Original article
DOI: 10.21683/1729-2646-2017-17-1-53-58
Dependability, vol. 17 no.1 2017 УДК 519.876.5:37.012.4:004.6

Parametric method of observation results processing with regard to missed data

Dmitri A. Nikilayev, *Obninsk Institute for Nuclear Power Engineering, Obninsk, Russia. e-mail: dafanday@gmail.com*

Dmitri A. Nikilayev

Abstract. *The matters of ensuring dependable and safe operation of NPP facilities is of significant relevance. That is due to the fact that the proportion of equipment at the end of assigned service life in the nuclear power industry is very high, thus dependability analysis of NPP elements and systems is required. In the process of dependability characteristics analysis a number of problems occur, i.e. evaluation of residual life of equipment, justification of life extension decisions. Also, it is required to provide spare parts for elements and systems, select maintenance strategies, etc. That increases the value of activities aimed at analyzing the dependability of nuclear power facilities and, subsequently, the requirement to develop the methods of analysis of statistical information on the operation of NPP elements, subsystems and systems for the purpose of identifying their performance parameters. At nuclear power plants, activities are organized to collect information on the operation of various facilities, i.e. failures and defects of system components, maintenance procedures, operating modes, storage conditions, etc. The information provided by the NPPs has a number of distinctive features. That is due to the following factors: presence of censorship of failure data, absence of sufficient service hours within the given observation interval and the limited volume of available data. All those factors cause an uncertainty in the resulting evaluations and, subsequently, lower that optimal accuracy on dependability characteristics calculation. In the process of evaluating the dependability of facilities in operation a certain part of facilities and systems often does not fail over the period of observation. In such situations statistical analysis of dependability is required that is based on the so-called right censored samples of which the distinctive feature consists in the fact that the inspected product does not fail within the period of observation. In some cases the operation times of specific facilities are unknown. For instance, at the initial stage of facility operation information on its performance was not collected, and the decision to collect data was taken later. In this case the required method must take into consideration the missing information that was not collected at the initial stage. The limited volume of information is due to the fact that the nuclear energy facilities fall into the category of highly dependable equipment. Failures are rare events. Therefore in order to increase the reliability of dependability indicators estimation all the available information must be used. Thus, taking into account all the available information enables more accurate results that can be used to calculate NPP facility service life. The purpose of this article is to show the application of the method of repeated sample and examine its efficiency. The main focus is on missed data that are to be recovered. The authors provide the results of evaluation of the exponential distribution law parameter subject to right censored and missed data. The suggested method of repeated sample is compared with the bootstrap method and mean substitution method. For evaluation of exponential distribution law parameter the authors suggest using the maximum likelihood method. Statistical characteristics calculation is provided. All the calculations and results are based on test cases.*

Keywords: *single substitution method, repeated sample method, bootstrap method, maximum likelihood method, censored data, missed data, data recovery.*

For citation: *Nikolayev D.A. Parametric method of observation results processing with regard to missed data. Dependability, 2017, vol. 17, no. 1, pp. 53-58. (in Russian) DOI: 10.21683/1729-2646-2017-17-1-53-58*

Introduction

The matters of ensuring dependable and safe operation of NPP facilities is of significant relevance. That is due to the fact that the proportion of equipment at the end of assigned service life in the nuclear power industry is very high, thus dependability analysis of NPP elements and systems is required. For that purpose the article looks into the state of the art of statistical analysis of information that includes time to failure, time to censoring and missed data. Taking into account all the available information enables more accurate results.

Thus, the goal is to develop the parametric approach to recovering operation time distribution density based on times to failure, times to censoring and missed data, which would allow identifying a facility's behavior pattern. Consequently, it is required to evaluate the parameter of recovered distribution density and identify the statistical indicators of dependability.

Data description

It should be noted that the object of this research is recoverable facilities of which the operability is to be restored in

case of failure. Before proceeding to the solution of the set goal let us define the types of data to be processed. The first and primary type of data is the time to failure.

In practical situations, particularly during inspections of operating facilities the information submitted to processing is extremely limited. In such cases the need arises to perform statistical analysis of dependability based on the particular samples of which the distinctive feature consists in the absence of information on operating times of the inspected facility. This type of information includes censored data. Censoring is an event that causes the interruption of product observation before the onset of the system event or the onset of the event at an unknown moment of time within a certain interval. We are focusing on cases when this interval may not be limited from the right, i.e. the sample is right censored [1, 2, 5]. The next type of data is the missed data.

Missed data are commonplace in analytical tasks and can significantly affect the conclusions that can be made based on such data.

Possible situations of application of observed information are given in Figure 1.

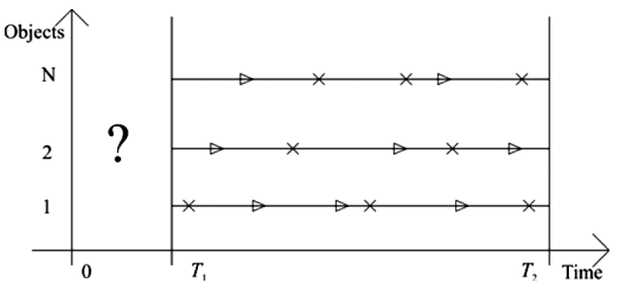

Figure 1. Event chart for data flow

This chart shows what happens to objects over the course of their operation. Data is shown as a continuous flow for a specific object, where × denotes a failure, while □ denotes right censoring. The chart also shows the interval $[0,T_1]$ in which data is not collected, i.e. within this interval observation is not conducted. Within the interval $[T_1,T_2]$ observations were conducted and information on each of the objects was registered. Therefore, the goal is to recover data for the first interval in order to be able to evaluate the parameter of the times-to-failure distribution law within the interval $[0,T_2]$.

Development of the method for missed data management using the repeated sample method

In solving the task of statistical evaluation of dependability indicators of elements and systems of special importance is the matter of collection and presentation of input information on the behavior of analyzed objects. The accuracy and integrity of input information conditions the accuracy of evaluation of the distribution density parameters and the results of dependability characteristics calculation.

As noted above, during statistical analysis operation it is often the case that within certain time intervals information of object behavior is missing. That causes the situation of missed data (Figure 1) which significantly complicates mathematical processing due to the presence of bias in primary statistical characteristics, e.g. mathematical expectation or variance.

Problem definition. Let N objects be under observation (figure 1). For each object, there is a set of data for a certain period of time. Over the time of object operation within the interval from 0 to T_1 information on its behavior was not recorded. Between moments T_1 and T_2 information was collected. Based on the results of observations within the interval $[T_1,T_2]$ for each object time to failure and times to censoring are recorded. The goal is to recover data for the interval $[0,T_1]$ that may constitute either failures or censored data and to evaluate the parameter of the times-to-failure distribution within the interval $[0,T_2]$.

It is suggested to solve the problem of missed data recovery by means of the repeated sample method.

The method consists in the following:

Based on the results of observation for the interval $[T_1,T_2]$ the distribution law parameter is calculated individually for times to failure and time to censoring using the maximum likelihood method.

The number n of operation times is evaluated for the interval $[0,T_1]$ for each type of operation time: $n = mT_1/(T_2-T_1)$, where m is the number of operation times for the interval $[T_1,T_2]$ in case of uniform data flow.

N operation times are modeled according to specified distribution law F(t) (Figure 2) for each type of operation times. I.e. on the probability axis we model the uniformly distributed random number γ_i. Let us perform bijective mapping onto time axis t.

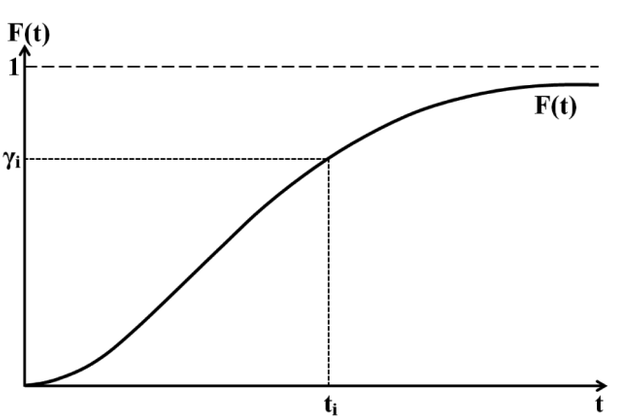

Figure 2. Method for recovery of missed data based on distribution function

Samples obtained within the intervals $[0,T_1]$ and $[T_1,T_2]$ are integrated.

Distribution law parameter of θ_i operation times is evaluated for the interval $[0,T_2]$, as well as mean square deviation σ_i is calculated using the maximum likelihood method.

Then, the distribution law parameter evaluation is calculated individually for times to failure and time to censoring for the interval $[0,T_2]$.

The evaluations obtained at step 6 are used for repeated modeling of operating times at the interval $[0,T_1]$. Steps 3 to 7 are repeated k times. The number of iterations k is defined by the researcher.

Upon completion of step 7 the average distribution law parameter is calculated:

$$\hat{\theta} = \frac{1}{k}\sum_{i=1}^{k}\theta_i \quad (1)$$

as well as the mean square deviation:

$$\hat{\sigma} = \frac{1}{k}\sum_{i=1}^{k}\sigma_i. \quad (2)$$

The obtained values $\hat{\theta}$ and $\hat{\sigma}$ are the result of application of the repeated sample method.

Test case

Step 1: As input information we use the information obtained as the result of modeling of random value in accordance with the exponential distribution law. The total number of data within the interval $[0,T_2]$ is 1000 operation times, out of which 581 are times to failure, 419 are censored data. The number is defined randomly. For data modeling, an exponential distribution was used with the following parameters: $\lambda_t = 0{,}003$ and $\lambda_c = 0{,}002$ for times to failure and times to censoring respectively. Out of the resulting set, data for each experiment (10%, 20%, …, 50%) was removed artificially. Figure 1 outlines the obtained set of data.

Step 2: Using the developed method we obtain the mean estimator of the exponential distribution law parameter $\hat{\lambda}$ and the average mean square deviation $\hat{\sigma}$. Dependability characteristics were calculated based on the method of maximum likelihood. The results are given in Table 1.

Table 1 shows not only the results of operation of the repeated sample method, but also the results obtained using such methods as mean substitution (single substitution) [3], [4] and bootstrapping [9].

The mean substitution method involves replacing missed data with the arithmetic average of sample calculated for the interval $[T_1,T_2]$ instead of each missed value within the interval $[0,T_1]$. Dependability characteristics are calculated using the method of maximum likelihood.

Bootstrapping is a practical computer-based method for researching probability distribution statistics based on repeated sampling by means of the Monte Carlo method based on the available samples. I.e. data for the interval $[T_1,T_2]$ is taken, out of which at each step of n consecutive iterations evenly distributed over the interval [1, n], a random element is retrieved that is then returned in the initial sample (i.e. can be retrieved again). Where n is the number of operation times within the interval $[0,T_1]$. The obtained elements make the data set for the interval $[0,T_1]$. Then, the sample for the interval $[0,T_2]$ is evaluated by means of the maximum likelihood method. The number of iterations for generation of a new sample is defined by the researcher. In our case the research included 1000 iterations. The mean estimator of the exponential distribution law parameter and the average mean square deviation are found using formulas (1) and (2).

The results of both method are given in Table 1.

Table 1. The results of application of the repeated sample method

Percentage of gaps	Gap-fill algorithm	Evaluation of the distribution law parameter for recovered sample, * 10^{-3}	Mean square deviation, * 10^{-3}
10%	Without recovery	2,908	0,110
	With arithmetic average	2,959	0,107
	Bootstrap value	2,909	0,104
	Repeated sample method	2,911	0,105
20%	Without recovery	2,874	0,116
	With arithmetic average	2,917	0,105
	Bootstrap value	2,877	0,103
	Repeated sample method	2,877	0,103
30%	Without recovery	2,877	0,125
	With arithmetic average	2,923	0,106
	Bootstrap value	2,881	0,104
	Repeated sample method	2,878	0,103
40%	Without recovery	2,894	0,136
	With arithmetic average	2,946	0,106
	Bootstrap value	2,895	0,104
	Repeated sample method	2,897	0,104
50%	Without recovery	2,963	0,151
	With arithmetic average	3,015	0,109
	Bootstrap value	2,966	0,106
	Repeated sample method	2,968	0,106

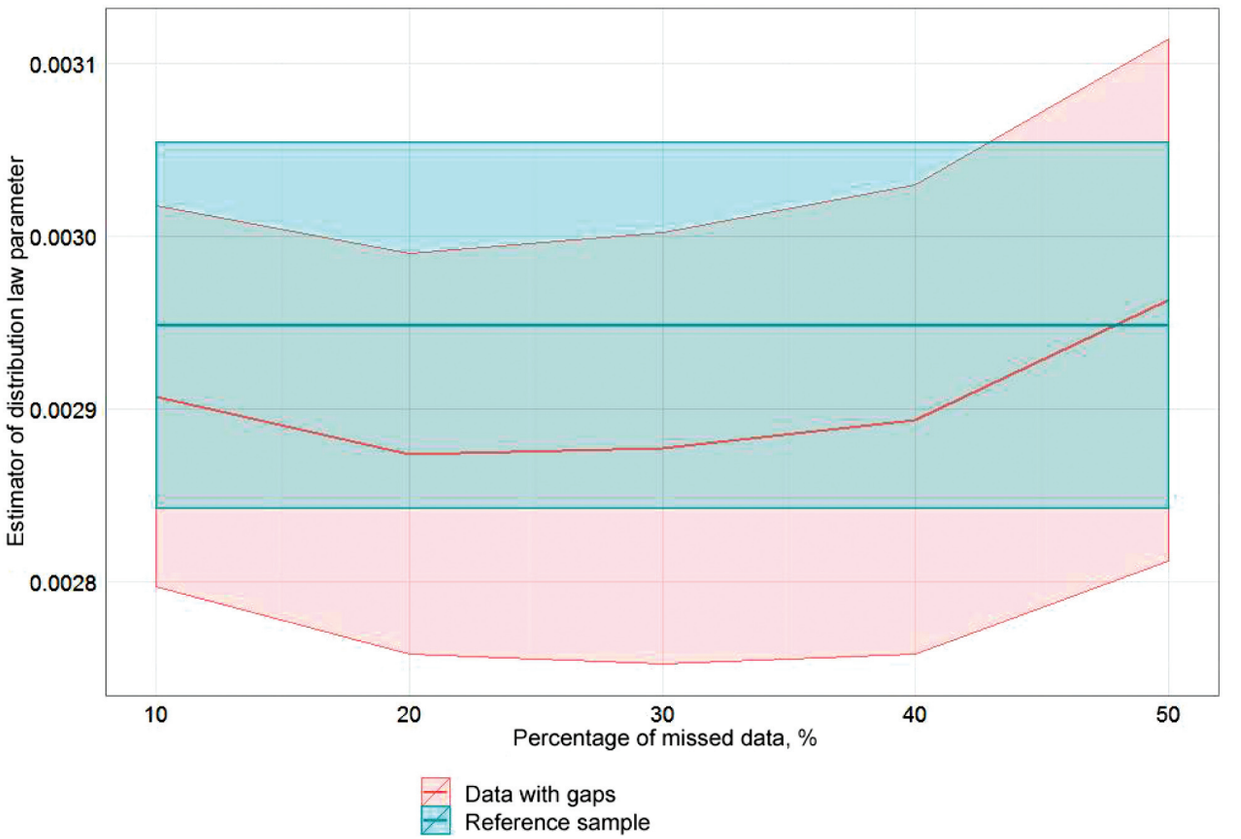

Figure 3. Estimator comparison of reference data and data with gaps

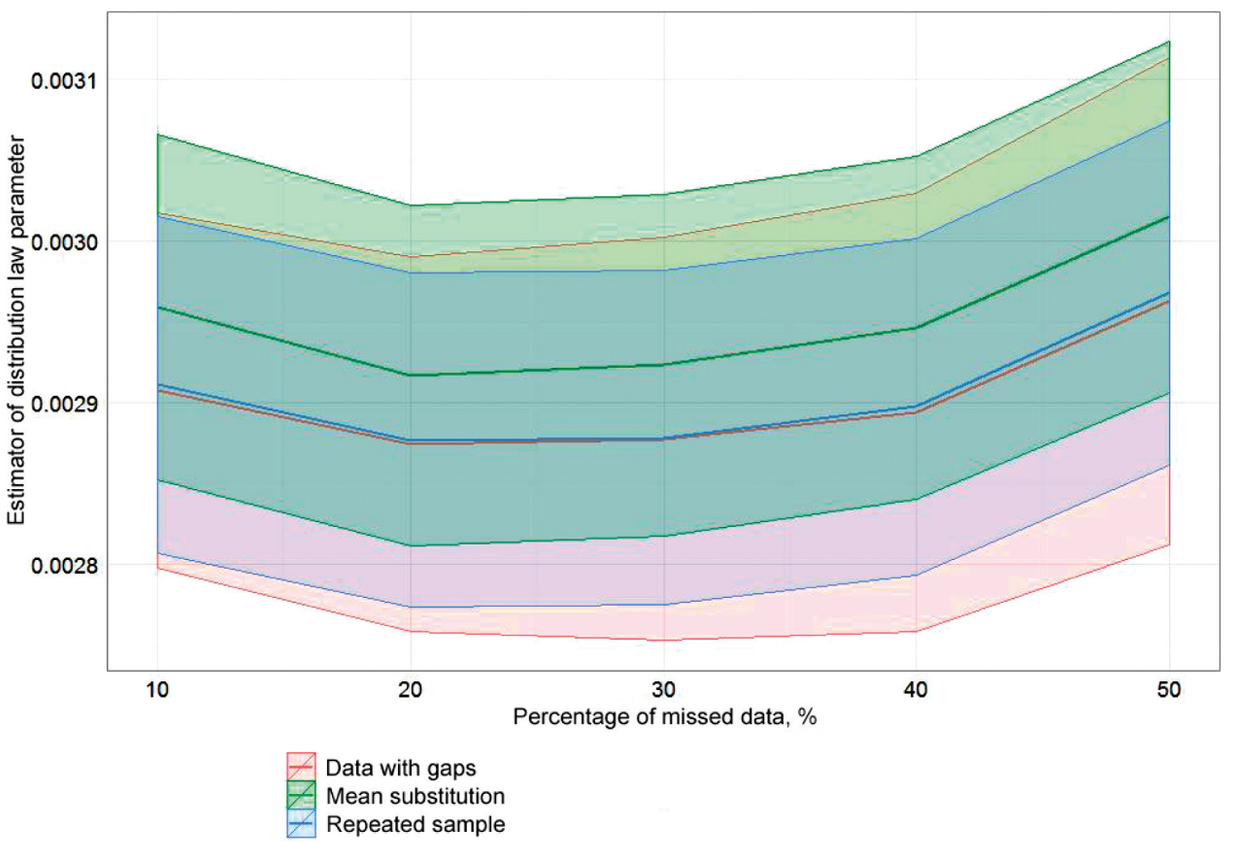

Figure 4. Comparison of estimators obtained by means of the repeated sample method and mean substitution

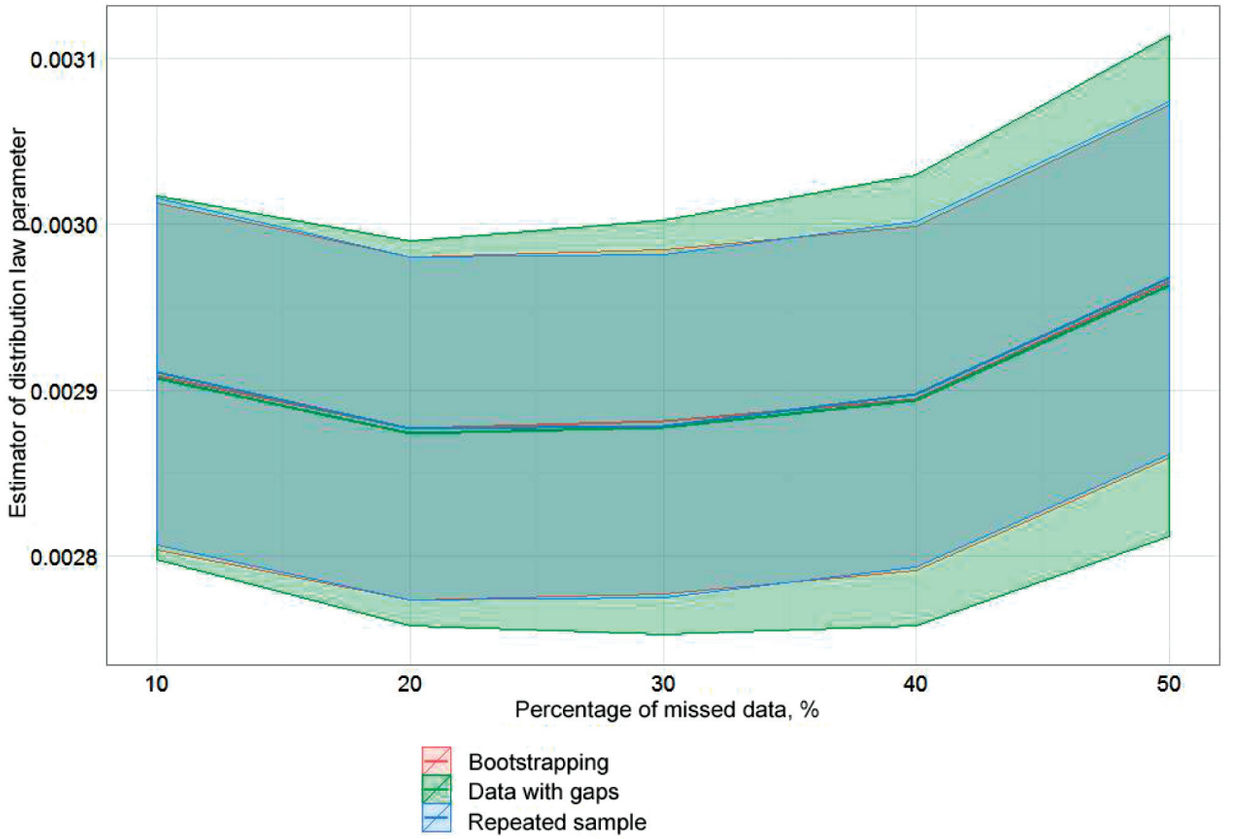

Figure 5. Comparison of estimators obtained by means of the repeated sample method and bootstrapping

In order to simplify the result comparison let us show our data in the following figures (3, 4, 5) and make conclusions for each.

Figure 3 shows the estimator of the distribution law parameter for the reference sample and the estimator for the data flow with gaps (10% ч 50%). As it can be seen in Figure 3 and Table 1, the higher the percentage of missed data the higher the mean square deviation. Therefore, it is required to recover missed data in order to reduce the deviation.

Let us analyze the resulting information in Figure 4. First, let us compare the distribution law parameter estimators. As it can be seen, the estimator obtained by means of the repeated sample method and the estimator of data with gaps match. This suggests that missed data recovery does not affect the estimator. The estimator obtained based on the mean substitution method has a bias. That is a natural consequence of the gaps being substituted with the means, i.e. the values in the middle of the sample. Therefore, the mean square deviation is biased. Let us consider the deviation for the repeated sample method. By using this method in gap recovery we managed to reduce the mean square deviation.

The results obtained by means of the repeated sample method match the results obtained by mans of bootstrapping (Figure 5). From here, two conclusions can be drawn. First, the proposed repeated sample method is not less accurate than the bootstrapping. Second, both methods fall within the class of sample modeling methods. The only difference is that the proposed method is a parametric one, while bootstrapping is a nonparametric one.

Conclusion

The activities described above have yielded the following main results and conclusions.

Various types of data were described. The parametric method for recovering missed data subject to censored information was developed and tested with a test case. Comparison with other methods was made (mean substitution and bootstrapping) and shown the efficiency of the proposed method.

Procedures were developed for recovery of information and evaluation of the exponential distribution law parameter: repeated sample value, mean substitution method, maximum likelihood method. The performed calculations allowed concluding that the proposed repeated sample method is as accurate as bootstrapping. It is also of note that the repeated sample is a parametric method, while bootstrapping is a nonparametric one.

The accuracy of distribution density recovery was researched. The results show that data recovery reduces uncertainty in the calculated indicators (parameters of the exponential distribution law) thereby indicating the requirement to take account of the missed data.

A comparison was made of the results of evaluation of information that was recovered using various methods: re-

peated sample, bootstrapping and mean substitution. It was shown that the mean substitution method causes a bias in the parameter of distribution law. At the same time, repeated sample and bootstrapping produced unbiased results.

References

1. Antonov AV. Sistemny analiz [System analysis]. Moscow: Vyshaya shkola; 2004. Russian.

2. Antonov AV, Nikulin MS. Statisticheskie modeli v teorii nadiozhnosti: Ouchebnoie posobie [Statistical models in the dependability theory: A study guide]. Moscow: Abris; 2012. Russian.

3. Zangiyeva IK. Reshenie problem nepolnykh dannykh massovykh oprosov [Solving the problem of incomplete data of mass survey]. Rossiyskaya sotsiologia zavtrashnego dnia [Russian social science tomorrow]. 2008: 84 – 95. Russian.

4. Zloba E, Iatskiv I. Statisticheskie metody vosstanovleniya propushhennykh dannykh [Statistical methods for recovery of missed data]. Computer Modeling & New Technologies. 2004; 6: 55 – 56. Russian.

5. Cox DR, Oakes D. Analiz dannykh tipa vremeni zhizni [Analysis of survival data]. Moscow: Financy i statistika; 1988. Russian.

6. Little RA, Rubin DB. Statisticheski analiz dannykh s propuskami [Statistical analysis with missed data]. Moscow: Financy i statistika; 1991. Russian.

7. Efron B. Netraditsionnie metody mnogomernogo statisticheskogo analiza [Unconventional methods of multivariate statistical analysis]. Moscow: Financy i statistika; 1988. Russian.

8. Bischl B, Mersmann O, Trautmann H. Resampling methods in model validation. Algorithm Engineering Report. 2010 Aug; 9: 14 – 31.

9. Meeker WQ, Escobar A. Statistical Methods for Reliability Data. New York: John Wiley & Sons, Inc.; 1998.

About the author

Dmitri A. Nikilayev, postgraduate, Obninsk Institute for Nuclear Power Engineering, 1 Studgorodok, 249040 Obninsk, Kaluga Oblast, Russia, e-mail: dafanday@gmail.com

Received on 01.06.2016