$$\Delta A = l_i \cdot R_i = l_i \cdot R/p_i$$

має всі лишки, окрім лишка по основі pi такими, що дорівнюють нулю. Але ΔA = li·Ri > P = R/pk тобто, навіть при li = 1, величина R/pi > R/pk по тій причині, що pk > pi.

Відтак, сума $\tilde{A}$ = A + ΔA > P, тобто викривлене число вийшло за межі робочого діапазону P і попало в діапазон [P, R).

Відомі алгоритми кодування-декодування як раз і використовують цей факт.

III Висновки

Застосування запропонованих узагальнених кодів дозволяє забезпечити виявлення та виправлення викривлень в b-розрядних узагальнених символах в кожному із базових кодових слів.

Застосування таких кодів, на погляд авторів, дозволить розв'язати проблему надійного забезпечення цілісності інформаційних об'єктів в умовах впливу пакетів викривлень значної тривалості.

Література: ***1****. НД ТЗІ 2.5-005-99 "Класифікація автоматизованих систем і стандартні функціональні профілі захищеності оброблюваної інформації від несанкціонованого доступу".* ***2****. Дубровский В. В. CDMA − взгляд глазами профессионала. //mailto:v_ dubrovskii@mail.ru.*

УДК 681.3.06

АНАЛИЗ ОДНОГО МЕТОДА ТЕСТИРОВАНИЯ СЛУЧАЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ОСНОВАННОГО НА КОНТЕКСТНОМ МОДЕЛИРОВАНИИ

Михаил Савчук, Виталий Шарапов

Физико-технический институт Национального технического университета Украины "КПИ"

Анотація: **Розробляється статистичний критерій тестування двійкових послідовностей на випадковість, що базується на правильності завбачення знаків за попередніми бітами. Знаходяться теоретичні та експериментальні розподіли статистик.**

Summary: **In the work the statistical criterion of testing binary sequences on the randomness, based on correctness prediction signs on the previous bits are offered. There are theoretical and experimental statistic distributions.**

Ключевые слова: **Контекстное моделирование, контекстное сжатие, случайные и псевдослучайные последовательности, тестирование случайных последовательностей, статистический критерий.**

I Постановка задачи. Основные определения

В настоящее время разработан ряд статистических тестов для проверки качества случайных и псевдослучайных последовательностей, например, пакет NIST Statistical Test Suite [1]. Однако задача построения новых критериев для оценивания характеристик таких последовательностей остается актуальной для многих применений и исследований, особенно, в области криптографической защиты информации. Новые и модифицированные критерии могут лучше "работать" при других параметрах последовательностей (например, малой длине) или лучше отличить определенные виды альтернативных гипотез. В [2] предложен метод тестирования последовательностей, основанный на контекстном моделировании [3]. В данной работе разрабатываются другие критерии, в основе которых лежит идея контекстного сжатия, проведен теоретический анализ, приведены алгоритмы для построения статистик и критических областей.

Пусть задана последовательность $\varepsilon = \varepsilon_1 \varepsilon_2 \ldots \varepsilon_n$, $\varepsilon_i \in \{0,1\}$, $i = 1, \ldots, n$, которую рассматриваем как реализацию некоторого случайного дискретного процесса. Необходимо проверить гипотезу H0 о том, что последовательность ε является реализацией последовательности независимых равновероятных двоичных знаков, т. е. случайные величины ε_i независимы в совокупности и $\forall i = 1, \ldots, n \; P(\varepsilon_i = 0) = P(\varepsilon_i = 0) = 1/2$.

Назовем контекстом $k_m(i)$ порядка m, $0 \le m < n$, символа ε_i, $i = m+1, \ldots, n,$ число, двоичная запись

которого $\varepsilon_{i-m1}\varepsilon_{i-m+1}\varepsilon_{i-m+2}\ldots\varepsilon_{i-1}$, т. е. есть набор предшествующих символу ε_i m знаков последовательности $\varepsilon\in\{0,1\}^n$. Таким образом, контекст любого символа есть некоторое число от 0 до 2^m-1. Если имеет место гипотеза H0 , то последовательность случайных величин $k_m(i)$, $i=m+1,\ldots,n,$ образуют по i однородную цепь Маркова с начальным распределением $p^{(m+1)}=(p_0^{(m+1)},p_1^{(m+1)},\ldots,p_{2^m-1}^{(m+1)})$ и матрицей переходных вероятностей $P=\left\|p_{ij}\right\|_0^{2^m-1}$, где $\forall j$ $p_j^{(m+1)}=1/2^m$, при $m\geq 2$ $p_{ij}=1/2$, если $j=2i \bmod 2^{m-1}$ или $j=2i \bmod 2^{m-1}+1$ и $p_{ij}=0$ в остальных случаях, а при $m=1$ $p_{ij}=1/2$ $\forall i,j\in\{0,1\}$. Нетрудно проверить, что при условии справедливости гипотезы H0 цепь Маркова $k_m(i)$ стационарна со стационарным распределением $p^{(m+1)}$ [4].

Обозначим $n_j(i)$, $i=m+1,\ldots,n$, – число символов среди $\varepsilon_{m+1},\varepsilon_{m+2},\ldots,\varepsilon_i$, которые имеют контекст j, $j=0,1,...,2^m-1$; $n_j(n)=n_j$, $\sum_{j=0}^{2^m-1}n_j=n-m$. Введем случайную функцию $Y_j(i)=\sum_{l=m+1}^{i}(2\varepsilon_l-1)\delta(j,k_m(l))$, где символ Кронекера $\delta(j,k)=1$, если $j=k$ и $\delta(j,k)=0$ при $j\neq k$. Случайная величина $Y_j(i)$ – разность между числом единиц и нулей, следовавших за контекстами с номером j до i-го символа включительно.

Назовем символ ε_i с контекстом j предсказуемым (правильно предсказуемым), если выполняется одно из двух условий: $\{Y_j(i-1)>0 \quad u\ \varepsilon_i=1\}$ или $\{Y_j(i-1)<0 \quad u \quad \varepsilon_i=0\}$. Т. е. при прогнозировании символа ε_i по преобладанию единиц или нулей, следовавших ранее за контекстом j, прогноз для предсказуемого символа будет правильным. Если имеет место одно из событий $\{Y_j(i-1)>0 \quad u \quad \varepsilon_i=0\}$ или $\{Y_j(i-1)<0 \quad u\ \varepsilon_i=1\}$ прогноз окажется неверным и символ ε_i называется неправильно предсказуемым. При равенстве соответствующих единиц и нулей, т. е. событии $\{Y_j(i-1)=0\}$ символ ε_i “не предсказывается” и считается непредсказуемым. Пусть r_i – случайная величина, равная числу всех предсказуемых символов среди $\varepsilon_{m+1},\varepsilon_{m+2},\ldots,\varepsilon_n$, которые следовали за контекстом j, $j=0,1,\ldots,2^m-1$, $\sum_{j=0}^{2^m-1}r_j=r$ число всех предсказуемых символов.

При тестировании последовательность ε просматривается от 1-го символа до n-го и вычисляются все пары чисел (n_j,r_j), $j=0,1,\ldots,2^m-1$. По вычисленному множеству пар строится статистический критерий с заданным уровнем значимости α проверки гипотезы H0: последовательность ε является последовательностью независимых равновероятных двоичных знаков. Для построения критерия и критических областей найдем распределение числа предсказуемых символов при условии, что имеет место гипотеза H0.

II Распределение вероятностей числа предсказуемых символов

Обозначим число предсказуемых символов среди первых x, $1\leq x\leq n_j$, символов с контекстом j, $j=0,1,\ldots,2^m-1$, последовательности ε через $S_j(x)$, тогда $S_j(n_j)=r_j$ общее число предсказуемых символов, которые имеют контекст j, а $S(n)=\sum_{j=0}^{2^m-1}S_j(n_j)=r$ общее число предсказуемых символов по

всем контекстам. Выберем в случайном процессе $Y_j(i)$ только ненулевые слагаемые, для которых $\delta(j,k_m(l))=1$, и перенумеруем значения аргумента, т. е. рассматриваем случайный процесс $Y_j(i)$ только в моменты времени, в которые соответствующий символ имеет контекст j, и нумеруем эти моменты начиная с 1. Обозначим построенный таким образом процесс через $Y_j(x)$, где x, $1\le x\le n_j$, – число встретившихся до момента i символов с контекстом j. Считаем по определению, что $Y_j(0)=0$. Очевидно, $\forall j\ \ Y_j(x)$ при справедливости гипотезы H0 представляет собой бернуллиевское случайное блуждание [5], в котором движению на единицу вверх будет соответствовать появление единичного символа с контекстом j, а движению на единицу вниз – появление нулевого символа с контекстом j в последовательности ε.

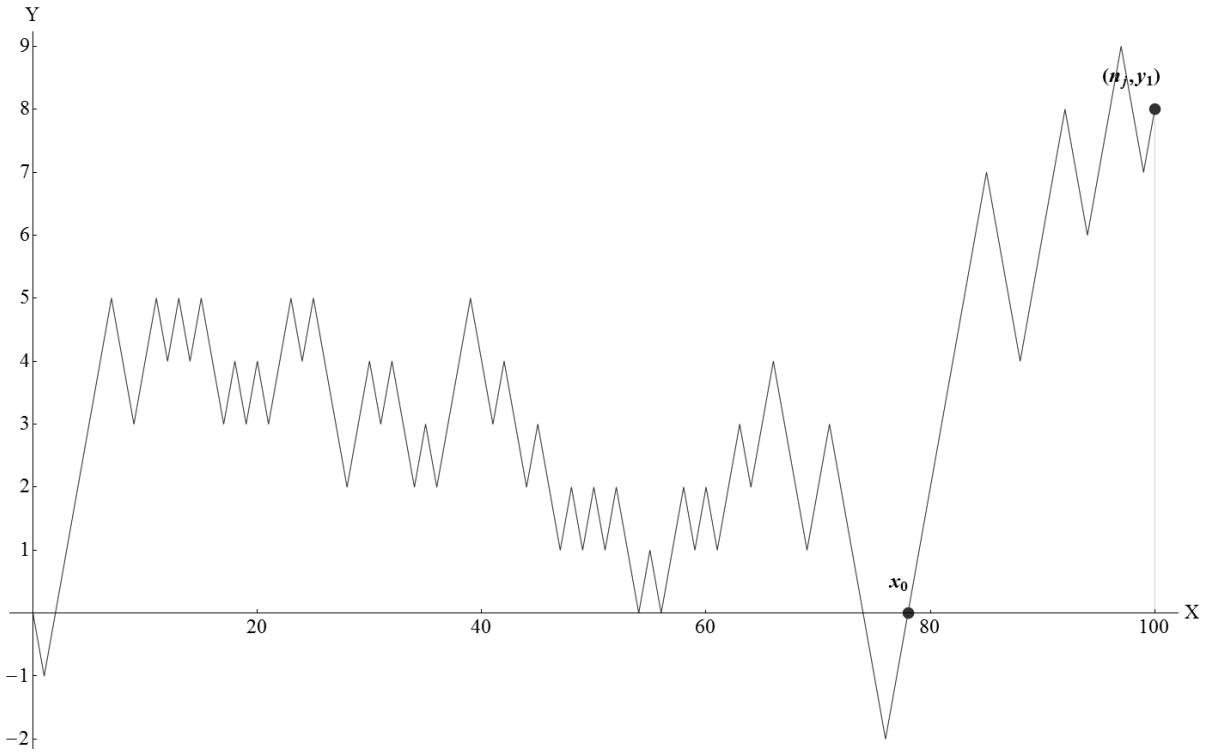

Рисунок 1 – Пример реализации случайного блуждания Бернулли

При этом точка с координатой (x, y) на графике означает, что среди первых x символов с контекстом j разность между встретившимся числом единиц и нулей будет равна y. Найдем распределение случайной величины r_j (которая может принимать значение от 0 до n_j) при условии, что общее число символов n_j с контекстом j фиксировано. Очевидно, при фиксированном n_j вероятностное распределение r_j не зависит от j, $j=0,1,\ldots,2^m-1$.

Для решения данной задачи воспользуемся теорией случайных блужданий, приведенной в [5]. Необходимо определить количество путей, которые ведут из точки (0, 0) в точку с координатами (nj, y1), где $y_1=Y_j(n_j)$. Обозначим $x_0=\max\left\{x: Y_j(x)=0,\quad 1\le x\le n_j\right\}$, т. е. x_0 – момент последнего пересечения блужданием оси x или ее касания. На рис. 1 видно, что удобно рассмотреть полный путь как объединение путей из точки (0, 0)в (x0, 0)и из (x0, 0) в точку (nj, y1).

Будем говорить, что предсказуемым символам с контекстом (nj, y1) соответствуют на графике случайного блуждания (nj, rj) предсказуемые (правильно предсказанные) точки, которые лежат на участках возрастания графика в положительной области оси ординат и участках убывания в отрицательной. Точка, следующая непосредственно за точкой, расположенной на оси абсцисс, является непредсказуемой. В частности, на участке пути из точки (x0, 0) в точку (nj, y1) правильно предсказанных

символов будет ровно $|y_1|-1$. Можно рассматривать только случаи, когда $y_1 > 0$ и учитывать симметрию со случаями $y_1 < 0$, т. е. в результате брать удвоенное количество путей. Для варианта, когда второй участок пути отсутствует ($x_0 = n_j$), будем считать, что после точки x_0 существует один единственный путь, который может быть выбран с вероятностью единица.

Рассмотрим детально пути из точки (0, 0) в (x0, 0). Количество таких путей составляет $\binom{x_0}{\frac{x_0}{2}}$ при общем количестве 2^{x_0}, при этом вероятность одного конкретного пути составляет $2^{-x_0}\binom{x_0}{\frac{x_0}{2}}$, причем x_0 может быть только четным числом. Если бы эти пути ни разу не пересекали ось абсцисс и не касались ее, то количество предсказуемых символов за весь путь составило бы $\frac{x_0}{2}-1$. Каждое пересечение оси уменьшает количество предсказуемых символов на единицу. Количество путей из точки (0, 0) в точку (x0, 0), которые ровно m раз (не считая крайних точек) пересекают ось абсцисс или касаются ее, равно $\frac{m+1}{x_0-m-1}\binom{x_0-m-1}{\frac{x_0}{2}}$, где $m \geq 0$.

От точки (0, 0) до точки (x0, 0) мы уже получили $\frac{x_0}{2}-m-1$ предсказуемых символов. Если общее число предсказуемых символов r_j, то необходимо получить $r_j-\left(\frac{x_0}{2}-m-1\right)$ предсказуемых символов на участке пути из точки $(x_0,0)$ в точку (n_j, y_1). Это означает, что y_1 должно быть равно $r_j-\left(\frac{x_0}{2}-m-1\right)+1$. Теперь осталось подсчитать количество последних путей, при условии что они не пересекают ось абсцисс и не касаются ее. Это можно сделать с помощью теоремы о баллотировке [5]. Общее число таких путей будет $\frac{2y_1-(n_j-x_0)}{n_j-x_0}\binom{n_j-x_0}{y_1}$

Из изложенного выше можно вывести формулу вычисления вероятности $P(n_j, r_j)$ того, что число предсказуемых символов, следующих за n_j контекстами с номером j, равно r_j:

$$P(n_j, r_j) = \sum_{\substack{i=2,\\ i \bmod 2=0}}^{n_j} \sum_{k=1}^{\frac{i}{2}} 2^{-i} \frac{k}{i-k}\binom{i-k}{\frac{i}{2}} p_1(n_j - i, r_j - (\tfrac{i}{2}-k)) + p_1(n_j, r_j), \qquad (1)$$

где
$$p_1(l,r) = \begin{cases} p_2(l, r+1), & \text{если } r > 0; \\ 1, & \text{если } r = 0 \text{ и } (l = 0 \text{ или } l = 1); \\ 0, & \text{иначе}, \end{cases}$$

$$p_2(l,q)=\begin{cases} 0, & \text{если } 2q<l \text{ или } l<q; \\ 1, & l=0 \text{ и } q\le l\le 2q; \\ 2^{-l+1}(\frac{2q}{l}-1)\binom{l}{q}, & q\le l\le 2q; \end{cases}$$

Вышеприведенная формула имеет силу при условии справедливости гипотезы H0.

Функция p1 вычисляет вероятность числа предсказуемых символов на участке пути из точки (x0, 0) в точку (nj, y1), а p2определяет вероятность сделать qшагов вверх на том же участке.

Обозначим функцию распределения количества предсказуемых символов r_j по некоторому контексту j при общем количестве появлений этого контекста n_j через $F_{n_j}(x)=P(r_j<x)$, где x – действительное число. Формула (1) позволяет вычислять функцию распределения $F_{n_j}(x)$ и необходимые квантили для произвольного числа символов с контекстом r_j при построении критических областей, но для очень больших n_j нужно воспользоваться асимптотическими приближениями.

III Алгоритм вычисления числа символов с заданным контекстом и числа предсказуемых символов

Анализ последовательностей, предложенный в пункте I, можно изложить в виде формального алгоритма следующим образом.

Алгоритм 1.

Входные данные.

- Двоичная последовательность ε;
- Порядок контекста *m*.

Выходные данные.

Пары (n_j, r_j), где n_j – количество раз, которое встретился контекст; r_j – количество правильных прогнозов, сделанных по контексту j.

Внутренние данные.

- i – текущий, обозреваемый в данный момент символ строки;
- j – текущий, обозреваемый в данный момент контекст.

[1.Начальная инициализация] $i \leftarrow m+1$; $(n_k, r_k) \leftarrow (0,0)$ для $k=0,\ldots,2^m-1$;

[2.Определяем контекст] $j \leftarrow k_m(i)$;

[3.Корректируем правильные прогнозы] Если $2\varepsilon_i - 1 = \mathrm{sgn}(Y_j(i))$, то $r_j \leftarrow r_j + 1$;

[4.Корректируем счетчик контекста] $n_j \leftarrow n_j + 1$;

[5.Переходим к следующему символу] $i \leftarrow i+1$;

[6.Проверяем условие завершения] Если i меньше или равно длине *n* последовательности ε, то перейти к шагу два.

[7.Алгоритм закончен] Значения пар (n_j, r_j) является результатом работы алгоритма.

IV Построение критерия тестирования

На основании изложенного в пунктах I – III предлагается критерий тестирования, который будет для произвольной двоичной строки ε давать ответ на вопрос: “Гипотеза H_0 о равновероятности и независимости двоичных знаков согласуется с наблюдаемой последовательностью $\varepsilon=\varepsilon_1\varepsilon_2\ldots\varepsilon_n$, $\varepsilon_i\in\{0,1\}$, $i=1,\ldots,n$, с выбранным уровнем значимости α или нет?” Опишем построение статистического критерия.

1. С помощью алгоритма 1 подсчитывается количество всех различных контекстов и количество

предсказуемых символов по каждому контексту, т. е. находятся пары (n_j, r_j), $j = 0,1,...,2^m - 1$. Количество контекстов, которые встретились хотя бы один раз, обозначим через U ($U \le 2^m - 1$).

2. Для каждого встретившегося контекста проверяется гипотеза H_{A0}, которая заключается в том, что количество предсказуемых символов согласуется с распределением $F_{n_j}(x)$. По уровню значимости α строим критическую область и область принятия гипотезы

$$a_{p_1} \le r_j \le a_{p2}, \tag{2}$$

где a_{p_1}, a_{p_2} – квантили распределения $F_{n_j}(x)$ уровня p_1, p_2 такие, что $p_1 + 1 - p_2 = \alpha$, а выражение $a_{p2} - a_{p_1}$ принимает минимальное значение. Подсчитывается число Q – количество контекстов, согласующихся с гипотезой H_{A0}, т. е. удовлетворяющих неравенству (2).

3. Проверяется гипотеза H_{B0}, о том, что величина Q будет иметь биномиальное распределение с вероятностью успеха $1-\alpha$ и числом испытаний U. Гипотеза H_{B0} должна согласоваться также с уровнем значимости α. Проверяется выполнимость неравенства

$$b_{p_1} \le Q \le b_{p2}, \tag{3}$$

где b_{p_1}, b_{p_2} – квантили уровня p_1, p_2 биномиального распределения с вероятностью успеха $1-\alpha$ и числом испытаний U такие, что $p_1 + 1 - p_2 = \alpha$, а выражение $b_{p_2} - b_{p_1}$ принимает минимальное значение. Для построения критической области для биномиального распределения и вычисления границ в неравенстве (3) удобно воспользоваться таблицами, приведенными в [6, 7]

4. Вывод: гипотеза H_0 согласуется с последовательностью $\varepsilon = \varepsilon_1 \varepsilon_2 \ldots \varepsilon_n$, $\varepsilon_i \in \{0,1\}$, с уровнем значимости α, если выполняется неравенство (3). В этом случае наблюдаемая последовательность не противоречит гипотезе. Если неравенство (3) не выполняется, то гипотеза H_0 отвергается как противоречащая статистическим данным.

Замечание. Так как распределение $F_{n_j}(x)$ статистики при даже не слишком больших n_j не сильно отличается от симметричного, то квантили a_{p_1} и a_{p_2} для упрощения построения критерия можно выбрать симметрично относительно математического ожидания.

Предложенный тест можно использовать для последовательностей любой длины, в том числе и для коротких последовательностей длиною порядка 50 бит. При этом рекомендуется, чтобы порядок контекста m и длина последовательности n были связанны следующим соотношением $m \le \lceil \log_2 n \rceil$.

По статистикам (n_j, r_j), $j = 0,1,...,2^m - 1$, можно строить другие варианты критерия, которые могут лучше “работать” при определенных альтернативных гипотезах.

V Таблицы распределений

Значения вероятностей $P(n_j, r_j)$ для некоторых длин последовательностей приводятся в таблице 1.

Таблица 1

$r_j \backslash n_j$	8	16	24	32	40	48
0	0.06250	0.003906	0.0002441	0.00001526	9.537×10^{-7}	5.960×10^{-8}
	0.1562	0.01758	0.001587	0.0001297	0.00001001	7.451×10^{-7}
	0.2031	0.04199	0.005432	0.0005760	0.00005460	4.813×10^{-6}
	0.2109	0.07275	0.01315	0.001787	0.0002064	0.00002144
	0.1875	0.1028	0.02536	0.004359	0.0006082	0.00007403
5	0.1250	0.1259	0.04147	0.008910	0.001490	0.0002114
	0.04688	0.1385	0.05980	0.01588	0.003161	0.0005197

	0.007812	0.1402	0.07801	0.02533	0.005960	0.001131
		0.1328	0.09377	0.03689	0.01019	0.002219
		0.1082	0.1053	0.04975	0.01604	0.003992
10		0.06812	0.1115	0.06280	0.02350	0.006657
		0.03259	0.1124	0.07487	0.03234	0.01039
		0.01147	0.1085	0.08489	0.04212	0.01528
		0.002808	0.09454	0.09206	0.05223	0.02132
		0.0004272	0.06963	0.09595	0.06198	0.02837
15		0.00003052	0.04312	0.09649	0.07069	0.03617
			0.02228	0.09396	0.07775	0.04435
			0.009490	0.08474	0.08276	0.05247
			0.003277	0.06740	0.08547	0.06007
			0.0008948	0.04718	0.08585	0.06671
20			0.0001860	0.02898	0.08405	0.07204
			0.00002766	0.01555	0.07737	0.07578
			2.623×10^{-6}	0.007245	0.06443	0.07781
			1.192×10^{-7}	0.002908	0.04849	0.07810
				0.0009950	0.03292	0.07674
25				0.0002862	0.02012	0.07162
				0.00006790	0.01104	0.06149
				0.00001294	0.005422	0.04853
				1.904×10^{-6}	0.002370	0.03519
				2.030×10^{-7}	0.0009175	0.02341
30				1.397×10^{-8}	0.0003122	0.01426
				4.657×10^{-10}	0.00009258	0.007947
					0.00002365	0.004037
					5.130×10^{-6}	0.001865
					9.261×10^{-7}	0.0007809
35					1.354×10^{-7}	0.0002950
					1.541×10^{-8}	0.0001000
					1.281×10^{-9}	0.00003026
					6.912×10^{-11}	8.107×10^{-6}
					1.819×10^{-12}	1.905×10^{-6}
40						3.881×10^{-7}
						6.753×10^{-8}
						9.834×10^{-9}
						1.166×10^{-9}
						1.082×10^{-10}
45						7.361×10^{-12}
						3.268×10^{-13}
						7.105×10^{-15}

На рис. 2 – 4 приводятся законы распределения вероятностей $P(n_j, r_j)$ для различных чисел появлений контекстов.

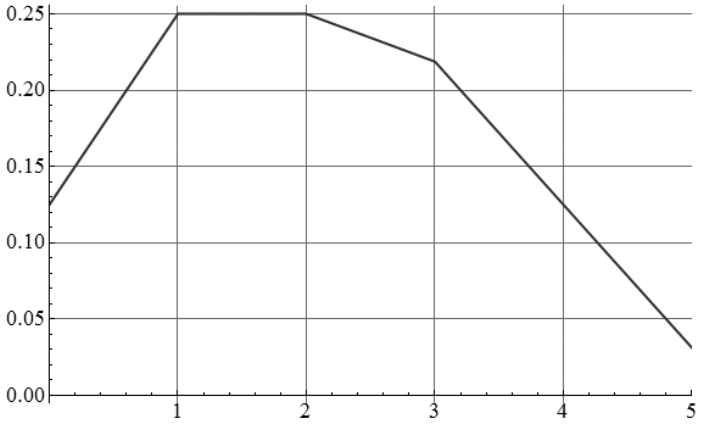

Рисунок 2 – Распределение r_j при $n_j = 6$

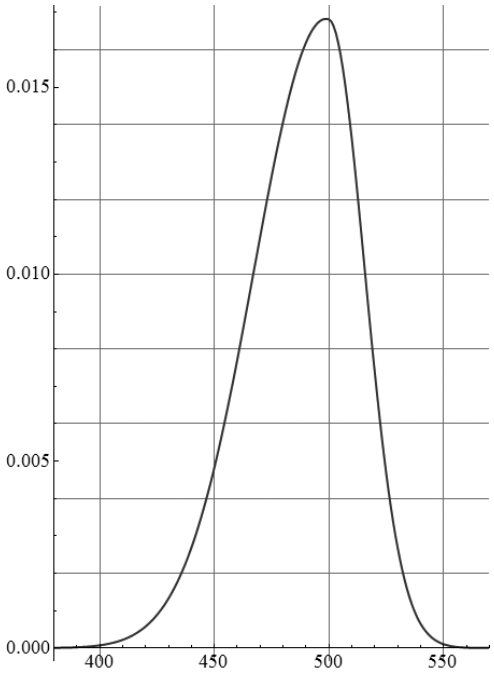

Рисунок 3 – Распределение r_j при $n_j = 100$

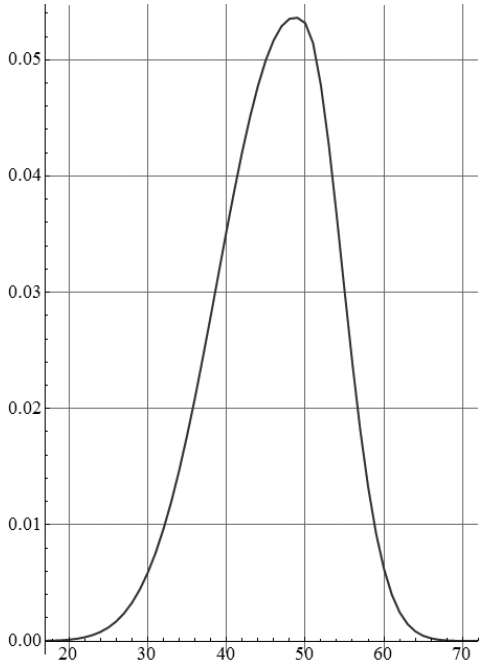

Рисунок 4 – Распределение r_j при $n_j = 1000$

Литература: ***1.*** *A Statistical Test Suite For Random And Pseudorandom Number Generators For Cryptographic Applications. NIST Special Publication 800-22.* ***2.*** *Шарапов В. Тестирование случайных и псевдослучайных последовательностей с использованием контекстного моделирования //Правове, нормативне та метрологічне забезпечення системи захисту інформації в Україні. Науково-технічний збірник. Випуск 2 (15). – К.: НИЦ "Тезис" НТУУ "КПИ", 2007. – С. 86 – 97.* ***3.*** *Ватолин Д., Ратушняк А., Смирнов М., Юкин В. Методы сжатия данных. Устройства архиваторов, сжатие изображений и видео. – М.: ДИАЛОГ-МИФИ, 2003. – 384 с.* ***4.*** *Коваленко И. Н., Гнеденко Б. В. Теория вероятностей. – К.: Выща школа, 1990. – 328 с.* ***5.*** *Феллер В. Введение в теорию вероятностей и ее приложения. В 2-х томах. Т.1: Пер. с англ.– М.: Мир, 1984. – 528 с.* ***6.*** *Шор Я. Б. Статистические методы анализа и контроля качества и надежности. – М.: Сов. Радио, 1962. – 552 с.* ***7.*** *Шор Я. Б., Кузьмин Ф. И. Таблицы для анализа и контроля надежности. – М.: Советское радио, 1968. – 284 с.*