



СТАТИСТИЧЕСКОЕ ТЕСТИРОВАНИЕ КРИПТОГРАФИЧЕСКИХ ГЕНЕРАТОРОВ: СОСТОЯНИЕ И ПЕРСПЕКТИВЫ

Ю.С. Харин

Kharin@bsu.by

НИИ прикладных проблем математики и информатики,
Белорусский государственный университет
Минск, БЕЛАРУСЬ

1. Введение
2. Недостатки существующих батарей тестов
3. Некорректность модели простой нулевой гипотезы
4. Математическая модель: сложная нулевая гипотеза
5. Логарифмическая статистика отношения правдоподобия
6. Решающее правило на основе логарифмической статистики отношения правдоподобия
7. Результаты компьютерных экспериментов
8. Заключение
9. Список литературы

В криптографии генерация случайности [1] сводится к задаче генерации равномерно распределенной двоичной случайной последовательности (РПС) $x_t \in \{0, 1\}$, $t = 1, 2, \dots$, которая определяется двумя базовыми свойствами:

C1) для любого n и любых индексов $1 \leq t_1 < \dots < t_n$ случайные величины $x_{t_1}, \dots, x_{t_n}$ независимы в совокупности;

C2) для любого номера t случайная величина x_t имеет равномерное распределение: $P\{x_t = a\} = 1/2$, $a \in \{0, 1\}$.

При этом неизбежно возникает сопутствующая проблема «статистического тестирования»: построить решающее правило (тест), позволяющий по наблюдаемой двоичной последовательности X длины T проверить с заданным уровнем значимости верна ли гипотеза H_0 об s -мерном равномерном распределении или нет. На текущий момент существуют десятки «батарей тестов» (т.е., комплексов тестов, основанных на различных статистиках) для проверки H_0 , $\overline{H_0}$, среди которых наиболее известны NIST STS [2], DieHarder [3], gJrand [4].

2. Недостатки существующих батарей статистического тестирования



Проведенный в [5] обзор существующих тестов выявляет следующие **недостатки**: 1) многие из тестов ориентированы на проверку лишь частных случаев альтернативы $\overline{H_0}$; 2) многие тесты не фиксируют семейство альтернатив и не имеют оценок мощности w_T ; 3) при включении нескольких тестов в «батарею» не удастся учесть вероятностную зависимость тестовых статистик. Особенно значимой критике и конструктивным предложениям подвергнута «батарея» тестов NIST STS [6]- [9], что привело к решению Национального института США о начале «ревизии» «батареи тестов» [10]. Для частичного преодоления указанных недостатков предлагается строить статистические тесты s -мерной равномерности для заданных семейств альтернатив $\overline{H_0}$, использовать энтропийные профили [11], строить асимптотические оценки мощностей известных NIST-тестов [12] и применять сложные марковские модели [13]- [15].

Существующие статистические тесты и построенные на их основе батареи обладают еще одним существенным недостатком, приводящим к следующему парадоксу: при увеличении длины $T \rightarrow +\infty$ наблюдаемых последовательностей, порождаемых реально существующими источниками случайности, вероятность принятия альтернативы $\overline{H}_0$ стремится к единице: $w_T \rightarrow 1$. Иначе говоря, «увеличением длины проверяемой последовательности T можно забраковать любой реальный источник случайности». Объяснение этого «парадокса» состоит в следующем. Реальные источники случайности – не идеальны, они отклоняются от гипотезы H_0 по своему вероятностному распределению на некоторую величину $\varepsilon \neq 0$. Все тесты, включаемые в батареи тестов, как известно, обладают оптимальным свойством проверки простой гипотезы H_0 против сложной альтернативы $\overline{H}_0$ – свойством состоятельности: $w_T \rightarrow 1$ при $T \rightarrow +\infty$. Таким образом, этот «парадокс» порожден некорректностью математической модели нулевой гипотезы H_0 . Для того, чтобы избежать указанного «парадокса», нулевая гипотеза должна быть сложной: она должна задавать величину допуска ε отклонений от простой гипотезы H_0 .

В связи с этим актуальна рассматриваемая в этом докладе задача построения семейства сложных нулевых гипотез H_0^ε , решающих правил заданного уровня значимости α_0 для проверки H_0^ε , $\overline{H}_0^\varepsilon$ и асимптотического анализа их вероятностных характеристик.

Пусть наблюдается двоичная случайная последовательность длины $T = n \cdot s$, разбитая на n последовательных s -грамм ($s \in \mathbb{N}$):

$$X = X_1^T = (x_1, \dots, x_T) \equiv (X_1^s, X_{s+1}^{2s}, \dots, X_{(n-1)s+1}^T) \in V^T, \quad V = \{0, 1\}; \quad (1)$$

эти s -граммы независимы в совокупности и одинаково распределены с некоторым распределением $p = (p_{J_1^s}) \in \mathcal{P}$:

$$\mathbf{P} \left\{ X_{(i-1)s+1}^{is} = J_1^s \right\} :: \mathbf{P} \left\{ x_{(i-1)s+1} = j_1, \dots, x_{is} = j_s \right\} = p_{J_1^s}, J_1^s = (j_k) \in V^s, i = 1, \dots, n, \quad (2)$$

$$\mathcal{P} = \left\{ p = (p_{J_1^s}) : p_{J_1^s} \geq 0, \quad J_1^s \in V^s, \quad \sum_{J_1^s \in V^s} p_{J_1^s} = 1 \right\} - \quad (3)$$

семейство (симплекс) всевозможных s -мерных вероятностных распределений на V^s .

Для упрощения обозначений биективно перейдем от мультииндекса $J_1^s = (j_1, \dots, j_s) \in V^s$ к одномерному индексу $k \in \{0, 1, \dots, 2^s - 1\}$:

$$k = \langle J_1^s \rangle :: \sum_{i=1}^s j_i \cdot 2^{i-1} \in \{0, 1, \dots, K-1\}, \quad K = 2^s;$$

$$J_1^s = (j_1, \dots, j_s) \Rightarrow k, \quad j_{s-i+1} = \left[\left(k - \sum_{l=s-i+2}^s j_l \cdot 2^{l-1} \right) / 2^{s-i} \right], \quad i = 1, 2, \dots, s.$$

В существующих «батареях тестов» [2, 3] подвергнутых многочисленным модификациям и критике [4] - [6], используются тесты статистической проверки **простой нулевой гипотезы**

$$H_0 = \{p = p_0\}, \quad p_0 = (p_{0k}), \quad p_{0k} \equiv \frac{1}{K}, \quad (4)$$

против сложной альтернативы $\bar{H}_0$.

Во избежание этого парадокса «неадекватности модели (4) нулевой гипотезы» введем в рассмотрение **сложную нулевую гипотезу** об s -мерной равномерности:

$$H_0^\varepsilon = \{p = (p_k) \in \mathcal{P}_0^\varepsilon\}, \quad (5)$$

$$\mathcal{P}_0^\varepsilon = \left\{ p \in \mathcal{P} : \|p - p_0\| = \sqrt{\sum_{k=0}^{K-1} \left(p_k - \frac{1}{K}\right)^2} \leq \varepsilon \right\} -$$

гипершар в R^K заданного радиуса ε с центром в точке p_0 s -мерного равномерного распределения; $\varepsilon \in (0, \sqrt{1 - 1/K})$ – достаточно малый параметр нулевой гипотезы, определяющий допустимые отклонения от простой нулевой гипотезы H_0 . Гипотеза H_0 является предельной по отношению к H_0^ε : $H_0^\varepsilon \xrightarrow{\varepsilon \rightarrow 0} H_0$.

Актуальными для теории и практики задачами являются построение статистических критериев проверки сложных гипотез H_0^ε и $H_1^\varepsilon = \bar{H}_0^\varepsilon$ заданного асимптотического уровня значимости $\alpha_0 \in (0, 1/2)$ на основе статистики отношения правдоподобия.

5. Логарифмическая статистика отношения правдоподобия



Логарифмическая статистика отношения правдоподобия для проверки сложных гипотез $H_0^\varepsilon, H_1^\varepsilon$ об s -мерной равномерности по наблюдаемой двоичной последовательности $X \in V^T$ имеет вид:

$$\lambda(X) = \begin{cases} \max_{p \in S_\varepsilon} l(X; p) - n \sum_{k=0}^{K-1} \hat{p}_k \ln \hat{p}_k, & \text{если } \|\hat{p} - p_0\| > \varepsilon, \\ 0, & \text{если } \|\hat{p} - p_0\| \leq \varepsilon, \end{cases} \quad (6)$$

$l(X; p) = n \sum_{k=0}^{K-1} \hat{p}_k \ln \hat{p}_k$ – логарифмическая функция правдоподобия (ЛФП).

$$\hat{p}_k = \frac{1}{n} \sum_{i=0}^n \mathbf{1} \left\{ \langle X_{(i-1)s+1}^{is} \rangle = k \right\} \geq 0,$$

$S_\varepsilon = \{p \in \mathcal{P} : \|p - p_0\| = \varepsilon\} \subset \mathcal{P}_0^\varepsilon$ – гиперсфера радиуса ε (см. рис. 1), $\mathbf{1}\{C\}$ – индикатор события C .

Максимум ЛФП на гиперсфере S_ε достигается в точке $\tilde{p}^*$:

$$\max_{p \in S_\varepsilon} l(X; p) = l(X; \tilde{p}^*).$$

Ближайшей к точке $\hat{p} \in \mathcal{P}$ точкой $p^{**} \in \mathcal{P}_0^\varepsilon$ является точка гиперсферы S_ε :

$$p^{**} = p_0 + \frac{\varepsilon}{\|\hat{p} - p_0\|} (\hat{p} - p_0). \quad (7)$$

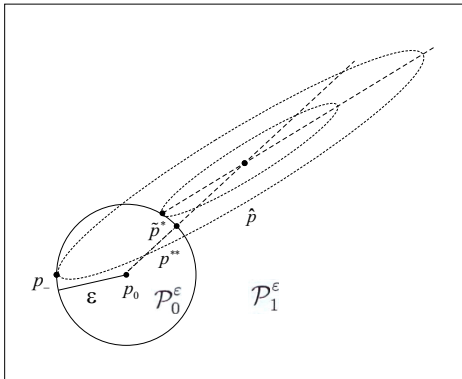


Рис. 1: Графическая схема соотношений между p^* и p^{**} , p_0 , $\hat{p}$, p_-

При $\epsilon \rightarrow 0$ справедливы следующие два асимптотических соотношения между точками $\tilde{p}^* = (\tilde{p}_k^*) \in \mathcal{P}_{(1)}$, $p^{**} = (p_k^{**})$ гиперповерхности S_ϵ ($k = 0, 1, \dots, K-1$):

$$\begin{aligned} \tilde{p}_k^* &= p_k^{**} + \epsilon^2 \left(K\hat{p}_k + (1 - K\hat{p}_k) \left(\hat{\epsilon}^2 + \frac{1 + K\hat{p}_k}{K\hat{\epsilon}^2} \right) \right) + O(\epsilon^3), \\ \tilde{p}_k^* &= p_k^{**} + O(\epsilon^3). \end{aligned} \tag{8}$$

6. Решающее правило на основе логарифмической статистики отношения правдоподобия



Решающее правило с асимптотическим ($n \rightarrow +\infty$) заданным уровнем значимости $\alpha_0 \in (0, 1/2)$ для проверки сложных гипотез $H_0^\varepsilon, \overline{H}_0^\varepsilon$ имеет вид [16]:

$$\text{принимается} \begin{cases} \text{гипотеза } H_0^\varepsilon, & \text{если } (\|\hat{p} - p_0\| - \varepsilon) \times (-2\lambda(X)/(nK))^{1/2} \leq \gamma_n, \\ \text{альтернатива } \overline{H}_0^\varepsilon, & \text{в противном случае,} \end{cases} \quad (9)$$

где $\Phi^{-1}(\cdot)$ – квантильная функция стандартного нормального закона.

Мощность теста (9) при альтернативе $H_1^{\varepsilon+} = \{\|p - p_0\| = \varepsilon_+\}$, $\varepsilon_+ > \varepsilon$ удовлетворяет выражению:

$$w_K(\varepsilon_+) = \Phi \left(\sqrt{\frac{n}{K}} \frac{K(\varepsilon_+ - \varepsilon)}{\sqrt{1 + K\varepsilon_+(e_K - \varepsilon_+)}} - \sqrt{\frac{K\varepsilon(e_K^+ - \varepsilon)}{1 + K\varepsilon_+(e_K - \varepsilon_+)}} \Phi^{-1}(1 - \alpha_0) \right), \quad (10)$$
$$e_K = \frac{1}{\varepsilon^3} \sum_{k=0}^{K-1} \left(p_k - \frac{1}{K} \right)^3.$$

Функция асимптотической мощности (10) монотонно возрастает при увеличении отклонения $\varepsilon_+ \geq \varepsilon$ K -мерного распределения от равномерного и имеет минимальное значение при $\varepsilon_+ = \varepsilon$, равное уровню значимости α_0 :

$$w_K(\varepsilon_+) \geq w_K(\varepsilon) = \alpha_0, \quad \varepsilon_+ \geq \varepsilon.$$

Мощность (10) теста монотонно возрастает при увеличении длительности наблюдений, и решающее правило (9) является состоятельным: $w_K(\varepsilon_+) \xrightarrow{n \rightarrow +\infty} 1, \varepsilon_+ > \varepsilon$.

Серия 1. При $s = 2$, $K = 2^s = 4$, $\varepsilon = 0.225$ имитировалось по 100 реализаций двоичных последовательностей согласно модели (1) – (3) при трех значениях длительности временного ряда $T \in \{400, 800, 1000\}$ и 16 значениях уровня отклонения от равномерного распределения $\|p - p_0\| \in \{0.071, 0.074, 0.113, 0.128, 0.137, 0.158, 0.165, 0.177, 0.183, 0.203, 0.224, 0.261, 0.331, 0.390, 0.423, 0.637\}$ для трех уровней значимости $\alpha_0 \in \{0.01, 0.05, 0.1\}$. По каждой выборке 100 реализаций временного ряда оценивалась вероятность P_1 принятия решения в пользу альтернативы $\bar{H}_0^\varepsilon$. Для иллюстрации полученных результатов на рис. 2, 3 представлены графики зависимости частотной оценки $\hat{P}_1$ от $\|p - p_0\|$ для $\alpha_0 = 0.1$ и $T = 400$; 1000 соответственно.

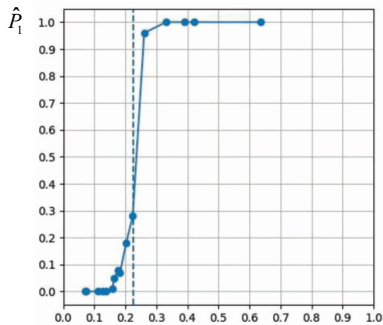


Рис. 2: $\alpha_0 = 0.1$, $T = 400$ $\|p - p_0\|$

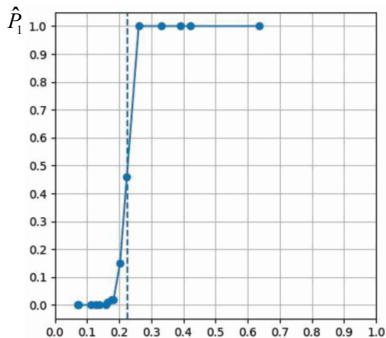


Рис. 3: $\alpha_0 = 0.1$, $T = 1000$ $\|p - p_0\|$

Серия 2. При $s = 8$, $K = 2^s = 256$, $\varepsilon = 0.9/K = 0.00352$ аналогично Серии 1 проводились эксперименты для $T \in \{2^{19}, 2^{20}, 2^{21}\}$, $\alpha_0 \in \{0.05, 0.1\}$. Графики зависимости $\hat{P}_1$ от ε представлены для $T = 2^{19}$; 2^{20} при $\alpha_0 = 0.05$ на рис. 4, 5 соответственно.

Рис. 2 – 5 иллюстрируют состоятельность построенного решающего правила (9). Видно также, что при увеличении длительности T наблюдения последовательности X график зависимости $\hat{P}_1$ от $\|p - p_0\|$ приближается к ступенчатой функции со ступенькой в точке $\|p - p_0\| = \varepsilon$: при $\|p - p_0\| < \varepsilon$ величина $P_1 < \alpha_0$, а при $\|p - p_0\| > \varepsilon$, величина P_1 приближается к единичному значению. Это иллюстрируется при сравнении рис. 4, 5: при увеличении длительности T график $\hat{P}_1$ становится ближе к ступенчатой функции в точке ε .

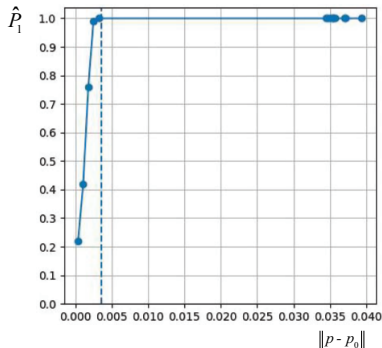


Рис. 4: $\alpha_0 = 0.05, T = 2^{19}$

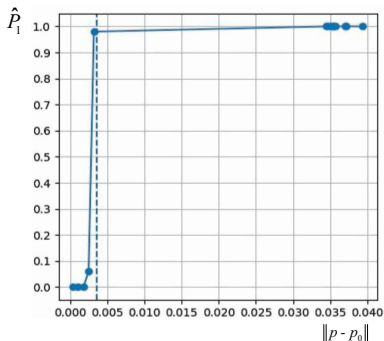


Рис. 5: $\alpha_0 = 0.05, T = 2^{20}$

При решении задачи статистической проверки сложных гипотез об s -мерном равномерном распределении вероятностей двоичных последовательностей получены следующие основные результаты:

1) предложена естественная для приложений модель сложной нулевой гипотезы H_0^ε об s -мерном равномерном распределении – множество K -мерных ($K = 2^s$) дискретных распределений вероятностей, являющихся точками гипершара радиуса ε с центром в точке $(1/K, \dots, 1/K)$, где ε – параметр, задающий допустимое отклонение от равномерного распределения;

2) разработан подход к построению статистического критерия проверки сложных гипотез H_0^ε , $H_1^\varepsilon = \overline{H_0^\varepsilon}$, состоящий из двух этапов: на первом этапе решается нетривиальная задача максимизации логарифмической функции правдоподобия на гиперсфере в R^K (задающей границу H_0^ε) и построено асимптотическое разложение решения этой задачи при $\varepsilon \rightarrow 0$ и фиксированной наблюдаемой последовательности X длины n ; на втором этапе, используя главный член этого разложения, строится критериальная статистика и находится ее асимптотическое распределение при $n \rightarrow +\infty$;

3) построен состоятельный статистический критерий (тест) для проверки сложных гипотез H_0^ε , H_1^ε с заданным асимптотическим уровнем значимости $\alpha_0 \in (0, 1/2)$ и для типичного случая проведен асимптотический анализ его мощности;

4) выполнены компьютерные эксперименты, иллюстрирующие работоспособность построенного решающего правила.

1. Алферов А.П., Зубов А.Ю., Кузьмин А.С., Черемушкин А.В. (2005). *Основы криптографии*. М. – 480 с.
2. NIST SP 800-22: Download Documentation and Software [Electronic resource].
3. Brown R. Dieharder: A Random Number Test Suite [Electronic resource].
4. Jones G. Gjr and random numbers official site [Electronic resource].
5. Харин Ю.С., Агиевич С.В., Васильев Д.В., Матвеев Г.В. (2013). *Криптология*. Минск: БГУ. – 512 с.
6. Zubkov A.M., Serov A.A. (2019). Testing the NIST Statistical Test Suite on artificial pseudorandom sequences. *Математические вопросы криптографии*. – Vol. 10, Вып. 2. – С. 89–96.
7. Харин Ю.С., Петлицкий А.И. (2007). Цепь Маркова s -го порядка с r частичными связями и статистические выводы о ее параметрах. *Дискретная математика*. – Том. 12, Вып. 2. – С. 109–130.
8. Савелов М.П. (2023). Предельное совместное распределение статистик критериев пакета NIST «Monobit Test», «Frequency Test within a Block» и обобщения критерия «Approximate Entropy Test». *Дискретная математика*. – Том. 35, Вып. 2. – С. 93–108.
9. Kowalska K.A. On the revision of NIST 800-22 Test Suite [Electronic resource].
10. Decision to Revise NIST SP 800-22 Rev. 1a [Electronic resource].
11. Палуха В.Ю., Харин Ю.С., Мальцев М.В. [и др.] (2022). Программный комплекс для энтропийного анализа дискретных последовательностей. *Информационные системы и технологии*. – Том. 1. – С. 102–107.
12. Волошко В.А., Трубей А.И. (2022). О мощности тестов многомерной дискретной равномерности, используемых для статистического анализа генераторов случайных последовательностей. *Журнал Белорусского государственного университета. Математика. Информатика*. – № 2. – С. 26–37.
13. Kharin Yu.S., Voloshko V.A., Medved E.A. (2018). Statistical estimation of parameters for binary conditionally nonlinear autoregressive time series. *Mathematical Methods of Statistics*. Vol. 27, 2. – P. 103–118.
14. Kharin Yu.S., Zhuk E.E. (1994). Robustness in statistical pattern recognition under contamination of training samples. *Int. Conf. on Pattern Recognition*. – Vol. 2. – P. 504–506.
15. Fokianos K., Fried R., Kharin Yu., Voloshko V. (2022). Statistical analysis of multivariate discrete-valued time series. *Journal of Multivariate Analysis*. – Vol. 188, 104805. – P. 504–506.
16. Харин Ю.С., Зубков А.М. (2024). О статистической проверке сложных гипотез об s -мерном равномерном распределении вероятностей двоичных последовательностей. *Дискретная математика*. – Том 36, Вып. 1. – С. 116–135.

БЛАГОДАРЮ ЗА ВНИМАНИЕ!