


Maleev E.A., Chepurko V.A.

ROOT ESTIMATION OF DENSITY FUNCTION USING INCOMPLETE DATA

The paper offers two modifications of the nonparametric root estimation of a density function in case of incomplete data in the form of grouped frequencies of failures. The first (integrated) method is connected with respective alteration of a likelihood function. The second (resampling) method of the restoration of failures is based on the iterative restoration of failure time points. The accuracy of the offered estimation methods has been investigated.

***Keywords**: psi-function, square-root method, likelihood function, pseudo-failures.*

The volume of information coming into processing from the facilities of nuclear power stations, as a rule, is limited. One is forced to face information which, alongside with the operating time of failed objects, contains the operating time of objects still working, but supervision over which functioning for the various reasons have been suspended. Besides it is often necessary to deal with the grouped information on failures in which operating time' information of failure objects is lost, and frequency of their occurrence is only known. The information of similar uncertainty is named as censored one.

As is generally known, methods of statistical information analysis are divided into parametric and nonparametric ones. For the analysis of failure data incoming from objects of nuclear power stations, it is more rational to use nonparametric methods which do not require describing a probability distribution by any parametrical law of distribution [1].

The most general characteristic describing the behavior of a one-dimensional random variable is its density function $f(t)$. The problem of estimating the density function of an observable random variable according to the finite number of its realizations at presence of uncertainty is one of the main problems in the statistical analysis, and that identifies the urgency of the present paper.

We know a variety of methods for estimating the density function of full and censored data: histogrammic, projective, nuclear, root estimations. All the methods have both advantages and disadvantages.

The method of histograms is simple in realization; however it is not too visual, and the histogram constructed using small samples, does not allow making correct conclusions. Lack of a projective estimation is in that at boundaries of ca onsidered interval it can take negative values whereas the density is nonnegative by definition. The quality of core estimation strongly depends on the choice of "kernel".

The root estimation represents a square of function expanded on the orthonormal basis and obviously specifies density. The estimation is well studied for full data. This paper considers the root estimation

for data having uncertainty in the time point of realization of the investigated attribute, i.e. for censored data.

For convenience, in the paper the root method of density estimation is divided into two methods: integrated and iterative one.

The integrated method of root estimation is a classical method of root estimation where the required density function $f_\xi(x)$ is found as a square of so called psi-function:

$$f_\xi(x) = |\psi(x)|^2. \tag{1}$$

Let

$$\psi(x) = \sum_{i=1}^{m} c_i \varphi_i(x),$$

where $\{\varphi_i(x)\}$ is an orthonormal system, $\{c_i\}$ are expansion coefficients to be estimated [2, 3]. In what follows it is assumed that the functions $\varphi_i(x)$, $\psi(x)$ and the coefficients c_i are real. The normalization condition $\int f_\xi(x)dx = 1$ causes the equality

$$\sum_{i,j=1}^{m} c_i c_j \int \varphi_i(x)\varphi_j(x)dx = \sum_{i=1}^{m} c_i^2 = 1. \tag{2}$$

Consequently, it is necessary to evaluate the *m-1* of independent coefficients. For the evaluation the maximum likelihood method is used. If the sampling is repeated $\vec{\xi} = (\xi_1, ..., \xi_p)$, then the likelihood function (LF) has the following form:

$$L_n(\vec{c}) = \prod_{k=1}^{p} \hat{f}_\xi(\xi_k) = \prod_{k=1}^{p} \left(\sum_{i=1}^{m} c_i \varphi_i(\xi_k) \right)^2. \tag{3}$$

Log-likelihood function (LLF):

$$l_n(\vec{c}) = \ln L_n(\vec{c}) = \sum_{k=1}^{p} \ln \hat{f}_\xi(\xi_k) = \sum_{k=1}^{p} \ln \left(\sum_{i=1}^{m} c_i \varphi_i(\xi_k) \right)^2 =$$

$$= \sum_{k=1}^{p} \ln \sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j \varphi_i(\xi_k)\varphi_j(\xi_k).$$

Its partial derivatives:

$$\frac{\partial l_n(\vec{c})}{\partial c_i} = \sum_{k=1}^{p} \frac{\partial}{\partial c_i} \ln \hat{f}_\xi(\xi_k) = \sum_{k=1}^{p} \frac{1}{\hat{f}_\xi(\xi_k)} \frac{\partial}{\partial c_i} \hat{f}_\xi(\xi_k) = \sum_{k=1}^{p} \frac{1}{\hat{f}_\xi(\xi_k)} \frac{\partial}{\partial c_i} \left(\sum_{j=1}^{m} c_j \varphi_j(\xi_k) \right)^2 =$$

$$= \sum_{k=1}^{p} \frac{2\varphi_i(\xi_k) \left(\sum_{j=1}^{m} c_j \varphi_j(\xi_k) \right)}{\hat{f}_\xi(\xi_k)} = 2 \sum_{k=1}^{p} \sum_{j=1}^{m} \frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi(\xi_k)} c_j.$$

Optimization problem arises:

$$L_n(\vec{c}) = \prod_{i=k}^{p} \hat{f}_\xi(\xi_k) = \prod_{i=k}^{p} \left(\sum_{i=1}^{m} c_i \varphi_i(\xi_k) \right)^2 \to \max_{\vec{c}}$$

with contingency type of equality: $\sum_{i=1}^{m} c_i^2 = 1$. Factors c_i are selected in such a way that LF was maximal, and their sum of squares is equal to 1.

For finding the maximal value of log-likelihood function $l_n(\vec{c})$ in view of the contingency (2) Lagrange's function is formed:

$$L(\vec{c}) = l_n(\vec{c}) + \lambda \left(1 - \sum_{i=1}^{m} c_i^2 \right).$$

The derivative of Lagrange's function is equated to zero:

$$\frac{\partial}{\partial c_i} L(\vec{c}) = 2 \sum_{k=1}^{p} \sum_{j=1}^{m} \frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi(\xi_k)} c_j - 2\lambda c_i = 0. \quad (4)$$

Multiplying both parts (4) by c_i and summarizing as to i, we shall obtain $\lambda = p$. Then we substitute it in (4):

$$c_i = \frac{1}{p} \sum_{k=1}^{p} \varphi_i(\xi_k) \frac{\sum_{j=1}^{m} c_j \varphi_j(\xi_k)}{\hat{f}_\xi(\xi_k)} = \frac{1}{p} \sum_{k=1}^{p} \varphi_i(\xi_k) \left(\sum_{j=1}^{m} c_j \varphi_j(\xi_k) \right)^{-1}.$$

Then for finding factors iterative numerical methods are used.

In case there are censored data, array elements of intervals $\overrightarrow{LR} = \left[(l_1, r_1); (l_2, r_2); ...; (l_s, r_s) \right]$ act as intervals for estimation description. The size of discontinuity of empirical density function in the beginning of each interval is proportional to the number of sample units (a random number of failures), got in the given interval $\vec{\nu} = (\nu_1, \nu_2, ..., \nu_s)$.

Likelihood function (3) for such data will take the following form:

$$L_n(\vec{c}) = \prod_{k=1}^{p} \hat{f}_\xi(\xi_k) \times \prod_{m=1}^{s} \left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m) \right)^{\nu_m},$$

that is, factor $\prod_{m=1}^{s} \left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m) \right)^{\nu_m}$ will be added to the likelihood function of full data which is responsible for censored data.

Partial derivatives of log-likelihood function will be equal to the following sums:

$$\frac{\partial l_n(\vec{c})}{\partial c_i} = 2\sum_{k=1}^{p}\sum_{j=1}^{m}\frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi(\xi_k)}c_j + \sum_{m=1}^{s}\nu_m\frac{\partial \ln\left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m)\right)}{\partial c_i}.$$

Let us consider individually

$$\frac{\partial \ln\left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m)\right)}{\partial c_i}. \tag{5}$$

If the estimation of density function

$$\hat{f}_\xi(x) = \left(\sum_{i=1}^{m} c_i\varphi_i(x)\right)^2, \tag{6}$$

then it is expedient to take integral as density function estimation:

$$\hat{F}_\xi(x) = \int_{-\infty}^{x}\hat{f}_\xi(u)du. \tag{7}$$

In the further calculations we shall assume, that the density function has the bearer – the segment [0,1]. For density with other bearers of distribution it is possible to make the necessary linear transformation of a random variable mapping set of values of a random variable in the segment [0,1], or it is possible to use other orthonormal bases.

$\varphi_k(x) = \sqrt{2}\sin(k\pi x),\ k = 1,2,...$ is known as Fourier orthonormal bases on the segment [0,1].

We shall find density function (7), using expansion (6).

$$\hat{F}_\xi(x) = \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j \int_0^x \varphi_i(u)\varphi_j(u)du = \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j 2\int_0^x \sin(i\pi u)\sin(j\pi u)du =$$

$$= x - \frac{1}{2\sqrt{2}\pi}\left[\sum_{i=1}^{m} c_i^2\frac{\varphi_{2i}(x)}{i} + \sum_{i=1}^{m}\sum_{j=1,\,j\neq i}^{m} c_i c_j\left(\frac{\varphi_{i+j}(x)}{i+j} - \frac{\varphi_{i-j}(x)}{i-j}\right)\right].$$

Partial derivatives of density function are equal:

$$\frac{\partial \hat{F}_\xi(x)}{\partial c_i} = \frac{1}{\sqrt{2}\pi}\left[2\sum_{j=1,\,j\neq i}^{m} c_j\left(\frac{\varphi_{i-j}(x)}{i-j} - \frac{\varphi_{i+j}(x)}{i+j}\right) - \frac{c_i\varphi_{2i}(x)}{i}\right] =$$

$$= \frac{\sqrt{2}}{\pi}\left[\sum_{j=1,\,j\neq i}^{m} \frac{c_j\varphi_{i-j}(x)}{i-j} - \sum_{j=1}^{m} \frac{c_j\varphi_{i+j}(x)}{i+j}\right].$$

After substitution of the obtained results (partial derivatives) in expression (5) we shall get the following equations.

$$\frac{\partial \ln\left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m)\right)}{\partial c_i} = \frac{\sqrt{2}}{\pi\left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m)\right)}\left[\sum_{j=1,\,j\neq i}^{m} \frac{c_j\left(\varphi_{i-j}(r_m) - \varphi_{i-j}(l_m)\right)}{i-j} -\right.$$

$$\left. - \sum_{j=1}^{m} \frac{c_j\left(\varphi_{i+j}(r_m) - \varphi_{i+j}(l_m)\right)}{i+j}\right].$$

Then

$$\frac{\partial l_n(\vec{c})}{\partial c_i} = 2\sum_{k=1}^{p}\sum_{j=1}^{m} \frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi(\xi_k)} c_j + \sum_{m=1}^{s} \frac{\sqrt{2}\nu_m}{\pi\left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m)\right)} \times$$

$$\times\left[\sum_{j=1,\,j\neq i}^{m} \frac{c_j\left(\varphi_{i-j}(r_m) - \varphi_{i-j}(l_m)\right)}{i-j} - \sum_{j=1}^{m} \frac{c_j\left(\varphi_{i+j}(r_m) - \varphi_{i+j}(l_m)\right)}{i+j}\right].$$

The necessary condition of an extremum, like for full data, is reduced to conditions of equality to zero of partial derivatives of Lagrange's function:

$$\frac{\partial}{\partial c_i} L(\vec{c}) = 2\sum_{k=1}^{p}\sum_{j=1}^{m} \frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi(\xi_k)} c_j - 2\lambda c_i + \sum_{m=1}^{s} \frac{\sqrt{2}\nu_m}{\pi\left(\hat{F}_\xi(l_m) - \hat{F}_\xi(r_m)\right)} \times$$

$$\times\left[\sum_{j=1,\,j\neq i}^{m} \frac{c_j\left(\varphi_{i-j}(r_m) - \varphi_{i-j}(l_m)\right)}{i-j} - \sum_{j=1}^{m} \frac{c_j\left(\varphi_{i+j}(r_m) - \varphi_{i+j}(l_m)\right)}{i+j}\right] = 0. \qquad (8)$$

Now we shall multiply both parts (8) on c_i, sum up on i, and finally obtain the equations for λ:

$$2(p-\lambda)+\sum_{m=1}^{s}\frac{\sqrt{2}\nu_m}{\pi\left(\hat{F}_\xi(r_m)-\hat{F}_\xi(l_m)\right)}\times\sum_{i=1}^{m}c_i\times$$

$$\times\left[\sum_{j=1,j\neq i}^{m}\frac{c_j\left(\varphi_{i-j}(r_m)-\varphi_{i-j}(l_m)\right)}{i-j}-\sum_{j=1}^{m}\frac{c_j\left(\varphi_{i+j}(r_m)-\varphi_{i+j}(l_m)\right)}{i+j}\right]=0.$$

The obtained solution is the following:

$$\lambda=p+\sum_{m=1}^{s}\frac{\nu_m}{\pi\sqrt{2}\left(\hat{F}_\xi(r_m)-\hat{F}_\xi(l_m)\right)}\times\sum_{i=1}^{m}c_i\times\left[\sum_{j=1,j\neq i}^{m}\frac{c_j\left(\varphi_{i-j}(r_m)-\varphi_{i-j}(l_m)\right)}{i-j}-\right.$$

$$\left.-\sum_{j=1}^{m}\frac{c_j\left(\varphi_{i+j}(r_m)-\varphi_{i+j}(l_m)\right)}{i+j}\right].$$

Iterative process for finding expansion factors:

$$c_i^{l+1}=\alpha c_i^l+\frac{1-\alpha}{2\lambda}\left\{2\sum_{k=1}^{p}\sum_{j=1}^{m}\frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi^l(\xi_k)}c_j^l+\sum_{m=1}^{s}\frac{\sqrt{2}\nu_m}{\pi\left(\hat{F}_\xi^{\,l}(r_m)-\hat{F}_\xi^{\,l}(l_m)\right)}\times\right.$$

$$\left.\times\left[\sum_{j=1,j\neq i}^{m}\frac{c_j^{\,l}\left(\varphi_{i-j}(r_m)-\varphi_{i-j}(l_m)\right)}{i-j}-\sum_{j=1}^{m}\frac{c_j^{\,l}\left(\varphi_{i+j}(r_m)-\varphi_{i+j}(l_m)\right)}{i+j}\right]\right\}.$$

Now we shall bring formulas for root estimation on any segment of localization. In arbitrary case (the random variable is distributed on a segment *[a, b]*) the basis will have the following form

$$\varphi_k(x)=\sqrt{\frac{2}{b-a}}\sin\left(k\pi\frac{x-a}{b-a}\right).$$

The first and the last order statistics can be taken as *[a, b]*, i.e. the minimum and maximum value.

In this case the iterative process for finding expansion factors will lead to the following expression (scheme):

$$c_i^{l+1} = \alpha c_i^l + \frac{1-\alpha}{2\lambda}\left\{2\sum_{k=1}^{p}\sum_{j=1}^{m}\frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi^l(\xi_k)}c_j^l + \sum_{m=1}^{s}\frac{\sqrt{2(b-a)}\nu_m}{\pi\left(\hat{F}_\xi^{\,l}(r_m)-\hat{F}_\xi^{\,l}(l_m)\right)}\times\right.$$

$$\left.\times\left[\sum_{j=1,j\neq i}^{m}\frac{c_j^{\,l}\left(\varphi_{i-j}(r_m)-\varphi_{i-j}(l_m)\right)}{i-j} - \sum_{j=1}^{m}\frac{c_j^{\,l}\left(\varphi_{i+j}(r_m)-\varphi_{i+j}(l_m)\right)}{i+j}\right]\right.$$

The next modification of the method for root estimation on censored data is based on a method of restoration of pseudo-failures. In the Western literature this procedure is named as Resampling-method. The principle of modeling of pseudo-observations is based on known property of monotonous function of distribution – the random variable $F_\xi(\xi)$ is evenly distributed in regular intervals distributed on the segment *[0; 1]*. Finding of the value of distribution function is carried out for each source with censored data in points of borders of censoration intervals, then in each of the described intervals of functions some number of points equal to the number of failures on the given interval is modeled. By means of interpolation an inverse mapping of simulated "pseudo-failures" is made onto the axis of operating time (Fig. 1) [4].

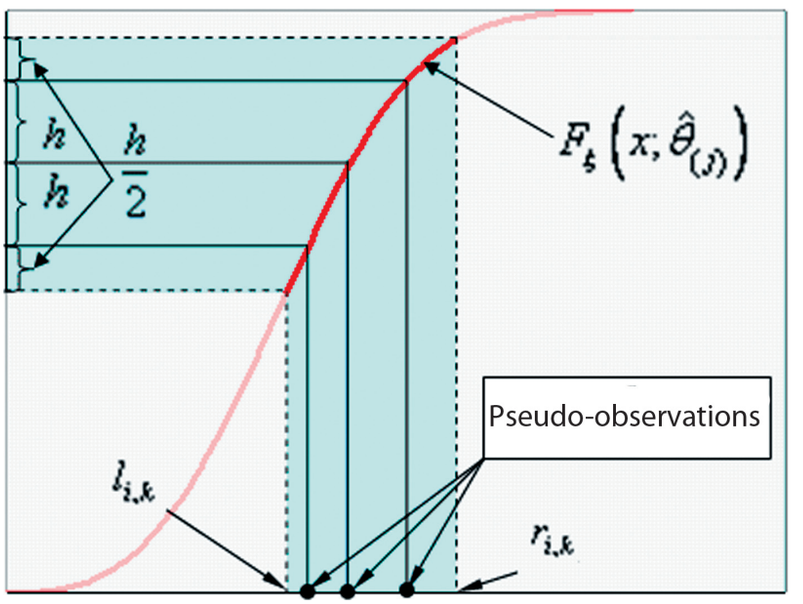

Fig. 1. Modeling of pseudo-observations on the segment of uncertainty

At the following step the restored failures of each source with the grouped information are collected in one array, and then collected data is processed similar to full data processing. They are used to describe the root (integrated) estimation under formula (1). The new distribution function is found out as the integral from this estimation. According to the new function, failures are restored again, and so on until iterative process will not converge.

Failures were restored according to the function F_{cp} representing the average value of empirical distribution functions of sources with censored data and function F_{np}, taken as integral of root estimation of full data density function.

Review of estimations. Estimations of density function were investigated using the example of Weibull's distribution law with parameter of the form $\alpha = 2$ and scale parameter $\lambda = 2$. Modeling of random variables was made by means of the method of inverse functions. Root estimation of density function was applied to the obtained samples, to check equivalence of estimation to the real density. Researches were carried out for five sources of information: one source with full information and four sources with censored data. Simulation scheme of censoration is shown in Fig. 2.

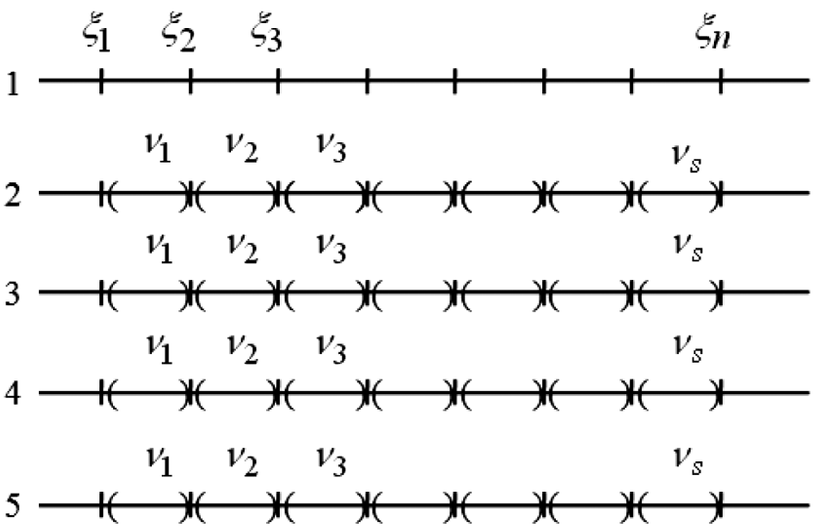

Fig. 2. Simulation of censored data

The accepted number of observations n for each source is 100. The length of censoration interval is set by default as 0,1.

Root estimations of censored information by an integrated method are shown in Fig. 3. Diagrams are presented for the number of harmonics m, equal to 2, 4 and 6.

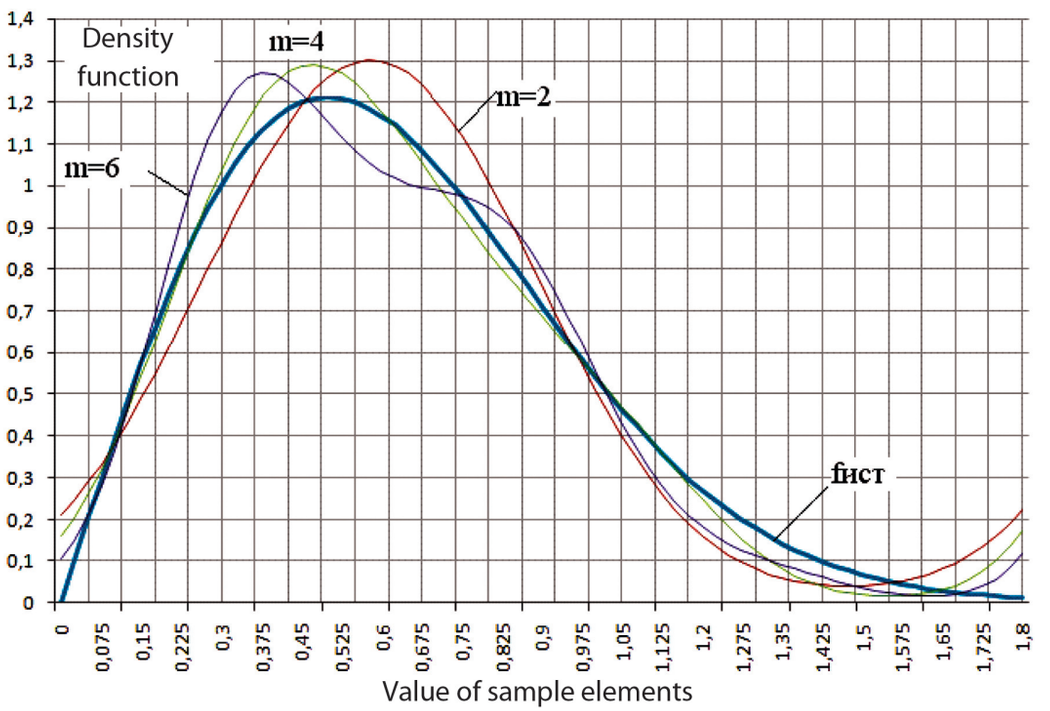

Fig. 3. Root estimation of density function of censored data for different numbers of harmonics

Estimation quality depends on the number of harmonics and the length of an interval of data grouping. During the study it was found out that the less the interval of grouping is, the more accurate is the estimation.

Diagrams of total error of estimation for different number of harmonics are shown in Fig. 4.

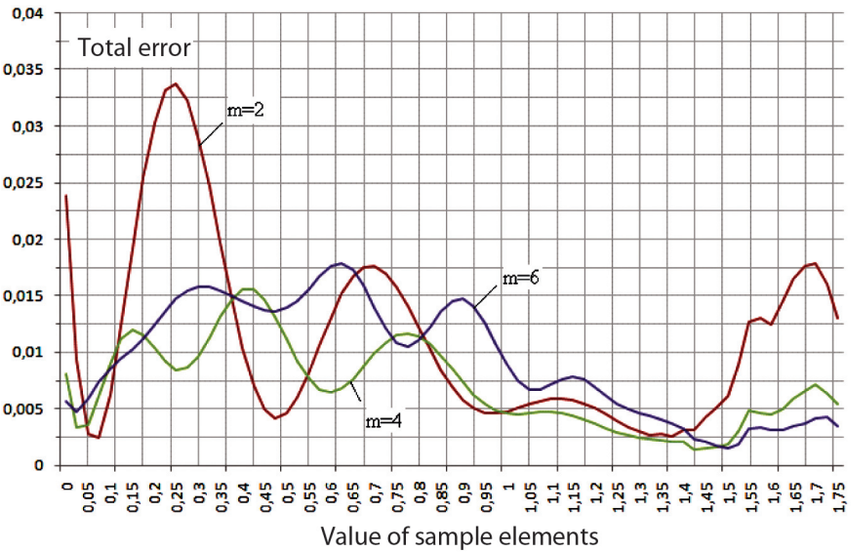

Fig. 4. Total error of root estimation of density function for censored data for different numbers of harmonics

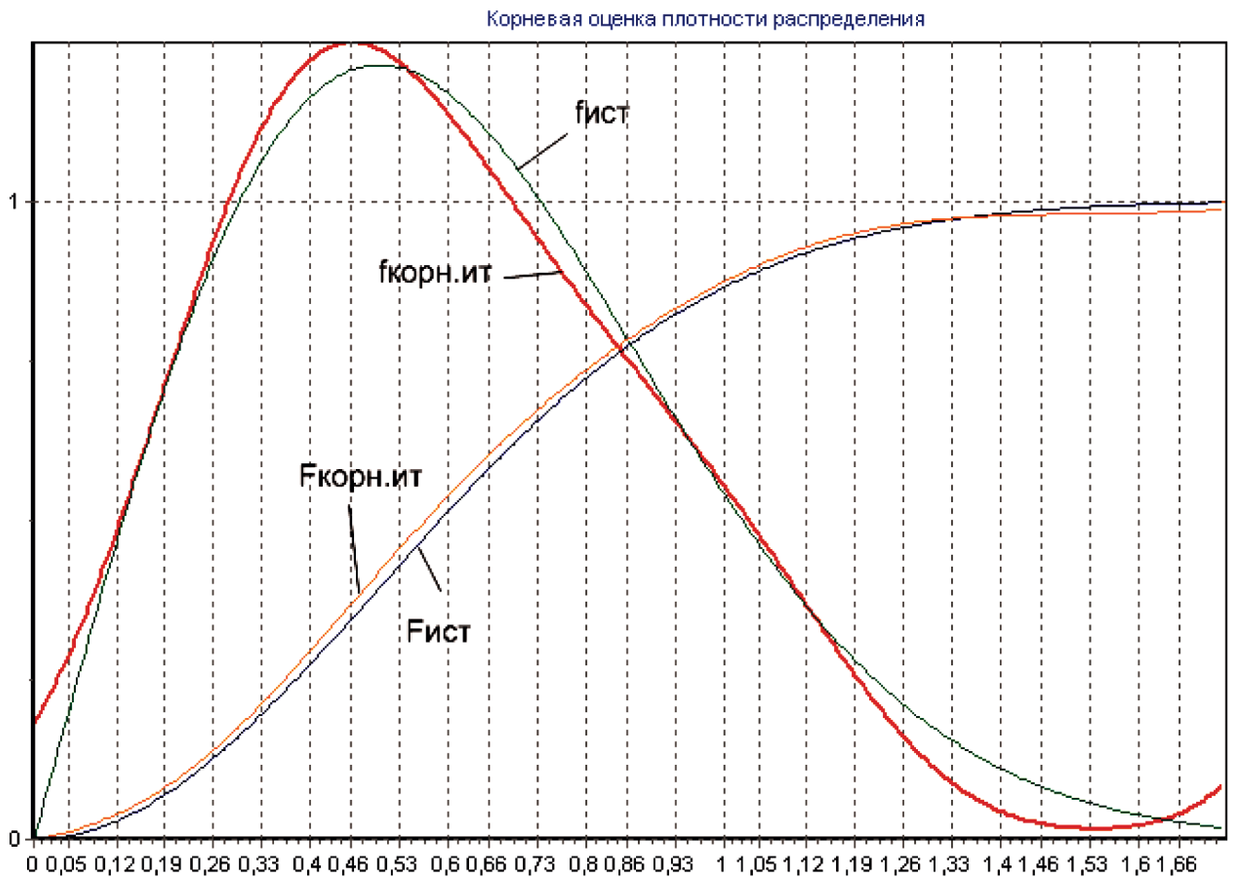

Fig. 5. Root estimation of density by the method of pseudo-refusals, with empirical functions of distribution of censored information taken into account

There is a number of harmonics at which the optimum estimation of density function is achieved. In our case $m = 4$ is the optimum number of harmonics.

The root estimation of density by a method of restoration of failures is presented in Fig. 5, the number of harmonics is $m = 3$.

In the figure *fucm* and *Fucm* are the true values of density and the function of distribution, while *fкорн.um* and *Fкорн.um* are their estimations by the method of restoration of failures.

The length of censoration interval strongly influences estimation quality. The less the length of censoration interval is, the more accurate is the estimation.

Estimation quality by iterative methods is inferior to the quality of the integrated method, however its doubtless advantage consists in the fact that actually it works with full data, and there is no necessity for application of complicated formulas. The method of iterative restoration of failures is more simple and convenient in application.

References

1. **Antonov A.V., Chepurko V.A.** Definition of nonparametric density function on the basis of censored data. Reliability. – M.: The Publishing house "Technology", 2005, №2. – p.3.

2. **Bogdanov U.I.** The primary task of statistical analysis of data: the root approach. – M: MIET, 2002. – 96 p.

3. **Kryanev A.V., Lukin G.V.** Mathematical methods of uncertain data processing. – M.: PHYSMATHLIT, 2003. – 216 p.

4. **Ershov A.N., Chepurko V.A.** Iterative estimation of parameters of the distribution law of a random variable at availability of censored data. Diagnostics and forecasting of complex systems' state: the collection of proceedings № 18 каф. Chair of Automatic Control Systems – Obninsk: Institute for Nuclear Power Engineering, 2009. p. 14-22.