УДК 519.2

Ю. С. ХАРИН, В. Ю. ПАЛУХА

СТАТИСТИЧЕСКИЕ ОЦЕНКИ ЭНТРОПИИ РЕНЬИ И ТСАЛЛИСА И ИХ ИСПОЛЬЗОВАНИЕ ДЛЯ ПРОВЕРКИ ГИПОТЕЗ О «ЧИСТОЙ СЛУЧАЙНОСТИ»

НИИ прикладных проблем математики и информатики
Белорусского государственного университета, Минск, Беларусь,
kharin@bsu.by, palukha@bsu.by

Предложен подход к построению состоятельных статистических оценок функционалов энтропии Реньи и Тсаллиса. Найдено асимптотическое распределение вероятностей построенных точечных оценок, построены интервальные оценки. На основе интервальных оценок разработано решающее правило для статистической проверки гипотез о «чистой случайности» наблюдаемой дискретной последовательности. Представлены результаты компьютерных экспериментов.

Ключевые слова: функционалы энтропии Реньи и Тсаллиса, асимптотически нормальное распределение вероятностей, точечные и интервальные статистические оценки, проверка гипотез.

Yu. S. KHARIN, U. Yu. PALUKHA

STATISTICAL ESTIMATORS OF RENYI AND TSALLIS ENTROPY AND THEIR USE FOR TESTING THE HYPOTHESES OF "PURE RANDOMNESS"

Research Institute for Applied Mathematics and Informatics of Belarusian State University, Minsk, Belarus,
kharin@bsu.by; palukha@bsu.by

An approach to the construction of consistent statistical estimators for Renyi and Tsallis entropy is considered. The asymptotic probability distribution of constructed point estimators is proved, and the interval estimators are constructed. On the basis of interval estimators the decision rule for the statistical testing of the hypotheses of "pure randomness" of the observed discrete sequence is developed. The results of computer experiments are presented.

Keywords: Renyi and Tsallis entropy, asymptotically normal probability distribution, statistical estimators, testing of hypotheses.

Введение. В теории информации наряду с энтропией Шеннона известны и другие функционалы информационной энтропии, которые оказываются полезными в ряде прикладных задач [1]. В криптологии и других приложениях часто возникает задача статистического оценивания энтропии, при этом необходимо знание вероятностных свойств построенных оценок. Общепринятым подходом к статистическому оцениванию энтропии является построение частотных оценок вероятностей элементов алфавита и подстановка полученных оценок в функционал энтропии вместо истинных значений вероятностей.

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 [1]:

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

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 [1] 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

Тип	Формула	$h(x)$	$\varphi_1(x)$	$\varphi_2(x)$	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	$w \equiv 1$

Задача построения статистической оценки энтропии Шеннона $H(p)$ рассмотрена в работах [2, 3]. В [2] изучено поведение математического ожидания оценки энтропии при различных способах построения частотных оценок вероятностей $\{p_k\}$. В [3] рассмотрено применение статистической оценки энтропии Шеннона для проверки гипотез о вероятностных свойствах наблюдаемой двоичной последовательности.

В последнее время внимание исследователей привлекают также функционалы энтропии Реньи $H_r(p)$ и Тсаллиса $S_r(p)$, изучаются методы построения статистических оценок этих функционалов. В [4] найдена оптимальная длина последовательности для построения оценки энтропии Реньи в зависимости от N и r, статья [5] посвящена построению оценок энтропии Шеннона, Реньи и Тсаллиса, сбалансированных по принципу «смещение – дисперсия». Стоит отметить, что функционал энтропии Шеннона является предельным значением функционалов Реньи и Тсаллиса при $r \to 1$ [5] и отличается от них наличием некоторых дополнительных свойств (например, аддитивности [6]).

В данной статье предлагается метод построения статистических оценок энтропии Реньи и Тсаллиса и доказываются вероятностные свойства полученных оценок. Построенные по реализации случайной последовательности оценки предлагается применять для статистической проверки гипотез о близости наблюдаемой последовательности к «чисто случайной» последовательности (т. е. равномерно распределенной случайной последовательности, далее – РРСП), что является актуальной задачей в приложениях, связанных с защитой информации, анализом генетических последовательностей [6].

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 $\{p_k : k = 1, \ldots, N\}$:

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

Введем в рассмотрение гипотезу H_* = {{x_t} является РРСП} = {{x_t} – н.о.р.с.в., $p_k = 1/N$, $k = 1, \ldots, N$} и альтернативу $\overline{H_*}$.

Следуя [7], 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 $\lambda_k = np_k$ [8]. 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 $\lambda = n/N$.

Рассмотрим подробнее 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 – функции от величины

$$P_r(P) = \sum_{k=1}^{N} p_k^r. \tag{4}$$

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

Известно [4], что статистическая оценка для (4) по подстановочному принципу $\widehat{P}_r(P) = \sum_{k=1}^{N} \hat{p}_k^r = \sum_{k=1}^{N} \left(\frac{v_k}{n} \right)^r$ является смещенной. Для построения асимптотически несмещенной оценки определим r-ю нисходящую факториальную степень x:

$$x^{\underline{r}} = x(x-1)\ldots(x-r+1) = \frac{x!}{(x-r)!} = \sum_{i=0}^{r} s(r,i) x^i, \tag{5}$$

где $s(r, i)$ – число Стирлинга первого рода [9]; по определению, при $x < r$ полагают $x^{\underline{r}} ::= 0$. В [4] предложена статистическая оценка для величины (4), которая основана на (5):

$$\widetilde{P}_r(P) = \sum_{k=1}^{N} \frac{v_k^{\underline{r}}}{n^r}. \tag{6}$$

Согласно [4], эта оценка (6) в асимптотике (3) удовлетворяет следующим асимптотическим свойствам:

$$E\{\widetilde{P}_r(P)\} \to P_r(P), \tag{7}$$

$$D\{\widetilde{P}_r(P)\} \to 0, \tag{8}$$

которые влекут асимптотическую несмещенность и состоятельность оценки (6) [10].

Положим

$$f_r(v) = v^{\underline{r}}, \tag{9}$$

$$Z_{n,r} = \sum_{k=1}^{N} f_r(v_k) = \sum_{k=1}^{N} v_k^{\underline{r}} = n^r \widetilde{P}_r(P). \tag{10}$$

Из [7] следует, что в асимптотике (3) при выполнении некоторых условий регулярности, которые мы сформулируем и проверим в дальнейшем, статистика (10) имеет асимптотически нормальное распределение:

$$\mathcal{L}\left\{ \frac{Z_{n,r} - \mu_{n,r}}{\sigma_{n,r}} \right\} \to \mathcal{N}_1(0,1),$$
$$\mu_{n,r} = \sum_{k=1}^{N} E\{v_k^{\underline{r}}\}, \tag{11}$$

$$\sigma_{n,r}^2 = \sum_{k=1}^{N} D\{v_k^r\} - \left(\sum_{k=1}^{N} \mathrm{cov}\{v_k, v_k^r\} \right)^2 \Big/ n, \tag{12}$$

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, $\mathrm{cov}\{\xi, \eta\}$ – ковариация случайных величин ξ и η.

Для нахождения параметров распределения вероятностей статистики (10) необходимо вычислить моменты $E\{v_k^r\}$ и $D\{v_k^r\}$. Для удобства опустим индекс *k* у величины v_k. Полагая, что случайная величина *v* распределена по закону Пуассона с параметром λ, т. е. $\mathcal{L}\{v\} = \Pi(\lambda)$, получим, согласно [4],

$$E\{v^r\} = \lambda^r. \tag{13}$$

Кроме того, согласно [11],

$$E\{v^r\} = \sum_{i=0}^{r} S(r,i)\lambda^i, \tag{14}$$

где $S(r, i)$ – число Стирлинга второго рода [9], а сумма в правой части этого равенства является многочленом Тоучарда. Приведем несколько свойств чисел Стирлинга, которые нам понадобятся в дальнейшем:

$$S(r,0) = s(r,0) = 0, \forall r \in \mathbb{N}; \quad S(0,0) = s(0,0) = 1; \tag{15}$$

$$S(r,1) = S(r,r) = s(r,r) = 1, \forall r \in \mathbb{N}. \tag{16}$$

Л е м м а 1. *Пусть* $\mathcal{L}\{v\} = \Pi(\lambda)$, *тогда*

$$D\{v^r\} = \lambda^r \left(E\{(v+r)^r\} - \lambda^r \right) = \lambda^r E\{(v+r)^r - v^r\}. \tag{17}$$

Д о к а з а т е л ь с т в о. Для математического ожидания квадрата статистики (10) справедлива формула [4]:

$$E\{(v^r)^2\} = \lambda^r E\{(v+r)^r\}. \tag{18}$$

Из свойства дисперсии $D\{v^r\} = E\{(v^r)^2\} - E^2\{v^r\}$, (13) и (18) вытекает

$$D\{v^r\} = \lambda^r E\{(v+r)^r\} - \lambda^{2r} = \lambda^r \left(E\{(v+r)^r\} - \lambda^r \right) = \lambda^r E\{(v+r)^r - v^r\},$$

что завершает доказательство леммы 1.

Л е м м а 2. *Пусть* $\mathcal{L}\{v\} = \Pi(\lambda)$, *тогда*

$$\mathrm{cov}\{v, v^r\} = r\lambda^r. \tag{19}$$

Д о к а з а т е л ь с т в о. Для математического ожидания произведения *v* и v^r справедливо

$$E\{vv^r\} = \sum_{i=0}^{+\infty} \frac{e^{-\lambda}\lambda^i}{i!} i i^r = \sum_{i=0}^{+\infty} \frac{e^{-\lambda}\lambda^i}{i!} i \frac{i!}{(i-r)!} = \sum_{i=r}^{+\infty} \frac{e^{-\lambda}\lambda^i}{(i-r)!} i = \lambda^r \sum_{i=0}^{+\infty} \frac{e^{-\lambda}\lambda^i}{i!} (i+r) =$$

$$= \lambda^r \left(\sum_{i=0}^{+\infty} \frac{e^{-\lambda}\lambda^i}{i!} i + r \sum_{i=0}^{+\infty} \frac{e^{-\lambda}\lambda^i}{i!} \right) = \lambda^r (\lambda + r). \tag{20}$$

Из (13), (20), свойств ковариации и распределения Пуассона имеем

$$\operatorname{cov}\{v, v^{\underline{r}}\} = E\{vv^{\underline{r}}\} - E\{v\}E\{v^{\underline{r}}\} = \lambda^r(\lambda + r) - \lambda\lambda^r = r\lambda^r,$$

что и требовалось доказать.

Вернемся к рассмотрению вероятностных свойств статистики (10). При истинной гипотезе H_* соотношения (11) и (12) преобразуются соответственно в

$$\mu_{n,r} = \sum_{k=1}^{N} E\{v_k^{\underline{r}}\} = NE\{v^{\underline{r}}\}, \tag{21}$$

$$\sigma_{n,r}^2 = ND\{v^{\underline{r}}\} - N^2 \operatorname{cov}^2\{v, v^{\underline{r}}\}/n = N\left(D\{v^{\underline{r}}\} - \operatorname{cov}^2\{v, v^{\underline{r}}\}/\lambda\right). \tag{22}$$

Т е о р е м а 1. *При истинной гипотезе* H_* *в асимптотике* (3) *статистика* (10) *имеет асимптотически нормальное распределение*:

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

$$\mu_{n,r} = N\lambda^r = n\lambda^{r-1}, \tag{23}$$

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

$$= n\lambda^{r-1}\left(\sum_{i=1}^{r} s(r,i)\sum_{j=0}^{i-1} C_i^j r^{i-j}\sum_{k=1}^{j} S(j,k)\lambda^k - r^2\lambda^{r-1} + r!\right). \tag{24}$$

Д о к а з а т е л ь с т в о. Сначала проверим выполнение условий теоремы 1 из [7].

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. $|f(v)| \le a\exp(bv)$. Поскольку (9) неотрицательна, то $|f(v)| = f(v) = v^{\underline{r}}$. Положим в условии $a = 1$, $b = r$ и рассмотрим отдельно два случая: $v = 0$ и $v \ge 1$. При $v = 0$ неравенство выполняется: $0 < 1$. Для $v \ge 1$ справедлива цепочка утверждений: $\ln v < v \Rightarrow r\ln v < rv \Leftrightarrow \ln v^r < rv \Rightarrow v^r < e^{rv}$. Поскольку $v^{\underline{r}} \le v^r$, то получаем $v^{\underline{r}} < e^{rv}$, что ведет к выполнению указанного условия.

4. $\limsup\limits_{n\to\infty} \sigma_{n,r}^2/n < \infty$. Из второй формулы в (24) следует, что $\sigma_{n,r}^2/n$ является многочленом от λ степени $2r - 3$, и значение этой величины конечно в силу того, что $\lambda < \infty$.

Подставив (13) в (21), получим (23). Для нахождения асимптотической дисперсии (24) проведем вспомогательные преобразования, воспользовавшись (13) и (14):

$$E\{(v+r)^{\underline{r}}\} - \lambda^r = E\{(v+r)^{\underline{r}} - v^{\underline{r}}\} = E\left\{\sum_{i=1}^{r} s(r,i)(v+r)^i - \sum_{i=1}^{r} s(r,i)v^i\right\} =$$

$$= \sum_{i=1}^{r} s(r,i)E\left\{(v+r)^i - v^i\right\} = \sum_{i=1}^{r} s(r,i)E\left\{\sum_{j=0}^{i} C_i^j v^j r^{i-j} - v^i\right\} =$$

$$= \sum_{i=1}^{r} s(r,i)\sum_{j=0}^{i-1} C_i^j r^{i-j} E\{v^j\} = \sum_{i=1}^{r} s(r,i)\sum_{j=0}^{i-1} C_i^j r^{i-j}\sum_{k=0}^{j} S(j,k)\lambda^k. \tag{25}$$

Применив (5) и (15), получим

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

$$= \sum_{i=1}^{r} s(r,i)\sum_{j=1}^{i-1} C_i^j r^{i-j}\sum_{k=1}^{j} S(j,k)\lambda^k + r^{\underline{r}} = \sum_{i=1}^{r} s(r,i)\sum_{j=1}^{i-1} C_i^j r^{i-j}\sum_{k=1}^{j} S(j,k)\lambda^k + r!. \tag{26}$$

Подставив (17) и (19) в (22) с учетом (25) и (26), приходим к (24):

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

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

что завершает доказательство теоремы 1.

Заметим, что в (25) имеется единственный одночлен λ^{r-1} при $i = r, j = i - 1, k = j$. Согласно (16), коэффициент при этом одночлене равен $s(r,r)C_r^{r-1}r^{r-r+1}S(r-1,r-1) = r^2$. Поэтому (24) можно представить в эквивалентном виде

$$\sigma_{n,r}^2 = N\lambda^r\left(\sum_{i=1}^{r-1} s(r,i)\sum_{j=1}^{i-1} C_i^j r^{i-j}\sum_{k=1}^{j} S(j,k)\lambda^k + \sum_{j=1}^{r-2} C_r^j r^{r-j}\sum_{k=1}^{j} S(j,k)\lambda^k + r^2\sum_{k=1}^{r-2} S(r-1,k)\lambda^k + r!\right). \quad (27)$$

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

$$\mu_{n,2} = n\lambda, \quad (28)$$

$$\sigma_{n,2}^2 = 2n\lambda. \quad (29)$$

Д о к а з а т е л ь с т в о. Справедливость (28) очевидно вытекает из (23) при $r = 2$. Вычислим дисперсию при $r = 2$, опираясь на (27):

$$\sigma_{n,2}^2 = N\lambda^2 \cdot 2! = 2n\lambda,$$

что и требовалось доказать.

Согласно таблице и (10), статистические оценки энтропии Реньи и Тсаллиса выражаются через $Z_{n,r}$, о чем свидетельствует следующая лемма.

Л е м м а 3. *Статистические оценки энтропии Реньи и Тсаллиса, построенные с использованием оценки* (6)*, выражаются через статистику* (10):

$$\widehat{H_r}(n, N) = \frac{1}{1-r}\ln\left(\sum_{k=1}^{N}\frac{\nu_k^r}{n^r}\right) = \ln n + \frac{1}{r-1}\left(\ln n - \ln Z_{n,r}\right), \quad (30)$$

$$\widehat{S_r}(n, N) = \frac{1}{r-1}\left(1 - \sum_{k=1}^{N}\frac{\nu_k^r}{n^r}\right) = \frac{1}{r-1}\left(1 - \frac{Z_{n,r}}{n^r}\right). \quad (31)$$

Д о к а з а т е л ь с т в о. В силу (10) и таблицы для оценки энтропии Реньи справедливо представление

$$\widehat{H_r}(n, N) = \frac{1}{1-r}\ln\left(\sum_{k=1}^{N}\frac{\nu_k^r}{n^r}\right) = \frac{1}{1-r}\left(\ln\frac{1}{n^r} + \ln Z_{n,r}\right) = \frac{1}{r-1}\left(r\ln n - \ln Z_{n,r}\right) = \ln n + \frac{1}{r-1}\left(\ln n - \ln Z_{n,r}\right).$$

Для оценки энтропии Тсаллиса аналогично имеем:

$$\widehat{S_r}(n, N) = \frac{1}{r-1}\left(1 - \sum_{k=1}^{N}\frac{\nu_k^r}{n^r}\right) = \frac{1}{r-1}\left(1 - \frac{1}{n^r}\sum_{k=1}^{N}\nu_k^r\right) = \frac{1}{r-1}\left(1 - \frac{Z_{n,r}}{n^r}\right).$$

Лемма 3 доказана.

Статистическая оценка энтропии Тсаллиса и ее свойства. Справедлива теорема об асимптотическом распределении вероятностей статистической оценки энтропии Тсаллиса.

Т е о р е м а 2. *В асимптотике* (3) *статистика* (31) *является состоятельной асимптотически несмещенной оценкой энтропии Тсаллиса и при истинной гипотезе* H_* *имеет асимптотически нормальное распределение*:

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

$$\mu_{S,r}=\frac{1}{r-1}\left(1-\frac{1}{N^{r-1}}\right), \tag{32}$$

$$\sigma^2_{S,r}=\frac{\lambda^{r-1}}{(r-1)^2n^{2r-1}}\left(\sum_{i=1}^{r}s(r,i)\sum_{j=1}^{i-1}C_i^j r^{i-j}\sum_{k=1}^{j}S(j,k)\lambda^k-r^2\lambda^{r-1}+r!\right). \tag{33}$$

Д о к а з а т е л ь с т в о. Асимптотическая несмещенность и состоятельность оценки (31) следует из (4), (7), (8), (10), (31) и таблицы.

Далее, как видно из формулы (31), статистическая оценка энтропии Тсаллиса является линейным преобразованием статистики (10), поэтому с учетом теоремы о линейном преобразовании нормально распределенной случайной величины [8] она также имеет асимптотически нормальное распределение. Для математического ожидания оценки (31) с учетом (3), (23) справедливо равенство

$$E\left\{\widehat{S_r}\right\}=E\left\{\frac{1}{r-1}\left(1-\frac{Z_{n,r}}{n^r}\right)\right\}=\frac{1}{r-1}\left(1-\frac{E\left\{Z_{n,r}\right\}}{n^r}\right)=\frac{1}{r-1}\left(1-\frac{N\lambda^r}{n^r}\right)=\frac{1}{r-1}\left(1-\frac{1}{N^{r-1}}\right).$$

Для дисперсии оценки (31) имеем

$$D\left\{\widehat{S_r}\right\}=D\left\{\frac{1}{r-1}\left(1-\frac{Z_{n,r}}{n^r}\right)\right\}=D\left\{\frac{Z_{n,r}}{(r-1)n^r}\right\}=\frac{D\{Z_{n,r}\}}{(r-1)^2n^{2r}}, \tag{34}$$

откуда с учетом (24) получаем (33). Теорема 2 доказана.

С л е д с т в и е 2. *При* $r=2$ *для математического ожидания и дисперсии асимптотического распределения оценки* (31) *справедливы выражения*:

$$\mu_{S,2}=1-\frac{1}{N}, \tag{35}$$

$$\sigma^2_{S,2}=\frac{2}{Nn^2}. \tag{36}$$

Д о к а з а т е л ь с т в о. Соотношение (35) следует из (32). Подставим (29) в (34), получим

$$D\left\{\widehat{S_2}\right\}=\frac{2n\lambda}{n^4}=\frac{2\lambda}{n^3}=\frac{2}{Nn^2},$$

что совпадает с (36).

Знание асимптотического распределения вероятностей точечной состоятельной оценки (31) позволяет построить интервальную оценку энтропии Тсаллиса:

$$\text{с вероятностью } 1-\varepsilon \text{ энтропия } S_r(P)\in(S_-,S_+),\ \ S_\pm=\mu_{S,r}\pm\sigma_{S,r}\Phi^{-1}\left(1-\frac{\varepsilon}{2}\right),$$

где $\Phi^{-1}(\cdot)$ – квантиль стандартного нормального закона [8].

Статистическая оценка энтропии Реньи и ее свойства. Справедлива следующая теорема об асимптотическом распределении вероятностей статистической оценки энтропии Реньи.

Т е о р е м а 3. *В асимптотике* (3) *статистика* (30) *является состоятельной оценкой энтропии Тсаллиса и при истинной гипотезе* H_* *имеет асимптотически нормальное распределение*:

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

$$\mu_{H,r} = \ln N, \tag{37}$$

$$\sigma^2_{H,r} = \frac{\sigma^2_{n,r}}{(r-1)^2 n^2 \lambda^{2r-2}}, \tag{38}$$

$\sigma^2_{n,r}$ – *дисперсия величины* (8).

Д о к а з а т е л ь с т в о. Состоятельность оценки (30) следует из (4), (7), (8), (10), (30), таблицы и теоремы о функциональном преобразовании сходящейся по вероятности случайной последовательности [12].

Далее, из (30) следует, что оценка энтропии Реньи является линейным преобразованием логарифма статистики (10). Воспользуемся теоремой 4.2.5 из [13] и тем, что линейное преобразование асимптотически нормальной величины [8] также имеет асимптотически нормальное распределение. В качестве функции $f(u)$ в теореме 4.2.5 из [13] выберем $f(u) = \ln u$. Тогда $\left(f'(u)\right)^2 = \frac{1}{u^2}$, откуда с учетом (23), вида формулы (30) и свойств дисперсии вытекает справедливость формулы (38).

Также из теоремы 4.2.5 из [13] и формулы (30) с учетом (23) следует

$$\mu_{H,r} = \ln n + \frac{1}{r-1}\left(\ln n - \ln \mu_{n,r}\right) = \ln n + \frac{1}{r-1}\left(\ln n - \ln n\lambda^{r-1}\right) =$$

$$= \ln n + \frac{1}{r-1}\left(\ln n - \ln n - (r-1)\ln\lambda\right) = \ln n - \ln\lambda = \ln N,$$

что завершает доказательство теоремы 3.

С л е д с т в и е 3. *При* $r = 2$ *для дисперсии асимптотического распределения вероятностей оценки* (30) *справедливо выражение*:

$$\sigma^2_{H,2} = \frac{2}{n\lambda}. \tag{39}$$

Д о к а з а т е л ь с т в о. Подставим (29) в (38), получим доказываемое: $\sigma^2_{H,2} = \frac{2n\lambda}{n^2\lambda^2} = \frac{2}{n\lambda}$.

Отметим, что при истинной гипотезе H_* $p_k = 1/N, k = 1, \ldots, N$, поэтому значение энтропии Реньи равно

$$H_r(P) = \frac{1}{1-r}\ln\left(\sum_{k=1}^{N} p_k^r\right) = \frac{1}{1-r}\ln\left(\sum_{k=1}^{N}\frac{1}{N^r}\right) = \ln N,$$

что совпадает с (37).

Знание асимптотического распределения точечной состоятельной оценки (30) позволяет построить интервальную оценку энтропии Реньи:

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

Проверка гипотезы о «чистой случайности» последовательности на основе оценок энтропии Реньи и Тсаллиса. Полученные интервальные оценки 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: $\widehat{h_r}(n,N)$ – статистическая оценка энтропии Тсаллиса (31) или Реньи (30), μ – асимптотическое математическое ожидание статистической оценки энтропии Тсаллиса (32) или Реньи (37), σ^2 – асимптотическая дисперсия статистической оценки энтропии Тсаллиса (33) или Реньи (38) при истинной гипотезе H_*. Вычислим для наблюдаемой последовательности статистику $\widehat{h_r}(n,N)$. Решающее правило, основанное на статистике $\widehat{h_r}(n,N)$, имеет вид

$$\begin{cases} H_*, & \text{если } t_- < \widehat{h_r}(n,N) < t_+; \\ \overline{H_*}, & \text{в противном случае,} \end{cases} \quad t_\pm = \mu \pm \sigma \Phi^{-1}\left(1-\frac{\varepsilon}{2}\right). \tag{40}$$

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. В первом эксперименте для проверки разработанного решающего правила (40) в качестве выходной последовательности использовалась псевдослучайная последовательность $\{y_\tau\}, \tau = 1,\ldots,T$, длиной $T = 2^{33}$ бит, полученная при помощи прореживающего генератора [6] с порождающим многочленом $x^{15} + x + 1$ и управляющим многочленом $x^{11} + x^2 + 1$. 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 $x_t = \sum_{j=1}^{s} 2^{j-1} X_j^{(t)} + 1$. Длина фрагмента *s* принимала значения $s \in \{11,\ldots,30\}$. Значения отклонения Δ оценки энтропии Реньи (30) при $r = 2$ от математического ожидания (37) и отнормированные нижние границы доверительных интервалов $-\sigma_{H,2}\Phi^{-1}\left(1-\frac{\varepsilon}{2}\right)$ на уровне значимости $\varepsilon = 0{,}05$ в зависимости от *s* представлены на рис. 1. Как видно, абсолютная величина

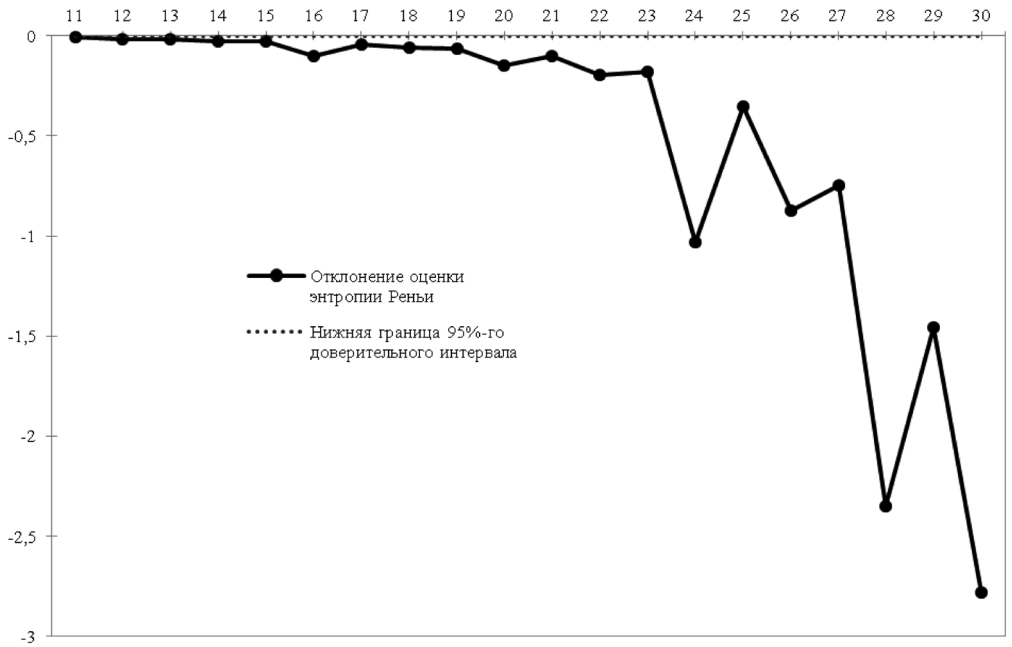

Рис. 1. Отклонение оценки энтропии Реньи от математического ожидания для $s \in \{11,\ldots,30\}$

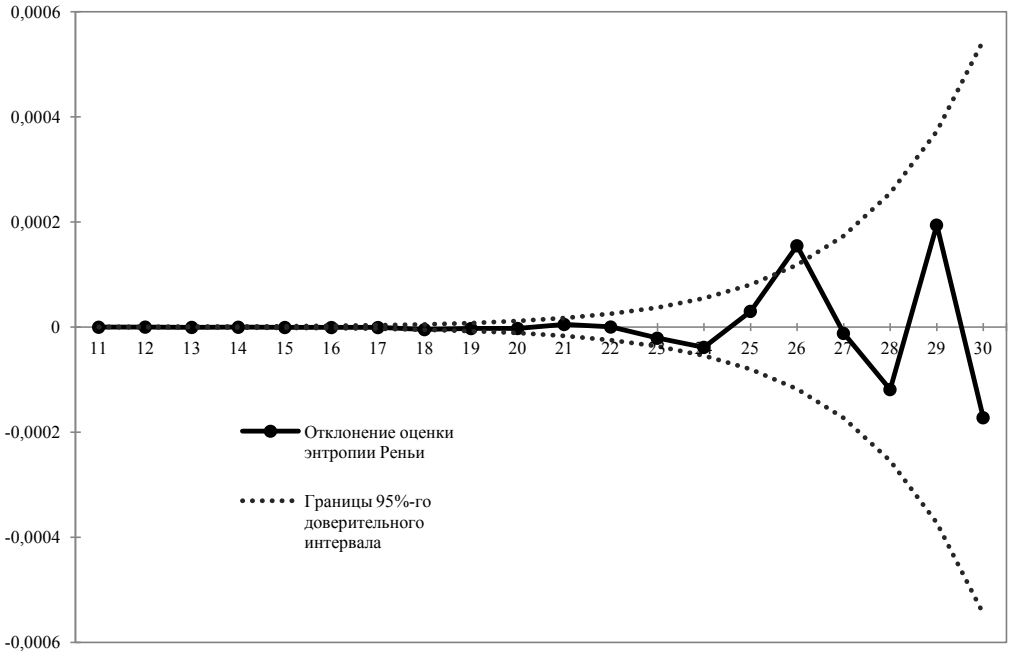

Рис. 2. Отклонение оценки энтропии Реньи от математического ожидания для $s \in \{11, \ldots, 30\}$

отклонения оценки значительно превышает границу доверительного интервала и гипотеза H_* уверенно отвергается начиная с $s = 16$.

Во втором эксперименте разработанное решающее правило (40) применено для анализа выходной двоичной последовательности реального физического генератора двоичной случайной последовательности [14] $\{y_\tau\}, \tau = 1, \ldots, T,$ длиной $T = 125 \cdot 2^{25}$ бит. Новая последовательность $\{x_t\}$ из алфавита мощности $N = 2^s$ формировалась так же, как и в первом эксперименте. На рис. 2 представлены значения отклонения оценки энтропии Реньи (30) при $r = 2$ от математического ожидания (37), а также отнормированные границы доверительных интервалов $\pm\sigma_{H,2}\Phi^{-1}\left(1-\frac{\varepsilon}{2}\right)$ на уровне значимости $\varepsilon = 0{,}1$ в зависимости от $s \in \{11, \ldots, 30\}$. Как видно, при значениях $s \le 25$ выходная последовательность генератора согласуется с моделью РРСП.

Заключение. Построены состоятельные, асимптотически нормально распределенные статистические оценки функционалов энтропии Реньи и Тсаллиса. Получены явные формулы для моментов построенных статистических оценок. Построено 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 использованной литературы

1. *Esteban, M. D.* A summary on entropy statistics / M. D. Esteban, D. Morales // Kybernetika. – 1995. – Vol. 31, N 4. – P. 337–346.

2. *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.

3. *Палуха, В. Ю.* Вероятностные свойства оценки многомерной энтропии выходных последовательностей криптографических генераторов / В. Ю. Палуха, Ю. С. Харин // Веб-программирование и Интернет-технологии WebConf-2015: Материалы 3-й Междунар. науч.-практ. конф., Минск, 12–14 мая 2015 г. – Минск: Изд. центр БГУ, 2015. – С. 146–147.

4. 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.

5. *Bonachela, J. A.* Entropy estimates of small data sets / J. A. Bonachela, H. Hinrichsen, M. A. Muñoz // J. Phys. A: Mathematical and Theoretical. – 2008. – Vol. 41, N 20. – 202001 (9 p).

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

7. *Holst, L.* Asymptotic normality and efficiency for certain goodness-of-fit tests / L. Holst // Biometrika. – 1972. – N 59. – P. 137–145.

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

9. *Энвин, А. Ю.* Дискретная математика: конспект лекций / А. Ю. Энвин. – Челябинск: Изд-во ЮУрГУ, 1998.

10. Математическая статистика: учеб. пособие / Моск. гос. ин-т электроники и математики; авт.-сост.: Н. Ю. Энатская, Е. Р. Хакимуллин. – Москва: МИЭМ, 2004. – Ч. 2.

11. *Riordan, J.* Moment recurrence relations for binomial, Poisson and hypergeometric frequency distributions / J. Riordan // Annals of Mathematical Statistics. – 1937. – Vol. 8, N 2. – P. 103–111.

12. *Боровков, А. А.* Теория вероятностей / А. А. Боровков. – М.: Эдиториал УРСС, 1999.

13. *Андерсон, Т.* Введение в многомерный статистический анализ / Т. Андерсон; пер. с англ. Ю. Ф. Кичатова, Е. С. Кочеткова, Н. С. Райбмана; под ред. Б. В. Гнеденко. – М.: Физматгиз, 1963.

14. 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.

Поступила в редакцию 26.05.2016