АНАЛИЗ СВОЙСТВ ПРОЦЕДУР МНОЖЕСТВЕННОЙ ПРОВЕРКИ ГИПОТЕЗ В ЗАДАЧАХ КРИПТОГРАФИИ

И. С. Никитина

НИИ прикладных проблем математики и информатики
Минск, Беларусь
E-mail:isdc@tut.by

В работе рассматриваются известные процедуры множественной проверки гипотез, обсуждаются трудности при построении процедур, предлагаются уточнения процедур на случай зависимости между статистиками критериев, а также исследуются условия устойчивости двухэтапных процедур к искажениям в распределении статистик первого этапа.

Ключевые слова: множественная проверка гипотез, процедура Бонферрони, двухэтапные процедуры множественной проверки гипотез, устойчивость.

ВВЕДЕНИЕ

Задача статистического тестирования бинарных последовательностей возникает в криптографии при оценке качества генераторов случайных чисел, криптографических примитивов и др. Например, на конкурсах алгоритмов шифрования AES, NESSIE одним из этапов исследования стойкости было статистическое тестирование выходных последовательностей криптоалгоритмов [1, 3]. При статистическом тести-

ровании качества бинарных последовательностей для принятия решения по результатам применения набора статистических критериев необходимо использование процедур множественной проверки гипотез (МПГ) для согласования уровней значимости критериев набора и заданного уровня значимости всей процедуры.

В процессе использования процедур МПГ на практике было обозначено несколько проблем. Во-первых, вероятность ошибки первого рода процедур, построенных для любого набора критериев, может быть гораздо меньше заданного уровня значимости из-за наличия зависимостей между статистиками критериев и тогда мощность процедуры будет заниженной [4]. Во-вторых, одновременное применение большого числа критериев приводит к установке близких к нулю уровней значимости критериев. Такие уровни значимости нежелательны в ситуациях, когда при построении критериев используются асимптотические распределения статистик: принятие решения происходит на хвостах функций распределения и возникающие погрешности могут приводить к ненадежным статистическим выводам. Помимо этого, сам факт использования асимптотических распределений вместо точных (что часто имеет место на практике в области криптографии из-за дискретности исходных данных) требует исследования устойчивости процедур к таким «искажениям».

В данной работе рассматриваются широко применяемые на практике процедуры множественной проверки гипотез, предлагаются их уточнения на случай зависимости между статистиками для увеличения мощности процедур, а также исследуется устойчивость процедур к искажениям в распределении статистик критериев.

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

Задача множественной проверки гипотез заключается в одновременной проверке нулевых гипотез $H_{0,1},...,H_{0,m}$ против соответствующих альтернатив $H_{1,1},...,H_{1,m}$ по выборке $X=(x_1,...,x_n)$ объема n с помощью m критериев $C_1,...,C_m$. Будем рассматривать процедуры множественной проверки гипотез, проверяющие объединенную гипотезу $H_0=\bigcap_{i=1}^{m}H_{0,i}$ против альтернативной гипотезы $H_1=\bigcup_{i=1}^{m}H_{1,i}$ с вероятностью ошибки первого рода, не превышающей заданный уровень значимости α:

$$a=\mathbf{P}\{\text{принять } H_1|H_0\}\le \alpha. \qquad (1)$$

В криптографии в качестве нулевой гипотезы рассматривается гипотеза о «чистой» случайности бинарной выборки:

$H_0:\{x_t\}$ – независимые, одинаково распределенные случайные величины,

$$\mathbf{P}\{x_t=0\}=\mathbf{P}\{x_t=1\}=0.5.$$

Одноэтапные процедуры МПГ проверяют нулевую гипотезу по одной выборке X, принимая решение обычно по набору P-значений критериев $P_1,...,P_m$ (статистик $S_1,...,S_m$). Наиболее известной является процедура Бонферрони:

$$\text{принимается } \begin{cases}\mathrm{H}_0, & \text{если } P_i\ge \alpha_c=\alpha/m, i=\overline{1,m},\\ \mathrm{H}_1, & \text{иначе.}\end{cases} \qquad (2)$$

Двухэтапные процедуры на первом этапе применяют критерии $C_1,...,C_m$ к K подвыборкам $X^{(1)},...,X^{(K)}$ (длины n_K каждая) выборки X для получения P-значений (статистик) первого этапа $\{P_i^{(1)},...,P_i^{(K)}\}$ $(\{S_i^{(1)},...,S_i^{(K)}\}), i=\overline{1,m}$. На втором

этапе проверяется гипотеза H_0 по наборам Р-значений первого этапа с помощью критериев второго этапа $C_1^{(2)},...,C_m^{(2)}$; итоговое решение принимается одноэтапной процедурой множественной проверки гипотез. Обычно критерии $C_i^{(2)}, i=\overline{1,m}$, являются критериями согласия распределения Р-значений $P_i^{(1)},...,P_i^{(K)}$ с равномерным на отрезке [0; 1] распределением. Перечислим известные двухэтапные процедуры множественной проверки гипотез:

1. Процедура хи-квадрат. В качестве критериев $C_i^{(2)}, i=\overline{1,m}$, используется критерий согласия хи-квадрат. Пусть задано разбиение отрезка [0;1] на M интервалов G_j, $j=\overline{1,M}$ ($[0;1]=\bigcup_{j=1}^{M} G_j$). Критерии второго этапа основаны на статистиках и Р-значениях следующего вида:

$$S_{\chi^2}(\{P_i^{(k)}\})=\sum_{j=1}^{M}\frac{(n_j(P_i^{(1)},...,P_i^{(K)})-Kp_j^{(0)})^2}{Kp_j^{(0)}},\ P_i=1-F_{\chi^2,M-1}(S_{\chi^2}(\{P_i^{(k)}\})),\ i=\overline{1,m}, \qquad (3)$$

где $n_j(P_i^{(1)},...,P_i^{(K)})$ – число P-значений, попавших в интервал G_j, $p_j^{(0)}$ – вероятность попадания Р-значений в интервал G_j при верной гипотезе H_0, $F_{\chi^2,M-1}(x)$ – функция хи-квадрат распределения с $M-1$ степенями свободы. Для принятия итогового решения используется процедура Бонферрони (2) с P-значениями $P_1,...,P_m$, вычисленными по формуле (3).

2. Процедура Бернулли. Критерии второго этапа процедуры Бернулли основаны на статистике частоты принятых на уровне значимости α_1 гипотез первого этапа:

$$S_B(\{P_i^{(k)}\})=\frac{\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbf{I}\{P_i^{(k)}\le\alpha_1\}-\sqrt{K}(1-\alpha_1)}{\sqrt{\alpha_1(1-\alpha_1)}},\ P_i=\Phi(S_B(\{P_i^{(k)}\})),\ i=\overline{1,m}, \qquad (4)$$

где $\Phi(x)$ – функция распределения стандартного нормального закона. Для принятия решения используется процедура Бонферрони с P-значениями (4).

3. Процедура с критерием Колмогорова. Критериями второго этапа процедуры являются критерии согласия Колмогорова со статистиками и P-значениями, равными

$$S_{Kolm}(\{P_i^{(k)}\})=\sqrt{K}\sup_{0\le x\le 1}\left|\frac{1}{K}\sum_{k=1}^{K}\mathbf{I}\{P_i^{(k)}\le x\}-x\right|,\ P_i=1-F(S_{Kolm}\{P_i^{(k)}\}),i=\overline{1,m}, \qquad (5)$$

где $F(x)$ – функция распределения Колмогорова. Для принятия решения используется процедура Бонферрони с P-значениями (5).

УТОЧНЕНИЕ ПРОЦЕДУР МНОЖЕСТВЕННОЙ ПРОВЕРКИ ГИПОТЕЗ НА СЛУЧАЙ ЗАВИСИМОСТИ МЕЖДУ КРИТЕРИЯМИ

Известно [4], что, выбирая $\alpha_c=\alpha/m$, для вероятности ошибки первого рода процедуры Бонферрони (2) справедлива оценка сверху $a\le\alpha$. Однако при наличии зависимости между статистиками такой выбор может привести к завышенной оценке сверху: $a/\alpha<<1$. В [5] рассмотрена задача нахождения оценки вероятности ошибки

первого рода a по α_c и уточнена верхняя граница для значения a в случае, если статистики критериев имеют стандартное нормальное распределение и их совместное распределение также является нормальным с известной ковариационной матрицей:

$$\mathrm{L}\{S = (S_1,...,S_m)'\} = N(0,\Sigma),\ \ \mathrm{Cov}\{S_i, S_j\} = \sigma_{ij},\ i,j = \overline{1,m}\,. \qquad (6)$$

Теорема 1 [5]. Для вероятности ошибки первого рода процедуры Бонферрони (2) с критериями, статистики которых имеют распределение (6), выполняется:

$$a \le \alpha_+ = G(\Delta(\alpha_c)),$$

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

в случае двусторонних критериев, а в случае односторонних:

$$G(\Delta(\alpha_c)) = 1 + (m-2)(1-\alpha_c) - \max_{1\le j\le m}\sum_{i\ne j}^{m}F_{ij}(\Delta,\Delta), \qquad (7)$$

где $F_{ij}(x)$ – функция распределения вектора (S_i, S_j), $\Delta = \Delta(\alpha_c)$ – порог критериев.

Задача нахождения уровней значимости α_c критериев по заданному уровню значимости процедуры Бонферрони, т. е. решение уравнения $\alpha_+(\alpha_c) = \alpha$, рассматривается в [5].

Аналогичная проблема учета зависимостей возникает и при использовании двухэтапных процедур, так как критерии первого этапа применяются к одним и тем же подвыборкам. Уточнение двухэтапной процедуры Бернулли производится с помощью теоремы 1 по формуле (7), так как на втором этапе используется процедура Бонферрони и статистики процедуры второго этапа S_B имеют в пределе совместное нормальное распределение, найденное в теореме 2.

Теорема 2 [5]. Статистики второго этапа процедуры Бернулли с двусторонними критериями первого этапа имеют в пределе $K - \propto$ совместное m-мерное нормальное распределение с нулевым вектором математического ожидания, единичными дисперсиями и ковариациями:

$$\mathrm{Cov}\{S_i, S_j\} = \sigma_{ij} = \frac{F_{ij}(\Delta_1,\Delta_1) + F_{ij}(-\Delta_1,-\Delta_1) - 2F_{ij}(-\Delta_1,\Delta_1) - (1-\alpha_1)^2}{\alpha_1(1-\alpha_1)},\ i,j = \overline{1,m},$$

где $F_{ij}(x)$ – маргинальная функция распределения вектора $(S_i^{(k)}, S_j^{(k)})$, $k = \overline{1,K}$.

Как уже было сказано, учет зависимостей между статистиками критериев приводит к построению более мощных процедур множественной проверки гипотез. С другой стороны, увеличение индивидуальных уровней значимости в некоторых случаях позволяет уменьшить погрешности, возникающие в результате использования предельных распределений статистик вместо точных. Вторым подходом к увеличению уровней значимости индивидуальных критериев в батарее является построение процедур множественной проверки гипотез со «свидетелями» [6].

УСТОЙЧИВОСТЬ ДВУХЭТАПНЫХ ПРОЦЕДУР К ИСКАЖЕНИЯМ

На практике, например, вследствие использования асимптотического распределения статистик критериев вместо точного, распределение Р-значений при истинной

нулевой гипотезе может отличаться от равномерного на отрезке [0;1] распределения. Будем считать, что истинное распределение принадлежит $\varepsilon(n)$-окрестности Леви равномерного распределения:

$$P_{\varepsilon(n)}(U[0;1]) = \{F \mid (\forall t)t - \varepsilon(n) \le F(t) \le t + \varepsilon(n)\},\ \varepsilon(n) \underset{n\to\infty}{\longrightarrow} 0. \qquad (8)$$

Для обозначения «искаженных» *P*-значений с распределением из окрестности (8) будем к нижнему индексу добавлять $\varepsilon(n)$. Легко показать, что одноэтапная процедура Бонферрони устойчива к таким искажениям. Ситуация усложняется в случае двухэтапных процедур. При возрастании числа выборок и фиксированных искажениях на первом этапе критерий второго этапа будет обнаруживать искажения и тогда вероятность ошибки первого рода процедуры может быть близка к 1. Исследуем устойчивость процедуры с критерием Колмогорова в случае, если распределение *P*-значений первого этапа $F_{\varepsilon(n_K)}(x)$ неизвестно и принадлежит окрестности (8).

Статистикой критерия Колмогорова при наличии искажений является статистика $\widetilde{S}_{Kolm}(\{P^{(k)}_{i,\varepsilon(n_K)}\}) = \sqrt{K}\sup_{0\le x\le 1}\left|\frac{1}{K}\sum_{k=1}^{K} I\{P^{(k)}_{i,\varepsilon(n_K)} \le x\} - F_{\varepsilon(n_K)}(x)\right|$, которая имеет распределение Колмогорова при истинной нулевой гипотезе, однако ее использование на практике затруднительно вследствие неизвестной функции распределения $F_{\varepsilon(n_K)}(x)$ «искаженных» *P*-значений. Поэтому на практике используется статистика

$$S_{Kolm}(\{P^{(k)}_{i,\varepsilon(n_K)}\}) = \sqrt{K}\sup_{0\le x\le 1}\left|\frac{1}{K}\sum_{k=1}^{K} I\{P^{(k)}_{i,\varepsilon(n_K)} \le x\} - x\right|.$$

Обозначим вероятность ошибки первого рода процедуры в этом случае:

$$a_{\varepsilon(n),K} = \mathbf{P}\{\text{принять } H_1 \text{ по } \{S_{Kolm}(\{P^{(k)}_{i,\varepsilon(n_K)}\})\} \mid H_0\}$$

и найдем условия сходимости вероятности ошибки первого рода при наличии искажений к вероятности ошибки первого рода процедуры без искажений (1).

Теорема 3 [2]. При наличии искажений (8), если $\sqrt{K}\varepsilon(n_K) \underset{n_K,K\to\infty}{\to} 0$, то

$$\left|S_{Kolm}(\{P^{(k)}_{i,\varepsilon(n_K)}\}) - \tilde{S}_{Kolm}(\{P^{(k)}_{i,\varepsilon(n_K)}\})\right| \xrightarrow{\text{п.н.}} 0 \text{ и } a_{\varepsilon(n),K} \to a \text{ при } n_K, K - \ \propto.$$

Аналогичный результат для процедур хи-квадрат и Бернулли приведен в [2].

ЗАКЛЮЧЕНИЕ

В работе рассматриваются известные процедуры множественной проверки гипотез. Предлагается уточнение процедуры Бонферрони на случай зависимости между статистиками для увеличения мощности процедуры, приводятся соотношения, связывающие параметры двухэтапных процедур для обеспечения устойчивости к искажениям в распределении *P*-значений первого этапа.

ЛИТЕРАТУРА

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

2. *Kostevich, A. L.* Robustness of Two-Level Testing Procedures under Distortions of First Level Statistics / A. L. Kostevich, I. S. Nikitina // The Eighth International Conference «Computer Data Analysis and Modeling: Complex Stochastic Data and Systems» (September 11–15, 2007, Minsk). Proceedings of int.conference. Минск: BSU. 2007. Vol. 1. P. 128–131.

3. NIST Special Publication 800-22. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications, 2000.

4. *Simes, R. J.* An improved Bonferroni procedure for multiple tests of significance / R. J. Simes // Biometrika. 1986. Vol. 73. P. 751–754.

5. *Костевич, А. Л.* Уточнение процедуры Бонферрони множественной проверки гипотез в случае зависимости между критериями / А. Л. Костевич, И. С. Никитина // Обозрение прикладной и промышленной математики. 2009. Т. 16. Вып. 3. С. 436–448.

6. *Милованова, И. С.* Об одном подходе к построению одноэтапных процедур множественной проверки гипотез / И. С. Милованова, А. Л. Костевич // X Респ. науч. конф. студентов и аспирантов высших учеб. заведений Респ. Беларусь «НИРС-2005». 14–16 февраля 2006 г. Материалы науч. конф. Минск, 2006. С. 183.