КОСТЕВИЧ А.Л., НИКИТИНА И.С.

УТОЧНЕНИЕ ПРОЦЕДУРЫ БОНФЕРРОНИ МНОЖЕСТВЕННОЙ ПРОВЕРКИ ГИПОТЕЗ В СЛУЧАЕ ЗАВИСИМОСТИ МЕЖДУ КРИТЕРИЯМИ

Аннотация

Построена уточненная процедура Бонферрони множественной проверки гипотез, учитывающая зависимости между статистиками критериев, и указано ее применение совместно с другими распространенными процедурами множественной проверки гипотез.

1. Введение

Рассмотрим задачу проверки по одной выборке "объединенной" гипотезы $\mathcal{H}_0$, являющейся пересечением m индивидуальных статистических гипотез $\mathcal{H}_{0,1}, \ldots, \mathcal{H}_{0,m}$: $\mathcal{H}_0 = \cap_{i=1}^{m} \mathcal{H}_{0,i}$. В общем случае, ввиду сложности параметрического описания гипотезы $\mathcal{H}_0$ и наличия пересекающихся или вложенных индивидуальных гипотез, построение одного статистического критерия для проверки "объединенной" гипотезы $\mathcal{H}_0$ является затруднительным. Поэтому используются процедуры множественной проверки гипотез, принимающие решение о гипотезе $\mathcal{H}_0$ по результатам проверок гипотез $\mathcal{H}_{0,1}, \ldots, \mathcal{H}_{0,m}$ против возможных альтернатив $\mathcal{H}_{1,1}, \ldots, \mathcal{H}_{1,m}$ с помощью индивидуальных статистических критериев [3].

Описанная задача множественной проверки гипотез возникает в различных областях, например, в медицине, генетике [6], а также в криптографии при анализе выходных последовательностей криптографических алгоритмов и генераторов случайных чисел для выявления альтернатив $\mathcal{H}_{1,1}, \ldots, \mathcal{H}_{1,m}$ для нулевой гипотезы $\mathcal{H}_0$ о том, что выборка порождена последовательностью независимых симметричных испытаний Бернулли.

Наиболее распространенной процедурой множественной проверки гипотез является процедура Бонферрони [3, 6]. Так, она использовалась на конкурсах NESSIE и AES при статистическом анализе выходных последовательностей криптографических алгоритмов для проверки "объединенной" гипотезы $\mathcal{H}_0$ [5].

Известно [11], что данная процедура обладает рядом недостатков: вероятность ошибки первого рода процедуры много меньше уровня значимости в случае значимой корреляции между статистиками используемых критериев и, как следствие, мощность процедуры является заниженной. В данной статье построим уточненную процедуру Бонферрони, учитывающую зависимости между статистиками критериев, и укажем ее применение в других процедурах множественной проверки гипотез.

2. Математическая модель

Пусть регистрируется выборка X. Пусть имеется m критериев $C_1, \ldots, C_m$ для проверки нулевых гипотез $\mathcal{H}_{0,1}, \ldots, \mathcal{H}_{0,m}$ против альтернатив $\mathcal{H}_{1,1}, \ldots, \mathcal{H}_{1,m}$ соответственно. Под множественной проверкой гипотез понимают проверку "объединенной" нулевой гипотезы $\mathcal{H}_0$ против "объединенной" альтернативы $\mathcal{H}_1$:

$$\mathcal{H}_0 = \bigcap_{i=1}^{m} \mathcal{H}_{0,i}, \qquad \mathcal{H}_1 = \bigcup_{i=1}^{m} \mathcal{H}_{1,i}.$$

Пусть для проверки гипотез $\mathcal{H}_{0,i}$ используются двусторонние критерии C_i, основанные на статистиках S_i, имеющих абсолютно непрерывные распределения:

$$\text{принимается } \begin{cases} \mathcal{H}_{0,i}, & \text{если } |S_i| \le \Delta_i, \\ \mathcal{H}_{1,i}, & \text{иначе,} \end{cases} \quad \text{или} \quad \begin{cases} \mathcal{H}_{0,i}, & \text{если } P_i \ge \alpha_i, \\ \mathcal{H}_{1,i}, & \text{иначе,} \end{cases} \tag{1}$$

где α_i — уровень значимости, $\Delta_i = \Delta(\alpha_i) = F^{-1}(1 - \alpha_i/2)$ — порог, $P_i = 2 - 2F(|S_i|)$ — P-значение критерия C_i, $F(\cdot)$ — функция распределения статистики S_i, $i = \overline{1,m}$. Известно, что при истинной гипотезе $\mathcal{H}_{0,i}$ P-значения P_i имеют равномерное на отрезке $[0,1]$ распределение.

Тогда процедура Бонферрони [3, с. 92] имеет следующий вид:

$$\text{принимается } \begin{cases} \mathcal{H}_0, & \text{если } P_i \ge \alpha_i \quad \text{для всех } i = \overline{1,m}, \\ \mathcal{H}_1, & \text{иначе.} \end{cases} \tag{2}$$

В качестве показателя эффективности процедуры множественной проверки гипотез будем рассматривать "обобщенную" вероятность ошибки первого рода [6]:

$$\varepsilon = \mathbf{P}\left\{\text{принять } \mathcal{H}_1 \mid \text{верна } \mathcal{H}_0\right\}, \tag{3}$$

требуя, чтобы для процедуры она не превосходила выбранного уровня значимости α: $\varepsilon \le \alpha$. Известно [3, с. 93], что для обобщенной вероятности ошибки первого рода процедуры Бонферрони (2) выполняется следующее неравенство:

$$\alpha_-^B \le \varepsilon \le \alpha_+^B,$$

$$\alpha_-^B = \max_{1 \le i \le m} \alpha_i, \qquad \alpha_+^B = \sum_{i=1}^{m} \alpha_i$$

и уровни значимости α_i выбирают из условия

$$\alpha_+^B = \alpha_+^B(\alpha_1, \ldots, \alpha_m) = \alpha. \tag{4}$$

Например, при $\alpha_i = \alpha_c$, $i = \overline{1,m}$, по заданному уровню α индивидуальные уровни α_c вычисляются как $\alpha_c = \alpha/m$. Однако в [11] отмечается, что в случае зависимости между статистиками критериев такой выбор может привести к завышенной верхней границе: $\varepsilon/\alpha \ll 1$. Как следствие, это приводит к снижению мощности процедуры и неадекватным статистическим выводам.

3. Уточнение процедуры Бонферрони в случае зависимости между критериями

Обозначим F_{ij} — совместная функция распределения статистик S_i, S_j критериев (1).

Теорема 1. *Для вероятности ошибки первого рода* (3) *процедуры Бонферрони* (2) *с критериями* (1) *справедлива следующая граница сверху:*

$$\varepsilon \le \sum_{i=1}^{m} \alpha_i - \max_{1\le j\le m} \sum_{i=1, i\ne j}^{m} \big(F_{ij}(-\Delta_i, -\Delta_j) - F_{ij}(-\Delta_i, \Delta_j) - \\ - F_{ij}(\Delta_i, -\Delta_j) + F_{ij}(\Delta_i, \Delta_j) + \alpha_i + \alpha_j - 1\big), \quad (5)$$

где α_i — уровень значимости, $\Delta_i = \Delta(\alpha_i)$ — порог критерия C_i, $i = \overline{1,m}$.

Доказательство. Доказательство основано на применении неравенства Бонферрони второго рода [9]

$$\mathsf{P}\left\{\cup_{i=1}^{m} A_i\right\} \le \sum_{i=1}^{m} \mathsf{P}\left\{A_i\right\} - \max_{1\le j\le m} \sum_{i=1, i\ne j}^{m} \mathsf{P}\left\{A_i, A_j\right\}$$

к вероятности ошибки первого рода процедуры: $\varepsilon = \mathsf{P}\left\{\cup_{i=1}^{m}\{P_i < \alpha_i\}\right\}$. Так как P_i распределены равномерно на отрезке $[0,1]$, то $\mathsf{P}\left\{P_i < \alpha_i\right\} = \alpha_i$, $i = \overline{1,m}$. Для вычисления слагаемых $\mathsf{P}\left\{P_i < \alpha_i, P_j < \alpha_j\right\}$ использовалось свойство многомерной функции распределения:

$$\mathsf{P}\left\{a_1 \le x_1 < b_1, a_2 \le x_2 < b_2\right\} = \Delta^{(1)}_{[a_1,b_1)} \Delta^{(2)}_{[a_2,b_2)} F(x_1, x_2),$$

где $\Delta^{(i)}_{[a_i,b_i)}$ — приращение функции распределения $F(x)$ по i-му аргументу на полуинтервале $[a_i, b_i)$. □

Рассмотрим семейство критериев, имеющих совместное нормальное распределение статистик с известной ковариационной матрицей:

$$\mathcal{L}\left\{(S_1, \ldots, S_m)' \mid \mathcal{H}_0\right\} = \mathcal{N}_m\left(0, \Sigma\right), \quad (6)$$
$$\Sigma = (\sigma_{ij})_{i,j=1}^{m}, \qquad \sigma_{ii} = 1, \quad \sigma_{ij} = \mathsf{Corr}\left\{S_i, S_j\right\}, \quad i, j = \overline{1,m},$$

и не являющихся линейно зависимыми ($|\sigma_{ij}| < 1$, $i, j = \overline{1,m}$, $i \neq j$).

Рассмотрим сначала задачу вычисления границы сверху для вероятности ошибки первого рода процедуры Бонферрони по заданным уровням значимости критериев α_c.

Теорема 2. *Для семейства критериев* (1) *с нормальным распределением статистик* (6) *в случае* $\alpha_i = \alpha_c$, $i = \overline{1,m}$, *граница сверху для вероятности ошибки первого рода* (3) *процедуры* (2) *равна:*

$$G(\Delta(\alpha_c)) = m\alpha_c - (m-1)\alpha_c^2 - \max_{1\le j\le m} \sum_{i=1, i\neq j}^{m} \left[\sum_{k=1}^{\infty} \frac{4\left(n_1^{(2k-1)}(\Delta \mid 0,1)\right)^2}{(2k)!} \sigma_{ij}^{2k} \right],$$

где $n_1^{(k)}(\Delta \mid 0,1)$ — (k)*-я производная плотности стандартного нормального закона, вычисленная в точке* $\Delta = \Phi^{-1}(1-\alpha_c/2)$, $\Phi(\cdot)$ — *функция распределения стандартного нормального закона.*

Доказательство. Доказательство основано на использовании в формуле (5) разложения функции распределения двумерного нормального закона [1]:

$$F_{ij}(a,b) = \Phi(a)\,\Phi(b) + \sum_{n=0}^{\infty} \frac{n_1^{(n)}(a \mid 0,1) n_1^{(n)}(b \mid 0,1)}{(n+1)!} \sigma_{ij}^{n+1}. \qquad (7)$$

Учитывая, что

$$n_1^{(n)}(-\Delta \mid 0,1) = n_1^{(n)}(\Delta \mid 0,1), \qquad n = 2k-2, k \in \mathbb{N},$$
$$n_1^{(n)}(-\Delta \mid 0,1) = -n_1^{(n)}(\Delta \mid 0,1), \quad n = 2k-1, k \in \mathbb{N},$$

получаем утверждение теоремы. □

Нахождение по заданному уровню значимости процедуры α уровней значимости индивидуальных критериев α_c с учетом зависимости между статистиками критериев сводится к решению следующего уравнения:

$$G(\Delta(\alpha_c)) - \alpha = 0. \qquad (8)$$

Теорема 3. *При* $m > 2$ *функция* $G(\Delta)$ *строго возрастает на отрезке* $[0;\Delta^*]$ *и строго убывает на промежутке* $[\Delta^*;+\infty)$, $G(0) = 1$ *и* $G(\Delta) \to 0$ *при* $\Delta \to \infty$, *где* $\Delta^* \in (0;+\infty)$. *При* $m = 2$ *функция* $G(\Delta)$ *строго убывающая и* $G(0) = 1$.

Доказательство. Запишем функцию G через двумерные функции рас-

пределения:

$$G(\Delta) = m - 1 - (m-2)(2 - 2\Phi(\Delta)) -$$
$$- \max_{1\le j\le m} \sum_{i=1, i\ne j}^{m} (F_{ij}(\Delta,\Delta) - 2F_{ij}(-\Delta,\Delta) + F_{ij}(-\Delta,-\Delta)) =$$
$$= \min_{1\le j\le m} \{m - 1 - (m-2)(2 - 2\Phi(\Delta)) -$$
$$- \sum_{i=1, i\ne j}^{m} (F_{ij}(\Delta,\Delta) - 2F_{ij}(-\Delta,\Delta) + F_{ij}(-\Delta,-\Delta))\} = \min_{1\le j\le m} G_j(\Delta).$$

Вычисляя производную функции G_j, получаем:

$$G_j'(\Delta) = n(\Delta \mid 0,1) \left(6m - 8 - 4 \sum_{i=1, i\ne j}^{m} [\Phi(k_{ij}\Delta) + \Phi(\Delta/k_{ij})] \right) =$$
$$= n(\Delta \mid 0,1) K(\Delta), \quad k_{ij} = \sqrt{(1-\sigma_{ij})/(1+\sigma_{ij})}$$

и

$$K'(\Delta) = -4 \sum_{i=1, i\ne j}^{m} [k_{ij} n(k_{ij}\Delta \mid 0,1) + n(\Delta/k_{ij} \mid 0,1)/k_{ij}] < 0 \quad \forall \Delta.$$

Легко убедиться, что $\forall m > 2$ выполняется $K(0) > 0$ и существует Δ такое, что $K(\Delta) < 0$. Поэтому, существует точка $\Delta_j^* : K(\Delta_j^*) = 0$. Следовательно, производная функции $G_j(\Delta)$ на промежутке $[0; \Delta_j^*)$ имеет положительный знак, а на $(\Delta_j^*; +\infty)$ — отрицательный, а сама функция $G_j(\Delta)$ на этих промежутках возрастает и убывает соответственно. Также заметим, что $G_j(0) = 1$ и $G_j(\Delta) \to 0$ при $\Delta \to \infty$.

Функции G_j являются гладкими по построению, следовательно, функция G является непрерывной, кусочно-гладкой функцией. Ее область определения состоит из конечного числа промежутков, на каждом из которых $G(\Delta) = G_j(\Delta)$. Следовательно, функция G обладает свойствами функций G_j, установленными в доказательстве теоремы.

Случай $m = 2$ рассматривается аналогично. □

Найденный в теореме 3 вид функции G делает возможным применение численных методов, в частности, метода половинного деления, для численного решения уравнения (8).

Полученные результаты легко обобщаются на случай использования в процедуре Бонферрони односторонних критериев вместо двусторонних.

4. Применения улучшенной процедуры Бонферрони

4.1. Самостоятельное применение. Теоремы 2, 3 позволяют строить уточненную процедуру Бонферрони для семейств критериев, имеющих совместное нормальное распределение статистик. Наиболее известными семействами таких критериев являются критерии серий, приведенные в FIPS 140–2 [7], и критерии поиска шаблонов, приведенные в SP 800-22 [10].

Рассмотрим модельный пример построения процедуры Бонферрони на основе четырех критериев поиска шаблонов `1`, `11`, `111`, `1111` для проверки гипотезы согласия бинарной выборки объема $n = 1000$ с моделью независимых симметричных испытаний Бернулли. Частоты встречаемости данных шаблонов сильно коррелированы: коэффициент корреляции близок к 0,9.

По заданному уровню значимости процедуры Бонферрони $\alpha = 0,05$ индивидуальные уровни значимости критериев традиционно выбираются равными $\alpha_c = 0,05/4 = 0,0125$. При таком выборе α_c верхняя граница (5) для вероятности ошибки первого рода ε с учетом корреляции между статистиками равна $0,0313$. Статистическая оценка для ε и 95%-й доверительный интервал, построенные методом имитационного моделирования, в этом случае равны $0,032$ и $[0,0166; 0,0474]$ соответственно.

Для обеспечения уровня значимости процедуры Бонферрони в $\alpha = 0,05$ индивидуальные уровни значимости критериев были вычислены с использованием теоремы 3 и оказались равными $\alpha_c = 0,02044$. Статистическая оценка для ε и 95%-й доверительный интервал, построенные методом имитационного моделирования, в этом случае равны $0,066$ и $[0,044; 0,088]$ соответственно.

Можно видеть, что результаты имитационных экспериментов согласуются с теоретическими результатами.

4.2. Применение в процедуре Холма. Процедура Холма [8] представляет собой модификацию процедуры Бонферрони на случай пошаговой проверки индивидуальных гипотез. Процедура Холма отвергает гипотезы $\mathcal{H}_{0,i_1}, \ldots, \mathcal{H}_{0,i_k}$ и принимает гипотезы $\mathcal{H}_{0,i_{k+1}}, \ldots, \mathcal{H}_{0,i_m}$, где $i_1, \ldots, i_m$ — индексы P-значений в вариационном ряду $(P_{i_1} \leq \cdots \leq P_{i_m})$ и k определяется пошаговыми сравнениями P-значений с порогами:

$$P_{i_1} < \alpha_{(1)}, P_{i_2} < \alpha_{(2)}, \ldots, P_{i_k} < \alpha_{(k)}, P_{i_{k+1}} \geq \alpha_{(k+1)}, \quad 0 \leq k \leq m;$$

либо отвергает все нулевые гипотезы, если $P_{i_j} < \alpha_{(j)}$ для всех $j = \overline{1,m}$, где $\alpha_{(j)} = \alpha/(m - j + 1)$ — пороги процедуры.

В случае известного попарного совместного распределения статистик для нахождения порогов может использоваться уточненная процедура Бонферрони. Применение данных порогов увеличивает мощность процедуры по сравнению с процедурой Холма.

4.3. Применение в процедуре Бернулли. Для применения процедуры Бернулли [10] исходную выборку X разбивают на K непересекающихся подвыборок $X^{(1)}, \ldots, X^{(K)}$, относительно которых предполагается, что они являются независимыми. На первом этапе ко всем подвыборкам применяются критерии C_i, $i = \overline{1,m}$, и по полученным P-значениям $P_{i,1}, \ldots, P_{i,K}$ вычисляются статистики второго этапа:

$$S_i^{(2)} = \sqrt{K}\frac{v_0 - \mu_0}{\sigma_0}, \quad v_0 = v_0(P_{i,1}, \ldots, P_{i,K}) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{I}\{P_{i,k} \geq \alpha_1\},$$

где v_0 — доля принятых гипотез первого этапа на уровне значимости α_1, $\mu_0 = 1 - \alpha_1$, $\sigma_0 = \sqrt{\alpha_1(1-\alpha_1)}$. Относительно статистик второго этапа $\{S_i^{(2)}\}$ известно, что они имеют совместное нормальное распределение с неизвестной ковариационной матрицей $\Sigma^{(2)}$ [10].

Итоговое решение о гипотезе $\mathcal{H}_0$ принимается на втором этапе процедурой Бонферрони без учета корреляции между статистиками второго этапа:

$$\text{принимается } \begin{cases} \mathcal{H}_0, & \text{если } P_i^{(2)} \geq \alpha_c, \quad i = \overline{1,m}, \\ \mathcal{H}_1, & \text{иначе,} \end{cases}$$

где $P_i^{(2)} = \Phi(S_i^{(2)})$ — P-значение, вычисленное по статистике второго этапа $S_i^{(2)}$, $i = \overline{1,m}$.

В случае использования фиксированного набора критериев для неизвестной матрицы $\Sigma^{(2)}$ может быть построена статистическая оценка методом имитационного моделирования. В случае, когда критерии первого этапа описываются (1), (6), матрица $\Sigma^{(2)}$ может быть найдена аналитически, а для принятия решения может быть использована уточненная процедура Бонферрони.

Теорема 4. *В случае использования на первом этапе двусторонних критериев* (1) *с совместным нормальным распределением статистик* (6) *вектор статистик второго этапа* $S^{(2)} = (S_1^{(2)}, \ldots, S_m^{(2)})'$ *имеет предельное при* $K \to \infty$ m*-мерное нормальное распределение с нулевым вектором математического ожидания и ковариационной матрицей* $\Sigma^{(2)} = (\sigma_{ij}^{(2)})_{i,j=1}^{m}$*:* $\sigma_{ii}^{(2)} = 1$,

$$\sigma_{ij}^{(2)} = \mathsf{Cov}\left\{S_i^{(2)}, S_j^{(2)}\right\} = \frac{4}{\alpha_1(1-\alpha_1)}\sum_{k=1}^{\infty}\frac{\left(n_1^{(2k-1)}(\Delta_1 \mid 0,1)\right)^2}{(2k)!}\sigma_{ij}^{2k},$$

где σ_{ij} *— коэффициент корреляции статистик первого этапа* $S_i^{(1)}$, $S_j^{(1)}$, $\Delta_1 = \Phi^{-1}(1 - \alpha_1/2)$. *Для вероятности ошибки первого рода процедуры*

Бернулли справедлива следующая граница сверху:

$$G_2(\Delta) = m\alpha_c - (m-1)\alpha_c^2 - \max_{1\le j\le m}\sum_{i=1,i\neq j}^{m}\sum_{k=0}^{\infty}\frac{\left(n_1^{(k)}(\Delta \mid 0,1)\right)^2}{(k+1)!}(\sigma_{ij}^{(2)})^{k+1},$$

причем функция $G_2(\Delta)$ является убывающей, $\Delta = \Phi^{-1}(\alpha_c)$.

Доказательство. Используя представление v_0 через индикаторные функции и учитывая, что P_{i,k_1}, P_{j,k_2} независимы при $k_1 \neq k_2$ и распределение P-значений не зависит от номера подвыборки, получим:

$$\sigma_{ij}^{(2)} = \mathsf{Cov}\left\{\sqrt{K}\frac{v_0(P_{i,1},\ldots,P_{i,K})-\mu_0}{\sigma_0}, \sqrt{K}\frac{v_0(P_{j,1},\ldots,P_{j,K})-\mu_0}{\sigma_0}\right\} =$$
$$= \frac{1}{\sigma_0^2}\mathsf{Cov}\left\{\mathbb{I}\{P_{i,1}\ge\alpha_1\}, \mathbb{I}\{P_{j,1}\ge\alpha_1\}\right\}.$$

Далее, находя ковариацию по определению, получим, что

$$\sigma_{ij}^{(2)} = \frac{1}{\sigma_0^2}\left(\mathsf{P}\left\{P_{i,1}\ge\alpha_1, P_{j,1}\ge\alpha_1\right\} - \mathsf{P}\left\{P_{i,1}\ge\alpha_1\right\}\mathsf{P}\left\{P_{j,1}\ge\alpha_1\right\}\right) =$$
$$= \frac{F_{ij}(-\Delta_1,-\Delta_1) - 2F_{ij}(-\Delta_1,\Delta_1) + F_{ij}(\Delta_1,\Delta_1) - (1-\alpha_1)^2}{\alpha_1(1-\alpha_1)}.$$

Используя представление двумерной нормальной функции распределения через ряд (7), получаем первое утверждение теоремы.

Доказательство второго утверждения аналогично доказательству теоремы 2. □

Рассмотрим модельный пример применения процедуры Бернулли на основе двух критериев поиска шаблонов `111`, `1111` для проверки гипотезы согласия бинарной выборки с моделью независимых симметричных испытаний Бернулли. На первом этапе к $K = 300$ подвыборкам объема $n = 10000$ применялись критерии поиска шаблонов на уровне значимости $\alpha_1 = 0,05$. Коэффициент корреляции между статистиками данных критериев равен $\rho = 0,93383$. Согласно теореме 4, коэффициент корреляции между статистиками второго этапа равен $\sigma_{12}^{(2)} = 0,648506$. Для того, чтобы процедура Бернулли имела уровень значимости 0,05, в качестве уровня значимости критерия второго этапа вместо традиционного значения $0,05/2 = 0,025$ должно использоваться $\alpha_c = 0,0295$.

4.4. Применение в процедуре на основе агрегированных статистик. Приведенные выше основные результаты получены для семейства критериев с совместным нормальным распределением статистик. Однако часто на практике используются критерии, имеющие другое распределение статистик, в частности, хи-квадрат распределение. В этом случае

исследование совместного распределения статистик является затруднительным и предлагается использовать следующую процедуру.

Выборка длины n разбивается на K непересекающихся подвыборок длины T каждая, относительно которых предполагается, что они являются независимыми. Критерии C_i, $i = \overline{1,m}$, применяются к подвыборкам и по полученным статистикам $S_i^{(1)}, \ldots, S_i^{(K)}$ вычисляются агрегированные статистики:

$$\tilde{S}_i = \frac{1}{\sqrt{K}} \left(\sum_{k=1}^{K} S_i^{(k)} - K\mu_i \right), \quad i = \overline{1,m},$$

где μ_i — математическое ожидание статистики $S_i^{(k)}$. Если статистики $S_i^{(1)}, \ldots, S_i^{(K)}$ удовлетворяют условиям центральной предельной теоремы, то агрегированные статистики будут иметь совместное нормальное распределение и может быть применена процедура Бонферрони следующего вида:

$$\text{принимается} \begin{cases} \mathcal{H}_0, & \text{если } \tilde{S}_i/\sqrt{\sigma_{ii}} \le \Delta = \Phi^{-1}(1-\alpha_c), \quad i = \overline{1,m}, \\ \mathcal{H}_1, & \text{иначе,} \end{cases}$$

где σ_{ii} — дисперсия статистики $\tilde{S}_i$. В качестве критериев C_i рассмотрим критерии $\chi^2(l)$, предназначенные для проверки гипотезы $\mathcal{H}_0$ о том, что выборка порождена последовательностью независимых симметричных испытаний Бернулли. Критерии $\chi^2(l)$ строятся по непересекающимся l-граммам длин $l \in \{l_1, \ldots, l_m\}$ и основаны на статистике $S^{(k)}_{\chi^2(l_j)} = \frac{2^{l_j}}{K_j} \sum_{i=1}^{2^{l_j}} \nu_{i,l_j}^2 - K_j$, где K_j — число отрезков длины l_j в подвыборке, ν_{i,l_j} — частота встречаемости i-й l_j-граммы в подвыборке:

$$\nu_{i,l_j} = \sum_{k=1}^{K_j} \mathbb{I}\{x_{k,l_j} = h_{i,l_j}\},$$

x_{k,l_j} — k-й отрезок выборки длины l_j, h_{i,l_j} — i-ая l_j-грамма. Агрегированные статистики имеют вид $\tilde{S}_{\chi^2(l_i)} = \frac{1}{\sqrt{K}} \left(\sum_{k=1}^{K} S^{(k)}_{\chi^2(l_i)} - K(2^{l_i} - 1) \right)$, $i = \overline{1,m}$.

Теорема 5. *При верной гипотезе $\mathcal{H}_0$ в асимптотике $K, T \to \infty$ предельным распределением вектора $(\tilde{S}_{\chi^2(l_1)}, \ldots, \tilde{S}_{\chi^2(l_m)})'$ является m-мерное нормальное распределение с нулевым вектором математического ожидания и ковариационной матрицей $\Sigma = (\sigma_{ij})_{i,j=1}^m$: $\sigma_{ii} = 2(2^{l_i} - 1)$,*

$$\sigma_{ij} = \frac{2}{A_i A_j} \left(\sum_{k=1}^{|A^{i,j}|-1} 2^{a^{i,j}_{k+1} - a^{i,j}_k} - A_i - A_j + 1 \right),$$

$i, j = \overline{1,m}$, $A_i = \frac{HOK(l_i,l_j)}{l_i}$, $a_k^{i,j}$ — *k-й по величине член множества*

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

Граница сверху для вероятности ошибки первого рода процедуры, основанной на агрегированных критериях, будет иметь вид:

$$\varepsilon \le m\alpha_c - (m-1)\alpha_c^2 - \max_{1\le j\le m} \sum_{i=1, i\ne j}^{m} \sum_{k=0}^{\infty} \frac{\left(n_1^{(k)}(\Delta \mid 0,1)\right)^2}{(k+1)!} \sigma_{ij}^{k+1}.$$

Доказательство. При фиксированной длине подвыборок T статистики $S^{(k)}_{\chi^2(l_i)}$ имеют некоторое распределение вероятностей с параметрами математического ожидания $\mu_i(T)$ и невырожденной ковариационной матрицей $\Sigma(T) = (\sigma_{ij}(T))_{i,j=1}^m$. Из многомерного случая центральной предельной теоремы [4] следует, что предельным в асимптотике $K \to \infty$ распределением вектора $(\tilde{S}_{\chi^2(l_1)}, \ldots, \tilde{S}_{\chi^2(l_m)})'$ будет m-мерное нормальное распределение с параметрами $\mu(T) = (\mu_1(T), \ldots, \mu_m(T))'$ и $\Sigma(T)$.

Известно [2], что при верной $\mathcal{H}_0$ выполняется $\mu_i(T) = \mu_i = 2^{l_i} - 1$, $i = \overline{1,m}$. Найдем теперь точное значение ковариации статистик $S^{(k)}_{\chi^2(l_i)}$, $S^{(k)}_{\chi^2(l_j)}$ при фиксированной длине T. Будем предполагать, что длина подвыборок T кратна НОК(l_i, l_j): $T = \text{НОК}(l_i, l_j)B$.

Найдем ковариацию при $i = 1$, $j = 2$,.для остальных i, j величины $\sigma_{ij}(T)$ находятся аналогично. Также далее опустим индекс k в $S^{(k)}_{\chi^2(l_1)}, S^{(k)}_{\chi^2(l_2)}$, обозначающий номер подвыборки. По определению:

$$\sigma_{12}(T) = \mathsf{E}\left\{S_{\chi^2(l_1)} S_{\chi^2(l_2)}\right\} - (2^{l_1} - 1)(2^{l_2} - 1). \tag{9}$$

Представим $\mathsf{E}\left\{S_{\chi^2(l_1)} S_{\chi^2(l_2)}\right\}$ в виде суммы четырех слагаемых:

$$\mathsf{E}\left\{S_{\chi^2(l_1)} S_{\chi^2(l_2)}\right\} = -\frac{K_2}{K_1} 2^{l_1} \mathsf{E}\left\{\sum_{i=1}^{2^{l_1}} \nu_{i,l_1}^2\right\} - \frac{K_1}{K_2} 2^{l_2} \mathsf{E}\left\{\sum_{j=1}^{2^{l_2}} \nu_{j,l_2}^2\right\} +$$
$$+ \frac{2^{l_1+l_2}}{K_1 K_2} \mathsf{E}\left\{\sum_{i=1}^{2^{l_1}} \sum_{j=1}^{2^{l_2}} \nu_{i,l_1}^2 \nu_{j,l_2}^2\right\} + K_1 K_2. \tag{10}$$

Каждое слагаемое будем находить, представляя частоты ν_{i,l_1} через ин-

дикаторные функции. Так,

$$\mathsf{E}\left\{\sum_{i=1}^{2^{l_1}}\nu_{i,l_1}^2\right\}=\sum_{i=1}^{2^{l_1}}\mathsf{E}\left\{\left(\sum_{k=1}^{K_1}\mathbb{I}\{x_{k,l_1}=h_{i,l_1}\}\right)^2\right\}=$$

$$=\sum_{i=1}^{2^{l_1}}\mathsf{E}\left\{\sum_{k=1}^{K_1}\mathbb{I}\{x_{k,l_1}=h_{i,l_1}\}+\sum_{\substack{k,l=1\\k\neq l}}^{K_1}\mathbb{I}\{x_{k,l_1}=h_{i,l_1},x_{l,l_1}=h_{i,l_1}\}\right\}=$$

$$=2^{l_1}K_1\frac{1}{2^{l_1}}+2^{l_1}K_1(K_1-1)\frac{1}{2^{2l_1}}=K_1\left(1+\frac{K_1-1}{2^{l_1}}\right).$$

Аналогично находим $\mathsf{E}\left\{\sum_{j=1}^{2^{l_2}}\nu_{j,l_2}^2\right\}=K_2\left(1+\frac{K_2-1}{2^{l_2}}\right)$.

Найдем $\mathsf{E}\left\{\sum_{i=1}^{2^{l_1}}\sum_{j=1}^{2^{l_2}}\nu_{i,l_1}^2\nu_{j,l_2}^2\right\}$:

$$\mathsf{E}\left\{\sum_{i=1}^{2^{l_1}}\sum_{j=1}^{2^{l_2}}\nu_{i,l_1}^2\nu_{j,l_2}^2\right\}=\sum_{i=1}^{2^{l_1}}\sum_{j=1}^{2^{l_2}}\mathsf{E}\left\{\sum_{k=1}^{K_1}\sum_{m=1}^{K_2}\mathbb{I}\{x_{k,l_1}=h_{i,l_1},x_{m,l_2}=h_{j,l_2}\}\right\}+ \tag{11}$$

$$+\sum_{i=1}^{2^{l_1}}\sum_{j=1}^{2^{l_2}}\mathsf{E}\left\{\sum_{\substack{k,l=1\\k\neq l}}^{K_1}\sum_{m=1}^{K_2}\mathbb{I}\{x_{k,l_1}=h_{i,l_1},x_{l,l_1}=h_{i,l_1},x_{m,l_2}=h_{j,l_2}\}\right\}+ \tag{12}$$

$$+\sum_{i=1}^{2^{l_1}}\sum_{j=1}^{2^{l_2}}\mathsf{E}\left\{\sum_{k=1}^{K_1}\sum_{\substack{m,n=1\\m\neq n}}^{K_2}\mathbb{I}\{x_{k,l_1}=h_{i,l_1},x_{m,l_2}=h_{j,l_2},x_{n,l_2}=h_{j,l_2}\}\right\}+ \tag{13}$$

$$+\sum_{i=1}^{2^{l_1}}\sum_{j=1}^{2^{l_2}}\mathsf{E}\left\{\sum_{\substack{k,l=1\\k\neq l}}^{K_1}\sum_{\substack{m,n=1\\m\neq n}}^{K_2}\mathbb{I}\{x_{k,l_1}=h_{i,l_1},x_{l,l_1}=h_{i,l_1}\}\times\right.$$

$$\left.\times\,\mathbb{I}\{x_{m,l_2}=h_{j,l_2},x_{n,l_2}=h_{j,l_2}\}\right\}. \tag{14}$$

Найдем слагаемое (11). Разделим его на две суммы, соответствующие множествам $B_1=\{k,m:x_{k,l_1}\cap x_{m,l_2}\}$ и $B_2=\{k,m:x_{k,l_1}\not\cap x_{m,l_2}\}$.

$$\sum_{i=1}^{2^{l_1}}\sum_{j=1}^{2^{l_2}}\mathsf{E}\left\{\sum_{B_1}\mathbb{I}\{x_{k,l_1}=h_{i,l_1},x_{m,l_2}=h_{j,l_2}\}\right\}=$$

$$=\sum_{B_1}2^{l_1}2^{l_2-|\Delta_{k,m}|}\frac{1}{2^{l_1+l_2-|\Delta_{k,m}|}}=||B_1||,$$

где $|\Delta_{k,m}|$ — длина пересечения отрезков x_{k,l_1} и x_{m,l_2}.

$$\sum_{i=1}^{2^{l_1}}\sum_{j=1}^{2^{l_2}} \mathsf{E}\left\{\sum_{B_2} \mathbb{I}\{x_{k,l_1}=h_{i,l_1}, x_{m,l_2}=h_{j,l_2}\}\right\} = \sum_{B_2} 2^{l_1}2^{l_2}\frac{1}{2^{l_1+l_2}} = ||B_2||.$$

Следовательно, слагаемое (11) равно $||B_1|| + ||B_2|| = K_1K_2$.

Аналогичными рассуждениями получаем, что слагаемое (12) равно $K_1K_2(K_1-1)2^{-l_1}$, слагаемое (13) — $K_1K_2(K_2-1)2^{-l_2}$ и слагаемое (14)

$$\frac{1}{2^{l_1+l_2}}K_1(K_1-1)K_2(K_2-1)+$$
$$+\frac{1}{2^{l_1+l_2}}2B(B-1)\left(\sum_{k=1}^{|A^{1,2}|-1} 2^{a^{1,2}_{k+1}-a^{1,2}_k} - (A_1+A_2-1)\right).$$

Подставляя найденные слагаемые в (10) и (9), получаем выражение для точной ковариации статистик критериев хи-квадрат:

$$\sigma_{12}(T) =$$
$$= \frac{2}{A_1A_2}\left(\sum_{k=1}^{|A^{1,2}|-1} 2^{a^{1,2}_{k+1}-a^{1,2}_k} - (A_1+A_2-1)\right)\frac{T/\mathrm{HOK}(l_1,l_2)-1}{T/\mathrm{HOK}(l_1,l_2)}.$$

Переходя к пределу $T\to\infty$, получаем условие теоремы.

Доказательство второго утверждения аналогично доказательству теоремы 2. □

Следствие. *Если l_i делит l_j, то $\sigma_{ij} = 2(2^{l_i}-1)$.*

Рассмотрим модельный пример применения процедуры на основе агрегированных статистик критериев $\chi^2(l)$, построенных по непересекающимся l-граммам длин 1, 2, 3, 4 для проверки гипотезы согласия бинарной выборки объема $n = 204\,800$ с моделью независимых симметричных испытаний Бернулли.

По заданному уровню значимости процедуры $\alpha = 0,2$ индивидуальные уровни значимости критериев согласно (4) выбираются равными $\alpha_c = 0,2/4 = 0,05$. При таком выборе α_c верхняя граница для вероятности ошибки первого рода ε с учетом корреляции между статистиками равна $0,0167674$. Статистическая оценка для ε и 95%-й доверительный интервал, построенные методом имитационного моделирования, в этом случае равны $0,1692$ и $[0,1613; 0,1746]$ соответственно.

Для обеспечения уровня значимости процедуры в $\alpha = 0.2$ индивидуальные уровни значимости критериев следует выбирать равными $\alpha_c = 0,0607$. Статистическая оценка для ε и 95%-й доверительный интервал равны $0,194$ и $[0,1878; 0,2019]$ соответственно.

Таким образом, результаты имитационных экспериментов согласуются с теоретическими результатами.

5. Заключение

В работе представлена уточненная процедура Бонферрони множественной проверки гипотез, учитывающая зависимости между статистиками нескольких распространенных на практике семейств критериев, рассмотрено ее применение для улучшения других процедур множественной проверки гипотез. Представлены данные численных экспериментов, иллюстрирующих полученные теоретические результаты.

Список литературы

[1] *Большев Л.Н., Смирнов Н. В.* Таблицы математической статистики. — М.: Наука, 1983. — 416 с.

[2] *Ивченко Г.И., Медведев Ю.И.* Математическая статистика. — М.: Высшая школа, 1984. — 248 с.

[3] *Кокс Д., Хинкли Д.* Теоретическая статистика. — М.: Мир, 1978.

[4] *Феллер В.* Введение в теорию вероятностей и ее приложения. В 2-х томах. Т.2. — М.: Мир, 1984. — 738 с.

[5] *Dichtl M.* Descriptions of General NESSIE Test Tools. NESSIE Document NES/DOC/SAG/WP2/023/2, 2001.

[6] *Dudoit S., van der Laan M. J.* Multiple Testing Procedures with Applications to Genomics. — N.Y.: Springer, 2007. — 590 p.

[7] FIPS Publication 140–2. Security Requirements for Cryptographic Modules. — National Institute of Standards and Technology, 2001.

[8] *Holm S.* A simple sequentially rejective multiple test procedure. Scand. J. Statist. 1979. Vol.6. P.65-70.

[9] *Kounias E.* Bounds for the probability of a union, with applications. Ann. Math. Statist. 1968. Vol. 39. P. 2154-2158.

[10] NIST special Publications 800-22. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. — National Institute of Standards and Technology, 2000.

[11] *Simes R.J.* An improved Bonferroni procedure for multiple tests of significance. Biometrika. 1986. Vol. 73. P.751-754.