ISSN 0002-3574 (print)
УДК 519.233.3:519.722:004.421.5

Поступила в редакцию 23.01.2017
Received 23.01.2017

В. Ю. Палуха

Белорусский государственный университет, Минск, Беларусь

СТАТИСТИЧЕСКИЕ ТЕСТЫ НА ОСНОВЕ ОЦЕНОК ЭНТРОПИИ ДЛЯ ПРОВЕРКИ ГИПОТЕЗ О РАВНОМЕРНОМ РАСПРЕДЕЛЕНИИ СЛУЧАЙНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ

Рассматривается актуальная в области защиты информации задача построения статистических тестов для проверки гипотезы о дискретном равномерном распределении («чистой случайности») выходных последовательностей криптографических генераторов. Для функционалов энтропии Шеннона, Реньи и Тсаллиса построены точечные статистические оценки на основе подстановочного принципа с использованием частотных статистик. Найдено асимптотическое распределение вероятностей полученных точечных оценок при справедливости гипотезы о «чистой случайности» в асимптотике, означающей, что количество наблюдаемых данных сравнимо с числом оцениваемых параметров. С использованием распределений вероятностей точечных оценок построены интервальные статистические оценки рассматриваемых функционалов информационной энтропии. На основе интервальных оценок разработаны решающие правила для статистической проверки гипотез о «чистой случайности» наблюдаемой дискретной последовательности. Представлены результаты компьютерных экспериментов, в которых разработанные статистические тесты применяются к выходной последовательности криптографического генератора. Выходная двоичная последовательность в этих экспериментах преобразовывалась к последовательности с алфавитом большей размерности путем объединения соседних s элементов в s-граммы.

Ключевые слова: функционалы энтропии Шеннона, Реньи и Тсаллиса, асимптотически нормальное распределение вероятностей, точечные и интервальные статистические оценки, проверка гипотез, криптографические генераторы случайных и псевдослучайных последовательностей.

U. Yu. Palukha

Belarusian State University, Minsk, Belarus

STATISTICAL TESTS BASED ON ENTROPY ESTIMATES FOR CHECKING THE HYPOTHESES OF THE UNIFORM DISTRIBUTION OF A RANDOM SEQUENCE

The actual information security problem of developing statistical tests of the hypothesis about a discrete uniform distribution (‘pure randomness’) of output sequences of cryptographic generators is considered. For the entropy functionals of Shannon, Renyi and Tsallis, the point statistical estimators based on the principle of ‘plug-in’ frequency statistics are constructed. The asymptotic probability distribution of the constructed point estimators is found when the ‘pure randomness’ hypothesis in asymptotics is valid, meaning that the number of observed data is comparable with the number of estimated parameters. With the use of the probability distributions of point estimators, the interval statistical estimators of considered information entropy functionals are constructed. On the basis of interval estimators, the decision rules for statistical testing of the hypothesis about the ‘pure randomness’ of the observed discrete sequence are developed. The results of computer experiments, in which the developed statistical tests are applied to the output sequence of cryptographic generators, are given. In these experiments, the output binary sequence was transformed to the sequence of alphabet with a larger dimension by combining the s neighboring elements in the s-grams.

Keywords: Shannon, Renyi and Tsallis entropy, asymptotically normal probability distribution, statistical estimators, hypotheses testing, cryptographic generators of random and pseudo-random sequences.

Введение. Генераторы случайных и псевдослучайных последовательностей являются одним из основных структурных элементов средств криптографической защиты информации (криптосистем). Стойкость криптосистем зависит от того, насколько близка генерируемая последовательность по своим свойствам к «чисто случайной», или равномерно распределенной случайной последовательности (РРСП) [1]. Для проверки качества криптографических генераторов (генераторов, используемых в криптосистемах) в смысле их близости по своим вероятностным свойствам к РРСП применяются статистические тесты, суть которых заключается в следующем.

Наблюдается выходная последовательность криптографического генератора и вводится гипотеза H_* о том, что последовательность является РРСП. Вычисляется некоторая статистика, распределение вероятностей которой при истинной гипотезе H_* известно. На основании значения статистики гипотеза H_* принимается либо отклоняется. В качестве тестовых статистик в настоящей статье предлагается использовать оценки функционалов информационной энтропии. Существуют различные функционалы энтропии (напр., в [2] приводятся формулы 23 функционалов), из них наиболее распространенными являются функционалы Шеннона, Реньи и Тсаллиса, для точечных статистических оценок которых в данной работе найдено асимптотическое распределение вероятностей при истинной гипотезе H_*. Полученное распределение вероятностей позволило построить и применить статистические тесты проверки гипотезы H_*, что подробно будет рассмотрено далее.

Математическая модель. Пусть на вероятностном пространстве (Ω, F, P) с множеством состояний $\Omega = \{\omega_1, \ldots, \omega_N\}$ определена случайная величина $x = x(\omega) = \omega$ с дискретным распределением вероятностей $p_k = P\{x = \omega_k\}$, $p_k \geq 0$, $\sum_{k=1}^{N} p_k = 1, k = 1, \ldots, N$. Определим функционал обобщенной энтропии согласно [2]:

$$H_{h,w}^{\varphi_1,\varphi_2}(P) = h\left(\frac{\sum_{k=1}^{N} w_k \varphi_1(p_k)}{\sum_{k=1}^{N} w_k \varphi_2(p_k)}\right), \tag{1}$$

где $w_k > 0$, $k = 1, \ldots, N$ – вес состояния ω_k, $\varphi_1 : [0,1) \to \mathbb{R}, \varphi_2 : [0,1) \to \mathbb{R}, h : \mathbb{R} \to \mathbb{R}$, – заданные функции.

В таблице приведены наиболее часто используемые [2] частные случаи функционала обобщенной энтропии (1), определяемые заданием функций $h(\cdot), \varphi_1(\cdot), \varphi_2(\cdot), \{w_k\}$, входящих в (1).

Основные функционалы энтропии

Basic entropy functionals

Тип Type	Формула Formula	$h(x)$	$\varphi_1(x)$	$\varphi_2(x)$	w_k
Энтропия Шеннона	$H(P) = -\sum_{k=1}^{N} p_k \ln p_k$	x	$-x\ln x$	x	$w \equiv 1$
Энтропия Реньи	$H_r(P) = \frac{1}{1-r}\ln\left(\sum_{k=1}^{N} p_k^r\right)$	$(1-r)^{-1}\ln x$	x^r	x	$w \equiv 1$
Энтропия Тсаллиса	$S_r(P) = \frac{1}{r-1}\left(1 - \sum_{k=1}^{N} p_k^r\right)$	$(1-r)^{-1}(x-1)$	x^r	x	$w \equiv 1$

Стоит отметить, что функционал энтропии Шеннона является предельным значением функционалов Реньи и Тсаллиса при $r \to 1$ [3] и отличается от них наличием некоторых дополнительных свойств (напр., аддитивности [1]). При истинной гипотезе H_* все три функционала достигают своего максимального значения.

Общепринятым подходом к статистическому оцениванию энтропии является построение частотных оценок вероятностей элементов алфавита и подстановка полученных оценок в функционал энтропии вместо истинных значений вероятностей. В данной статье предлагается метод построения статистических оценок энтропии Шеннона, Реньи и Тсаллиса и приводятся вероятностные свойства этих оценок в асимптотике, которая чаще встречается на практике и означает, что количество наблюдаемых данных сравнимо с числом оцениваемых параметров. С помощью полученных точечных оценок строятся интервальные статистические оценки энтропии, которые служат основой для разработки решающих правил для статистической проверки гипотез о близости наблюдаемой последовательности к РРСП.

Построение статистических оценок энтропии на основе частотных оценок вероятностей. Пусть имеется случайная последовательность $\{x_t : t = 1, \ldots, n\}$ объема n из распределения вероятностей $\{p_k\}$. Построим частотные оценки распределения вероятностей $\{p_k : k = 1, \ldots, N\}$:

$$\hat{p}_k = \frac{v_k}{n}, \quad v_k = \sum_{t=1}^{n} I\{x_t = \omega_k\} \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}, \quad I\{x_t = \omega_k\} = \begin{cases} 1, x_t = \omega_k; \\ 0, x_t \neq \omega_k. \end{cases} \tag{2}$$

Введем в рассмотрение гипотезу $H_* = \{\{x_t\}$ является РРСП$\} = \{\{x_t\}$ – независимые одинаково распределенные случайные величины, $p_k = 1/N$, $k = 1, \ldots, N\}$ и альтернативу $\overline{H_*}$.

Следуя [4], будем полагать, что имеет место схема серий. В таком случае вектор $(v_1, \ldots, v_N)^T$, составленный из частот v_k из (2), имеет полиномиальное распределение вероятностей Pol(n, N, p_1, …, p_N), а каждая из компонент распределена по биномиальному закону $Bi(n, p_k)$. Рассмотрим асимптотику:

$$n, N \to \infty, n/N \to \lambda, 0 < \lambda < \infty, \tag{3}$$

которая отличается от классической ($n \to \infty$, $N < \infty$) тем, что длительность наблюдения n и число значений N растут синхронно. В асимптотике (3) для распределения вероятностей статистик $\{v_k\}$ справедлива аппроксимация законом Пуассона $\Pi(\lambda_k)$ с параметром $\lambda_k = np_k$ [5]. При истинной гипотезе H_* все элементарные вероятности равны: $p_k = 1/N$, $k = 1, \ldots, N$, поэтому все частоты $\{v_k\}$ имеют одинаковый параметр распределения Пуассона $\lambda = n/N$.

В [4] доказана теорема об асимптотически нормальном распределении статистик, являющихся функциями от частот v_k, которую кратко можно переформулировать следующим образом. Пусть $f(\cdot): \mathbb{N}_0 \to \mathbb{R}$ – некоторая функция; $Z = \sum_{k=1}^{N} f(v_k)$, где v_k, $k = 1, \ldots, N$ – частоты с совместным полиномиальным распределением, аппроксимированные законом Пуассона в асимптотике (3). Тогда при выполнении ряда условий регулярности статистика Z имеет асимптотически нормальное распределение $\mathcal{L}\left\{\frac{Z-\mu}{\sigma}\right\} \to \mathcal{N}_1(0,1)$:

$$\mu = \sum_{k=1}^{N} E\{f(v_k)\}, \tag{4}$$

$$\sigma^2 = \sum_{k=1}^{N} D\{f(v_k)\} - \left(\sum_{k=1}^{N} \operatorname{cov}\{v_k, f(v_k)\} \right)^2 \Big/ n, \tag{5}$$

где $\mathcal{N}_1(0,1)$ – стандартный одномерный нормальный закон распределения вероятностей с нулевым математическим ожиданием и единичной дисперсией, $E\{\xi\}$ и $D\{\xi\}$ – соответственно математическое ожидание и дисперсия случайной величины ξ, $\operatorname{cov}\{\xi,\eta\}$ – ковариация случайных величин ξ и η. При истинной гипотезе H_* соотношения (4) и (5) преобразуются соответственно:

$$\mu = \sum_{k=1}^{N} E\{f(v_k)\} = NE\{f(v)\}, \tag{6}$$

$$\sigma^2 = ND\{f(v)\} - N^2 \operatorname{cov}^2\{v, f(v)\}/n = N\left(D\{f(v)\} - \operatorname{cov}^2\{v, f(v)\}/\lambda\right). \tag{7}$$

Для применения результатов из [4] к доказательству вероятностных свойств статистических оценок энтропии необходимо выразить оценки энтропийных функционалов через частоты.

Статистическая оценка энтропии Шеннона. В качестве функции f возьмем $f(v) = v \ln v$. Статистическая оценка энтропии Шеннона $\hat{H}(n, N)$ линейно выражается через Z [6]:

$$\hat{H} = \hat{H}(n, N) = -\sum_{k=1}^{N} \hat{p}_k \ln \hat{p}_k = -\sum_{k=1}^{N} \frac{v_k}{n} \ln \frac{v_k}{n} = \ln n - \frac{1}{n} Z. \tag{8}$$

В работе [6] теорема об асимптотическом распределении вероятностей статистики (8) сформулирована автором без доказательства, поэтому приведем его здесь.

Т е о р е м а 1. *В асимптотике* (3) *статистика* (8) *при истинной гипотезе* H_* *имеет асимптотически нормальное распределение* $\mathcal{L}\left\{\dfrac{\hat{H}-\mu_H}{\sigma_H}\right\}\to\mathcal{N}_1(0,1)$:

$$\mu_H=\ln n-e^{-\lambda}\sum_{k=1}^{+\infty}\frac{\ln(k+1)\lambda^k}{k!}, \tag{9}$$

$$\sigma_H^2=\frac{e^{-\lambda}}{n}\sum_{k=1}^{+\infty}\frac{(k+1)\lambda^k}{k!}\ln^2(k+1)-\frac{e^{-2\lambda}}{N}\left(\sum_{k=1}^{+\infty}\frac{\ln(k+1)\lambda^k}{k!}\right)^2-\frac{e^{-2\lambda}}{n}\left(\sum_{k=1}^{+\infty}\ln(k+1)\frac{\lambda^k}{k!}(k+1-\lambda)\right)^2. \tag{10}$$

Д о к а з а т е л ь с т в о. Сначала проверим выполнение условий теоремы 1 из [4].

1. $n, N\to\infty$, $n/N\to\lambda, 0<\lambda<\infty$ – это выполнено в силу условия теоремы.

2. $Np_k\le C<\infty, \forall N,k$. Поскольку $p_k=1/N, k=1,\ldots,N$, то $Np_k\equiv 1$.

3. $|f(v)|\le a\exp(bv)$. Положим в условии a = 1, b = 2 и рассмотрим отдельно два случая: $v=0$ и $v\ge 1$. При $v=0$ мы полагаем $0\ln 0=0$, поэтому неравенство выполняется: $0<1$. При $v\ge 1$ функция $f(v)$ неотрицательна, и $|f(v)|=f(v)=v\ln v$. Справедлива цепочка утверждений $\ln v<v\Leftrightarrow v<e^v\Rightarrow\ln v<v<e^v\Rightarrow\ln v<e^v$. Домножим обе части последнего неравенства на $v<e^v$, получим $v\ln v<e^{2v}$, что ведет к выполнению указанного условия.

4. $\lim\limits_{n\to\infty}\sup\sigma^2/n<\infty$. Справедливость выражения

$$\lim_{n\to\infty}\sup e^{-\lambda}\sum_{k=1}^{+\infty}\frac{(k+1)\lambda^k}{k!}\ln^2(k+1)-e^{-2\lambda}\lambda\left(\sum_{k=1}^{+\infty}\frac{\ln(k+1)\lambda^k}{k!}\right)^2-e^{-2\lambda}\left(\sum_{k=1}^{+\infty}(k+1-\lambda)\frac{\ln(k+1)\lambda^k}{k!}\right)^2<+\infty$$

следует из вида слагаемых, в которых сумма, имеющая порядок $O(e^{\lambda})$, умножается на $e^{-\lambda}$, и конечности λ, которая вытекает из (3).

Теперь мы можем применить для Z теорему 1 [4], осталось найти параметры нормального распределения вероятностей (6), (7). Обозначим $\mu_0=E\{v\ln v\}$. Поскольку $v\sim\Pi(\lambda)$, то

$$\mu_0=E\{v\ln v\}=\sum_{k=0}^{+\infty}k\ln k\frac{e^{-\lambda}\lambda^k}{k!}=e^{-\lambda}\sum_{k=2}^{+\infty}\ln k\frac{\lambda^k}{(k-1)!}=e^{-\lambda}\lambda\sum_{k=1}^{+\infty}\frac{\ln(k+1)\lambda^k}{k!}. \tag{11}$$

Отсюда следует выражение для математического ожидания Z:

$$\mu=NE\{v\ln v\}=N\lambda e^{-\lambda}\sum_{k=1}^{+\infty}\frac{\ln(k+1)\lambda^k}{k!}=ne^{-\lambda}\sum_{k=1}^{+\infty}\frac{\ln(k+1)\lambda^k}{k!}. \tag{12}$$

Вычислим дисперсию Z:

$$\sigma^2=N\left(E\left\{v^2\ln^2 v\right\}-\mu_0^2\right)-N^2\left(E\left\{v^2\ln v\right\}-E\{v\}\mu_0\right)^2/n. \tag{13}$$

Математическое ожидание квадрата в уменьшаемом в (13) равно

$$E\left\{v^2\ln^2 v\right\}=\sum_{k=0}^{+\infty}k^2\ln^2 k\frac{e^{-\lambda}\lambda^k}{k!}=e^{-\lambda}\sum_{k=2}^{+\infty}\frac{k\lambda^k}{(k-1)!}\ln^2 k=e^{-\lambda}\lambda\sum_{k=1}^{+\infty}\frac{(k+1)\lambda^k}{k!}\ln^2(k+1). \tag{14}$$

Отсюда с учетом (11) получаем

$$N\left(E\left\{v^2\ln^2 v\right\}-\mu_0^2\right)=N\left(e^{-\lambda}\lambda\sum_{k=1}^{+\infty}\frac{(k+1)\lambda^k}{k!}\ln^2(k+1)-e^{-2\lambda}\lambda^2\left(\sum_{k=1}^{+\infty}\frac{\ln(k+1)\lambda^k}{k!}\right)^2\right)=$$

$$=e^{-\lambda}n\sum_{k=1}^{+\infty}\frac{(k+1)\lambda^k}{k!}\ln^2(k+1)-e^{-2\lambda}n\lambda\left(\sum_{k=1}^{+\infty}\frac{\ln(k+1)\lambda^k}{k!}\right)^2. \tag{15}$$

Рассмотрим вычитаемое в (13). Математическое ожидание произведения равно

$$E\left\{v^2 \ln v\right\} = \sum_{k=0}^{+\infty} k^2 \ln k \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=2}^{+\infty} k \ln k \frac{\lambda^k}{(k-1)!} = e^{-\lambda}\lambda \sum_{k=1}^{+\infty} (k+1) \frac{\ln(k+1)\lambda^k}{k!}. \tag{16}$$

Отсюда с учетом (11) получаем

$$E\left\{v^2 \ln v\right\} - E\{v\}\mu_0 = e^{-\lambda}\lambda \sum_{k=1}^{+\infty} (k+1) \frac{\ln(k+1)\lambda^k}{k!} - e^{-\lambda}\lambda^2 \sum_{k=1}^{+\infty} \frac{\ln(k+1)\lambda^k}{k!} =$$

$$= e^{-\lambda}\lambda \sum_{k=1}^{+\infty} (k+1-\lambda) \frac{\ln(k+1)\lambda^k}{k!}.$$

Следовательно,

$$\frac{N^2}{n}\left(E\left\{v^2 \ln v\right\} - E\{v\}\mu_0\right)^2 = \frac{N^2}{n} e^{-2\lambda}\lambda^2 \left(\sum_{k=1}^{+\infty} (k+1-\lambda) \frac{\ln(k+1)\lambda^k}{k!}\right)^2 =$$

$$= e^{-2\lambda} n \left(\sum_{k=1}^{+\infty} (k+1-\lambda) \frac{\ln(k+1)\lambda^k}{k!}\right)^2. \tag{17}$$

Окончательно, подставив (15) и (17) в (13), получим

$$\sigma^2 = e^{-\lambda} n \sum_{k=1}^{+\infty} \frac{(k+1)\lambda^k}{k!} \ln^2(k+1) - e^{-2\lambda} n\lambda \left(\sum_{k=1}^{+\infty} \frac{\ln(k+1)\lambda^k}{k!}\right)^2 -$$

$$-e^{-2\lambda} n \left(\sum_{k=1}^{+\infty} (k+1-\lambda) \frac{\ln(k+1)\lambda^k}{k!}\right)^2. \tag{18}$$

Из (8) следует, что статистическая оценка энтропии Шеннона является линейным преобразованием статистики Z и, значит, также асимптотически нормально распределена. Выразим математическое ожидание и дисперсию оценки (8) через математическое ожидание и дисперсию статистики Z:

$$E\left\{\hat{H}(n, N)\right\} = E\left\{\ln n - \frac{1}{n} Z\right\} = \ln n - \frac{1}{n} E\{Z\}, \tag{19}$$

$$D\left\{\hat{H}(n, N)\right\} = D\left\{\ln n - \frac{1}{n} Z\right\} = \frac{1}{n^2} D\{Z\}. \tag{20}$$

Подставив (12) в (19), получим (9); подставив (18) в (20), получим (10). Теорема 1 доказана.

В [7] рассмотрено поведение математического ожидания оценки энтропии Шеннона двоичной последовательности, которая разбивается на фрагменты длины s, и при этом $N = 2^s$.

Знание асимптотического распределения точечной оценки (8) позволяет построить интервальную оценку энтропии Шеннона: с вероятностью $1 - \varepsilon$ оценка энтропии

$$\hat{H}(P) \in (H_-, H_+), \quad H_\pm = \mu_H \pm \sigma_H \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right), \tag{21}$$

где $\Phi^{-1}(\cdot)$ – квантиль стандартного нормального закона [5].

Недостатком полученной точечной оценки является наличие смещения, что продемонстрировано в [7]. Поэтому далее рассмотрим построение статистических оценок функционалов энтропии Реньи и Тсаллиса.

Статистические оценки энтропии Реньи и Тсаллиса. Будем рассматривать функционалы энтропии Реньи и Тсаллиса с параметром $r \in \{2, 3, \ldots\}$. Как видно из таблицы, функционалы

объединяет общая функция $\varphi_1(x) = x^r$. Аргументом функции является вероятность p_k. Видно также, что энтропии Реньи и Тсаллиса являются функциями от величины

$$P_r(P) = \sum_{k=1}^{N} p_k^r. \tag{22}$$

Следовательно, возникает задача статистического оценивания величины $P_r(P)$.

Известно [8], что статистическая оценка для (22) $\widehat{P_r}(P) = \sum_{k=1}^{N} \hat{p}_k^r = \sum_{k=1}^{N} \left(\frac{v_k}{n} \right)^r$, построенная по подстановочному принципу, является смещенной. Для получения асимптотически несмещенной оценки определим r-ю нисходящую факториальную степень x:

$$x^{\underline{r}} = x(x-1)\ldots(x-r+1) = \frac{x!}{(x-r)!} = \sum_{i=0}^{r} s(r,i)x^i, \tag{23}$$

где $s(r, i)$ – число Стирлинга первого рода [9]; по определению, при $x < r$ полагают $x^{\underline{r}} ::= 0$. В [8] предложена статистическая оценка для величины (22), которая основана на (23):

$$\widetilde{P_r}(P) = \sum_{k=1}^{N} \frac{v_k^{\underline{r}}}{n^r}, \tag{24}$$

и является асимптотически несмещенной и состоятельной [10].

Положим $f_r(v) = v^{\underline{r}}$,

$$Z_r = \sum_{k=1}^{N} f_r(v_k) = \sum_{k=1}^{N} v_k^{\underline{r}} = n^r \widetilde{P_r}(P). \tag{25}$$

Справедлива лемма [10] о распределении статистики (25).

Л е м м а. *При истинной гипотезе H_* в асимптотике* (3) *статистика* (25) *имеет асимптотически нормальное распределение* $\mathcal{L}\left\{ \frac{Z_r - \mu_r}{\sigma_r} \right\} \to \mathcal{N}_1(0,1)$:

$$\mu_r = N\lambda^r = n\lambda^{r-1},$$

$$\sigma_r^2 = N\lambda^r \left(\sum_{i=1}^{r} s(r,i) \sum_{j=0}^{i-1} C_i^j r^{i-j} \sum_{k=1}^{j} S(j,k)\lambda^k - r^2\lambda^{r-1} + r! \right) =$$

$$= n\lambda^{r-1} \left(\sum_{i=1}^{r} s(r,i) \sum_{j=0}^{i-1} C_i^j r^{i-j} \sum_{k=1}^{j} S(j,k)\lambda^k - r^2\lambda^{r-1} + r! \right),$$

где $S(j,k)$ – число Стирлинга второго рода [9].

С л е д с т в и е 1. *При $r = 2$ для параметров асимптотически нормального распределения вероятностей случайной величины Z_2 справедливы выражения*

$$\mu_2 = n\lambda, \quad \sigma_2^2 = 2n\lambda.$$

Статистические оценки энтропии Реньи и Тсаллиса выражаются через Z_r [10]:

$$\widehat{H_r}(n, N) = \frac{1}{1-r} \ln \left(\sum_{k=1}^{N} \frac{v_k^{\underline{r}}}{n^r} \right) = \ln n + \frac{1}{r-1} \left(\ln n - \ln Z_r \right), \tag{26}$$

$$\widehat{S_r}(n, N) = \frac{1}{r-1} \left(1 - \sum_{k=1}^{N} \frac{v_k^{\underline{r}}}{n^r} \right) = \frac{1}{r-1} \left(1 - \frac{Z_r}{n^r} \right). \tag{27}$$

Справедливы теоремы, доказанные автором настоящей статьи совместно с Ю. С. Хариным [10], об асимптотическом распределении вероятностей статистических оценок энтропии Реньи и Тсаллиса, которые опираются на [4, 5] и позволяют построить интервальные оценки. Приведем также следствия из теорем для наиболее употребительного на практике случая $r = 2$.

Т е о р е м а 2. *В асимптотике* (3) *статистика* (26) *является состоятельной оценкой энтропии Реньи и при истинной гипотезе* H_* *имеет асимптотически нормальное распределение*:

$$\mathcal{L}\left\{\frac{\widehat{H_r}-\mu_{H,r}}{\sigma_{H,r}}\right\}\to\mathcal{N}_1(0,1),$$

$$\mu_{H,r}=\ln N, \tag{28}$$

$$\sigma^2_{H,r}=\frac{\sum\limits_{i=1}^{r}s(r,i)\sum\limits_{j=0}^{i-1}C_i^j r^{i-j}\sum\limits_{k=1}^{j}S(j,k)\lambda^k-r^2\lambda^{r-1}+r!}{(r-1)^2 n\lambda^{r-1}}. \tag{29}$$

С л е д с т в и е 2. *При* $r = 2$ *для дисперсии асимптотического распределения вероятностей оценки* (26) *справедливо выражение*

$$\sigma^2_{H,2}=\frac{2}{n\lambda}.$$

Отметим, что при истинной гипотезе H_* $p_k = 1/N, k = 1, \ldots, N,$ поэтому значение энтропии Реньи равно $H_r(P)=\frac{1}{1-r}\ln\left(\sum\limits_{k=1}^{N}p_k^r\right)=\frac{1}{1-r}\ln\left(\sum\limits_{k=1}^{N}\frac{1}{N^r}\right)=\ln N,$ что совпадает с (28).

Знание асимптотического распределения точечной состоятельной оценки (26) позволяет построить интервальную оценку энтропии Реньи: с вероятностью 1 – ε энтропия

$$H_r(P)\in(H_-,H_+),\quad H_\pm=\mu_{H,r}\pm\sigma_{H,r}\Phi^{-1}\left(1-\frac{\varepsilon}{2}\right). \tag{30}$$

Т е о р е м а 3. *В асимптотике* (3) *статистика* (27) *является состоятельной асимптотически несмещенной оценкой энтропии Тсаллиса и при истинной гипотезе* H_* *имеет асимптотически нормальное распределение* $\mathcal{L}\left\{\frac{\widehat{S_r}-\mu_{S,r}}{\sigma_{S,r}}\right\}\to\mathcal{N}_1(0,1)$:

$$\mu_{S,r}=\frac{1}{r-1}\left(1-\frac{1}{N^{r-1}}\right), \tag{31}$$

$$\sigma^2_{S,r}=\frac{\lambda^{r-1}}{(r-1)^2 n^{2r-1}}\left(\sum_{i=1}^{r}s(r,i)\sum_{j=1}^{i-1}C_i^j r^{i-j}\sum_{k=1}^{j}S(j,k)\lambda^k-r^2\lambda^{r-1}+r!\right). \tag{32}$$

С л е д с т в и е 3. *При* $r = 2$ *для математического ожидания и дисперсии асимптотического распределения оценки* (27) *справедливы выражения*

$$\mu_{S,2}=1-\frac{1}{N},$$

$$\sigma^2_{S,2}=\frac{2}{Nn^2}.$$

Знание асимптотического распределения вероятностей точечной состоятельной оценки (27) позволяет построить интервальную оценку энтропии Тсаллиса: с вероятностью 1 – ε энтропия

$$S_r(P) \in (S_-, S_+), \;\; S_\pm = \mu_{S,r} \pm \sigma_{S,r} \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right). \tag{33}$$

Проверка гипотезы о «чистой случайности» последовательности на основе оценок энтропии. Полученные интервальные оценки (21), (30) и (33) позволяют построить решающее правило для проверки гипотез о том, является ли наблюдаемая последовательность генератора «чисто случайной», т. е. РРСП: H_* и $\overline{H_*}$. Пусть $\varepsilon \in (0,1)$ – заданный уровень значимости. Введем обозначения: $\hat{h}$ – статистическая оценка энтропии Шеннона (8), Реньи (26) или Тсаллиса (27), μ_h – асимптотическое математическое ожидание статистической оценки энтропии Шеннона (9), Реньи (28) или Тсаллиса (31), σ_h^2 – асимптотическая дисперсия статистической оценки энтропии Шеннона (10), Реньи (29) или Тсаллиса (32) при истинной гипотезе H_*. Вычислим для наблюдаемой последовательности статистику $\hat{h}$. Решающее правило, основанное на статистике $\hat{h}$, имеет вид

$$\begin{cases} H_*, & \text{если } t_- < \hat{h} < t_+; \\ \overline{H_*}, & \text{в противном случае,} \end{cases} \qquad t_\pm = \mu_h \pm \sigma_h \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right). \tag{34}$$

В случае принятия решения о справедливости гипотезы H_* можно сделать вывод о том, что на уровне значимости ε исследуемый процесс по своим энтропийным свойствам неотличим от «чисто случайной» последовательности на основе наблюдаемой реализации длиной не более n.

Результаты компьютерных экспериментов. Разработанное решающее правило (34) применено для анализа выходной двоичной последовательности реального физического генератора двоичной случайной последовательности [11] $\{y_\tau\}, \tau = 1, \ldots, T,$ длиной $T = 125 \cdot 2^{25}$ бит. Выходная последовательность «нарезалась» на непересекающиеся подряд идущие фрагменты длины s (s-граммы): $X^{(t)} = (X_j^{(t)}) = (y_{(t-1)s+1}, \ldots, y_{ts}) \in \{0,1\}^s, \;\; t = 1, \ldots, n = [T/s]$. Из полученных s-грамм формировалась новая последовательность $\{x_t\}$ из алфавита мощности $N = 2^s$ по правилу $x_t = \sum_{j=1}^{s} 2^{j-1} X_j^{(t)} + 1$.

На рис. 1 представлены значения отклонений оценки энтропии Шеннона (8) от математического ожидания (9), деленных на границы доверительных интервалов: $\dfrac{\widehat{H} - \mu_H}{\sigma_H \Phi^{-1}(1-\varepsilon/2)}$, на уровне значимости ε = 0,05 в зависимости от $s \in \{5, \ldots, 24\}$. Выход за полосу (–1; 1) означает непопада-

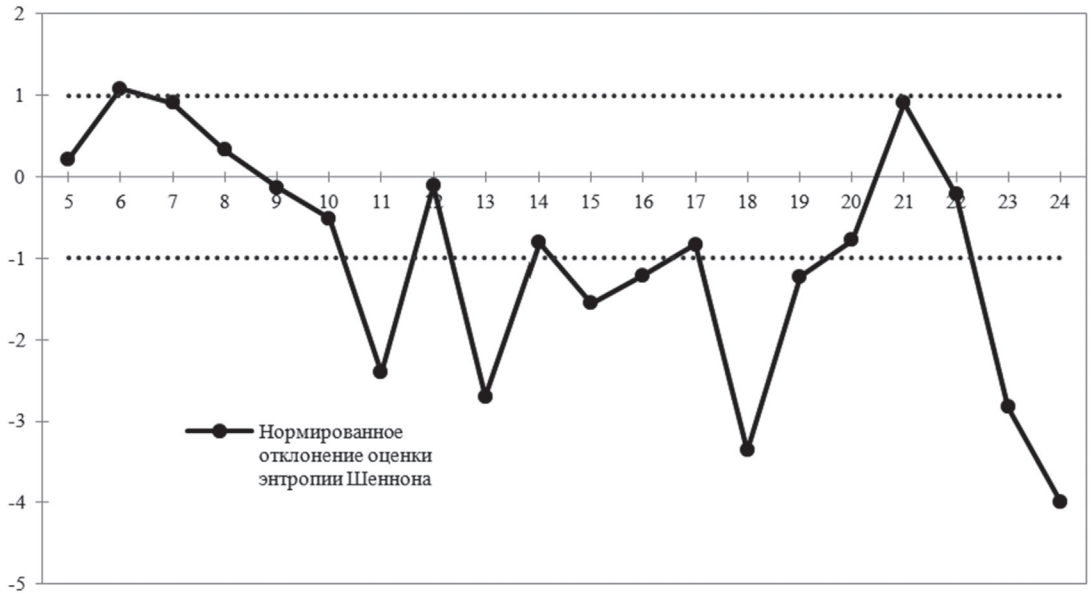

Рис. 1. Отклонение оценки энтропии Шеннона от математического ожидания для $s \in \{5, \ldots, 24\}$

Fig. 1. Deviation of the Shannon entropy estimate from expectation for $s \in \{5, \ldots, 24\}$

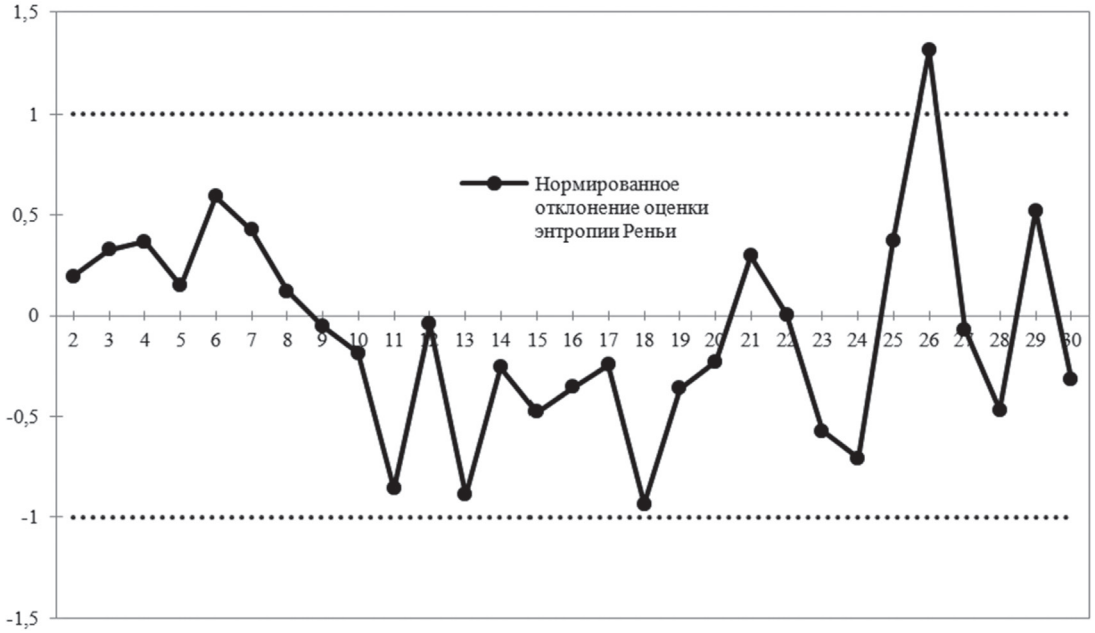

Рис. 2. Отклонение оценки энтропии Реньи от математического ожидания для $s \in \{2, \ldots, 30\}$

Fig. 2. Deviation of the Renyi entropy estimate from expectation for $s \in \{2, \ldots, 30\}$

ние в доверительный интервал и отклонение гипотезы H_*. Как видно, на многих значениях порядка s гипотеза H_* отклоняется, что свидетельствует о том, что выходная последовательность генератора отличается от РРСП.

На рис. 2 представлены значения отклонений оценки энтропии Реньи (26) при $r = 2$ от математического ожидания (28), деленных на границы доверительных интервалов: $\dfrac{\widehat{H_r} - \mu_{H,r}}{\sigma_{H,r}\Phi^{-1}(1-\varepsilon/2)}$, на уровне значимости $\varepsilon = 0{,}1$ в зависимости от $s \in \{2, \ldots, 30\}$, и доверительная полоса $(-1; 1)$. Как видно, при значениях $s \leq 25$ выходная последовательность генератора согласуется с моделью РРСП.

Вычисление оценки энтропии Тсаллиса и применение решающего правила (34) на ее основе дает аналогичный энтропии Реньи результат. Построенный график практически не отличается от графика, представленного на рис. 2, поэтому в данной статье не приводится. Такое поведение оценок энтропии Реньи и Тсаллиса можно объяснить наличием зависимости от одной и той же величины (24).

В то же время решающие правила на основе оценок энтропии Шеннона и Реньи дали различные результаты. Это означает, что данные тесты необходимо применять в комплексе, так как один из них может выявить отклонения от РРСП, которые не выявил другой, и наоборот.

Заключение. Построены асимптотически нормально распределенные статистические оценки функционалов энтропии Шеннона, Реньи и Тсаллиса. Получены явные формулы для моментов построенных статистических оценок. Сформулировано решающее правило, основанное на этих оценках, для проверки гипотезы о том, является ли наблюдаемая последовательность равномерно распределенной случайной последовательностью. Проведены компьютерные эксперименты, иллюстрирующие свойства построенных статистических оценок и решающих правил.

Список использованных источников

1. Криптология / Ю. С. Харин [и др.]. – Минск: БГУ, 2013. – 512 с.

2. Esteban, M. D. A summary on entropy statistics / M. D. Esteban, D. Morales // Kybernetika. – 1995. – Vol. 31, № 4. – P. 337–346.

3. Bromiley, P. A. Shannon Entropy, Renyi Entropy, and Information [Electronic resource] / P. A. Bromiley, N. A. Thacker, E. Bouhova-Thacker. – Mode of access: http://www.tina-vision.net/docs/memos/2004-004.pdf. – Date of access: 08.04.2016.

4. Holst, L. Asymptotic normality and efficiency for certain goodness-of-fit tests / L. Holst // Biometrika. – 1972. – Vol. 59, № 1. – P. 137–145.

5. Харин, Ю. С. Теория вероятностей, математическая и прикладная статистика / Ю. С. Харин, Н. М. Зуев, Е. Е. Жук. – Минск: БГУ, 2011. – 463 с.

6. Палуха, В. Ю. Вероятностные свойства статистической оценки многомерной энтропии в задачах защиты информации / В. Ю. Палуха // XVII Респ. науч.-практ. конф. молодых ученых: сб. материалов, Брест, 15 мая 2015 г.: в 2 ч. – Брест: БрГУ, 2015. – Ч. 1. – С. 57–59.

7. Палуха, В. Ю. Энтропийные характеристики двоичных последовательностей в криптографии / В. Ю. Палуха, Ю. С. Харин // Комплексная защита информации: материалы XX науч.-практ. конф., Минск, 19–21 мая 2015 г. – Минск: РИВШ, 2015. – С. 99–102.

8. Estimating Renyi Entropy of Discrete Distributions [Electronic resource] / J. Acharya, [et al.]. – Mode of access: http://arxiv.org/pdf/1408.1000v3.pdf. – Date of access: 08.04.2016.

9. Энвин, А. Ю. Дискретная математика / А. Ю. Энвин. – Челябинск: Изд-во ЮУрГУ, 1998. – 176 с.

10. Харин, Ю. С. Статистические оценки энтропии Реньи и Тсаллиса и их использование для проверки гипотез о «чистой случайности» / Ю. С. Харин, В. Ю. Палуха // Вес. Нац. акад. навук Беларусі. Сер. фіз.-мат. навук. – 2016. – № 2. – С. 37–47.

11. Speedtest-500MB.bin [Electronic resource] // Humboldt Berlin University, Faculty of Mathematics and Natural Sciences, Department of Physics. – Mode of access: http://qrng.physik.hu-berlin.de/files/speedtest-500MB.bin. – Date of access: 08.04.2016.

References

1. Kharin Yu.S., Agievich S.V., Vasil'ev D.V., Matveev G.V. *Cryptology.* Minsk, Belarusian State University, 2013. 512 p. (In Russian).

2. Esteban M.D., Morales D.A summary on entropy statistics. *Kybernetika*, 1995, vol. 31, no. 4, pp. 337–346.

3. Bromiley P.A., Thacker N.A., Bouhova-Thacker E. *Shannon Entropy, Renyi Entropy, and Information*. Available at: http://www.tina-vision.net/docs/memos/2004-004.pdf. (accessed 8 April 2016).

4. Holst L. Asymptotic normality and efficiency for certain goodness-of-fit tests. *Biometrika*, 1972, vol. 59, no. 1, pp. 137–145. Doi: 10.2307/2334624

5. Kharin Yu.S., Zuev N.M., Zhuk E.E. *Probability Theory, Mathematical and Applied Statistics*. Minsk, Belarusian State University, 2011. 463 p. (In Russian).

6. Palukha V.Yu. The probability properties of the multivariate entropy estimator in information security tasks. *XVII Respublikanskaya nauchno-prakticheskaya konferentsiya molodykh uchenykh. Sbornik materialov. Ch. 1* [Proceedings of XVII Republican Young Scientists Scientific and Practical Conference. Part 1]. Brest, Brest State University, 2015, pp. 57–59. (In Russian).

7. Palukha V.Yu., Kharin Yu.S. Entropy characteristics of binary sequences in cryptography. *Kompleksnaya zashchita informatsii: materialy XX nauchno-prakticheskoi konferentsii* [Complex Information Protection. Proceedings of XX Scientific and Practical Conference]. Minsk, Republican Institute for Higher Education, 2015, pp. 99–102. (In Russian).

8. Acharya J., Orlitsky A., Suresh A.T., Tyagi H. *Estimating Renyi Entropy of Discrete Distributions*. Available at http://arxiv.org/pdf/1408.1000v3.pdf. (accessed 8 April 2016).

9. Envin A.Yu. *Discrete Mathematics*. Cheliabinsk, South Ural State University, 1998. 176 p. (In Russian).

10. Kharin Yu.S., Palukha U.Yu. Statistical estimates of Rényi and Tsallis entropy and their use for testing the 'pure randomness hypotheses. *Vestsi Natsyyanalnai akademii navuk Belarusi. Seryya fizika-matematychnykh navuk* [Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics Series], 2016, no. 2, pp. 37–47. (In Russian).

11. Speedtest-500MB.bin. *Humboldt Berlin University, Faculty of Mathematics and Natural Sciences, Department of Physics*. Available at http://qrng.physik.hu-berlin.de/files/speedtest-500MB.bin. (accessed 8 April 2016).

Информация об авторе

Палуха Владимир Юрьевич – аспирант кафедры математического моделирования и анализа данных факультета прикладной математики и информатики, Белорусский государственный университет; младший научный сотрудник НИИ прикладных проблем математики и информатики БГУ (пр. Независимости, 4, 220030, г. Минск, Республика Беларусь). E-mail: palukha@bsu.by

Information about the author

Palukha Uladzimir Yur'evich – Postgraduate of the Department of Mathematical Modeling and Data Analysis, Faculty of Applied Mathematics and Computer Science, Belarusian State University; Junior Researcher of the Research Institute for Applied Problems of Mathematics and Informatics; (4, Nezavisimosti Ave., 220030, Minsk, Republic of Belarus). E-mail: palukha@bsu.by

Для цитирования

Палуха, В. Ю. Статистические тесты на основе оценок энтропии для проверки гипотез о равномерном распределении случайной последовательности // Вес. Нац. акад навук Беларусі. Сер фіз.-мат. навук. – 2017. – № 1. – С. 79–88.

For citation

Palukha U.Yu. Statistical tests based on entropy estimates for checking the hypotheses of the uniform distribution of a random sequence. *Vestsi Natsyianal'nai akademii navuk Belarusi. Seryia fizika-matematychnykh navuk* [Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics series], 2017, no. 1, pp. 79–88. (In Russian).