ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 30 МАТЕМАТИКИ Выпуск 1
2023

В. Г. Михайлов, **В. И. Круглов** (Москва, МИ РАН им. В. А. Стеклова). **Точное вычисление распределения статистики Лемпеля–Зива и построение критериев согласия для неравновероятных случайных двоичных последовательностей.**

УДК 519.233.3

Резюме: Пусть гипотеза H_p состоит в том, что элементы последовательности $X_1, \dots, X_n$ независимы и имеют одинаковое распределение: $\mathbf{P}\{X_i = 1\} = p$, $\mathbf{P}\{X_i = 0\} = 1 - p$, где $p \in (0, 1)$. Предлагаются два критерия согласия для гипотезы H_p, основанные на точном вычислении распределений статистик Лемпеля–Зива.

Для каждого критерия последовательность длины $n = mrT$ делится на блоки длины T, для которых строятся статистики Лемпеля–Зива $W_1(T), \dots, W_{mr}(T)$.

Первый критерий при $r = 2$ основан на статистике $\widetilde{W}(2mT) = (W_1 + \dots + W_m) - (W_{m+1} + \dots + W_{2m})$, распределение которой симметрично относительно нуля.

Статистикой второго критерия является величина $\widetilde{\chi}^2(mrT) = \max_{1 \leqslant k \leqslant m} \chi^2_{(k)}(T)$, где $\chi^2_{(1)}(T), \dots, \chi^2_{(m)}(T)$ — значения статистик хи-квадрат, построенных по наборам $(W_{1,1}(T), \dots, W_{1,r}(T)), \dots, (W_{m,1}(T), W_{m,2}(T), \dots, W_{m,r}(T))$ соответственно.

Для статистик обоих критериев найдены предельные распределения, для статистики первого критерия приведена оценка скорости сходимости к предельному нормальному распределению.

Ключевые слова: критерий Лемпеля–Зива, тестирование генераторов случайных чисел, статистические критерии, вычисление распределений.

1. Введение. Критерий Лемпеля–Зива долгое время входил в набор тестов НИСТ — известный набор статистических критериев для проверки качества случайных и псевдослучайных двоичных последовательностей, разработанный в Национальном Институте Стандартов и Технологии США (NIST) (см. [5]). Позже по ряду причин он был исключен из него (см. [6], [7]).

Статистика критерия Лемпеля–Зива определяется следующим образом. Двоичная последовательность $X_1, X_2, \dots, X_T$ разбивается на подпоследовательности (слова) так, чтобы каждое следующее слово было наименьшим, не совпадающим ни с одним из предыдущих слов. В качестве первого берется пустое слово. Статистикой критерия Лемпеля–Зива является число слов $W(T)$, полученных в результате такого разбиения последовательности $X_1, X_2, \dots, X_T$.

Примеры:

Двоичная последовательность 011101101011 из 12 чисел разбивается на 7 слов $\varnothing, (0), (1), (11), (01), (10), (101)$ и остаток 1, который не считается, так как совпадает с третьим словом.

Двоичная последовательность 010101101010 из 12 чисел разбивается на 7 слов $\varnothing, (0), (1), (01), (011), (010), (10)$ без остатка.

Гипотеза H_p заключается в том, что наблюдаемая последовательность $X_1, X_2, \ldots, X_T$ является реализацией независимых бернуллиевских случайных величин, для которых вероятность единицы равна p, а вероятность нуля равна $1-p$.

Критерий согласия с гипотезой H_p, основанный на статистике Лемпеля–Зива $W(T)$, имеет вид

$$\bigl\{|W(T) - \mathbf{E}W(T)| < C\sqrt{\mathbf{D}W(T)}\bigr\} \Rightarrow H_p \text{ принимается}$$

$$\bigl\{|W(T) - \mathbf{E}W(T)| \geqslant C\sqrt{\mathbf{D}W(T)}\bigr\} \Rightarrow H_p \text{ отвергается,}$$

где C — критический уровень.

В указанном виде тест Лемпеля–Зива предлагался в [5] для проверки гипотезы $H_{0.5}$. Для вычисления значений $\mathbf{E}W(T)$ и $\mathbf{D}W(T)$ предлагалось пользоваться тем, что при $p = 0.5$ и $T \to \infty$

$$\mathbf{E}W(T) \sim \frac{T}{\log_2 T}, \quad \mathbf{D}W(T) \sim \frac{T(C_D + \delta\log_2(T))}{\log_2^3 T}, \tag{1}$$

где $C_D = 0.26600\ldots$ — константа и $\delta(\cdot)$ — медленно меняющаяся функция с нулевым средним, $|\delta(\cdot)| < 10^{-6}$.

Также в [5] отмечалось, что при $p = 0.5$ и $T \to \infty$ статистика Лемпеля–Зива $W(T)$ асимптотически нормальна. Однако, из-за отсутствия достаточной информации о скорости сходимости распределения статистики $W(T)$ к предельному распределению, критерий Лемпеля – Зива было рекомендовано применять лишь на длинных двоичных последовательностях, а именно, когда $T \geqslant 10^6$ (см. [5]). Этот недостаток упоминается как одна из причин исключения критерия из пакета НИСТ (см. [7]).

В работе В. Г. Михайлова [1] был предложен способ точного вычисления распределения величины $W(T)$ при бернуллиевских гипотезах H_p, устроенный следующим образом. При работе алгоритма Лемпеля–Зива последовательно составляется список слов из нулей и единиц. Будем называть этот список словарем и обозначать его $H(n)$, если словарь составлен из n слов. Определим случайную величину $S(n)$, равную сумме длин всех слов, содержащихся в словаре $H(n)$.

Теорема. *Величина $S(n)$ связана со статистикой $W(T)$ равенством*

$$\bigl\{W(T) < n\bigr\} = \bigl\{S(n) > T\bigr\}.$$

Случайная величина $S(n)$ принимает неотрицательные целочисленные значения, при $n = 0, 1, 2$ распределение $S(n)$ задается равенствами

$$\mathbf{P}\{S(0) = 0\} = \mathbf{P}\{S(1) = 0\} = \mathbf{P}\{S(2) = 1\} = 1,$$

а при $n \geq 2$ равенствами

$$\mathbf{P}\{S(n+1) = 0\} = \mathbf{P}\{S(n+1) = 1\} = \cdots = \mathbf{P}\{S(n+1) = n-1\} = 0$$

и рекуррентными формулами

$$\mathbf{P}\{S(n+1) = n + r\} = \\ = \sum\nolimits_{k=0}^{n} (1-p)^k p^{n-k} C_n^k \sum\nolimits_{l=0}^{r} \mathbf{P}\{S(k) = l\}\mathbf{P}\{S(n-k) = r-l\}.$$

По вычисленным распределениям случайных величин $S(n)$ распределение случайной величины $W(T)$ может быть вычислено в соответствии с формулами

$$\mathbf{P}\{W(T)=n\}=\sum\nolimits_{k=0}^{T}\big(\mathbf{P}\{S(n)=k\}-\mathbf{P}\{S(n+1)=k\}\big).$$

2. Вычисленные распределения статистики Лемпеля–Зива. Приведем примеры распределений статистики Лемпеля–Зива $W(T)$ при $p=0.5$ и $T=1000, 2000$.

n	$T=1000$	n	$T=2000$
169	0.0007	300	0.0012
170	0.0089	301	0.0103
171	0.0648	302	0.0564
172	0.2457	303	0.1848
173	0.4098	304	0.3317
174	0.2361	305	0.2915
175	0.0330	306	0.1088
176	0.0006	307	0.0143
		308	0.0005

Распределения $W(T)$, вычисленные для $T=1000, 2000, 3000, \ldots, 8000$ при $p=0.5$, приводятся в [4]. Можно отметить, что при указанных T вычисленные значения $\mathbf{E}W(T)$ и $\mathbf{D}W(T)$ существенно отличаются от главных частей формул (1), рекомендованных НИСТ для приближенного вычисления значений центрировки и нормировки распределения $W(T)$.

T	$\mathbf{E}W_T$	$T/\log_2 T$	$\mathbf{D}W_T$	$0.266T/\log_2^3 T$
1000	172.899	100.343	0.96268	0.26874
2000	304.220	182.385	1.34154	0.40345
3000	425.627	259.723	1.65301	0.51781
4000	541.309	334.286	1.92859	0.62103
5000	653.046	406.910	2.18096	0.71686
6000	761.811	478.059	2.41656	0.80727
7000	868.213	548.025	2.63918	0.89348
8000	972.665	617.008	2.85136	0.97628

Распределения $W(T)$, вычисленные для $T=1000$ и $p=0.1, 0.5, 0.9$, приведены в [3].

3. Критерий типа хи-квадрат. Пусть задана выборка $X_1, \ldots, X_n$, состоящая из $n=mrT$ нулей и единиц. Разобьем эту выборку на mr непересекающихся блоков длиной T и для каждой из них вычислим величину $W(T)$, получив, таким образом, mr значений статистики Лемпеля–Зива:

$$W_{1,1}(T), W_{1,2}(T), \ldots, W_{1,r}(T),$$

$$\ldots$$

$$W_{m,1}(T), W_{m,2}(T), \ldots, W_{m,r}(T).$$

Принцип построения критерия не зависит от параметров p и T. В качестве примера рассмотрим значения $T=1000$ и $p=0.1$ или, что эквивалентно, $p=0.9$. Соответствующее распределение $W(T)$ выглядит следующим образом:

n	$\mathbf{P}\{W(T)=n\}$	n	$\mathbf{P}\{W(T)=n\}$	n	$\mathbf{P}\{W(T)=n\}$
89	0.0001293	100	0.03624	111	0.04782
90	0.0002615	101	0.04736	112	0.03549
91	0.0005096	102	0.0592	113	0.02478
92	0.0009567	103	0.07069	114	0.01625
93	0.00173	104	0.08054	115	0.009986
94	0.00301	105	0.08744	116	0.005744
95	0.00504	106	0.09032	117	0.003086
96	0.008113	107	0.08865	118	0.001545
97	0.01255	108	0.08254	119	0.0007202
98	0.01863	109	0.0728	120	0.0003117
99	0.02654	110	0.06072	121	0.000125

Множество возможных значений величины $W(T)$ разобьем на $N = 12$ интервалов

$$\Delta_1 = \{0, \dots, 100\}, \Delta_2 = \{101\}, \Delta_3 = \{102\}, \Delta_4 = \{103\},$$

$$\Delta_5 = \{104\}, \Delta_6 = \{105\}, \Delta_7 = \{106\}, \Delta_8 = \{107\},$$

$$\Delta_9 = \{108\}, \Delta_{10} = \{109\}, \Delta_{11} = \{110\}, \Delta_{12} = \{111, 112, \dots\},$$

тогда, в соответствии с вычисленным распределением $W(T)$, вероятности $p_j^0 = \mathbf{P}\{W(T) \in \Delta_j\}$ попадания значения $W(T)$ в эти интервалы равны

$$p_1^0 = 0.113825, p_2^0 = 0.0473614, p_3^0 = 0.0592027, p_4^0 = 0.0706947,$$

$$p_5^0 = 0.0805426, p_6^0 = 0.0874356, p_7^0 = 0.0903182, p_8^0 = 0.0886453,$$

$$p_9^0 = 0.0825413, p_{10}^0 = 0.072801, p_{11}^0 = 0.0607218, p_{12}^0 = 0.145911.$$

Для каждого из $k = 1, \dots, m$ и соответствующих значений $W_{k,1}(T)$, $W_{k,2}(T), \dots, W_{k,r}(T)$, вычислим величины $v_{k,1}(T), v_{k,2}(T), \dots, v_{k,12}(T)$, где величина $v_{k,j}$ равна количеству величин $W_{k,i}(T)$, попавших в интервал Δ_j. Построим статистики $\chi_k^2(rT) = \sum_{j=1}^N \frac{(v_{k,j} - np_j^0)^2}{np_j^0}$, где, напомним, $N = 12$, и найдем статистику $\widetilde{\chi}^2(mrT) = \max_{1 \leqslant k \leqslant m} \chi_k^2(T)$. Для уровня значимости α определим квантиль $C(N-1, \alpha^{1/m})$ равенством $\chi_{N-1}^2(C(N-1, \alpha^{1/m})) = \alpha^{1/m}$, где $\chi_{N-1}^2(x)$ — функция распределения хи-квадрат с $N-1$ степенью свободы, и зададим критерий правилами

$$\{\widetilde{\chi}^2(mrT) < C(N-1, \alpha^{1/m})\} \Rightarrow H_p \text{ принимается},$$

$$\{\widetilde{\chi}^2(mrT) \geqslant C(N-1, \alpha^{1/m})\} \Rightarrow H_p \text{ отвергается}.$$

Теорема 2. *Пусть выполнена гипотеза H_p о том, что последовательность $X_1, \dots, X_{mrT}$ является реализацией независимых бернуллиевских случайных величин, для которых вероятность успеха равна p. Пусть параметры m, T и N фиксированы. Тогда при $r \to \infty$*

$$\mathbf{P}\{\widetilde{\chi}^2(mrT) < x\} \to 1 - \big(1 - \chi_{N-1}^2(x)\big)^m$$

и

$$\mathbf{P}\{\widetilde{\chi}^2(mrT) \geqslant C(N-1, \alpha^{1/m})\} \to \alpha.$$

4. Критерий с суммированием. Разобьем выборку $X_1, \ldots, X_{2mT}$ на $2m$ непересекающихся блоков длины T каждый и для каждого такого блока найдем значение статистики $W(T)$. Обозначим полученные величины $W_1, W_2, \ldots, W_{2m}$ и построим по ним статистику $\widetilde{W}(2mT) = (W_1 + W_2 + \cdots + W_m) - (W_{m+1} + W_{m+2} + \cdots + W_{2m})$.

Распределение статистики $\widetilde{W}(2mT)$ симметрично относительно нуля, $\mathbf{E}\widetilde{W}(2mT) = 0$ и $\mathbf{D}\widetilde{W}(2mT) = 2m\mathbf{D}W(T)$.

Критерий согласия с гипотезой H_p, основанный на статистике $\widetilde{W}(2mT)$, имеет вид

$$\left\{|\widetilde{W}(2mT)| < C\sqrt{\mathbf{D}\widetilde{W}(2mT)\}}\right\} \Rightarrow H_p \text{ принимается,}$$

$$\left\{|\widetilde{W}(2mT)| \geqslant C\sqrt{\mathbf{D}\widetilde{W}(2mT)}\right\} \Rightarrow H_p \text{ отвергается.}$$

Здесь C — критический уровень, который в силу следствия из теоремы 3 может быть выбран по правилу $2(1 - \Phi(C)) \approx \varepsilon$. Однако, для вычисленного при заданных значениях m и T точного распределения статистики $\widetilde{W}(2mT)$ мы можем точно вычислить вероятности ошибок первого и второго уровня для каждой константы C.

С целью вычисления точного распределения случайной величины $\widetilde{W}(2mT)$ перепишем ее в виде $\widetilde{W}(2mT) = \sum_{i=1}^{m}(W_i(T) - W_{i+m}(T)) = \sum_{i=1}^{m} V_i(2T)$, где $V_i(2T) = W_i(T) - W_{i+m}(T)$, тогда случайная величина $\widetilde{W}(2mT)$ состоит из m независимых и одинаково распределенных слагаемых $V_i(2T)$.

Если вероятности $\mathbf{P}\{W(T) = k\}$ вычислены для $0 \leq k \leq b$, то не зависящие от $i = 1, \ldots, m$ вероятности $\mathbf{P}\{V_i(2T) = k\}$ для $-b \leq k \leq b$ могут быть вычислены по формуле $\mathbf{P}\{V_i(2T) = k\} = \sum \mathbf{P}\{W(T) = l\}\mathbf{P}\{W(T) = l - k\}$, в которой суммирование проводится по всем таким l, что $\max(0, k) \leq l \leq \min(b, b + k)$.

Распределения $V(2T)$ и величины $\mathbf{E}|V(2T)|$, $\mathbf{D}V(2T)$, $\mathbf{E}|V(2T)|^3$, вычисленные при $T = 1000, 2000, \ldots, 6000$ и при различных значениях вероятности p, приведены в [3].

После вычисления распределения $V(2T)$ распределение случайной величины $\widetilde{W}(2mT)$ может быть вычислено как m-кратная свертка распределения $V(2T)$.

4.1. О точности нормальной аппроксимации для $\widetilde{W}(2mT)$. Точность нормальной аппроксимации распределения статистики $\widetilde{W}(2mT)$ может быть оценена с помощью неравенства Берри–Эссеена.

Теорема 3. *Для функции распределения величины $\widetilde{W}(2mT)$ выполнено неравенство*

$$\sup_{-\infty<x<\infty}\left|\mathbf{P}\left\{\widetilde{W}(2mT)\Big/\sqrt{\mathbf{D}\widetilde{W}(2mT)} < x\right\} - \Phi(x)\right| \leqslant \frac{C_1\mathbf{E}|V(2T)|^3}{(2m\mathbf{D}W(T))^{3/2}}. \qquad (2)$$

Как указано в [2], для константы C_1 справедлива оценка $C_1 \leqslant 0.4774$.

Следствие. *Если $m \to \infty$, то $\mathbf{P}\left\{\widetilde{W}(2mT)\Big/\sqrt{\mathbf{D}\widetilde{W}(2mT)} < x\right\} \to \Phi(x)$ при всех $-\infty < x < \infty$.*

Значения правой части неравенства (2) были вычислены для $T = 1000, 2000, \ldots, 6000$, $m = 1000, 2000$ и $p = 0.1, 0.2, \ldots, 0.9$. Эти значения существенно зависят от числа блоков m и практически не зависят от размера блока T или вероятности p: при всех рассмотренных значениях T и p значение правой части (2) приблизительно равно $2.41 \cdot 10^{-5}$ при $m = 1000$ и приблизительно равно $8.5 \cdot 10^{-6}$ при $m = 2000$.

4.2. Критерий с суммированием: точные значения. Для выборки $X_1, \ldots, X_{2mT}$ вычислим статистику $\widetilde{W}(2mT)$ и при выбранном критическом уровне l будем проверять гипотезу H_p следующим образом:

$$\left\{|\widetilde{W}(2mT)| < l\right\} \Rightarrow H_p \text{ принимается}, \quad \left\{|\widetilde{W}(2mT)| \geqslant l\right\} \Rightarrow H_p \text{ отвергается}.$$

Тогда вероятность отвергнуть гипотезу H_p, если она верна, равна

$$\alpha_l = \mathbf{P}\{H_p \text{ отвергли}|H_p \text{ верна}\} = \mathbf{P}\left\{|\widetilde{W}(2mT)| \geqslant l|H_p\right\} = \\ = 1 - \mathbf{P}\left\{|\widetilde{W}(2mT)| < l|H_p\right\}.$$

Приведем значения вероятностей α_l при различных значениях критического уровня l для равновероятного распределения $p = 0.5$, размера блока $T = 1000$ и $m = 10$.

l	α_l
1	0.9090
2	0.7317
3	0.5678
4	0.4238

l	α_l
5	0.3038
6	0.2089
7	0.1376
8	0.0867

l	α_l
9	0.0522
10	0.0300
11	0.0165
12	0.0086

l	α_l
13	0.0043
14	0.0020
15	0.0009
16	0.0004

Аналогичные критерии могут быть построены тем же способом для проверки гипотезы H_p при любых значениях $p \in (0, 1)$.

СПИСОК ЛИТЕРАТУРЫ

1. *Михайлов В. Г.* Формулы для вычисления распределений статистики Лемпеля–Зива и связанных с ней статистик. — Обозрение прикл. и промышл. матем., 2007, т. 15, в. 3, с. 461–473.
2. *Тюрин И. С.* Уточнения остаточного члена в теореме Ляпунова. — Теория вероятн. и ее примен., 2011, т. 56, в. 3, с. 808–811.
3. *Kruglov V. I.* Two Lempel–Ziv goodness-of-fit tests for nonequiprobable random binary sequences. — Матем. вопр. криптогр., 2023, т. 14, в. 2, с. 97–110.
4. *Mikhailov V. G., Kruglov V. I.* Two variants of Lempel–Ziv test for binary sequences. — Матем. вопр. криптогр., 2022, т. 13, в. 3, с. 93–106.
5. *Rukhin A. et al.* A statistical test suite for random and pseudorandom number generators for cryptographic applications. — NIST Special Publication 800-22, 2000.
6. *Rukhin A. et al.* A statistical test suite for random and pseudorandom number generators for cryptographic applications. — NIST Special Publication 800-22r1a, 2010.
7. https://crypto.stackexchange.com/questions/129/why-did-nist-remove-the-lempel-ziv-compression-test-from-the-statistical-test-su

УДК 519.212.2

Mikhailov V. G., ***Kruglov V. I.*** (Moscow, Steklov Mathematical Institute of RAS)**. Exact computation of Lempel–Ziv statistics distributions and construction of goodness-of-fit tests for nonequiprobable random binary sequences.**

Abstract: Consider the hypothesis H_p that elements of the sequence $X_1, \ldots, X_n$ are independent and identically distributed: $\mathbf{P}\{X_i = 1\} = p$, $\mathbf{P}\{X_i = 0\} = 1 - p$, where $p \in (0, 1)$. We propose two goodness-of-fit tests for the hypothesis H_p that are based on the possibility of exact computation of Lempel–Ziv statistics distributions.

For each test a sequence of length $n = mrT$ is divided into blocks of length T, for these blocks Lempel–Ziv statistics $W_1(T), \ldots, W_{mr}(T)$ are computed.

The first test for $r = 2$ is based on the statistic $\widetilde{W}(2mT) = (W_1 + \ldots + W_m) - (W_{m+1} + \cdots + W_{2m})$, its distribution is symmetric about zero.

The statistic of the second test is $\widetilde{\chi}^2(mrT) = \max_{1 \leqslant k \leqslant m} \chi^2_{(k)}(T)$, where $\chi^2_{(1)}(T), \ldots, \chi^2_{(m)}(T)$ are values of chi-square statistics computed for $(W_{1,1}(T), \ldots, W_{1,r}(T)), \ldots, (W_{m,1}(T), W_{m,2}(T), \ldots, W_{m,r}(T))$ correspondingly.

For statistics of both tests limit distributions are found, for the statistic of the first test the rate of convergence to the limit normal distribution is given.

Keywords: Lempel – Ziv test, RNG testing, statistical test, computation of distributions.