Теория вероятностей и математическая статистика

Probability theory and mathematical statistics

УДК 519.226

О МОЩНОСТИ ТЕСТОВ МНОГОМЕРНОЙ ДИСКРЕТНОЙ РАВНОМЕРНОСТИ, ИСПОЛЬЗУЕМЫХ ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА ГЕНЕРАТОРОВ СЛУЧАЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

В. А. ВОЛОШКО[1)]***, А. И. ТРУБЕЙ***[1)]

[1)]*Научно-исследовательский институт прикладных проблем математики и информатики БГУ, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Получена асимптотика мощностей статистических тестов многомерной дискретной равномерности в условиях контигуального сближения альтернатив. Рассмотрены две версии теста многомерной дискретной равномерности – по пересекающимся отрезкам (входит в состав батареи тестов NIST SP 800-22) и по непересекающимся отрезкам. Нулевой гипотезе H_0 соответствует так называемая чистая случайность наблюдаемой последовательности, т. е. независимость и одинаковое равномерное распределение ее элементов. Альтернатива H_1 предполагается цепью Маркова некоторого произвольного фиксированного конечного порядка.

Ключевые слова: мощность теста; тест многомерной дискретной равномерности; контигуальные альтернативы; нецентральное распределение хи-квадрат; генератор случайных чисел; цепь Маркова.

Образец цитирования:
Волошко ВА, Трубей АИ. О мощности тестов многомерной дискретной равномерности, используемых для статистического анализа генераторов случайных последовательностей. *Журнал Белорусского государственного университета. Математика. Информатика.* 2022;1:26–37.
https://doi.org/10.33581/2520-6508-2022-1-26-37

For citation:
Voloshko VA, Trubey AI. On the power of tests of multidimensional discrete uniformity used for statistical analysis of random number generators. *Journal of the Belarusian State University. Mathematics and Informatics.* 2022;1:26–37. Russian.
https://doi.org/10.33581/2520-6508-2022-1-26-37

Авторы:
Валерий Анатольевич Волошко – кандидат физико-математических наук; заведующий сектором компьютерного анализа данных.
Антон Иванович Трубей – заведующий научно-исследовательской лабораторией прикладной информатики.

Authors:
Valeriy A. Voloshko, PhD (physics and mathematics); head of the sector of computer data analysis.
valeravoloshko@yandex.ru
https://orcid.org/0000-0002-9693-0688
Anton I. Trubey, head of the laboratory of applied informatics.
trubeia@mail.ru

ON THE POWER OF TESTS OF MULTIDIMENSIONAL DISCRETE UNIFORMITY USED FOR STATISTICAL ANALYSIS OF RANDOM NUMBER GENERATORS

***V. A. VOLOSHKO*[a], *A. I. TRUBEY*[a]**

[a]*Research Institute for Applied Problems of Mathematics and Informatics, Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
Corresponding author: V. A. Voloshko (valeravoloshko@yandex.ru)

In this paper, we obtained the asymptotic power values for the statistical tests of multidimensional discrete uniformity under conditions of contiguous convergence of alternatives. Two versions of the test are considered, namely, with overlapping blocks (included in the NIST SP 800-22 test suit) and with non-overlapping blocks. The null hypothesis H_0 is related to the so-called pure randomness of the observed sequence, i. e. independence and the same uniform distribution of its elements. An alternative H_1 is assumed to be a Markov chain of some arbitrary fixed finite order.

Keywords: power of a test; test of multidimensional discrete uniformity; contiguous alternatives; non-central chi-squared distribution; random number generator; Markov chain.

Введение

Многие криптографические задачи (например, генерация ключей, анализ стойкости криптографических алгоритмов) требуют применения статистических критериев для обнаружения большого числа отклонений от гипотетической модели. Теория вероятностей описывает и анализирует случайность с помощью абстрактных математических объектов, моделируя ее случайными величинами и случайными процессами. Математическая статистика посредством экспериментов связывает эти абстрактные математические модели с реально существующими генераторами случайных последовательностей. Данные эксперименты могут быть использованы для оценки параметров, описывающих модели, или для проверки гипотез, проистекающих из моделей. Разработчику генераторов следует провести исследования в целях обнаружения отклонений двоичной последовательности от модели независимых симметричных испытаний Бернулли, а также минимизации вероятности ошибки второго рода для применяемых тестов.

Общие сведения о проверке статистических гипотез

С каждым из используемых для проверки гипотезы H_0 тестов связана определенная статистика S, которая в соответствии с некой мерой измеряет отклонение в наблюдаемом процессе от ситуации, отвечающей H_0. В силу случайности извлекаемых выборок случайными оказываются и значения статистики S, вычисляемые по этим выборкам. При справедливости гипотезы H_1 статистика S подчиняется некоторому распределению $F_{S|H_i}(z) = P\{S \le z | H_i\}$.

Схема проверки гипотезы заключается в следующем. Область значений статистики S разбивается на два подмножества. Одно из них представляет собой критическую область, попадание в которую при справедливости H_0 маловероятно. При попадании вычисленного по выборке $x_1, x_2, \ldots, x_n$ значения статистики S в критическую область проверяемая гипотеза H_0 отклоняется (отвергается). В противном случае нет оснований для отклонения гипотезы H_0.

С проверкой статистических гипотез связывают вероятности ошибок двух видов. С одной стороны, справедливая гипотеза H_0 может быть отклонена, и этим будет совершена ошибка первого рода. При проверке гипотез, как правило, задают вероятность ошибки первого рода α (уровень значимости), допуская тем самым возможность отклонения H_0 и совершения такой ошибки. С другой стороны, может быть справедлива некоторая конкурирующая гипотеза H_1. Если при справедливости H_1 в процессе проверки гипотеза H_0 не была отклонена, то этим самым совершена ошибка второго рода. Ее вероятность будем обозначать β.

Обычно, используя тесты для проверки гипотез, не рассматривают конкретную конкурирующую гипотезу. В таком случае при проверке гипотез о случайности можно считать, что конкурирующая гипотеза связана с наличием, например, какого-то отклонения от случайности (неравновероятности или зависимости между знаками).

Если же гипотеза H_1 задана и связана с наличием конкретного отклонения от случайности (к примеру, неравновероятности заданного типа), то задание величины α для используемого теста проверки гипотез определяет и вероятность ошибки второго рода β. Так, в случае правостороннего теста вида

$$\begin{cases} H_0,\ S \le S_{1-\alpha}, \\ H_1,\ S > S_{1-\alpha}, \end{cases} \tag{1}$$

где $S_{1-\alpha} = F^{-1}_{S|H_0}(1-\alpha)$, вероятность ошибки второго рода

$$\beta = F_{S|H_1}\left(S_{1-\alpha}\right). \tag{2}$$

Величину $1 - \beta$ принято называть мощностью статистического теста. Чем выше мощность теста при заданном значении α, тем лучше он различает гипотезы H_0 и H_1. Особенно важно, чтобы используемый тест хорошо различал близкие конкурирующие гипотезы.

Очевидно, что при проверке любой статистической гипотезы желательно применять наиболее мощный критерий, который для заданной вероятности ошибки первого рода α обеспечивает минимальную вероятность ошибки второго рода β относительно любой конкурирующей гипотезы H_1. Лучше всего использовать равномерно наиболее мощный критерий, который для любого заданного α обеспечивает минимальное значение β. Однако существование такого критерия для проверки конкретной гипотезы H_0 является редчайшим исключением.

Имеющиеся наборы тестов, в том числе NIST SP 800-22 [1], предлагают современные статистические тесты для генераторов (псевдо)случайных последовательностей, позволяющие обнаружить отклонения двоичной последовательности от модели независимых симметричных испытаний Бернулли. Одной из основных целей данных статистических тестов является минимизация вероятности ошибки второго рода, т. е. минимизация вероятности принятия последовательности, созданной генератором, в качестве удовлетворительной, в то время как генератор в действительности был некачественным. Вероятности α и β связаны друг с другом, а также с размером n тестируемой последовательности таким образом, что если два из этих значений указаны, то третье значение определяется автоматически.

На практике обычно выбираются уровень значимости α, тип альтернативы H_1 и приемлемая вероятность ошибки второго рода β (вероятность решения, что некачественный генератор произвел в действительности случайную последовательность). Затем вычисляется достаточный размер последовательности n, обеспечивающий выбранные значения α и β. При этом тесты и альтернативы H_1 взаимосвязаны: для каждого типа альтернативы существуют свои оптимальные тесты, которые лучше других тестов отличают нулевую гипотезу H_0 от данного типа альтернативы H_1. Таким образом, батареи тестов позволяют охватить некоторый конечный набор альтернатив. Реальная, возникающая на практике альтернатива при этом в случае неверной нулевой гипотезы неизвестна, поэтому реальное значение вероятности ошибки второго рода β не может быть вычислено, а могут быть определены только значения β, относящиеся к гипотетическим альтернативам H_1.

Далее рассмотрены статистические тесты многомерной дискретной равномерности по непересекающимся и по пересекающимся отрезкам. Для этих тестов в условиях бернуллиевских и марковских (рекуррентных) альтернатив H_1, контигуально сближающихся с нулевой гипотезой H_0, получены асимптотические выражения для вероятностей ошибки второго рода β.

Тест многомерной дискретной равномерности по непересекающимся *L*-отрезкам

Введем обозначения: Z – произвольное конечное множество; $\mathbb{Z}$ – множество целых чисел; $U(Z)$ – равномерное распределение вероятностей $p(z) \equiv |Z|^{-1}$, $z \in Z$; $\Lambda\{\xi\}$ – распределение вероятностей случайной величины $\xi \in Z$. Дивергенция Кульбака – Лейблера двух распределений вероятностей p, q на Z имеет вид

$$\mathrm{KL}(p; q) = \sum_{z \in Z} p(z) \ln \frac{p(z)}{q(z)} \ge 0.$$

Энтропия Шеннона определяется по формуле

$$H(p) = -\sum_{z \in Z} p(z) \ln p(z) \ge 0.$$

Для последовательности $\mathbf{z} = (z_1, \ldots, z_n) \in Z^n$ элементов множества Z введем следующие функции и величины:

1) функцию частот

$$f = f_z : Z \to \mathbb{R},\ f = \psi(z),\ f_z = \sum_{i=1}^{n} 1\{z_i = z\},$$

где $1\{A\}$ – индикаторная функция события A;

2) статистику хи-квадрат относительно равномерного распределения вероятностей

$$\chi^2(z_1, \ldots, z_n) = \sum_{z \in Z} \frac{\left(f_z - n/|Z|\right)^2}{n/|Z|},\ f = \psi(z);$$

3) последовательность пересекающихся L-грамм (слов длины L) в закольцованной последовательности $\mathbf{z}$

$$\varphi_L(\mathbf{z}) = \left((z_1, z_2, \ldots, z_L), (z_2, z_3, \ldots, z_{L+1}), \ldots, (z_n, z_1, \ldots, z_{L-1})\right) \in \left(Z^L\right)^n;$$

4) последовательность непересекающихся L-грамм в последовательности $\mathbf{z}$ (в этом случае предполагается, что L делит n, $n' = \frac{n}{L}$)

$$\varphi_L(\mathbf{z}) = \left((z_1, \ldots, z_L), (z_{L+1}, \ldots, z_{2L}), \ldots, (z_{n-L+1}, \ldots, z_n)\right) \in \left(Z^L\right)^{n'}.$$

Одним из классических тестов, предложенных Д. Кнутом, является тест многомерной дискретной равномерности по непересекающимся L-отрезкам (далее – МДРН(L)-тест). Данный тест предназначен для проверки гипотезы согласия наблюдаемой последовательности L-векторов с L-мерным дискретным равномерным распределением. Предположим, что имеем двоичную последовательность

$$\mathbf{x} = (x_1, \ldots, x_n) \in V^n,\ V = \{0, 1\},$$

длины $n \in \mathbb{N}$, кратной L ($n' = \frac{n}{L}$). Статистика хи-квадрат МДРН(L)-теста

$$S = \chi^2(\varphi_L(\mathbf{x})) = \sum_{z \in V^L} \frac{\left(f_z - n'/2^L\right)^2}{n'/2^L}, \tag{3}$$

где $f = \psi(\varphi_L(\mathbf{x}))$ – частоты встречаемости среди n' непересекающихся L-грамм в последовательности $\mathbf{x}$. Своим названием статистика хи-квадрат МДРН(L)-теста (как и все прочие статистики хи-квадрат) обязана тому свойству, что при истинной нулевой гипотезе H_0 с ростом n она сходится по распределению к закону хи-квадрат [2]:

$$F_{S|H_0}(\cdot) \underset{n \to \infty}{\overset{D}{\to}} F_{\chi^2_d}(\cdot),$$

где в случае МДРН(L)-теста число степеней свободы $d = 2^{L-1}$.

Мы будем рассматривать ситуацию контигуального сближения альтернатив, когда с ростом n альтернатива H_1 сходится к гипотезе H_0 таким образом, что статистика хи-квадрат S при истинной альтернативе H_1 сходится по распределению к нецентральному распределению хи-квадрат с некоторым параметром нецентральности $\lambda > 0$:

$$F_{S|H_1}(\cdot) \underset{n \to \infty}{\overset{D}{\to}} F_{\chi^2_{d,\lambda}}(\cdot).$$

При $\lambda = 0$ $F_{\chi^2_{d,0}}(\cdot) = F_{\chi^2_d}(\cdot)$. Тогда из выражений (1) и (2) получаем соотношение, связывающее асимптотические вероятности ошибок первого и второго рода:

$$\beta = F_{\chi^2_{d,\lambda}}\left(F^{-1}_{\chi^2_d}(1-\alpha)\right). \tag{4}$$

Как видим, соотношение (4) зависит от двух параметров – числа степеней свободы d и параметра нецентральности λ. Число степеней свободы d зависит только от вида статистического теста и не зависит от вида альтернативы H_1. Приведенные далее результаты основаны на формуле (4) и на вычислении параметра нецентральности λ для двух видов тестов хи-квадрат (по непересекающимся и по пересекающимся отрезкам) в случае марковских альтернатив H_1.

Определение 1. Будем говорить, что марковская альтернатива H_1 контигуально сближается с нулевой гипотезой H_0, если при истинной альтернативе H_1 последовательность $\{x_i\}$ представляет собой

стационарную цепь Маркова некоторого фиксированного порядка $s \in \mathbb{N}$, не зависящего от n, и переходные вероятности цепи Маркова $\{x_i\}$ удовлетворяют следующей асимптотике при $n \to \infty$:

$$\sum_{q=(q_1,\ldots,q_s)\in V^s} \delta^2(q) = O\left(n^{-1}\right),$$

$$\delta(q) = P\left\{x_i = 0 \middle| x_{i-1} = q_1, \ldots, x_{i-s} = q_s\right\} - \frac{1}{2}. \tag{5}$$

Для краткости s будем называть порядком марковской альтернативы H_1.

Введем функцию энтропии L-вектора элементов тестируемой последовательности в случае истинной альтернативы H_1 и связанные величины:

$$\mathrm{h}(L) = \mathrm{H}\left(\Lambda\left\{x_1, \ldots, x_L \middle| H_1\right\}\right),\ L > 0,\ \mathrm{h}(0) = 0,$$

$$\Delta\mathrm{h}(L) = \mathrm{h}(L) - \mathrm{h}(L-1),\ L > 0,$$

$$\mathrm{h}^* = \lim_{L\to\infty} \Delta\mathrm{h}(L) = \lim_{L\to\infty} \frac{\mathrm{h}(L)}{L}.$$

Величина h^* называется удельной энтропией случайной последовательности $\{x_i\}$ при истинной альтернативе H_1.

Теорема 1. *Пусть марковская альтернатива H_1 контигуально сближается с нулевой гипотезой H_0. Тогда МДРН(L)-тест имеет асимптотическую ошибку второго рода (4) с числом степеней свободы $d = 2^L - 1$ и параметром нецентральности вида*

$$\lambda = 2\lim_{n\to\infty} n\left(\ln 2 - \frac{\mathrm{h}(L)}{L}\right). \tag{6}$$

Доказательство. Для применения формулы (4) достаточно доказать, что статистика хи-квадрат (3) при истинной альтернативе H_1 имеет асимптотическое нецентральное распределение хи-квадрат с параметром нецентральности (6). При истинной марковской альтернативе H_1 некоторого порядка s последовательность непересекающихся L-грамм $\varphi_L(\mathbf{x})$ длины $n' = \frac{n}{L}$ также представляет собой цепь Маркова некоторого порядка s'. Нормированные частоты $\hat{\pi}_z = \frac{f_z}{n'}$, $z \in V^L$, $f = \psi\left(\varphi_L(\mathbf{x})\right)$, представляют собой состоятельные асимптотически нормальные [3; 4] статистические оценки вероятностей L-грамм $\pi_z = P\left\{(x_1, \ldots, x_L) = z \middle| H_1\right\}$. При контигуальном сближении альтернатив распределение $\pi = (\pi_z)_{z\in V^L}$ сходится к равномерному закону $U\left(V^L\right)$, и касательные пространства точек π и $U\left(V^L\right)$ в пространстве вероятностных распределений на V^L совпадают, равно как и асимптотические ковариационные матрицы Σ_0 и Σ_1 оценок $\hat{\pi} = \left(\hat{\pi}_z\right)_{z\in V^L}$ в случае нулевой гипотезы H_0 и альтернативы H_1 соответственно. Точке $U\left(V^L\right)$ в ее собственном касательном пространстве отвечает нулевой вектор, а близкой к $U\left(V^L\right)$ точке p – вектор, который будем обозначать $v(p)$. Статистика хи-квадрат (3) МДРН(L)-теста представляет собой квадрат нормы вектора $\sqrt{n'}\cdot v\left(\hat{\pi}\right)$ в стандартной метрике Фишера [3]. При нулевой гипотезе H_0 вектор $\sqrt{n'}\cdot v\left(\hat{\pi}\right)$ имеет асимптотическое центрированное стандартное нормальное распределение, а при альтернативе H_1 ввиду равенства асимптотических ковариационных матриц $\Sigma_0 = \Sigma_1$ – асимптотическое нецентрированное стандартное нормальное распределение со средним $\mu = \sqrt{n'}\cdot v(\pi)$, и, следовательно, статистика хи-квадрат S имеет асимптотическое нецентральное распределение хи-квадрат с параметром нецентральности

$$\lambda = \lim_{n'\to\infty} \|\mu\|^2 = 2\lim_{n'\to\infty} n'\cdot \mathrm{KL}\left(\pi; U\left(V^L\right)\right) =$$

$$= 2\lim_{n'\to\infty} n'\left(\ln\left|V^L\right| - \mathrm{H}\left(\Delta\left(x_1, \ldots, x_L \middle| H_1\right)\right)\right) = 2\lim_{n\to\infty} n\left(\ln 2 - \frac{\mathrm{h}(L)}{L}\right).$$

Теорема доказана.

Тест многомерной дискретной равномерности по пересекающимся L-отрезкам

Тест многомерной дискретной равномерности по пересекающимся L-отрезкам (далее – МДРП(L)-тест) предназначен для проверки гипотезы согласия наблюдаемой последовательности L-векторов с L-мерным дискретным равномерным распределением. Данный тест позволяет обнаруживать отклонения от L-мерного дискретного равномерного распределения типа рекуррентной марковской зависимости порядка $s < L$ (меньшего, чем размерность вектора L).

Статистика хи-квадрат МДРП(L)-теста имеет вид

$$S = \chi^2\left(\varphi_L(\mathbf{x})\right) - \chi^2\left(\varphi_{L-1}(\mathbf{x})\right) = \sum_{z \in V^L} \frac{\left(f_z - n/2^L\right)^2}{n/2^L} - \sum_{z \in V^{L-1}} \frac{\left(g_z - n/2^{L-1}\right)^2}{n/2^{L-1}}, \tag{7}$$

где $f = \psi\left(\varphi_L(\mathbf{x})\right)$ и $g = \psi\left(\varphi_{L-1}(\mathbf{x})\right)$ – частоты встречаемости среди n пересекающихся L-грамм и $(L-1)$-грамм соответственно в закольцованной последовательности $\mathbf{x}$.

Теорема 2. *Пусть марковская альтернатива H_1 контигуально сближается с нулевой гипотезой H_0. Тогда МДРП(L)-тест имеет асимптотическую ошибку второго рода (4) с числом степеней свободы $d = 2^{L-1}$ и параметром нецентральности вида*

$$\lambda = 2 \lim_{n \to \infty} n\left(\ln 2 - \Delta h(L)\right). \tag{8}$$

Доказательство. Как и в теореме 1, для применения формулы (4) достаточно доказать, что статистика хи-квадрат (7) при истинной альтернативе H_1 имеет асимптотическое нецентральное распределение хи-квадрат с параметром нецентральности (8). Обозначим

$$p^k = \left(p_z^k\right)_{z \in V^k},\ p_z^k = \frac{P\left\{\left(x_1, \ldots, x_k, x_{k+1}\right) = (z, 0)\right\}}{P\left\{\left(x_1, \ldots, x_k\right) = z\right\}}.$$

Для цепи Маркова порядка s вектор p^s – стандартный параметр модели (вектор переходных вероятностей в нулевой знак) и система координат в пространстве $\mathrm{MC}(s)$ цепей Маркова порядка s. При этом вектор p^k, $k < s$, представляет точку в подпространстве $\mathrm{MC}(k) \subset \mathrm{MC}(s)$ цепей Маркова меньшего порядка k:

$$p^k = \wp_{s \to k}\left(p^s\right),$$

где $\wp_{s \to k}$ – оператор m-проекции [3] экспоненциального семейства $\mathrm{MC}(s)$ [5] на свое экспоненциальное подсемейство $\mathrm{MC}(k)$. На уровне касательных пространств m-проекция действует как ортогональная проекция: можно считать [3], что касательное пространство T^s в точке p^s в пространстве $\mathrm{MC}(s)$ содержит в качестве подпространства касательное пространство T^k в точке p^k в пространстве $\mathrm{MC}(k)$, и оператор $\wp_{s \to k} : T^s \to T^k$ – ортогональный проектор в метрике Фишера.

Не ограничивая общности, можем считать, что $s > L$. Рассмотрим стандартные статистические оценки параметров цепи Маркова порядка k [4]:

$$\hat{p}^k = \left(\hat{p}_z^k\right)_{z \in V^k},\ \hat{p}_z^k = \frac{f_{(z,0)}}{g_z},\ f = \psi\left(\varphi_{k+1}(\mathbf{x})\right),\ g = \psi\left(\varphi_k(\mathbf{x})\right).$$

При истинной марковской альтернативе H_1 порядка s оценка $\hat{p}^s$ состоятельна, асимптотически нормальна и эффективна [4] (достигает границы Крамера – Рао), и оценка $\hat{p}^{L-1} = \wp_{s \to L-1}\left(\hat{p}^s\right)$ благодаря описанным выше проективным свойствам отображения $\wp$ обладает аналогичными свойствами асимптотической нормальности и эффективности. Пусть точка $q^k = \left(q_z^k \equiv \frac{1}{2}\right)_{z \in V^k}$ отвечает чисто случайной последовательности (нулевой гипотезе H_0). Как и в доказательстве теоремы 1, при контигуальном сближении альтернатив касательные пространства точек p^{L-1} и q^{L-1} совпадают, равно как и асимптотические ковариационные матрицы оценки $\hat{p}^{L-1}$ при истинной нулевой гипотезе H_0 и истинной альтернативе H_1 соответственно. Поэтому в касательном пространстве точки q^{L-1} вектор $\sqrt{n} \cdot v\left(\hat{p}^{L-1}\right)$ имеет асимптотическое нецентрированное стандартное нормальное распределение со средним $\mu = \sqrt{n} \cdot v\left(p^{L-1}\right)$, и квадрат нормы

$$S' = \left\| \sqrt{n} \cdot v\left(\hat{p}^{L-1} \right) \right\|^2$$

имеет асимптотическое нецентральное распределение хи-квадрат с параметром нецентральности

$$\lambda' = \lim_{n \to \infty} \|\mu\|^2 = 2 \lim_{n \to \infty} n \cdot \mathrm{KL}\left(p^{L-1}; q^{L-1} \right). \tag{9}$$

Метрика Фишера в пространстве $\mathrm{MC}(L-1)$ представляет собой разность [3; 5] метрик Фишера в пространствах распределений на L-граммах и на $(L-1)$-граммах, а именно информационная матрица Фишера

$$I\left(p^{L-1} \right) = I_L\left(p^{L-1} \right) - I_{L-1}\left(p^{L-1} \right),$$

где $I_k\left(p^{L-1} \right)$ – информационная матрица Фишера распределения k-грамм $\left(x_i \right)_{i=1}^k$ относительно параметра p^{L-1}. Поэтому $S' = S$ (разность (7) есть разность квадратов касательных векторов в метрике Фишера между L-граммами и $(L-1)$-граммами, чьи распределения задаются цепью Маркова порядка $L-1$ с параметром $\hat{p}^{L-1}$). Аналогично распадается в разность дивергенция Кульбака – Лейблера в формуле (9):

$$\mathrm{KL}\left(p^{L-1}; q^{L-1} \right) = \mathrm{KL}\left(\Lambda\left\{ x_1, \ldots, x_L \middle| H_1 \right\}; U\left(V^L \right) \right) - \mathrm{KL}\left(\Lambda\left\{ x_1, \ldots, x_{L-1} \middle| H_1 \right\}; U\left(V^{L-1} \right) \right) =$$
$$= L \ln 2 - \mathrm{h}(L) - (L-1)\ln 2 + \mathrm{h}(L-1) = \ln 2 - \Delta\mathrm{h}(L),$$

что в объединении с выражением (9) дает $\lambda = \lambda'$ (см. формулу (8)). Таким образом, мы доказали, что $S' = S$ имеет асимптотическое нецентральное распределение хи-квадрат с параметром нецентральности $\lambda' = \lambda$. Теорема доказана.

Вычисления в терминах отклонения переходных вероятностей марковской альтернативы

Пределы (6), (8) могут быть выражены в терминах функции (5) отклонений переходных вероятностей марковской альтернативы H_1 от чистой случайности.

Лемма 1. *Пусть марковская альтернатива H_1 порядка $s \geq 0$ контигуально сближается с нулевой гипотезой H_0. Тогда в обозначениях определения 1 имеет место следующее асимптотическое соотношение:*

$$\lim_{n \to \infty} n\left(\ln 2 - \Delta\mathrm{h}(L) \right) = 2\rho_{\min\{s, L-1\}}, \; L \geq 1, \tag{10}$$

$$\rho_r = \lim_{n \to \infty} n \cdot \mathrm{M}_2\left\{ \mathrm{E}\left\{ \delta(q) \middle| q_1, \ldots, q_r \right\} \right\}, \; 0 \leq r < s,$$

где $\mathrm{M}_2\{\xi\} = \mathrm{E}\left\{ \xi^2 \right\}$ для случайной величины $\xi \in \mathbb{R}$, случайный двоичный вектор $q = (q_1, \ldots, q_s) \in V^s$ равномерно распределен.

Доказательство. Используем обозначения доказательства теоремы 2. Как отмечалось в указанном доказательстве, $\Delta\mathrm{h}(r+1)$ есть удельная энтропия цепи Маркова порядка r с заданным распределением $(r+1)$-грамм, поэтому при $r \geq s$ эта разность равна удельной энтропии марковской альтернативы H_1 порядка s, а при $r < s$ – удельной энтропии проекции $\wp_{s \to r}(H_1)$. Разность $\ln 2 - \Delta\mathrm{h}(r+1)$ равна дивергенции Кульбака – Лейблера между чисто случайной цепью Маркова (гипотеза H_0) и проекцией $\wp_{s \to r}(H_1)$. При малых δ для $s = r$ эта дивергенция асимптотически эквивалентна [3] величине $\frac{1}{2} \sum_{q, q' \in V^s} \delta(q)\delta(q') I_{q, q'}$, где $I \in \mathbb{R}^{2^s \times 2^s}$ – информационная матрица Фишера цепи Маркова порядка s относительно параметра δ в точке $\delta \equiv 0$. Матрица I диагональна [4] со всеми диагональными элементами $I_{q, q} \equiv 2^{2-s}$, откуда получаем соотношение (10) для $L > s$. Ортогональный проектор $\wp_{s \to r} : T^s \to T^r$ относительно единичной с точностью до множителя матрицы Фишера I действует на функцию $\delta(q)$ усреднением по $q_{r+1}, \ldots, q_s$ при фиксированных $q_1, \ldots, q_r$:

$$\wp_{s \to r}(\delta) = \tilde{\delta} : V^r \to \mathbb{R}, \; \tilde{\delta}(q_1, \ldots, q_r) = \mathrm{E}\left\{ \delta(q) \middle| q_1, \ldots, q_r \right\},$$

откуда по аналогии со случаем $r = s$ при $r < s$ получаем

$$\lim_{n\to\infty} n\big(\ln 2 - \Delta h(r+1)\big) = 2\lim_{n\to\infty} n\cdot \mathrm{M}_2\big\{\tilde{\delta}(q_1, \ldots, q_r)\big\},$$

что эквивалентно соотношению (10) для $L \le s$. Лемма доказана.

Следствие. *Пусть марковская альтернатива H_1 порядка $s \ge 0$ контигуально сближается с нулевой гипотезой H_0. Тогда величина (6) в обозначениях леммы 1 имеет вид*

$$\lambda = 4\frac{\omega_{L-1}}{L},\ \omega_r = \sum_{i=0}^{r} \rho_{\min\{s,\, i\}}. \tag{11}$$

Величина (8) соответствует выражению

$$\lambda = 4\rho_{\min\{s,\, L-1\}}. \tag{12}$$

Доказательство. С использованием тождества (10) выражение (12) прямо получается из параметра нецентральности (8), а выражение (11) – из параметра нецентральности (6), представленного как деленная на L частичная сумма левых частей тождества (10). Следствие доказано.

Таким образом, для вычисления асимптотической вероятности ошибки второго рода МДРН(L)-теста и МДРП(L)-теста при контигуальном сближении марковской альтернативы H_1 с нулевой гипотезой H_0 нужно вычислить асимптотическое значение параметра нецентральности λ и подставить его в формулу (4). В тех простых случаях, когда марковская альтернатива H_1 допускает явное выражение функции $h(L)$, значение λ для обоих тестов может быть вычислено непосредственно по формулам (6) и (8). В общем случае следует использовать формулы (11) и (12). Приведем примеры вычисления параметра нецентральности λ обоими способами.

Рассмотрим альтернативу H_1, представляющую собой цепь Маркова порядка s с одной частичной связью MC$(s, 1)$ [6], так что переходная вероятность $P\big\{x_t = 0 \big| x_{t-1}, \ldots, x_{t-s}\big\} = \frac{1}{2} + (-1)^{x_{t-s}}\delta$ зависит только от x_{t-s}, δ – малое отклонение ($|\delta| \ll 1$). Тогда цепь Маркова $\{x_t\}$ распадается на s независимых цепей Маркова первого порядка $\{x_{st+k}\}_{t\in\mathbb{Z}}$, $k = 1, \ldots, s$, члены которых чередуются в $\{x_t\}$. К отрезку $x_1, \ldots, x_L$ длины L последовательности $\{x_t\}$ применим взаимно однозначное (а следовательно, сохраняющее энтропию) преобразование $(x_1, \ldots, x_L) \to (y_1, \ldots, y_L)$, где $y_i = x_i$ при $i \le s$ и $y_i = x_i \oplus x_{i-s}$ при $i > s$. Нетрудно видеть, что $\{y_i\}$ независимы в совокупности, причем $P\{y_i = 0\} = \frac{1}{2}$ при $i \le s$ и $P\{y_i = 0\} = \frac{1}{2} + \delta$ при $i > s$. Таким образом,

$$h(L) = \mathrm{H}\big(\Lambda\{x_1, \ldots, x_L\}\big) = \mathrm{H}\big(\Lambda\{y_1, \ldots, y_L\}\big) = \sum_{i=1}^{L} \mathrm{H}\big(\Lambda\{y_i\}\big) = \begin{cases} L\ln 2,\ L \le s, \\ s\ln 2 + (L-s)h^*,\ L > s, \end{cases}$$

где удельная энтропия

$$h^* = -\left(\frac{1}{2}+\delta\right)\ln\left(\frac{1}{2}+\delta\right) - \left(\frac{1}{2}-\delta\right)\ln\left(\frac{1}{2}-\delta\right) = \ln 2 - 2\delta^2 + O\big(\delta^4\big).$$

Отсюда при $L \le s$ имеем $\lambda = 0$ (неразличимость гипотез H_0 и H_1) как для МДРН(L)-теста (формула (6)), так и для МДРП(L)-теста (формула (8)). При $L > s$ для МДРН(L)-теста $\lambda = 4\left(1 - \frac{s}{L}\right)\kappa$, для МДРП$(L)$-теста $\lambda = 4\kappa$, где $\kappa = \lim\limits_{n\to\infty} n\delta^2$. Эти формулы верны и в случае $s = 0$, когда марковская альтернатива MC$(0, 1)$ превращается в схему независимых испытаний Бернулли (на рис. 1–5 обозначена как Ber).

Теперь рассмотрим альтернативу H_1, представляющую собой цепь Маркова порядка s с почти идеальными переходными вероятностями, нарушенными только для одной из 2^s предысторий:

$$P\big\{x_t = 0 \big| x_{t-1}, \ldots, x_{t-s}\big\} = \begin{cases} \frac{1}{2},\ (x_{t-1}, \ldots, x_{t-s}) \ne \hbar, \\ \frac{1}{2} + \delta,\ (x_{t-1}, \ldots, x_{t-s}) = \hbar, \end{cases}$$

где $\hbar = (\hbar_1, \ldots, \hbar_s) \in V^s$ – некоторая выделенная «плохая» предыстория. В этом случае функция (5) принимает вид $\delta(q) = \delta \cdot 1\{q = \hbar\}$. Для вычислений согласно лемме 1 полагаем q чисто случайным двоичным s-вектором. Обозначим $q' = (q_i)_{i=1}^{r}$, $q'' = (q_i)_{i=r+1}^{s}$, $\hbar' = (\hbar_i)_{i=1}^{r}$, $\hbar'' = (\hbar_i)_{i=r+1}^{s}$. Тогда

$$\mathrm{E}\{\delta(q)|q'\} = \mathrm{E}\{\delta \cdot 1\{q' = \hbar'\} \cdot 1\{q'' = \hbar''\}|q'\} = \delta \cdot 1\{q' = \hbar'\} \cdot \mathrm{E}\{1\{q'' = \hbar''\}|q'\} =$$

$$= \delta \cdot 1\{q' = \hbar'\} \cdot \mathrm{E}\{1\{q'' = \hbar''\}\} = \delta \cdot 1\{q' = \hbar'\} \cdot P\{q'' = \hbar''\} = 2^{r-s}\delta \cdot 1\{q' = \hbar'\},$$

$$\mathrm{M}_2\{\mathrm{E}\{\delta(q)|q'\}\} = 2^{2r-2s}\delta^2 P\{q' = \hbar'\} = 2^{r-2s}\delta^2,$$

$$\rho_r = \lim_{n\to\infty} n \cdot \mathrm{M}_2\{\mathrm{E}\{\delta(q)|q'\}\} = 2^{r-2s}\kappa,$$

$$\omega_r = 2^{-2s}\kappa \cdot \begin{cases} 2^{r+1} - 1, & r \le s, \\ (r - s + 2)2^s - 1, & r > s. \end{cases}$$

Подстановка найденных выражений в формулы (11) и (12) дает значения λ для МДРН(L)-тестов и МДРП(L)-тестов соответственно.

Численные примеры

На рис. 1–5 приведены графики зависимости вероятности ошибки второго рода β МДРН(L)-теста и МДРП(L)-теста от размера выборки n и от параметра δ отклонения марковской альтернативы H_1 от нулевой гипотезы H_0. Для вычислений применялась формула (4) с подстановкой значений параметра нецентральности λ, найденных по формулам (6), (8), (11), (12). Оценки сверху для используемой в соотношении (4) функции $F_{\chi^2_{d,\lambda}}(\cdot)$ приведены в [7]. Вычисление этой функции встроено во многие среды, например в Python и Wolfram.

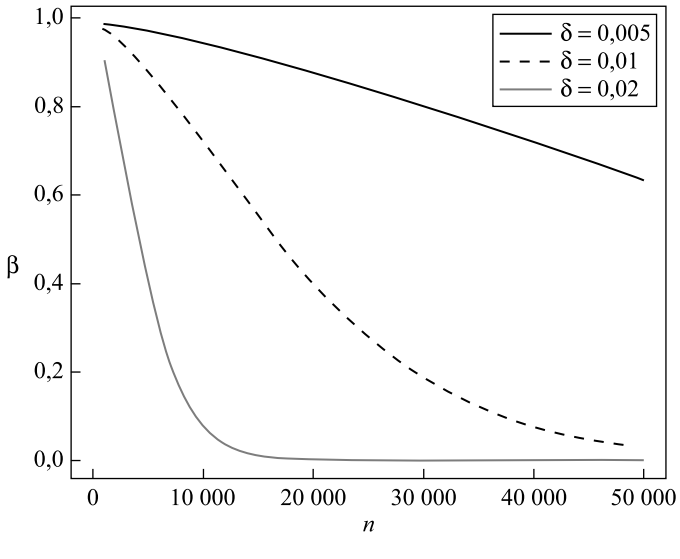

Рис. 1. Зависимость вероятности ошибки второго рода β от размера выборки n для теста *Monobit* (МДРН(1)):
$H_1 = \mathrm{Ber}$, $\alpha = 0{,}01$, $\delta \in \{0{,}005,\ 0{,}01,\ 0{,}02\}$

Fig. 1. Type II error probability β plotted against the length n of the binary sequence for the *Monobit* test:
$H_1 = \mathrm{Ber}$, $\alpha = 0.01$, $\delta \in \{0.005,\ 0.01,\ 0.02\}$

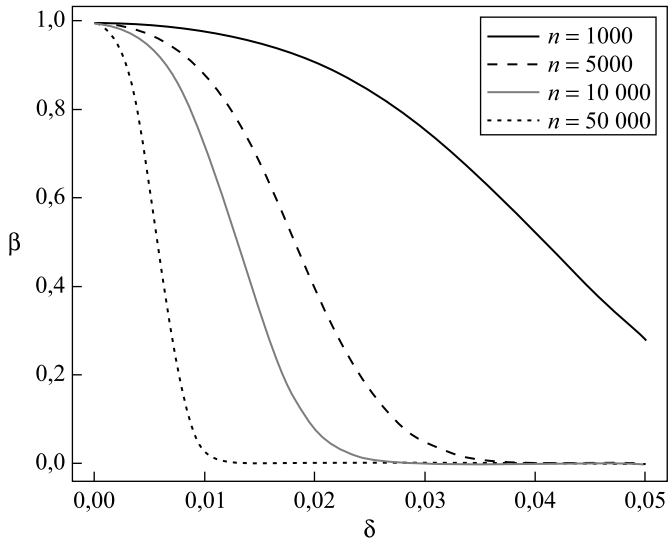

Рис. 2. Зависимость вероятности ошибки второго рода β от параметра δ для теста *Monobit* (МДРН(1)):
$H_1 = \mathrm{Ber},\ \alpha = 0{,}01,\ n \in \{1000,\ 5000,\ 10\ 000,\ 50\ 000\}$

Fig. 2. Type II error probability β plotted against the parameter δ for the *Monobit* test:
$H_1 = \mathrm{Ber},\ \alpha = 0.01,\ n \in \{1000,\ 5000,\ 10\ 000,\ 50\ 000\}$

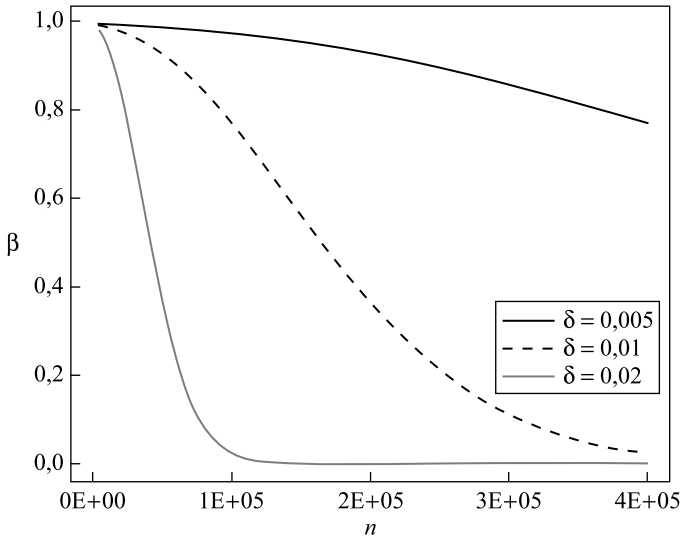

Рис. 3. Зависимость вероятности ошибки второго рода β от размера выборки *n* для теста МДРН(4):
$H_1 = \mathrm{MC}(3,\ 1),\ \alpha = 0{,}01,\ \delta \in \{0{,}005,\ 0{,}01,\ 0{,}02\}$

Fig. 3. Type II error probability β plotted against the length *n* of the binary sequence for the test of multidimensional discrete uniformity by non-overlapping 4-tuples:
$H_1 = \mathrm{MC}(3,\ 1),\ \alpha = 0.01,\ \delta \in \{0.005,\ 0.01,\ 0.02\}$

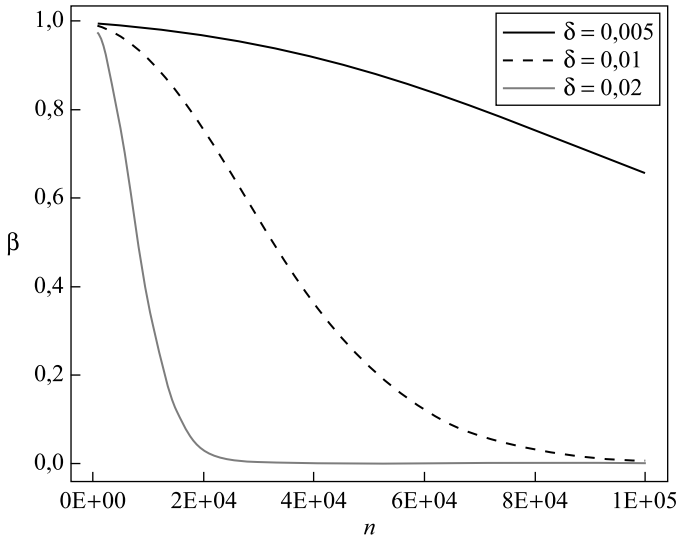

Рис. 4. Зависимость вероятности ошибки второго рода β
от размера выборки *n* для теста МДРП(4):
$H_1 = \mathrm{MC}(s, 1)$, $0 \le s \le 3$, $\alpha = 0{,}01$, $\delta \in \{0{,}005,\ 0{,}01,\ 0{,}02\}$

Fig. 4. Type II error probability β plotted against the length *n* of the binary sequence
for the test of multidimensional discrete uniformity by overlapping 4-tuples:
$H_1 = \mathrm{MC}(s, 1)$, $0 \le s \le 3$, $\alpha = 0.01$, $\delta \in \{0.005,\ 0.01,\ 0.02\}$

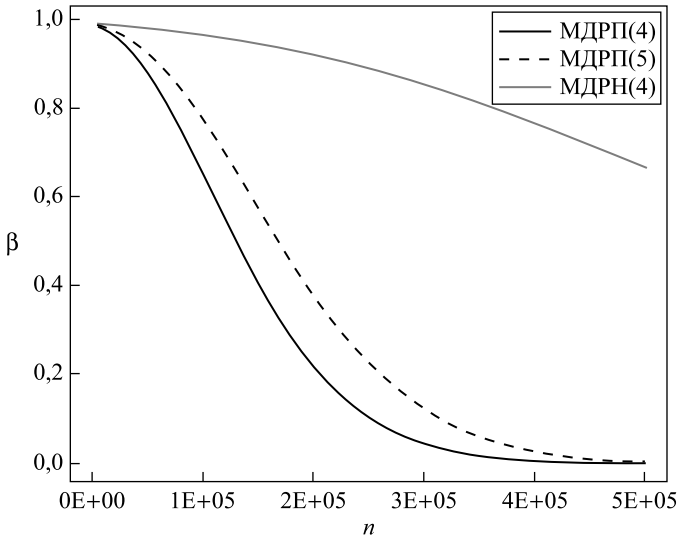

Рис. 5. Зависимость вероятности ошибки второго рода β
от размера выборки *n* для тестов МДРП(4), МДРП(5) и МДРН(4):
$H_1 = \mathrm{MC}(3, 1)$, $\alpha = 0{,}01$, $\delta = 0{,}005$

Fig. 5. Type II error probability β plotted against the length *n* of the binary sequence
for the tests of multidimensional discrete uniformity by overlapping 4-tuples (МДРП(4)),
by overlapping 5-tuples (МДРП(5)), and by non-overlapping 4-tuples (МДРН(4)):
$H_1 = \mathrm{MC}(3, 1)$, $\alpha = 0.01$, $\delta = 0.005$

Библиографические ссылки

1. Rukhin A, Soto J, Nechvatal J, Smid M, Barker E, Leigh S, et al. *A statistical test suite for random and pseudorandom number generators for cryptographic applications: NIST SP 800-22. Revision 1a* [Internet]. Gaithersburg: National Institute of Standards and Technology; 2010 [cited 2021 September 20]. 131 p. Available from: https://csrc.nist.gov/publications/detail/sp/800-22/rev-1a/final.

2. Трубей АИ, Палуха ВЮ, Пирштук ИК, Мальцев МВ, Ращеня НА. Методика тестирования случайных последовательностей на основе статистического расстояния и закона повторного логарифма. В: *Проблемы защиты информации (с грифом «Секретно»). Номер 16.* Минск: БГУ; 2020. с. 64–94.

3. Amari S, Nagaoka H. *Methods of information geometry.* Harada D, translator. Providence: American Mathematical Society; 2000. 206 p. (Translations of mathematical monographs; volume 191). Co-published by the Oxford University Press.

4. Billingsley P. Statistical methods in Markov chains. *The Annals of Mathematical Statistics.* 1961;32(1):12–40. DOI: 10.1214/aoms/1177705136.

5. Hayashi M, Watanabe S. Information geometry approach to parameter estimation in Markov chains. *The Annals of Statistics.* 2016;44(4):1495–1535. DOI: 10.1214/15-AOS1420.

6. Харин ЮС, Петлицкий АИ. Цепь Маркова *s*-го порядка с *r* частичными связями и статистические выводы о ее параметрах. *Дискретная математика.* 2007;19(2):109–130. DOI: 10.4213/dm26.

7. Волошко ВА, Вечерко ЕВ. Новые верхние границы для функции нецентрального хи-квадрат распределения. *Журнал Белорусского государственного университета. Математика. Информатика.* 2020;1:70–74. DOI: 10.33581/2520-6508-2020-1-70-74.

References

1. Rukhin A, Soto J, Nechvatal J, Smid M, Barker E, Leigh S, et al. *A statistical test suite for random and pseudorandom number generators for cryptographic applications: NIST SP 800-22. Revision 1a* [Internet]. Gaithersburg: National Institute of Standards and Technology; 2010 [cited 2021 September 20]. 131 p. Available from: https://csrc.nist.gov/publications/detail/sp/800-22/rev-1a/final.

2. Trubey AI, Palukha VYu, Pirshtuk IK, Mal'tsev MV, Rashcheny NA. [A technique for testing random sequences based on statistical distance and the law of the iterated logarithm]. In: *Problemy zashchity informatsii (s grifom «Sekretno»). Nomer 16* [Problems of information security (with the heading «Secret»). Number 16]. Minsk: Belarusian State University; 2020. p. 64–94. Russian.

3. Amari S, Nagaoka H. *Methods of information geometry.* Harada D, translator. Providence: American Mathematical Society; 2000. 206 p. (Translations of mathematical monographs; volume 191). Co-published by the Oxford University Press.

4. Billingsley P. Statistical methods in Markov chains. *The Annals of Mathematical Statistics.* 1961;32(1):12–40. DOI: 10.1214/aoms/1177705136.

5. Hayashi M, Watanabe S. Information geometry approach to parameter estimation in Markov chains. *The Annals of Statistics.* 2016;44(4):1495–1535. DOI: 10.1214/15-AOS1420.

6. Kharin YS, Petlitskii AI. [A Markov chain of order *s* with *r* partial connections and statistical inference on its parameters]. *Diskretnaya matematika.* 2007;19(2):109–130. Russian. DOI: 10.4213/dm26.

7. Voloshko VA, Vecherko EV. New upper bounds for noncentral chi-square cdf. *Journal of the Belarusian State University. Mathematics and Informatics.* 2020;1:70–74. DOI: 10.33581/2520-6508-2020-1-70-74.

Получена 18.10.2021 / исправлена 14.02.2022 / принята 14.02.2022.
Received 18.10.2021 / revised 14.02.2022 / accepted 14.02.2022.