BAYESIAN METHODS OF STATISTICAL HYPOTHESIS TESTING FOR SOLVING DIFFERENT PROBLEMS OF HUMAN ACTIVITY

K.J. Kachiashvili*, M. A. Hashmi**, A. Mueed**

Abdul Salam School of Mathematical Sciences of GC University*
68-B, New Muslim Town Lahore, Pakistan
I. Vekua Institute of Applied Mathematics of the Tbilisi State University*
0186 University Street 2, Tbilisi, Georgia
Abdul Salam School of Mathematical Sciences of GC University**
68-B, New Muslim Town Lahore, Pakistan

(Received: 23.02.09; accepted: 10.09.09)

Abstract

The methods of mathematical statistics by their nature are universal in the sense that the same methods can be used for solving the problems of absolutely different nature. The same mathematical methods successfully solve a great diversity of problems from different areas of knowledge. For illustration of this fact, in this work, the formalization of three absolutely different problems from different areas of knowledge is given (air defense, the environment monitoring, sustainable development of production). They show that, despite their absolutely different nature and character at first sight, the formalization reduces to identical mathematical tasks which could be solved by using the same methods of mathematical statistics. For solving of these tasks, unconditional and conditional Bayesian methods of testing of many hypotheses are used, which gives the opportunities of decision-making with certain significance level of criterion.

Key words and phrases: air defence, Bayesian methods, the environment monitoring, hypotheses testing, sustainable development.

AMS subject classification: 62C10.

1 Introduction

The modern level of development of science and engineering has been achieved due to wide application of mathematics for research and designing. As a means of penetration in the essence of the investigated phenomenon and revealing the basic laws, which these phenomena are governed by, mathematics has long been necessary practically in any area of knowledge. Their further advance is impossible without application of mathematical methods and means. The penetration of mathematics into different areas of knowledge was considerably promoted by development of computer science [1]. Due to it, practical resolution of many complex mathematical tasks and their use in man's activity became possible. For example, without wide spread use of modern computers, it would have been impossible to achieve the available level of development in space research, nuclear engineering, communication, mechanical engineering, transport etc., because of complexity of the problems in these fields of knowledge [2]. Going deeper into other fields of science and promoting their development, mathematics as a science experiences favorable influence from their party, as they bring up new problems to it, which requires the development of new approaches, methods and techniques. For the needs of practical tasks, a lot of sections of modern mathematics have emerged, such as the theory of optimum control, the theory of random processes and fields, the theory of decision-making, planning of experiments etc. [2]. Modern mathematics is a complex and varied science consisting of many sections. All mathematics can be divided into two integrally connected parts: determined and stochastic, which are dialectically interconnected as the link from simple to complex, from common to individual. At the present stage, the development of mathematics occurs in the direction of expansion by creation of new sections as well as by detailed elaboration, deepening and improvement of the methods of resolution of the tasks from already generated sections of mathematics. In our opinion, the development in both directions is necessary, requiring urgent application of expert efforts.

The mathematical methods are universal by their nature in the sense that they could be applied to solving the problems of completely different origin. Most of non-experts in this area think that, for solution of his task, it is necessary to use special methods which are different from the methods used for the solution of other tasks. However it happens seldom enough. The same methods of mathematics solve successfully diverse problems from different fields of knowledge. For illustration of this fact, in this work the formalization of three absolutely different problems from different areas of knowledge is given (air defense, the environment monitoring, sustainable development of production), which show that, in spite of their absolutely different nature and character at first sight, the formalization reduces to identical mathematical tasks which can be solved by using the same mathematical methods.

2 Detection and Tracking of Moving Objects on the Basis of Radiolocation Information (RLI)

Let us consider the problems of detection and tracking of a number of objects being in multidimensional space on the basis of radiolocation information (RLI). These problems are mathematically identical. In the first case, the space point which the position of the object corresponds to means the point of physical space where the object is and, in the second case, the parameters of trajectories which the objects follow [3]. For simplicity of exposition, let us suppose that each device measures one co-ordinate of the object.

Let M $(M > 1)$ objects be at the points $\theta^j = (\theta_1^j, \theta_2^j, \ldots, \theta_q^j)$, $q = 1, \ldots, M$ of q-dimensional physical space, $x_i^j, i = 1, \ldots, n$, be the measured value of the θ^j point detected by the i-th measuring device (MD):

$$x_i^j = a_i(\theta^j) + \varepsilon_i, \quad i = 1, \ldots, n;$$

$\varepsilon_1, \ldots, \varepsilon_n$ are the independent errors of measurement, which are normally distributed random values with zero average and σ^2 variance.

It is unknown which measurements have been detected from the same object by different MD, i.e. $a_i(\theta^j) = a_i^j$. Among a_i^j, $j = 1, \ldots, M$; $i = 1, \ldots, n$ which belong to the same objects, there are $n - q$ independent equations of connection [3]:

$$\Psi_k(a_1, a_2, \ldots, a_n) = 0, \quad k = 1, \ldots, n - q, \tag{2.1}$$

a concrete form of which depends on the type and the geometry of disposition of MDs. From (2.1), it is obvious that for the measurement system to be able of defining the location of the objects in q-dimensional physical space, it is necessary to satisfy the condition $q < n$.

As the measurements are realized with errors, equation (2.1), in general case, does not satisfy any sequence $x_1, \ldots, x_n$. The value $\eta_k = \Psi_k(x_1, \ldots, x_n)$ is random which, in accordance with (2.1), can be presented as follows:

$$\eta_k \approx \sum_{i=1}^{n} (x_i - a_i) \frac{\partial \Psi_k}{\partial x_i},$$

where $\frac{\partial \Psi_k}{\partial x_i}$ is the partial derivative.

At normality of measurement errors, η_k obeys normal distribution with zero average $(E(\eta_k) = 0$) and variance

$$V(\eta_k) = \sigma^2 \sum_{i=1}^{n} \left(\frac{\partial \Psi_k}{\partial x_i} \right)^2 .$$

If $t_j = (x_1^{j_1}, \ldots, x_n^{j_n})$ are the measurement values of the same point, then, with high chosen probability P_{t_j}, the following inequality will be satisfied:

$$\left| \frac{\Psi_k(t_j)}{\sqrt{V(\eta_k(t_j))}} \right| \equiv \left| \frac{\Psi_k(x_1^{j_1}, \ldots, x_n^{j_n})}{\sqrt{\sigma^2 \sum\limits_{i=1}^{n} (\frac{\partial \Psi_k}{\partial x_i})^2}} \right| \leq h \left(P_{t_j} \right), \tag{2.2}$$

where $h\left(P_{t_j}\right)$ is some threshold which depends on the chosen certainty P_{t_j} and is defined by ratio $h = \Phi^{-1}\left(\frac{1+P_{t_j}}{2}\right)$. Here $\Phi^{-1}(.)$ is the inverse function of standard normal distribution.

The fulfillment of condition (2.2) for all $k = 1, \dots, n-q$, is the necessary condition (with P_{t_j} probability) of belonging of measurements $x_1^{j_1}, \dots, x_n^{j_n}$ to the same object. If we suppose the distribution of errors being truncate normal, then, for some h_1, conditions (2.2) could be fulfilled with probability one.

Let's designate:

$$G_j = \left\{ t: \ \left| \frac{\Psi_k(t)}{\sqrt{V(\eta_k(t))}} \right| \le h\left(P_{t_j}\right); \ \ \forall k: \ k \in (1, \dots, n-q) \right\},$$

$$j = 1, \dots, N,$$

where G_j are the areas from n-dimensional space of measurement, where the objects could be present; N is the quantity of such areas.

In practice N- the number of areas G_j is always greater than M- the number of detected objects. This is caused by measurement errors and by presence, in the measurement space, of "symmetrical" points in reference MD which give identical measurement information even at the absence of measurement errors.

The problem consists in optimal selection of those areas where the objects from total number of G_j areas are. The optimality of decision rule consists in the minimization of false decisions about the presence of objects in the given areas at restrictions on the number of incorrectly rejected true decisions (conditional task of optimization) or in the minimization of the total error of incorrectly taken decisions of both types (unconditional task of optimization).

Let $H_i, \ i = 1, \dots, S, \ S \le C_N^M$ be the hypothesis supposing the presence of the objects in the areas $G_{i_1}, \dots, G_{i_M}$. It is implied that the same object could not be present in two or more different areas simultaneously (or a stronger restriction $G_i \bigcap G_j = 0$ at $i \ne j \ i, j \in (1, \dots, N)$).

Let us designate: $p(H_i)$ a priori probabilities of hypotheses; x are the measurement results; X is the space of values of x; $p(x|H_i)$ is the probability distribution of x under the condition that hypothesis H_i is true; $\delta(x) = \{\delta_1(x), \dots, \delta_n(x)\}$ is the decision rule, herewith

$$\delta_j(x) = \begin{cases} 1, & \textit{if the decision about the presence of the object in } G_j \textit{ is made}; \\ 0, & \textit{on the contrary}; \end{cases}$$

$\Gamma_j = \{x: \ \delta_j(x) = 1\}$, i.e. Γ_j is the set of such x for which the decision about the presence of the object in G_j is made, $\Gamma_j \subseteq X$. It is obvious that $\delta_j(x)$ is defined by the set of areas Γ_j, i.e. $\delta(x) = \{\Gamma_1, \ldots, \Gamma_N\}$. The problem consists in taking the decision about the validity one of statistical hypotheses H_i, $i = 1, \ldots, S$, on the basis of measurement results x.

3 Formalization of the problem of Identification of River Water Emergency Pollution Sources

The problem of monitoring and controlling the water environment condition includes the problem of identification of pollution sources in order to take measures on their elimination [4]. The latter problem is especially actual for city conditions, when the number of pollution sources is rather great and there is no possibility to control each of them separately. Identification of pollution sources has not only ecological and technological effects, but a significant economical effect as well. The economical effect is reached by minimization of technical facilities, in particular, the measurement facilities, needed for the stand alone control of each pollution source.

Let's consider the problem of identification of the river water pollution sources located between two controlled ranges by means of automated systems [5]. The proposed algorithms are built with the assumption that the pollution sources have either different composition of waste water or (at the same composition) different ratios of ingredients.

Let the river water condition be controlled by M automated stations on the section under consideration. Each of the stations controls $m-$ physical chemical parameters. Let's denote the water quality index at the $j-$th station, i.e. in the $j-t$h controlled range of the river at t_N moment, by $\hat{X}_j(t_N) = \{\hat{x}_{jp}(t_N)\}$, $j = 1, \ldots, M$; $p = 1, \ldots, m$.

The symbol over x indicates that not exact values of the controlled parameters but their estimations are known in the $j-$th range.

Let pollution take place at t_N moment in the $j-$th range, i.e.

$$\hat{X}_j(t_N) = \{\hat{x}_{jp}(t_N)\} \in \overline{\Gamma},$$

where $\overline{\Gamma}_j = R - \Gamma_j$; $R - m$ is dimensional parametric space; Γ_j m is dimensional region of the unpolluted water in the $j-$th range,

$$\Gamma_j = \left\{\hat{x}_{jp}(t) : \mu_{jp}^1 < \hat{x}_{jp}(t) \leq \mu_{jp}^2; \forall p \in (1, ..., m)\right\};$$

$(\mu_{jp}^1, \mu_{jp}^2]$ is region of the unpolluted by parameter p water in the $j-$th range.

In formation of water quality in the $j-$th range $\hat{X}_j(t_N)$ are the following participants: the $(j-1)-$th range, $\hat{X}_{j-1}(t_N - \tau_{j-1})$, where τ_{j-1} is the time for water to run from the $(j-1)-$th range to the j-th one; $K-$ controlled objects with the known concentrations of the substances being released $Z_{j-1,k}(t_N - \tau_k)$, $k = 1, ..., K$, where τ_k is the time for water to run from the $k-$th controlled object to the $j-$th range; $R-$ uncontrolled objects, which in the normal mode of operation release concentrations $Y_{j-1,r}(t_N - \tau_r)$, $r = 1, ..., R$, and in the emergency mode may have additional releases $\Delta Y_{j-1,r}(t_N - \tau_r)$, $r = 1, ..., R$, where τ_r is the time for water to run from the $r-$th uncontrolled object to the $j-$th range.

Other uncontrolled factors are called "noise". Let's denote their influence on the quality of water in the $j-$th range by $\overline{X}_0^j(t) = \{\overline{x}_{0p}^j(t)\}$, $p = 1, ..., m$.

After introducing of denotations the model of water quality formation in the $j-$th range assumes the following form:

$$\hat{X}_j(t_N) = F_j\Big[\hat{X}_{j-1}(t_N - \tau_{j-1}), \lambda_{j-1}; Z_{j-1,k}(t_N - \tau_k), \alpha_{j-1,k}(k = 1, \ldots, K);$$

$$Y_{j-1,r}(t_N - \tau_r), \beta_{j-1,r}(r = 1, \ldots, R)\Big] + \overline{X}_0^j(t_N), \tag{3.1}$$

where F_j is the known operator corresponding to the process of formation of water quality in the $j-th$ range; τ, λ, α, β are parameters characterizing the time of running to the $j-th$ range and the peculiarities of formation of water quality in it.

If there is a pollution, i.e. when $\hat{X}_0^j(t_N) \in \overline{\Gamma}_j$, the model of water quality formation takes the following form

$$\hat{X}_j(t_N) = F_j\Big[\hat{X}_{j-1}(t_N - \tau_{j-1}), \lambda_{j-1}; Z_{j-1,k}(t_N - \tau_k), \alpha_{j-1,k}$$

$$(k = 1, \ldots, K); Y_{j-1,r_1}(t_N - \tau_{r_1}), \beta_{j-1,r_1}(r_1 \in R'); Y_{j-1,r_2} +$$

$$\Delta Y_{j-1,r_2}, \beta_{j-1,r_2}(r_2 \in R'')\Big] + \overline{X}_0^j(t_N), \tag{3.2}$$

where $R' \cup R'' = R$, $R' \cap R'' = 0$, division of set R into subsets R' and R'' being unknown.

The task consists in dividing the set R into subsets R' and R'' at the moment of pollution detection. Upon detection of pollution in the $j-$th range by means of operator F_j from (1.1), the concentrations in the $j-$th range are determined with the assumption that the emergency release was made by one or two, etc., or uncontrolled objects, i.e. are calculated $m-$dimensional points $X_j^{i_1,\ldots,i_r}(t_N)$, where $i_j \in (1, ..., R)$; $i_{j_1} \neq i_{j_2}$; r indicates the number of uncontrolled objects, which are suspected in the simultaneous emergency release. The number of points $X_j^{i_1,\ldots,i_r}(t_N)$ for each population r out of R objects is equal to C_R^r, and the total number of all the points is $\sum\limits_{\gamma=1}^{R} C_R^\gamma = 2^R - 1$. It is necessary to decide which population r of uncontrolled objects made the emergency release, i.e. it is necessary to test hypotheses

$$H_i : M\left(\hat{X}_j\left(t_N\right)\right) = X_j^{i_1,\dots,i_r}\left(t_N\right), \ i = 1, 2, ..., 2^R - 1. \tag{3.3}$$

4 Formalization of the problem of sustainable development of production

Let the technological process be characterized by parameters $b = (b_1, \dots, b_m)$. Depending on the values of these parameters, as a result of realization of technological process, the specified quality (including the quantity) of production which is characterized by the values of corresponding parameters $a = (a_1, \dots, a_n)$ is obtained. As a rule, $n \neq m$, and these parameters, as a matter of fact, differ from each other. Between the parameters of technological process b and the quality of production a, there are functional relations

$$a_i = f_i\left(b_{i_1}, b_{i_2}, \dots, b_{i_{m_i}}; c_{i_1}^i, c_{i_2}^i, \dots, c_{i_{k_i}}^i\right), \tag{4.1}$$
$$1 \le m_i \le m, \ \ i = 1, \dots, n;$$

where m_i is the number of parameters of the technological process which the value of the parameter a_i of the index of production quality depends on; $c_i = (c_1^i, c_2^i, \dots, c_{k_i}^i)$ are the parameters of this dependence; k_i is their number.

Dependencies (4.1) define the values of indicators of the finished product quality depending on the values of technological process parameters.

In a real situation, as a rule, dependencies (4.1) are regression dependencies at a passive or an active experiment, i.e. generally, in real situation instead of (4.1) there are the dependencies:

$$\begin{gathered} a_i = f_i\left(b_{i_1} + \delta_{i_1}, \cdots, b_{i_{m_i}} + \delta_{i_{m_i}}; c_{i_1}^i, c_{i_2}^i, \dots, c_{i_{k_i}}^i\right) + \varepsilon_i, \\ 1 \le m_i \le m, \qquad i = 1, \dots, n, \end{gathered} \tag{4.2}$$

where $\varepsilon_i, \delta_{i_j}$ are the random variables with certain probability characteristics. As a rule, the normal approximation of these distributions is acceptable. The dependence or the independence among them is possible.

The problem of identification of dependence (4.2) at different values of characteristics is very important and widely discussed in the literature [6-9]. This problem has also been considered in the work of the author [10]. Below, dependencies (4.1) are supposed to be given.

By controllable parameters, the production can have one of the set S qualities. Each state of quality is defined by belonging of controllable parameters $a = (a_1, \ldots, a_n)$ of the finished product to corresponding areas $A^i, i = 1, \ldots, S,$ from parametrical space $\mathbb{R}^{\Bbbk}$. As a rule, the i-th quality of production is defined by fulfillment of the condition:

$$A^i = \left\{a : a_{ij}^{'} \leq a_j \leq a_{ij}^{''}; \forall j : \ j \in (1, \ldots, n)\right\}, i = 1, \ldots, S, \tag{4.3}$$

where S is the total number of possible states of production quality which the finished product can have.

To each state of production quality, there corresponds the certain area in parametrical space of technological process which is defined by relations:

$$b_j = \phi_j \left(a_{j_1}, a_{j_2}, \ldots, a_{jQ_j}; d_{j_1}, d_{j_2}, \ldots, d_{jR_j}\right), j = 1, \ldots, m, \tag{4.4}$$

where Q_j is the number of indices of production quality which the parameter of technological process b_j influences; $d_j = \left(d_{j_1}, d_{j_2}, \ldots, d_{jR_j}\right)$ are the dependencies parameters; R_j is the number of these parameters.

The kind of functional dependence ϕ_j and its parameters d_j can be determined by solution of the system of equations (4.1) in relation to parameters b_j if such a solution exists, or they can be obtained by identification of this dependence on the basis of experimental data [10].

Thus, to each area A^i from the parametrical space of production quality, area B^i in the parametrical space of technological process corresponds. Functional dependencies f_i, $i = 1, \ldots, n$, reflect area B^i in area A^i, and functional dependencies $\phi_j, j = 1, \ldots, m$, reflect area A^i in area B^i . At monotony of f_i, $i = 1, \ldots, n$, and $n \geq m$, mapping of A^i in B^i, i.e. functional dependencies $\phi_j, j = 1, \ldots, m$, can be identically determined by solving system of equations (4.1) provided that it exists. As a rule, for real technological processes, the functions f_i, $i = 1, \ldots, n$, are monotonous even in the certain sub area of their definition area, and system of equations (4.1) has a simple solution [11]. At $n < m$, additional conditions can be found for mutual uniqueness of mappings f and ϕ .

Thus, we consider the case when, at monotony of functions $f_i, i = 1, \ldots, n$, it is always possible to find the conditions of mutual uniqueness of mappings f and ϕ.

In that case, to areas A^i determined by relations (4.3), there correspond areas B^i determined by the formulae:

$$B^i = \left\{b : b_{ij}^{'} \leq b_j \leq b_{ij}^{''}; \forall j : \ j \in (1, \ldots, m)\right\}, i = 1, \ldots, S, \tag{4.5}$$

where boundary values $b_{ij}^{'}$ also $b_{ij}^{''}$ are defined as follows:

$$\begin{aligned} b_{ij}^{'} &= \min_{\{a\in A^i\}} \phi_j \left(a_{j_1}, a_{j_2}, \dots, a_{jQ_j}; d_{j_1}, d_{j_2}, \dots, d_{jR_j}\right), \\ b_{ij}^{''} &= \max_{\{a\in A^i\}} \phi_j \left(a_{j_1}, a_{j_2}, \dots, a_{jQ_j}; d_{j_1}, d_{j_2}, \dots, d_{jR_j}\right), \end{aligned} \tag{4.6}$$

$$i = 1, \dots, S.$$

Let there be available Θ technological processes. Generally $\Theta \neq S$. At each j-th technological process, the area of possible values of its parameters $B_t^j, j = 1, \dots, \Theta$, can intersect one or several (in the limit, all) areas $B^i, i = 1, \dots, S,$ determined by relations (4.5) and (4.6). We shall designate these intersections as follows:

$$B_t^{j,l_k} = B_t^j \cap B^{l_k}, l_{_k} \in (1, \dots, S), k = 1, \dots, S_j, \tag{4.7}$$

i.e., when its parameter values belong to area B_t^{j,l_k} , the j-th technological process can provide the $l_{_k} \in (1, \dots, S)$ quality of finished product and the quantity of such qualities is equal to $S_j \leq S$.

For the j-th technological process, for supporting the values of the parameters in area B_t^{j,l_k} , it is necessary to determine the expenses by the relations:

$$\begin{aligned} E_{j,l_k} = \Psi_j \left(b_1, b_2, \dots, b_m; e_1^j, e_2^j, \dots, e_{q_j}^j\right), b \in B_t^{j,l_k}, \\ l_k \in (1, \dots, S), \ k = 1, \dots, S_j, \ j = 1, \dots, \Theta, \end{aligned} \tag{4.8}$$

where Ψ_j is the function defining the expenses size at the j-th technological process; $e_1^j, e_2^j, \dots, e_{q_j}^j$ are the parameters of this function; q_j is the quantity of these parameters.

If by

$$\begin{aligned} I_{j,l_k} = \psi_j \left(a_1, a_2, \dots, a_n; \theta_1^j, \theta_2^j, \dots, \theta_{p_j}^j\right), a \in A^{l_k}, \\ l_k \in (1, \dots, S), \ k = 1, \dots, S_j, \ j = 1, \dots, \Theta, \end{aligned} \tag{4.9}$$

where $\theta_1^j, \theta_2^j, \dots, \theta_{p_j}^j$ are the parameters of corresponding functional dependencies, we designate the income from the sale of the production of $l_k \in (1, \dots, S)$ quality $(k = 1, \dots, S_j)$ obtained by the j-th technological process (i.e. ψ_j is the function defining the income at the j-th technological process, and p_j is the quantity of its parameters), then the amount of possible profit obtained by the j-th technological process can be calculated by the relation:

$$G_{j,l_k} = I_{j,l_k} - E_{j,l_k}, \tag{4.10}$$

$$l_k \in (1,\dots,S)\,, k=1,\dots,S_j, j=1,\dots,\Theta.$$

The decision concerning the quality of production to be made on the basis of the measured values $x=(x_1,\dots,x_n)$ of the parameters $a=(a_1,\dots,a_n)$. As a rule, the measured values x contain random errors just as because of the essence of the considered technological process (at which, as a rule, the values of corresponding parameters fluctuate randomly) and the method of control (the finished products cannot be absolutely homogeneous, i.e. the quality parameters fluctuate randomly and the quantity of controllable products is limited) so because of the random character of measurement errors. Therefore each decision about the quality of production on the basis of is accompanied by a certain risk of being erroneous. The problem consists in the choosing such a mode of operation from all given technological processes and such values of the parameters of the technological process in the given mode of operation which will provide the maximum profit at the minimum risk, i.e. with the minimum average probability of obtaining the production undesirable quality and making the erroneous decision concerning the production quality at the given likelihood characteristics of random distortions.

The j-th technological process ($j=1,\dots,\Theta$) can ensure obtaining the products of $S_j \le S$ qualities by choosing the corresponding values of the parameters. We shall designate the probability distribution law of measurement results of production quality parameters on the basis of which the decision is made at the supposition that the production has the i-th quality by $p(x/H_i)$, where $H_i : a \in A^{l_i},\ \ l_i \in (1,\dots,S,\ ),\ i=1,\dots,S_j$, is the supposition (or that is the same - the hypothesis) that manufactured production on the whole has the i-th quality. The problem consists in the following: for each technological process in the corresponding areas of production quality $A^{l_i},\ \ l_i \in (1,\dots,S,\ ),\ i=1,\dots,S_j$ there are defined such values of parameters $a=(a_1,a_2,\dots,a_n)$, i.e. such n-dimensional points, that the averaged risk of obtaining the production of some quality at other planned quality was minimum and the profit resulting from the realization of the corresponding mode of technological process at corresponding values of parameters was maximum. For choosing the optimum (in the sense of the above mentioned) technological process and the corresponding mode of operation, it is necessary to define the solving rules of taking the optimum decisions about the production quality and to calculate the corresponding value of average risk:

$$r\left(\delta(x)\right)=\sum_{i=1}^{S}\rho\left(H_i,\delta(x)\right)p\left(H_i\right)$$

$$= \sum_{i=1}^{S} p(H_i) \int_{\mathbb{R}^{\Bbbk}} L\left(H_i, \delta(x)\right) p\left(x/H_i\right) dx, \tag{4.11}$$

where $H_i,\ i = 1, \ldots, S$, is the hypothesis that the production quality is in the i-th state, i.e. $a_i \in A^i,\ i = 1, \ldots, S$; $p(H_i)$ is a priori probability of the H_i hypothesis; $\delta(x) = \{\delta_1(x), \delta_2(x), \ldots, \delta_S(x)\}$ is the solving rule which, to each vector of observation x, assigns a certain decision, i.e. a certain hypothesis, where

$$\delta_j(x) = \begin{cases} 1, & \textit{if hypothesis } H_i \textit{ is accepted};\\ 0, & \textit{on the contrary.} \end{cases}$$

Thus, from the above mentioned, it is obvious that three absolutely different problems can be formalized as identical mathematical problems of testing of many hypotheses. Obviously, besides these problems, there are some other ones which by formalization, could be reduced to the same problem of mathematical statistics of testing of many hypotheses. As an example, let us note the problem of detection of the earthquake center by registered seismological waves and a lot of others.

Depending on the available a priori information and the aim, for solution of these problems different methods of statistical hypothesis testing could be used [12-20]. Among these methods, at availability suitable a priori information, the most universal and perfect methods are unconditional and conditional Bayesian methods of many-hypotheses tasting, which allow decision-making with certain significance level of criterion [12-15,17]. Below we give solution of conditional and unconditional Bayesian problems of many-hypotheses testing.

5 Statement and Solution of Conditional and Unconditional Bayesian Problems of Many-Hypotheses Testing

The essence of Bayesian problem of testing many hypotheses is as follows. On the basis of measured n-dimensional point $x = (x_1, x_2, ..., x_n)$, it is necessary to accept one of $H_i,\ i = 1, ..., S$, hypotheses. For simplicity of representation, below we shall use the following denotations: $x-$ the n-dimensional measured point in X^n- space of measurement; $a_i = (a_1^i, a_2^i, ..., a_n^i)$ is the mathematical expectation of measurement point $x = (x_1, x_2, ..., x_n)$ on condition that hypothesis H_i is true, i.e. $E(x|H_i) = a_i$; $p(H_i)$ is the a priori probability of hypothesis H_i; $p(x|H_i)$ is the probability distribution of x on condition that hypothesis H_i is true; $D = \{d\}$ is a set of solutions, where $d = \{d_1, ..., d_S\}$, it being so that

$$d_i = \begin{cases} 1, & \mathit{if\, hypothesis\ H_i\, is\ accepted,} \\ 0, & \mathit{otherwise;} \end{cases}$$

$\delta(x) = \{\delta_1(x), \delta_2(x), ..., \delta_S(x)\}$ is the decision function that associates each observation vector x with a certain decision

$$x \xrightarrow{\delta(x)} d \in D.$$

$E_i = \{x : \delta_i(x) = 1\}$ is the region of hypotheses H_i acceptance. As hypotheses H_i, $i = 1, ..., S$, are non-interceptable, $E_i \cap E_j = 0$ and $\bigcup_{i=1}^{S} E_i = X^n$, where S is the number of hypotheses.

Let hypothesis H_i be true. We introduce loss function $L(H_i, \delta(x))$. Then the risk, corresponding to hypotheses H_i, is determined in the following way:

$$\rho(H_i, \delta(x)) = \int_{X^n} L(H_i, \delta(x)) \cdot p(x|H_i)dx.$$

For each decision rule $\delta(x)$, the risk function is [17]:

$$r(\delta(x)) = \sum_{i=1}^{S} \rho(H_i, \delta(x))p(H_i) = \sum_{i=1}^{S} P(H_i) \int_{X^n} L(H_i, \delta(x))p(x|H_i)dx. \tag{5.1}$$

The problem consists in finding of such decision rule $\delta^*(x)$, i.e. such E_i, $i = 1, ..., S$, hypotheses H_i acceptance regions, for which we would have

$$r(\delta^*(x)) = \min_{\{\delta(x)\}} r(\delta(x)). \tag{5.2}$$

5.1 Stepwise Loss Function

Let's consider the case, when the losses for falsely accepted hypotheses are identical, while those for correctly made decisions are equal to zero, i.e.

$$L(H_i, H_j) = \begin{cases} C \; npu \;\; i \neq j, \\ 0 \; npu \;\; i = j. \end{cases} \tag{5.3}$$

The risk function being

$$r(\delta(x)) = C\left(1 - \sum_{i=1}^{S} p(H_i) \int_{E_i} p(x|H_i)dx\right). \tag{5.4}$$

The minimum in (5.4) is reached by solving the following problem

$$\sum_{i=1}^{S} p(H_i) \int_{E_i} p(x|H_i)dx \Rightarrow \max_{\{E_i\}}. \tag{5.5}$$

It is seen from here, that we may consider $C = 1$ without any loss of generality. The solution of problem (5.5) has the following form:

$$E_i = \Big\{x : p(H_i)p(x|H_i) > p(H_j)p(x|H_j); \forall j : \\ j \in (1, ..., i-1, i+1, ..., S)\Big\}. \tag{5.6}$$

5.2 Non-stepwise Loss Function

Let's rewrite expression (5.1) in the following way:

$$r(\delta(x)) = \sum_{j=1}^{S} \sum_{\substack{i=1,\\ i\neq j}}^{S} L(H_i, H_j)p(H_i) \int_{E_j} p(x|H_i)dx. \tag{5.7}$$

It is not difficult to make sure that optimal region E_j of hypothesis H_j acceptance, which minimizes the risk function (5.7), assumes the following form:

$$E_j = \left\{x : \sum_{i=1}^{S} L(H_i, H_j)p(H_i)p(x|H_i) < \sum_{i=1}^{S} L(H_i, H_k)p(H_i)p(x|H_i); \\ \forall k : k \in (1, ..., j-1, j+1, ..., S)\right\},\ j = 1, ..., S. \tag{5.8}$$

5.3 Conditional Bayesian Problem of Many-Hypotheses Testing

If, for some reason or other, it is difficult to define loss function $L(H_i, \delta(x))$, or it is required to have the guaranteed decision concerning errors of first and second kind, e.g. to have the guaranty that the error probability of omitting true decisions would not exceed the prescribed level, then, instead of unconditional Bayesian problem [17], it is necessary to solve conditional problem of optimization with respect to losses, caused by the made decision [17].

The problem is stated as follows: it is necessary to find a decision rule $\delta(x)$ such that the mean number of false decisions would be minimized

$$r(\delta(x)) = \sum_{i=1}^{S} p(H_i) \sum_{\substack{j=1\\ j\neq i}}^{S} \int_{E_j} p(x|H_i)dx \Rightarrow \min_{\{E_j\}} \tag{5.9}$$

at restrictions imposed on the mean probability of correctly made decisions

$$\sum_{i=1}^{S} p(H_i) \int_{E_i} p(x|H_i)dx \geq 1 - \alpha, \tag{5.10}$$

where $1 - \alpha$ specified probability.

Solution of problem (5.9), (5.10) has the following form

$$E_i = \left\{ x : \sum_{\substack{j=1, \\ j\neq i}}^{S} p(H_j)p(x|H_j) < \lambda \cdot p(H_i)p(x|H_i) \right\}, \ i = 1, ..., S. \tag{5.11}$$

References

1. Tiurin, I. N., Makarov, A. A. The statistical analysis of the data on the computer. INFRA-M, Moscow, 1998.

2. Gnedenko, B.V. Mathematics and scientific knowledge. Znanie, Moscow, 1983.

3. Kuzmin, S.Z. The basis of digital processing of radio-location information. Sovetskoe Radio, Moscow, 1974.

4. Primak, A.V., Kafarov, V.V., Kachiashvili, K.J. System analysis of air and water quality control. Naukova Dumka, Kiev, 1991.

5. Kachiashvili, K.J., Melikdzhanian, D.I. Identification of River Water Excessive Pollution Sources. International Journal of Information Technology & Decision Making, World Scientific Publishing Company, 5, 2, 2006, 397-417.

6. Cook, R.D., Ni, L. Using intraslice covariances for improved estimation of the central subspace in regression. Biometrika, 93, 1, 2006, 65-74.

7. Krzanowski, W.J., Marriott, F.H.C. Multivariate Analysis, Part 1: Distributions, Ordination and Inference. Arnold, London, 1994.

8. Krzanowski, W.J., Marriott, F.H.C. Multivariate Analysis, Part 2: Classification, Covariance Structures and Repeated Measures. Arnold, London, 1995.

9. Stoica, P., Viberg, M. Maximum likelihood parameter and rank estimation in reduced-rank multivariate linear regression. IEEE Transaction Signal Processing, 44, 1996, 3069-3078.

10. Kachiashvili, K.J., Melikdzhanian, D.I., Methodology of nonlinear regressions identification by modified method of least squares. Zavadskaia Laboratoria, 5, 2000, 157-164.

11. Lopez-Granados F., Jurado-Exposito M., Atenciano S. and others. Spatial variability of agricultural soil parameters in southern Spain. Plant and Soil 246, 2002, 97-105.

12. Berger, J.O. Statistical Decision Theory and Bayesian Analysis. Springer, New York, 1985.

13. De Groot, M. Optimal statistical decisions. McGraw-Hill Book Company, 1970.

14. De Groot, M. Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. J. American Statistical Association, 68, 1973, 966-969.

15. Dickey, J. Is the tail area useful as an approximate Bayes factor. J. American Statistical Association, 72, 1977, 138-142.

16. Hwang, J.T., Casella, G., Robert, C., Wells, M., et al. Estimation of Accuracy in Testing. The Annals of Statistics, 20, 1, 1992, 490-509.

17. Kachiashvili, K.J. Generalization of Bayesian Rule of Many Simple Hypotheses Testing. International Journal of Information Technology & Decision Making, World Scientific Publishing Company, 2, 1, 2003, 41 - 70.

18. Lehmann, E.L. Testing Statistical hypotheses. Wiley, New York, 1986.

19. Lindley, D.V. Making Decisions. Wiley, New York, 1985.

20. Meinshausen, N., Buhlmann, P. Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. Biometrika, 92, 4, 2005, 893-907.