Processing of Experimental Data by Means of Nonparametric Statistics

V Syzrantsev a, K Syzrantseva b
Tyumen Industrial University, 38 Volodarskogo Str., Tyumen, 625000, Russia

E-mail a v_syzrantsev@mail.ru, b kv.syzr@gmail.com (corresponding author)

Abstract. The article is dedicated to development of algorithms for recovery of the unknown function of the experimental data distribution density on the basis of nonparametric statistics, i.e. Parzen-Rosenblatt estimates. The results of the algorithms are illustrated by examples of gas pipeline pressures processing and the number of cycles prior to samples failure during their durability tests.

Introduction

When assessing strength reliability of oil and gas equipment elements and facilities, stresses and their limit values obtained during an experiment or calculation in the form of random variable $x_i, i = 1, N$ samples are used. Consequently, determining failure-free operation probability of the facility under study requires knowledge of the functions of the above mentioned random variables distribution density.

In the practice of processing experimental data for solving strength reliability problems it is traditionally accepted [1, 2] that required function $P(X)$ is known up to the finite number of the parameters. In this case, the problem of density recovery is well-posed and its solution uses methods of parametric statistics [2-10]. However, in general case, the class of functions to which $P(X)$ can belong is substantially wider than those discussed in papers [2, 4, 6, 7, 9, 10] and many others.

To date, in the theories of probability and mathematical statistics over a hundred of random variables distribution laws have been studied. At the same time, the experience of processing experimental data on parts loading and durability during machines and mechanisms operation indicates that the distribution of random variables of operating and limit stresses, the number of cycles prior to parts destruction do not follow the standard laws described in the theory of parametric statistics. As examples, Figure 1 shows histograms of pressure in a gas pipeline; figure 2 shows histograms of the number of cycles prior to the samples destruction at different stress levels obtained during the samples durability tests.

To study the processes containing random variables not described by standard distributions nonparametric methods have been developed in the theory of mathematical statistics [11-15]. When processing experimental data in problems of recovery of unknown function of random variables distribution density the methods of nonparametric statistics based on Parzen –Rosenblatt estimates are mostly used [16-27], or the decomposition method with respect to basis functions [28, 29].
Figure 1. Histograms of pressure in different sections of gas pipeline daily fixed per operating year

Figure 2. Histograms of the number of cycles ($N$) prior to the samples destruction at different stress levels ($\sigma_a$) obtained during the samples durability tests
Nonparametric Estimates of the Distribution Density Function

Based on the physical meaning of $X$, we present the only requirement – its continuity and consider the problem of $P(X)$ recovery based on the existing sample $x_i, i = 1, N$ by means of nonparametric statistics [23, 24, 25].

It is known [4, 5, 7, 16, 23, 24], that the probability distribution density is related to the probability distribution function by the ratio:

$$\int_{-\infty}^{\infty} P(x)dx = F(y),$$

presented in the form of:

$$\int_{-\infty}^{\infty} \theta(y - x) \cdot P(x)dx = F(y), \quad \theta(s) = \begin{cases} 1, & \text{if } s \geq 0 \\ 0, & \text{if } s < 0 \end{cases}$$

where $\theta(s)$ is a function of the unit step (Heaviside function).

The optimal nonparametric estimate at each point $y$ for theoretical distribution function $F(y)$ is [23] empirical cumulative probability distribution function $F_N(y)$. If value $y$ exceeds $k$ of the sampling elements $x_i, i = 1, N$ by volume $N$ function $F_N(y)$ looks like:

$$F_N(y) = \frac{1}{N} \sum_{i=1}^{N} \theta \cdot (y - x_i).$$

With the growth of sample volume $N$ function $F_N(y)$ with probability of one uniformly approaches $F(y)$:

$$\Pr\left\{\lim_{N \to \infty} \sup \left| F_N(y) - F(y) \right| = 0\right\} = 1.$$  

In statistical processing the data of experimental stress studies the right part of equation (2) - distribution function $F(y)$ - is replaced by empirical distribution function $F_N(y)$ obtained on the basis of a sample of limited volume stresses. Therefore, the solution of the equation (2) will always be approximate. To recover the distribution density function by solving the equation (2) within the theory of nonparametric statistics, special procedures have been developed [16, 23, 24], providing equation solutions sequence convergence (2) to the required probability density $P(X)$ with increasing $N$ and taking into account the incorrect formulation of the problem (2), associated with the requirement to differentiate the inexact right part of the equation (2). Based on the physical meaning of the distribution density function, it has only positive values on the entire axis of change $X$ from $-\infty$ to $+\infty$ and fulfills the following condition:

$$\int_{-\infty}^{+\infty} P(x)dx = 1.$$  

Under the condition of continuity of function $P(X)$ the solution (2) is the only one.

Algorithms for recovery of unknown distribution density function

The methods of distribution density function estimation proposed by Parzen and Rosenblatt [16] use a smoothed empirical distribution function in the form of:
where \( G(t) \) is a steadily non-decreasing function from 0 to 1 of its argument; in this case \( G(t)=1-G(-t) \), i.e. \( G(t) \) is a symmetric function relative to zero; \( h_N \) is a fuzziness parameter (bandwidth).

After differentiation (6) we have:

\[
P_N(y) = F'_N(y) = \frac{1}{N \cdot h_N} \sum_{i=1}^{N} G\left( \frac{y-x_i}{h_N} \right) = \frac{1}{N \cdot h_N} \sum_{i=1}^{N} K\left( \frac{y-x_i}{h} \right),
\]

(7)

where \( K(t) = G'(t) \) is distribution density \( G(t) \) or kernel function (kernel).

The theoretical studies of the function (6) show that displacement and variation of the estimate (7) depend on kernel type \( K(t) \) and the value of fuzziness parameter \( h_N \). Various dependencies are used as kernel functions [23, 25]. The most common of them, used in processing experiments data by assessing loading and durability of mechanical engineering products, are presented in Table 1.

Recovery of distribution density function by means of the Parzen – Rosenblatt method based on (7) is performed in two stages. At the first stage, kernel function \( K(t) \) is selected from among the known ones (Table 1 or others). At the second stage, the problem of determining an optimal value of fuzzing parameter \( h_N \) is solved.

**Table 1. Kernel functions**

| Normal kernel | Laplace kernel | Fisher kernel |
|---------------|---------------|--------------|
| \( K_1(t) = \frac{1}{\sqrt{2 \cdot \pi}} \exp \left( -\frac{t^2}{2} \right) \) | \( K_2(t) = \frac{1}{2} \exp(-|t|) \) | \( K_3(t) = \frac{1}{2 \cdot \pi} \left( \sin \left( \frac{t}{2} \right) \right) \) |
| Cauchy kernel | Logistics kernel | Epanechnikov kernel |
| \( K_4(t) = \frac{1}{\pi} \frac{1}{1+t^2} \) | \( K_5(t) = \frac{e^{-|t|}}{(1+e^{-|t|})^2} \) | \( K_6(t) = \frac{3}{4\sqrt{5}} \left( 1-\frac{t^2}{5} \right) , |t| \leq \sqrt{5} \) |
| Uniform kernel | Triangular kernel No.1 | Quadratic kernel |
| \( K_7(t) = \frac{1}{2} , |t| \leq 1 \) | \( K_8(t) = 1 - |t| , |t| \leq 1 \) | \( K_9(t) = \frac{3}{4} (1-\frac{t^2}{4}) , |t| \leq 1 \) |
| Parabolic kernel | Triangular kernel No.2 | Trigonometric kernel |
| \( K_{10}(t) = \frac{9}{8} \left( 1-\frac{5}{3}t^2 \right) , |t| \leq 1 \) | \( K_{11}(t) = \frac{1}{\sqrt{6}} - \frac{|t|}{6} , |t| \leq \sqrt{6} \) | \( K_{12}(t) = 0.54 + 0.46 \cos(\pi \cdot t) , |t| \leq 1 \) |

In physical meaning, use of function (7) allows describing a random variable, the discrete values of which are recorded in the course of experimental studies, with continuous function which provides a possibility of its application to solve numerous practical problems of differential and integral methods.

To select the best one from among a finite number of functions \( K(t) \), it is required to have a selection criterion. Information functional of the form [23, 26, 27]:
\[ J = \int \ln[K(t)]P(t)dt = \int \ln[K(t)]dF(t), \] 

(8)

the maximum value of which fulfills condition \( K(t) = P(t) \) can be adopted as such a criterion.

Then the search for optimal \( h_N^* \) and \( K(t) \in K = \{K_1(t),...K_{12}(t)\} \) comes down to solving the following problem:

\[ (h_N^*, K^*(t)) = \arg \max_{h, K(t)} J_N(h_N, K(t)) = \arg \max_{h, K(t)} \left\{ \frac{1}{N} \sum_{j=1}^{N} \ln \left( \frac{1}{(N-1)h_N} \sum_{j=1}^{N-1} K \left( \frac{x_i - x_j}{h_N} \right) \right) \right\} \] 

(9)

Determination of optimal value \( h_N^* \), as shown in papers [23, 29], is a more difficult task than a recovery of distribution density, since optimal value \( h_N^* \) depends on unknown distribution density, especially, its unknown derivatives. In practical applications density estimates are often required in particular areas; for example, in prediction of resource and reliability, first of all, density estimate in the distribution tail area is very important. Therefore, in solving this problem, the algorithms providing determination of optimal values of parameter \( h_N^* \) on the basis of only available sample \( x_i, i = 1, N \) of random variable \( X \) are required.

For some functions \( K(t) \) in Table 1, the dependences for calculating estimates of optimal value \( h_N^* \), based on different estimates of sample \( x_i, i = 1, N \) have been obtained. For example, when using kernel function as normal distribution (Table 1):

\[ K \left( \frac{y - x_i}{h_N} \right) = \frac{1}{\sqrt{2 \cdot \pi} h_N} \exp \left[ -0.5 \left( \frac{y - x_i}{h_N} \right)^2 \right] \] 

(10)

optimal value of parameter \( h_N^* \) («bandwidth»), is defined by formula:

\[ h_N^* = D_N \cdot N^{-0.2}, \] 

(11)

where \( D_N \) is sampling variance, calculated based on the available values sample \( x_i, i = 1, N \):

\[ D_N^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left( x_i - \frac{1}{N} \sum_{i=1}^{N} x_i \right)^2. \] 

(12)

As a result, to estimate the density with kernel (10) and fuzziness parameter (11) based on (7) we have the following expression:

\[ P_N(y) = \frac{1}{\sqrt{2 \cdot \pi} \cdot N \cdot h_N} \sum_{i=1}^{N} \exp \left[ -0.5 \left( \frac{y - x_i}{h_N^*} \right)^2 \right]. \] 

(13)

Analyzing (13), it is easy to see that the implementation of the Parzen – Rosenblatt method involves two steps of calculation. First, rough characterization of the sample, sampling variance (12), is calculated, which is further used to correct the estimation of distribution density (13) via fuzziness parameter \( h_N^* \). Since value \( D_N \) is sensitive to emissions and does not reflect the nature of the density function change (unimodal, multimodal) the information about the distribution density obtained with \( D_N \) may not be sufficient to solve the problem correctly with the method under consideration.
To solve the problem of recovery of distribution density unknown function using (9) on the basis of the kernel functions presented in Table 1, a set of programs in the MathCad system has been developed. The search for function maximum (9) for each kernel function $K(t)$ from Table 1 by value $h_N$ is carried out with the numerical method taking into account the initial value $h_N$ calculated by expression (11).

As an illustration of operation of the created programs Figures 1 and 2 show the results of the recovery of unknown functions of random variables distribution density: the pipeline pressure (Figure 1) and the number of cycles prior to the samples failure during durability tests at fixed stress levels (Figure 2). The distribution density functions recovered according to the developed algorithms are further used to solve the problems of pipeline reliability assessment and to predict the durability of parts the material of which corresponds to the samples material.

This work was supported by grant (project N9.6355.2017/BCh) of government order of Ministry of education and science of Russian Federation for the period 2017–2019 in Tyumen Industrial University.

**Conclusion**

It has been shown that the developed algorithms for solving the problem of recovery of the unknown distribution density function based on the Parzen-Rosenblatt estimates and of finite set of kernel functions ensure efficient processing of the experimental data regardless of the complexity of the distribution law. The examples of determination of distribution density unknown functions for experimental data, the description of which by means of standard laws considered in the theory of parametric statistics is not possible, have been given.

**References**

[1] N.G. Burago, A.B. Zhuravlev, I.S. Nikitin, Models of multiaxial fatigue fracture and service life estimation of structural elements, Mech. of Solids. 46 no.6 (2011) 828-838.

[2] V.P. Kogaev, I.V. Gadolina, Analysis of machine parts in unsteady-state loading conditions, Mashinovedenie 5 (1991) 45-50.

[3] Johnson Norman L., Kotz Samuel & Balakrishnan N., Continuous univariate distributions, Vol. 1,third ed., Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, 1994.

[4] Norman L. Jonson, Fred C.Leon, Statistics and experimental design in Engineering and the Physical Sciences, Volume I Methods of data processing, John Wiley&Sons, New York, London, Sydney, Toronto, 1977.

[5] Norman L. Jonson, Fred C.Leon, Statistics and experimental design in Engineering and the Physical Sciences, Volume II Methods for planning experiments, John Wiley&Sons, New York, London, Sydney, Toronto, 1977.

[6] M.N. Stepnov, A.C. Naumkin, A computational and experimental method for plotting multicycle fatigue curves for structural elements with concentrated stress, Journal of Machinery Manufacture and Reliability. 41 iss. 1 (2012) 34-38.

[7] B.V. Gnedenko, The Theory of Probability, Physics Today. 15 iss. 10 (1962) 62-64.

[8] Z. Zhang, Domains of attraction on countable alphabets, Bernoulli. 24 iss.2 (2018) 873-894.

[9] W. Weibull W, A statistical distribution function of wide applicability, J. Appl. Mech.-Trans. 18 iss.3 (1951) 293–297.

[10] Yu.K. Belyayev, B.V. Gnedenko, I.A. Ushakov, Mathematical problems in queueing and reliability theory, Engineering cybernetics. 21 iss.6 (1983) 62-69.

[11] R.V. Hogg, Adaptive robust procedures: A partial review and some suggestions for future applications and theory, J. of the American Statistical Association. 69 iss.348 (1974) 909-923.

[12] R.S. Singh, A. Ullah, Nonparametric recursive estimation of a multivariate, marginal and conditional DGP with an application to specification of econometric models, Communications in Statistics - Theory and Methods. 15 iss.2 (1986) 3489-3513.

[13] Y. Liu, J. Zhang, X. Zhao, A new nonparametric screening method for ultrahigh-dimensional
survival data, Computational Statistics and Data Analysis. 119 (2018) 74-85.
[14] Baptista de Souza D, Chanussot J, Favre A-C, Borgnat P, A nonparametric test for slowly-varying non stationarities, Signal Processing. 143 (2018) 241-252.
[15] P. Dunn-Rankin, F. Wilcoxon, The true distributions of the range of rank totals in the two-way classification, Psychometrika. 31 iss. 4 (1966) 573-580.
[16] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, AMS. 27 (1956) 832-835.
[17] G.M. Koshkin, Smooth Nonparametric Estimation of the Failure Rate Function and its First Two Derivatives, Russian Physics Journal. 59 iss.6 (2016) 833-844.
[18] V.I. Smagin, G.M. Koshkin, K.S. Kim, Control algorithms for discrete delayed systems with unknown inputs and model parameters using nonparametric technique, Proceedings of the International Conference on Information and Digital Technologies, IDT, 2017, Art.No. 8024321, 350-354.
[19] V.I. Smagin, G.M. Koshkin, K.S. Kim, Control strategies for discrete delayed systems with unknown input using nonparametric algorithms, Proceedings of the International Conference on Information and Digital Technologies, 2016, Art.No. 7557162, 133-137.
[20] V.I. Smagin, G.M. Koshkin, Kalman filtering and control algorithms for systems with unknown disturbances and parameters using nonparametric technique, 20th International Conference on Methods and Models in Automation and Robotics, 2015, Art.No.7283881, 247-251.
[21] A.S. Porshnev, A.S. Koposov, Quantitative estimation peculiarities of densities blur parameter of random sequences distribution in Parzen-Rosenblatt approximation, In the World of Scientific Discoveries, Series B, vol. 2, №1 (2014) 120-129.
[22] A.S. Porshnev, A.S. Koposov, D.A. Bozhilkin, Features of information flows in the backbone Internet-channel: The analysis of the statistical characteristics of the relationship between the number of packets and the time, 9th International Conference on Application of Information and Communication Technologies, 2015, Art.No. 7338596, 437-440.
[23] I.V. Rymar, V.A. Simahin Nonparametric robust estimates of shift and scale parameters, Proceedings of the International Society for Optical Engineering. Volume 6160 I (2006) Art.No 61600.
[24] V.N. Syzrantseva, Ya.P. Nevelev, S.L. Golofast, Calculation of equipment durability based on the methods of the distribution-free statistics, Nauka, Novosibirsk, 2008. (in russian)
[25] K.V. Syzrantseva, Predicting the failure of gear transmissions by nonparametric statistics, Russian Engineering Research. 29 (2009) 1206-1208. DOI: 10.3103/S1068798X09120028.
[26] K. Syzrantseva, V. Syzrantseva, Determination of parameters of endurance limit distribution law of material by the methods of nonparametric statistics and kinetic theory of high-cycle fatigue, Key Engineering Materials. 736 (2017) 52-57. DOI: 10.4028/www.scientific.net/KEM.736.52.
[27] V. Syzrantseva, K. Syzrantseva, Reliability Estimation of Machine Parts with Complicated Geometry on a Base of Methods of Nonparametric Statistics, J. Eng. Appl. Sci. 11 (2016) 204-209.
[28] V. Vapnik, R. Izmailkov, Statistical inference problems and their rigorous solutions: In memory of Alexey Chervonenkis, Lecture Notes in Computer Science. 9047 (2015) 33-71.
[29] V. Vapnik, An overview of statistical learning theory, IEEE Transactions on Neural Networks. 10 iss.5 (1999) 988-999.