ТЕОРЕТИЧЕСКОЕ ИССЛЕДОВАНИЕ СВОЙСТВ СТАТИСТИЧЕСКОГО ТЕСТА “СТОПКА КНИГ”*

А. И. Пестунов
Институт вычислительных технологий СО РАН, Новосибирск, Россия
e-mail: an24@ngs.ru

The theoretical investigation of the “Bookstack” test is carrying out in this paper. It is shown that for the wide class of alternative hypothesizes this test can check samples when their size is $O(\sqrt{S})$, where S is the length of the alphabet which produces the sample. The sample size becomes the crucial factor when S is rather large. Many other tests, for example chi-square test, are not applicable in such cases because they require samples of the $O(S)$ size.

Введение

В современной информатике широко используются случайные числа: они применяются в криптографии, методах Монте-Карло, численном моделировании и т. д. Как известно, не существует идеально случайных чисел, поэтому на практике используются псевдослучайные. Это числа, порожденные с помощью некоторого датчика. Однако не все последовательности псевдослучайных чисел обладают удовлетворительными статистическими свойствами. Исследовать качество таких последовательностей призваны различные статистические критерии (тесты). Наиболее распространенной задачей критериев является проверка выборки на предмет равномерности распределения, т. е. порождаются ли буквы алфавита равновероятно и независимо или нет.

В [1] предложен новый статистический тест “Стопка книг”, эффективность которого исследовалась экспериментально. В частности, с его помощью проверялись датчики, приведенные в [2]. Они были исследованы автором с помощью мощного спектрального теста [3] и прошли его успешно, но тест “Стопка книг” выявил отклонения от случайности в последовательностях чисел, порожденных этими датчиками. Кроме того, с помощью этого теста впервые удалось найти существенные статистические недостатки у блокового шифра MARS с уменьшенным числом раундов [4], а также генераторов псевдослучайных чисел RANDU [5] и RC4 [6].

В настоящей работе приводится теоретическое исследование свойств теста “Стопка книг” и доказывается, что для достаточно широкого класса альтернативных гипотез он позволяет проверять выборки уже при длине порядка $\sqrt{S}$, где S — количество слов в алфавите, из которого сделана выборка. Размер выборки становится решающим фактором, когда S очень велико, например, 2^{32} или 2^{64}. Именно с такими алфавитами приходится

*Работа выполнена при финансовой поддержке Президентской программы “Ведущие научные школы РФ” (грант № НШ-2314.203.1) и Фонда содействия отечественной науке “Лучшие аспиранты РАН”.

сталкиваться в современной криптографии и защите информации. Многие тесты, в частности критерий хи-квадрат, просто неприменимы в подобных случаях, так как размер тестируемой с их помощью выборки должен быть порядка S.

1. Проверка выборки с помощью теста "Стопка книг"

Тест "Стопка книг" является критерием согласия, поэтому вначале приведем описание такого типа критериев. Пусть имеется выборка $X = (x_1, x_2, \ldots, x_N)$ из алфавита $A = \{a_1, a_2, \ldots, a_S\}$. Рассмотрим гипотезу H_0, которая заключается в том, что элементы выборки независимы и

$$\mathbf{P}(x_n = a_i) = p^0 = 1/S, \quad n = 1, \ldots, N, \quad i = 1, \ldots, S.$$

Другими словами, элементы выборки имеют равномерное распределение, т. е. все буквы алфавита порождаются независимо и с равными вероятностями. Под критерием согласия с гипотезой H_0 понимается некоторая функция от выборки $\pi(X)$, такая, что

$$\pi(X) = \begin{cases} H_0, & \text{т. е. принимаем} H_0; \\ \neg H_0, & \text{т. е. отвергаем} H_0. \end{cases}$$

Критерии характеризуются прежде всего ошибками первого и второго рода. Ошибка первого рода — это вероятность отвергнуть гипотезу H_0, если она верна. Ошибка второго рода — это вероятность принять гипотезу H_0, если она неверна. Величина $(1 - \alpha)$ называется уровнем значимости критерия, $(1 - \beta)$ — мощностью.

Теперь опишем тест "Стопка книг" с теми параметрами, которые будут использоваться в доказательстве. Перед тестированием выборки в алфавите A фиксируется произвольный порядок, который меняется после анализа каждого выборочного элемента x_n следующим образом: буква x_n получает номер 1; номера тех букв, которые были меньше номера этой буквы, увеличиваются на 1; у остальных букв номера не меняются. Формально эту процедуру можно описать так: пусть $\omega^n(a)$ — это номер буквы $a \in A$ после анализа элементов $x_1, x_2, \ldots, x_{n-1}$, тогда

$$\omega^{n+1}(a) = \begin{cases} 1, & \text{если} \quad x_n = a; \\ \omega^n(a) + 1, & \text{если} \quad \omega^n(a) < \omega^n(x_n); \\ \omega^n(a), & \text{если} \quad \omega^n(a) > \omega^n(x_n). \end{cases}$$

Такая конструкция похожа на стопку книг, если считать, что номер книги совпадает с ее положением в стопке. Книга извлекается из стопки, после чтения кладется наверх, и ее номер становится первым. Книги, которые первоначально были над ней, двигаются вниз, а остальные остаются на месте.

В отличие от многих других тестов, например критерия хи-квадрат, здесь подсчитывается не частота встречаемости букв в выборке, а частота встречаемости номеров букв при описанном упорядочивании. Перед тестированием множество всех номеров $\{1, \ldots, S\}$ разбивается на две непересекающиеся части: $A_1 = \{1, 2, \ldots, [\sqrt{S}]\}$ и $A_2 = \{[\sqrt{S}]+1, \ldots, S\}$. Затем по выборке $(x_1, x_2, \ldots, x_N)$ подсчитывается ν_N — количество номеров $\omega^n(x_n)$, принадлежащих подмножеству A_1, т. е. количество попаданий букв в "верхнюю часть" "стопки книг". Число $(N - \nu_N)$, очевидно, равно количеству попаданий в "нижнюю часть". Далее вычисляется статистика

$$x^2 = \frac{(\nu_N - NP_1)^2}{NP_1} + \frac{((N - \nu_N) - N(1 - P_1)^2)}{N(1 - P_1)}, \; P_1 = |A_1|/S,$$

и если x^2 меньше критического уровня $\chi^2_{1,1-\alpha}$, то гипотеза H_0 принимается, иначе — отвергается. Величина $\chi^2_{1,1-\alpha}$– квантиль распределения хи-квадрат уровня значимости $(1-\alpha)$ с одной степенью свободы. Таким образом, тест "Стопка книг" будет выглядеть так:

$$\pi_{bs}(X) = \begin{cases} H_0, & \text{если } x^2 < \chi_{1,1-\alpha}; \\ \neg H_0, & \text{если иначе.} \end{cases}$$

Если H_0 верна, то $\mathbf{P}(\omega^n(x_n) \in A_1) = P_1$, а $\mathbf{P}(\omega^n(x_n) \in A_2) = 1 - P_1$, поэтому x^2 с ростом N приближается к распределению хи-квадрат с одной степенью свободы. Отсюда следует, что при достаточном объеме выборки (должно быть $NP_1 > 8$ [7]) уровень значимости критерия π_{bs} составляет $(1-\alpha)$.

2. Описание критерия хи-квадрат

Критерий хи-квадрат является критерием согласия с гипотезой H_0. Для проверки выборки X нужно определить величины z_i, означающие, сколько раз встретилась буква a_i в выборке. После этого вычисляется статистика

$$\chi^2 = \sum_{i=1}^{S} \frac{(z_i - Np^0)^2}{Np^0},$$

и критерий хи-квадрат будет выглядеть следующим образом:

$$\pi_{\chi^2}(X) = \begin{cases} H_0, & \text{если } \chi^2 < \chi_{S-1,1-\alpha}; \\ \neg H_0, & \text{если иначе.} \end{cases}$$

Здесь $\chi_{S-1,1-\alpha}$ — это квантиль распределения хи-квадрат уровня значимости $(1-\alpha)$ с $(S-1)$ степенями свободы. Критерий основан на следующем свойстве статистики χ^2 — с ростом N она сходится к распределению хи-квадрат с $(S-1)$ степенями свободы. Однако есть одно существенное требование — для достижения заданного уровня значимости $(1-\alpha)$ необходим достаточно большой объем выборки. Точнее, должно выполняться соотношение $Np^0 > 8$ [7]. Другие авторы рекомендуют $Np^0 > 5$ или $Np^0 > 10$. В любом случае это означает, что применять критерий можно, если объем выборки пропорционален длине алфавита, т. е. $N = O(S)$. При работе с современными алгоритмами, например блочными шифрами, размер алфавита может быть более 2^{32}, и в этих условиях использование критерия хи-квадрат становится практически невозможным из-за ограниченности времени и памяти. Нетрудно вычислить, что одна выборка, состоящая из 2^{32} 4-битных слов, занимает порядка 20 Гбайт.

3. Теоретический анализ теста "Стопка книг"

В этой части будет показано, что для достаточно широкого класса альтернативных гипотез тест "Стопка книг" позволяет проверять выборки на длине, пропорциональной $\sqrt{S}$. Точнее, для того чтобы обеспечить заданные уровень значимости и мощность, необходима выборка размера $N = O(\sqrt{S})$.

Рассмотрим некоторую перестановку индексов $\sigma(t), t \in \{1, \ldots, S\}$, и соответствующую ей простую гипотезу $H^{\gamma,\delta}_{\sigma(t)}$ с параметрами γ и δ. Она заключается в том, что элементы выборки X независимы и

$$p_i = \mathbf{P}(x_n = a_{\sigma(t)}) = \begin{cases} 1/S(1+\delta), & \text{если} \quad 1 \le i \le \gamma; \\ 1/S(1-\delta), & \text{если} \quad \gamma S + 1 \le i \le 2\gamma; \\ 1/S, & \text{если} \quad 2\gamma S + 1 \le i \le . \end{cases}$$

Гипотеза говорит о том, что некоторые буквы выпадают немного чаще, а другие — немного реже. Подобный вариант альтернативной гипотезы использовался, например, в [8]. Теперь определим сложную гипотезу $\mathcal{H}^{\gamma,\delta}$ как множество $\{H^{\gamma,\delta}_{\sigma(t)}\}$ со всевозможными перестановками $\sigma(t)$. Вместо гипотезы $\neg H_0$ возьмем ее сужение $\mathcal{H}^{\gamma,\delta}$, однако она достаточно обширна: к такому виду можно привести любую гипотезу, говорящую о независимости, но неравновероятности появления букв. Критерий π_{bs} преобразуется к виду

$$\pi^1_{bs}(X) = \begin{cases} H_0, & \text{если } x^2 < \chi_{1,1-\alpha}; \\ \mathcal{H}^{\gamma,\delta}, & \text{иначе.} \end{cases}$$

Теорема. *Для любых α и β из интервала $(0,1)$ существует константа $C > 0$ такая, что при объеме выборки $N = C[\sqrt{S}]$ ошибки первого и второго рода критерия $\pi^1_{bs}(X)$ асимптотически при $S \to \infty$ не превосходят α и β соответственно.*

Введем обозначения:

$$\tilde{B}_n = \{x_{n-[\sqrt{S}]}, \ldots, x_{n-1}\}, \quad \tilde{\xi}_n = \begin{cases} 1, & \text{если} \quad x_n \in \tilde{B}_n, \\ 0, & \text{если} \quad x_n \notin \tilde{B}_n, \end{cases} \qquad n \in \{[\sqrt{S}]+1, \ldots, N\},$$

$$\tilde{\nu}_N = \sum_{n=[\sqrt{S}]+1}^{N} \tilde{\xi}_n.$$

Смысл этих величин состоит в следующем: $\tilde{B}_n$ — множество букв, встретившихся среди последних $[\sqrt{S}]$ элементов выборки; $\tilde{\xi}_n$ — индикатор попадания очередного выборочного значения x_n в $\tilde{B}_n$; $\tilde{\nu}_N$ — количество таких попаданий после обработки всей выборки.

Для доказательства теоремы нам понадобятся три леммы.

Лемма 1. *Величины $\tilde{\nu}_N$ и ν_N связаны отношением $\tilde{\nu}_N \le \nu_N$.*

Лемма 2. *Если $S \to \infty, C' > 1$, $N = C'[\sqrt{S}]$ и верна гипотеза $\mathcal{H}^{\gamma,\delta}$, то*

$$\mathbf{E}\tilde{\nu}_N = (C'-1)(1+2\gamma\delta^2) + o(1).$$

Лемма 3. *Если выполнены условия леммы 1, то $\mathbf{D}\tilde{\nu}_N \le 10(C'-1)$.*

Доказательство. Положим

$$C = \left(\frac{\sqrt{\chi^2_{1,1-\alpha}} + \sqrt{10/\beta}}{2\gamma\delta^2} \right)^2 + 1$$

и покажем, что это и будет искомая константа. Прямым вычислением нетрудно установить, что $C \ge 11$, поэтому для $N = C[\sqrt{S}]$ условие $NP_1 > 8$ выполнено. Отсюда заключаем, что если верна гипотеза H_0, то величина x^2 имеет распределение хи-квадрат с одной степенью свободы и ошибка первого рода критерия π^1_{bs} равна α.

Теперь достаточно показать, что если верна гипотеза $\mathcal{H}^{\gamma,\delta}$, то

$$\mathbf{P}(x^2 < \chi^2_{1,1-\alpha}) < \beta + o(1). \tag{1}$$

Это будет означать, что при таком объеме выборки ошибка критерия π^1_{bs} второго рода асимптотически при $S \to \infty$ не превосходит β.

Обозначим $\nu^{cr}_{N,1-\alpha} = C + \sqrt{C\chi^2_{1,1-\alpha}}$. Величину x^2 можно преобразовать к виду $x^2 = (\nu_N - NP_1)^2/(NP_1(1-P_1))$. Учитывая, что $NP_1 = C[\sqrt{S}]^2/S$, получим

$$\mathbf{P}(x^2 < \chi^2_{1,1-\alpha}) \le \mathbf{P}(\nu_N < \nu^{cr}_{N,1-\alpha}).$$

Воспользуемся леммой 1 и продолжим оценку

$$\mathbf{P}(x^2 < \chi^2_{1,1-\alpha}) \le \mathbf{P}(\tilde{\nu}_N < \nu^{cr}_{N,1-\alpha}). \tag{2}$$

Обозначим $\Delta = \mathbf{E}\tilde{\nu}_N - \nu^{cr}_{N,1-\alpha}$. Используя очевидные преобразования и затем неравенство Чебышева, можно записать

$$\mathbf{P}(\tilde{\nu}_N < \nu^{cr}_{N,1-\alpha}) \le \mathbf{P}(|\mathbf{E}\tilde{\nu}_N - \tilde{\nu}_N| > \Delta) \le \frac{\mathbf{D}\tilde{\nu}_N}{\Delta^2}.$$

Применяя этот результат к (2), имеем

$$\mathbf{P}(x^2 < \chi^2_{1,1-\alpha}) \le \frac{\mathbf{D}\tilde{\nu}_N}{\Delta^2}. \tag{3}$$

Воспользуемся леммой 2, подставив значение C вместо C', и вычислим

$$\Delta = \frac{1}{2\gamma\delta^2}\left(\sqrt{\frac{10\chi^2_{1,1-\alpha}}{\beta}} + \frac{10}{\beta}\right) + o(1).$$

Непосредственной подстановкий вместо Δ и C их значений легко убедиться в том, что $10(C-1) = \beta\Delta^2 + o(1)$, поэтому из леммы 3 получим (подставив C вместо C')

$$\frac{\mathbf{D}\tilde{\nu}_N}{\Delta^2} \le \beta + o(1).$$

Применив этот результат к (3), получим (1). □

4. Доказательства лемм 1–3

Доказательство леммы 1. Обозначим через ξ_n индикатор попадания номера буквы x_n в “верхнюю часть” “стопки книг”, т. е.

$$\xi_n = \begin{cases} 1, & \text{если} \quad \omega^n(x_n) \in A_1; \\ 0, & \text{если} \quad \omega^n(x_n) \in A_2. \end{cases}$$

Количество таких попаданий ν_N равно, очевидно, сумме индикаторов $\sum\limits_{n=1}^{N} \xi_n$. Теперь пусть B_n — это множество тех букв, номера которых принадлежат A_1 после обработки

$(x_1, \ldots, x_{n-1})$, т. е. B_n — это состояние "верхней части" "стопки книг" после обработки $(n-1)$ элемента выборки. Тогда ξ_n можно представить так:

$$\xi_n = \begin{cases} 1, & \text{если} \quad x_n \in B_n; \\ 0, & \text{если} \quad x_n \notin B_n. \end{cases}$$

Если $n > [\sqrt{S}]$, то множество B_n состоит из всех элементов, встретившихся среди $(x_{n-[\sqrt{S}]}, \ldots, x_{n-1})$, т. е. элементов множества $\tilde{B}_n$, и (так как среди них возможны повторения) из некоторых элементов, встретившихся ранее. Из этих рассуждений можно заключить, что $\tilde{B}_n \subseteq B_n$ и $\tilde{\xi}_n \leq \xi_n$. Как это было показано ранее, величина $\tilde{B}_n$ аналогично B_n представляется через сумму $\tilde{\xi}_n$, поэтому $\tilde{\nu}_N \leq \nu_N$. □

Для краткости обозначим $K = [\sqrt{S}]$.

Доказательство леммы 2. Истинность $\mathcal{H}^{\gamma,\delta}$ означает истинность одной из $H^{\gamma,\delta}_{\sigma(t)}$, поэтому необходимо рассматривать всевозможные перестановки $\sigma(t)$. Заметим, однако, что в дальнейших выкладках $\sigma(t)$ влияет только на порядок слагаемых в сумме, что, конечно, не отражается на результате. Рассмотрим $n \in \{K+1, \ldots, N\}$, так как именно эти индексы фигурируют в определении $\tilde{\nu}_N$. Применим формулу полной вероятности к определению $\tilde{B}_n$:

$$\mathbf{P}(x_n \in \tilde{B}_n) = \sum_{i=1}^{S} p_i \mathbf{P}(x_n \in \tilde{B}_n | x_n = a_i) = \sum_{i=1}^{S} p_i \mathbf{P}\left[\bigcup_{j=1}^{K} (x_{n-j} = a_i)\right] =$$

$$= \sum_{i=1}^{S} p_i \left[1 - \mathbf{P}\left(\bigcap_{j=1}^{K} (x_{n-j} = a_i)\right)\right] = \sum_{i=1}^{S} p_i (1 - (1-p_i)^K) =$$

$$= \sum_{i=1}^{S} p_i \left[1 - \left(1 - Kp_i + \frac{K(K-1)}{2} p_i^2 - \frac{K(K-1)(K-2)}{6} p_i^3 + \ldots\right)\right].$$

Из определения гипотезы $H^{\gamma,\delta}_{\sigma(t)}$ следует, что $p_i = o(1/S)$, значит, предыдущая формула дает

$$\mathbf{P}(x_n \in \tilde{B}_n) = K \sum_{i=1}^{S} p_i^2 + o\left(\frac{1}{\sqrt{S}}\right).$$

Подставляем сюда значения p_i:

$$\mathbf{P}(x_n \in \tilde{B}_n) = \frac{1}{K}(1 + 2\gamma\delta^2) + o\left(\frac{1}{\sqrt{S}}\right).$$

Теперь осталось применить эту формулу к определениям $\tilde{\nu}_N$ и $\tilde{\xi}_n$, воспользоваться тем, что математическое ожидание суммы равно сумме математических ожиданий, и подставить $N = C'K$:

$$\mathbf{E}\tilde{\nu}_N = \sum_{n=K+1}^{N} \mathbf{E}\tilde{\xi}_n = \sum_{n=K+1}^{N} \mathbf{P}(x_n \in \tilde{B}_n) = (C'-1)(1+2\gamma\delta^2) + o(1).$$ □

Доказательство леммы 3. По формуле дисперсии суммы случайных величин

$$\mathbf{D}\tilde{\nu}_N = (N-K)\mathbf{D}\tilde{\xi}_{K+1} + 2\sum_{m=K}^{N-1} \sum_{n=m+1}^{N} \mathbf{Cov}(\tilde{\xi}_m, \tilde{\xi}_n).$$

При $|m-n| > K$ события $\{x_n \in \tilde{B}_n\}$ и $\{x_m \in \tilde{B}_m\}$ независимы, так как порождены независимыми случайными величинами, поэтому $\mathbf{Cov}(\tilde{\xi}_n, \tilde{\xi}_m) = 0$ и

$$\mathbf{D}\tilde{\nu}_N \le (N-K)\mathbf{D}\tilde{\xi}_{K+1} + 2\sum_{m=K}^{N-1}\sum_{n=m+1}^{\min(N,m+K+1)} |\mathbf{Cov}(\tilde{\xi}_m, \tilde{\xi}_n)|. \tag{4}$$

Теперь нужно оценить ковариацию и дисперсию из правой части этой формулы. По определению ковариации получаем

$$\mathbf{Cov}(\tilde{\xi}_m, \tilde{\xi}_n) = \mathbf{P}(\tilde{\xi}_m = 1, \tilde{\xi}_n = 1) - \mathbf{P}(\tilde{\xi}_m = 1)\mathbf{P}(\tilde{\xi}_n = 1),$$

значит,

$$|\mathbf{Cov}(\tilde{\xi}_m, \tilde{\xi}_n)| \le \max\{\mathbf{P}(\tilde{\xi}_m = 1, \tilde{\xi}_n = 1), \mathbf{P}(\tilde{\xi}_m = 1)\mathbf{P}(\tilde{\xi}_n = 1)\}. \tag{5}$$

Учитывая, что

$$\mathbf{P}(\tilde{\xi}_m = 1, \tilde{\xi}_n = 1) = \mathbf{P}(x_n \in \tilde{B}_n, x_m \in \tilde{B}_m) = \mathbf{P}\left(\bigcup_{i=n-K-1}^{n-1}\{x_n = x_i\} \cap \bigcup_{j=m-K-1}^{m-1}\{x_m = x_j\}\right) =$$

$$= \mathbf{P}\left(\bigcup_{i=n-K-1}^{n-1}\bigcup_{j=m-K-1}^{m-1}\{x_m = x_j\} \cap \{x_n = x_i\}\right) \le \sum_{i=n-K-1}^{n-1}\sum_{j=m-K-1}^{m-1}\mathbf{P}(x_m = x_j, x_n = x_i),$$

получаем

$$\mathbf{P}(\tilde{\xi}_m = 1, \tilde{\xi}_n = 1) \le K^2 \max_{i,j}\{\mathbf{P}(x_m = x_j, x_n = x_i)\}. \tag{6}$$

Поскольку $m < n$ (см. неравенство (4)) и $j < m$ (см. вывод формулы (6)), то $j < n$ и для оценки $\mathbf{P}(x_m = x_j, x_n = x_i)$ достаточно рассмотреть три случая.

1. $i \ne m$ и $i \ne j$, тогда события $\{x_n = x_i\}$ и $\{x_m = x_j\}$ независимы. Они порождены независимыми случайными величинами, поэтому $\mathbf{P}(x_n = x_i, x_m = x_j) = \mathbf{P}(x_n = x_i)\mathbf{P}(x_m = x_j)$ и из определения гипотезы $H^{\gamma,\delta}_{\sigma(t)}$ следует, что

$$\mathbf{P}(x_n = x_i, x_m = x_j) \le \frac{(1+\delta)^2}{S^2}.$$

2. $i = m$, тогда $\mathbf{P}(x_n = x_i, x_m = x_j) = \mathbf{P}(x_n = x_m, x_j = x_m)$. Применяем формулу полной вероятности, рассматривая возможные значения x_m:

$$\mathbf{P}(x_n = x_m, x_j = x_m) = \sum_{k=1}^{S}\mathbf{P}(x_m = a_k)\mathbf{P}(x_n = a_k, x_j = a_k).$$

Величины x_n и x_j независимы, поэтому $\mathbf{P}(x_n = a_k, x_j = a_k) = \mathbf{P}(x_n = a_k)\mathbf{P}(x_j = a_k)$ и

$$\mathbf{P}(x_n = x_m, x_j = x_m) = \sum_{k=1}^{S} p_k^3 = \frac{1 + 6\gamma\delta^2}{S^2}.$$

3. Случай $i = j$ аналогичен предыдущему, поэтому формула (6) и эти случаи приводят к тому, что

$$\mathbf{P}(\tilde{\xi}_n = 1, \tilde{\xi}_m = 1) = K^2\frac{(1+\delta)^2}{S^2}. \tag{7}$$

Поскольку $|\tilde{B}_n| \le K$ и $\mathbf{P}(x_n = a_i) \le (1+\delta)/S$ (следует из определения гипотезы $H^{\gamma,\delta}_{\sigma(t)}$), то

$$\mathbf{P}(x_n \in \tilde{B}_n) \le \frac{K}{S}(1+\delta), \tag{8}$$

поэтому

$$\mathbf{P}(\tilde{\xi}_n = 1)\mathbf{P}(\tilde{\xi}_m = 1) \le K^2 \frac{(1+\delta)^2}{S^2},$$

значит, из (5), (7) и этой формулы следует, что

$$|\mathbf{Cov}(\tilde{\xi}_m, \tilde{\xi}_n)| \le K^2 \frac{(1+\delta)^2}{S^2}. \tag{9}$$

Таким образом, мы оценили ковариацию из формулы (4), и нам осталось оценить $\mathbf{D}\tilde{\xi}_{K+1}$. Применим определение дисперсии случайной величины и получим $\mathbf{D}\tilde{\xi}_{K+1} = \mathbf{P}(\tilde{\xi}_{K+1} = 1)(1 - \mathbf{P}(\tilde{\xi}_{K+1} = 1))$. Используя (8), продолжаем $\mathbf{D}\tilde{\xi}_{K+1} \le K(1+\delta)/S$. Теперь, с помощью (9) оценку для дисперсии (4) можно записать как

$$\mathbf{D}\tilde{\nu}_N \le \frac{(N-K)K}{S}(1+\delta) + 2\frac{(N-K)K^3}{S^2}(1+\delta)^2.$$

Так как $N = C'K, K^2 \le S$ и $\delta < 1$, то

$$\mathsf{D}\tilde{\nu}_N \le 10(C'-1). \qquad \square$$

Список литературы

[1] Рябко Б.Я., Пестунов А.И. "Стопка книг" как новый статистический тест для случайных чисел // Пробл. передачи информации. 2004. Т. 40, вып. 1. С. 73–78.

[2] L'Ecuyer P. Tables of linear congruential generators of different sizes and good lattice structure // Math. of Comp. 1999. Vol. 68. P. 249–260.

[3] Кнут Д.Э. Искусство программирования. Т. 2: Получисленные алгоритмы. М.: Изд. дом "Вильямс", 2000.

[4] Pestunov A. Statistical Analysis of the MARS Block Cipher // Cryptology ePrint Archive. Report 2006/217. 2006. `http://eprint.iacr.org/2006/217.`

[5] Ryabko B., Monarev V. Using information theory approach for randomness testing // J. of Statistical Planning and Reference. 2005. Vol. 133, N 1. P. 95–110.

[6] Doroshenko S., Ryabko B. The experimental distinguishing attack on RC4 // Cryptology ePrint Archive. Report 2006/070. 2006. `http://eprint.iacr.org/2006/070.`

[7] Боровков А.А. Математическая статистика. М.: Наука. Гл. ред. физ.-мат. лит., 1984.

[8] Ryabko B., Stognienko V., Shokin Yu. A new test for randomness and its application to some cryptographic problems // J. of Statistical Planning and Reference. 2004. Vol. 123, N 2. P. 365–376.

Поступила в редакцию 18 сентября 2006 г.