SAMPLE ESTIMATION OF DISTRIBUTION PARAMETERS IF UPPER AND LOWER BOUNDS OF RANDOM VARIABLE ARE KNOWN

V.O. Barannik
Kharkiv National University of Municipal Economy, Kharkov, Ukraine
E-mail: v_barannik@ukr.net

The point and interval distribution parameter estimators are obtained by direct numerical approximation of the definition integral with the use of upper and lower bounds of distributed random variable. Like in Bayesian estimation, the distribution parameters are treated as random variables, and their uncertainty is described as a distribution. The Monte Carlo procedure is involved to get the posteriori parameter distributions and the correspondent confidence interval limits.

PACS: 02.50.Ng

INTRODUCTION

Sample methods are widely used for the investigations of population properties from which the samples are drawn in order to get, partially, knowledge about the distribution parameters. Different parametric and nonparametric approaches are used for these purposes depending on pertinent information and the size of the sample that is available. Most parametric statistical methods assume an underlying distribution in the derivation of their results. The consequences of specifying the wrong distribution may prove very costly. If such distribution does not hold, then the confidence levels of the confidence intervals (or of hypotheses tests) may be completely off. Non-parametric or distribution-free methods do not assume an underlying distribution. One of them, the bootstrap was introduced by Efron [1] on the base of sampling generation of statistics by taking repeated replication with replacement from the sample available.

Though bootstrap spread widely in statistical sciences within a couple of decades due to its high practical efficiency [2], from the very beginning Rubin [3], introducing the operationally and inferentially similar Bayesian bootstrap, pointed out one significant drawback of this approach. Strictly speaking, the probability of appearance of any value of continuously distributed random variable is equal to zero, so that there is no reasonable argument to attach the finite probability of appearance to any figure of the sample available. It may be done if these figures represent definite intervals from the set of the random variable distribution. From this standpoint the question, how to relate the sample values with intervals of random variable distribution and correspondent probabilities, deserves special attention.

In any case it seems to be unreasonable to neglect any reliable quantitative information about population density if it exists. In this respect one can say that the main property of real population is that any measurable property x is always confined having the upper and lower bounds. Then, it can be supposed that introducing bounds, if they are known, to statistical enhances would significantly change the properties of distribution parameter assessments. It enables to consider another statistical approach to the distribution parameter estimation that uses Monte Carlo procedure like bootstrap but has different theoretical background.

1. PROBLEM FORMULATION

We consider the random value x having unknown continuously differentiable probability density function (**pdf**) $\rho(x)$ defined on the local set of real line, so that $x_{\min} \le x \le x_{\max}$, where $x_{\min}$ and $x_{\max}$ are known lower and upper set bounds respectively. Then, let $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ be the simple random sample from the continuous population.

It is required to estimate the distribution parameter U, that can be defined as definite integral

$$U = \int_{x_{\min}}^{x_{\max}} u(x)\rho(x)dx, \qquad (1)$$

where $u(x)$ is continuously differentiable generator for the parameter U.

2. PROBLEM ANALYSIS

We introduce the cumulative distribution function (**cdf**) into consideration in a usual way

$$f(x) = \int_{x_{\min}}^{x} \rho(x)dx,$$

so that integral (1) can be presented as following

$$U = \int_0^1 u[x(f)]df, \qquad (2)$$

where $x(f)$ is inverse **cdf**.

We consider the random sample to be ordered from the bottom to the top so that the correspondent value of both generator and **cdf** can be matched to every sample element:

$$x_0 = x_{\min} \le x_1 \le x_2 \le ... \le x_n \le x_{n+1} = x_{\max}; \; u_i = u(x_i);$$
$$f_i = f(x_i); \; i = 0,1,...,n+1;$$
$$f_0 = 0 \le f_1 \le f_2 \le ... \le f_n \le f_{n+1} = 1. \qquad (3)$$

Then integral (2) can be approximated according to the trapezoidal rule:

$$U = \sum_{i=1}^{n+1} a_i(u)\Delta f_i - O\left(\frac{1}{12}\sum_{i=1}^{n+1}\frac{d^2u}{df_{i-1}^2}(\Delta f_i)^3\right), \qquad (4)$$

where $a_i(u) = (u_{i-1} + u_i)/2$ and $\Delta f_i = f_i - f_{i-1}$.

Equation (4) contains set (3) of **cdf** unknown values. At the same time the posteriori **pdf** of these values is known to be independent on $\rho(x)$ and can be presented as:

$$\rho(f_1, f_2, ..., f_n) = n!\prod_{i=1}^{n+1} H(f_i - f_{i-1}), \qquad (5)$$

where $H(\cdot)$ is Heaviside function. It means that every random set of n figures, satisfying condition (3), is equally probable and can be considered to be likely true set. Distribution (5) enables to define different mathematical expectations, for instance:

$$\left\langle(\Delta f_i)^m\right\rangle = n!\int_0^1 df_1\int_{f_1}^1 df_2 ...\int_{f_{n-1}}^1 df_n \left(\Delta f_i\right)^m = \frac{n!m!}{(n+m)!}, \quad (6)$$

where m is positive integer.

Naturally, we introduce the point estimator $\langle U\rangle$ of distribution parameter U as the expected value (4) on distribution (5) that gives

$$\langle U\rangle = \frac{1}{n+1}\sum_{i=1}^{n+1} a_i(u) - O\left(\frac{1}{2(n+1)(n+2)(n+3)}\sum_{i=1}^{n+1}\frac{d^2u}{df_{i-1}^2}\right).(7)$$

From this point and further the errors of numerical approximation (4) and (7) will be ignored, being smaller on order of magnitude. In particular, for the point estimator of distribution mean we have got simple equation

$$\langle X\rangle = \frac{1}{n+1}\sum_{i=1}^{n+1} a_i(x) = \frac{1}{n+1}\left(\frac{x_0+x_{n+1}}{2}+\sum_{i=1}^{n} x_i\right). \quad (8)$$

It means that if there is no sample available $(n=0)$ then the half-sum of the random value bounds can be taken as distribution mean estimation. If random sample is available then half-sum should be added to the sample as independent value. Point estimator (8) is asymptotically unbiased, but if $\rho(x)$ is symmetrical relatively to the centre $(x_{\min}+x_{\max})/2$, then it is simply unbiased. Obviously, point estimator (8) is consistent because if $n\to\infty$ the sum (8) converges to definition integral (2) where $u(x)=x$. The same conclusions are justified for the general point estimator (4).

It should be emphasized that according to (4) and (5) we treat the distribution parameters as random variables (like in Bayesian estimation), and their uncertainty is described as a posteriori distribution. The Monte Carlo method [4, 5] is applied to obtain this distribution. In accordance to the Monte Carlo procedure K set of uniformly distributed on the interval [0,1] random figures: $\tilde{f}_1^{(k)}, \tilde{f}_2^{(k)},...,\tilde{f}_n^{(k)}$; $k=1,2,...,K$, should be generated and ordered from the bottom to the top. On the each ordered set the correspondent likely value of distribution parameter (random estimates) can be calculated as

$$\tilde{U}_k = \sum_{i=1}^{n+1} a_i(u)\Delta f_i^{(k)}, \quad (9)$$

and also be ordered as $U_1 \le U_2 \le ... \le U_K$.

At last, if the degree of confidence P is chosen, the lower B_L and upper B_H limits of correspondent confidence interval are defined accordingly to their places taken up in the ordered set

$$B_L = U_{K(1-P)/2}, \; B_H = U_{K(1+P)/2},$$

and, if the sample size K is sufficiently large, then the point estimation can be calculated in a simple way as

$$\langle U\rangle = \bar{U} = \frac{1}{K}\sum_{k=1}^{K} U_k \, . \quad (10)$$

Besides, different graphic presentations of simulated data, like histogram or **pdf** diagram, can be also applied.

3. DISCRETE DISTRIBUTION

Here we consider the discrete ordered population of size N:

$$x_1, x_2, ..., x_N\,, \quad (11)$$

having bounds $x_0 = x_{\min}$, $x_{N+1} = x_{\max}$, and let $x_{r(1)}, x_{r(2)}, ..., x_{r(n)}$ be the ordered simple random sample of size n drown from the population (11) without replacement; the set $1 \le r(1) < r(2) < ... < r(n) \le N$ being the order numbers of the sample elements in the ordered population.

For the purpose of better compatibility with definition integral (2) we introduce the discrete distribution parameter U as following

$$U = \frac{1}{N+1}\sum_{j=1}^{N+1} a_j(u)\,. \quad (12)$$

Expression (12) is directly related with common definition $U_C = \frac{1}{N}\sum_{j=1}^{N} u_j$ as

$$U = \frac{N}{N+1}U_C + \frac{u_0+u_{N+1}}{2}, \quad (13)$$

providing, on the other hand, for the better convergence to the definition integral if $N\to\infty$.

Then equation (4), after substitution $f_i = r(i)/(N+1)$, can be taken as the distribution parameter estimator

$$U \approx \frac{1}{N+1}\sum_{i=1}^{n+1} a_{r(i)}(u)\Delta r(i), \quad (14)$$

where $\Delta r(i) = r(i) - r(i-1)$; $r(0)=0$; $r(n+1) = N+1$, and $r(i)$ is random positive integer variable distributed on the set $i, i+1, ..., N-n+i$.

The total number of ordered samples of size n without replacement from the population is $\binom{N}{n}$ and number of samples with fixed value of $r(i)$ is $\binom{r(i)-1}{i-1}\binom{N-r(i)}{n-i}$, so that following identity for the binomial coefficients takes place

$$\binom{N}{n} = \sum_{r(i)=i}^{N-n+i}\binom{r(i)-1}{i-1}\binom{N-r(i)}{n-i}, \quad (15)$$

and probability distribution of $r(i)$ can be defined as

$$\mathrm{P}\left[r(i)\right] = \binom{N}{n}^{-1}\binom{r(i)-1}{i-1}\binom{N-r(i)}{n-i}.$$

The mathematical expectation $\langle r(i)\rangle$ can be calculated now as

$$\langle r(i)\rangle = \sum_{r(i)=i}^{N-n+i}\mathrm{P}\left[r(i)\right]r(i) = \frac{N+1}{n+1}i\,. \quad (16)$$

If to define the point estimator for the discrete distribution parameter as mathematical expectation of (14) then it will be the same as (6):

$$\langle U\rangle \approx \frac{1}{n+1}\sum_{i=1}^{n+1}\frac{u_{i+1}+u_i}{2}, \quad (17)$$

where i is index of the element in ordered sample.

The Monte Carlo procedure can be also applied to find both posteriori distribution of the parameter and confidence interval limits. For this purposes K random samples of n positive integers should be drown from the set $1, 2, ..., N$ without replacement, ordered from the bottom to the top to represent $r^{(k)}(i)$, and substituted to (14) providing for the finding of K likely true values of the distribution parameter. This set of simulated data is the basis for the application of different statistical models to represent the properties of parameter distribution.

4. SIMULATIONS

As a final result we have got the clear and sufficiently simple method for the point and interval estimations of distribution parameters on the simple random samples if the random value bounds are known. For the purpose of demonstration of the method potential to treat the samples of small size we consider the following case example. Table 1 contains twenty figures that were generated from an uniform distribution on interval [0,1] representing the sample with "unknown" distribution of random variable having bounds: $x_{\min} = 0$, $x_{\max} = 1$.

Table 1

Random sample from uniform distribution on [0,1]

0.7475	0.3275	0.9443	0.2467
0.6789	0.5683	0.7703	0.0315
0.3239	0.2536	0.748	0.7319
0.2539	0.1412	0.0205	0.2221
0.6789	0.067	0.976	0.6882

The following distribution parameters are estimated: the distribution mean X ($X = 1/2$ is the true value), variance D ($D = 1/12$), the third $CM3$ ($CM3 = 0$) and fourth $CM4$ ($CM4 = 1/80$) central moments. Generator for the central moment of order m is as following

$$u_m(x) = \left(x - \sum_{i=1}^{n+1} a_i(x)\Delta f_i\right)^m . \qquad (18)$$

Three sub-samples are chosen from the Table 1: The sample 1 contains first three figures ($n = 3$) from the first column, the sample 2 contains first ten figures ($n = 10$) from the first and second columns and the sample 3 contains all figures ($n = 20$). According to (8) the point estimation of distribution mean is for the sample 1: $\langle X\rangle = 0.563$; for the sample 2: $\langle X\rangle = 0.413$; and for the sample 3: $\langle X\rangle = 0.472$.

The point estimator for the variance D can be derived from (5) and (18) as

$$\langle D\rangle = \langle x^2\rangle - \langle X\rangle^2 - \frac{1}{12}\left[\frac{1}{n+1}\sum_{i=1}^{n+1} a_i^2(x) - \langle X\rangle^2\right], \qquad (19)$$

$$\langle x^2\rangle = \frac{1}{n+1}\sum_{i=1}^{n+1} a_i\left(x^2\right).$$

The correspondent values of expected variance are for the sample 1: $\langle D\rangle = 0.755$; for the sample 2: $\langle D\rangle = 0.0655$, and for the sample 3: $\langle D\rangle = 0.0955$.

There is no urgent necessity to derive analitical formulas for the point estimators of the cenral moments of the higher orders. They can be calculated much more easier by the use (10) if the Monte Carlo set of simulated values U_k is available and sufficiently large. In order to get the necessary probabilistic entity $K = 10^6$ generations of draws from the distribution (5) were made and the resultant point estimations and confidence interval limits were obtained. These results are recorded in Table 2, demonstrating method efficiency.

Table 2

Point estimates and confidence limits for the central moments ($P = 0.95$)

n	X_L	X_H	D_L	D_H	$\overline{CM3}$
3	0.311	0.722	0.028	0.126	0.008
10	0.271	0.565	0.036	0.104	0.006
20	0.342	0.604	0.067	0.129	0.001
n	$CM3_L$	$CM3_H$	$\overline{CM4}$	$CM4_L$	$CM4_H$
3	-0.027	0.015	0.015	0.0038	0.029
10	-0.0094	0.02	0.01	0.0037	0.02
20	-0.017	0.021	0.016	0.0087	0.025

The table cells for the $\overline{CM3}$ show small values, however its confidence intervals contain zero so that its expected values $\langle CM3\rangle$ can be taken as equal to zero. It is evident that true values of evaluated parameters fall into the correspondent confidence intervals. Much more interesting is the fact that interval estimation precision, measured as the confidence interval width and provided with the sample of extremely low size $(n = 3)$, is comparable to that for the sample of significantly larger size $(n = 20)$.

The large number $K = 10^6$ of statistical trials provides for the smooth shape of two posteriori probability density functions presented at the Fig. 1.

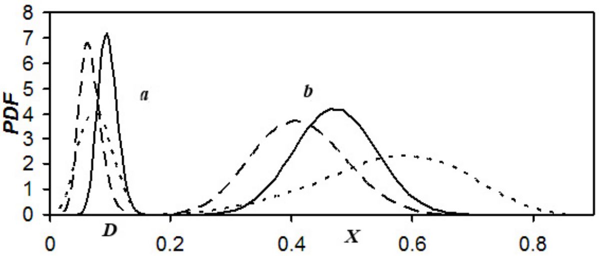

*Fig. 1. Posteriori **PDF**s of the variance D (a) and mean X (b) on the sample of size $n = 3$ (dotted line), $n = 10$ (dashed line), and $n = 20$ (solid line) drown from the continuous population*

To illustrate some properties of the estimator (14) we consider the figures at the Table 1 to be the population, having bounds $x_{\min} = 0$ and $x_{\max} = 1$; the first free figures from the first column being the random sample without replacement. The four common estimates for the distribution parameters of the population are:

$$X_C = \frac{1}{20}\sum_{i=1}^{20}\tilde{x}_i = 0.471,$$

$$\sigma_C = \sqrt{\frac{1}{20}\sum_{i=1}^{20}\left(\tilde{x}_i - X_C\right)^2} = 0.304,$$

$$\gamma_{\tilde{N}} = \frac{1}{20}\sum_{i=1}^{20}\left(\tilde{x}_i - X_C\right)^3 \Big/ \sigma_C^3 = 0.034,$$

$$\beta_C = \frac{1}{20}\sum_{i=1}^{20}\left(\tilde{x}_i - X_C\right)^4 \Big/ \sigma_C^4 = 1.603.$$

In accordance to (12) the correspondent values of above mentioned parameters are:

$$X = 0.472\,,\ \sigma = 0.312\,,\ \gamma = 0.038\,,\ \beta = 1.585\,.$$

These values are somewhat differ from the common estimates but differences will be diminished if the size of population rises.

To get the posteriori distributions of any parameter U we use following approximation of its likely value $\tilde{U}_k$ on every statistical trial $k = 1, 2, ..., K$:

$$\tilde{U}_k = \frac{1}{N+1}\sum\nolimits_{i=1}^{n+1} a_{r(i)}(u)\Delta r^{(k)}(i)\,,$$

where $\Delta r^{(k)}(i) = r^{(k)}(i) - r^{(k)}(i-1)$ and

$$1 \le r^{(k)}(1) < r^{(k)}(2) < ... < r^{(k)}(n) \le N\,,$$

are the ordered set of the positive integers drown from the random variable uniformly distributed on the array $1, 2, ..., N$. Then the likely values of the mean, standard deviation, skewness and kurtosis are calculated as

$$\tilde{X}_k = \frac{1}{N+1}\sum\nolimits_{i=1}^{n+1} a_{r(i)}(x)\Delta r^{(k)}(i)\,,$$

$$\tilde{\sigma}_k = \left\{\frac{1}{N+1}\sum\nolimits_{i=1}^{n+1} a_{r(i)}\left[\left(x - \tilde{X}_k\right)^2\right]\Delta r^{(k)}(i)\right\}^{1/2},$$

$$\tilde{\gamma}_k = \frac{1}{N+1}\sum\nolimits_{i=1}^{n+1} a_{r(i)}\left[\left(x - \tilde{X}_k\right)^3\right]\Delta r^{(k)}(i)\Big/\tilde{\sigma}_k^3\,,$$

$$\tilde{\beta}_k = \frac{1}{N+1}\sum\nolimits_{i=1}^{n+1} a_{r(i)}\left[\left(x - \tilde{X}_k\right)^4\right]\Delta r^{(k)}(i)\Big/\tilde{\sigma}_k^4\,.$$

After $K = 10^6$ statistical trials the confidence ($P = 0.95$) intervals for the mentioned statistics are:

$$0.392 \le X \le 0.602\,;\ 0.282 \le \sigma \le 0.443\,;$$
$$-1.42 \le \gamma \le 0.071;\ 1.11 \le \beta \le 3.36\,.$$

The histograms of the correspondent posteriori distributions for these parameters are presented at the Figs. 2, 3.

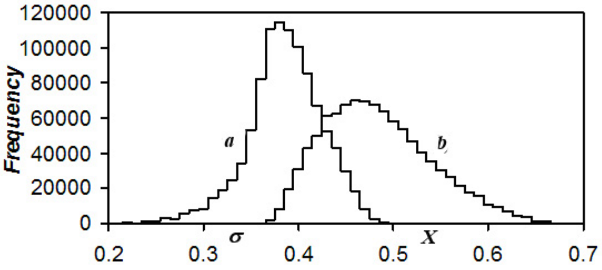

Fig. 2. Histograms of standard deviation σ (a) and mean X (b) posteriori distributions on the sample of size $n = 3$ drown from the discrete population

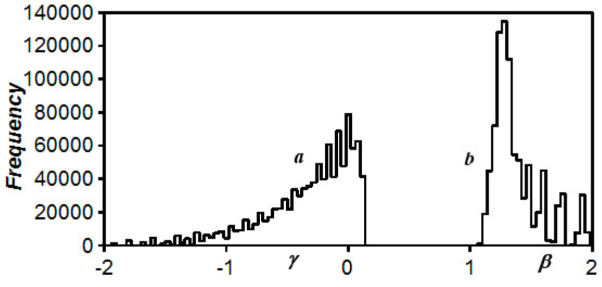

Fig. 3. Histograms of skewness γ (a) and kurtosis β (b) posteriori distributions on the sample of size $n = 3$ drown from the discrete population

The small size of discrete population is also the reason that histograms of skewness and kurtosis show distributions having several modes.

5. DISCUSSION

There are two approaches to statistical assessment of **pdf** parameters**.** In classical estimation these parameters are considered "fixed but unknown" whereas the values of the sample are random. In particular, that means that any element $\tilde{x}_i$ of the simple random sample is random value having determinate statistical weight $1/n$ (probability of appearance). The bootstrap method is the example of efficient interval estimator having that background.

For the Bayesian approach [6] it is assumed that after the sample extraction from the population any sample element $\tilde{x}_i$ is the determinate value and posteriori conditional **pdf** of the estimated parameter $\rho\left(X\middle|\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\right)$ can be defined if the likelihood $\rho\left(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\middle|X\right)$ and priori **pdf** $\rho(X)$ are known. In this paper we also consider any sample element $\tilde{x}_i$ to be the determinate value. But there is no necessity to attract any working hypothesis about distribution of $\tilde{x}_i$. Instead, the upper and lower bounds of random variable should be known to be included into the biased point estimator (7) or (17). Actually, there is also possibility, for instance, to introduce the unbiased point estimator for $\langle X \rangle$ instead (7) if to replace the generator $u(x) = x$ by

$$u(x) = \frac{1}{n+1}\left[(n+1)x - \frac{x_{\min} + x_{\max}}{2}\right],$$

but undesirable consequence of such substitution will be the confidence interval widening, which can be perceptible especially for the samples of extremely small size.

The formal numerical approximation of the definition integral (2) or finite sum (12) enables to realize the Monte Carlo procedure and to get the correct inclusion of the random variable bounds to the every statistical trial (8) or (14) providing for the possibility to find the confidence interval limits for different distribution parameters. Prescribed bounds of the random variable ensure the width of the confidence interval to be as narrow as possible under conditions given. Simulation studies show the remarkable efficiency of the considered method even for sample size as small as 3.

The practical attractiveness of the described approach is stipulated for the circumstance that some measurable properties of the physical, biological and social populations have known bounds. For instance, if the population proportion is estimated then there are obvious bounds 0 and 1 of the random indications. Sometimes the available resources don’t allow to carry out the large-scale sample observations so that only small-size samples can be obtained. Furthermore, the special options could be envisaged in the frames of the sampling plan in order to find appropriate population elements and to estimate the measured random variable bounds. These are just the cases, when the described approximation method could be applied.

REFERENCES

1. B. Efron. Bootstrap methods: another look at the jackknife // *The Annals of Statistics*. 1979, v. 7, № 1, 1979, p. 1-26.
2. B. Efron. The bootstrap and modern statistics // *Journal of the American Association*. 2000, v. 95, № 452, p. 1293-1296.
3. D.B. Rubin. The Bayesian bootstrap // *The Annals of Statistics*. 1981, v. 9, № 1, p. 130-134.
4. N. Metropolis, S. Ulam. The Monte Carlo Method // *Journal of the American Statistical Association*. 1949, v. 44, № 247, p. 335-341.
5. D.P. Kroese, T. Taimre, Z.I. Botev. *Handbook of Monte Carlo Methods*. New York: John Wiley & Sons, 2011.
6. J.A. Bernardo, Adrian F.M. Smith. *Bayesian Theory*. New York: John Wiley & Sons, 1994.

Article received 20.06.2015

ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ ПО ВЫБОРКЕ С ИЗВЕСТНЫМИ ВЕРХНЕЙ И НИЖНЕЙ ГРАНИЦАМИ ИЗМЕНЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

В.А. Баранник

Предлагается способ точечной и интервальной оценки параметров распределения случайной величины с известными границами области ее изменения посредством численной аппроксимации определяющего интеграла. Аналогично методу Байеса параметры распределения интерпретируются как случайные переменные, и их неопределенность выражается в терминах распределений. Для нахождения апостериорного распределения параметра или границ доверительного интервала используется метод Монте-Карло.

ОЦІНКА ПАРАМЕТРІВ РОЗПОДІЛУ ЗА ВИБІРКОЮ З ВІДОМИМИ ВЕРХНЬОЮ ТА НИЖНЬОЮ ГРАНИЦЯМИ ЗМІНЮВАННЯ ВИПАДКОВОЇ ВЕЛИЧИНИ

В.О. Баранник

Пропонується спосіб точкової та інтервальної оцінки параметрів розподілу випадкової величини з відомими границями її змінювання з використанням числової апроксимації визначаючого інтеграла. Аналогічно до методу Байєса параметри розподілу розглядаються як випадкові величини, а їх невизначеність виражається в термінах розподілу. Для отримання апостеріорного розподілу параметра або границь довірчого інтервалу застосовується метод Монте-Карло.