

# ОБ УТОЧНЕНИИ ПРОЦЕДУРЫ БОНФЕРРОНИ МНОЖЕСТВЕННОЙ ПРОВЕРКИ ГИПОТЕЗ В СЛУЧАЕ ХИ-КВАДРАТ КРИТЕРИЕВ

И. С. Никитина

*Учреждение БГУ «Научно-исследовательский институт  
прикладных проблем математики и информатики»*

*Минск, Беларусь*

*E-mail: isdc@tut.by*

В статье приводится уточнение вероятности ошибки первого рода процедуры Бонферрони множественной проверки гипотез для критерия согласия Пирсона и критерия поиска шаблона, которые имеют хи-квадрат распределение статистик. Для учета зависимости между хи-квадрат статистиками найдены ковариации агрегированных статистик и использовано неравенство Бонферрони второго порядка, уточняющее верхнюю границу для вероятности ошибки первого рода процедуры.

*Ключевые слова:* процедура Бонферрони, хи-квадрат статистика, неравенство Бонферрони второго порядка.

## ВВЕДЕНИЕ

Многие задачи криптографии, например генерация ключей [6], предварительный анализ стойкости криптографических примитивов [8], требуют применения статистических критериев для обнаружения большого числа отклонений от гипотетической модели (например, от модели независимых симметричных испытаний Бернулли). Каждый статистический критерий предназначен, как правило, для проверки гипотезы  $H_0$  о чистоте случайности бинарной выборки и позволяет обнаруживать только определенные типы отклонений от  $H_0$ . Поэтому для проверки набора гипотез и (или) для обнаружения широкого класса альтернатив и увеличения точности обнаружения альтернативных гипотез при статистическом тестировании последовательностей следует использовать набор статистических критериев, а для принятия итогового решения — процедуры множественной проверки гипотез. Наиболее популярной процедурой множественной проверки гипотез является процедура Бонферрони [9]:

$$\text{процедура} \begin{cases} \text{принимает } H_0, \text{ если все критерии принимают } H_0; \\ \text{отвергает } H_0, \text{ если хотя бы один критерий отвергает } H_0. \end{cases} \quad (1)$$

Определение вероятности ошибки первого рода

$$\alpha = P\{\text{отвергнуть } H_0 \mid H_0 \text{ истинна}\}$$

процедуры — сложная задача, вовлекающая уровни значимости критериев  $\alpha_1, \dots, \alpha_m$ , совместное распределение статистик критериев, которое на практике известно редко.

Вследствие этого для определения вероятности ошибки первого рода либо пользуются эмпирическими методами, либо строят оценки с помощью различных неравенств.

Например, согласно неравенству Бонферрони  $\alpha \leq \alpha_1 + \dots + \alpha_m$ . Это неравенство не использует информации о зависимостях между критериями, поэтому в общем случае данная оценка сверху является грубой [9].

Задача построения более точных оценок была нами изучена для семейств критериев с нормальным распределением статистик [2]. На практике наиболее часто применяются наборы критериев Д. Кнута [1], конкурсов AES [10] и NESSIE [5], в которых большое число тестов получены из критерия согласия Пирсона и критерия поиска шаблона при различном задании параметров критериев — длин и значений шаблонов — и имеют статистики с хи-квадрат распределением. Учет зависимости между данными статистиками требует нахождения их совместного распределения вероятностей, что является затруднительным. Поэтому в данной статье предлагается использовать агрегированные статистики, полученные суммированием статистик, построенным по подвыборкам. Это позволяет аппроксимировать совместное распределение агрегированных статистик многомерным нормальным распределением и строить уточненные оценки вероятности ошибки первого рода для процедуры Бонферрони. Данный подход с использованием агрегированных статистик может быть применен к набору критериев с произвольным распределением статистик при ограниченной информации об их совместном распределении: математическом ожидании и ковариационной матрице статистик.

Статья организована следующим образом. В первом подразделе описана методика уточнения процедуры Бонферрони с использованием ограниченной информации о совместном распределении статистик — математического ожидания, дисперсий и ковариаций. Во втором и третьем подразделах приводятся параметры совместного распределения статистик критериев шаблонов по непересекающимся отрезкам и статистик критериев хи-квадрат. В четвертом подразделе приведены результаты численных экспериментов.

## МЕТОДИКА УТОЧНЕНИЯ ПРОЦЕДУРЫ БОНФЕРРОНИ

Пусть для проверки гипотезы  $H_0$  по выборке  $X$  длины  $n$  используются  $m$  критериев  $C_1, \dots, C_m$  со статистиками  $S_1, \dots, S_m$ . Пусть при истинной нулевой гипотезе известно математическое ожидание  $\mu_1, \dots, \mu_m$  и ковариационная матрица  $\Sigma = (\sigma_{ij})_{i,j=1}^m$  статистик критериев, которая является невырожденной.

Будем применять критерии по следующей схеме. Разобьем исходную выборку на  $K$  подвыборок  $X^{(1)}, \dots, X^{(K)}$  длины  $N$  каждая. Обозначим  $S_i^{(k)}$  — статистика, полученная применением критерия  $C_i$  к подвыборке  $X^{(k)}$ . Построим агрегированные статистики:

$$\tilde{S}_i = \frac{S_i^{(1)} + \dots + S_i^{(K)} - K\mu_i}{\sqrt{K}}. \quad (2)$$

**Утверждение.** Предельным распределением при  $K \rightarrow \infty$  вектора  $\tilde{S} = (\tilde{S}_1, \dots, \tilde{S}_m)'$  при истинной гипотезе  $H_0$  является  $m$ -мерное нормальное распределение с нулевым вектором математического ожидания и ковариационной матрицей  $\Sigma$ .

**Доказательство.** Утверждение прямо следует из применения многомерного случая центральной предельной теоремы [4] к вектору  $\tilde{S} = (\tilde{S}_1, \dots, \tilde{S}_m)'$ . ■

Теперь, зная распределение агрегированных статистик при нулевой гипотезе, могут быть построены критерии (односторонние или двусторонние) на их основе:

$$\tilde{C}_i^{\text{одн}} \begin{cases} \text{принимает } H_0 \text{ на уровне значимости } \alpha_i, \text{ если } \tilde{S}_i / \sqrt{\sigma_{ii}} < \Delta_i^{\text{одн}} = \Phi^{-1}(1 - \alpha_i); \\ \text{иначе.} \end{cases} \quad (3)$$

$$\tilde{C}_i^{\text{дв}} \begin{cases} \text{принимает } H_0 \text{ на уровне значимости } \alpha_i, \text{ если } |\tilde{S}_i / \sqrt{\sigma_{ii}}| < \Delta_i^{\text{дв}} = \Phi^{-1}(1 - \alpha_i / 2); \\ \text{иначе.} \end{cases} \quad (4)$$

Для процедуры (1), построенной на основе агрегированных статистик (2) и критериев (3), (4), возможно уточнение оценки вероятности ошибки первого рода за счет использования информации о попарном совместном распределении статистик.

**Теорема 1.** Для вероятности ошибки первого рода процедуры (1) с критериями (3) или (4) справедливы следующие оценки сверху соответственно:

$$\alpha \leq \sum_{i=1}^m \alpha_i - \max_{1 \leq j \leq m} \sum_{i=1, i \neq j}^m (F_{ij}(\Delta_i^{\text{одн}}, \Delta_j^{\text{одн}}) + \alpha_i + \alpha_j - 1), \quad (5)$$

$$\alpha \leq \sum_{i=1}^m \alpha_i - \max_{1 \leq j \leq m} \sum_{i=1, i \neq j}^m (g_{ij}(\Delta_i^{\text{дв}}, \Delta_j^{\text{дв}}) + \alpha_i + \alpha_j - 1),$$

$$g_{ij}(\Delta_i^{\text{дв}}, \Delta_j^{\text{дв}}) = F_{ij}(-\Delta_i^{\text{дв}}, -\Delta_j^{\text{дв}}) - F_{ij}(-\Delta_i^{\text{дв}}, \Delta_j^{\text{дв}}) - F_{ij}(\Delta_i^{\text{дв}}, -\Delta_j^{\text{дв}}) + F_{ij}(\Delta_i^{\text{дв}}, \Delta_j^{\text{дв}}),$$

где  $F_{ij}$  — функция распределения двумерного нормального закона с нулевым математическим ожиданием, единичными дисперсиями и ковариацией  $\rho_{ij} = \sigma_{ij} / \sqrt{\sigma_{ii}\sigma_{jj}}$ , которая является маргинальной функцией распределения вектора  $(\tilde{S}_i / \sqrt{\sigma_{ii}}, \tilde{S}_j / \sqrt{\sigma_{jj}})'$ .

**Доказательство.** Проведем доказательство для случая односторонних критериев в предположении истинности нулевой гипотезы.

Вероятность ошибки первого рода можно представить в виде  $\alpha = P\left\{\bigcup_{i=1}^m \{\tilde{S}_i / \sqrt{\sigma_{ii}} \geq \Delta_i^{\text{одн}}\}\right\}$ . Применив к этому выражению неравенство Бонферрони второго порядка [7], получаем:

$$\alpha \leq \sum_{i=1}^m P\{\tilde{S}_i / \sqrt{\sigma_{ii}} \geq \Delta_i^{\text{одн}}\} - \max_{1 \leq j \leq m} \sum_{i=1, i \neq j}^m P\{\tilde{S}_i / \sqrt{\sigma_{ii}} \geq \Delta_i^{\text{одн}}, \tilde{S}_j / \sqrt{\sigma_{jj}} \geq \Delta_j^{\text{одн}}\}.$$

Так как величины  $\tilde{S}_i / \sqrt{\sigma_{ii}}$  имеют стандартное нормальное распределение, то

$$P\{\tilde{S}_i / \sqrt{\sigma_{ii}} \geq \Delta_i^{\text{одн}}\} = 1 - \Phi(\Delta_i^{\text{одн}}),$$

$$P\{\tilde{S}_i / \sqrt{\sigma_{ii}} \geq \Delta_i^{\text{одн}}, \tilde{S}_j / \sqrt{\sigma_{jj}} \geq \Delta_j^{\text{одн}}\} = F_{ij}(+\infty, +\infty) - F_{ij}(+\infty, \Delta_j^{\text{одн}}) - F_{ij}(\Delta_i^{\text{одн}}, +\infty) + F_{ij}(\Delta_i^{\text{одн}}, \Delta_j^{\text{одн}}) = 1 - \Phi(\Delta_j^{\text{одн}}) - \Phi(\Delta_i^{\text{одн}}) + F_{ij}(\Delta_i^{\text{одн}}, \Delta_j^{\text{одн}}).$$

Утверждение теоремы получается, подставив значения порогов  $\Delta_i^{\text{одн}}$ .

Доказательство для случая двусторонних критериев проводится аналогично. ■

**Замечание.** Значения двумерной функции распределения  $F_{ij}$  легко могут быть вычислены с помощью следующего разложения в ряд [3]:

$$F_{ij}(a, b) = \Phi(a)\Phi(b) + \sum_{n=0}^{+\infty} \frac{\phi^{(n)}(a)\phi^{(n)}(b)}{(n+1)!} \rho_{ij}^{n+1},$$

где  $\phi^{(n)}$  —  $n$ -ая производная плотности стандартного нормального распределения.

## КРИТЕРИИ КВАДРАТОВ ЧАСТОТ ШАБЛОНОВ ПО НЕПЕРЕСЕКАЮЩИМСЯ ОТРЕЗКАМ

Рассмотрим критерии со статистиками

$$S_{h,l_j} = \frac{(v_{h,l_j} - K_j/2^{l_j})^2}{K_j/2^{l_j}(1-1/2^{l_j})}, \quad v_{h,l_j} = \sum_{k=1}^{K_j} I\{x_{k,l_j} = h_{l_j}\},$$

где  $v_{h,l_j}$  — частота встречаемости шаблона  $h_{l_j}$  длины  $l_j$  в выборке,  $K_j$  — число отрезков длины  $l_j$  в выборке,  $x_{k,l_j}$  —  $k$ -й отрезок длины  $l_j$  в выборке. При исследовании распределения двух статистик шаблонов длины  $l_1$  и  $l_2$  будем предполагать, что длина выборки кратна наименьшему общему кратному длин шаблонов, то есть  $n = \text{ВНОК}(l_1, l_2)$ .

Легко видеть, что при истинной гипотезе  $H_0$  статистики  $S_{h,l_j}$  имеют предельное при  $n \rightarrow \infty$  хи-квадрат распределение с 1 степенью свободы, а совместное распределение статистик для различных шаблонов неизвестно. В следующей теореме приводятся параметры совместного распределения статистик критериев.

**Теорема 2.** При истинной гипотезе  $H_0$  статистики  $S_{h,l_j}$  имеют следующие параметры распределения:  $E\{S_{h,l_j}\} = 1$ ,

$$\sigma_{ij}(n) = \text{Cov}\{S_{h,l_1}, S_{h,l_2}\} = \frac{2}{A_1 A_2 (2^{l_1} - 1)(2^{l_2} - 1)} \left( \sum_{k=1}^{|A^{1,2}|-1} 2^{a_{k+1}^{1,2} - a_k^{1,2}} p_k + 1 - A_1 - A_2 \right)^2 + O(n),$$

где  $A_i = \frac{\text{НОК}(l_1, l_2)}{l_i}$ ,  $i = 1, 2$ ,  $a_k^{i,j}$  —  $k$ -й по величине член множества

$$A^{i,j} = \{0, l_j, 2l_j, \dots, A_i l_j, l_j, 2l_j, \dots, (A_j - 1)l_j\}, \quad a_k^{i,j} < a_{k+1}^{i,j}, \quad |A^{i,j}| = A_i + A_j,$$

$p_k$  — индикатор совпадения подстроки шаблона  $h_{l_1}$ , начинающейся с индекса  $a_k^{1,2} \bmod l_1 + 1$  и заканчивающейся индексом  $(a_{k+1}^{1,2} - 1) \bmod l_1 + 1$ , с подстрокой шаблона  $h_{l_2}$  с индекса  $a_k^{1,2} \bmod l_2 + 1$  по  $(a_{k+1}^{1,2} - 1) \bmod l_2 + 1$ . Предельные при  $n \rightarrow \infty$  параметры распределения имеют следующий вид:

$$\sigma_{ij}(n) \xrightarrow{n \rightarrow \infty} \sigma_{ij} = \frac{2}{A_1 A_2 (2^{l_1} - 1)(2^{l_2} - 1)} \left( \sum_{k=1}^{|A^{1,2}|-1} 2^{a_{k+1}^{1,2} - a_k^{1,2}} p_k + 1 - A_1 - A_2 \right)^2,$$

$$\sigma_{ii} = D\{S_{h,l_i}\} = 2.$$

**Доказательство.** Найдем математическое ожидание статистики

$$E\{S_{h,l_j}\} = \frac{1}{K_j/2^{l_j}(1-1/2^{l_j})} E\{(v_{h,l_j} - K_j/2^{l_j})^2\}.$$

Раскроем скобки под знаком математического ожидания, учтем, что  $E\{v_{h,l_j}\} = K_j/2^{l_j}$  и представим частоты в виде индикаторных функций:

$$E\{S_{h,l_j}\} = \frac{1}{K_j/2^{l_j}(1-1/2^{l_j})} \left( E\left\{ \sum_{k=1}^{K_j} \sum_{l=1}^{K_j} I\{x_{k,l_j} = h_{l_j}\} I\{x_{l,l_j} = h_{l_j}\} \right\} - K_j^2/2^{2l_j} \right).$$

Разделив сумму по  $k, l$  на две, если  $k = l$  и если  $k \neq l$ , получаем, что

$$E\{S_{h,j}\} = \frac{1}{K_j/2^{l_j}(1-1/2^{l_j})} \left( K_j/2^{l_j} + K_j(K_j-1)/2^{2l_j} - K_j^2/2^{2l_j} \right) = 1.$$

Ковариацию будем находить по определению:

$$\text{Cov}\{S_{h,j_1}, S_{h,j_2}\} = E\{S_{h,j_1}S_{h,j_2}\} - E\{S_{h,j_1}\}E\{S_{h,j_2}\} = E\{S_{h,j_1}S_{h,j_2}\} - 1.$$

Раскрыв скобки под знаком математического ожидания, получаем, что

$$\begin{aligned} \text{Cov}\{S_{h,j_1}, S_{h,j_2}\} &= \frac{1}{K_1K_2/2^{l_1+l_2}(1-1/2^{l_1})(1-1/2^{l_2})} \left( E\{v_{h,j_1}^2 v_{h,j_2}^2\} - 2K_2/2^{l_2} E\{v_{h,j_1}^2 v_{h,j_2}\} - \right. \\ &\quad \left. - 2K_1/2^{l_1} E\{v_{h,j_1} v_{h,j_2}^2\} + 4K_1K_2/2^{l_1+l_2} E\{v_{h,j_1} v_{h,j_2}\} + K_1^2/2^{2l_1} E\{v_{h,j_2}^2\} + \right. \\ &\quad \left. + K_2^2/2^{2l_2} E\{v_{h,j_1}^2\} - 3K_1^2K_2^2/2^{2l_1+2l_2} \right) \end{aligned}$$

Найдем полученные слагаемые:

$$\begin{aligned} E\{v_{h,j_1}^2 v_{h,j_2}^2\} &= E\left\{ \sum_{k=1}^{K_1} \sum_{l=1}^{K_1} \sum_{m=1}^{K_2} \sum_{n=1}^{K_2} \mathbf{I}\{x_{k,j_1} = h_{l,j_1}\} \mathbf{I}\{x_{l,j_1} = h_{l,j_1}\} \mathbf{I}\{x_{m,j_2} = h_{j,j_2}\} \mathbf{I}\{x_{n,j_2} = h_{j,j_2}\} \right\} = \\ &= E\left\{ \sum_{k=1}^{K_1} \sum_{m=1}^{K_2} \mathbf{I}\{x_{k,j_1} = h_{l,j_1}\} \mathbf{I}\{x_{m,j_2} = h_{j,j_2}\} \right\} + \\ &\quad + E\left\{ \sum_{k=1}^{K_1} \sum_{l=1, l \neq k}^{K_1} \sum_{m=1}^{K_2} \mathbf{I}\{x_{k,j_1} = h_{l,j_1}\} \mathbf{I}\{x_{l,j_1} = h_{l,j_1}\} \mathbf{I}\{x_{m,j_2} = h_{j,j_2}\} \right\} + \\ &\quad + E\left\{ \sum_{k=1}^{K_1} \sum_{m=1}^{K_2} \sum_{n=1, n \neq m}^{K_2} \mathbf{I}\{x_{k,j_1} = h_{l,j_1}\} \mathbf{I}\{x_{m,j_2} = h_{j,j_2}\} \mathbf{I}\{x_{n,j_2} = h_{j,j_2}\} \right\} + \\ &\quad + E\left\{ \sum_{k=1}^{K_1} \sum_{l=1, l \neq k}^{K_1} \sum_{m=1}^{K_2} \sum_{n=1, n \neq m}^{K_2} \mathbf{I}\{x_{k,j_1} = h_{l,j_1}\} \mathbf{I}\{x_{l,j_1} = h_{l,j_1}\} \mathbf{I}\{x_{m,j_2} = h_{j,j_2}\} \mathbf{I}\{x_{n,j_2} = h_{j,j_2}\} \right\} \end{aligned}$$

Найдем первую сумму. Разделим ее на две группы слагаемых:

$$\begin{aligned} E\left\{ \sum_{k=1}^{K_1} \sum_{m=1}^{K_2} \mathbf{I}\{x_{k,j_1} = h_{l,j_1}\} \mathbf{I}\{x_{m,j_2} = h_{j,j_2}\} \right\} &= \sum_{k=1}^{K_1} \sum_{m: x_{k,j_1} \cap x_{m,j_2}} E\left\{ \mathbf{I}\{x_{k,j_1} = h_{l,j_1}\} \mathbf{I}\{x_{m,j_2} = h_{j,j_2}\} \right\} + \\ &\quad + \sum_{k=1}^{K_1} \sum_{m \in \{1, \dots, K_2\} \setminus \{m: x_{k,j_1} \cap x_{m,j_2}\}} E\left\{ \mathbf{I}\{x_{k,j_1} = h_{l,j_1}\} \mathbf{I}\{x_{m,j_2} = h_{j,j_2}\} \right\} \end{aligned}$$

Первая группа слагаемых состоит из  $B$  одинаковых групп слагаемых, соответствующих отрезкам выборки длины  $\text{НОК}(l_1, l_2)$ . Число пересекающихся пар  $x_{k,j_1}, x_{m,j_2}$  равно  $B(A_1 + A_2 - 1) = K_1 + K_2 - B$ , следовательно, число непересекающихся равно  $K_1K_2 - (K_1 + K_2 - B)$ . Тогда

$$\begin{aligned} E\left\{ \sum_{k=1}^{K_1} \sum_{m=1}^{K_2} \mathbf{I}\{x_{k,j_1} = h_{l,j_1}\} \mathbf{I}\{x_{m,j_2} = h_{j,j_2}\} \right\} &= B \sum_{k=1}^{\lfloor A^{1,2} \rfloor - 1} 2^{a_{k,A}^{1,2} - a_k^{1,2}} p_k \frac{1}{2^{l_1+l_2}} + \\ &\quad + (K_1K_2 - K_1 - K_2 + B) \frac{1}{2^{l_1+l_2}}. \end{aligned}$$

Аналогичным образом находим остальные слагаемые, входящие в  $E\{v_{h,j_1}^2 v_{h,j_2}^2\}$ , а потом и в  $E\{S_{h,j_1}S_{h,j_2}\}$ . Далее, собрав слагаемые одного порядка малости, получаем утверждение теоремы. ■

Агрегированные статистики (2) для критериев квадратов частот шаблонов имеют вид  $\tilde{S}_{h,j} = \frac{S_{h,j}^{(1)} + \dots + S_{h,j}^{(K)} - K}{\sqrt{K}}$  и, согласно теореме 1, имеют предельное при  $K, N \rightarrow \infty$  нормальное распределение с нулевым вектором математического ожидания и ковариационной матрицей  $\Sigma = (\sigma_{ij})_{i,j=1}^m$ , найденной в теореме 2. Эти параметры могут использовать-

ся в теореме 1 для уточнения вероятности ошибки первого рода процедуры Бонферрони с односторонними критериями (3) со статистиками  $S_{h,j_i}$ .

## КРИТЕРИИ ХИ-КВАДРАТ

Приведем аналогичные результаты для критериев хи-квадрат со статистиками  $S_{\chi^2,j_i} = \sum_{j=1}^{2^{l_i}} \frac{(v_{h,j_i} - K_i / 2^{l_i})^2}{K_i / 2^{l_i}}$ .

**Теорема 3.** При истинной гипотезе  $H_0$  статистики  $S_{\chi^2,j_i}$  имеют следующие параметры распределения:

$$E\{S_{\chi^2,j_i}\} = 2^{l_i} - 1, \quad \sigma_{ij}(n) = \text{Cov}\{S_{\chi^2,j_i}, S_{\chi^2,j_l}\} = \frac{2}{A_i A_j} \left( \sum_{k=1}^{|A^{l_i,l_j}|-1} 2^{a_{k+1}^{l_i,l_j} - a_k^{l_i,l_j}} + 1 - A_i - A_j \right) \frac{B-1}{B}.$$

Предельные при  $n \rightarrow \infty$  параметры распределения имеют следующий вид:

$$\sigma_{ij}(n) \xrightarrow{n \rightarrow \infty} \sigma_{ij} = \frac{2}{A_i A_j} \left( \sum_{k=1}^{|A^{l_i,l_j}|-1} 2^{a_{k+1}^{l_i,l_j} - a_k^{l_i,l_j}} + 1 - A_i - A_j \right), \quad \sigma_{ii} = D\{S_{\chi^2,j_i}\} = 2(2^{l_i} - 1).$$

**Доказательство.** Проведем доказательство для предельного значения  $\sigma_{12}$ .

Легко убедиться непосредственным вычислением, что

$$\text{Cov}\{S_{\chi^2,j_1}, S_{\chi^2,j_2}\} = \frac{(2^{l_1} - 1)(2^{l_2} - 1)}{2^{l_1+l_2}} \sum_{i=1}^{2^{l_1}} \sum_{j=1}^{2^{l_2}} \text{Cov}\{S_{h,j_1}, S_{h,j_2}\}.$$

Тогда, используя результаты теоремы 2, получаем

$$\sigma_{12} = \frac{2}{2^{l_1+l_2} A_1 A_2} \sum_{i=1}^{2^{l_1}} \sum_{j=1}^{2^{l_2}} \left( \sum_{k=1}^{|A^{1,2}|-1} 2^{a_{k+1}^{1,2} - a_k^{1,2}} p_k + 1 - A_1 - A_2 \right)^2.$$

Раскрыв квадрат, и учитывая, что из  $2^{l_1+l_2}$  пар шаблонов индикатор  $p_k = 1$  только для  $2^{l_1+l_2} - (a_{k+1}^{1,2} - a_k^{1,2})$  пар, получаем:

$$\begin{aligned} \sigma_{12} &= \frac{2}{2^{l_1+l_2} A_1 A_2} \sum_{i=1}^{2^{l_1}} \sum_{j=1}^{2^{l_2}} \left( \left( \sum_{k=1}^{|A^{1,2}|-1} 2^{a_{k+1}^{1,2} - a_k^{1,2}} p_k \right)^2 + 2(1 - A_1 - A_2) \left( \sum_{k=1}^{|A^{1,2}|-1} 2^{a_{k+1}^{1,2} - a_k^{1,2}} p_k \right) + \right. \\ &\quad \left. + (1 - A_1 - A_2)^2 \right) = \\ &= \frac{2}{2^{l_1+l_2} A_1 A_2} \left( \sum_{i=1}^{2^{l_1}} \sum_{j=1}^{2^{l_2}} \left( \sum_{k=1}^{|A^{1,2}|-1} 2^{a_{k+1}^{1,2} - a_k^{1,2}} p_k + \sum_{k=1}^{|A^{1,2}|-1} \sum_{l=1, l \neq k}^{|A^{1,2}|-1} 2^{(a_{k+1}^{1,2} - a_k^{1,2}) + (a_{l+1}^{1,2} - a_l^{1,2})} p_k p_l \right) \right) + \\ &\quad + \frac{2}{2^{l_1+l_2} A_1 A_2} \left( 2(1 - A_1 - A_2) \left( \sum_{k=1}^{|A^{1,2}|-1} 2^{a_{k+1}^{1,2} - a_k^{1,2}} 2^{l_1+l_2 - (a_{k+1}^{1,2} - a_k^{1,2})} \right) + 2^{l_1+l_2} (1 - A_1 - A_2)^2 \right). \end{aligned}$$

Так же можно убедиться, что произведение индикаторов  $p_k p_l = 1$  для  $2^{l_1+l_2} - (a_{k+1}^{1,2} - a_k^{1,2}) - (a_{l+1}^{1,2} - a_l^{1,2})$  пар шаблонов. Поэтому

$$\begin{aligned} \sigma_{12} &= \frac{2}{2^{l_1+l_2} A_1 A_2} \left( \sum_{k=1}^{|A^{1,2}|-1} 2^{2(a_{k+1}^{1,2} - a_k^{1,2}) + l_1+l_2 - (a_{k+1}^{1,2} - a_k^{1,2})} + \right. \\ &\quad \left. + \sum_{k=1}^{|A^{1,2}|-1} \sum_{l=1, l \neq k}^{|A^{1,2}|-1} 2^{(a_{k+1}^{1,2} - a_k^{1,2}) + (a_{l+1}^{1,2} - a_l^{1,2}) + l_1+l_2 - (a_{k+1}^{1,2} - a_k^{1,2}) - (a_{l+1}^{1,2} - a_l^{1,2})} \right) + \end{aligned}$$

$$\begin{aligned}
& + \frac{2}{2^{l_1+l_2} A_1 A_2} \left( 2(1-A_1-A_2) \left( \sum_{k=1}^{|A_1|^{l_2}-1} 2^{l_1+l_2} \right) + 2^{l_1+l_2} (1-A_1-A_2)^2 \right) = \\
& = \frac{2}{A_1 A_2} \left( \sum_{k=1}^{|A_1|^{l_2}-1} 2^{a_{k+1}^{l_2}-a_k^{l_2}} + (A_1+A_2-1)(A_1+A_2-2) + 2(1-A_1-A_2)(A_1+A_2-1) + (1-A_1-A_2)^2 \right) = \\
& = \frac{2}{A_1 A_2} \left( \sum_{k=1}^{|A_1|^{l_2}-1} 2^{a_{k+1}^{l_2}-a_k^{l_2}} - (A_1+A_2-1) \right).
\end{aligned}$$

Обобщенные статистики (2) для критериев хи-квадрат имеют вид  $\tilde{S}_{\chi^2, l_i} = \frac{S_{\chi^2, l_i}^{(1)} + \dots + S_{\chi^2, l_i}^{(K)} - K(2^{l_i} - 1)}{\sqrt{K}}$  и, согласно теореме 1, имеют предельное при  $K, N \rightarrow \infty$  нормальное распределение с нулевым вектором математического ожидания и ковариационной матрицей  $\Sigma = (\sigma_{ij})_{i,j=1}^m$ , найденной в теореме 3.

Процедура проверки нулевой гипотезы батарей критериев хи-квадрат строится из односторонних критериев. При построении итогового решения используются предельные значения дисперсий  $\sigma_{ii} = 2(2^{l_i} - 1)$ , а для вычисления оценки сверху для вероятности ошибки первого рода (5) используются корреляции между статистиками  $\sigma_{ij} / \sqrt{\sigma_{ii} \sigma_{jj}}$ , входящие в функции распределения  $F_{ij}$ .

## ЧИСЛЕННЫЙ ЭКСПЕРИМЕНТ

Проводилось тестирование бинарной выборки объема  $n = 204800$  бит с помощью одновременного применения четырех критериев хи-квадрат с длинами  $l_1 = 1, l_2 = 2, l_3 = 3, l_4 = 4$ . Опираясь на теорему 3, была найдена корреляционная матрица статистик критериев:

|        |        |        |        |
|--------|--------|--------|--------|
| 1      | 0.5774 | 0.378  | 0.2582 |
| 0.5774 | 1      | 0.291  | 0.4472 |
| 0.378  | 0.291  | 1      | 0.1789 |
| 0.2582 | 0.4472 | 0.1789 | 1      |

С помощью теоремы 1 была найдена оценка сверху для вероятности ошибки первого рода процедуры  $\alpha \leq \alpha_+ = 0.167674$  для значений  $\alpha_i = 0.05, i = \overline{1,4}$ , в то время как оценка сверху с помощью неравенства Бонферрони первого порядка равна 0.2.

Был проведен численный эксперимент, в ходе которого оценивалась вероятность ошибки первого рода процедуры Бонферрони по 12169 независимым повторениям процедуры и значению  $K = 100$ . Полученная оценка вероятности ошибки первого рода  $\hat{\alpha} = 0.16797$ , 95-процентный доверительный интервал для этой оценки равен  $[0.161325717; 0.174609857]$ .

Оценка  $\hat{\alpha}$ , полученная в ходе численного эксперимента, вследствие зависимостей между статистиками критериев существенно меньше верхней границы, полученной из неравенства Бонферрони. Как видно, использование результатов теоремы 1 позволяет учесть эти зависимости при оценке вероятности ошибки первого рода: величина  $\alpha_+$  по-

падает в доверительный интервал, построенный для практической оценки вероятности ошибки первого рода.

## ЗАКЛЮЧЕНИЕ

В данной работе рассмотрена задача уточнения вероятности ошибки первого рода процедуры Бонферрони за счет учета зависимостей между статистиками критериев хи-квадрат. Для ее решения предложена методика построения агрегированных статистик с известным предельным распределением, найдены необходимые для их вычисления параметры совместного распределения статистик критериев хи-квадрат и критериев поиска шаблонов. Данные, полученные в ходе численных экспериментов, согласуются с теоретическими результатами.

## ЛИТЕРАТУРА

1. Кнут Д. Искусство программирования для ЭВМ. Т.3. Сортировка и поиск / Д. Кнут. М.: Издательский дом «Вильямс», 2000. 824 с.
2. Костевич А.Л., Милованова И.С. Улучшенная процедура Бонферрони множественной проверки гипотез для семейства критериев с нормальным распределением статистик / А.Л. Костевич, И.С. Милованова // Материалы VI Международной Петрозаводской конференции «Вероятностные методы в дискретной математике», 2004.
3. Прохоров Ю.В. Вероятность и математическая статистика: Энциклопедия / Ю.В. Прохоров. М.: Большая Российская Энциклопедия, 1999. 910 с.
4. Феллер В. Введение в теорию вероятностей и ее приложения. В 2-х томах. Т.2 / В. Феллер. М.: Мир, 1984. 738 с.
5. Dichtl M. Descriptions of General NESSIE Test Tools / M. Dichtl. NESSIE Document NES/DOC/SAG/WP2/023/2, 2001.
6. FIPS 140-2. Security Requirements for Cryptographic Modules, 2001.
7. Kounias E., Bounds for the probability of a union, with applications / E. Kounias. // Ann. Math. Statist. 1968. Vol. 39. P. 2154-2158.
8. NIST Special Publication 800-22 A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications, 2000.
9. Simes R.J. An improved Bonferroni procedure for multiple tests of significance / R.J. Simes // Biometrika. 1986. Vol. 73. P.751-754.
10. Soto J. Randomness Testing of the AES Candidate Algorithms /J. Soto. NIST AES report. 1999.