Method for constructing estimates of accuracy of measuring equipment based on Bayesian scientific approach

Rustam Khayrullin 1,*

1 Moscow State University of Civil Engineering, 129337, Moscow, Russian Federation

Abstract. Before putting new unique samples of technical systems into commercial operation, as well as before introducing new technologies into production, as a rule, all kinds of tests are carried out. Small and very small volume of statistical data during testing is a characteristic feature of unique and small-scale products and technical systems. Therefore, the problem of constructing effective statistical estimates with a limited amount of statistical information is an important practical problem. The article proposes the development of the Bayesian approach to the construction of point and interval estimates of the parameters of the known distribution laws. The joint use of a priori and posterior information in the processing of statistical data of a limited volume can significantly increase the reliability of the result. As an example, we consider two most typical distribution laws that arise when testing new unique samples of measuring devices and equipment: normal distribution with an unknown average value and a known dispersion, as well as with an unknown average value and an unknown dispersion. It is shown that for these cases, the parameters of the distribution laws themselves are random variables and obey the normal law and gamma normal law. Recalculation formulas are obtained to refine the parameters of these laws, taking into account a posteriori information. If these formulas are applied several times successively, the process of self-learning of the system or self-tuning of the system occurs. Thus, the proposed scientific approach can find application in the development of intelligent self-learning and self-turning systems.

1 Introduction

The Bayesian scientific approach is widely used to create effective statistical estimates in various fields of activity [1-12]. Radio engineering, classification theory, machine learning, the creation of self-learning and self-tuning systems are just some of the areas where the Bayesian approach is effectively used. This paper describes the application of the Bayesian approach to the problems of constructing effective statistical estimates of accuracy of new and innovation measurement devices and instruments [13-15]. The algorithms for constructing the distribution density function of the average value and the distribution function of the mean square deviation are described. The developed algorithm is based on

* Corresponding author: zrkzrk@list.ru

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).
taking account the available statistical data together with, a priori information about the process or object under investigation. The results of the paper are used in determining the accuracy class of measuring instruments according to the results of acceptance tests, of state tests or tests directed to confirm the type of measuring device and instrument [13-15].

2 Materials and Methods

2.1 The logic scheme of Bayesian scientific approach

Suppose \( \Theta = (\Theta_1, \Theta_2, \ldots, \Theta_s)^T \) - random vector-parameter that involved in the description of the distribution law, \( S \) - dimension of \( \Theta \). It is required to construct the best, in a certain sense, statistical estimation \( \hat{\Theta} \) of this vector of parameters from the available \( k \) - dimensional observations \( \mathbf{X} = (x_{i1}, x_{i2}, \ldots, x_{i(k)}) \), \( i = 1, 2, \ldots, n \). The superscript \( T \), hereinafter, means the operation of transposing a vector. Uppercase letters will denote vector quantities, lowercase letters will be used to denote the one-dimensional (possible or observed) values of the random variables being analyzed, and we will denote matrices (vectors whose vectors are also components).

A priori information is a probability distribution function of the analyzed unknown parameter. It is assumed that this information was obtained prior to the collection of statistical data. As new statistics become available, the distribution function is refined. Under certain assumptions, there is a transition from an a priori distribution to an aposteriori distribution using the Bayes formula [1]

\[
\tilde{\rho}(\Theta|x_1, x_2, \ldots, x_n) = \frac{\rho(\Theta)L(x_1, x_2, \ldots, x_n|\Theta)}{\int \rho(\Theta)L(x_1, x_2, \ldots, x_n|\Theta)d\Theta},
\]

(1)

where the likelihood function \( L(X_1, X_2, \ldots, X_n|\Theta) \) is determined by the relation

\[
L(x_1, x_2, \ldots, x_n|\Theta) = f(x_1|\Theta) \cdot f(x_2|\Theta) \cdot \ldots \cdot f(x_n|\Theta)
\]

(2)

according to the Maximum Likelihood Method (MLM) [1].

The general logical scheme of the Bayesian method for estimating the distribution parameter values is presented in Fig. 1.

---

**Fig. 1.** General logic scheme of the Bayesian approach in statistical estimation.

We describe in accordance with the scheme the main steps of its implementation. A priori information about the parameter \( \Theta \) is based on the history of the functioning of the process under study, as well as on theoretical provisions about its essence and specificity.
This a priori information should be presented in the form of function of density of a priori distribution law $\rho(\Theta)$ of the parameter $\Theta$. Let additional statistics appear in the measurement result

$$X = (x_1, x_2, \ldots, x_n).$$

(3)

In accordance with the law of probability distribution $f(X|\Theta)$. It is assumed that observations (2) with a fixed one are independent. The calculation of the posterior distribution is carried out using a formula (1)

The construction of Bayesian point and interval estimates is based on knowledge of the posterior distribution law $\bar{\rho}(\theta|x_1, x_2, \ldots, x_n)$ that defined jointly by a priori information and empirical data (2). In particular, as the Bayesian point estimates $\hat{\Theta}^{(E)}$ we will use the average value of this distribution [5]

$$\hat{\Theta}^{(E)} = E(\Theta|x_1, x_2, \ldots, x_n) = \int \Theta \bar{\rho}(\Theta|x_1, x_2, \ldots, x_n) d\Theta.$$

(4)

To calculation the Bayesian confidence interval or Bayesian confidence area for a parameter, it takes:

A) In case of one-dimensional distribution law to calculate according formula (1) the function $\bar{\rho}(\theta|x_1, x_2, \ldots, x_n)$ of the posterior distribution law of the parameter $\theta$. Then using the given confidence probability $P_0$, determine $0,5 \cdot (100 + P_0)\%$ - point and $0,5 \cdot (100 - P_0)\%$ - point of this law that give the corresponding boundaries of the desired interval estimate;

B) In case of a multidimensional distribution law to calculate according formula (1) the function $\bar{\rho}(\Theta|x_1, x_2, \ldots, x_n)$ of the posterior law of distribution of the multidimensional parameter $\Theta$, and then, based on a given confidence probability $P_0$, to construct the corresponding $P_0\%$ - level lines, which give the boundaries of the confidence area.

In the practical implementation of the Bayesian approach, a meaningful solution of two main issues (I1, and I2) is necessary:

I1. How to choose a general form (i.e., a parametric family of $\rho(\Theta; D)$ - a priori distribution of the estimated parameter (here $D$ is the vector of parameters)?

I2. How to choose the numerical values $D_0$ of the parameters $D$ that determine the specific form of the a priori distribution $\rho(\Theta; D)$?

Let us consider two typical problems:

Problem 1. The analyzed population has a normal distribution with an unknown mean value and a known dispersion $D_0^2 = \sigma_0^2$. (we will denote hereinafter a similar fact in the form $\xi \in N(\theta, \sigma_0^2)$, where $\xi$ - is the observed random variable:

Problem 2. $\xi \in N(\theta_1; 1/\theta_2)$, where the average value $\theta_1 = E\xi$ and the accuracy parameter $\theta_2 = 1/D_0\xi$ are unknown, that is $\Theta = (\theta_1, \theta_2)$. 


In [1] it was shown that the conjugate law for Problem 1 is the law of normal distribution with an unimportant mean value, and for Problem 2 the conjugate law is the gamma-normal distribution.

Note that the general form of the posterior distribution law calculated by formula (1) is determined, accurate to the normalizing constant, only by the numerator of the right-hand side of this formula. Therefore, below when analyzing equalities that are accurate up to the normalizing constant we will use the sign “≈”.

According the [1] for the Problem 1:

\[ \rho(x|\theta) \approx \exp\left( -\frac{n}{2\sigma^2_0} (\bar{x} - \theta)^2 \right) \]  

The right-hand side of this relation is (up to a normalizing factor independent of \( \theta \)), the density of the normal distribution with average value \( \bar{x} \) and dispersion \( \sigma^2_0 / n \).

Consequently, the family of conjugate a priori distribution laws of an unknown average value of \( \theta \) a normally distributed general population (for a known variance \( \sigma^2_0 = \sigma^2_0 \)) itself belongs to the class of normal distribution laws [1].

Note that for the Problem 2 the right-hand side is (up to a normalizing factor independent of \( \theta \) and \( h \)) the density of two-dimensional gamma-normal distribution law [1]:

\[ \rho(\theta, h) \approx (\lambda_0 h)^{1/2} \cdot \exp\left( -\frac{\lambda_0 h}{2} (\theta - \theta_0)^2 \right) \cdot h^{a-1} \cdot \exp(- \beta h) \]

(6)

with parameters \( \lambda_0 = n, \theta_0 = \bar{x}, \alpha = \frac{n-1}{2}, \beta = \frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \).

Consequently, the set of conjugate a priori distribution laws of a two-dimensional parameter belongs to the class of two-dimensional gamma-normal distribution laws [1].

2.2  Method for calculation specific parameter values in conjugate priory distribution laws

Using as a priori laws the probability distributions associated with the observed general population allows us to determine their general form, i.e., it defines a whole set of a priori distribution laws \( \{ \rho(\Theta; D) \} \). However, when implementing the Bayesian approach, we must operate with a specific a priori distribution law, which requires knowledge of the numerical values \( D_0 \) of the parameters \( D \).

Investigation have shown that in a wide class of problems we can proceed from the fact that we know the a priori mean values of the estimated parameter \( \Theta = (E \theta_1, E \theta_2, ..., E \theta_s)^T \) and their root mean square errors \( \Delta_1 = \sqrt{D\theta_1}, \Delta_2 = \sqrt{D\theta_2}, ..., \Delta_s = \sqrt{D\theta_s} \). Then the parameters of the a priori distribution law can be determined by the method of moments [1].
Since the calculation of parameters for the Problem 1 is obviously, we describe the procedure for calculating parameters only for the Problem 2. From the properties of two-dimensional gamma-normal distribution law it follows that the partial a priori distribution of a parameter $h$ is a gamma distribution law with parameters $\alpha$ and $\beta$. Therefore, using the given values of $h_0 = Eh$ and $\Delta^2_h = Dh$, we compose two equations with respect to $\alpha$ and $\beta$: $Eh = \alpha / \beta = h_0$, $Dh = \alpha / \beta^2 = \Delta^2_h$. Then we obtain the solution: $\alpha = h_0^2 / \Delta^2_h$ and $\beta = h_0 / \Delta^2_h$.

2. The value of the shear parameter in the distribution (6) is equal to $\theta_0$ and

$$\Delta^2_{\theta} = \frac{\beta}{\alpha_0\alpha} \cdot \frac{\alpha}{\alpha - 1}, \quad \lambda_0 = \frac{1}{\Delta^2_{\theta}} \cdot \frac{\alpha}{\alpha - 1}.$$  

2.3 Method for recalculatio of parameter values

Let $\{ \rho(\Theta; D) \}$ be a set of a priori distribution laws associated with the likelihood function $L(x_1, x_2, ..., x_n | \Theta)$ of our observations; $D(d_1, d_2, ..., d_q)$ - the vector of parameters on which the conjugate a priori distribution $\rho(\Theta; D)$ depends; $D_0$ - given (known) values of the parameters $D$ in the analyzed case.

For the Problem 1: $d_1 = \theta_0$, $d_2 = \Delta^2_0$,

$$\tilde{d}_1(x_1, x_2, ..., x_n) = E(\theta | x_1, x_2, ..., x_n) = \left( \frac{1}{\sigma^2_0 / n} \cdot \bar{x} + \frac{1}{\Delta^2_0} \cdot \theta_0 \right) \left( \frac{1}{\sigma^2_0 / n} + \frac{1}{\Delta^2_0} \right)^{-1},$$  \hspace{1cm} (7)

$$\tilde{d}_2(x_1, x_2, ..., x_n) = D(\theta | x_1, x_2, ..., x_n) = \left( \frac{1}{\sigma^2_0 / n} + \frac{1}{\Delta^2_0} \right)^{-1}.$$  \hspace{1cm} (8)

Note that the average value $\tilde{d}_1$ and dispersion $\tilde{d}_2$ of a posterior normal distribution law are the weighted average values of a priori and sample mean and variances, respectively.

For the Problem 2 when implementing the general scheme for converting a priori parameters into a posteriori parameters ones in this case, one should take into account the representation of the likelihood function $L$ in the form (6); a priori density form of two-dimensional gamma-normal distribution (6) too. And the vector of parameters is $D_0 = (\lambda_0; \theta_0; \alpha; \beta)$,

$$\tilde{\lambda}_0 = \lambda_0 + n, \quad \tilde{\theta}_0 = \frac{\lambda_0 \theta_0 + n \bar{x}}{\lambda_0 + n}, \quad \tilde{\alpha} = \alpha + \frac{n}{2},$$
$$\tilde{\beta} = \beta + \frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{1}{2} (\bar{x} - \theta_0)^2 \left( \frac{1}{n} + \frac{1}{\lambda_0} \right).$$  \hspace{1cm} (9)
3 Results

Example 1. Suppose that we have the following information about the analyzed data:

(a) The statistical data is distributed normally with an unknown average value \( \theta \) and known dispersion \( \sigma_0^2 = 0.28 \);

(b) There are measurement results of \( n = 10 \) randomly selected values of the measured parameter \( y \) (see Table 1).

Table 1. Statistical data.

| \( i \) | \( X_i \) |
|---|---|
| 1 | 0.5 |
| 2 | 1.2 |
| 3 | 0.3 |
| 4 | 0.8 |
| 5 | 0.4 |
| 6 | 2.1 |
| 7 | 0.7 |
| 8 | 0.2 |
| 9 | 0.9 |
| 10 | 0.4 |

(c) From history and from results of previous measurements it is known that the a priori values of the mean \( E\theta = \theta_0 = 0.60 \) and dispersion \( D\theta = \Delta_0^2 = 0.03 \). It is required, using the conjugate a priori distribution of the parameter, to obtain Bayesian point and interval (with confidence level \( P_\theta = 0.95 \)) estimates of the average value of a random variable and compare them with the corresponding estimates obtained by MLM.

Example 2. In the conditions of Example 1, it is assumed that the a priori value of the variance \( \sigma_0^2 \) (accuracy parameter \( p = 1/\sigma_0^2 \)) is unknown. It is required, using the conjugate a priori distribution of the parameters \( \theta \) and \( h \), to obtain a Bayesian point estimate and construct the corresponding confidence (with a confidence level \( P_\theta = 0.95 \)) estimates of the average value and standard deviation of the quantity and compare them with the corresponding estimates of the MLM. Also build two-dimensional confidence areas corresponding to confidence levels \( P_{(\theta,k)} = 0.95 \) and \( P_{(\theta,k)} = 0.98 \). Table 2 shows the point estimates and confidence intervals based on the Bayesian approach and the MLM.

Table 2. Comparison of the Bayesian approach and the maximum likelihood method

| | Bayesian approach | | Maximum likelihood method |
|---|---|---|---|
| | Average value | Dispersion | Average value | Dispersion |
| The mathematical expectation of the average value | Confiden ce Interval of the average value | The mathematical expectation of the dispersion | Confiden ce Interval of the dispersions |
| Example 1 | 0.691 | [0.451; 0.931] | 0.0145 | No |
| Example 2 | 0.697 | [0.363; 1.187] | 0.14 | No |

It can be seen that the application of the Bayesian approach allows one to construct more accurate and reliable estimates. Fig. 2 shows a general view of gamma - normal distribution law.
Fig. 2. General view of gamma - normal distribution.

Fig. 3 shows confidence areas (in the plane of parameters $(\theta; h)$ - accuracy and average value) for confidence probabilities $P_{(\theta; h)} = 0.95$ and $P_{(\theta; h)} = 0.975$. It can be seen that the areas are drop-shaped. These confidence areas are level lines of the function of two variables $\rho(\theta; h)$ by the plane $\rho = const$. The values of $const$ are selected in such a way that the volume of the upper cut-off part of the figure that sets the distribution density $\rho(\theta; h)$ is 0.95 and 0.975, respectively. Note that with increasing $n$ confidence areas will become more and more similar to ellipses, since the gamma - normal distribution will tend to a two-dimensional normal law. We also note that currently in the scientific works of other authors the methods for constructing confidence regions which have the shapes of ellipses, rectangles, ellipsoids are described and implemented.

Fig. 3. Confidence area in the plane “average value-accuracy”.

4 Discussion

Modern innovative projects lead to the need to develop new measurement technical means and devices with specified technical, metrological and operational characteristics. The above characteristics of a new creating product are detailed in the relevant technical specifications of the development of a new product. Before the introduction of these technical means and devices or before the state acceptance of these technical means and devices, a whole test cycle is carried out. The goal of testing these devices is to confirm the specified characteristics defined in the technical projects. To achieve this goal the results of testing are carefully analyzed, including processed by statistical methods.
The algorithms and results obtained in the article are aimed at methodological support of the problems described above. The algorithms are based on taking account the available statistical data together with a priori information about the process or object under study.

It is important that the in the paper the obtained distribution function law has an analytical form. We note that in most practical problems it is possible to construct a distribution function only by numerical methods.

The developed in this paper method can be used to create self-learning and self-tuning systems. For this, it is necessary to consistently apply formulas (7) - (9).

5 Conclusions

The use of a priori information about the unknown value of the parameter and the application of the Bayesian approach made it possible to refine the estimates and, in particular, narrow the interval estimate in comparison with MLM from 1.5 to 2 times.

Note that the developed Bayesian scientific approach can yield significant gains in accuracy with limited sample sizes compared to the traditional approach (MLM – method). With an increase in the dimension of additional data or the arrival of a large number of different samples (statistical information) both approaches, due to their consistency, will give more and more similar results.

References

1. S.A. Ayvazyan Applied Econometrics, vol. 1 (9), pp. 93-130 (2008).
2. B.V. Vishnyakov, A.I. Egorov Theory and control systems, vol. 3, pp. 124-132 (2013).
3. D. Kropotov, D. Vetrov Proc. of 7-th Open German-Russian Workshop on Pattern Recognition and Image Understanding (2007).
4. K. Yochan, P. Jinkyun. Reliability Engineering & System Safety, vol. 189, pp. 210-217 (2019).
5. D. Duyguİçen International Journal of Approximate Reasoning, vol. 108, pp. 76-88 (2019).
6. H. Yang, J. Jintao Ke, J. Jieping Ye Transportation Research Part C: Emerging Technologies, vol. 96, pp. 22-37 (2018).
7. M. Ković, C. Butenweg Engineering Structures, vol. 197, 109435 (2019).
8. E. Lavrik, U. Frankenfeld et al., Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, vol. 935, pp. 167-172 (2019).
9. L. Wang, L. Zhao et al., Signal Processing, vol. 166, 107255 (2019).
10. J. Geweke, G. Durham Journal of Econometrics, vol. 210, Iss. 1, pp. 4-25 (2019).
11. D. Michael, A. Thomas, A. Adcock Ocean Engineering, vol. 158, pp. 221-231 (2018).
12. B. Vishnyakov, A. Egorov Theory and control systems, vol. 3, pp. 124-132 (2013).
13. R.Z. Khayrullin, A.A. Safonov, Scientific Review, vol. 19, pp. 167-170 (2017).
14. S. Francisco J. Guzmán et al., Analytica Chimica Acta, vol. 1059, pp. 1-15 (2019).
15. W. Gao H. Haitjema et al., CIRP Annals, vol. 68, Iss. 2 (2019).