ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 28 МАТЕМАТИКИ Выпуск 1
2021

В. О. Миронкин, **Н. А. Юрасов** (Москва, Национальный исследовательский университет «Высшая школа экономики»). **К вопросу о применимости статистик меры «хи-квадрат» в задаче определения параметров блочных алгоритмов шифрования.**

УДК 519.248

Резюме: Изучаются вопросы, связанные с возможностью использования статистик меры «хи-квадрат» для определения параметров базового блочного алгоритма зашифрования, функционирующего в режиме шифрования ECB. Сформулированы условия, при которых применение статистик указанного типа позволяет определить размер блока соответствующего алгоритма. Предложены рекомендации по выбору некоторых параметров кодировки знаков открытого текста в условиях использования режима шифрования ECB.

Ключевые слова: Блочный алгоритм шифрования, базовый блочный алгоритм зашифрования, режим шифрования, статистика, критерий, гипотеза, распределение «хи-квадрат», режим ECB.

Введение

Хорошо известно, что блочные алгоритмы шифрования должны обеспечивать неотличимость шифртекста, представленного в двоичном виде, от реализации схемы независимых равновероятных испытаний Бернулли [1]. При этом указанное требование, очевидным образом, распространяется на все известные режимы шифрования [2].

В научной периодике по данному направлению имеется спектр работ, подтверждающих выполнение указанного требования как для отечественных, так и для иностранных стандартизированных решений (например, [3, 4]). При этом, как правило, соответствующее обоснование строится на проверке сформированных с использованием блочного алгоритма шифрования двоичных последовательностей (далее — шифртекста) при помощи пакетов статистического тестирования типа NIST STS, DieHard, TestU01, RaBiGeTe и др. Примечательно, что ни один из указанных пакетов не использует статистики меры «хи-квадрат» [5], в ряде случаев позволяющие выявить зависимости знаков шифртекста, обусловленные принципом рассеивания, априори заложенным в каждый базовый блочный алгоритм зашифрования [2].

Настоящая работа посвящена исследованию возможности применения статистик указанного типа при решении задачи определения параметров базового блочного алгоритма зашифрования в режиме шифрования ECB [6] с использованием особенностей кодирования знаков открытого текста (далее — кодировки).

1. О зависимости знаков шифртекста

Напомним, что *рассеиванием* называется распространение влияния каждого отдельного знака произвольного блока открытого текста сразу на несколько знаков соответствующего блока шифртекста, которые позволяют нивелировать статистические свойства исходного открытого текста в процессе шифрования.

Перейдем к формальному описанию данного принципа. Для произвольного блока $(x_1, x_2, \ldots, x_m)$ закодированного открытого текста $\overline{x}$, где $m \geqslant 2$ — размер блока базового алгоритма зашифрования, соответствующий блок $(y_1, y_2, \ldots, y_m) =$

© Редакция журнала «ОПиПМ», 2021 г.

$E_k(x_1, x_2, \ldots, x_m)$ шифртекста $\overline{y}$ может быть представлен в следующем виде:

$$\begin{cases} y_1 = \psi_1\left(x_1, \ldots, x_m\right), \\ \ldots \\ y_m = \psi_m\left(x_1, \ldots, x_m\right), \end{cases} \tag{1}$$

где отображения $\psi_1, \ldots, \psi_m \colon \{0,1\}^m \to \{0,1\}$ (координатные функции отображения $E_k \colon \{0,1\}^m \to \{0,1\}^m$, k — некоторый ключ зашифрования) существенно зависят от переменных $x_1, \ldots, x_m$ и однозначно определяются отображением E_k (или что то же самое, ключом k).

З а м е ч а н и е 1. Здесь и далее ключ зашифрования k полагаем произвольным фиксированным. Поэтому с целью упрощения обозначений зависимость отображений $\psi_1, \ldots, \psi_m$ от величины k по тексту работы указывать не будем.

Из (1) и того факта, что переменные $x_1, \ldots, x_m$ являются существенными для каждого из отображений $\psi_1, \ldots, \psi_m$, следует, что для некоторых распределений на множестве $\{0,1\}^m$ блоков закодированного открытого текста случайные величины $y_u = \psi_u(x_1, \ldots, x_m)$ и $y_v = \psi_v(x_1, \ldots, x_m)$, где $u, v \in \{1, \ldots, m\}$, могут быть статистически зависимыми (далее — зависимыми).

Так, например, если на указанном множестве $\{0,1\}^m$ задано равновероятное распределение, то индуцированное на множестве блоков шифртекста распределение, очевидно, также является равновероятным, а в случае сбалансированности E_k по выходу [2] и соответствующей равновесности [2] отображений $\psi_1, \ldots, \psi_m$, знаки в каждом из блоков шифртекста являются независимыми (как попарно, так и в совокупности). В случае же неравновероятного распределения на $\{0,1\}^m$ соответствующие случайные величины y_u и y_v в общем случае являются зависимыми.

Указанная неравновероятность (даже при условии равновероятном распределении знаков алфавита A исходного открытого текста) может быть обусловлена выполнением следующего соотношения:

$$|A| < 2^t, \tag{2}$$

где t — длина кодового слова, полученного в результате преобразования отдельного знака исходного открытого текста в бинарный вид, $|A|$ — мощность алфавита исходного открытого текста.

Одновременно с этим отметим, что в режиме ECB каждый блок закодированного открытого текста $\overline{x}$ шифруется независимо.

Таким образом, в условиях неравновероятности на множестве блоков закодированного открытого текста $\overline{x}$ или при выполнении (2) (далее — Условия) шифртекст $\overline{y}$ обладает следующими статистическими свойствами:

- в рамках каждого блока знаки шифртекста $\overline{y}$ в общем случае являются зависимыми;
- знаки шифртекста $\overline{y}$, принадлежащие разным блокам, являются независимыми.

Отметим, что указанные свойства могут быть использованы при определении размера блока базового блочного алгоритма зашифрования в режиме шифрования ECB.

Для этого выполним предварительную группировку знаков шифртекста $\overline{y}$, а именно, последовательно для каждого $k = 2, 3, \ldots$ (здесь k — опробуемое значение размера блока базового алгоритма зашифрования) запишем друг под другом отрезки шифртекста $\overline{y}$ длины k (см. рис. 1).

$$y_1, y_2, \ldots, y_k, y_{k+1}, \ldots, y_{2k}, y_{2k+1}, \ldots, y_{3k}, y_{3k+1}, \ldots, y_n$$
$$\downarrow$$
$$\begin{pmatrix} y_1 & \cdots & y_8 & \cdots & y_{k-7} & \cdots & y_k \\ y_{k+1} & \cdots & y_{k+8} & \cdots & y_{2k-7} & \cdots & y_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{(\frac{n}{k}-1)k+1} & \cdots & y_{(\frac{n}{k}-1)k+8} & \cdots & y_{n-7} & \cdots & y_n \end{pmatrix}$$

Рис. 1. Преобразование шифртекста $\overline{y}$

З а м е ч а н и е 2. Не ограничивая общности, будем считать, что $k|n$ (в противном случае отбросим последние $n - [\frac{n}{k}]k$ знаков шифртекста $\overline{y}$).

Далее для каждого $k = 2, 3, \ldots$ проведем статистический анализ сгруппированных данных с использованием статистик меры «хи-квадрат».

2. Статистика $S_k^{(1)}$ проверки независимости

Для каждой строки с номером $i \in \{1, \ldots, \frac{n}{k}\}$ представленной на рис. 1 матрицы положим

$$X_i = \left(y_{(i-1)k+1}, \ldots, y_{(i-1)k+l}\right), \quad Y_i = \left(y_{ik-l+1}, \ldots, y_{ik}\right),$$

где $l \in \{1, \ldots, \left[\frac{k}{2}\right]\}$, и будем рассматривать $X_1, \ldots, X_{\frac{n}{k}}$ как реализации случайной величины X, распределение которой индуцировано совместным распределением первых l знаков в каждом из блоков размера k шифртекста $\overline{y}$, а $Y_1, \ldots, Y_{\frac{n}{k}}$ — как реализации случайной величины Y, распределение которой индуцировано совместным распределением последних l знаков тех же блоков соответственно.

З а м е ч а н и е 3. Для проведения статистического анализа сгруппированного шифртекста $\overline{y}$ было выбрано значение $l = 8$ и рассчитаны значения статистики $S_k^{(1)}$ для $k = 16, 17, \ldots$.

При выполнении Условий зависимость между величинами X, Y становится существенной при $k \in \{16, 32, \ldots, \frac{m}{2}, m\}$, поскольку для соответствующих значений k в матрице, изображенной на рис. 1, первые и последние 8 столбцов в общем случае являются зависимыми (для каждого из указанных значений k по построению каждая из пар восьмиграмм $(X_1, Y_1), \ldots, \left(X_{\frac{n}{k}}, Y_{\frac{n}{k}}\right)$ принадлежит отдельному блоку шифртекста истинного размера m). Таким образом, статистика (см. [5])

$$S_k^{(1)} = \frac{n}{k} \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(\nu_{ij} - k\frac{\nu_{i*} \cdot \nu_{*j}}{n}\right)^2}{\nu_{i*} \cdot \nu_{*j}}$$

с высокой вероятностью имеет локальные максимумы в указанных значениях k. Здесь ν_{ij} — элементы таблицы сопряженности [5] признаков X, Y, ν_{i*} — сумма элементов таблицы в i-й строке, ν_{*j} — сумма элементов таблицы в j-м столбце, а r, s — число строк и столбцов матрицы соответственно. При этом справедливы следующие ограничения на размеры таблицы сопряженности:

$$r \leqslant \min\{|A|, 2^l\}, \quad s \leqslant \min\{|A|, 2^l\}.$$

При проведении статистических исследований был использован массив открытых текстов на русском, английском, французском и немецком языках. Каждый экземпляр открытого текста имел длину 10^6 знаков и не подвергался никакой дополнительной обработке. Для преобразования указанных экземпляров в двоичные последовательности использовалась 8, 16, 32-битная кодировки [7, 8].

Приведем некоторые результаты расчета статистики $S_k^{(1)}$, полученные для используемых экземпляров. Так, в частности, на рис. 2–4 представлена функциональная

зависимость статистики $S_k^{(1)}$ от опробуемых значений k для шифртекста, полученного при зашифровании закодированного романа Теодора Драйзера «Финансист» (далее — текст№ 1) (объем — 10^6 символов, язык — английский) с использованием алгоритма AES (размер блока — 128 бит) в режиме ECB.

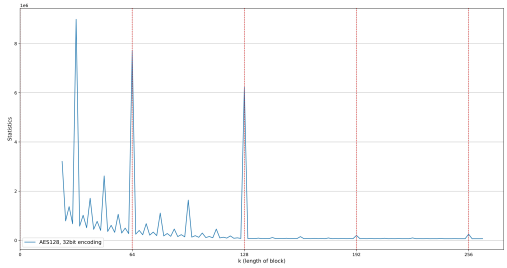

Рис. 2. Значения $S_k^{(1)}$ для AES и текста № 1 при 32-битной кодировке

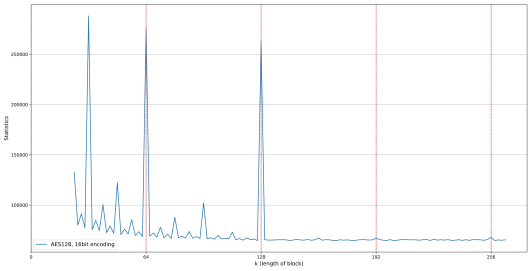

Рис. 3. Значения $S_k^{(1)}$ для AES и текста № 1 при 16-битной кодировке

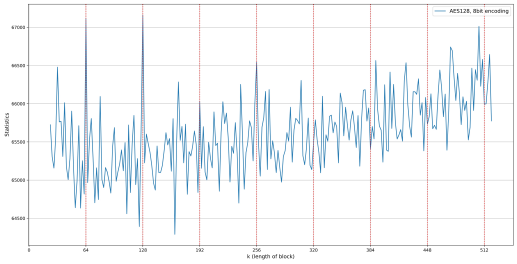

Рис. 4. Значения $S_k^{(1)}$ для AES и текста № 1 при 8-битной кодировке

В свою очередь, аналогичные графики получены для шифртекста, сформированного при зашифровании закодированного романа-эпопеи Михаила Шолохова «Тихий Дон» (далее — текст № 2) того же объема (язык — русский) с использованием алгоритма AES в режиме ECB (рис. 5–7).

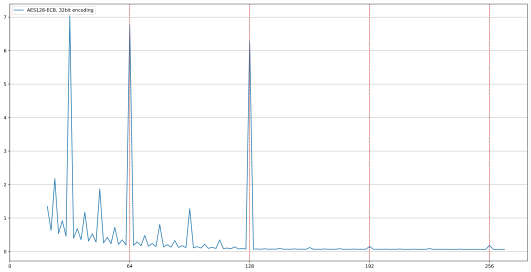

Рис. 5. Значения $S_k^{(1)}$ для AES и текста № 2 при 32-битной кодировке

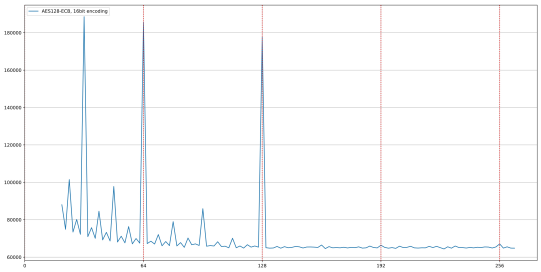

Рис. 6. Значения $S_k^{(1)}$ для AES и текста № 2 при 16-битной кодировке

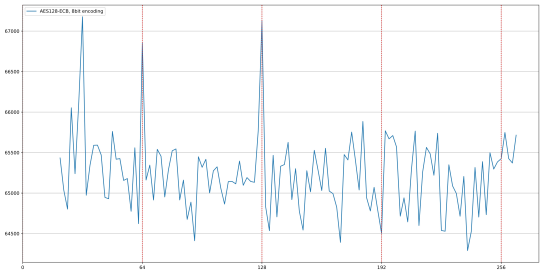

Рис. 7. Значения $S_k^{(1)}$ для AES и текста № 2 при 8-битной кодировке

З а м е ч а н и е 4. Красными пунктирными линиями на всех представленных графиках обозначены опробуемые значения k, кратные истинному значению m размера блока базового алгоритма зашифрования. Обратим внимание, что даже при 8-битной кодировке четко выражены всплески статистики $S_k^{(1)}$ при $k \in \{32, 64, 128\}$ (см. рис. 4 и 7), что полностью соответствует приведенным выше рассуждениям.

Для шифра CAST (за счет меньшего размера блока — 64 бита) на тех же входных данных истинный размер блока m достоверно определяется и для 8-битной кодировки (рис. 8).

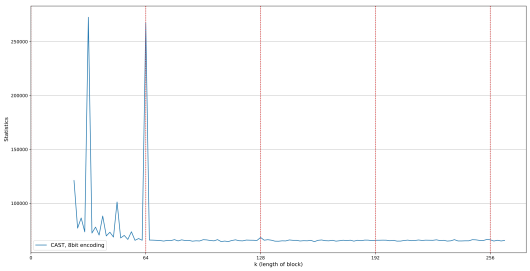

Рис. 8. Значения $S_k^{(1)}$ для CAST и текста № 1 при 8-битной кодировке

3. Статистика $S_k^{(2)}$ проверки однородности

Снова обратимся к рис. 1. Теперь сформированные столбцы будем рассматривать как k отдельных последовательностей $\overline{z}_1, \ldots, \overline{z}_k$. Заметим, что при подобном разбиении каждая из последовательностей $\overline{z}_i$, $i = 1, \ldots, k$, в равной степени содержит информацию о каждом из блоков шифртекста $\overline{y}$. Рассмотрим статистику [9]

$$S_k^{(2)} = n \sum_{i=0}^{1} \sum_{j=1}^{k} \frac{\left(\nu_{ij} - \frac{\nu_{i*} \cdot \nu_{*j}}{n}\right)^2}{\nu_{i*} \cdot \nu_{*j}},$$

где ν_{ij} — число элементов последовательности $\overline{z}_j$, $j \in \{1, \ldots, k\}$, равных $i \in \{0, 1\}$, $\nu_{i*} = \sum\limits_{l=1}^{k} \nu_{ij}$, $\nu_{*j} = \sum\limits_{l=0}^{1} \nu_{ij}$.

С использованием указанной статистики были проведены аналогичные описанным выше исследования для тех же блочных алгоритмов шифрования.

Функциональная зависимость $S_k^{(2)}$ от опробуемых значений k (при шифровании текста № 1) для шифров AES и CAST приведена ниже (рис. 9–11 и 12, 13 соответственно).

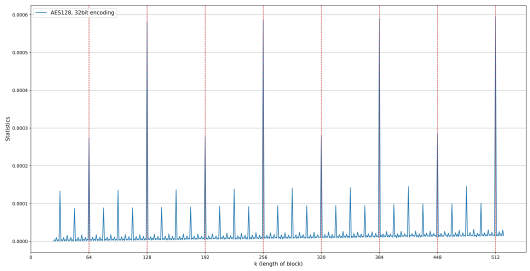

Рис. 9. Значения $S_k^{(2)}$ для AES и текста № 1 при 32-битной кодировке

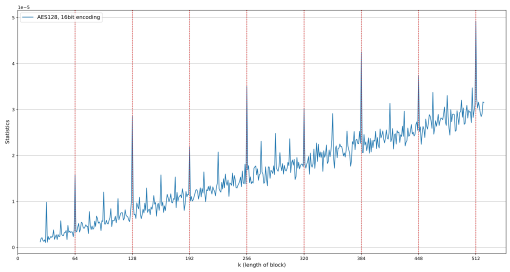

Рис. 10. Значения $S_k^{(2)}$ для AES и текста № 1 при 16-битной кодировке

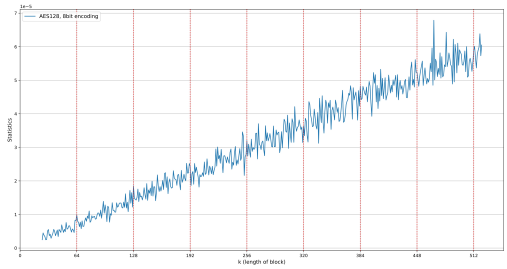

Рис. 11. Значения $S_k^{(2)}$ для AES и текста № 1 при 8-битной кодировке

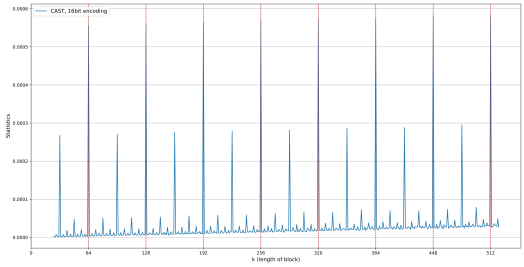

Рис. 12. Значения $S_k^{(2)}$ для CAST и текста № 1 при 16-битной кодировке

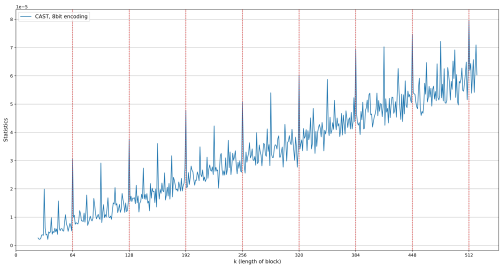

Рис. 13. Значения $S_k^{(2)}$ для CAST и текста №1 при 8-битной кодировке

В заключение отметим, что при справедливости гипотезы о независимости знаков шифртекста $\overline{y}$ статистика $S_k^{(1)}$ имеет при $n \to \infty$ асимптотическое распределение $\chi^2((r-1)(s-1))$, а при справедливости гипотезы об одинаковой распределенности знаков шифртекста $\overline{y}$ статистика $S_k^{(2)}$ — распределение $\chi^2(k-1)$, что позволяет строить на их основе статистические критерии проверки соответствующих гипотез.

Вывод. При наличии неравновероятности на множестве $\{0,1\}^m$ блоков закодированного открытого текста $\overline{x}$ или при выполнении (2) анализ поведения статистик $S_k^{(1)}$ и $S_k^{(2)}$, построенных на основе шифртекста $\overline{y}$, сформированного с использованием базового алгоритма зашифрования в режиме ECB, позволяет определить размер блока m в случаях использования $\frac{m}{4}, \frac{m}{2}, m$-битных кодировок открытого текста, записанного на русском, английском, французском и немецком языках. В частности, для базовых блочных алгоритмов зашифрования с размером блока 64 бита указанная возможность имеет место для широко используемых на практике 8,16-битных кодировок.

Рекомендации. В условиях необходимости использования произвольного базового блочного алгоритма зашифрования с размером блока m в режиме ECB во избежание потенциально возможности определения значения m (на основе статистик $S_k^{(1)}$ и $S_k^{(2)}$) следует использовать t-битную кодировку открытого текста, где $t \in \{\frac{m}{8}, \frac{m}{16}, \ldots\}$.

СПИСОК ЛИТЕРАТУРЫ

1. *Перов А. А.* Универсальный метод построения решающих правил с использованием сверточных нейронных сетей для анализа генераторов псевдослучайных последовательностей на основе итеративных блочных шифров: диссертация на соискание ученой степени кандидата технических наук. Красноярск: СФУ, 2020, 153 с.// *Perov A. A.* Universal'nyj metod postroeniya reshayushchih pravil s ispol'zovaniem svertochnyh nejronnyh setej dlya analiza generatorov psevdosluchajnyh posledovatel'nostej na osnove iterativnyh blochnyh shifrov: dissertaciya na soiskanie uchenoj stepeni kandidata tekhnicheskih nauk. Krasnoyarsk: SFU, 2020, 153 p. (in Russian)
2. *Погорелов Б. А.*, *Сачков В. Н.* Словарь криптографических терминов. М.: МЦНМО, 2006, 94 с.// *Pogorelov B. A.*, *Sachkov V. N.* Slovar' kriptograficheskih terminov. M.: MCNMO, 2006, 94 p. (in Russian)
3. *Ключарев П. Г.* О статистическом тестировании блочных шифров. — Математика и математическое моделирование, 2018, т. 5, с. 35–56.// *Klyucharev P. G.* O statisticheskom testirovanii blochnyh shifrov. — Matematika i matematicheskoe modelirovanie, 2018, 5, p. 35–56. (in Russian)
4. *Soto J.*, *Bassham L.* Randomness testing of the advanced encryption standard finalist candidates. National Bureau of Standards, 2000, NIST IR 6483, p. 1–14.

5. *Ивченко Г. И.*, *Медведев Ю. И.* Введение в математическую статистику. — Изд. 2-е, испр. и доп. М.: ЛЕНАНД, 2017, 608 с. // *Ivchenko G. I.*, *Medvedev Yu. I.* Vvedenie v matematicheskuyu statistiku. — Izd. 2-e, ispr. i dop. M.: LENAND, 2017, 608 p. (in Russian)
6. *Лось А. Б.*, *Нестеренко А. Ю.*, *Рожков М. И.* Криптографические методы защиты информации: учебник для академического бакалавриата. М.: Издательство Юрайт, 2018, 473 с. // *Los A. B.*, *Nesterenko A. Yu.*, *Rozhkov M. I.* Kriptograficheskie metodi zashchiti informacii: uchebnik dlya akademicheskogo bakalavriata. M.: Izdatelstvo Yurajt, 2018, 473 p. (in Russian)
7. *Духин А. А.* Теория информации: Учебное пособие М.: Гелиос АРВ, 2007, 248 с. // *Duhin A. A.* Teoriya informacii: Uchebnoe posobie M.: Gelios ARV, 2007, 248 p. (in Russian)
8. *Чечета С. И.* Введение в дискретную теорию информации и кодирования: учебное издание. М.: МЦНМО, 2011, 224 с. // *Checheta S. I.* Vvedenie v diskretnuyu teoriyu informacii i kodirovaniya: uchebnoe izdanie. M.: MCNMO, 2011, 224 p. (in Russian)
9. *Айвазян С. А.*, *Енюков И. С.*, *Мешалкин Л. Д.* Прикладная статистика: Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983, 471 с. // *Ajvazyan S. A.*, *Enyukov I. S.*, *Meshalkin L. D.* Prikladnaya statistika: Osnovy modelirovaniya i pervichnaya obrabotka dannyh. M.: Finansy i statistika, 1983, 471 p. (in Russian)

UDC 519.248

Mironkin V. O., Yurasov N. A. (National Research University Higher School of Economics, Moscow)**. On the question of the applicability of the statistics of the measure "chi-square"in the problem of determining the parameters of block encryption algorithms**

Abstract: The possibility of using the statistics of the measure "chi-square"to identify the parameters of the basic block encryption algorithm operating in the ECB encryption mode is studied. Conditions which make it possible the use of statistics of the specified type to identify the block size of the corresponding algorithm are formulated. Recommendations on the choice of some parameters of the encoding of plaintext characters in the conditions of using the ECB encryption mode are offered.

Keywords: Block encryption algorithm, basic block encryption algorithm, encryption mode, statistics, criterion, hypothesis, chi-squared distribution, ECB mode.