УДК 004.056

ПРИМЕНЕНИЕ ТЕСТОВ НА ОСНОВЕ ЗАКОНА ПОВТОРНОГО ЛОГАРИФМА ДЛЯ ОЦЕНКИ КАЧЕСТВА СЛУЧАЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

А.И. ТРУБЕЙ, В.Ю. ПАЛУХА, И.К. ПИРШТУК, А.А. ОРЛОВ
Учреждение БГУ «НИИ прикладных проблем математики и информатики»
г. Минск, Республика Беларусь

Введение. Существующие батареи статистического тестирования не охватывают некоторые основополагающие законы случайности. Существуют две фундаментальные предельные теоремы о случайных двоичных последовательностях – это центральная предельная теорема и закон повторного логарифма (ЗПЛ). Несколько тестов в батарее NIST SP800-22 включают центральную предельную теорему, в то время как ни один тест в батарее не охватывает закон повторного логарифма.

В докладе описывается методика принятия решений о качестве последовательностей с применением закона повторного логарифма, в которой в качестве статистического расстояния применяется статистика хи-квадрат согласия. Представлены результаты экспериментов по тестированию генераторов псевдослучайных последовательностей, в том числе последовательностей, сгенерированных линейным конгруэнтным генератором, а также разработанным сотрудниками НИИ ППМИ стандартом СТБ 34.101.47-2012 (в режиме счетчика). Идея тестирования с применением закона повторного логарифма (ЗПЛ-тестирования) была впервые предложена в статье [1] и развита в статье [2]. В работе [3] приведена двухэтапная процедура проверки гипотез с применением закона повторного логарифма для частного случая теста многомерной дискретной равномерности по непересекающимся отрезкам (при $L = 1$) – теста Монобит.

1 Тест многомерной дискретной равномерности по непересекающимся отрезкам. Тест многомерной дискретной равномерности по непересекающимся отрезкам (МДРН-тест) является одним из классических тестов, предложенных Кнутом. МДРН-тест предназначен для проверки гипотезы H_0 согласия наблюдаемой последовательности L-векторов с L-мерным дискретным равномерным распределением.

Пусть имеется двоичная последовательность:

$$X = \{x_1, x_2, ..., x_n\}$$

Разбиваем последовательность X на непересекающиеся отрезки длиной L:

$X_t = \{x_{L(t-1)+1}, ..., x_{Lt}\}, 1 \le t \le k$, где $k = \left[n / L \right]$ – число отрезков разбиения.

Проверку гипотез с применением закона повторного логарифма для МДРН-теста будем осуществлять в зависимости от длины L (на основании аппроксимации нормальным распределением схемы независимых испытаний Бернулли и распределения хи-квадрат).

1.1. Аппроксимация нормальным распределением схемы независимых испытаний Бернулли. При $1 \le L < L_0$ по наблюдаемой последовательности X_t строим 2^L вспомогательных двоичных последовательностей:

$$Y_v = \{y_1^{(v)}, ..., y_t^{(v)}, ..., y_k^{(v)}\}, 1 \le t \le k, \text{ где:}$$

$$y_t^{(v)} = 1, \text{ если } X_t = v; v \in \{0,1\}^L;$$

$$y_t^{(v)} = 0, \text{ если } X_t \ne v; v \in \{0,1\}^L.$$

Таким образом мы получим схему независимых испытаний Бернулли, где $p = 1/2^L$ – вероятность положительного исхода в одном испытании, $1 - p$ – вероятность отрицательного исхода. Вычислим 2^L статистик теста: $S_v(k) = \sum_{t=1}^{k} y_t^{(v)}$ (далее индекс v для удобства опустим).

Математическое ожидание и дисперсия статистики имеют следующий вид:

$$E\{S(k)\} = kp = \frac{k}{2^L};\ D\{S(k)\} = kp(1-p) = \frac{k}{2^L}\left(1-\frac{1}{2^L}\right).$$

Тогда при $k \to \infty$ статистика

$$S(k)^* = \frac{S(k)-kp}{\sqrt{kp(1-p)}} = \frac{\left\{S(k)-\frac{k}{2^L}\right\}}{\sqrt{\frac{k}{2^L}\left(1-\frac{1}{2^L}\right)}} \quad (1)$$

распределена асимптотически нормально по закону $N(0,1)$.

Согласно закону повторного логарифма (в общем виде) справедлива формула [3]:

$$\limsup_{k\to\infty} \frac{S(k)^*}{\sqrt{2\ln\ln k}} = \limsup_{k\to\infty} \frac{\frac{S(k)-kp}{\sqrt{kp(1-p)}}}{\sqrt{2\ln\ln k}} = 1. \quad (2)$$

В соответствии с вышеизложенным, для МДРН-теста при проверке гипотез с применением закона повторного логарифма будем использовать статистики:

$$S_{зпл}(k) = \frac{S(k)^*}{\sqrt{2\ln\ln k}} = \frac{\frac{S(k)-kp}{\sqrt{kp(1-p)}}}{\sqrt{2\ln\ln k}} = \frac{\frac{\left\{S(k)-\frac{k}{2^L}\right\}}{\sqrt{\frac{k}{2^L}\left(1-\frac{1}{2^L}\right)}}}{\sqrt{2\ln\ln k}}. \quad (3)$$

Нетрудно заметить, что при $L = 1$ формула (3) превращается в формулу для теста Монобит:

$$S_{зпл}(n) = \frac{S(n)^*}{\sqrt{2\ln\ln n}} = \frac{2S(n)-n}{\sqrt{2n\ln n\ln n}}. \quad (4)$$

Для МДРН-теста меру μ_k^U можно рассчитать следующим образом:

$$\mu_k^U\{(-\infty, z]\} = \Phi\left(z\sqrt{2\ln\ln k}\right) = \sqrt{2\ln\ln k}\int_{-\infty}^{z}\phi\left(s\sqrt{2\ln\ln k}\right)ds. \quad (5)$$

Таким образом, чтобы оценить качество генератора G с применением закона повторного логарифма для МДРН-теста при $1 \le L < L_0$, необходимо:

1. Осуществить генерацию набора $\mathrm{R} \in \Sigma^n$ из $m = 10000$ последовательностей возможно большей длины.

2. Разбить последовательности на непересекающиеся отрезки длиной L.

3. На первом этапе двухэтапной процедуры проверки гипотез вычислить значения статистики $S_{зпл}(k)$ по всем m последовательностям.

4. На втором этапе сравнить между собой вероятностные меры $\mu_k^{\mathrm{R}_k}$ и μ_k^U. Для сравнения будем использовать следующую статистику χ^2 согласия [3]:

$$\chi^2\left(v \in \{0,1\}^L\right) = \sum_{j=1}^{|B|}\frac{\left[v_k^{\mathrm{R}_k}(I_j) - mp_k^{\mathrm{U}}(I_j)\right]^2}{mp_k^{\mathrm{U}}(I_j)}, \quad (6)$$

где $v_k^{\mathrm{R}_k}(I_j)$ – частоты попадания значений статистики $S_{зпл}(k)$ в интервал I_j по всем m последовательностям; $p_k^{\mathrm{U}}(I_j)$ – теоретические вероятности попадания $S_{зпл}(k)$ в интервал I_j.

Полагаем, что генератор G прошел тестирование по МДРН-тесту с применением ЗПЛ, если P-значения статистик $\chi^2\left(v \in \{0,1\}^L\right)$ согласия для всех $v \in \{0,1\}^L$ превышают заданный уровень значимости α, то есть, $P_v \ge \alpha$.

1.2. Аппроксимация нормальным распределением распределения хи-квадрат. При длинах $L \geq L_0$ процесс принятия решения можно оптимизировать и вместо 2^L статистик использовать только одну статистику. С этой целью для $v \in \{0,1\}^L$ вычисляем частоты встречаемости по всем отрезкам:

$$\nu_v^L = \sum_{t=1}^{k} I\left\{ X_t^L = v \right\} \; v \in \{0,1\}^L .$$

Вычисляем статистику:

$$S(k) = \chi_l^2(k) = \sum_{v \in \{0,1\}^L} \frac{\left(\nu_v^L - \frac{k}{2^L} \right)^2}{\frac{k}{2^L}}, \text{ где } l = 2^L - 1 . \quad (7)$$

В силу центральной предельной теоремы, при большом числе степеней свободы распределение случайной величины $S(k) = \chi_l^2(k)$ может быть аппроксимировано нормальным распределением. Более точно, при $l \to \infty$:

$$\mathrm{L}\left\{ S(k)^* \right\} \to \mathrm{N}\,(0,1), \text{где } S(k)^* = \frac{S(k) - l}{\sqrt{2l}} . \quad (8)$$

При этом следует учитывать, что график плотности распределения хи-квадрат, в отличие от биномиального распределения, не является симметричным. Математическое ожидание больше моды этого распределения, потому что правый хвост «тяжелее» левого. Поэтому полученный в результате нормировки график плотности распределения хи-квадрат (8) также не будет симметричным. Однако с увеличением числа степеней свободы графики становятся все более симметричными, то есть значения моды и математического ожидания постепенно сближаются. Необходимо экспериментально определить длину отрезка L_0, для которого при $L \geq L_0$ это будет давать аппроксимацию, достаточную для практических целей.

В таблице 1 приведены значения математических ожиданий M статистики хи-квадрат и их p-значений в зависимости от длин отрезков L.

Таблица 1

Зависимость *p*-значений математических ожиданий статистики хи-квадрат от длин отрезков *L*

Длина отрезка L	5	6	7	8	9	10	11	12
$M = 2^L - 1$	31	63	127	255	511	1023	2047	4095
p-значения	0,4662	0,4763	0,4833	0,4882	0,4916	0,4941	0,4958	0,4971

На рисунке 1 приведены графики плотности распределения хи-квадрат для различного числа степеней свободы.

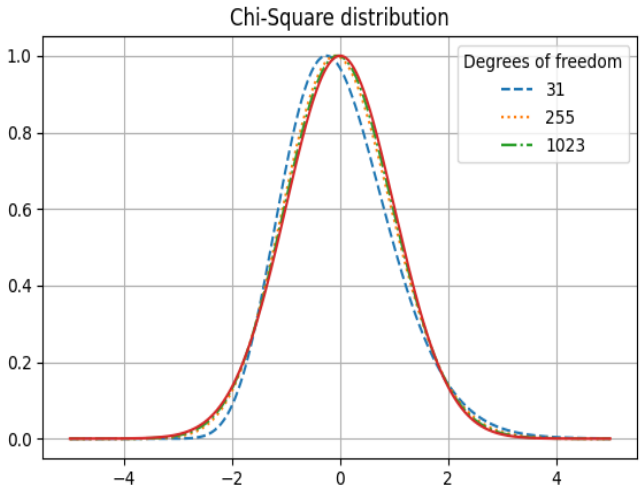

Рис. 1. Графики плотности нормированного распределения хи-квадрат для L = 5, 8, 10, 12

Из таблицы 1 и рисунка 1 видно, что уже при $L \geq 10$ левый и правый «хвосты» распределения хи-квадрат становятся практически симметричными.

Следовательно, для МДРН-теста при $L \geq 10$ можно использовать статистику:

$$S_{зпл}(k) = \frac{S(k)^*}{\sqrt{2\ln\ln k}} = \frac{\frac{S(k)-l}{\sqrt{2l}}}{\sqrt{2\ln\ln k}} = \frac{\frac{\left\{S(k)-\left(2^L-1\right)\right\}}{\sqrt{2\left(2^L-1\right)}}}{\sqrt{2\ln\ln k}}. \quad (9)$$

Полагаем, что генератор G прошел тестирование по МДРН-тесту с применением закона повторного логарифма, если P-значение статистики χ^2 согласия превышает заданный уровень значимости α, то есть, $P_v \geq \alpha$.

2. Экспериментальные результаты. Для проверки гипотезы проведено тестирование 10 000 последовательностей, выработанных соответственно линейным конгруэнтным генератором (ЛКГ) и стандартом СТБ 34.101.47-2012 (в режиме счетчика).

ЛКГ определяется рекуррентным соотношением $X_{n+1} = aX_n + c\left(\mathrm{mod} m\right)$, где X_n – последовательность псевдослучайных чисел, m – модуль, $a, c < m$.

Для любого начального значения X_0 последовательность имеет вид $X_0, X_1, \ldots, X_i, \ldots$ где X_i – двоичное представление целого числа X_i.

Результаты сравнительного тестирования по МДРН-тесту с применением закона повторного логарифма приведены в таблице 2.

Таблица 2

Результаты тестирования по МДРН-тесту последовательностей, выработанных ЛКГ и СТБ 34.101.47-2012 (в режиме счетчика), при L = 12

Объем, GB	ЛКГ			СТБ 34.101.47-2012		
	Степ. своб.	χ^2	P-знач.	Степ. своб.	χ^2	P-знач.
5	41	57.55	0.0446	41	27.30	0.9503
10	41	139.95	$9.8*10^{-13}$	41	45,25	0.2991

Из таблицы 2 видно, что для последовательностей (объемом 5 и 10 GB), сгенерированных в соответствии с СТБ 34.101.47-2012 (в режиме счетчика), выполняется гипотеза H_0 согласия с моделью независимых симметричных испытаний Бернулли на уровне значимости $\alpha = 0.05$. В то время как для последовательности, вырабатываемой линейным конгруэнтным генератором, гипотеза H_0 на данном уровне значимости не выполняется.

Гистограммы частот выборок, полученных с применением ЗПЛ по МДРН-тесту для линейного конгруэнтного генератора и алгоритма генерации псевдослучайных последовательностей в соответствии с СТБ 34.101.47-2012, приведены на рисунках 2, 3.

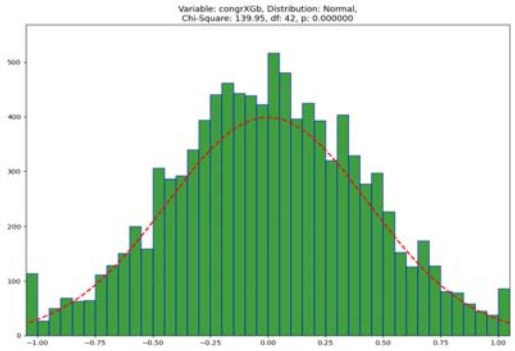

Рис. 2. Гистограмма частот линейного конгруэнтного генератора, 10 GB

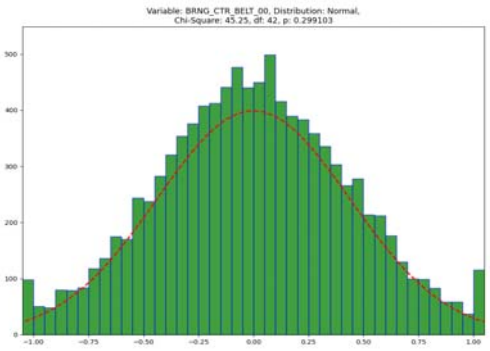

Рис. 3. Гистограмма частот алгоритма СТБ 34.101.47-2012, 10 GB

Список литературы

1. Wang, Y. Linear complexity versus pseudorandomness: on Beth and Dai's result / Y. Wang // Proc. Asiacrypt. – 1999. – P. 288–298.

2. Wang, Y. On statistical distance based testing of pseudorandom sequences and experiments with PHP and Debian OpenSSL / Y. Wang, T. Nicol // Computers & Security. 2015. – Vol. 53. – P. 44–64.

3. Трубей, А. И. Методика тестирования случайных последовательностей на основе статистического расстояния и закона повторного логарифма / А. И.Трубей [и др.] // Проблемы защиты информации : сб. науч. статей.