საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 4, № 2, 2010
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 4, no. 2, 2010

Mathematics

About Using Sequential Analysis Approach for Testing Many Hypotheses

Kartlos Kachiashvili*, Muntazim Abbas Hashmi**

* *I. Vekua Institute of Applied Mathematics, Tbilisi State University;Abdus Salam School of Mathematical Sciences of GC University, Lahore, Pakistan*
** *Abdus Salam School of Mathematical Sciences of GC University, Lahore, Pakistan*

(Presented by Academy Member Elizbar Nadaraya)

ABSTRACT. New sequential method of testing many hypotheses based on special properties of decision-making areas in the conditional Bayesian task of testing many hypotheses is offered. The results of research of the properties of this method are given. They show the consistency, simplicity and optimality of the obtained results in the sense of the chosen criterion, which consists in the upper restriction of the probability of the error of one kind and the minimization of the probability of the error of the second kind. The examples of testing of hypotheses for the case of the sequential independent sample from the multidimensional normal law of probability distribution with correlated components are cited. They show the high quality of the offered methods. *© 2010 Bull. Georg. Natl. Acad. Sci.*

Key words: *conditional Bayesian tasks, decision rule, hypotheses testing, sequential analysis.*

1. Introduction

Sequential methods were first developed by Wald [1, 2] and Barnard [3]. The historical development of this subject is nicely described in Ghosh & Sen [4]. The properties of optimality of this method were investigated in Wald [1, 2, 5], and so did many other authors: Girshick [6, 7], Ghosh [8], Siegmund [9] and others). Some time later the development of Bayesian sequential procedures started: Arrow, Blackwell, & Girshick [10], Ray [11], Barnard [12], Anscombe [13], Berger [14] and others). The essence of these procedures consists in minimization of the risk which is defined as the average cost of taking observations plus the average loss resulting from erroneous decisions. A number of sequential criteria for testing many hypotheses are known. Their logical development are sequentially rejective multiple test methods, which include a modified Bonferroni procedure with a greater power than the Bonferroni procedure, offered by Holm [15, 16]. In Shiryaev [17] it was shown that the search for Bayesian decision rules could be reduced to solving a problem of optimal stopping for the Markov random function constructed in a special manner. In the work of Bartroff [18], multistage tests of simple hypotheses are described. Using a loss function, which is a linear combination of sampling costs and error probabilities, these tests are shown to minimize the integrated risk to second order as the costs per stage and per observation approach zero.

The methods of sequential analysis described in the above-mentioned works (Wald's method and the method based on the Bayesian approach) are quite simple, graphic and convenient for practical realization, but unfortunately, only for the case of two hypotheses. For an arbitrary number of hypotheses, the problem becomes significantly complex, and it has not been solved completely in the sense of classical statements of both the sequential criterion

based on the sequential probability ratio test (Wald statement) and the minimization of the sum of Bayesian risk calculated for sequentially incoming observation results and the cost of obtaining of the same results of the experiment. However, there are different possible procedures offered both by Wald [1, 2], and other authors (see, for example, Berger [14]) for solving the problem for an arbitrary number of hypotheses, but, as a rule, they do not possess the optimal properties in the scope of the chosen criteria or these properties are still not completely investigated.

Below we offer new methods of sequential analysis for testing many hypotheses, which are based on the specific properties of decision-making areas in conditional Bayesian problems of testing many hypotheses [19, 20]. For simplicity and clarity of the offered sequential method, let us briefly describe one of the mentioned conditional Bayesian problems of testing many hypotheses and the properties of their decision-making areas.

2. Conditional Bayesian problem of testing many hypotheses

Let us consider *n*-dimensional random observation vector $x^T = (x_1,...,x_n)$ with probability distribution density $p(x,\theta) = p(x_1,...,x_n;\theta_1,...,\theta_k)$, given on σ-algebra of Borel set of space R^n $(x \in R^n)$, which is called the sample space. By $\theta^T = (\theta_1,...,\theta_k)$ is designated the vector of parameters of distribution. In general, $n \neq k$. Let in *k*-dimensional parametrical space Θ^k be given *S* possible values of considered parameters $\theta^{i^T} = (\theta_1^i,...,\theta_k^i)$, *i*=1,...,*S*, i.e. $\theta^i \in \Theta^k$; $\forall i : i = 1,...,S$. On the basis of $x^T = (x_1,...,x_n)$ it is necessary to make the decision namely by which distribution $p(x,\theta^i)$, *i*=1,...,*S*, the sample *x* is born.

Let us introduce designations: $H_i : \theta = \theta^i$ is the hypothesis that the sample $x^T = (x_1,...,x_n)$ is born by distribution $p(x,\theta^i) = p(x_1,...,x_n;\theta_1^i,...,\theta_k^i) \equiv p(x \mid H_i)$, *i*=1,...,*S*; $p(H_i)$ is the a priori probability of hypothesis H_i.

Conditional Bayesian task has the following form (see [19, 20]). Find such a decision rule, i.e. such decision-making areas $\Gamma_1, \Gamma_2,..., \Gamma_S$ that

$$r_\delta = \sum_{i=1}^{S} p(H_i) \cdot \sum_{j=1; j\neq i}^{S} \int_{\Gamma_j} p(x \mid H_i)dx \Rightarrow \min_{\{\Gamma_j\}}, \tag{2.1}$$

at restrictions

$$1 - \sum_{i=1}^{S} p(H_i) \int_{\Gamma_i} p(x \mid H_i)dx \leq \alpha . \tag{2.2}$$

The solution of this problem is

$$\Gamma_j = \left\{ x : \sum_{i=1; i\neq j}^{S} p(H_i)p(x \mid H_i) < \lambda \cdot p(H_j)p(x \mid H_j) \right\}, \quad j = 1,...,S, \tag{2.3}$$

where λ is determined so that in (2.2) the equality takes place.

Task (2.1), (2.2) is one of possible formulations of the conditional Bayesian problem. In a similar manner, we can introduce and solve a set of other conditional Bayesian tasks which we omit here for conciseness.

The investigation of the properties of the decision-making areas (2.3) shows that, if λ=1, then for decision-making areas there take place $\Gamma_i \bigcap \Gamma_j = 0$ and $\bigcup_{i=1}^{S} \Gamma_i = R^n$, where R^n is the observation space, i.e. on the basis of any observation result *x* one of the tested hypotheses H_i, *i*=1,...,*S,* is accepted without fail.

At violation of this condition, depending on the values of undetermined Lagrange coefficient (the value of which

is determined by the significance level of the criterion, i.e. by the value of α), for the considered task in observation space R^n, the subareas of intersection of some (or, in a particular case, all) decision-making areas (let us call these areas the *ambiguous areas of decision*) could exist concurrently with the subareas which do not belong to any of the decision-making areas (let us call these areas the *impossible areas of decision*). In particular, at λ>1, there takes place $\bigcup_{i=1;i\neq j}^{S} \Gamma_i \ni \overline{\Gamma}_j$. This is available only if area Γ_j of acceptance of hypothesis H_j intersects with one or more (in the limit, with all) areas of acceptance of other hypotheses. At λ <1, there takes place $\bigcup_{i=1;i\neq j}^{S} \Gamma_i \in \overline{\Gamma}_j$. Thus, in the observation space R^n, there are such subareas which do not belong to any area of acceptance of the tested hypotheses.

Thus, the situation is similar to the sequential analysis in the case when, on the basis of present observation results, it could be impossible to make a decision (with the given probabilities of errors) about the validity of one of the hypotheses from the considered set. Therefore, in the considered task, if there emerges the situation of impossibility of making an ambiguous or any decision for the given significance level, we shall continue the observations until such an opportunity appears. For this reason, let us determine the expressions for the areas of acceptance of each of the tested hypotheses and of rejection of any of the tested hypotheses on the basis of the given number of sequentially obtained observation results. Thus, on the basis of the above-considered conditional Bayesian task, let us determine the methods of sequential analysis for testing many hypotheses. For clarity let us call this method *the sequential analysis method of Bayesian type*.

3. The method of sequential analysis of Bayesian type

Let us suppose that there is an opportunity of obtaining repeated observations. In order to introduce the method of sequential analysis for an arbitrary number of hypotheses on the basis of the above-considered conditional Bayesian task, let us use the designations introduced by Wald [1, 2]. Let R_m^n be the sampling space of all possible samples of m independent n-dimensional observation vectors $x=(x_1,...,x_n)$. Let us split R_m^n into S+1 disjoint subareas $R_{m,1}^n$, $R_{m,2}^n,...,R_{m,S}^n$, $R_{m,S+1}^n$. Let $p(x^1,...,x^m \mid H_i)$ be the total probability distribution density of m independent n-dimensional observation vectors. Then $p(x^1,...,x^m \mid H_i)=p(x^1 \mid H_i)\cdots p(x^m \mid H_i)$.

Let us determine the following decision rule. If the matrix of observation results $\mathbf{x}=(x^1,...,x^m)$ belongs to the subarea $R_{m,i}^n$, $i=1,...,S$, then hypothesis H_i is accepted, and, if $\mathbf{x}=(x^1,...,x^m)$ belongs to the subarea $R_{m,S+1}^n$, the decision is not made, and the observations go on until one of the tested hypotheses is accepted.

Areas $R_{m,i}^n$, $i=1,...,S+1$, are determined in the following way: $R_{m,i}^n$, $i=1,...,S$, is such a part of acceptance area Γ_i^m of hypothesis H_i which does not belong to any other area Γ_j^m, $j=1,...,i-1,i+1,...,S$; $R_{m,S+1}^n$ is such a part of sampling space R_m^n which belongs simultaneously to more than one area Γ_i^m, $i=1,...,S$, or it does not belong to any of these areas. Here the index m (m=1,2,...) points to the fact that the areas are determined on the basis of m sequential observation results.

Let us designate the population of subareas of intersections of acceptance areas Γ_i^m of hypotheses H_i (i=1,...,S) in conditional Bayesian task of hypothesis testing with the areas of acceptance of other hypotheses H_j, $j=1,...,S;\ j\neq i,$ by I_i^m. By $E_m^n=R_m^n-\bigcup_{i=1}^{S}\Gamma_i^m$, we designate the population of areas of space R_m^n which do not belong to any of hypotheses acceptance areas. Then the decision acceptance areas in the method of sequential

analysis of Bayesian type are determined in the following way.

At $\lambda>1$,

$$R_{m,i}^n = \Gamma_i^m / I_i^m,\ i = 1,...,S;$$

$$R_{m,S+1}^n = \bigcup_{i=1}^{S} I_i^m .$$

At $\lambda<1$

$$R_{m,i}^n = \Gamma_i^m,\ i = 1,...,S;$$

$$R_{m,S+1}^n = E_m^n .$$

At $\lambda=1$

$$R_{m,i}^n = \Gamma_i^m,\ i = 1,...,S;$$

$$R_{m,S+1}^n = \varnothing .$$

Here areas Γ_i^m, I_i^m, E_m^n, $i = 1,...,S$, are defined on the basis of decision-making areas (2.3) in conditional Bayesian task.

4. Consistency and uniqueness of sequential analysis method of Bayesian type

Let us designate: M_1 and M_2 are the first and the second methods of testing of statistical hypotheses; α_i, β_i are the probabilities of errors of the first and the second kinds, respectively, corresponding to the methods M_i, i=1,2, at the identical number of observations. Let us introduce the following definition.

Definition. The method of testing of statistical hypotheses M_1 *rigorously surpasses* the method M_2 if there take place $\alpha_1<\alpha_2$ and $\beta_1<\beta_2$.

For clarity, from here on, by α_1 and β_1, we shall designate the probabilities of errors of the first and the second kinds for sequential method of Bayesian type, and, by α and β, the same quantities for conditional Bayesian task. The theorems confirming the consistency and uniqueness of the offered sequential analysis method of Bayesian type are given below without proving.

Proposition 1. *If the probability distribution* $p(\mathbf{x} \mid H_i),\ i = 1,...,S$, *is such that an increase in the sample size m entails a decrease in the entropy concerning distribution parameters* θ *about which the hypotheses are formulated, then infinitely increasing number of repeated observations, i.e.* $m \to \infty$ *in the sequential analysis method of Bayesian type, entails infinite decreasing probabilities of errors of the first and the second kinds, i.e.* $\alpha_1 \to 0$ *and* $\beta_1 \to 0$.

Lemma 1. *In the conditions of theorem 1, at increasing divergence* $J(H_i, H_j)$ *between tested hypotheses* H_i *and* H_j, $i, j = 1,...,S;\ i \neq j$, *Lagrange coefficient* λ *in solution (2.3) decreases, and, in the limit, at* $\min\limits_{\{i,j\}} J(H_i, H_j) \to \infty$, $\lambda \to 0$ *takes place for the given* α.

Hereinafter we shall suppose that probability distributions $p(x \mid H_i),\ i = 1,...,S$, are such that increasing information causes a decrease in the entropy relative to parameter θ which the hypotheses are formulated about.

Proposition 2. *For any given sample size m and as small errors of the first and the second kinds* α' *and* β' *as one likes, there always exists such a positive value* J^* *that, if the divergence between tested hypotheses is more than that value, i.e.* $\min\limits_{\{i,j\}} J(H_i, H_j) > J^*$, $\alpha_1(J) < \alpha'$ *and* $\beta_1(J) < \beta'$ *hold true, i.e. the method of sequential analysis of Bayesian type rigorously surpasses the criterion with errors of the first and the second kinds equal to* α' *and* β', *respectively.*

Proposition 3. *For any value of* α *in conditional Bayesian task there always exists such an integer* m^* *that if*

the number of repeated observations m, in the method of sequential analysis of Bayesian type, is more than this value, i.e. m> m^*, there will be accepted one of the tested hypotheses with the probability equal to unity.

5. Experimental research

To illustrate the correctness of the above-mentioned results and showing the quality of the offered methods in practice, let us bring the calculation results of some examples for the cases when sequentially accepted observation results are normally distributed independent random variables. In the example where the number of hypotheses is equal to two are also given the results of operation of the Wald method with parameters: α=0.05 and β=0.05. Consequently, the decision-making thresholds for this criterion are equal to B=0.05263 and A=19.

Example 1. Tested hypotheses: $H_1 : a_1^1 = 1, a_2^1 = 1$, $H_2 : a_1^2 = 4, a_2^2 = 4$. A priori probabilities of hypotheses: $p(H_1) = 0.5$, $p(H_2) = 0.5$. The significance level of the criterion in conditional Bayesian task is α=0.05. The parameters of sequentially incoming observation results as a two-dimensional normally distributed random vector with the mathematical expectation a=(4;4) and the covariance matrix $W = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

To save room, let us describe the obtained results without going into details. Totally there were generated 60 sequentially incoming observation results on the basis of which, in the sequential Wald criterion decisions were made 36 times, of these 18 decisions were made on the basis of 1 observation result, 14 - on the basis of 2 observation results, 3 - on the basis of 3 observation results and 1 decision - on the basis of 5 observation results. In a sequential method of Bayesian type, decisions were made 52 times, of which 47 decisions were made on the basis of 1 observation result, 3 - on the basis of 2 observation results, 1 - on the basis of 3 observation results and 1 - on the basis of 4 observation results. All decisions are correct. The average number of observation results necessary for decision-making is equal to: in the Wald criterion - $\overline{n}_W = 1.6666(6)$; in the sequential method of Bayesian type - $\overline{n}_B = 1.1538$. The average probabilities of errors of the first and the second kinds in sequential method of Bayesian type at decision-making are equal to: on the basis of one observation - $\alpha' = 0.00469$ and $\beta' = 0.05$ ($\lambda = 0.13246$); on the basis of two observations - $\alpha' = 6.65 \cdot 10^{-6}$ and $\beta' = 0.05$ ($\lambda = 0.00029$); on the basis of three observations - $\alpha' = 5.85 \cdot 10^{-9}$ and $\beta' = 0.05$ ($\lambda = 3.34 \cdot 10^{-7}$); on the basis of four observations $\alpha' = 3.9 \cdot 10^{-12}$ and $\beta' = 0.05$ ($\lambda = 2.7 \cdot 10^{-10}$).

Example 2. Tested hypotheses: $H_1 : a_1^1 = 1, a_2^1 = 1$, $H_2 : a_1^2 = 4, a_2^2 = 4$, $H_3 : a_1^3 = 8, a_2^3 = 8$ and $H_4 : a_1^4 = 12, a_2^4 = 12$. A priori probabilities of hypotheses: $p(H_1) = 1/4$, $p(H_2) = 1/4$, $p(H_3) = 1/4$, $p(H_4) = 1/4$. The significance level of the criterion in conditional Bayesian task is α=0.05. The parameters of sequentially incoming observation results as a two-dimensional normally distributed random vector with the mathematical expectation a=(4;4) and the covariance matrix $W = \begin{pmatrix} 10 & 9 \\ 9 & 10 \end{pmatrix}$.

To save room, let us describe the obtained results without going into details. Totally there were generated 40 sequentially incoming observation results on the basis of which, in sequential method of Bayesian type, decisions were made 15 times; from them 8 decisions were made on the basis of 2 observation results, 4 - on the basis of 3 observation results and 3 - on the basis of 4 observation results. All decisions are correct except of one case when, instead of the second hypothesis, the first one is accepted on the basis of two, the sixteenth and the seventeenth, observation results (arithmetic mean of these observation results on the basis of which the decision is made is equal to $(x_{16} + x_{17})/2 = (0.0389;1.2279)$). The average number of observation results necessary for decision-making is equal to $\overline{n}_B = 2.66(6)$. The average probabilities of errors of the first and the second kinds in sequential method of

Bayesian type at hypotheses testing are equal to: on the basis of two observations - $\bar{\alpha}_m = 0.128$ and $\beta' = 0.05$ ($\lambda = 2.455$); on the basis of three observations - $\bar{\alpha}_m = 0.052$ and $\beta' = 0.05$ ($\lambda = 1.925$) and on the basis of four observations - $\bar{\alpha}_m = 0.00866(6)$ and $\beta' = 0.05$ ($\lambda = 1.465$), respectively.

6. Conclusion

From the above-mentioned it is obvious that the offered new method of sequential analysis of many hypotheses is convenient, unified and clear for use with the purpose of hypotheses testing for any number of tested hypotheses. In these methods the criterion of optimality is a restriction from above of the probability of error of one kind and minimization of the probability of error of the second kind. The adduced examples demonstrate high quality of the offered methods at testing hypotheses in different situations which differ both by the divergence between the hypotheses and the number of tested hypotheses. The offered sequential methods are quite reliable, they do not need a great number of observation results for hypotheses testing and each decision made is accompanied by calculated values of the probabilities of errors of both kinds.

მათემატიკა

მიმდევრობითი ანალიზის მიდგომის გამოყენების შესახებ მრავალი ჰიპოთეზის შემოწმების ამოცანებში

ქ. ყაჭიაშვილი*, მ. ა. ჰაშმი**

** ი.ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტის ი. ვეკუას გამოყენებითი მათემატიკის ინსტიტუტი; ლაჰორის სახელმწიფო კოლეჯ-უნივერსიტეტის აბდუ სალამის სახ. მათემატიკურ მეცნიერებათა სკოლა, პაკისტანი*
*** ლაჰორის სახელმწიფო კოლეჯ-უნივერსიტეტის აბდუ სალამის სახ. მათემატიკურ მეცნიერებათა სკოლა, პაკისტანი*

(წარმოდგენილია აკადემიის წევრის ე. ნადარაიას მიერ)

შემოთავაზებულია მრავალი ჰიპოთეზის შემოწმების ახალი მეთოდი, რომელიც დაფუძნებულია პირობით ბაიესის ამოცანაში გადაწყვეტილებების მიღების არეების განსაკუთრებულ თვისებებზე. მოცემულია მეთოდის თვისებების გამოკვლევის შედეგები. ისინი გვიჩვენებენ მიღებული შედეგების ძალმოსილობას, სიმარტივეს და ოპტიმალურობას არჩეული კრიტერიუმის თვალსაზრისით, რომელიც მდგომარეობს ერთი ტიპის შეცდომების ალბათობების ზემოდან შეზღუდვაში და მეორე ტიპის შეცდომების ალბათობების მინიმიზაციაში. მოყვანილია ჰიპოთეზების შემოწმების მაგალითები მიმდევრობით მიღებული კორელირებულ კომპონენტებიანი ნორმალურად განაწილებული დამოუკიდებელი ამონარჩევისათვის. ისინი ადასტურებენ შემოთავაზებული მეთოდის მაღალ ხარისხს.

REFERENCES

1. *A. Wald* (1947a), Sequential Analysis. Wiley, NY.
2. *A. Wald* (1947b), Econometrica, 15: 279-313.
3. *G.A. Barnard* (1946), Suppl. J. Roy. Statist. Soc., **8**, 1.

4. *B.K. Ghosh and P.K. Sen* (eds.) (1991), Handbook of Sequential Analysis. New York: Dekker.
5. *A. Wald and J. Wolfowitz* (1948), Ann. Math. Statist. 19: 326-339.
6. *M.A. Girshick* (1946a), Ann. Math. Statist., **17**, 2: 123-143.
7. *M.A. Girshick* (1946b), Ann. Math. Statist., **17**, 3: 282-298.
8. *B.K. Ghosh* (1970), Sequential Tests of Statistical Hypotheses. Massachusetts: Addison-Wesley, Reading.
9. *D. Siegmund* (1985), Sequential Analysis. Springer Series in Statistics. New York: Springer-Verlag.
10. *K.J. Arrow, D. Blackwell & M.A. Girshick* (1949), Econometrica, **17** (2): 213-244.
11. *S.N. Ray* (1965), Ann. Math. Statist., **36**, 3: 859-878.
12. *G.A. Barnard* (1947), J. Amer. Statist. Assoc., 42: 658-669.
13. *F.J. Anscombe* (1963), J. Amer. Statist. Assoc., 58: 365-383.
14. *J.O. Berger* (1985), Statistical Decision Theory and Bayesian Analysis. New York: Springer.
15. *S. Holm* (1977), Statistical Research Report, University of Umea (Sweden): Institute of Mathematics and Statistics.
16. *S. Holm* (1979), Scand. J. Statist., 6: 65-70.
17. *A.N. Shiryaev* (2008), Optimal Stopping Rules. Berlin: Springer-Verlag.
18. *J. Bartroff* (2007), Ann. Stat., **35**, 5: 2075-2105.
19. *K.J. KachiashviIi* (1989), Bayesian algorithms of many hypothesis testing. Tbilisi.
20. *K.J. Kachiashvili* (2003), International Journal of Information Technology & Decision Making, World Scientific Publishing Company, **2**, 1: 41-70.
21. *S. Kullback* (1978), Informatin Theory and Statistics. Wiley.
22. *W. Feller* (1968, 1971), An Introduction to the Theory of Probability and its Applications. V. 1, 3d ed. and V. 2, 2nd ed. New York: Wiley.

Received May, 2010