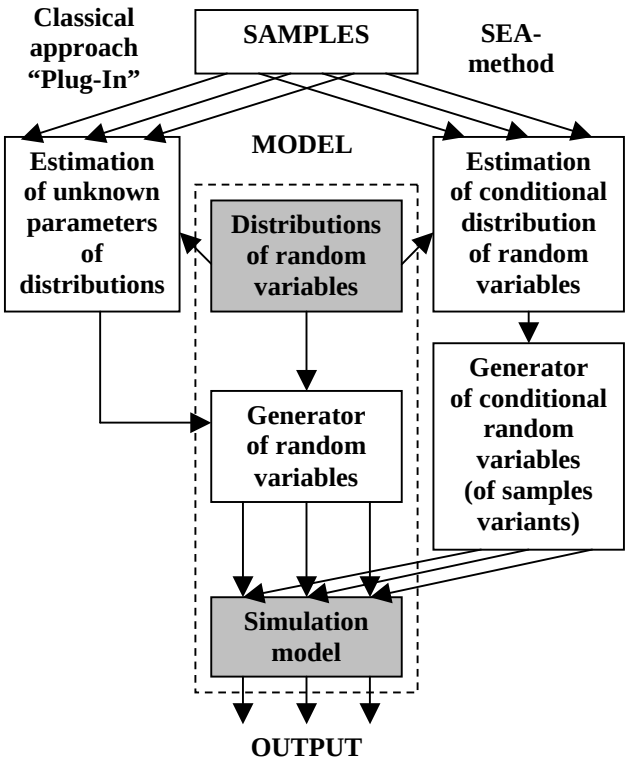

Figure 1. The comparison of classical approach and SEA-method

Note that according (3) the conditional distribution of the sample $X \mid \boldsymbol{S}_n = \boldsymbol{s}$ has a uniform distribution in the n-measure simplex (4). In principle this allows us to generate the conditional distributions of the random values with exponential distribution on condition that the sufficient statistics (sum) is fixed. But this method is inefficient. Often for generation the conditional distributions "special ways" can be used.

For example, for the exponential case (Engen and Lillegard 1997), the following way is supposed. Let the value $\boldsymbol{s}$ of the $\boldsymbol{S}_n$ is fixed. Firstly we generate n exponential distributed random variable $X^0_1, X^0_2, ..., X^0_n$ with parameter $\lambda = 1$. Secondly we calculate their sum: $S^0_n = \sum_{i=1}^{n} X_i^0$. Then random variables of interest $X^*_1, X^*_2, ..., X^*_n$ may be calculated by the formula $X^*_i = \frac{X_i^0 \cdot s}{S_n^0}$, $i = 1, 2, ..., n$.

PROBLEM SETTING

Let's examine the following statistical problem of the inventory control that has been considered earlier (Afanasyeva 2005). Initially there are k articles of a commodity in the warehouse. In the course of time the number of articles can be increased (because of supply some new ones), or decreased (because of selling some of them). The selling of articles are realized according

to the Puason flow with the intensity μ. Time intervals between supply of the articles have the exponential distribution with the parameter λ, shifted with the variable δ:

$$f(x) = \begin{cases} \lambda e^{-\lambda(x-\delta)}, & x > \delta, \\ 0, & otherwise. \end{cases}$$

If in the moment of receiving the order for a certain article we have it in the warehouse then this order is performed. Otherwise, the order is rejected and is not examined again.

The value of δ is known while the intensities λ and μ are unknown. We have two samples: the sample $X_1, X_2, ..., X_{n_a}$ of the fixed time intervals between the supply (of the size n_a) and the sample $Y_1, Y_2, ..., Y_{n_d}$ of the time intervals between the sale (of the size n_d). It is necessary to estimate the probability $P(i, k, t)$ that during the given time interval $(0, t)$ i refusals for orders will take place. Besides, it is necessary to estimate the optimal size of the initial stock $k^*(t)$ which maximizes the expected general income. This income is equal to the difference of cost values of the sold articles from the warehouse minus expenses for supply and storage of the initial stock of articles. The income from the sale of one article is c_d, but the expenses per one initially available article are c_a.

PROBLEM SOLVING

We will solve this problem with the help of the SEA-method. In our case the exponential distribution is taking place. The complete sufficient statistics for it are the sum of the sample elements and the sample size. Let's define such sums as A for the articles supply and D for articles sale. Then the simulation of the considered process in the given time t is applied. The time intervals between the supply and selling of the articles are generated according to the appropriate conditional distributions given fixed values (A, n_a) and (D, n_d).

Let's describe the procedure of generation of random variables with respect to supply process. So, as the initial we have the sufficient statistics (A, n_a), calculated with n_a intervals between articles supply. First n_a random variables are generated according to the exponential distribution with the parameter 1. Let's define them as $X_1^0, X_2^0, ..., X_{n_a}^0$. Then let's calculate their sum

$$A^0 = \frac{1}{n_a} \sum_{i=1}^{n_a} X_i^0 . \qquad (5)$$

The needed values of the intervals $X_1^*, X_2^*, ..., X_{n_a}^*$ between articles sale are calculated by the formula

$$X_i^* = X_i^0 (A - \delta n_a) \frac{1}{A^0} + \delta , \; i = 1, 2, \ldots n_a. \qquad (6)$$

In the same way the intervals $Y_1^*, Y_2^*, ..., Y_{n_d}^*$ between the articles supply are generated (for $\delta = 0$). The simulation of the process of articles supply and sale is realized with the m runs.

Let's describe the algorithm of the simulation within one run. For doing this let's introduce the following notations: A_i and D_i – are the moments of the supply of the i-th article and of the i-th demand arise:

$$A_i = \sum_{j=1}^{i} X_j^* , \; D_i = \sum_{j=1}^{i} Y_j^* . \qquad (7)$$

Let's define Na and Nd as the next numbers of supply of the articles and the next number of the articles demands, which take place for the present moment. The factual number of articles in the warehouse at present moment is denoted by S for the current run. Let's R define the number of demands, which were rejected for the current run. The modeling algorithm is such for one run.

Initial date: The sufficient statistics $A = \sum_{i=1}^{n_a} X_i$,

$D = \sum_{i=1}^{n_d} Y_i$, δ, k, t.

Output date: R – number of rejected demands at the current run, $C = (Nd - R) \cdot c_d - k \cdot c_a$ – the value of the income.

Algorithm:

Step 1. To generate exponential distributed with parameter 1 random variables $X_1^0, X_2^0, ..., X_{n_a}^0$ and $Y_1^0, Y_2^0, ..., Y_{n_d}^0$.

Step 2. To form the values $\{A_i\}$ and $\{D_i\}$ with respect to formulas (5) – (7), to take $Na = 1$, $Nd = 1$, $R = 0$, $S = k$.

Step 3. If $A_{Na} < D_{Nd}$, then take $Na = Na + 1$, $S = S + 1$; otherwise to take $Nd = Nd + 1$, $S = S - 1$; if $S < 0$, then take $S = 0$, $R = R + 1$.

Step 4. If $t > min\{A_{Na}, D_{Nd}\}$, then go to the step 3, otherwise – the end.

In the end of this run the number of rejected demands R and the value of the income $C = (N(d\text{-}1) - R) \cdot c_d - k \cdot c_a$ are remembered.

Then we repeat steps 1 – 4 r times. According to the gotten results of the runs, the frequencies of different values of rejected demands are calculated and the average income as well. When changing the initial value of the stock k, we can estimate the optimal value of the initial stock $k^*(t)$ with which the average income will be maximum.

NUMERICAL EXAMPLE

In this article a numerical example is considered. Let input data have the following values $n_a = 9$, $n_d = 12$, $A = 10$, $D = 8$, $\delta = 0.5$, $c_a = 2$, $c_d = 3$, $t = 6$. In first we set $k = 3$ and we have to investigate a dependence of average income C as a function of simulated runs (see *Table 1*). From the data of this table we can conclude that the average value of C^* (as an estimator of C) doesn't change significantly in any way starting from the 200 run. So, we can set the number of runs $r = 200$.

Table 1. The mean values of C as function of the number of runs r

r	$r = 10$	$r = 50$	$r = 100$	$r = 150$
C^*	15,10	13,86	14,27	14,16
r	$r = 200$	$r = 250$	$r = 300$	$r = 400$
C^*	14,14	14,13	14,12	14,12

Initially we have $k = 3$ articles in the warehouse. Next we wish to estimate the probability $P(i, 3, 6)$ that during the given time interval (0, 6) i refusals for orders will take place. The frequencies of rejected demands R are considered in the *Table* 2. The frequency polygon of rejected demands R is shown on the *Figure2*.

Table 2. The frequencies of rejected demands R

i	0	1	2	3
$P^*(i,3,6)$	0,15	0,20	0,21	0,19
i	4	5	6	7
$P^*(i,3,6)$	0,15	0,08	0,02	0,01

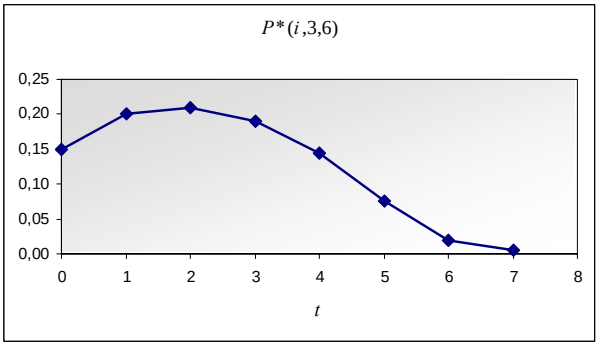

Figure2. The frequency polygon of rejected demands R

Next we wish to determinate the initial stock k, for which the maximum value of C was gotten. The corresponding arithmetical mean $C^*(k)$ (as a function of k) is calculated by formula.

$$C^*(k) = \frac{1}{r}\sum_{i=1}^{r} C^{(i)}(k). \qquad (8)$$

where $C^{(i)}(k)$ is the value of C, that has been gotten in the i-th run.

Table 3. The mean values of C^* as function of the initial stock k

k	0	1	2	3	4
$C^*(k)$	10,67	12,13	13,64	14,23	14,78
k	5	6	7	8	9
$C^*(k)$	14,34	12,44	11,05	9,54	7,57

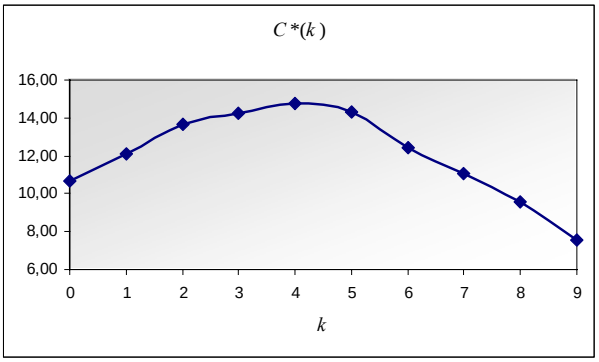

Figure 3. The mean values of C^* as a function of the initial stock k

According to the obtained results we can conclude that $k = 4$ is the optimal value of the initial stock $k^*(t)$ with which the average income will be maximum.

CONCLUSION

Firstly we considered the various approaches to tackle the problem solving: the classical "Plug-In" approach, the "Resampling" approach and the Sufficient Empirical Averaging method. Next we described the *SEA*-method in details. After that we considered some problems of the inventory control and described one method of their solving. Numerical examples end our paper. Gotten results showed that the Sufficient Empirical Averaging method allows solving various practical problems efficiently.

ACKNOWLEDGEMENTS

I'm really grateful to Alexander Andronov for his help in writing this article.

BIBLIOGRAPHY

Afanasyeva, H. 2005. "Resampling-Approach to a Task of Comparision of Two Renewal Processes". In *Proceedings of the 12th International Conference on Analytical and Stochastic Modelling Techniques and Applications*. (Riga, Latvia, June 1 – 4), Khalid Al-Begain, Gunter Bolch, Miklos Telek (Eds.). ASMTA 2005, Riga, 94 – 99.

Andronov, A. and Merkuryev, Yu. 2002. "Use of a Resampling Approach to Systems Simulation." In *Proceedings of the 16th European Simulation Multiconference.* (Darmstadt, Germany, June 3-5). SPS, 150 – 155.

Andronov, A., Zhukovskaya, C. and Chepurin, E.V. 2005. "On Application of the Sufficient Empirical Averaging Method to System Simulation". In *Proceedings of the 12th International Conference on Analytical and Stochastic Modelling Techniques and Applications*. (Riga, Latvia, June 1 – 4), Khalid Al-Begain, Gunter Bolch, Miklos Telek (Eds.). ASMTA 2005, Riga, 144 – 150.

Chepurin, E.V. 1994. "The Statistical Methods in Theory of Reliability*." Obozrenije Prikladnoj i Promishlennoj Matematiki, Ser. Verojatnost i Statistika*, Vol.1*,* N 2, Moscow, 279 – 330. (In Russian.)

Chepurin, E.V. 1995. "The Statistical Analysis of the Gauss Data Based on the Sufficient Empiric Averaging Method." *Proceeding of the Russian University of People's Friendship. Series Applied Mathematics and Informatics*, N 1, Moscow, 112 – 125. (In Russian.)

Chepurin, E.V. 1999. "On Analitic-Computer Methods of Statistical Inferences of Small Size Data Samples." In *Proceedings of the International Conference Probabilistic Analysis of Rare Events*. (Riga, Latvia, June 28 – July 3), V.V. Kalashnikov and A.M. Andronov (Eds.). Riga Aviation University, Riga, 180 – 194.

Cox, D.R., Hinkley, D.V. 1974. *Theoretical Statistics*. Chapman and Hall, London.

Engen, S. and Lillegard, M. (1997). *Stochastic Simulations Conditioned of Sufficient Statistics.* Biometrica, Vol.84, N 1, pp. 235 – 240.

Lehman, E.L. 1983. *Theory of Point Estimation*. John Wiley and Sons, New York.

CATHERINE ZHUKOVSKAYA was born in 1970, in Riga, Latvia. She received the MS degree at the Aviation University of Riga in 1996. She is currently doctoral student at the Institute of Transport Machine Technology at Riga Technical University. Her research interest includes stochastic process, modeling and simulation, statistical optimization. Her e-mail address is: kat_zuk@hotmail.com