УДК 004.056.5:519.254

СРАВНЕНИЕ СТАТИСТИК ТЕСТОВ СЕРИЙ И АППРОКСИМИРОВАННОЙ ЭНТРОПИИ

Н.Г. КИЕВЕЦ, А.И. КОРЗУН

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220013, Беларусь

Поступила в редакцию 27 декабря 2013

Показывается, что выражения для тестовых статистик серий и аппроксимированной энтропии являются эквивалентными при рассмотрении пересекающихся серий одинаковых длин и при длинах исследуемых последовательностей $n \to \infty$. Экспериментально подтверждено совпадение результатов, полученных при тестировании последовательностей различных длин.

Ключевые слова: тестовая статистика, тест серий, тест аппроксимированной энтропии.

Введение

На сегодняшний день остается открытой проблема зависимости статистических тестов, используемых для тестирования генераторов случайных чисел (ГСЧ) [1]. Применение зависимых тестов приводит к неоправданному увеличению временных затрат и может привести к неверным выводам о качестве генератора [2].

Одной из наиболее используемых систем тестирования ГСЧ является система NIST [3]. Данная система включает тест серий и тест аппроксимированной энтропии, тестовая статистика которых вычисляется на основе частот появления в исследуемой последовательности пересекающихся серий различной длины.

В данной работе приводится сравнение тестовых статистик теста серий и теста аппроксимированной энтропии для выявления зависимости тестов и подтверждение полученных выводов экспериментальными данными.

Анализ статистик теста серий и аппроксимированной энтропии

Целью применения статистических тестов является принятие либо отклонение выдвинутой гипотезы о случайности исследуемой последовательности. В каждом тесте рассчитывается статистика, на основе которой определяется значение вероятности P, характеризующей приемлемость выдвинутой гипотезы. В системе NIST для прохождения теста необходимо выполнение условия: $P \geq 0{,}01$. В тестах серий и аппроксимированной энтропии рассчитываются статистики χ^2, имеющие распределение «хи-квадрат». При таком распределении статистики вероятность P рассчитывается по формуле:

$$P = \frac{\int_{\chi^2/2}^{\infty} t^{\frac{K}{2}-1} e^{-t} dt}{\int_{0}^{\infty} t^{\frac{K}{2}-1} e^{-t} dt}, \quad (1)$$

где K – число степеней свободы распределения.

Система NIST содержит два теста серий. В данной статье рассматривается тест, в котором рассчитывается статистика:

$$\chi^2 = \frac{2^{m+1}}{n}\sum_{k=0}^{2^{m+1}-1} w_k^2 - \frac{2^m}{n}\sum_{i=0}^{2^m-1} v_i^2, \qquad (2)$$

где $(m+1)$ и m – длины рассматриваемых пересекающихся серий, n – длина исследуемой последовательности, w_k – число появлений в последовательности серии k-го вида длины $(m+1)$, v_i – число появлений в последовательности серии i-го вида длины m.

В тесте аппроксимированной энтропии рассчитывается статистика: $\chi^2 = 2n(\ln 2 - ApEn(m))$, (3)

где

$ApEn(m) = \varphi^{(m)} - \varphi^{(m+1)}$ – (4)

аппроксимированная энтропия, $\varphi^{(m)} = \sum_{i=0}^{2^m-1}\frac{v_i}{n}\ln\left(\frac{v_i}{n}\right)$, $\varphi^{(m+1)} = \sum_{k=0}^{2^{m+1}-1}\frac{w_k}{n}\ln\left(\frac{w_k}{n}\right)$.

Заменим в (4) величины $\varphi^{(m)}$ и $\varphi^{(m+1)}$ эквивалентными соотношениями при $n\to\infty$ [4]:

$\varphi^{(m)} \sim -m\ln 2 + \frac{2^m}{2n}\sum_{i=0}^{2^m-1} Z_i^2$, где $Z_i = \sqrt{n}\left(\frac{v_i}{n} - \frac{1}{2^m}\right)$;

$\varphi^{(m+1)} \sim -(m+1)\ln 2 + \frac{2^{m+1}}{2n}\sum_{k=0}^{2^{m+1}-1} Y_k^2$, где $Y_k = \sqrt{n}\left(\frac{w_k}{n} - \frac{1}{2^{m+1}}\right)$.

Получаем выражение для аппроксимированной энтропии:

$$ApEn(m) = -m\ln 2 + \frac{2^m}{2n}\sum_{i=0}^{2^m-1} Z_i^2 + (m+1)\ln 2 - \frac{2^{m+1}}{2n}\sum_{k=0}^{2^{m+1}-1} Y_k^2 = -m\ln 2 + \frac{2^m}{2n}\sum_{i=0}^{2^m-1} n\left(\frac{v_i}{n} - \frac{1}{2^m}\right)^2 +$$

$$+m\ln 2 + \ln 2 - \frac{2^{m+1}}{2n}\sum_{k=0}^{2^{m+1}-1} n\left(\frac{w_k}{n} - \frac{1}{2^{m+1}}\right)^2 = \ln 2 + \frac{2^m}{2n}\sum_{i=0}^{2^m-1}\frac{1}{n}\left(v_i - \frac{n}{2^m}\right)^2 - \frac{2^{m+1}}{2n}\sum_{k=0}^{2^{m+1}-1}\frac{1}{n}\left(w_k - \frac{n}{2^{m+1}}\right)^2.$$

Подставив полученное выражение для $ApEn(m)$ в (2), имеем:

$$\chi^2 = 2n\left(\ln 2 - \left(\ln 2 + \frac{2^m}{2n}\sum_{i=0}^{2^m-1}\frac{1}{n}\left(v_i - \frac{n}{2^m}\right)^2 - \frac{2^{m+1}}{2n}\sum_{k=0}^{2^{m+1}-1}\frac{1}{n}\left(w_k - \frac{n}{2^{m+1}}\right)^2\right)\right) =$$

$$= 2n\left(\frac{2^{m+1}}{2n}\sum_{k=0}^{2^{m+1}-1}\frac{1}{n}\left(w_k - \frac{n}{2^{m+1}}\right)^2 - \frac{2^m}{2n}\sum_{i=0}^{2^m-1}\frac{1}{n}\left(v_i - \frac{n}{2^m}\right)^2\right) = \frac{2^{m+1}}{n}\sum_{k=0}^{2^{m+1}-1}\left(w_k - \frac{n}{2^{m+1}}\right)^2 -$$

$$-\frac{2^m}{n}\sum_{i=0}^{2^m-1}\left(v_i - \frac{n}{2^m}\right)^2 = \frac{2^{m+1}}{n}\sum_{k=0}^{2^{m+1}-1}\left({w_k}^2 - \frac{2nw_k}{2^{m+1}} + \left(\frac{n}{2^{m+1}}\right)^2\right) - \frac{2^m}{n}\sum_{i=0}^{2^m-1}\left({v_i}^2 - \frac{2n}{2^m} + \left(\frac{n}{2^m}\right)^2\right) =$$

$$= \frac{2^{m+1}}{n}\sum_{k=0}^{2^{m+1}-1}{w_k}^2 - \frac{2^{m+1}}{n}\cdot\frac{2n}{2^{m+1}}\sum_{k=0}^{2^{m+1}-1} w_k + \frac{2^{m+1}}{n}\cdot 2^{m+1}\cdot\left(\frac{n}{2^{m+1}}\right)^2 - \frac{2^m}{n}\sum_{i=0}^{2^m-1}{v_i}^2 + \frac{2^m}{n}\cdot\frac{2n}{2^m}\sum_{i=0}^{2^m-1} v_i -$$

$$-\frac{2^m}{n}\cdot 2^m\cdot\left(\frac{n}{2^m}\right)^2 = \frac{2^{m+1}}{n}\sum_{k=0}^{2^{m+1}-1}{w_k}^2 - 2\sum_{k=0}^{2^{m+1}-1} w_k + n - \frac{2^m}{n}\sum_{i=0}^{2^m-1}{v_i}^2 + 2\sum_{i=0}^{2^m-1} v_i - n.$$

Так как $\sum_{k=0}^{2^{m+1}-1} w_k = n$ и $\sum_{i=0}^{2^m-1} v_i = n$, то получаем выражение для χ^2:

$$\chi^2 = \frac{2^{m+1}}{n}\sum_{k=0}^{2^{m+1}-1}{w_k}^2 - 2n + n - \frac{2^m}{n}\sum_{i=0}^{2^m-1}{v_i}^2 + 2n - n = \frac{2^{m+1}}{n}\sum_{k=0}^{2^{m+1}-1}{w_k}^2 - \frac{2^m}{n}\sum_{i=0}^{2^m-1}{v_i}^2. \qquad (5)$$

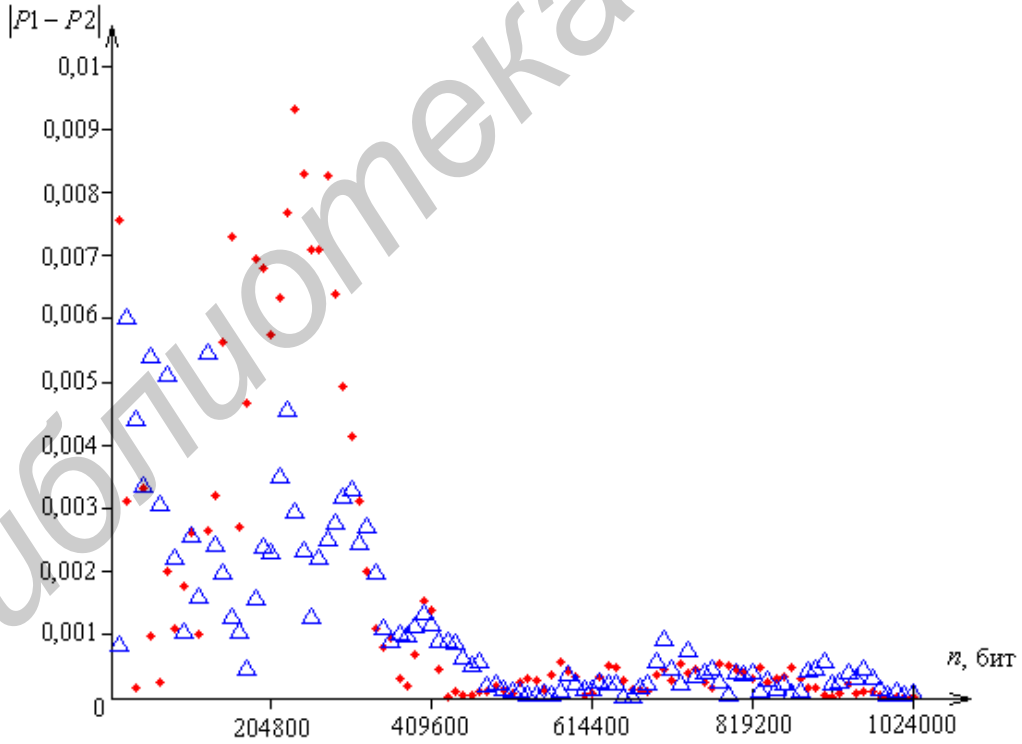
Получилось, что выражение (5) для χ^2 теста аппроксимированной энтропии совпадает с выражением (2) для χ^2 теста серий. Так как для обоих тестов величины K равны и $K = 2^m$, получаем равные значения P, что видно из формулы (1).

Таким образом, показано, что при $n \to \infty$ значения вероятностей P для теста аппроксимированной энтропии и теста серий совпадают, что говорит об их зависимости.

Экспериментальные исследования

С целью подтверждения правильности принятых допущений проведены экспериментальные исследования. Проверялось условие $|P1 - P2| \to 0$ при $n \to \infty$, где $P1$ - значение вероятности, полученное для теста серий; $P2$ – значение вероятности, полученное для теста аппроксимированной энтропии. Для этого из двух электронных пластиковых карт (ЭПК) с операционной системой MINOS с помощью аппаратно-программного комплекса [5] извлечено по 1000 ключей длиной 1024 бит.

Для каждой из ЭПК №1 и №2 сформировано 1000 последовательностей a_i, где i – номер последовательности. При этом i-я последовательность сформирована из i ключей, представленных в двоичном коде и записанных в поток данных побитно. Полученные последовательности протестированы с использованием MATLAB [6]. Задан параметр $m = 3$. Для достижения наглядности на рисунке представлено каждое десятое значение $|P1 - P2|$. Видно практически полное совпадение значений $P1$ и $P2$ для n>500000 бит. Заметим, что даже при относительно небольших значениях n $|P1 - P2|$ не превышает 1% от максимально возможного значения, равного 1.



Зависимость величины $|P1 - P2|$ от длины тестируемой последовательности n:
♦ – для последовательностей из ЭПК №1, △ – для последовательностей из ЭПК №2

В табл. 1 содержатся значения вероятностей $P1$, $P2$ и разности $|P1 - P2|$ для каждой 50-й последовательности. Таблица позволяет увидеть, как изменяются $P1$ и $P2$ с ростом n. Отчетливо видна тенденция приближения значений $P1$ и $P2$ при увеличении n. Тенденция характерна для последовательностей, выработанных генераторами различных карт.

Таблица 1. **Значения вероятностей *P*1, *P*2 и разностей |*P*1-*P*2|, полученных при тестировании последовательностей, длины которых кратны 51200 битам**

i	*n*, бит	Последовательности из ЭПК №1			Последовательности из ЭПК №2		
		*P*1	*P*2	|*P*1-*P*2|	*P*1	*P*2	|*P*1-*P*2|
50	51200	0,7111	0,7121	0,0010	0,2219	0,2272	0,0054
100	102400	0,1940	0,1966	0,0026	0,7877	0,7902	0,0026
150	153600	0,2893	0,2966	0,0073	0,3343	0,3355	0,0013
200	204800	0,3607	0,3665	0,0057	0,2723	0,2746	0,0023
250	256000	0,1333	0,1404	0,0071	0,0736	0,0749	0,0013
300	307200	0,4354	0,4396	0,0041	0,2786	0,2819	0,0033
350	358400	0,3404	0,3414	0,0009	0,5725	0,5734	0,0009
400	409600	0,4130	0,4116	0,0014	0,6240	0,6252	0,0011
450	460800	0,6164	0,6164	0,0001	0,7383	0,7388	0,0005
500	51200	0,5342	0,5341	0,0001	0,6745	0,6746	0,0001
550	563200	0,5753	0,5749	0,0004	0,7490	0,7489	0,0000
600	614400	0,6085	0,6084	0,0001	0,5319	0,5318	0,0001
650	665600	0,5776	0,5775	0,0001	0,4233	0,4234	0,0000
700	716800	0,6556	0,6553	0,0003	0,3216	0,3220	0,0005
750	768000	0,5893	0,5891	0,0002	0,4425	0,4430	0,0005
800	819200	0,7549	0,7546	0,0003	0,5677	0,5673	0,0004
850	870400	0,6357	0,6352	0,0005	0,5711	0,5710	0,0001
900	921600	0,8916	0,8916	0,0000	0,7058	0,7060	0,0002
950	972800	0,8085	0,8084	0,0001	0,5854	0,5857	0,0003
1000	1024000	0,9691	0,9691	0,0000	0,6798	0,6797	0,0001

В табл. 2 и 3 представлены значения $P1$, $P2$ и $|P1-P2|$ для последовательностей, длины которых отличаются на 4096 бит. Из таблиц видно, что увеличение 54-й последовательности на 45056 бит привело к более существенному изменению вероятностей $P1$ и $P2$, чем увеличение 954-й последовательности на то же количество бит. Это объясняется тем, что чем длиннее последовательность, тем незначительнее меняет ее добавление некоторого количества бит. Табл. 2 и 3 показывают, что значения $|P1-P2|$ на порядок меньше для длинных последовательностей с номерами от 954 до 998, чем для относительно коротких последовательностей с номерами от 54 до 98.

Таблица 2. **Значения вероятностей *P*1, *P*2 и разностей |*P*1-*P*2|, полученных при тестировании последовательностей длиной от 55296 до 1003523 бит**

i	*n*, бит	Последовательности из ЭПК №1			Последовательности из ЭПК №2		
		*P*1	*P*2	|*P*1-*P*2|	*P*1	*P*2	|*P*1-*P*2|
54	55296	0,8295	0,8295	0,0000	0,3527	0,3561	0,0033
58	59392	0,8041	0,8035	0,0006	0,3172	0,3201	0,0028
62	63488	0,8206	0,8212	0,0006	0,3108	0,3149	0,0041
66	67584	0,6635	0,6648	0,0013	0,3638	0,3684	0,0046
70	71680	0,6594	0,6614	0,0020	0,3825	0,3876	0,0051
74	75776	0,5070	0,5098	0,0028	0,4841	0,4878	0,0037
78	79848	0,4387	0,4395	0,0008	0,5987	0,6012	0,0025
82	83944	0,2208	0,2231	0,0023	0,5573	0,5599	0,0026
86	88040	0,3061	0,3084	0,0022	0,7140	0,7155	0,0014
90	92136	0,3209	0,3227	0,0018	0,7921	0,7931	0,0010
94	96208	0,3178	0,3197	0,0019	0,7127	0,7148	0,0021
98	100352	0,2481	0,2517	0,0035	0,7702	0,7723	0,0021

Таблица 3. **Значения вероятностей *P*1, *P*2 и разностей |*P*1-*P*2|, полученных при тестировании последовательностей длиной от 97896 до 1021952 бит**

i	*n*, бит	Последовательности из ЭПК №1			Последовательности из ЭПК №2		
		*P*1	*P*2	$\lvert P1\text{-}P2\rvert*10^4$	*P*1	*P*2	$\lvert P1\text{-}P2\rvert*10^3$
954	976896	0,7825	0,7825	0,2002	0,6182	0,6185	0,2769
958	980992	0,8166	0,8167	0,5935	0,6134	0,6137	0,3073
962	985088	0,8721	0,8722	0,4031	0,6486	0,6486	0,0149
966	989184	0,8905	0,8905	0,2937	0,6629	0,6631	0,1488
970	993280	0,9091	0,9091	0,0211	0,7031	0,7031	0,0462
974	997376	0,9337	0,9337	0,1770	0,7018	0,7019	0,0613
978	1001472	0,9371	0,9372	0,1981	0,7056	0,7057	0,1165
982	1005568	0,9411	0,9411	0,3784	0,6879	0,6880	0,0747
986	1009664	0,9385	0,9385	0,1474	0,6591	0,6590	0,0699
990	1013760	0,9464	0,9465	0,2898	0,7237	0,7237	0,0493
994	1017856	0,9627	0,9627	0,2722	0,6961	0,6961	0,0344
998	1021952	0,9771	0,9771	0,0612	0,7000	0,6999	0,0930

На рисунке видна область значений $|P1-P2|$ при сравнительно небольших длинах последовательностей, в пределах которой значение $|P1-P2|$ может быть больше значения 0,01. Для более детального рассмотрения этой области значений был проведен эксперимент. Из ЭПК MINOS извлечено 1152 ключа длиной 1024 бит. Полученный массив ключей разбит на 48 групп по 24 ключа. Из ключей каждой группы, представленных в двоичном коде, сформирована битовая последовательность. Получено 48 последовательностей, каждая длиной 24576 бит. В табл. 2 представлены значения $P1$ и $P2$ для всех последовательностей. Все значения $|P1-P2|$ принадлежат диапазону $[1{,}3400\cdot10^{-5}; 0{,}0213]$.

Таблица 4. **Значения вероятностей *P*1 и *P*2, полученных при тестировании 48 последовательностей длиной 24576 бит**

№	*P*1	*P*2	№	*P*1	*P*2	№	*P*1	*P*2	№	*P*1	*P*2
1	0,0051	0,0051	13	0,7979	0,7978	25	0,8678	0,8648	37	0,5647	0,5703
2	0,3817	0,3917	14	0,9244	0,9223	26	0,5701	0,5682	38	0,5069	0,5139
3	0,2143	0,2113	15	0,8709	0,8707	27	0,5424	0,5434	39	0,9875	0,9876
4	0,7966	0,8008	16	0,1032	0,1013	28	0,2342	0,2350	40	0,0993	0,0999
5	0,3879	0,3848	17	0,0169	0,0197	29	0,8167	0,8169	41	0,4887	0,4954
6	0,2719	0,2740	18	0,1748	0,1755	30	0,9330	0,9324	42	0,5459	0,5482
7	0,8588	0,8587	19	0,6879	0,6838	31	0,0173	0,0183	43	0,4717	0,4717
8	0,1949	0,1889	20	0,5769	0,5744	32	0,3509	0,3477	44	0,6026	0,6088
9	0,4957	0,4912	21	0,3100	0,3113	33	0,2418	0,2446	45	0,6871	0,6880
10	0,9507	0,9513	22	0,6068	0,6120	34	0,6294	0,6291	46	0,1072	0,1076
11	0,5089	0,5096	23	0,3000	0,2984	35	0,8213	0,8216	47	0,7865	0,7864
12	0,0885	0,0867	24	0,0407	0,0377	36	0,1313	0,1526	48	0,9268	0,9271

Заключение

Полученные выражения для тестовых статистик тестов серий и аппроксимированной энтропии и результаты экспериментальных исследований по тестированию последовательностей различных длин показали, что тесты являются эквивалентными при рассмотрении серий одинаковых длин и при длинах исследуемых последовательностей $n\to\infty$. На основании того, что выявлена высокая степень зависимости двух тестов, целесообразно исключить один из них из системы тестирования.

COMPARISON OF SERIAL AND APPROXIMATE ENTROPY TEST STATISTICS

N.G. KIYEVETS, A.I. KORZUN

Abstract

It is shown that mathematical expressions for serial and approximate test statistics are equivalent on conditions that concerned overlapping series have the same lengths and observable sequence length $n \to \infty$. The test result coincidence is experimentally confirmed.

Список литературы

1. Statistical testing of random number generators. [Электронный ресурс]. – Режим доступа: http://infosec.pku.edu.cn/~tly/oldversion/nist-nissc-1999/papers/p24.pdf. – Дата доступа: 13.11.2012.
2. *Turan M.S., Doğanaksoy A., Boztaş S.*// Proceedings of the 5th International Conference «Sequences and Their Applications – SETA 2008 ». September, 14–18, 2008. P. 18–29.
3. A statistical test suite for random and pseudorandom number generators for cryptographic applications. [Электронный ресурс]. – Режим доступа: http://csrc.nist.gov/publications/nistpubs/800-22-rev1a/SP800-22rev1a.pdf. – Дата доступа: 13.11.2012.
4. *Rukhin A.* // J.of Applied Probability. 2000. Vol. 37. P. 88–100.
5. *Киевец Н.Г., Корзун А.И.* // Электроника инфо. 2013. №6 (96). С. 158–160.
6. *Бондаренко В.Ф., Дубовец В.Д.* MatLab. Основы работы и программирования, компьютерная математика. Минск, 2010.