METHOD OF PROCESSING EXPERIMENTAL DATA WITH A MULTIMODAL DISTRIBUTION LAW

Goroshko A., Zembytska M.
Khmelnytskyi National University, Ukraine

Abstract. *The article considers the problem of identifying technological processes based on experimental data using statistical methods. It is shown that in cases where the result is influenced by several independent factors, the distribution of product parameters can be multimodal. A model for describing the sample in the form of a linear mixture of normal densities is proposed, which allows to correctly reproduce the behavior of complex technological processes. Methods for estimating mixture parameters are presented, in particular, maximum likelihood, the method of moments and the method of least squares. It is shown that the use of such models increases the accuracy of forecasting and assessing the reliability of products. The results obtained can be used for product quality control and optimization of standards.*

Keywords: *mixture of normal distributions, process identification, statistical modeling, multimodality, quality control.*

The problem of studying distribution laws is often addressed in the identification of technological processes [1], the development of regulatory documentation, quality control of manufactured products [2, 3], product life forecasting, and a number of other quality assurance tasks where the values of controlled quantities are determined by testing prototypes followed by processing the experimental data using mathematical statistics [4].

The essence of the probabilistic method proposed by the authors for processing experimental data subject to multimodal distribution laws is as follows. Let a certain parameter of an object be considered as a random variable X, each sample of whose realizations can be represented as a union of n subsamples. Each subsample is a sample x_i from the general population of realizations of the random variable with probability density $f_i(x, \mu_i, S_i)$, where μ_i i S_i are the mathematical expectation and standard deviation of the i-th subsample.

If the probability that X takes on values belonging to x_i is equal to ρ_i, then for subsequent processing of statistical data, it is proposed to represent the probability density of X as a linear combination of unimodal probability densities f_i, where the weighting coefficients $\rho_i \geq 0$, $i = 1, 2, \ldots, n$ are related by the condition $\sum_{i=1}^{n} \rho_i = 1$. In particular, the desired histogram is represented as a linear combination of Gaussian probability density functions of the form:

$$f\left(x,\mu_1,\mu_2,...,\mu_N;S_1,S_2,...,S_N;\rho_1,\rho_2,...,\rho_N\right)=\sum_{i=1}^{n}\frac{\rho_i}{S_i\sqrt{2\pi}}\exp\left(-\frac{\left(x-\mu_i\right)^2}{2S_i^2}\right). \quad (1)$$

To further process the experimental results, it is necessary to determine the unknown parameters, for example, by applying interpolation on a certain point set, according to which the unknown parameters must be sought based on the condition that the values of function (1) at certain points coincide with the values of the approximating function, the graph of which forms a smooth curve around the constructed histogram. To uniquely determine the $3n$ unknown parameters, the number of points in the set must be no less than $3n$–1. To find the unknowns μ_i, S_i and ρ_i, it is necessary to compose and solve a system of equations of the form:

$$\begin{cases} F(x)=\sum_{i=1}^{n}\frac{\rho_i}{S_i\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{\left(x_j-\mu_i\right)^2}{2S_i^2}\right)dx, \\ j=1,2,...,3n-1,\ \sum_{i=1}^{n}\rho_i=1. \end{cases}, \quad (2)$$

where μ_i, S_i, ρ_i – constant but unknown parameters of the distribution of the i-th subsample and its weighting coefficient.

These parameters can also be found using the least squares method and the method of moments. These estimates should be refined by maximizing the maximum likelihood function.

$$W=\prod_{j=1}^{3n-1}\sum_{i=1}^{n}\frac{\rho_i}{S_i\sqrt{2\pi}}\exp\left(-\frac{\left(x_j-\mu_i\right)^2}{2S_i^2}\right), \quad (3)$$

equating its partial derivatives with respect to the desired parameters to zero. In any case, the problem is reduced to a computer solution of a system of transcendental equations.

Obtaining a probability distribution law for the parameter under study in the form (1) allows us to proceed to solving an important practical problem: assigning an acceptable value for this parameter with a certain reliability. For example, when studying the mechanical strength of resistors, they are tested by applying various types of loads, measuring the values of those loads that lead to failure of the resistor body.

As is well known, the dispersion of the values of the parameter under study depends on the adopted manufacturing method. The boundaries of the

dispersion intervals are determined by the distribution laws of the parameter, which is considered a random variable representing the sum of random variables, each of which is caused by one of the irreducible factors. When the magnitude of each component in the sum described previously is small compared to its magnitude, according to the central limit theorem [5], the distribution of the sum is close to normal. If one or more dominant factors appear among the specified factors, then the corresponding terms prevail in the sum, and the distribution law of the sum becomes multimodal. Further steps to assign an acceptable value for the parameter under study can be carried out in two ways.

1. A subsample with the minimum μ_i value is considered. Clearly, the permissible parameter value for products in this subgroup is minimal, meaning these products are more likely to fail under operating conditions than others. The assigned permissible parameter value for such products can be adopted for the entire batch. In this case, further processing of extreme data can only occur for the specified normally distributed subsample of values with distribution parameters μ_i, S_i. If it is possible to divide the original sample of products into subsamples united by one of the dominant causes of value dispersion, then similar processing operations for the experimental data should be performed for each subsample.

2. Certain parameters allow us to write down the cumulative distribution function (1), which, like a Gaussian random variable, can be specified in a table using a computer as follows. For each value of X, which varies with a certain numerical interval, for example, 0.1, according to the table of the distribution function of the normalized normal distribution $\Phi^x(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{x^2}{2}\right) dx$ it is possible to determine the probabilities

$$\gamma_i = \frac{1}{S_i\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(x-M_i)^2}{2S_i^2}\right) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-M_i)/S_i} \exp\left(-\frac{x^2}{2}\right) dx$$

for all $i = 1,2,...,n$ and further the values of the integral function $F^x(x) = \sum_{i=1}^{n} \rho_i \gamma_i$.

The function $F^x(x)$ will be defined in a table. The resulting table allows one to determine not only the value of the function $F^x(x)$ from the values of x, but also, conversely, to determine the value of the argument from the given values of the function. Thus, for a given confidence probability,

one can determine the desired permissible value of the parameter $[X]$ from a relation of the form

$$\gamma = P\{X < [X]\} = F^x([X]) = \sum_{i=1}^{n} \frac{\rho_i}{S_i\sqrt{2\pi}} \int_{-\infty}^{[X]} \exp\left(-\frac{(x-M_i)^2}{2S_i^2}\right) dx,$$

compiled on the basis of the definition of the integral distribution function of a random variable parameter with distribution density (1).

Firstly, the second method is more accurate, as it takes into account the distribution functions of all subsamples. Secondly, it is more versatile, as it can be used to solve the problem posed in the case of an arbitrary distribution, provided a table of the dependence of the confidence probability and the argument of the cumulative distribution function of the variable being studied is first created. Furthermore, the second method for assigning tolerances for a multimodal parameter distribution extends to both the special case and the unimodal distribution. Moreover, assigning a tolerance using the cumulative distribution function in this special case can even serve as a supplement and refinement to the method for solving a similar problem for a unimodal parameter distribution, described earlier under the assumption that the true value of the measured variable coincides with its mathematical expectation. In order not to require the fulfillment of the last condition and to obtain a value of the permissible parameter value that is smaller and, in this sense, more reliable than that calculated with the specified assumption, one should take the left end of the confidence interval for S as the standard deviation and apply the tolerance assignments using the integral distribution function for a unimodal law with the mathematical expectation and variance obtained in the manner specified earlier.

References

1. Pintelon, R., & Schoukens, J. (2012). System identification: a frequency domain approach. John Wiley & Sons.
2. Montgomery, D. C. (2020). Introduction to statistical quality control. John wiley & sons.
3. Meeker, W. Q., & Hong, Y. (2014). Reliability meets big data: opportunities and challenges. Quality engineering, 26(1), 102–116.
4. McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. Annual review of statistics and its application, 6(1), 355–378.
5. DasGupta, A. (2008). Asymptotic theory of statistics and probability. New York: Springer. Vol. 180, pp. 40–41.