Optimal sequential tests for multiple hypotheses when sampling from a Bernoulli population

Andrey Novikov

Metropolitan Autonomous University, Mexico City, Mexico

ARTICLE HISTORY
Compiled December 13, 2022

ABSTRACT
In this paper we deal with the problem of sequential testing of multiple hypotheses. We are interested in minimising a weighted average sample number under restrictions on the error probabilities. A computer-oriented method of construction of optimal sequential tests is proposed. For the particular case of sampling from a Bernoulli population we develop a whole set of computer algorithms for optimal design and performance evaluation of sequential tests and implement them in the form of computer code written in R programming language. The tests we obtain are exact (neither asymptotic nor approximate). Extensions to other distribution families are discussed. A numerical comparison with other known tests (of MSPRT type) is carried out.

KEYWORDS
sequential analysis; hypothesis testing; optimal stopping; optimal sequential tests; multiple hypotheses; SPRT; MSPRT

AMS CLASSIFICATION
62L10, 62L15, 62F03, 60G40, 62M02

1. Introduction

The problem of testing multiple hypotheses is one of the oldest problems in sequential analysis. It has been treated from the Bayesian viewpoint in Blackwell and Girshick (1954), Baum and Veeravalli (1994), Tartakovsky, Nikiforov, and Basseville (2015), among others. Here are some opinions about the structure of optimal Bayesian tests: “the optimal test has a very complex structure ... that makes implementation very impractical” (Baum and Veeravalli 1994), “...extremely difficult problem, so that approximations are needed” (Tartakovsky, Nikiforov, and Basseville 2015).

For these reasons, approximate solutions of the problem have been proposed. One of the widely used is the multi-hypothesis sequential probability ratio test (MSPRT) by Armitage (1950), due to its simplicity. Tartakovsky, Nikiforov, and Basseville (2015) showed that the MSPRT is asymptotically optimal, as error probabilities go to 0.

Liu, Gao, and Li (2016) studied methods of calculating performance characteristics of the MSPRT.
In Novikov (2009b) we proposed a method of construction of optimal sequential multi-hypothesis test not directly based on the Bayesian paradigm. In this paper, we generalize this method for a more general context, and apply it for construction exact (not asymptotically) optimal tests for binary (Bernoulli) observations. Also we derive computational formulas for evaluation of performance characteristics of the proposed and other multi-hypothesis tests based on sufficient statistics (in particular, MSPRTs). Based on computer implementation of the derived formulas, we make numerical comparisons of optimal tests with MSPRTs for the case of three hypotheses in the Bernoulli model.

In Section 2, we generalise the method of Novikov (2009b) for minimisation of weighted average sample number calculated at arbitrary parameter points.

In Section 3, we derive computational formulas for the Bernoulli model.

Numerical results are presented in Section 4.

Section 5 is a brief list of the results and suggestions for further work.

2. Some general results

We assume that independent and identically distributed observations \(X_1, X_2, \ldots, X_n, \ldots\) are potentially available to the statistician on the one-by-one basis, providing us with information about the unknown distribution of the data. Let us denote it \(P_\theta\), where \(\theta\) is some parameter identifying the distribution in a unique manner. We are concerned with the problem of distinguishing between a finite number of simple hypotheses \(H_1 : \theta = \theta_1, H_2 : \theta = \theta_2, \ldots, H_k : \theta = \theta_k\).

A sequential hypothesis test is a pair \(\langle \psi, \phi \rangle\) consisting of a stopping rule \(\psi = (\psi_1, \psi_2, \ldots, \psi_n, \ldots)\) and a (terminal) decision rule \(\phi = (\phi_1, \phi_2, \ldots, \phi_n, \ldots)\), where \(\psi_n = \psi_n(x_1, \ldots, x_n)\) and \(\phi_n = \phi_n(x_1, \ldots, x_n)\) are some measurable functions of observations governing the process of statistical decision-making. In their simplest (non-randomized) form they are functions taking values in \(\{0, 1\}\) and \(\{1, 2, \ldots, k\}\), respectively, indicating where to stop and which decision to take: the process is stopped when \(\psi_n(x_1, \ldots, x_n) = 1\) and hypothesis \(H_i\) is accepted whenever \(\phi_n(x_1, \ldots, x_n) = i, i = 1, \ldots, k\). If \(\psi_n(x_1, \ldots, x_n) = 0\) the process continues to the next step by taking an additional observation.

Randomised versions provide more flexibility designing the tests and define \(\psi_n\) and \(\phi_n\) as probability distributions on \(\{0, 1\}\) and \(\{1, 2, \ldots, k\}\), respectively, given the data \((x_1, \ldots, x_n)\) observed at any step. In this case, the final decision is taken according to the respective distribution, with the aid of an additional randomisation.

For simplicity, we will treat the problem in the non-randomised terms, and only present the key results in the general randomised form.

Let

\[ s_n^\psi = s_n^\psi(x_1, \ldots, x_n) = (1 - \psi_1(x_1)) \cdots (1 - \psi_{n-1}(x_1, \ldots, x_{n-1})) \psi_n(x_1, \ldots, x_n) \]

\((s_1^\psi(x_1) = \psi_1(x_1)\) by definition).

Then the stopping time \(\tau_\psi\) (number of steps of the decision process described above) is defined as the random variable taking value \(n\) whenever \(s_n^\psi = 1, n = 1, 2, \ldots,\) and
the average sample number corresponding to this is

\[ E_{\theta} \tau_\psi = \sum_{n=1}^{\infty} n E_{\theta} s_n = \sum_{n=1}^{\infty} n E_{\theta} s_n (X_1, \ldots, X_n). \]

where \( E_{\theta} \) is the symbol of mathematical expectation with respect to \( P_{\theta} \). This average, by definition, is infinite if \( P_{\theta} (\tau_\psi < \infty) < 1 \).

Other characteristics of a sequential test \( \langle \psi, \phi \rangle \) are the error probabilities defined as

\[ \alpha_{ij} (\psi, \phi) = \sum_{n=1}^{\infty} E_{\theta} s_n I_{\{\phi_n = j\}}, \quad 1 \leq i \neq j \leq k. \]

Another natural way to define error probabilities is less detailed:

\[ \alpha_i (\psi, \phi) = \sum_{n=1}^{\infty} E_{\theta_i} s_n I_{\{\phi_n \neq i\}} = \sum_{j \neq i} \alpha_{ij} (\psi, \phi), \quad 1 \leq i \leq k. \]

In the case of two hypotheses the definitions are equivalent.

For \( k = 2 \), the classical result of [Wald and Wolfowitz (1948)] states that the sequential probability ratio test (SPRT) minimises both \( E_{\theta_1} \tau_\psi \) and \( E_{\theta_2} \tau_\psi \) in the class of sequential tests \( \langle \psi, \phi \rangle \) such that

\[ \alpha_{1} (\psi, \phi) \leq \alpha_1, \quad \alpha_2 (\psi, \phi) \leq \alpha_2, \]

where \( \alpha_1 \) and \( \alpha_2 \) are the error probabilities of the SPRT.

No direct generalisations of this result exist for \( k > 2 \).

Let \( \theta_i, i = 1, \ldots, K \) be some parametric points and \( \gamma_i, i = 1, \ldots, K \) some non-negative numbers such that \( \sum_{i=1}^{K} \gamma_i = 1, K \geq 1. \)

In this paper, we are concerned with the problem of minimisation of a weighted average sample number

\[ C_{\gamma, \phi} (\psi) = \sum_{i=1}^{K} \gamma_i E_{\theta_i} \tau_\psi \]  \hfill (1)

under conditions that

\[ \alpha_{ij} (\psi, \phi) \leq \alpha_{ij}, \quad 1 \leq i, j \leq k, \quad i \neq j \]

or that

\[ \alpha_i (\psi, \phi) \leq \alpha_i, \quad 1 \leq i \leq k \]

where \( \alpha_{ij} \) and \( \alpha_i \) are some positive numbers.

The most natural context for this problem is to consider \( \theta_i = \theta_i, i = 1, \ldots, K, k = K \), assuming that \( \gamma_i \) are a priori probabilities of the hypotheses. In this setting, it is the problem of minimising the unconditional average sample number under restrictions
on the conditional probabilities of errors, and it is closely related with the Bayesian problem.

On the other hand, the well-known modified Kiefer-Weiss problem (see, for example, Lorden [1980]) also easily embeds into this scheme by taking \( \gamma_1 = 1, K = 1, \) and \( \vartheta_1 \) between the hypothesized values \( \theta_1 \) and \( \theta_2, \) being \( k = 2.\)

In this section, we adapt general results of Novikov (2009b) to the present context. First of all, it is easy to see, by usual arguments of the Lagrange multipliers method, that the problem of minimisation under conditions (2) reduces to a problem of minimisation of

\[
L(\psi, \phi) = C_{\gamma, \vartheta}(\psi) + \sum_{1 \leq i \neq j \leq k} \lambda_{ij} \alpha_{ij}(\psi, \phi),
\]

where \( \lambda_{ij} \geq 0 \) are some constant multipliers. Respectively, the problem of minimisation under conditions (3) reduces to minimisation of

\[
L(\psi, \phi) = C_{\gamma, \vartheta}(\psi) + \sum_{1 \leq i \leq k} \lambda_i \alpha_i(\psi, \phi),
\]

with multipliers \( \lambda_i, i = 1, \ldots, k. \)

It is easy to see that (5) is a particular case of (4) with \( \lambda_{ij} = \lambda_i \) for all \( j = 1, 2, \ldots, k, \) \( j \neq i, \) so in what follows we focus on the minimisation of (4).

Now, in a very standard way, it can be shown that there is a universal decision rule that minimises \( L(\psi, \phi) \) whatever fixed \( \psi. \)

It is worth noting that \( L(\psi, \phi) \) implicitly depends on the Lagrange multipliers, therefore all the constructions below will also (implicitly) depend on \( \lambda_{ij}, \) as well as on other elements of problem setting, like \( \theta_i \) and \( \vartheta_i, \) etc.

Let us assume that \( P_\theta \) is absolutely continuous with respect to a \( \sigma \)-finite measure \( \mu \) and denote \( f_\theta \) its Radon-Nikodym derivative. Also denote \( f^n_\theta = f^n_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n f_\theta(x_i), \) and let \( f^n_* = \sum_{i=1}^K \gamma_i f^n_\theta. \)

Define

\[
v_n = \min_{1 \leq j \leq k} \sum_{i \neq j} \lambda_{ij} f^n_\theta.
\]

Then the optimal decision rule is:

\[
\phi_n = j \quad \text{whenever} \quad \sum_{i \neq j} \lambda_{ij} f^n_\theta = v_n.
\]

In case there are several \( j \) for which the last equality in (7) holds, the value of \( \phi_n \) may be obtained by an arbitrary randomisation between them.

With this definition,

\[
L(\psi) = \inf_{\phi} L(\psi, \phi) = \sum_{n=1}^{\infty} \int s^n_\psi(n f^n_* + v_n) \, d\mu^n,
\]

and we have an optimal stopping problem of minimizing (8) over stopping rules \( \psi. \)
The problem is first resolved in the class of truncated tests, i.e. those not taking more than a finite number $N$ of observations. Let $S^N$ be the set of all such stopping rules that $(1 - \psi_1) \ldots (1 - \psi_N) \equiv 0$.

Let us define operator $\mathcal{I}_n$ in the following way.

For any measurable non-negative $v = v(x_1, \ldots, x_n)$ let

$$\mathcal{I}_n v = (\mathcal{I}_n v)(x_1, \ldots, x_{n-1}) = \int v(x_1, \ldots, x_n) d\mu(x_n)$$

Now, starting from

$$V_N^N \equiv v_N$$

define recursively over $n = N, N - 1, \ldots, 2$

$$V_{n-1}^N = \min\{v_{n-1}, f_n^{n-1} + \mathcal{I}_n V_n^N\}$$

Then for any $\psi \in S^N$

$$L_N(\psi) \geq 1 + \mathcal{I}_1 V_1^N,$$  \hspace{1cm} (9)

and there is an equality in (9) if for all $n = 1, 2, \ldots, N - 1$

$$\psi_n = I_{\{v_n \leq f_n^* + \mathcal{I}_n V_{n+1}^N\}}$$  \hspace{1cm} (10)

Thus, this is a stopping rule minimising $L_N(\psi)$ in $S^N$.

Any $\psi_n$ may be arbitrarily randomised between samples $(x_1, \ldots, x_n)$ for which there is an equality in the inequality under the indicator function in (10).

Under very mild conditions (see Novikov 2009b), the optimal non-truncated tests can be found passing to the limit as $N \to \infty$.

First of all, it is easy to see that $V_{n+1}^N \leq V_n^N$, so there exists $V_n = \lim_{N \to \infty} V_n^N$, $n = 1, 2, \ldots$

Then it follows from (9) that

$$L(\psi) \geq 1 + \mathcal{I}_1 V_1,$$  \hspace{1cm} (11)

and the right-hand side in (11) is attained if

$$\psi_n = I_{\{v_n \leq f_n^* + \mathcal{I}_n V_{n+1}\}}$$  \hspace{1cm} (12)

for all $n = 1, 2, \ldots$

In this way, we obtain tests $\langle \psi, \phi \rangle$ with $\psi$ satisfying (12) and $\phi$ satisfying (7) which minimise the Lagrangian function $L(\psi, \phi)$.

In particular, we obtain in this way optimal sequential tests for $k$ hypotheses in the Bayesian set-up, as explained above. Also, optimal tests in the modified Kiefer-Weiss problem can be obtained in this way (Novikov, Novikov, and Farkhshatov 2022).

Now, to obtain optimal sequential tests in the conditional setting (i.e. those minimizing $C_{\gamma, \vartheta}$ under conditions (3)) we need to find Lagrangian multipliers $\lambda_{ij}$, $1 \leq i \neq j \leq k$, providing a test (7) - (12) for which equalities in (2) hold. Respectively, minimisation of $C_{\gamma, \vartheta}$ under conditions (3) reduces to finding $\lambda_i$, $i = 1, \ldots, k$ such that
for the test in (7) - (12), with \( \lambda_{ij} = \lambda_i \) for \( 1 \leq j \neq i \leq k \), for which there are all equalities in (3).

In no way one can be sure such \( \lambda_{ij} \) exist for any combination of \( \alpha_{ij} \) (not even in the classical case of two hypotheses). On the other hand, every combination of \( \lambda_{ij} \) employed in (7) - (12), produces an optimal test in the conditional setting, if one takes its error probabilities as \( \alpha_{ij} \) in (2) (or, respectively, as \( \alpha_i \) in (3)).

3. Optimal sequential tests for sampling from a Bernoulli population

In this section, we apply the general results of Section 2 to the model of Bernoulli observations.

3.1. Construction of optimal tests

We apply the results of Section 2 to the model of observations following a Bernoulli distribution, in which case \( f_\theta(x) = \theta^x(1-\theta)^{1-x}, x = 0,1 \), and \( f_n^\theta(x_1,\ldots,x_n) = \theta^{s_n}(1-\theta)^{n-s_n} \) with \( s_n = \sum_{i=1}^n x_i \).

Let

\[
g^\theta_n(s) = \binom{n}{s} \theta^s(1-\theta)^{n-s}, \quad 0 \leq s \leq n
\]

be the probability mass function corresponding to the sufficient statistic \( S_n = \sum_{i=1}^n X_i \) (binomial distribution with parameters \( n \) and \( \theta \)). Define

\[
u_n = u_n(s) = \min_{1 \leq j \leq k} \sum_{i \neq j} \lambda_{ij}g^\theta_{ij}(s), \quad 0 \leq s \leq n
\]

and let

\[
g^\gamma_n(s) = \sum_{i=1}^K \gamma_ig^\theta_{ij}(s), \quad 0 \leq s \leq n
\]

Starting from

\[
U^N_N(s) = u_N(s), \quad 0 \leq s \leq N,
\]

define recursively for \( n = N - 1, N - 2, \ldots, 1 \)

\[
U^N_n(s) = \min\{u_n(s), g^n_n(s) + U^N_{n+1}(s)\frac{n+1-s}{n+1} + U^N_{n+1}(s+1)\frac{s+1}{n+1}\}, \quad 0 \leq s \leq n. \tag{14}
\]

**Proposition 3.1.** For all \( m = 1, 2, \ldots, N \)

\[
U^N_m(s_m) = \binom{m}{s_m}V^N_m(x_1,\ldots,x_m) \tag{15}
\]

where \( s_m = \sum_{i=1}^m x_i \).
Proof. By induction over \( m = N, N - 1, \ldots, 1 \).

It follows from \((13)\) that \((15)\) is satisfied for \( m = N \). Let us suppose that \((15)\) holds for some \( m \leq N \). Then for \( n = m - 1 \)

\[
U_n^n(s_n) = \min \{ u_n(s_n), g^n_*(s_n) + U_{n+1}^n(s_n) \frac{n - s_n}{n+1} + U_{n+1}^n(s_n + 1) \frac{s_n + 1}{n+1} \}
\]

\[
= \min \left\{ \left( n \atop s_n \right) u_n(s_n), \left( n \atop s_n \right) f^n_s(s_n) + \left( n \atop s_n \right) V_{n+1}^n(x_1, \ldots, x_n, 0) \frac{n + 1 - s_n}{n + 1} \right. \\
\left. + \left( n + 1 \atop s_n + 1 \right) V_{n+1}^n(x_1, \ldots, x_n, 1) \frac{s_n + 1}{n + 1} \right\} = \left( n \atop s_n \right) V_{n+1}^n(x_1, \ldots, x_n),
\]

that is, \((15)\) also holds for \( m - 1 \). □

It follows from Proposition \(3.1\) that the optimal decision rule \((7)\) can be expressed in terms of the sufficient statistic \( s_n \):

\[
\phi_n = j \quad \text{whenever} \quad \sum_{i:i \neq j} \lambda_{ij} g^n_{\theta_i}(s_n) = u_n(s_n), \quad (16)
\]

and also the optimal truncated stopping rule \((10)\):

\[
\psi_n = I_{\{u_n \leq g^n* + J_{n+1}U_{n+1}^n\}}(s_n), \quad (17)
\]

for \( n = 1, 2, \ldots, N - 1 \), and the optimal non-truncated one as

\[
\psi_n = I_{\{u_n \leq g^n + J_{n+1}U_{n+1}^n\}}(s_n) \quad (18)
\]

with \( U_n = \lim_{N \to \infty} U_n^n \) for all \( n = 1, 2, \ldots, \).

Formulas \((16)-(17)\) suggest a computational algorithm for evaluating the elements of optimal sequential test: start from step \( N \) calculating \( \phi_N \) for all \( 0 \leq s \leq N \) (which is based on weighted sums of binomial probabilities with parameters \( N \) and \( \theta_i, \ i = 1, 2, \ldots, k \), according to \((16)\)), and recurrently use \((11)\) for steps \( n = N - 1, N - 2, \ldots, 1 \) to calculate \( U_n^n(s) \) for all \( 0 \leq s \leq n \), marking those \( s \) for which

\[
u_n(s) > g^n_*(s) + J_{n+1}U_{n+1}^n(s)
\]

as belonging to the continuation region (by virtue of \((17)\)); for all other \( s \) storing the terminal decision based on \((16)\) as that corresponding to \( s \).

This provides a test which has an exact optimality property (neither asymptotic nor approximate) whatever be \( k \geq 2, \theta_1, \ldots, \theta_k, \gamma_1, \ldots, \gamma_K, \vartheta_1, \ldots, \vartheta_K, K \geq 1, N \geq 2 \) and Largange multipliers \( \lambda_{ij} \geq 0, 1 \leq i \neq j \leq k \).

Making \( N \) large enough we approximate the optimal non-truncated test corresponding to \((18)\).

3.2. Evaluation of performance characteristics

We derive in this part computational formulas for performance characteristics of sequential multiple hypothesis tests for the Bernoulli model.

Let \( \langle \psi, \phi \rangle \) be any sequential multiple hypothesis test based on sufficient statistics: \( \psi_n = \psi_n(s_n), \phi_n = \phi_n(s_n) \) with \( \psi \in \mathcal{S}^N \). The test \( \langle \psi, \phi \rangle \) is arbitrary but will be held
Therefore, (21) equals
\[ g_\theta^n(s; \theta) I_{\{\phi_{n-1} = j\}}(s), \quad s = 0, 1, \ldots, N, \]
and, recursively over \( n = N - 1, N - 2, \ldots, 1, \)

\[ a_j^n(s; \theta) = g_\theta^n(s) \psi_n(s) I_{\{\phi_{n-1} = j\}}(s) \]
\[ + \left( a_j^{n+1}(s; \theta) \frac{n + 1 - s}{n + 1} + a_j^{n+1}(s + 1; \theta) \frac{s + 1}{n + 1} \right) (1 - \psi_n(s)), \]

\( s = 0, 1, \ldots, n, \quad j = 1, \ldots, k. \)

Then the probability to accept hypothesis \( H_j, \) given that the true parameter is \( \theta, \) can be calculated as \( a_j^0(\theta) = a_j^1(0; \theta) + a_j^1(1; \theta). \) In particular, \( a_{ij}(\psi, \phi) = a_j^0(\theta_i), \) \( i \neq j. \)

**Proof.** Let us denote \( A_j^n = A_j^n(\psi, \phi) \) the event meaning that hypothesis \( H_j \) is accepted at or after step \( n \) (following the rules of the test \( \langle \psi, \phi \rangle), \) \( n = 1, 2, \ldots, N. \)

Let us first prove, by induction over \( n = N, N - 1, \ldots, 1, \) that

\[ a_j^n(S_n; \theta) = P_\theta(A_j^n | X_1, \ldots, X_n) g_\theta^n(S_n) \]  

(20)

For \( n = N, \) (20) follows from (19) and the definition of the decision rule \( \phi. \)

Let us suppose now that (20) holds for some \( n \leq N. \) Then

\[ a_j^{n-1}(S_{n-1}; \theta) = g_\theta^{n-1}(S_{n-1}) \psi_{n-1}(S_{n-1}) I_{\{\phi_{n-1} = j\}}(S_{n-1}) \]
\[ + \left[ a_j^n(S_{n-1}; \theta) \frac{n - S_{n-1}}{n} + a_j^n(S_{n-1} + 1; \theta) \frac{S_{n-1} + 1}{n} \right] (1 - \psi_{n-1}). \]  

(21)

But, by the supposition,

\[ \left[ a_j^n(S_{n-1}; \theta) \frac{n - S_{n-1}}{n} + a_j^n(S_{n-1} + 1; \theta) \frac{S_{n-1} + 1}{n} \right] (1 - \psi_{n-1}) \]
\[ = \left[ P_\theta(A_j^n | X_1, \ldots, X_{n-1}, X_n = 0) g_\theta^n(S_{n-1}) \frac{n - S_{n-1}}{n} \right. \]
\[ + \left. P_\theta(A_j^n | X_1, \ldots, X_{n-1}, X_n = 1) g_\theta^n(S_{n-1} + 1) \frac{S_{n-1} + 1}{n} \right] (1 - \psi_{n-1}) \]
\[ = \left[ P_\theta(A_j^n | X_1, \ldots, X_{n-1}, X_n = 0) (1 - \theta) \right. \]
\[ + \left. P_\theta(A_j^n | X_1, \ldots, X_{n-1}, X_n = 1) \theta \right] (1 - \psi_{n-1}) g_\theta^{n-1}(S_{n-1}) \]
\[ = P_\theta(A_j^n \{ \psi_{n-1} = 0 \}) | X_1, \ldots, X_{n-1} \) g_\theta^{n-1}(S_{n-1}) \]

Therefore, (21) equals

\[ (\psi_{n-1} I_{\{\phi_{n-1} = j\}} + P_\theta(A_j^n \{ \psi_{n-1} = 0 \}) | X_1, \ldots, X_{n-1} \) g_\theta^{n-1}(S_{n-1}) \]
\[ = P_\theta(\{ \psi_{n-1} = 1 \} \{ \phi_{n-1} = j \} + A_j^n \{ \psi_{n-1} = 0 \}) | X_1, \ldots, X_{n-1} \) g_\theta^{n-1}(S_{n-1}) \]
\[ = P_\theta(A_j^{n-1} | X_1, \ldots, X_{n-1}) g_\theta^{n-1}(S_{n-1}). \]
For any stopping rule $\tau$ following the stopping rule $s$, thus, Proposition 3.3.

Now that (20) is proved, we apply it for $n = 0$, \( n = 0 \), and have

$$a_j^1(1; \theta) = P_\theta(A_j^1|X_1 = 1)\theta \quad \text{and} \quad a_j^1(0; \theta) = P_\theta(A_j^1|X_1 = 0)(1 - \theta).$$

Thus,

$$a_j^1(0; \theta) + a_j^1(1; \theta) = P_\theta(A_j^1|X_1 = 1)\theta + P_\theta(A_j^1|X_1 = 0)(1 - \theta) = P_\theta(A_j^1) = a_j^0(\theta).$$

\[ \square \]

In an analogous way, characteristics of sample number can be treated.

**Proposition 3.3.** For any stopping rule $\psi$ define for any $m \geq 1$

$$b_n^m(s; \theta) = g_\theta^m(s)(1 - \psi_m(s)), \ s = 0, 1, \ldots, m, \quad (22)$$

and, recursively over $n = m - 1, m - 2, \ldots, 1$,

$$b_n^m(s; \theta) = b_{n+1}^m(b_n^m(s; \theta)\frac{n+1-s}{n+1} + b_{n+1}^m(s; \theta)\frac{s+1}{n+1}) (1 - \psi_n(s)), \quad (23)$$

$s = 0, 1, \ldots, n$. Then $P_\theta(\tau_\psi > m) = b_1^m(0; \theta) + b_1^m(1; \theta)$.

**Proof.** Let us denote $B_n^m = B_n^m(\psi), \ n = 1, 2, \ldots, m$, the event meaning that the test following the stopping rule $\psi$ does not stop at any step between $n$ and $m$, inclusively.

Let us first prove, by induction over $n = m, m-1, \ldots, 1$, that

$$b_n^m(S_n; \theta) = P_\theta(B_n^m|X_1, \ldots, X_n)g_\theta^n(S_n) \quad (24)$$

For $n = m$, (24) follows from (22).

Let us suppose now that (24) holds for some $n \leq m$. Then

$$b_{n-1}^m(S_{n-1}; \theta) = \left(b_n^m(S_{n-1}; \theta)\frac{n-S_{n-1}}{n} + b_n^m(S_{n-1} + 1; \theta)\frac{S_{n-1}+1}{n}\right)(1 - \psi_{n-1})$$

$$= \left[P_\theta(B_n^m|X_1, \ldots, X_{n-1}, X_n = 0)g_\theta^n(S_{n-1})\frac{n-S_{n-1}}{n}\right.$$

$$+ P_\theta(B_n^m|X_1, \ldots, X_{n-1}, X_n = 1)g_\theta^n(S_{n-1} + 1)\frac{S_{n-1}+1}{n}] (1 - \psi_{n-1})$$

$$= \left[P_\theta(B_n^m|X_1, \ldots, X_{n-1}, X_n = 0)(1 - \theta)$$

$$+ P_\theta(B_n^m|X_1, \ldots, X_{n-1}, X_n = 1)\theta\right](1 - \psi_{n-1})g_\theta^{n-1}(S_{n-1})$$

$$= P_\theta(B_n^m(1 - \psi_{n-1})|X_1, \ldots, X_{n-1})g_\theta^{n-1}(S_{n-1})$$

$$= P_\theta(B_{n-1}^m|X_1, \ldots, X_{n-1})g_\theta^{n-1}(S_{n-1})$$

Now that (23) is proved, we apply it for $n = 1$ and have

$$b_1^m(1; \theta) = P_\theta(B_1^m|X_1 = 1)\theta \quad \text{and} \quad b_1^m(0; \theta) = P_\theta(B_1^m|X_1 = 0)(1 - \theta),$$
thus,
\[ b_m^n(0; \theta) + b_m^n(1; \theta) = P_\theta(B_m^n|X_1 = 1)\theta + P_\theta(B_m^n|X_1 = 0)(1-\theta) = P_\theta(B_m^n) = P_\theta(\tau_\psi > m). \]

\[ \square \]

It follows from Proposition 3.3 that if \( \psi \in S^N \), then
\[ E_\theta \tau_\psi = \sum_{m=1}^{N} P_\theta(\tau_\psi \geq m) = 1 + \sum_{m=1}^{N-1} (b_m^n(0; \theta) + b_m^n(1; \theta)). \]  \hspace{1cm} (25)

If a stopping rule \( \psi \) is not truncated, we can use (25) to approximate \( E_\theta \tau_\psi \), noting that \( E_\theta \min\{\tau_\psi, N\} \to E_\theta \tau_\psi \), as \( N \to \infty \), by the theorem of monotone convergence, and \( \min\{\tau_\psi, N\} \) corresponds to the truncated rule \( \psi^N = (\psi_1, \ldots, \psi_{N-1}, 1, \ldots) \in S^N \).

Applying (25) to \( \psi^N \) we see that \( E_\theta \min\{\tau_\psi, N\} = 1 + \sum_{m=1}^{N-1} (b_m^n(0; \theta) + b_m^n(1; \theta)) \), thus
\[ E_\theta \tau_\psi = 1 + \sum_{m=1}^{\infty} (b_m^n(0; \theta) + b_m^n(1; \theta)). \]

Dealing with expectations, a more direct way to evaluate (25) is incorporating the summation in (25) into the inductive evaluations in (23). This is done in the following

**Proposition 3.4.** For a stopping rule \( \psi \), define
\[ c_N^n(s; \theta) = g_N^n(1 - \psi_N(s)), \quad s = 0, 1, \ldots, N, \]
and, recursively over \( n = N - 1, N - 2, \ldots, 1, \)
\[ c_n^n(s; \theta) = \left( g_n^n(s) + c_{n+1}^n(s; \theta) \frac{n + 1 - s}{n + 1} + c_{n+1}^n(s + 1; \theta) \frac{s + 1}{n + 1} \right) (1 - \psi_n(s)), \]
\( s = 0, 1, \ldots, n. \) Then
\[ E_\theta \min\{\tau_\psi, N + 1\} = 1 + c_1^N(0; \theta) + c_1^N(1; \theta) \]  \hspace{1cm} (26)

Again, passing to the limit in (26), as \( N \to \infty \), we obtain
\[ E_\theta \tau_\psi = 1 + \lim_{N \to \infty} (c_1^N(0; \theta) + c_1^N(1; \theta)) \]

4. Applications. Numerical results

In this Section we want to apply the theoretical results of the preceding Sections to construction of optimal tests in the Bernoulli model and numerically compare the performance of these with that of widely-used multi-hypothesis sequential probability ratio tests (MSPRT) by Armitage \((1950)\).

The MSPRT is applicable for the case when \( \vartheta_i = \theta_i, \quad i = 1, 2, \ldots, k = K. \) The idea of the MSPRT is to simultaneously run \( k(k-1)/2 \) SPRTs for each pair of the hypothesized values, stopping only when all the SPRTs decide in favour of a certain hypothesis.
Let $A_{ij} > 1$ be some constants, $1 \leq i \neq j \leq k$. Then the stopping time of the MSPRT (let us denote it $\tau^*$) is defined as

$$
\min\{n \geq 1 : \text{there is } i \text{ such that } f^n_\theta(x_1, \ldots, x_n) \geq A_{ij} f^n_\theta(x_1, \ldots, x_n) \text{ for all } j \neq i\}
$$

(27)

in which case hypothesis $H_i$ is accepted. Armitage (1950) showed that the MSPRT stops with probability one under each $H_i$, and that

$$
\alpha_{ij}^* \leq 1/A_{ji}, \quad 1 \leq i \neq j \leq k
$$

(28)

where $\alpha_{ij}^*$ is the error probability of MSPRT (27).

For $k = 2$ the MSPRT is an ordinary SPRT and (28) are the very well known Wald’s inequalities for its error probabilities.

To get numerical results we consider a particular case of $k = 3$ hypotheses for the parameter of success $\theta$ of the Bernoulli distribution, with $\theta_1 = 0.3$, $\theta_2 = 0.4$ and $\theta_3 = 0.5$.

First of all, we will be interested in calculating the performance characteristics of the MSPRT in this particular case.

It is easy to see that the rules of the MSPRT are based on the sufficient statistics $S_n, n = 1, 2, \ldots$, so the formulas of Subsection 3.2 apply for the truncated version of the MSPRT. Strictly speaking, the terminal decision at the last step, when the MSPRT is truncated at time $N$, is not defined. But we will calculate the exact probability that MSPRT does not come to a decision at any earlier stage, and make the probability of this so small (choosing $N$ large enough) that any concrete decision one can take in the last step will not affect the numerical values of the error probabilities, nor those of the average sample number under any one of the hypotheses.

In Tartakovsky, Nikiforov, and Basseville (2015), asymptotic formulas are obtained for the average sample number of the MSPRT, so we consider this example a good opportunity to juxtapose the exact and the asymptotic values of the corresponding numerical characteristics, calculated in various practical scenarios. We use the thresholds $A_{ji} = (k - 1)/\alpha$ which make the MSPRT in (27) asymptotically optimal, as $\max_i \{\alpha_i\} = \alpha \to 0$ (see Tartakovsky, Nikiforov, and Basseville 2015, Section 4.3.1).

The results of evaluations are presented in Table 1, where $\alpha_{ij}^*, E_{\theta}^\tau, \tau^*$ are the evaluated characteristics of the MSPRT, and $R_i$ the respective ratio between $E_{\theta}^\tau$ and the asymptotic expression for it (according to Tartakovsky, Nikiforov, and Basseville 2015, p. 196), $i = 1, 2, 3$.

Now, let us numerically compare the optimal multi-hypothesis test with the MSPRT, provided both have the same levels of error probabilities $\alpha_i = \alpha$, $i = 1, 2, 3$. To this end, we numerically find the Lagrange multipliers $\lambda_i$ providing the best approximation of the error probabilities of the test (7)-(10) to $\alpha$, with respect to the distance

$$
\max_i \{\alpha_i(\psi, \phi) - \alpha/\alpha\}.
$$

We use $\vartheta_i = \theta_i$ and $\gamma_i = 1/3$, for $i = 1, 2, 3$ as a criterion of minimization in (11), i.e. we evaluate the Bayesian tests with the “least informative” prior distribution. The results of fitting are presented in Table 2 (upper block).

As a competing MSPRT we take the test (27), with $A_{ij}$ defined as $A_{ij} = A_j$ for all $1 \leq j \neq i \leq 3$, and carry out the same fitting procedure as above, with respect to $A_1, A_2, A_3$. The results are presented in the middle block of Table 2.
In the lower block of Table 2 we placed the ratios $R_i$ between the average sample number of the MSPRT ($E_{\theta_i, \tau^*}$) and that of the respective Bayesian test ($E_{\theta_i, \tau}$), under each one of the hypotheses.

The results show an astonishingly high efficiency of the MSPRT, especially for small $\alpha$. This would not be so surprising for two hypotheses, because in this case any MSPRT is in fact an SPRT, and any Bayesian test is an SPRT, too (see Wald and Wolfowitz 1948), so fitting numerically both tests to given error probabilities should give a relative efficiency of about 100%. But we see that largely the same happens for three hypotheses, at least in the case of equal error probabilities we are examining.

Question arises whether there exist Bayesian tests “essentially” outperforming MSPRTs, in the case of three hypotheses. The answer is “yes”, as the following numerical example suggests.

Rather straightforwardly, we found a Bayesian test, corresponding to very “unbalanced” weights $\gamma = (0.01, 0.01, 0.98)$, and an MSPRT having the same error probabilities: $\alpha_1 = 0.0051$, $\alpha_2 = 0.089$, $\alpha_3 = 0.068$.

These correspond to Lagrangian multipliers of $\lambda_1 = 200$, $\lambda_2 = 500$, $\lambda_3 = 200$ for the Bayesian test and the thresholds $\log(A_1) = 4.90$, $\log(A_2) = 3.00$, $\log(A_3) = 1.69$ for the MSPRT, respectively.

Accordingly, we obtained $E_{\theta_1, \tau} = 320.1$, $E_{\theta_2, \tau} = 258.5$, $E_{\theta_3, \tau} = 101.3$ for the Bayesian test, and $E_{\theta_1, \tau^*} = 140.0$, $E_{\theta_2, \tau^*} = 239.1$, $E_{\theta_3, \tau^*} = 134.4$ for the MSPRT. Respectively, the weighted costs evaluated to $C_{\gamma, \theta}(\tau) = 105.07$ and $C_{\gamma, \theta}(\tau^*) = 135.46$, that is, nearly 29% higher for the MSPRT in comparison to the Bayesian test.

5. Conclusions and further work

In this paper, we proposed a numerical method of construction of optimal sequential multi-hypothesis tests, which includes both conditional and Bayesian setting, and both finite and infinite horizon.

For the particular case of sampling from a Bernoulli population, we developed a computational scheme for evaluating the optimal tests and calculating the numerical characteristics of sequential tests based on sufficient statistics. A numerical evaluation of the widely-used multi-hypothesis sequential probability ratio test is carried out for the case of three simple hypotheses about the parameter of the Bernoulli distribution, and a numerical comparison is made with the asymptotic expressions for the average

| $\alpha$ | $\alpha_1^*$ | $\alpha_2^*$ | $\alpha_3^*$ | $E_{\theta_1, \tau^*}$ | $E_{\theta_2, \tau^*}$ | $E_{\theta_3, \tau^*}$ | $R_1$ | $R_2$ | $R_3$ |
|---------|------------|------------|------------|----------------|----------------|----------------|-----|-----|-----|
| 0.1     | 0.026091   | 0.089375   | 0.029442   | 134.5          | 211.8          | 142.5          | 1.2  | 1.85| 1.26 |
| 0.05    | 0.013039   | 0.045834   | 0.014829   | 169.4          | 264.9          | 180            | 1.22 | 1.78| 1.23 |
| 0.025   | 0.006498   | 0.022826   | 0.007467   | 203.5          | 313.2          | 216.2          | 1.19 | 1.71| 1.2  |
| 0.01    | 0.002575   | 0.009172   | 0.002981   | 247.4          | 372.4          | 262.7          | 1.16 | 1.63| 1.16 |
| 0.005   | 0.001291   | 0.004596   | 0.001504   | 280            | 414.1          | 297.4          | 1.14 | 1.57| 1.15 |
| 0.002   | 0.0005     | 0.00184    | 0.000594   | 322.8          | 458.5          | 342.8          | 1.12 | 1.52| 1.13 |
| 0.001   | 0.000248   | 0.00092    | 0.000296   | 355.1          | 508.8          | 376.9          | 1.11 | 1.48| 1.11 |
| 0.0005  | 0.000123   | 0.00046    | 0.000147   | 387.2          | 548.5          | 411            | 1.1  | 1.45| 1.1  |
| 5E-07   | 1.14E-07   | 4.6E-07    | 1.47E-07   | 707.1          | 928.5          | 749.5          | 1.04 | 1.29| 1.05 |
| 5E-09   | 1.1E-09    | 4.6E-09    | 1.46E-09   | 920.3          | 1175.5         | 975.2          | 1.04 | 1.24| 1.04 |

Table 1. Average Sample Number: MSPRT vs. asymptotic
|   | 0.1  | 0.05 | 0.025 | 0.01  | 0.005 | 0.002 | 0.001 | 0.0005 |
|---|------|------|-------|-------|-------|-------|-------|--------|
| $\alpha$ | 2.21 | 2.70 | 3.00  | 3.27  | 3.63  | 3.93  | 4.22  |        |
| $\log(\lambda_1)$ | 2.56 | 2.55 | 3.13  | 3.52  | 3.81  | 4.20  | 4.51  | 4.80   |
| $\log(\lambda_2)$ | 2.27 | 2.78 | 2.75  | 3.10  | 3.37  | 3.75  | 4.04  | 4.34   |
| $\log(\lambda_3)$ | 113.5 | 160.7 | 194.4 | 242.0 | 276.1 | 319.9 | 352.8 | 385.0   |
| $E_{\theta_1 \tau}$ | 136.4 | 189.4 | 238.4 | 298.4 | 340.9 | 395.6 | 435.7 | 475.4   |
| $E_{\theta_2 \tau}$ | 116.2 | 156.6 | 204.7 | 255.1 | 301.2 | 347.2 | 372.4 | 406.6   |
| $E_{\theta_3 \tau}$ | 1.67  | 2.37  | 3.07  | 3.96  | 5.52  | 6.21  | 6.89  |        |
| $E_{\theta_1 \tau^*}$ | 2.56  | 2.55  | 3.13  | 3.52  | 3.81  | 4.20  | 4.51  | 4.80   |
| $E_{\theta_2 \tau^*}$ | 2.27  | 2.78  | 2.75  | 3.10  | 3.37  | 3.75  | 4.04  | 4.34   |
| $E_{\theta_3 \tau^*}$ | 113.5 | 160.7 | 194.4 | 242.0 | 276.1 | 319.9 | 352.8 | 385.0   |
| $R_1$ | 0.979 | 0.955 | 0.989 | 0.992 | 0.992 | 0.994 | 0.995 |        |
| $R_2$ | 0.998 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |        |
| $R_3$ | 1.018 | 1.043 | 1.011 | 1.007 | 1.007 | 1.004 | 1.005 | 1.005   |

Table 2. Efficiency of the MSPRT

sample number of the asymptotically optimal MSPRT.

For a series of error probabilities we evaluated the average sample number of the Bayesian test and compared it with that of the MSPRT having the same error probabilities, in which case the MSPRT exhibited a very high efficiency. On the other hand, we found a numerical example where the MSPRT is substantially less efficient than the optimal Bayesian test.

A very immediate extension of this work could be developing computational algorithms for construction and performance evaluation of optimal sequential multi-hypothesis tests for other parametric families, first of all for one-parameter exponential families (cf. Novikov and Farkhshatov 2022).

Another possible application is an extension of the Kiefer-Weiss (Kiefer and Weiss 1957) problem to the case of multiple hypotheses: under restrictions on error probabilities, minimize the maximum value of the average sample number, $\sup_{\theta} E_{\theta \tau}$ (cf. Novikov, Novikov, and Farkhshatov 2022).

Another expected application is an extension of sequentially planned tests for two hypotheses (Novikov 2009a) to the case of multiple hypotheses (Novikov 2022).

Acknowledgements

The author gratefully acknowledges a partial support of SNI by CONACyT (Mexico) for this work.

References

Armitage, P. 1950. “Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis.” *Journal of the Royal Statistical Society B* 12: 137–144.

Baum, C. W., and V. V. Veeravalli. 1994. “A Sequential Procedure for Multihypothesis Testing.” *IEEE Transactions on Information Theory* 40 (6): 1994–2007.
Blackwell, D., and M. A. Girshick. 1954. *Theory of games and statistical decisions*. John Wiley and Sons, Inc.

Kiefer, J., and L. Weiss. 1957. “Some properties of generalized sequential probability ratio tests.” *Annals of Mathematical Statistics* 28: 57–75.

Liu, Y., Y. Gao, and X. Rong Li. 2016. “Operating Characteristic and Average Sample Number of Binary and Multi-Hypothesis Sequential Probability Ratio Test.” *IEEE Transactions on Signal Processing* 64 (12): 3167–3179.

Lorden, G. 1980. “Structure of sequential tests minimizing an expected sample size.” *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 51 (3): 291–302.

Novikov, A. 2009a. “Optimal Sequential Multiple Hypothesis Testing in Presence of Control Variables.” *Kybernetika* 45 (3): 507–528.

Novikov, A. 2009b. “Optimal Sequential Multiple Hypothesis Tests.” *Kybernetika* 45 (2): 309–330.

Novikov, A. 2022. “Optimal design and performance evaluation of sequentially planned hypothesis tests.” [https://arxiv.org/abs/2210.07203](https://arxiv.org/abs/2210.07203).

Novikov, A., and F. Farkhshatov. 2022. “Design and performance evaluation in Kiefer-Weiss problems when sampling from discrete exponential families.” *Sequential Analysis* 41 (04): 417–434.

Novikov, A., A. Novikov, and F. Farkhshatov. 2022. “A computational approach to the Kiefer-Weiss problem for sampling from a Bernoulli population.” *Sequential Analysis* 41 (02): 198–219. [http://www.tandfonline.com/doi/abs/10.1080/07474946.2022.2070212](http://www.tandfonline.com/doi/abs/10.1080/07474946.2022.2070212)

Tartakovsky, A. G., I. V. Nikiforov, and M. Basseville. 2015. *Sequential analysis: hypothesis testing and changepoint detection*. Boca Raton, Florida: Chapman & Hall/CRC Press.

Wald, A., and J. Wolfowitz. 1948. “Optimum character of the sequential probability ratio test.” *Annals of Mathematical Statistics* 19 (3): 326–339.