УДК 519.2

СТАТИСТИЧЕСКАЯ ПОСЛЕДОВАТЕЛЬНАЯ ПРОВЕРКА ГИПОТЕЗ О ПАРАМЕТРАХ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ СЛУЧАЙНЫХ БИНАРНЫХ ДАННЫХ

А. Ю. ХАРИН[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Рассматривается актуальная математическая задача компьютерного анализа данных – задача статистической последовательной проверки простых гипотез о параметрах распределения вероятностей наблюдаемых бинарных данных. Эта задача решается для двух моделей наблюдений: схемы независимых испытаний и однородной цепи Маркова. Выведены легко интерпретируемые и удобные для компьютерной реализации явные выражения статистик последовательных тестов (статистических критериев). Разработан подход для вычисления характеристик эффективности решающих правил – вероятностей ошибочных решений и математических ожиданий случайного числа наблюдений, необходимых для обеспечения требуемой точности. Получены асимптотические разложения для указанных характеристик эффективности при «засорениях» распределения вероятностей наблюдаемых данных.

Ключевые слова: случайные бинарные данные; простые гипотезы; статистический последовательный тест; вероятность ошибки; математическое ожидание случайного числа наблюдений; «засорения»; асимптотические разложения.

Благодарность. Работа выполнена при финансовой поддержке Министерства образования Республики Беларусь в рамках государственной программы научных исследований «Конвергенция-2025».

Образец цитирования:
Харин АЮ. Статистическая последовательная проверка гипотез о параметрах распределений вероятностей случайных бинарных данных. *Журнал Белорусского государственного университета. Математика. Информатика.* 2021;2:60–66 (на англ.).
https://doi.org/10.33581/2520-6508-2021-2-60-66

For citation:
Kharin AYu. Statistical sequential hypotheses testing on parameters of probability distributions of random binary data. *Journal of the Belarusian State University. Mathematics and Informatics.* 2021;2:60–66.
https://doi.org/10.33581/2520-6508-2021-2-60-66

Автор:
Алексей Юрьевич Харин – доктор физико-математических наук, доцент; заведующий кафедрой теории вероятностей и математической статистики факультета прикладной математики и информатики.

Author:
Alexey Yu. Kharin, doctor of science (physics and mathematics), docent; head of the department of probability theory and mathematical statistics, faculty of applied mathematics and computer science.
kharinay@bsu.by

STATISTICAL SEQUENTIAL HYPOTHESES TESTING ON PARAMETERS OF PROBABILITY DISTRIBUTIONS OF RANDOM BINARY DATA

A. Yu. KHARIN[a]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*

An important mathematical problem of computer data analysis – the problem of statistical sequential testing of simple hypotheses on parameters of probability distributions of observed binary data – is considered in the paper. This problem is being solved for two models of observation: for independent observations and for homogeneous Markov chains. Explicit expressions of the sequential tests statistics are derived, transparent for interpretation and convenient for computer realisation. An approach is developed to calculate the performance characteristics – error probabilities and mathematical expectations of the random number of observations required to guarantee the requested accuracy for decision rules. Asymptotic expansions for the mentioned performance characteristics are constructed under «contamination» of the probability distributions of observed data.

Keywords: random binary data; simple hypotheses; statistical sequential test; error probability; mathematical expectation of the random number of observations; «contamination»; asymptotic expansions.

Acknowledgements. This work was supported by Ministry of Education of the Republic of Belarus within the state research program «Convergence-2025».

Introduction

Data become one of the active drivers of the world economy, and computer data analysis becomes an essential part of the modern life. Efficiency of data analysis defines the success in a growing spectrum of fields. Binary data is a very important class of data for several reasons: 1) binary data is natural for computer processing; 2) binary data describes many situations in terms of «presence – absence», «positive – negative», etc.; 3) binary data can be used for description of a significantly rich family of observations, if considered by groups. Classical methods of statistical analysis are often not applied to such data, as those methods assumptions are usually not satisfied for the binary data models, or they are not effective.

As deterministic approach has a limited potential to describe the processes, real-life data is usually considered to be random, and probabilistic models are used. In these models, an important problem that often appears, is a problem of discrimination between two typical situations on the probability distributions of random data. These two typical situations can be formulated in terms of simple hypotheses on parameters of the probability distribution, and the problem turns to the problem of statistical testing of two simple hypotheses [1].

In many cases, especially in statistical quality control, in automatic warning systems, in personalised medicine, in financial decision making, it is important to use the minimal number of observations that guarantee the requested accuracy [2]. Sequential statistical tests [3] follow this principle with the assumption that the number of observations to be used is defined through the observation process, depend on observations themselves, and thus is a random variable. Due to the complex structure of sequential decision rules, usually theoretical analysis of their performance characteristics – error probabilities and mathematical expectations of the random number of required observations – is problematic [4].

In the paper we develop an approach to calculate and analyse the performance characteristics of sequential tests for binary data. In practice the factual probability distribution of data often deviates from the hypothetical one – the hypothetical probability distribution is «contaminated» [5; 6], so we also consider this situation here, and analyse deviations of the performance characteristics under distortion.

Results for the model of independent binary random observations

Denote by $\mathbf{N}$, $\mathbf{Z}$, $\mathbf{R}$ correspondently sets of: positive integer, integer, and real numbers; $\mathbf{Z}_+ = \mathbf{N} \cup \{0\}$. Let independent identically distributed K-dimensional binary random vectors

$$x_t = (x_{ti}) \in U = \left\{u^1, \ldots, u^{2^K}\right\} = \left\{ \begin{pmatrix} u_1 \\ \vdots \\ u_K \end{pmatrix}, u_i \in B = \{0, 1\}, i = 1, \ldots, K \right\}, t \in \mathbf{N}, \tag{1}$$

be observed in a probability space (Ω, F, P) with independent components. The probability distribution of random vectors (1) depends on the unknown value of the parameters vector $\theta = \begin{pmatrix} p_1 \\ \vdots \\ p_K \end{pmatrix} \in \Theta = \{\theta_0, \theta_1\}$, where $\theta_0 = \begin{pmatrix} p_1^0 \\ \vdots \\ p_K^0 \end{pmatrix}$, $\theta_1 = \begin{pmatrix} p_1^1 \\ \vdots \\ p_K^1 \end{pmatrix}$, $\theta_0 \neq \theta_1$. Such a model is often used in practice to decide in favour of two possible alternative typical situations. The components of the vector of parameters $p_i \in \{p_i^0,\ p_i^1\}$, $p_i^0,\ p_i^1 \in (0, 1)$, mean the probabilities of the random event $\{x_{ti} = 1\}$, $t \in \mathbf{N}$, $i = 1, \ldots, K$.

Suppose the following assumption is satisfied for the probability distribution of binary random vectors (1):

$$P(u; \theta) = P_\theta\{x_t = u\} = a^{-J(u;\, \theta)},\ t \in \mathbf{N},\ u \in U, \tag{2}$$

where $a \in \mathbf{R}$, $a > 1$; with $J(u; \theta): U \times \Theta \to \mathbf{Z}_+$ being a function that satisfies the following condition:

$$\sum_{u \in U} a^{-J(u;\, \theta)} = 1. \tag{3}$$

There are two simple hypotheses considered on the parameters vector value θ of probability distribution (2):

$$H_0\colon \theta = \theta_0,\ H_1\colon \theta = \theta_1. \tag{4}$$

Denote the accumulated log-likelihood ratio test statistic:

$$\Lambda_n = \Lambda_n(x_1, \ldots, x_n) = \sum_{t=1}^{n} \lambda_t, \tag{5}$$

where

$$\lambda_t = \lambda(x_t) = \log_a\left(\frac{P(x_t; \theta_1)}{P(x_t; \theta_0)}\right) \tag{6}$$

is the log-likelihood ratio for the binary observation vector x_t.

Theorem 1. *For the model (1)–(3) the sequence of statistics (5) for hypotheses (4) is a homogeneous Markov chain with discrete time, and it has the form*

$$\Lambda_n = \Lambda_n(x_1, \ldots, x_n) = \sum_{k=1}^{2^K} n_k\left(J(u^k; \theta_0) - J(u^k; \theta_1)\right) \in \mathbf{Z},\ n \in \mathbf{N}, \tag{7}$$

where n_k denotes the number of the vectors that equal u^k observed within n binary random vectors $\{x_1, \ldots, x_n\}$, $\sum_{k=1}^{2^K} n_k = n$.

P r o o f. The Markov property [7] of the random sequence (5) follows from the independence of its increments (6) due to the independence of random observations (1). Expression (7) is derived by equivalent transformations of (5), (6) under the assumption (2), (3).

Corollary 1. *If under theorem 1 conditions $K = 1$, then test statistic (5) is*

$$\Lambda_n = \Lambda_n(x_1, \ldots, x_n) = n_0\left(J(0; p^0) - J(0; p^1)\right) + (n - n_0)\left(J(1; p^0) - J(1; p^1)\right),\ n \in \mathbf{N},$$

where n_0 denotes the number of the observations equal to 0.

Using statistics (7), the sequential probability ratio test [1] for hypotheses (4) is constructed as follows: the decision after n observations made ($n = 1, 2, \ldots$) is

$$d = d(n) = \mathbf{1}_{[C_+,\, +\infty)}(\Lambda_n) + 2 \cdot \mathbf{1}_{(C_-,\, C_+)}(\Lambda_n), \tag{8}$$

where $\mathbf{1}_D(\cdot)$ denotes the indicator function of a set D. Decisions $d = 0$ and $d = 1$ mean the observation process termination and acceptance of correspondently H_0 or H_1 after n binary random vectors observed; $d = 2$ means that the $(n + 1)$ vector should be observed; $C_-, C_+ \in \mathbf{Z}$, $C_- < C_+$ are parameters of the decision rule (8) called thresholds; in practice they are often calculated according to [3]:

$$C_- = \left[\log_a \left(\frac{\beta_0}{1 - \alpha_0} \right) \right], \ C_+ = \left[\log_a \left(\frac{1 - \beta_0}{\alpha_0} \right) \right], \tag{9}$$

where α_0, β_0 are the admissible values of error type I (to accept H_1, when H_0 is true) and error type II (H_1 is true, H_0 is accepted) probabilities; $[\cdot]$ means the integer part of an argument. With thresholds (9) the factual values of error probabilities of type I and II may differ from α_0, β_0, and the problem of the factual values calculation of the performance characteristics is open for the sequential tests.

Introduce the notation: $\delta_{i,\,j}$ for the Kronecker delta; $\mathbf{I}_k$ for the identity matrix of size k; $\mathbf{0}_{m \times n}$ for the zero-matrix of size $(m \times n)$; $\mathbf{1}(\cdot)$ for the unit step function; $\mathbf{1}_k$ for the k-vector column with components equal 1. Denote by $t^{(k)}$ the expected value of the random number of observations (sample size) provided the true hypothesis is H_k, $k \in \{0, 1\}$, and by α, β the factual values of error type I and II probabilities for test (8); $N = C_+ - C_-$; let $P^{(k)} = \left(p_{ij}^{(k)}\right) = \begin{pmatrix} \mathbf{I}_2 & \vdots & \mathbf{0}_{2 \times N} \\ --- & \vdots & --- \\ R^{(k)} & \vdots & Q^{(k)} \end{pmatrix}$ be the matrix of size $(N + 2) \times (N + 2)$, with blocks $R^{(k)}$, $Q^{(k)}$ defined by

$$p_{ij}^{(k)} = \begin{cases} \sum\limits_{u \in U} \delta_{J(u;\, \theta_0) - J(u;\, \theta_1),\, j - i} P(u;\, \theta_k), \ i, j \in (C_-, C_+), \\ \sum\limits_{u \in U} \mathbf{1}\big(C_- - i + J(u;\, \theta_1) - J(u;\, \theta_0)\big) P(u;\, \theta_k), \ i \in (C_-, C_+), \ j = C_-, \\ \sum\limits_{u \in U} \mathbf{1}\big(J(u;\, \theta_0 - J(u;\, \theta_1)) + i - C_+\big) P(u;\, \theta_k), \ i \in (C_-, C_+), \ j = C_+. \end{cases}$$

Denote

$$\pi^{(k)} = \left(\pi_i^{(k)}\right), \ \pi_i^{(k)} = \sum_{u \in U} \delta_{J(u;\, \theta_0) - J(u;\, \theta_1),\, i} P(u;\, \theta_k), \ i \in \{C_- + 1, \ldots, C_+ - 1\},$$

$$\pi_{C_+}^{(k)} = \sum_{i \ge C_+} \sum_{u \in U} \delta_{J(u;\, \theta_0) - J(u;\, \theta_1),\, i} P(u;\, \theta_k), \ \pi_{C_-}^{(k)} = \sum_{i \le C_-} \sum_{u \in U} \delta_{J(u;\, \theta_0) - J(u;\, \theta_1),\, i} P(u;\, \theta_k),$$

$$S^{(k)} = \mathbf{I}_N - Q^{(k)}, \ B^{(k)} = \left(S^{(k)}\right)^{-1} R^{(k)},$$

let $W_{(i)}$ means the i column of the matrix W.

Theorem 2. *If under the model (1)–(6),* $\left|S^{(k)}\right| \ne 0$, $k \in \{0, 1\}$, *then the performance characteristics of sequential decision rule (8) are calculated in the explicit form:*

$$t^{(k)} = \left(\pi^{(k)}\right)' \left(S^{(k)}\right)^{-1} \mathbf{1}_N + 1, \ \alpha = \left(\pi^{(0)}\right)' B_{(2)}^{(0)} + \pi_{C_+}^{(0)}, \ \beta = \left(\pi^{(1)}\right)' B_{(1)}^{(1)} + \pi_{C_-}^{(1)}.$$

Proof is based on the theory of finite homogeneous Markov chains with discrete time. The sequence

$$\zeta_n = C_- \cdot \mathbf{1}_{(-\infty,\, C_-]}(\Lambda_n) + C_+ \cdot \mathbf{1}_{[C_+,\, +\infty)}(\Lambda_n) + \Lambda_n \cdot \mathbf{1}_{(C_-,\, C_+)}(\Lambda_n)$$

is a homogeneous Markov chain with $N + 2$ states, and C_-, C_+ are the absorbing states.

The situation, where the probability distribution of data is «contaminated» is considered in the next section for a more general model, where the observed binary data form a Markov chain instead of being independent.

Results for the model of observations forming a homogeneous binary Markov chain

Suppose binary random vectors $x_1, x_2, \ldots$ forming a homogeneous Markov chain are being observed, taking values in the set $U = \left\{u^1, \ldots, u^{2^K}\right\} = \left\{ \begin{pmatrix} u_1 \\ \vdots \\ u_K \end{pmatrix}, \ u_i \in B = \{0, 1\}, \ i = 1, \ldots, K \right\}$. To simplify notation, introduce the set $V = \{0, 1, \ldots, M-1\}$, $M = 2^K$, one-to-one corresponding to the set U. Denote for the observed Markov chain the initial probabilities vector $\pi = (\pi_i)$, $i \in V$, and the one-step transition probabilities matrix $P = (p_{ij})$, $i, j \in V$:

$$P\{x_1 = i\} = \pi_i,\ P\{x_n = j | x_{n-1} = i\} = p_{ij},\ i, j \in V.$$

As in the case of independent binary random vectors, consider two simple alternative hypotheses on the parameter values of the Markov chain:

$$H_0:\ \pi = \pi^{(0)},\ P = P^{(0)};\ H_1:\ \pi = \pi^{(1)},\ P = P^{(1)}, \tag{10}$$

where $\pi^{(0)} = \left(\pi_i^{(0)}\right)$, $\pi^{(1)} = \left(\pi_i^{(1)}\right)$ are two given values for initial probabilities vector, $P^{(0)} = \left(p_{ij}^{(0)}\right) \neq P^{(1)} = \left(p_{ij}^{(1)}\right)$ are given matrices of one-step transition probabilities for the correspondent hypotheses.

For construction of the sequential test for this model of data, denote

$$\lambda_1 = \log \frac{\pi_{x_1}^{(1)}}{\pi_{x_1}^{(0)}},\ \lambda_k = \log \frac{p_{x_{k-1}, x_k}^{(1)}}{p_{x_{k-1}, x_k}^{(0)}},\ k > 1,\ \Lambda_n = \sum_{k=1}^{n} \lambda_k,\ n \in \mathbf{N}. \tag{11}$$

The sequential test for the considered model and hypotheses (10) is constructed according to the decision rule (8), with replacing (5)–(7) by (11). According to this test, for defined thresholds (see, e. g., (9)) C_-, $C_+ \in \mathbf{R}$, $C_- < 0$, $C_+ > 0$, hypothesis H_0 is accepted after n observations, if $\Lambda_n \le C_-$, hypothesis H_1 is accepted, if $\Lambda_n \ge C_+$, the test is stopped in both those cases, otherwise the test is proceeded, and the $(n+1)$ binary random vector should be observed. The sequence of $(K+1)$ dimensional random vectors $(\Lambda_n, x_n)'$, $n \in \mathbf{N}$, is a Markov chain by the definition:

$$P\{\Lambda_n, x_n | \Lambda_{n-1}, \Lambda_{n-2}, \ldots, \Lambda_1, x_{n-1}, x_{n-2}, \ldots, x_1\} = P\{\Lambda_n, x_n | \Lambda_{n-1}, x_{n-1}\}.$$

Suppose $\pi^{(0)}$, $P^{(0)}$, $\pi^{(1)}$, $P^{(1)}$ be satisfying the following assumption:

$$\exists a \in \mathbf{R},\ m_i, m_{ij} \in \mathbf{Z},\ i, j \in V:\ \log \frac{\pi_i^{(1)}}{\pi_i^{(0)}} = m_i a,\ \log \frac{p_{ij}^{(1)}}{p_{ij}^{(0)}} = m_{ij} a. \tag{12}$$

Without loss of generality, suppose for test (5), (8), (11) the thresholds C_-, $C_+ \in \mathbf{Z}$, and denote that $t^{(k)}$ is the expected sample size till one of the hypotheses is accepted, provided H_k is true, $k \in \{0, 1\}$;

$$W^{(k)} = \left(w_{ij}^{(k)}\right) = \begin{pmatrix} \mathbf{I}_2 & \vdots & \mathbf{0}_{2 \times MN} \\ --- & \vdots & --- \\ R^{(k)} & \vdots & Q^{(k)} \end{pmatrix}$$

is the matrix of the size $(MN+2)(MN+2)$, with blocks $R^{(k)}$, $Q^{(k)}$ defined by their elements ($s, t \in V$):

$$w_{Mi+s\ Mj+t}^{(k)} = \delta_{m_{st},\ j-i} p_{st}^{(k)},\ i, j \in (C_-, C_+),$$

$$w_{Mi+s\ Mj+t}^{(k)} = \begin{cases} \sum\limits_{t \in V} \mathbf{1}(i + m_{st} - C_+),\ j = C_+,\ i \in (C_-, C_+), \\ \sum\limits_{t \in V} \mathbf{1}(C_- - i - m_{st}),\ j = C_-,\ i \in (C_-, C_+); \end{cases} \tag{13}$$

as in the case of independent observations, denote the matrices

$$S^{(k)} = \mathbf{I}_{MN} - Q^{(k)},\ B^{(k)} = \left(S^{(k)}\right)^{-1} R^{(k)},\ k \in \{0, 1\};$$

the vectors

$$\omega^{(k)} = \left(\omega_i^{(k)}\right),\ i \in \{MC_- + 1, \ldots, MC_+ - 1\}:\ \omega_{Mi+s}^{(k)} = \delta_{m_s, i} \pi_s^{(k)},\ i \in (C_-, C_+), \tag{14}$$

and the two prior probabilities of absorption

$$\omega_{MC_-}^{(k)} = \sum_{s \in V} \mathbf{1}(C_- - m_s) \pi_s^{(k)},\ \omega_{MC_+}^{(k)} = \sum_{s \in V} \mathbf{1}(m_s - C_+) \pi_s^{(k)}. \tag{15}$$

Theorem 3. *For the observation model of a homogeneous binary vector Markov chain described above, if* $\left|S^{(k)}\right| \neq 0,\ k \in \{0, 1\}$, *and (12) holds, then the performance characteristics of sequential test (5), (8), (11) are calculated in the explicit form:*

$$t^{(k)} = \left(\omega^{(k)}\right)' \left(S^{(k)}\right)^{-1} \mathbf{1}_{MN} + 1,\ \alpha = \left(\omega^{(0)}\right)' B^{(0)}_{(2)} + \omega^{(0)}_{MC_+},\ \beta = \left(\omega^{(1)}\right)' B^{(1)}_{(1)} + \omega^{(1)}_{MC_-}.$$

P r o o f. Introduce the random sequence

$$\xi_n = MC_- \cdot \mathbf{1}_{(-\infty,\, C_-]}\left(\frac{\Lambda_n}{a}\right) + MC_+ \cdot \mathbf{1}_{[C_+,\, +\infty)}\left(\frac{\Lambda_n}{a}\right) + \left(\frac{\Lambda_n}{a} M + x_n\right) \cdot \mathbf{1}_{(C_-,\, C_+)}\left(\frac{\Lambda_n}{a}\right),\ n \in \mathbf{N},$$

that is a homogeneous Markov chain with $MN + 2$ states; two of them ($\xi_n = MC_-$ and $\xi_n = MC_+$) are absorbing. The one-step transition probabilities matrix is defined by (13), the vector of transient states initial probabilities is (14), and the initial probabilities of absorbing states are calculated in (15).

Consider now the situation that is often in practice, when the hypothetical model described above is «contaminated» in terms of the probability distribution of observations via «contamination» of the initial probabilities vector and of the one-step transition probabilities matrix. Suppose instead of the hypothetical values, the factual (distorted) vector of the initial probabilities and the factual one-step transition probabilities matrix are

$$\bar{\pi}^{(k)} = (1 - \varepsilon)\pi^{(k)} + \varepsilon\tilde{\pi}^{(k)},\ \bar{P}^{(k)} = (1 - \varepsilon)P^{(k)} + \varepsilon\tilde{P}^{(k)},\ k = 0, 1, \tag{16}$$

where $\tilde{\pi}^{(k)}$ and $\tilde{P}^{(k)}$ are the initial probabilities vector and the one-step transition probabilities matrix for the «contaminating» Markov chain, $P^{(k)} \neq \tilde{P}^{(k)}$, $k = 0, 1$, and $\varepsilon \in \left[0, \frac{1}{2}\right)$ is the probability of «contamination» (also called «contamination» level).

Denote for the «contaminated» model (16): $\tilde{W}^{(k)},\ \tilde{Q}^{(k)},\ \tilde{R}^{(k)},\ \tilde{\omega}^{(k)},\ \tilde{\omega}^{(k)}_{MC_\pm},\ k = 0, 1$, analogously to the hypothetical case, replacing the hypothetical probability distribution by the «contaminating» one. Denote that $\tilde{S}^{(k)} = \mathbf{I}_{MN} - Q^{(k)} - \varepsilon\left(\tilde{Q}^{(k)} - Q^{(k)}\right)$.

Theorem 4. *If the hypothetical model of binary vector homogeneous Markov chain observations described above is distorted according to (16), and assumption (12) is satisfied also for the distorted model, then the error type I and II probabilities* $\bar{\alpha}, \bar{\beta}$, *and the conditional mathematical expectations of the observation number* $\bar{t}^{(k)}$, $k = 0, 1$, *deviate from the hypothetical performance characteristics by the values of order* $O(\varepsilon)$:

$$\bar{\alpha} - \alpha = \varepsilon\Bigg(\left(\omega^{(0)}\right)'\left(\left(S^{(0)}\right)^{-1}\left(\left(\tilde{Q}^{(0)} - Q^{(0)}\right)\left(S^{(0)}\right)^{-1} R^{(0)} + \tilde{R}^{(0)} - R^{(0)}\right)\right)_{(2)} +$$

$$+ \left(\tilde{\omega}^{(0)} - \omega^{(0)}\right)' B^{(0)}_{(2)} + \tilde{\omega}^{(0)}_{MC_+} - \omega^{(0)}_{MC_+}\Bigg) + O\left(\varepsilon^2\right),$$

$$\bar{\beta} - \beta = \varepsilon\Bigg(\left(\omega^{(1)}\right)'\left(\left(S^{(1)}\right)^{-1}\left(\left(\tilde{Q}^{(1)} - Q^{(1)}\right)\left(S^{(1)}\right)^{-1} R^{(1)} + \tilde{R}^{(1)} - R^{(1)}\right)\right)_{(1)} +$$

$$+ \left(\tilde{\omega}^{(1)} - \omega^{(1)}\right)' B^{(1)}_{(1)} + \tilde{\omega}^{(1)}_{MC_-} - \omega^{(1)}_{MC_-}\Bigg) + O\left(\varepsilon^2\right),$$

$$\bar{t}^{(k)} - t^{(k)} = \varepsilon\left(\left(\tilde{\omega}^{(k)} - \omega^{(k)}\right)' + \left(\omega^{(k)}\right)'\left(S^{(k)}\right)^{-1}\left(\tilde{Q}^{(k)} - Q^{(k)}\right)\right)\left(S^{(k)}\right)^{-1} \mathbf{1}_{MN} + O\left(\varepsilon^2\right).$$

P r o o f. Under «contamination» (16) the initial probabilities vector and the one-step transition probabilities matrix of the random sequence ξ_n have the correspondent mixture form, and the rest of the proof is derived by equivalent transformations.

Conclusion

An approach to calculate and analyse the performance characteristics of sequential tests for binary data for two models is proposed: for the independent binary vectors, and for the homogeneous binary vector Markov chains. The situation, where the factual model of data deviates from the hypothetical assumptions, is considered in the paper, and correspondent differences in performance characteristics of the sequential tests are analysed asymptotically with respect to the «contamination» level. The results can be applied to construct robust sequential tests for binary random data [8], for the model of random sequences with a trend [9]. The results are also potentially applicable to the case of more than two hypotheses [10], complex hypotheses [11] and to the analysis of truncated sequential tests [12].

Библиографические ссылки

1. Mukhopadhyay N, de Silva B. *Sequential methods and their applications.* Boca Raton: CRC Press; 2009. 409 p.
2. Lai TL. *Sequential analysis: some classical problems and new challenges. Statistica Sinica.* 2001;11:303–408.
3. Wald A. Sequential analysis. New York: John Wiley and Sons; 1947. 212 p.
4. Айвазян СА. Сравнение оптимальных свойств критериев Неймана – Пирсона и Вальда. *Теория вероятностей и ее применения.* 1959;4(1):86–93.
5. Huber PJ, Ronchetti EM. *Robust statistics.* New York: Wiley; 2009. 354 p.
6. Maevskii VV, Kharin YuS. Robust regressive forecasting under functional distortions in a model. *Automation and Remote Control.* 2002;63(11):1803–1820. DOI: 10.1023/A:1020959432568.
7. Kemeny JG, Snell JL. *Finite Markov Chains.* New York: D. Van Nostrand Co.; 1960. 210 p.
8. Харин АЮ. *Робастность байесовских и последовательных статистических решающих правил.* Минск: БГУ; 2013. 207 с.
9. Kharin A, Tu TT. Performance and robustness analysis of sequential hypotheses testing for time series with trend. *Austrian Journal of Statistics.* 2017;46(3–4):23–36. DOI: 10.17713/ajs.v46i3-4.668.
10. Tu TT, Kharin AYu. Sequential probability ratio test for many simple hypotheses on parameters of time series with trend. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;1:35–45. DOI: 10.33581/2520-6508-2019-1-35-45.
11. Kharin AY. An approach to asymptotic robustness analysis of sequential tests for composite parametric hypotheses. *Journal of Mathematical Sciences.* 2017;227(2):196–203. DOI: 10.1007/s10958-017-3585-z.
12. Харин АЮ, Ту ТТ. О вычислении вероятностей ошибок усеченного последовательного критерия отношения вероятностей. *Журнал Белорусского государственного университета. Математика. Информатика.* 2018;2018(1):68–76.

References

1. Mukhopadhyay N, de Silva B. *Sequential methods and their applications.* Boca Raton: CRC Press; 2009. 409 p.
2. Lai TL. *Sequential analysis: some classical problems and new challenges. Statistica Sinica.* 2001;11:303–408.
3. Wald A. Sequential analysis. New York: John Wiley and Sons; 1947. 212 p.
4. Aivazian SA. Comparison of optimal properties of the tests of Neyman – Pearson and Wald. *Teoriya veroyatnostei i ee primeneniya.* 1959;4(1):86–93. Russian.
5. Huber PJ, Ronchetti EM. *Robust statistics.* New York: Wiley; 2009. 354 p.
6. Maevskii VV, Kharin YuS. Robust regressive forecasting under functional distortions in a model. *Automation and Remote Control.* 2002;63(11):1803–1820. DOI: 10.1023/A:1020959432568.
7. Kemeny JG, Snell JL. *Finite Markov Chains.* New York: D. Van Nostrand Co.; 1960. 210 p.
8. Kharin AY. *Robastnost' baiesovskikh i posledovatel'nykh statisticheskikh reshayushchikh pravil* [Robustness of Bayesian and sequential statistical decision rules]. Minsk: Belarusian State University; 2013. 207 p. Russian.
9. Kharin A, Tu TT. Performance and robustness analysis of sequential hypotheses testing for time series with trend. *Austrian Journal of Statistics.* 2017;46(3–4):23–36. DOI: 10.17713/ajs.v46i3-4.668.
10. Tu TT, Kharin AY. Sequential probability ratio test for many simple hypotheses on parameters of time series with trend. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;1:35–45. DOI: 10.33581/2520-6508-2019-1-35-45.
11. Kharin AY. An approach to asymptotic robustness analysis of sequential tests for composite parametric hypotheses. *Journal of Mathematical Sciences.* 2017;227(2):196–203. DOI: 10.1007/s10958-017-3585-z.
12. Kharin AY, Tu TT. On error probability calculation for the truncated sequential probability ratio test. *Journal of the Belarusian State University. Mathematics and Informatics.* 2018;2018(1):68–76. Russian.

Received 25.05.2021 / revised 28.06.2021 / accepted 28.06.2021.