УДК 004.056.55:519.233.33 Вестник СПбГУ. Сер. 1. Т. 3 (61). 2016. Вып. 4

О НЕКОТОРЫХ СТАТИСТИЧЕСКИХ СВОЙСТВАХ ПРЕОБРАЗОВАНИЯ «BOOK STACK»

А. В. Бзикадзе, В. В. Некруткин

Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

В статье изучаются статистические свойства так называемого преобразования «Book Stack», предложенного Б. Я. Рябко (Пробл. передачи инф., т. 16, вып. 4, 1980) в качестве процедуры сжатия информации. Это же преобразование было использовано Б. Я. Рябко и А. И. Пестуновым (Пробл. передачи инф., т. 40, вып. 1, 2004) при построении одноименного статистического теста. Тест предназначен для проверки гипотезы $\mathbb{H}_0$ о том, что имеющаяся «входная» повторная выборка соответствует дискретному равномерному распределению с известным носителем. При этом предлагается проверять эту гипотезу не для «входной» выборки, а для новой, полученной с помощью «Book Stack»-преобразования. Тем самым возникает естественная задача сравнения результатов применения одного и того же статистического теста к «входной» и «выходной» выборкам. При выполнении нулевой гипотезы эти процедуры являются эквивалентными, однако при отклонениях от $\mathbb{H}_0$ это, вообще говоря, уже не так. Результаты сравнения критериев, конечно, зависят от класса рассматриваемых альтернатив. В статье рассматривается естественная альтернатива, состоящая в том, что исходная повторная выборка соответствует дискретному, но не равномерному распределению с фиксированным носителем. При этом показано, что некоторые стандартные критерии для проверки гипотезы $\mathbb{H}_0$ оказываются более мощными при их применении к «входной» выборке, чем к преобразованной. В частности, это имеет место для критерия отношения правдоподобия и (с некоторыми формальными ограничениями) к критерию χ^2. Библиогр. 14 назв.

Ключевые слова: сжатие информации, «Book Stack»-преобразование, проверка статистических гипотез, дискретное равномерное распределение.

1. Введение и постановка задачи. В статье [1] предложен новый статистический критерий (*Book Stack-тест*, в дальнейшем BS-тест), предназначенный для проверки гипотезы о том, что наблюдаемая повторная независимая выборка взята из дискретного равномерного распределения.

Хотя в [1] целью применения этого теста объявляется проверка качества генераторов псевдослучайных чисел, сама идея критерия, несомненно, имеет более общий характер. После [1] опубликован еще ряд статей (например, [2–6]), так или иначе связанных с изучением применимости BS-теста.

В основе теста лежит одноименное преобразование, введенное в [7] в качестве простой и наглядной процедуры сжатия информации. В англоязычной литературе (см. [8]) более распространено название *Move-to-Front*. BS-преобразование приобрело достаточно широкую популярность и ныне используется в некоторых утилитах для сжатия данных — в основном как один из промежуточных шагов (см, например, [9]).

Дадим формальное описание BS-преобразования в удобных для дальнейшего повествования терминах. Пусть $\mathbb{S} = \{1, 2, \ldots, S\}$ и $\mathfrak{S}_S$ — множество всевозможных перестановок чисел от 1 до S. Для любого $x \in \mathbb{S}$ и любой перестановки $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_S)^{\mathrm{T}} \in \mathfrak{S}_S$ обозначим через $i_0 = i_0(\alpha, x)$ такой индекс, что $\alpha_{i_0} = x$.

DOI: 10.21638/11701/spbu01.2016.402

Тогда будем иметь

$$f(\alpha,x)[i]\stackrel{\text{def}}{=}\begin{cases}x & \text{при } i=1,\\ \alpha_{i-1} & \text{при } 1<i\leqslant i_0,\\ \alpha_i & \text{при } i>i_0.\end{cases}\tag{1}$$

Тем самым получим отображение

$$f=\Big(f(\alpha,x)[1],\ldots,f(\alpha,x)[S]\Big)^{\mathrm{T}}\colon\ \mathfrak{S}_S\times\mathbb{S}\mapsto\mathfrak{S}_S,$$

которое и называется BS-преобразованием. Переход от BS-преобразования к различным вариантам BS-теста происходит следующим образом. Рассмотрим последовательность случайных величин $\{\eta_i\}_{i\geqslant 1}$, предполагая, что $\eta_i\in\mathbb{S}$ при любом i.

Введем последовательность векторов $\{\Xi_n\in\mathfrak{S}_S\}_{n\geqslant 0}$ так, что для $i\geqslant 1$ выполняется

$$\Xi_i=(\Xi_i[1],\Xi_i[2],\ldots,\Xi_i[S])^{\mathrm{T}}=f(\Xi_{i-1},\eta_i),\tag{2}$$

а Ξ_0 — это (вообще говоря, случайный) S-мерный вектор, принимающий значения во множестве перестановок $\mathfrak{S}_S$. Предполагается, что случайный вектор Ξ_0 не зависит от $\{\eta_i\}_{i\geqslant 1}$.

Наконец, определим последовательность $\{\xi_i\}_{i\geqslant 1}$, где $\xi_i\in\mathbb{S}$ задается как решение уравнения

$$\eta_i=\Xi_{i-1}[\xi_i].\tag{3}$$

Заметим, что для любого $i\geqslant 1$ это решение существует и единственно, так как Ξ_{i-1} является некоторой перестановкой чисел $1,2,\ldots,S$, а $\eta_i\in\mathbb{S}$.

Нетрудно показать, что случайные величины $\{\xi_i\}_{i\geqslant 1}$ являются независимыми и равномерно распределенными на множестве $\mathbb{S}$ (последнее будет обозначаться как $\xi_i\in\mathrm{U}_S$), тогда и только тогда, когда последовательность $\{\eta_i\}_{i\geqslant 1}$ обладает таким же свойством (это сразу же следует из того, что при фиксированном Ξ_0 для любого $n\geqslant 1$ отображение $\{\eta_i\}_{i=1}^n\mapsto\{\xi_i\}_{i=1}^n$ является биекцией).

Пусть теперь нулевая гипотеза $\mathbb{H}_0$ относительно $\{\eta_i\}_{i\geqslant 1}$ состоит в том, что эти случайные величины независимы и $\eta_i\in\mathrm{U}_S$. Общая идея статистических тестов, основанных на BS-преобразовании, состоит в том, что гипотеза $\mathbb{H}_0$ проверяется с помощью случайных величин $\{\xi_i\}_{i\geqslant 1}$, а не исходных $\{\eta_i\}_{i\geqslant 1}$.

Для проверки гипотезы $\mathbb{H}_0$ существует много статистических критериев (см., например, [10]), среди которых наиболее популярными являются критерий χ^2 и критерий отношения правдоподобия. Заметим, что в оригинальной постановке BS-тест использует критерий χ^2 со специальным видом разбиения множества $\mathbb{S}$.

Конечно, о качестве критерия нельзя говорить, не рассматривая некоторый класс альтернатив. Здесь мы ограничимся естественной для статистики альтернативой $\mathbb{H}_1$, состоящей в том, что случайные величины $\{\eta_i\}_{i\geqslant 1}$ являются независимыми и одинаково распределенными, но распределение $\mathcal{P}\stackrel{\text{def}}{=}\mathcal{L}(\eta_i)$ не совпадает с U_S (здесь и далее $\mathcal{L}(\delta)$ обозначает распределение случайной величины δ со значениями в произвольном измеримом пространстве).

В настоящей работе показано, что в условиях альтернативы $\mathcal{P}\neq\mathrm{U}_S$ критерии отношения правдоподобия и χ^2, примененные к последовательности $\{\xi_i\}_{i\geqslant 1}$, будут при

больших объемах выборки (и при некоторых дополнительных условиях) менее мощными, чем такие же критерии, примененные к исходной последовательности $\{\eta_i\}_{i\geqslant 1}$. Аналогичный факт оказывается верным и для некоторых других критериев.

Общий ход рассуждений следующий. В разделе 2 доказывается, что последовательность Ξ_i является однородной марковской цепью (ОМЦ), причем при условии $p_i > 0$ для любого i эта цепь является эргодической. Отсюда (раздел 3) выводится условие сходимости вероятности $\mathrm{P}(\xi_n = k)$ при всех k от 1 до S к некоторым предельным значениям $s_k > 0$. Более того, для последовательности $\{\xi_i\}_{i\geqslant 1}$ выполняется закон больших чисел: если положить

$$\tau_k^{(\xi)} = \tau_k^{(\xi)}(n) = \sum_{j=1}^{n} \mathbb{I}_k(\xi_j), \tag{4}$$

(здесь и далее $\mathbb{I}_x$ обозначает индикатор одноточечного множества $A = \{x\}$), то будем иметь $\tau_k^{(\xi)}/n \xrightarrow{\mathrm{P}} s_k$ при $n \to \infty$.

В разделе 4 показано, что распределение, определяемое вероятностями s_k, расположено «ближе» к равномерному распределению U_S, чем $\mathcal{P}$. На этом и основаны финальные рассуждения раздела 5, относящиеся к мощностям критериев.

Введем еще несколько обозначений. Пусть задан вектор $(x_1, \ldots, x_n)^{\mathrm{T}} \in \mathbb{S}^n$ и перестановка $\alpha \in \mathfrak{S}_S$. Тогда при $f^{(1)}(\alpha, x_1) \stackrel{\mathrm{def}}{=} f(\alpha, x_1)$ для любого $2 \leqslant j \leqslant n$ определим рекуррентным образом

$$f^{(j)}(\alpha, x_j) \stackrel{\mathrm{def}}{=} f\left(f^{(j-1)}(\alpha, x_{j-1}), x_j\right). \tag{5}$$

Кроме того, для любого $m \geqslant 1$ и любых $\overline{\ell} = (\ell_0, \ldots, \ell_m)$, $\overline{k} = (k_0, \ldots, k_m) \in \mathbb{S}^{m+1}$ выполняется

$$\mathcal{F}_m(\overline{\ell}, \overline{k}) \stackrel{\mathrm{def}}{=} \left\{\alpha \in \mathfrak{S}_S \,|\, \alpha_{\ell_0} = k_0,\ f^{(j)}(\alpha, k_{j-1})[\ell_j] = k_j \ \text{ для } j \in 1:m\right\}. \tag{6}$$

Для $m = 0$ и любых $\ell, k \in \mathbb{S}$ справедливо

$$\mathcal{F}_0(\ell, k) \stackrel{\mathrm{def}}{=} \{\alpha \in \mathfrak{S}_S \,|\, \alpha_\ell = k\}. \tag{7}$$

2. Предельное поведение последовательности $\{\xi_i\}_{i\geq 1}$. Как уже говорилось, мы везде дальше будем предполагать, что $\{\eta_i\}_{i\geqslant 1}$ являются независимыми одинаково распределенными случайными величинами с распределением

$$\mathrm{P}(\eta_i = k) = p_k, \quad k \in \mathbb{S}. \tag{8}$$

Лемма 1. *Последовательность $\{\Xi_n\}_{n\geq 0}$, определенная в (2), образует ОМЦ с фазовым пространством $\mathfrak{S}_S$ и матрицей переходных вероятностей $\mathbf{P}_S = \left(p_{\alpha\beta}^{(S)}\right)$ размерности $(S!) \times (S!)$ такой, что при $\alpha, \beta \in \mathfrak{S}_S$ выполняется*

$$p_{\alpha\beta}^{(S)} = \mathrm{P}(f(\alpha, \eta_1) = \beta),$$

где f определено в (1). Более того, если $p_k > 0$ для любого $k \in \mathbb{S}$, эта марковская цепь является эргодической.

Доказательство. Первое утверждение непосредственно следует из представления (2) и того факта, что Ξ_{i-1} и η_i независимы (см. [11, гл. 1, § 12]). Если $p_k > 0$ для всех k, то, как нетрудно видеть, матрица $\mathbf{P}_S^k$ при $k \geqslant S$ имеет все положительные элементы. Тем самым утверждение доказано. □

Замечание 1. Если все p_k положительны, то, согласно лемме 1, у марковской цепи Ξ_n существует единственное стационарное распределение $\pi = \pi_S$, которое является решением системы $\pi \mathbf{P}_S = \pi$. Можно показать, что это стационарное распределение задается вектором-строкой

$$\pi_S = (\pi_{1,2,\ldots,S}, \pi_{1,2,\ldots,S,S-1}, \ldots, \pi_{S,S-1,\ldots1}) \in \mathbb{R}^{S!}, \tag{9}$$

где

$$\pi_{i_1,i_2,\ldots,i_s} = \prod_{k=1}^{S-1} p_{i_k} \Big/ \prod_{k=1}^{S-2} \left(1 - \sum_{j=1}^{k} p_{i_j}\right). \tag{10}$$

Поскольку явный вид π_S в дальнейшем не используется, мы опускаем доказательство этого факта.

Перейдем к изучению стационарного поведения последовательности $\{\xi_i\}_{i\geqslant1}$.

Предложение 1. *Пусть начальное распределение марковской цепи $\{\Xi_i\}_{i\geqslant0}$ — стационарное распределение π_S, определенное в* (9), (10). *Тогда последовательность $\{\xi_i\}_{i\geqslant1}$ является стационарной в узком смысле. Точнее, в обозначениях $\overline{\ell} = (l_0, \ldots, l_m), \overline{k} = (k_0, \ldots, k_m) \in \mathbb{S}^{m+1}$ и $\overline{\xi}_i = (\xi_i, \ldots, \xi_{i+m})$ имеет место равенство*

$$\mathrm{P}\big(\overline{\xi}_i = \overline{\ell}\big) = \sum_{k_0,\ldots,k_m} \prod_{j=0}^{m} p_{k_j} \sum_{\alpha\in\mathcal{F}_m(\overline{\ell},\overline{k})} \pi_\alpha, \tag{11}$$

где множество $\mathcal{F}_m(\overline{\ell},\overline{k})$ определено в (6).

Доказательство. Обозначим $\overline{\eta}_i = (\eta_i, \ldots, \eta_{i+m})$. Тогда будем иметь

$$\mathrm{P}\big(\overline{\xi}_i = \overline{\ell}\big) = \sum_{\overline{k},\,\alpha} \mathrm{P}\big(\overline{\xi}_i = \overline{\ell} \,\big|\, \overline{\eta}_i = \overline{k}, \Xi_{i-1} = \alpha\big)\, \mathrm{P}\big(\overline{\eta}_i = \overline{k}, \Xi_{i-1} = \alpha\big).$$

Так как распределение Ξ_{i-1} совпадает со стационарным, η_i и Ξ_{i-1} независимы и $\mathrm{P}\big(\overline{\eta}_i = \overline{k}\big) = \prod_{j=0}^{m} p_{k_j}$, то выполняется $\mathrm{P}\big(\overline{\eta}_i = \overline{k}, \Xi_{i-1} = \alpha\big) = \pi_\alpha \prod_{j=0}^{m} p_{k_j}$. С другой стороны, по определению множества $\mathcal{F}_m(\overline{\ell},\overline{k})$ справедливо

$$\mathrm{P}\big(\overline{\xi}_i = \overline{\ell} \,\big|\, \overline{\eta}_i = \overline{k}, \Xi_{i-1} = \alpha\big) = \begin{cases} 1 & \text{при } \ \alpha \in \mathcal{F}_m(\overline{\ell},\overline{k}), \\ 0 & \text{иначе.} \end{cases}$$

Тем самым равенство (11) доказано. Поскольку правая его часть не зависит от i, то и стационарность последовательности $\{\xi_i\}_{i\geqslant1}$ тоже установлена. □

Следствие 1. *Так как* $\mathcal{F}_0(\ell,k)=\{\alpha\in\mathfrak{S}_S\,|\,\alpha_\ell=k\}$, *то в условиях предложения 1 для любых* $j\in\mathbb{S}$ *и* $i\geqslant 1$ *выполняется*

$$s_j\stackrel{\text{def}}{=}\mathrm{P}(\xi_i=j)=\sum_{k=1}^{S}p_k\sum_{\substack{\alpha\in\mathfrak{S}_S\\ \alpha_j=k}}\pi_\alpha. \tag{12}$$

Распределение с вероятностями (12) далее будем обозначать $\mathcal{R}$.

3. Закон больших чисел для частот последовательности $\{\xi_i\}_{i\geqslant 1}$**.** Здесь, как и раньше, предполагается, что вероятности p_k являются положительными.

Теорема 1. *Для любого начального распределения* $\mathcal{L}(\Xi_0)$ *справедливы соотношения*

$$\mathrm{P}(\xi_n=k)\underset{n\to+\infty}{\longrightarrow}s_k \tag{13}$$

и

$$\frac{\tau_k^{(\xi)}}{n}\underset{n\to+\infty}{\overset{\mathrm{P}}{\longrightarrow}}s_k, \tag{14}$$

где s_k *— вероятности, определенные в* (12)*, а частоты* $\tau_k^{(\xi)}$ *введены в* (4)*.*

Доказательство. Поскольку конечная ОМЦ $\{\Xi_i\}_{i\geqslant 0}$ является эргодической, то имеют место неравенства

$$|p_{\alpha\beta}^{(S)}(n)-\pi_\beta|\leqslant c\rho^n\quad\text{и}\quad\bigl|\mathrm{P}(\Xi_k=\beta)-\pi_\beta\bigr|\leqslant c\rho^k, \tag{15}$$

где $0\leqslant\rho<1$, $c>0$, $\beta\in\mathfrak{S}_S$, $k\geqslant 0$, а π_β определена в (10).

Используя (15), оценим сверху величину $\bigl|\mathrm{P}(\Xi_i=\alpha,\Xi_m=\beta)-\pi_\alpha\pi_\beta\bigr|$ для $\alpha,\beta\in\mathfrak{S}_S$ и $1\leqslant i<m$:

$$\bigl|\mathrm{P}(\Xi_i=\alpha,\Xi_m=\beta)-\pi_\alpha\pi_\beta\bigr|=$$
$$=\bigl|\mathrm{P}(\Xi_i=\alpha)\,p_{\alpha\beta}^{(S)}(m-i)-\pi_\alpha\pi_\beta\bigr|\leqslant C\begin{cases}\rho^i & \text{при}\quad 2i\leqslant m,\\ \rho^{m-i} & \text{при}\quad 2i>m,\end{cases}$$

здесь и далее C обозначает некоторую абсолютную постоянную.

Отсюда при фиксированном $k\in\mathbb{S}$, используя формулу полной вероятности и определение (3) случайных величин ξ_n, получим при $n\to\infty$

$$\mathrm{E}\,\mathbb{I}_k(\xi_n)=\sum_{j=1}^{S}\mathrm{E}\,\mathbb{I}_j(\eta_n)\sum_{\substack{\alpha\in\mathfrak{S}_S,\\ \alpha_k=j}}\mathrm{E}\,\mathbb{I}_\alpha(\Xi_{n-1})=\sum_{j=1}^{S}p_k\sum_{\substack{\alpha\in\mathfrak{S}_S,\\ \alpha_k=j}}\bigl(\pi_\alpha+\mathrm{O}(\rho^n)\bigr)=$$
$$=s_k+\mathrm{O}(\rho^n)\to s_k. \tag{16}$$

Тем самым (13) доказано. Аналогичным образом можно убедиться в том, что выполняется неравенство

$$\mathrm{E}\mathbb{I}_k(\xi_i)\mathbb{I}_k(\xi_j)-s_k^2\leqslant C\begin{cases}\rho^{i-1} & \text{при } 2i\leqslant j,\\ \rho^{j-i} & \text{при } i<j<2i,\end{cases}$$

и, следовательно, неравенство

$$\sum_{i,j=1}^{n} \bigl(\mathrm{E}\,\mathbb{I}_k(\xi_i)\mathbb{I}_k(\xi_j) - s_k^2\bigr) \leqslant Cn. \tag{17}$$

Покажем теперь, что $\mathrm{E}\bigl(\tau_k^{(\xi)}/n - s_k\bigr)^2 \to 0$ при $n \to \infty$, откуда и будет следовать (14). Для фиксированного $k \in \mathbb{S}$ имеем

$$n^2\mathrm{E}\left(\frac{\tau_k^{(\xi)}}{n - s_k}\right)^2 = \mathrm{E}\left(\sum_{i=1}^{n}(\mathbb{I}_k(\xi_i) - s_k)\right)^2 = \mathrm{E}\left(\sum_{i=1}^{n}(\mathbb{I}_k(\xi_i) - s_k)\right)\left(\sum_{j=1}^{n}(\mathbb{I}_k(\xi_j) - s_k)\right) =$$

$$= \sum_{i,j=1}^{n} \bigl(\mathrm{E}\,\mathbb{I}_k(\xi_i)\mathbb{I}_k(\xi_j) - s_k\mathrm{E}\,\mathbb{I}_k(\xi_j) - s_k\mathrm{E}\,\mathbb{I}_k(\xi_i) + s_k^2\bigr) =$$

$$= \sum_{i,j=1}^{n} \Bigl(\mathrm{E}\,\mathbb{I}_k(\xi_i)\mathbb{I}_k(\xi_j) - s_k^2 + \Delta\bigl(\rho^{\min\{i,j\}}\bigr)\Bigr), \tag{18}$$

где $\Delta\bigl(x\bigr) \leqslant Cx$, а последнее равенство следует из (16).

Используя (17), нетрудно убедиться в том, что левая часть (18) имеет вид $\mathrm{O}(n)$ при $n \to \infty$. Тем самым второе утверждение тоже доказано. □

4. Эффект выравнивания вероятностей. Здесь мы покажем, что в случае, когда распределение $\mathcal{P} = \mathcal{L}(\eta_i)$ отличается от равномерного U_S, предельное распределение $\mathcal{R}$ последовательности $\{\xi_i\}_{i\geqslant 1}$ оказывается «ближе» к равномерному, чем распределения исходной последовательности $\{\eta_i\}_{i\geqslant 1}$. В качестве меры близости $\mathcal{Q} = (q_1, \ldots, q_S)$ к равномерному распределению U_S рассматриваются следующие характеристики:

а) двоичная энтропия

$$\mathcal{H}_2(\mathcal{Q}) = -\sum_{i=1}^{S} q_i \log_2 q_i$$

(чем она больше, тем ближе распределение $\mathcal{Q}$ к равномерному);

б) $\rho_1(Q, \mathrm{U}_S) \stackrel{\text{def}}{=} \sum_{k=1}^{S} |q_k - 1/S|$, что представляет собой удвоенное расстояние по вариации между Q и U_S;

в) $\rho_2(Q, \mathrm{U}_S) \stackrel{\text{def}}{=} \sqrt{\sum_{k=1}^{S}(q_k - 1/S)^2}$;

г) $\rho_\infty(Q, \mathrm{U}_S) \stackrel{\text{def}}{=} \max_k |q_k - 1/S|$.

Рассмотрим сначала энтропии распределений $\mathcal{P}$ и $\mathcal{R}$.

Теорема 2. *Пусть случайные величины $\{\eta_i\}_{i\geqslant 1}$ независимы и одинаково распределены с распределением $\mathcal{P} \neq \mathrm{U}_S$, задаваемым набором положительных вероятностей $\{p_i\}_{i=1}^{S}$. Обозначим через $\mathcal{R} = (s_1, \ldots, s_S)$ предельное распределение случайных величин ξ_n при $n \to \infty$. Тогда будет справедливым неравенство $\mathcal{H}_2(\mathcal{R}) > \mathcal{H}_2(\mathcal{P})$.*

Доказательство. Согласно следствию 1, имеем

$$s_i = \sum_{j=1}^{S} p_j a_{ij}, \quad \text{где} \;\; a_{ij} = \sum_{\substack{\alpha\in\mathfrak{S}_S \\ \alpha_i = j}} \pi_\alpha. \tag{19}$$

Кроме того, в силу формулы полной вероятности справедливы равенства

$$\sum_{i=1}^{S} a_{ij} = \sum_{i=1}^{S} \sum_{\substack{\alpha\in\mathfrak{S}_S \\ \alpha_i=j}} \pi_\alpha = 1 \quad \text{и} \quad \sum_{j=1}^{S} a_{ij} = \sum_{j=1}^{S} \sum_{\substack{\alpha\in\mathfrak{S}_S \\ \alpha_i=j}} \pi_\alpha = 1.$$

Тем самым матрица $\mathbf{A} = (a_{ij})$ является бистохастической, и из [12, гл. 2, лемма 1] сразу же следует $\mathcal{H}_2(\mathcal{R}) \geqslant \mathcal{H}_2(\mathcal{P})$, причем для доказательства строгого неравенства достаточно установить, что выполняется $\max_k p_k > \max_k s_k$.

Будем считать, что числа p_i упорядочены по убыванию: $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_{S-1} \geqslant p_S$. Так как не все вероятности p_i одинаковы, существует $\ell \geqslant 2$ такое, что $p_1 > p_\ell$.

Покажем, что $p_1 > s_k$ для любого k. Согласно равенству (19), будем иметь

$$p_1 - s_k = p_1 - \sum_{j=1}^{S} a_{kj}p_j = p_1 - a_{k1}p_1 - \sum_{j=2}^{S} a_{kj}p_j =$$

$$= p_1(1 - a_{k1}) - \sum_{j=2}^{S} a_{kj}p_j > p_1(1 - a_{k1}) - p_1 \sum_{j=2}^{S} a_{kj} = p_1 \left(1 - a_{k1} - \sum_{j=2}^{S} a_{kj}\right) = 0,$$

так как матрица $\mathbf{A} = (a_{ij})$ является бистохастической. Таким образом, и строгое неравенство тоже доказано. □

Рассмотрим теперь другие меры отклонения распределений от U_S.

Теорема 3. *При $q = 1, 2$ и ∞ имеют место неравенства $\rho_q(\mathcal{R}, \mathrm{U}_S) \leqslant \rho_q(\mathcal{P}, \mathrm{U}_S)$, причем выполняется $\rho_\infty(\mathcal{R}, \mathrm{U}_S) < \rho_\infty(\mathcal{P}, \mathrm{U}_S)$.*

Доказательство. Введем обозначения $\mathbf{p} = (p_1, \ldots, p_S)^{\mathrm{T}}$, $\mathbf{s} = (s_1, \ldots, s_S)^{\mathrm{T}}$ и $\mathbf{e} = (1, \ldots, 1)^{\mathrm{T}}/S$. Тогда система (12) запишется в виде $\mathbf{s} = \mathbf{A}\mathbf{p}$, где $\mathbf{A} = (a_{ij})_{i,j=1}^{S}$ — бистохастическая матрица с элементами a_{ij}, определенными в (19). Поскольку $\mathbf{e} = \mathbf{A}\mathbf{e}$, то $\mathbf{A}(\mathbf{p} - \mathbf{e}) = \mathbf{s} - \mathbf{e}$, и поэтому выполняется неравенство

$$\|\mathbf{s} - \mathbf{e}\|_q \leqslant \|\mathbf{A}\|_q \|\mathbf{p} - \mathbf{e}\|_q,$$

где $\|\cdot\|_q$ — гельдерова векторная норма порядка q, а $\|\mathbf{A}\|_q$ — соответствующая подчиненная матричная норма. Поскольку матрица $\mathbf{A}$ бистохастическая, то по определению соответствующих матричных норм имеем

$$\|\mathbf{A}\|_1 = \max_{1\leqslant j\leqslant S} \sum_{i=1}^{S} |a_{ij}| = \|\mathbf{A}\|_\infty = \max_{1\leqslant i\leqslant S} \sum_{j=1}^{S} |a_{ij}| = 1 \quad \text{и} \quad \|\mathbf{A}\|_2 = \sqrt{\lambda},$$

где λ — максимальное по модулю собственное число матрицы $\mathbf{A}^{\mathrm{T}}\mathbf{A}$.

Из доказательства теоремы 2 следует $\max_k p_k > \max_k s_k$. Совершенно аналогично доказывается $\min_k p_k < \min_k s_k$, и поэтому нужное нам утверждение доказано для $q = 1$ и $q = \infty$.

Заметим, что при $q = 2$ матрица $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ также является бистохастической, и поэтому ее максимальное собственное число равно 1, что завершает доказательство. □

5. Статистические приложения. Рассмотрим теперь применение полученных результатов к проверке статистических гипотез. Оно основано на следующем простом факте, не требующем специального доказательства.

Лемма 2. *Пусть* $\{\beta_{1n}\}_{n\geqslant 1}$ *и* $\{\beta_{2n}\}_{n\geqslant 1}$ *— две последовательности случайных величин, причем* $\beta_{1n} \xrightarrow{\mathrm{P}} b_1$ *и* $\beta_{2n} \xrightarrow{\mathrm{P}} b_2$ *при* $n \to \infty$*, где* b_1, b_2 *— некоторые постоянные. Тогда, если* $b_1 > b_2$*, выполняется* $\mathrm{P}(\beta_{1n} > \beta_{2n}) \to 1$ *при* $n \to \infty$.

Пусть $\eta_1, \ldots, \eta_n$ повторная независимая выборка из распределения $\mathcal{P}$, сосредоточенного на $\mathbb{S}$ и такого, что $p_k = \mathrm{P}(\eta_i = k) \neq 0$ для всех $1 \leqslant k \leqslant S$. Поставим задачу проверки гипотезы $\mathbb{H}_0$, состоящей в справедливости равенства $\mathcal{P} = \mathrm{U}_S$. Как уже говорилось во введении, ее можно проверять как с помощью исходной выборки $\eta_1, \ldots, \eta_n$, так и с помощью $\xi_1, \ldots, \xi_n$.

Для этой цели применим критерий отношения правдоподобия. Зададим при каждом $k \in \mathbb{S}$ величину $\tau_k^{(\xi)}$ формулой (4), аналогичным образом положим $\tau_k^{(\eta)} = \sum_{j=1}^n \mathbb{I}_k(\eta_j)$ и рассмотрим статистики

$$G_n^2(\eta) = 2\sum_{k=1}^S \tau_k^{(\eta)} \ln\left(\frac{S\tau_k^{(\eta)}}{n}\right) \quad \text{и} \quad G_n^2(\xi) = 2\sum_{k=1}^S \tau_k^{(\xi)} \ln\left(\frac{S\tau_k^{(\xi)}}{n}\right).$$

При выполнении нулевой гипотезы обе статистики асимптотически имеют распределение χ^2 с $(S-1)$-й степенью свободы (результат восходит к [13]). На этом факте и основан критерий отношения правдоподобия, отвергающий нулевую гипотезу при больших значениях $G_n^2(\eta)$ или $G_n^2(\xi)$.

В то же время нетрудно видеть (например, [14]), что имеет место

$$\widehat{\mathcal{H}}_n(\mathcal{P}) \stackrel{\mathrm{def}}{=} -\sum_{k=1}^S \left(\frac{\tau_k^{(\eta)}}{n}\right) \log_2\left(\frac{\tau_k^{(\eta)}}{n}\right) = \log_2 S - \frac{G_n^2(\eta)}{2n\ln 2},$$

и для выборочной энтропии $\widehat{\mathcal{H}}_n(\mathcal{R})$ выполняется аналогичное тождество. Это означает, что гипотеза $\mathbb{H}_0$ отвергается при слишком маленьких значениях выборочной энтропии.

Поскольку имеем $\widehat{\mathcal{H}}_n(\mathcal{P}) \xrightarrow{\mathrm{P}} \mathcal{H}(\mathcal{P})$ и $\widehat{\mathcal{H}}_n(\mathcal{R}) \xrightarrow{\mathrm{P}} \mathcal{H}(\mathcal{R})$, а также $\mathcal{H}(\mathcal{R}) > \mathcal{H}(\mathcal{P})$ при $\mathcal{P} \neq \mathrm{U}_S$, отсюда сразу же следует, что при альтернативе $\mathcal{P} \neq \mathrm{U}_S$ критерий отношения правдоподобия, примененный к $\xi_1, \ldots, \xi_n$, будет при больших n иметь меньшую мощность, чем такой же критерий, примененный к $\eta_1, \ldots, \eta_n$.

Аналогичные рассуждения можно применить к критерию, основанному на метрике ρ_∞ (для получения статистик этого критерия достаточно заменить q_k на $\tau_k^{(\eta)}$ и $\tau_k^{(\eta)}$ в формуле для $\rho_\infty(\mathcal{Q}, \mathrm{U}_S)$).

С некоторыми оговорками такой же вывод можно сделать относительно критерия χ^2. В этом случае вместо статистик $G_n^2(\eta)$ и $G_n^2(\xi)$ мы имеем дело с

$$\chi_n^2(\eta) = \sum_{k=1}^S \frac{\left(\tau_k^{(\eta)} - n/S\right)^2}{n/S} \quad \text{и} \quad \chi_n^2(\xi) = \sum_{k=1}^S \frac{\left(\tau_k^{(\xi)} - n/S\right)^2}{n/S},$$

причем нулевая гипотеза отвергается, если статистика χ^2_n оказывается слишком большой. Заметим, что при $n \to \infty$ выполняется

$$\frac{\chi^2_n(\eta)}{(Sn)} = \sum_{k=1}^{S} \left(\frac{\tau_k^{(\eta)}}{n} - \frac{1}{S} \right)^2 \xrightarrow{\mathrm{P}} \rho_2(\mathcal{P}, \mathrm{U}_S).$$

Для статистики $\chi^2_n(\xi)/(Sn)$ имеет место аналогичная сходимость к $\rho_2(\mathcal{R}, \mathrm{U}_S)$, причем справедливо неравенство $\rho_2(\mathcal{P}, \mathrm{U}_S) \geqslant \rho_2(\mathcal{R}, \mathrm{U}_S)$. Поэтому можно ожидать, что при больших n с вероятностью, близкой к 1, значение статистики $\chi^2_n(\eta)$ будет больше, чем значение статистики $\chi^2_n(\xi)$.

Конечно, такой вывод является математически обоснованным только для распределений $\mathcal{P}$, удовлетворяющих строгому неравенству $\rho_2(\mathcal{P}, \mathrm{U}_S) > \rho_2(\mathcal{R}, \mathrm{U}_S)$, однако это неравенство для любых $\mathcal{P} \neq \mathrm{U}_S$ удалось проверить (с помощью символьных вычислений) лишь при $S = 2, 3$.

При $S > 3$ проведены отдельные компьютерные эксперименты с некоторым набором распределений $\mathcal{P} \neq \mathrm{U}_S$, в каждом из них зарегистрировано неравенство $\rho_2(\mathcal{P}, \mathrm{U}_S) > \rho_2(\mathcal{R}, \mathrm{U}_S)$.

Выводы, аналогичные случаю метрики ρ_2, можно сделать и для критерия, основанного на метрике ρ_1.

Тем самым, если входные случайные величины $\{\eta_i\}_{i \geqslant 1}$ независимы и одинаково распределены с $\mathcal{L}(\eta_i) = \mathcal{P}$, то идея применения критериев с использованием последовательности $\{\xi_i\}_{i \geqslant 1}$ (вместо $\{\eta_i\}_{i \geqslant 1}$) для проверки гипотезы $\mathcal{P} = \mathrm{U}_S$ против альтернативы $\mathcal{P} \neq \mathrm{U}_S$, скорее всего, является малопродуктивной.

6. Благодарности. Авторы благодарят рецензентов за ценные замечания, повлекшие значительное улучшение качества статьи.

Литература

1. *Рябко Б. Я., Пестунов А. И.* «Стопка книг» как новый статистический тест для случайных чисел // Проблемы передачи информации. 2004. Т. 40, вып. 1. С. 73–78.

2. *Монарев В. А., Рябко Б. Я.* Экспериментальный анализ генераторов псевдослучайных чисел при помощи нового статистического теста // Ж. вычисл. матем. и матем. физ. 2004. Т. 44, вып. 5. С. 812–816.

3. *Ryabko B., Stognienko V., Shokin Yu.* A new test for randomness and its application to some cryptographic problems // Journal of Statistical Planning and Inference. 2004. Vol. 123, N 2. P. 365–376.

4. *Ryabko B., Monarev V.* Using information theory approach to randomness testing // Journal of Statistical Planning and Inference. 2005. Vol. 133. P. 95–110.

5. *Doroshenko S. et al.* On ZK-crypt, Book Stack, and statistical tests // IACR Cryptology ePrint Archive. 2006. P. 1–8. URL: http://eprint.iacr.org/2006/196.pdf (дата обращения: 27.08.2016).

6. *Doroshenko S., Ryabko B.* The experimental distinguishing attack on RC4 // IACR Cryptology ePrint Archive. 2006. P. 1–4. URL: http://eprint.iacr.org/2006/070.pdf (дата обращения: 27.08.2016).

7. *Рябко Б. Я.* Сжатие данных с помощью стопки книг // Проблемы передачи информормации. 1980. Т. XVI, вып. 4. С. 16–20.

8. *Bentley J., Sleator D, Tarjan R., Wei V.* A locally adaptive data compression scheme // Commun. ACM. 1986. Vol. 29, N 4. P. 320–330.

9. *Seward J.* bzip2 and libbzip2, version 1.0.5: A program and library for datacompression, 2007. URL: http://www.bzip.org/1.0.5/bzip2-manual-1.0.5.pdf (дата обращения: 27.08.2016).

10. *Read T., Cressie N.* Goodness-of-Fit statistics for discrete multivariate data. New York: Springer-Verlag, 1988. 224 p.

11. *Ширяев А. Н.* Вероятность. Т. 1. М.: Изд-во МЦНМО, 2004. 520 с.

12. *Файнстейн А.* Основы теории информации. М.: ИИЛ, 1960. 144 с.

13. *Hoeffding W.* Asymptotically optimal test for multinomial distributions // Ann. Math. Stat. 1965. Vol. 36, N 2. P. 269–401.

14. *L'Ecuyer P., Compagner A., Cordeau J-F.* Entropy tests for random number generation // Les Cahiers du GERAD, no. G-96-41, 1996, P. 1–22.

Статья поступила в редакцию 5 марта 2016 г.

Сведения об авторах

Бзикадзе Андрей Важевич — студент; seryrzu@gmail.com

Некруткин Владимир Викторович — кандидат физико-математических наук, доцент; v.nekrutkin@spbu.ru, vnekr@statmod.ru

ON SOME STATISTICAL PROPERTIES OF THE "BOOK STACK" TRANSFORMATION

Andrey V. Bzikadze, Vladimir V. Nekrutkin

St. Petersburg State University, Universitetskaya nab., 7–9, St. Petersburg, 199034, Russian Federation; seryrzu@gmail.com, v.nekrutkin@spbu.ru, vnekr@statmod.ru

This paper is devoted to the statistical properties of the so-called "Book Stack" transformation. It was proposed by B. Ryabko (Probl. Inf. Trans, **16**(4), 1980) as the method of data compression. B. Ryabko and A. Pestunov used (Probl. Inf. Trans, **40**(1), 2004) the transformation to construct the statistical test of the same name. The test is used to saying the null hypothesis that the "input" pure random sample is taken from the discrete uniform distribution with the given support. The idea of B. Ryabko and A. Pestunov is to verify this hypothesis with the help of the "output" sample which is obtained by the "Book Stack" transformation. Thus, arises the natural problem of comparing the results of the same criterion for "input" and "output" samples. Under the null hypothesis, procedures are equivalent, but if the null hypothesis fails, the results are generally different. Obviously, these results depend on the class of alternatives. The natural alternative is that the "input" pure random sample is taken from a discrete, but non-uniform, distribution with fixed support. We prove that several standard criteria applied to the "input" sample are more powerful than the same criteria applied to the "output" sample. In particular, this effect occurs for the likelihood ratio criterion and, with some formal restrictions, to the chi-square criterion. Refs 14.

Keywords: data compression, "Book Stack" transformation, discrete uniform distribution, statistical hypothesis.

References

1. Ryabko B., Pestunov A., ""Book Stack" as a new statistical test for random numbers", *Probl. Inform. Trans.* **40**(1), 66–71 (2004).

2. Monarev A., Ryabko B., "Experimental analysis of pseudorandom number generators by means of a new statistical test", *Computational Mathematics and Mathematical Physics* **44**(5), 766–770 (2004).

3. Ryabko B., Stognienko V., Shokin Yu. "A new test for randomness and its application to some cryptographic problems", *Journal of Statistical Planning and Inference* **123**(2), 365–376 (2004).

4. Ryabko B., Monarev V., "Using information theory approach to randomness testing", *Journal of Statistical Planning and Inference* **133**(1), 95–110 (2005).

5. Doroshenko S. et al., "On ZK-crypt, Book Stack, and statistical Tests", *IACR Cryptology ePrint Archive*, 1–8 (2006). Available at: http://eprint.iacr.org/2006/196.pdf (accessed 27.08.2016).

6. Doroshenko S., Ryabko B., "The experimental distinguishing attack on RC4", *IACR Cryptology ePrint Archive*, 1–4 (2006). Available at: http://eprint.iacr.org/2006/070.pdf (accessed 27.08.2016).

7. Ryabko B., "Data compression by means of a "Book Stack"", *Probl. Inf. Trans.* **16**(4), 265–269 (1980).

8. Bentley J., Sleator D., Tarjan R., Wei V., "A locally adaptive data compression scheme", *Commun. ACM* **29**(4), 320–330 (1986).

9. Seward J., *bzip2 and libbzip2, version 1.0.5: A program and library for datacompression* (2007). Available at: http://www.bzip.org/1.0.5/bzip2-manual-1.0.5.pdf (accessed 27.08.2016).

10. Read T., Cressie N., *Goodness-of-Fit statistics for discrete multivariate data* (Springer-Verlag, New York, 1988, 224 p.).

11. Shiryaev A. N., *Probability* (2nd edition, Springer-Verlag, Berlin-Heidelberg-New York, 1996, 620 p.).

12. Feinstein A., *Foundation of information theory* (McGraw-Hill Electrical and Electronic Engineering Series, Literary Licensing, LLC, 2013, 144 p.).

13. Hoeffding W., "Asymptotically optimal test for multinomial distributions", *Ann. Math. Stat.* **36**(2), 269–401 (1965).

14. L'Ecuyer P., Compagner A., Cordeau J-F., "Entropy tests for random number generation", *Les Cahiers du GERAD* no. G-96-41, 1–22, 1996.

Для цитирования: Бзикадзе А. В., Некруткин В. В. О некоторых статистических свойствах преобразования «Book Stack» // Вестник Санкт-Петербургского университета. Серия 1. Математика. Механика. Астрономия. 2016. Т. 3 (61). Вып. 4. С. 533–543. DOI: 10.21638/11701/spbu01.2016.402

For citation: Bzikadze A. V., Nekrutkin V. V. On some statistical properties of the "Book Stack" transformation. *Vestnik of Saint Petersburg University. Series 1. Mathematics. Mechanics. Astronomy*, 2016, vol. 3 (61), issue 4, pp. 533–543. DOI: 10.21638/11701/spbu01.2016.402

ХРОНИКА

18 мая 2016 г. на заседании секции теоретической механики им. проф. Н. Н. Поляхова в Санкт-Петербургском Доме Ученых РАН выступил кандидат физ.-мат. наук, доцент В. В. Чистяков (Университет ИТМО, Военно-космическая академия им. А. Ф. Можайского) с докладом на тему «О кинетике вращения динамически неуравновешенного ротатора в условиях комбинированного трения».

Краткое содержание доклада:

Изучается динамика как свободного, так и вынужденного вращения астатического ротатора вокруг неглавной оси инерции при действии сил сухого трения в опорах оси, а также сил аэрогидродинамического сопротивления. Для первых предполагается справедливость законов Кулона—Амонтона, для вторых — квадратичный закон Рэлея. Как аналитически, так и численно решены динамические уравнения для различных типов вынуждающих усилий: постоянный момент, синусоидальная нагрузка, прямоугольные импульсы и т. д. Показано, что в общем случае динамическое уравнение содержит иррациональность в правой части, что приводит к излому в кинетике угловой скорости. В частном случае это уравнение характеризуется разрывной правой частью. Соответствующая кинетика изменения угла и угловой скорости утрачивает свойство периодичности и проявляет признаки стохастического поведения.