Math-Net.Ru

Общероссийский математический портал

А. М. Зубков, Д. О. Меньшенин, Свойства статистики критерия симметричности Шекели и Мори в случае двоичных векторов, *Матем. заметки*, 2012, том 91, выпуск 4, 551–562

DOI: <https://doi.org/10.4213/mzm9325>

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением <http://www.mathnet.ru/rus/agreement>

Параметры загрузки:

IP: 54.147.182.235

26 апреля 2023 г., 08:30:32

Свойства статистики критерия симметричности Шекели и Мори в случае двоичных векторов

А. М. Зубков, Д. О. Меньшенин

В статье изучаются свойства статистики критерия симметричности распределения в евклидовом пространстве, предложенной Шекели и Мори, для класса дискретных распределений, сосредоточенных на множестве вершин d -мерного куба. Выводятся точные и асимптотические (при $d \rightarrow \infty$) формулы для первых моментов статистики, доказаны предельные теоремы, рассмотрены примеры, дающие представление о зависимости качества критерия от вида распределения.

Библиография: 6 названий.

1. Введение. Пусть $X, X_1, X_2, \dots$ – независимые одинаково распределенные в $\mathbb{R}^d$ векторы и $\|X\|$ обозначает евклидову норму вектора X . В [1] было показано, что

$$E(\|X_1 + X_2\| - \|X_1 - X_2\|) \geq 0$$

и что $E(\|X_1 + X_2\| - \|X_1 - X_2\|) = 0$ тогда и только тогда, когда случайный вектор X имеет симметричное распределение (т.е. распределения X и $-X$ совпадают).

На основании этого свойства в [1] было предложено использовать последовательность статистик

$$T_n = T_n(X_1, \dots, X_n) = \frac{\sum_{1 \leq i < j \leq n} (\|X_i + X_j\| - \|X_i - X_j\|)}{\sum_{1 \leq i \leq n} \|X_i\|}$$

для проверки гипотезы о симметричности распределения вектора X против класса всех несимметричных альтернатив. Согласно [1] если

$$E\|X\| < \infty \quad \text{и} \quad x(\alpha) = \left(\Phi^{-1}\left(\frac{\alpha}{2}\right) \right)^2$$

($\Phi(x)$ – функция стандартного нормального распределения, а $\Phi^{-1}(y)$ – обратная функция), то

$$\sup_{H_0} \lim_{n \rightarrow \infty} P\{1 + T_n \geq x(\alpha)\} = \alpha, \quad (1)$$

где H_0 – множество всех симметричных вероятностных распределений на $\mathbb{R}^d$.

Равенство в (1) достигается на симметричных двухточечных распределениях: если

$$P\{X_1 = a\} = P\{X_1 = -a\} = \frac{1}{2} \quad \text{для некоторого } a \in \mathbb{R}^d \setminus \{0\}$$

и ξ_n — число таких $i \in \{1, \dots, n\}$, что $X_i = a$, то ξ_n имеет биномиальное распределение с параметрами $(n, 1/2)$ и

$$T_n(X_1, \dots, X_n) = \frac{2}{n}(C_{\xi_n}^2 + C_{n-\xi_n}^2 - \xi_n(n - \xi_n)) = \left(\frac{\xi_n - n/2}{\sqrt{n/4}}\right)^2 - 1.$$

Поэтому при $x > 0$ в силу теоремы Муавра–Лапласа при $n \rightarrow \infty$

$$P\{T_n(X_1, \dots, X_n) + 1 \geq x\} = P\left\{\left|\frac{\xi_n - E\xi_n}{\sqrt{D\xi_n}}\right| \geq \sqrt{x}\right\} \rightarrow 2\Phi(-\sqrt{x}),$$

что соответствует (1).

В настоящей работе получены явные оценки и предельные распределения для случайных величин $\|X_i + X_j\| - \|X_i - X_j\|$ и $T_n(X_1, \dots, X_n)$ в случае, когда распределение X сосредоточено на множестве

$$V_d = \{U = (u_1, \dots, u_d) : u_j \in \{-1, +1\}, j = 1, \dots, d\}$$

вершин d -мерного куба. Эти результаты дают представление о качестве критерия, основанного на статистике $T_n(X_1, \dots, X_n)$. Отметим, что если распределение случайного вектора X сосредоточено на V_d , то $P\{\|X\| = \sqrt{d}\} = 1$ и поэтому T_n является U -статистикой.

2. Основные результаты. Если $X = (x_1, \dots, x_d)$ имеет равномерное распределение на V_d , то $x_1, \dots, x_d$ независимы и $P\{x_i = -1\} = P\{x_i = 1\} = 1/2$. Если вектор Y имеет такое же (симметричное) распределение и не зависит от X , то

$$E(\|X + Y\| - \|X - Y\|) = 0. \quad (2)$$

ТЕОРЕМА 1. Если случайные векторы

$$X = (x_1, \dots, x_d), \quad Y = (y_1, \dots, y_d)$$

независимы и равномерно распределены на V_d , $d \geq 2$, то

$$\begin{aligned} D(\|X + Y\| - \|X - Y\|) &= \frac{1}{2^{d-3}} \sum_{m=0}^d C_d^m \left(\frac{d}{2} - \sqrt{m(d-m)} \right) = 2 + \frac{\theta_d}{d}, \quad \theta_d \in [1, 6], \\ P\{(\|X + Y\| - \|X - Y\|) \leq t\} &\rightarrow \Phi\left(\frac{t}{\sqrt{2}}\right), \quad d \rightarrow \infty. \end{aligned}$$

Если векторы $X, X_1, X_2, \dots$ независимы и имеют равномерное распределение на V_d , то

$$ET_n = \frac{1}{n\sqrt{d}} C_n^2 E(\|X_1 + X_2\| - \|X_1 - X_2\|) = 0$$

и для произвольного $a \in V_d$

$$\mathbb{E}(\|a + X\| - \|a - X\|) = 0; \quad (3)$$

отсюда, используя свойства условного математического ожидания, нетрудно вывести, что для всех $1 \leq i < j$, $1 \leq k < l$, $(i, j) \neq (k, l)$,

$$\text{Cov}(\|X_i + X_j\| - \|X_i - X_j\|, \|X_k + X_l\| - \|X_k - X_l\|) = 0.$$

Таким образом, согласно теореме 1

$$\begin{aligned} DT_n &= \frac{1}{n^2 d} \mathbb{D} \left(\sum_{1 \leq i < j \leq n} (\|X_i + X_j\| - \|X_i - X_j\|) \right) \\ &= \frac{1}{n^2 d} C_n^2 \mathbb{D}(\|X_1 + X_2\| - \|X_1 - X_2\|) = \frac{1 + o(1)}{d} \left(1 + \frac{\theta_d}{2d} \right), \quad n \rightarrow \infty. \end{aligned}$$

Далее, в силу (3) U -статистика T_n является вырожденной. Применяя результат [2] (см. также [3]) к нашему случаю, получаем, что если $d = \text{const}$ и $n \rightarrow \infty$, то распределение T_n сходится к распределению суммы $\sum_{k=1}^{2^d} c_k \nu_k^2 - 1$, где $\nu_1, \nu_2, \dots$ — независимые случайные величины со стандартным нормальным распределением,

$$c_k = \frac{\lambda_k}{2^{d+1} \sqrt{d}}, \quad c_k \geq 0, \quad \sum c_k = 1,$$

и коэффициенты λ_k суть собственные значения линейного оператора

$$S: f(x) \rightarrow \mathbb{E}(\|X + x\| - \|X - x\|)f(X)$$

в $L^2(V_d)$ (см. [1]). Если X имеет равномерное распределение на V_d , то матрица оператора S имеет вид $(s_{ij})_{i,j=0,\dots,2^d-1}$, где

$$s_{ij} = \frac{1}{2^d} (\|A(i) + A(j)\| - \|A(i) - A(j)\|),$$

а векторы $A(j) = (a_0(j), \dots, a_{d-1}(j)) \in \{0, 1\}^d$ определяются соотношениями

$$\sum_{k=0}^{d-1} a_k(j) 2^k = j, \quad j \in \{0, 2^d - 1\}.$$

Матрица (s_{ij}) симметрична относительно главной и побочной диагоналей.

Общую формулу для собственных чисел оператора S найти не удалось. При $d = 2$ оператор S имеет 2 собственных значения, равных $4\sqrt{2}$, и 2 нулевых собственных значения. Следовательно, согласно [1] U -статистика T_n при $n \rightarrow \infty$ сходится по распределению к $\nu_1^2/2 + \nu_2^2/2 - 1$. При $d = 3$ оператор S имеет 4 нулевых собственных значения и 4 отличных от нуля:

$$\lambda_1 = \lambda_2 = \lambda_3 = -4 + 4\sqrt{3} + 4\sqrt{2}, \quad \lambda_4 = 12 + 4\sqrt{3} - 12\sqrt{2};$$

в этом случае

$$c_1 = c_2 = c_3 = 0,309786\dots, \quad c_4 = 0,070640\dots$$

При $d = 4$ оператор S имеет 8 нулевых и 8 отличных от нуля собственных значений:

$$\lambda_1 = \dots = \lambda_4 = 16 - 8\sqrt{3}, \quad \lambda_5 = \dots = \lambda_8 = 8\sqrt{3};$$

в этом случае

$$c_1 = \dots = c_4 = 0,216506\dots, \quad c_5 = \dots = c_8 = 0,033493\dots$$

В [4] доказано, что если $Q_n = \sum_{i=1}^n \lambda_i \nu_i^2$, где все λ_i неотрицательны, $\sum_{i=1}^n \lambda_i = 1$, а $\nu_1, \nu_2, \dots$ – независимые случайные величины со стандартным нормальным распределением, то при любом натуральном n и любом $x > 1.53640\dots$

$$P\{Q_n \leq x\} \geq P\{\nu_1^2 \leq x\}.$$

Эти результаты можно использовать при построении критериев согласия с гипотезой о равномерности распределения и независимости знаков, порождаемых генераторами случайных или псевдослучайных двоичных последовательностей.

Рассмотрим теперь распределения статистик $\|X_1 + X_2\| - \|X_1 - X_2\|$ и T_n при некоторых неравномерных распределениях на V_d .

ТЕОРЕМА 2. Если при каждом $d = 1, 2, \dots$ случайные векторы $X = (x_1, \dots, x_d)$, $Y = (y_1, \dots, y_d)$ со значениями в V_d независимы, одинаково распределены, имеют независимые компоненты,

$$P\{x_j = 1\} = \frac{1 + \varepsilon_j^{(d)}}{2}, \quad P\{x_j = -1\} = \frac{1 - \varepsilon_j^{(d)}}{2}, \quad |\varepsilon_j^{(d)}| \leq 1, \quad j = 1, \dots, d,$$

и для некоторого $\delta > 0$ для всех достаточно больших d

$$a_d \stackrel{\text{def}}{=} \frac{1}{d} \sum_{j=1}^d (\varepsilon_j^{(d)})^2 < 1 - \delta, \quad b_d \stackrel{\text{def}}{=} \frac{1}{d} \sum_{j=1}^d (\varepsilon_j^{(d)})^4 < a_d,$$

то при $d \rightarrow \infty$

$$\left| E(\|X + Y\| - \|X - Y\|) - \frac{2a_d\sqrt{2d}}{\sqrt{1-a_d} + \sqrt{1+a_d}} \right| = O\left(\frac{1}{\sqrt{d}}\right), \quad (4)$$

а распределение разности $\|X + Y\| - \|X - Y\|$ асимптотически нормально с параметрами

$$\left(\frac{2a_d\sqrt{2d}}{\sqrt{1-a_d} + \sqrt{1+a_d}}, (1-b_d) \frac{1 + \sqrt{1-a_d^2}}{1-a_d^2} \right).$$

Если $X = (x_1, \dots, x_d)$, $X_1, X_2, \dots$ – независимые случайные векторы и $a_d > 0$, то их распределение не симметрично, а U -статистика

$$T_n = \frac{1}{n\sqrt{d}} \sum_{1 \leq i < j \leq n} (\|X_i + X_j\| - \|X_i - X_j\|)$$

невырождена; тогда согласно [5] распределение T_n при $n \rightarrow \infty$, $d = \text{const}$ асимптотически нормально с параметрами

$$\left(\frac{nE(\|X_1 + X_2\| - \|X_1 - X_2\|)}{2\sqrt{d}}, \frac{4nD(E\{\|X_1 + X_2\| - \|X_1 - X_2\| \mid X_1\})}{d} \right).$$

Для конечного d и фиксированных $\varepsilon_1^{(d)}, \dots, \varepsilon_d^{(d)}$ параметры предельного нормального распределения принимают конкретные значения. Точные или простые асимптотические формулы для

$$D(\mathbb{E}\{\|X_1 + X_2\| - \|X_1 - X_2\| \mid X_1\})$$

получить не удалось.

Теоремы 1 и 2 были опубликованы в [6]; здесь приводятся уточненные и более простые доказательства.

В следующей теореме рассматривается равномерное распределение на подмножестве

$$G_d = \left\{ (v_1, \dots, v_d) \in V_d : \sum_{k=0}^d I\{v_k = 1\} \equiv 0 \pmod{2} \right\}$$

множества вершин V_d . Если d нечетно, то при любом $(v_1, \dots, v_d) \in V_d$ множеству G_d принадлежит ровно один из векторов $(v_1, \dots, v_d)$ и $-(v_1, \dots, v_d)$. Поэтому при нечетном d такое распределение случайного вектора X с формальной точки зрения является “абсолютно несимметричным” (носители распределений X и $-X$ не пересекаются). Функции распределения и математические ожидания статистики $\|X + Y\| - \|X - Y\|$ в случаях, когда X и Y имеют равномерные распределения на G_d или на всем V_d , очень слабо различаются, хотя расстояние по вариации между ее распределениями в этих двух случаях близко к $1/2$.

ТЕОРЕМА 3. Если случайные векторы $X = (x_1, \dots, x_d)$, $Y = (y_1, \dots, y_d)$ независимы и равномерно распределены на G_d , $d \geq 2$, то для любого $t \in (-\infty, \infty)$

$$P\{(\|X + Y\| - \|X - Y\|) \leq t\} \rightarrow \Phi\left(\frac{t}{\sqrt{2}}\right), \quad d \rightarrow \infty,$$

$$\mathbb{E}(\|X + Y\| - \|X - Y\|) = \frac{1 - (-1)^d}{2^{d-1}} \sum_{l=0}^d (-1)^{l+1} C_d^l \sqrt{l} < \frac{1}{2^{d-2}}.$$

Кроме того, если случайные векторы $X^* = (x_1^*, \dots, x_d^*)$, $Y^* = (y_1^*, \dots, y_d^*)$ независимы и равномерно распределены на V_d , то расстояние по вариации между распределениями $\|X + Y\| - \|X - Y\|$ и $\|X^* + Y^*\| - \|X^* - Y^*\|$ равно $1/2$.

Последняя теорема настоящей работы дает оценку среднего значения $\mathbb{E}(\|X + Y\| - \|X - Y\|)$ для независимых одинаково распределенных на V_d случайных векторов по множеству $2^d!$ распределений, получающихся из исходного распределения $P = \{p_W, W \in V_d\}$ всеми возможными перестановками набора вероятностей по множеству вершин куба V_d .

Для взаимно однозначного отображения (перестановки) $\sigma: V_d \rightarrow V_d$ обозначим через P_σ распределение $\{p_{\sigma(W)}, W \in V_d\}$ и через $\mathbb{E}_\sigma$ математическое ожидание для независимых случайных векторов X и Y , имеющих распределение P_σ . Пусть Σ_d обозначает множество всех перестановок из 2^d элементов.

ТЕОРЕМА 4. Если случайные векторы $X = (x_1, \dots, x_d)$, $Y = (y_1, \dots, y_d)$ независимы и одинаково распределены на V_d , $d \geq 2$, то справедливо равенство

$$\frac{1}{2^d!} \sum_{\sigma \in \Sigma_d} \mathbb{E}_\sigma(\|X_1 + X_2\| - \|X_1 - X_2\|) = 2\sqrt{d} \left(\sum_{w \in V_d} p_w^2 - \frac{1}{2^d - 1} \left(1 - \sum_{w \in V_d} p_w^2 \right) \right).$$

Рассмотренные примеры показывают, в частности, что величина $E(\|X + Y\| - \|X - Y\|)$ и расстояние по вариации между распределениями X и $-X$ не являются эквивалентными мерами несимметричности распределения.

3. Доказательства.

ДОКАЗАТЕЛЬСТВО ТЕОРЕМЫ 1. В силу (2)

$$\begin{aligned} D(\|X + Y\| - \|X - Y\|) &= E \left(\sqrt{\sum_{j=1}^d (x_j + y_j)^2} - \sqrt{\sum_{j=1}^d (x_j - y_j)^2} \right)^2 \\ &= 2E \sum_{j=1}^d (x_j^2 + y_j^2) - 2E \sqrt{\sum_{j=1}^d (x_j + y_j)^2} \sqrt{\sum_{j=1}^d (x_j - y_j)^2}. \end{aligned}$$

В нашем случае $E \sum_{j=1}^d (x_j^2 + y_j^2) = 2d$. Рассмотрим множества

$$A_m = \{(U, W) \in V_d \times V_d : |\{k : u_k = w_k\}| = m\}, \quad m = 0, \dots, d.$$

Если $(U, W) \in A_m$, то

$$\sqrt{\sum_{j=1}^d (u_j + w_j)^2} \sqrt{\sum_{j=1}^d (u_j - w_j)^2} = 4\sqrt{m(d-m)}.$$

Множество A_m содержит $2^d C_d^m$ элементов, множество $V_d \times V_d$ всех возможных пар (U, W) состоит из 2^{2d} элементов, пара (X, Y) имеет равномерное распределение на $V_d \times V_d$. Следовательно,

$$E \sqrt{\sum_{j=1}^d (x_j + y_j)^2} \sqrt{\sum_{j=1}^d (x_j - y_j)^2} = \frac{1}{2^{d-2}} \sum_{m=0}^d C_d^m \sqrt{m(d-m)}$$

и

$$\begin{aligned} D(\|X + Y\| - \|X - Y\|) &= 4d - \frac{1}{2^{d-3}} \sum_{m=0}^d C_d^m \sqrt{m(d-m)} \\ &= \frac{1}{2^{d-3}} \sum_{m=0}^d C_d^m \left(\frac{d}{2} - \sqrt{m(d-m)} \right) = 8E \left(\frac{d}{2} - \sqrt{\xi_d(d-\xi_d)} \right), \end{aligned}$$

где случайная величина $\xi_d = |\{k : x_k = y_k\}|$ имеет биномиальное распределение $\text{Bin}(d, 1/2)$.

Нетрудно проверить, что

$$\begin{aligned} E \left(\xi_d - \frac{d}{2} \right)^2 &= \frac{d}{4}, \quad E \left(\xi_d - \frac{d}{2} \right)^4 = \frac{d(3d-2)}{16}, \\ 1 - \frac{z}{2} - \frac{z^2}{2} &\leq \sqrt{1-z} \leq 1 - \frac{z}{2} - \frac{z^2}{8} \quad \text{при } 0 \leq z \leq 1. \end{aligned}$$

Таким образом,

$$\begin{aligned} \frac{d}{2} - \sqrt{z(d-z)} &= \frac{d}{2} \left(1 - \sqrt{1 - \left(\frac{z-d/2}{d/2} \right)^2} \right) = \frac{d}{2} \left(\frac{1}{2} \frac{(z-d/2)^2}{(d/2)^2} + \theta \frac{(z-d/2)^4}{(d/2)^4} \right) \\ &= \frac{(z-d/2)^2}{d} + 8\theta \frac{(z-d/2)^4}{d^3}, \quad \theta \in \left[\frac{1}{8}, \frac{1}{2} \right], \end{aligned}$$

и

$$\begin{aligned} D(\|X_1 + X_2\| - \|X_1 - X_2\|) &= 8E \left(\frac{d}{2} - \sqrt{\xi_d(d - \xi_d)} \right) \\ &= 8E \left(\frac{(\xi_d - d/2)^2}{d} + 8\theta \frac{(\xi_d - d/2)^4}{d^3} \right) = 2 + \frac{\theta_d}{d}, \quad \theta_d \in [1, 6]. \end{aligned}$$

(Область значений θ_d здесь выбрана так, чтобы равенство имело место для всех $d \geq 2$.)

Для доказательства асимптотической нормальности заметим, что

$$\|X + Y\| - \|X - Y\| = 2(\sqrt{\xi_d} - \sqrt{d - \xi_d}) \leq 2\sqrt{d}.$$

Так как функция $u_d(z) = \sqrt{z} - \sqrt{d - z}$ возрастает на отрезке $[0, d]$, то

$$P\{\|X + Y\| - \|X - Y\| \leq t\} = P\{2u_d(\xi_d) \leq t\}.$$

Легко проверить, что

$$k_d(t) \stackrel{\text{def}}{=} \max \left\{ x : u_d(x) \leq \frac{t}{2} \right\} = \frac{d}{2} + \frac{t}{2} \sqrt{\frac{d}{2}} \sqrt{1 - \frac{t^2}{8d}}.$$

Следовательно,

$$\begin{aligned} P\{\|X + Y\| - \|X - Y\| \leq t\} &= P\{\xi_d \leq k_d(t)\} \\ &= P\left\{ \frac{\xi_d - d/2}{\sqrt{d}/2} \leq \frac{t}{\sqrt{2}} \sqrt{1 - \frac{t^2}{8d}} \right\} \rightarrow \Phi\left(\frac{t}{\sqrt{2}}\right), \quad d \rightarrow \infty, \end{aligned}$$

для любого фиксированного $t \in (-\infty, \infty)$ в силу предельной теоремы Муавра–Лапласа. Теорема 1 доказана.

ДОКАЗАТЕЛЬСТВО ТЕОРЕМЫ 2. Для любых $X = (x_1, \dots, x_d), Y = (y_1, \dots, y_d) \in V_d$ справедливы равенства

$$\begin{aligned} \|X + Y\|^2 &= 4\xi_d, \quad \|X - Y\|^2 = 4(d - \xi_d), \\ \|X + Y\| - \|X - Y\| &= 2(\sqrt{\xi_d} - \sqrt{d - \xi_d}) = 2\sqrt{d} u_1\left(\frac{1}{d} \xi_d\right), \end{aligned} \quad (5)$$

где

$$\xi_d \stackrel{\text{def}}{=} \sum_{j=1}^d I\{x_j = y_j\} \quad \text{и} \quad u_1(z) = \sqrt{z} - \sqrt{1 - z}.$$

При условиях теоремы 2 индикаторы $\chi_j = I(x_j = y_j)$, $j = 1, \dots, d$, независимы и

$$P\{\chi_j = 1\} = \frac{1}{2}(1 + (\varepsilon_j^{(d)})^2), \quad j = 1, \dots, d.$$

Поэтому

$$\begin{aligned} \mathbb{E} \frac{1}{d} \xi_d &= \frac{1+a_d}{2}, & \mathbb{D} \frac{1}{d} \xi_d &= \frac{1-b_d}{4d}, \\ \mathbb{E}(\|X+Y\| - \|X-Y\|) &= 2\sqrt{d} \mathbb{E} u_1\left(\frac{1}{d} \xi_d\right). \end{aligned} \quad (6)$$

Так как $|u_1(z)| \leq 1$ при $z \in [0, 1]$ и все производные $u_1(z)$ ограничены на любом интервале $(\varepsilon, 1-\varepsilon)$, $\varepsilon > 0$, то существует такая функция $C(\varepsilon) < \infty$, $\varepsilon \in (0, 1/2)$, что в формуле Тейлора

$$u_1(z) = u_1(s) + u'_1(s)(z-s) + c(s, z)(z-s)^2 \quad (7)$$

коэффициент $c(s, z)$ удовлетворяет неравенству $|c(s, z)| < C(\varepsilon)$ при любых $s \in (\varepsilon, 1-\varepsilon)$, $z \in [0, 1]$. Используя разложение (7) с

$$x = \frac{1}{d} \xi_d, \quad s = \frac{1}{d} \mathbb{E} \xi_d = \frac{1+a_d}{2}$$

и учитывая, что

$$\frac{1}{d} \mathbb{E} \xi_d \in \left[\frac{1}{2}, 1 - \frac{1}{2} \delta \right)$$

по условию теоремы 2, получаем

$$\begin{aligned} & \left| \mathbb{E} u_1\left(\frac{1}{d} \xi_d\right) - u_1\left(\frac{1}{d} \mathbb{E} \xi_d\right) \right| \\ &= \left| \mathbb{E} u'_1\left(\frac{1}{d} \xi_d\right) \left(\frac{1}{d} \xi_d - \frac{1}{d} \mathbb{E} \xi_d\right) + c\left(\frac{1}{d} \mathbb{E} \xi_d, \frac{1}{d} \xi_d\right) \left(\frac{1}{d} \xi_d - \frac{1}{d} \mathbb{E} \xi_d\right)^2 \right| \\ &\leq C\left(\frac{1}{2} \delta\right) \mathbb{D}\left(\frac{1}{d} \xi_d\right) < \frac{1}{4d} C\left(\frac{1}{2} \delta\right). \end{aligned}$$

Используя эту оценку, (6) и равенство

$$u_1\left(\frac{1}{d} \mathbb{E} \xi_d\right) = \sqrt{\frac{1+a_d}{2}} - \sqrt{\frac{1-a_d}{2}} = \frac{a_d \sqrt{2}}{\sqrt{1+a_d} + \sqrt{1-a_d}},$$

получаем соотношение (4):

$$\left| \mathbb{E}(\|X+Y\| - \|X-Y\|) - \frac{2a_d \sqrt{2d}}{\sqrt{1+a_d} + \sqrt{1-a_d}} \right| = O\left(\frac{1}{\sqrt{d}}\right), \quad d \rightarrow \infty.$$

Из соотношений

$$\mathbb{E} \frac{1}{d} \xi_d = \frac{1+a_d}{2}, \quad \mathbb{D} \frac{1}{d} \xi_d = \frac{1-b_d}{4d}, \quad \mathbb{E} |\xi_d - \mathbb{E} \xi_d|^3 = \sum_{j=1}^d \mathbb{E} |\chi_j - \mathbb{E} \chi_j|^3 < \frac{d}{8} \quad (8)$$

и из теоремы Ляпунова следует, что при выполнении условий теоремы 2 случайные величины $(1/d)\xi_d$ при $d \rightarrow \infty$ асимптотически нормальны с параметрами

$$\left(\frac{1}{2}(1+a_d), \frac{1}{4d}(1-b_d) \right).$$

Так как производная функции $u_1(z) = \sqrt{z} - \sqrt{1-z}$ строго положительна и ограничена на отрезке $[1/2, 1-\delta]$, то случайная величина

$$u_1\left(\frac{1}{d}\xi_d\right) = \sqrt{\frac{1}{d}\xi_d} - \sqrt{1 - \frac{1}{d}\xi_d}$$

асимптотически нормальна с параметрами

$$\begin{aligned} & \left(u\left(\frac{1}{d}\mathbb{E}\xi_d\right), \left(u'\left(\frac{1}{d}\mathbb{E}\xi_d\right)\right)^2 \mathbb{D}\frac{1}{d}\xi_d\right) \\ &= \left(\frac{a_d\sqrt{2}}{\sqrt{1-a_d} + \sqrt{1+a_d}}, \frac{1 + \sqrt{1-a_d^2}}{1-a_d^2} \frac{1-b_d}{4d}\right). \end{aligned} \quad (9)$$

Отсюда следует, что случайная величина

$$\|X+Y\| - \|X-Y\| = 2\sqrt{d}u_1\left(\frac{1}{d}\xi_d\right)$$

асимптотически нормальна с параметрами

$$\left(\frac{2a_d\sqrt{2d}}{\sqrt{1-a_d} + \sqrt{1+a_d}}, \frac{1 + \sqrt{1-a_d^2}}{1-a_d^2} (1-b_d)\right).$$

Теорема 2 доказана.

ДОКАЗАТЕЛЬСТВО ТЕОРЕМЫ 3. Между множеством

$$G_d = \left\{ (v_1, \dots, v_d) \in V_d : \sum_{k=0}^d I\{v_k = 1\} \equiv 0 \pmod{2} \right\}$$

и множеством V_{d-1} можно установить однозначное соответствие, сопоставляя вектору $(v_1, \dots, v_{d-1}, v_d) \in G_d$ вектор $(v_1, \dots, v_{d-1}) \in V_{d-1}$. Поэтому если вектор $X = (x_1, \dots, x_d)$ имеет равномерное распределение на G_d , то вектор $(v_1, \dots, v_{d-1})$ имеет равномерное распределение на V_{d-1} , т.е. случайные величины $x_1, \dots, x_{d-1}$ независимы и принимают значения $+1$ и -1 с вероятностями $1/2$, а $x_d = (-1)^{1+x_1+\dots+x_{d-1}}$.

Значит, если случайные векторы X и Y независимы и имеют равномерное распределение на G_d , то

$$\xi_{d-1} = \sum_{j=1}^{d-1} I\{x_j = y_j\}$$

имеет биномиальное распределение с параметрами $(d-1, 1/2)$ и

$$\chi_d \stackrel{\text{def}}{=} I\{x_d = y_d\} = I\{\xi_{d-1} \equiv d-1 \pmod{2}\}.$$

Следовательно,

$$\mathbb{P}\{\xi_{d-1} + \chi_d = v\} = \begin{cases} \frac{1}{2^{d-1}}(C_{d-1}^{v-1} + C_{d-1}^v) = \frac{1}{2^{d-1}} C_{d-1}^v, & v \equiv d \pmod{2}, \\ 0, & v \not\equiv d \pmod{2}; \end{cases} \quad (10)$$

здесь считается, что $C_{d-1}^{-1} = C_{d-1}^d = 0$.

Аналогично (5)

$$\|X + Y\|^2 = 4(\xi_{d-1} + \chi_d), \quad \|X - Y\|^2 = 4(d - (\xi_{d-1} + \chi_d))$$

и отсюда

$$\|X + Y\| - \|X - Y\| = 2(\sqrt{\xi_{d-1} + \chi_d} - \sqrt{d - (\xi_{d-1} + \chi_d)}) = 2\sqrt{d}u_1\left(\frac{1}{d}(\xi_{d-1} + \chi_d)\right).$$

Случайная величина $(1/d)(\xi_{d-1} + \chi_d)$ при $d \rightarrow \infty$ асимптотически нормальна с параметрами $(1/2, 1/(4d))$. Отсюда и из рассуждений в конце доказательства теоремы 2 следует, что разность $\|X + Y\| - \|X - Y\|$ при $d \rightarrow \infty$ асимптотически нормальна с параметрами $(0, 2)$.

Перейдем к оценке $E(\|X + Y\| - \|X - Y\|)$. При четном d равномерное распределение на G_d симметрично, следовательно,

$$E(\|X + Y\| - \|X - Y\|) = 0.$$

Если d нечетно, то согласно (10)

$$\begin{aligned} E(\|X + Y\| - \|X - Y\|) &= \frac{1}{2^{d-2}} \sum_{v=0}^{(d-1)/2} C_d^{2v+1} (\sqrt{2v+1} - \sqrt{d - (2v+1)}) \\ &= \frac{1}{2^{d-2}} \sum_{k=0}^d (-1)^{d-k+1} C_d^k \sqrt{d-k} = \frac{\sqrt{d}}{2^{d-2}} \sum_{k=0}^d (-1)^{d-k+1} C_d^k \sqrt{1 - \frac{k}{d}}. \end{aligned}$$

Так как

$$\sqrt{1-x} = 1 - \frac{x}{2} - \sum_{m=2}^{\infty} \frac{1}{m!} \frac{(2m-2)!}{2^{2m-1}(m-1)!} x^m, \quad |x| < 1,$$

и $\sum_{k=0}^d (-1)^k C_d^k k^n = 0$ при любом $n = 0, 1, \dots, d-1$, то

$$\begin{aligned} E(\|X + Y\| - \|X - Y\|) &= \frac{\sqrt{d}}{2^{d-2}} \sum_{k=0}^d (-1)^{d-k+1} C_d^k \left(1 - \frac{k}{2d} - \sum_{m=2}^{\infty} \frac{1}{m!} \frac{(2m-2)!}{2^{2m-1}(m-1)!} \frac{k^m}{d^m}\right) \\ &= \frac{\sqrt{d}}{2^{d-2}} \sum_{m=d}^{\infty} \frac{1}{m! d^m} \frac{(2m-2)!}{2^{2m-1}(m-1)!} \sum_{k=0}^d (-1)^{d-k} C_d^k k^m. \end{aligned}$$

Сумму $\sum_{k=0}^{d-1} (-1)^{d-k-1} C_d^k k^m$ можно интерпретировать как формулу включения-исключения для количества упорядоченных наборов $(\alpha_1, \dots, \alpha_m) \in \{1, \dots, d\}^m$, не содержащих хотя бы одно из значений $1, \dots, d$. Поэтому

$$0 \leq \sum_{k=0}^{d-1} (-1)^{d-k-1} C_d^k k^m \leq d^m \quad \text{и} \quad 0 \leq \sum_{k=0}^d (-1)^{d-k} C_d^k k^m \leq d^m.$$

Отсюда и из неравенств $1 - 1/(2k) < \sqrt{(k-1/2)/(k+1/2)}$,

$$\frac{(2m-2)!}{2^{2m-1}m!(m-1)!} = \frac{1}{2m} \prod_{k=1}^{m-1} \left(1 - \frac{1}{2k}\right) < \frac{1}{2m} \sqrt{\frac{1}{2m-1}} < \frac{1}{2\sqrt{m-1}} - \frac{1}{2\sqrt{m}}$$

следует, что

$$\sum_{m=d}^{\infty} \frac{1}{m! d^m} \frac{(2m-2)!}{2^{2m-1}(m-1)!} \sum_{k=0}^d (-1)^{k+1} C_d^k k^m \leq \sum_{m=d}^{\infty} \frac{1}{m!} \frac{(2m-2)!}{2^{2m-1}(m-1)!} < \frac{1}{\sqrt{2(d-1)}}.$$

Таким образом,

$$\mathbb{E}(\|X+Y\| - \|X-Y\|) \leq \frac{1}{2^{d-2}} \sqrt{\frac{d}{2(d-1)}} < \frac{1}{2^{d-2}} \quad \text{при } d \geq 2.$$

Чтобы найти расстояние по вариации $\rho(S_d, S_d^*)$, заметим, что если $d = 2r + 1$, то

$$2\rho(S_d, S_d^*) = \sum_{k=0}^{(d-1)/2} |p_{d-1,2k+1} + p_{d-1,2k} - p_{d,2k+1}| + \sum_{k=0}^{(d-1)/2} p_{d,2k} = \frac{1}{2} + \frac{1}{2} = 1,$$

а если $d = 2r$, то $2\rho(S_d, S_d^*) = 1$. Следовательно, $\rho(S_d, S_d^*) = 1/2$, и последнее утверждение теоремы доказано.

ДОКАЗАТЕЛЬСТВО ТЕОРЕМЫ 4. Обозначим $B_d = \{0, 1\}^d$. Тогда распределению $P = \{p_W, W \in V_d\}$ можно сопоставить набор вероятностей

$$\{p_{\mathbf{k}}, \mathbf{k} \in B_d\}, \quad \text{где } p_{\mathbf{k}} = P\{X = 2\mathbf{k} - \mathbf{1}\}, \quad \mathbf{k} \in B_d, \quad X \in V_d, \quad \mathbf{1} = (1, 1, \dots, 1).$$

Пусть $\oplus$ обозначает покоординатное сложение по модулю 2, и $|\mathbf{k}| = |(k_1, \dots, k_d)| = \sum_{i=1}^d k_i$. Если случайные векторы X и Y независимы и имеют распределение $\{p_{\mathbf{k}}, \mathbf{k} \in B_d\}$ на V_d , то

$$\mathbb{E}(\|X+Y\| - \|X-Y\|) = 2 \sum_{\mathbf{k} \in B_d} \sum_{\mathbf{m} \in B_d} p_{\mathbf{k}} p_{\mathbf{m}} (\sqrt{d - |\mathbf{k} \oplus \mathbf{m}|} - \sqrt{|\mathbf{k} \oplus \mathbf{m}|}).$$

Обозначим $P_{\sigma} = \{p_{\sigma(\mathbf{k})}, \mathbf{k} \in B_d\}$, тогда согласно введенным ранее обозначениям:

$$\begin{aligned} & \frac{1}{2^d!} \sum_{\sigma \in \Sigma_d} \mathbb{E}_{\sigma}(\|X+Y\| - \|X-Y\|) \\ &= \frac{2}{2^d!} \sum_{\sigma \in \Sigma_d} \sum_{\mathbf{k} \in B_d} \sum_{\mathbf{m} \in B_d} p_{\sigma(\mathbf{k})} p_{\sigma(\mathbf{m})} (\sqrt{d - |\mathbf{k} \oplus \mathbf{m}|} - \sqrt{|\mathbf{k} \oplus \mathbf{m}|}) \\ &= \frac{2}{2^d!} \sum_{\mathbf{k} \in B_d} \sum_{\mathbf{m} \in B_d} (\sqrt{d - |\mathbf{k} \oplus \mathbf{m}|} - \sqrt{|\mathbf{k} \oplus \mathbf{m}|}) \sum_{\sigma \in \Sigma_d} p_{\sigma(\mathbf{k})} p_{\sigma(\mathbf{m})} \\ &= 2 \frac{2^d(2^d - 1)!}{2^d!} \sum_{\mathbf{k}=\mathbf{m} \in B_d} (\sqrt{d - |\mathbf{k} \oplus \mathbf{m}|} - \sqrt{|\mathbf{k} \oplus \mathbf{m}|}) \sum_{\mathbf{j} \in B_d} p_{\mathbf{j}}^2 \\ & \quad + 2 \frac{2^d(2^d - 2)!}{2^d!} \sum_{\mathbf{k} \neq \mathbf{m} \in B_d} (\sqrt{d - |\mathbf{k} \oplus \mathbf{m}|} - \sqrt{|\mathbf{k} \oplus \mathbf{m}|}) \sum_{\substack{\mathbf{i}, \mathbf{j} \in B_d \\ \mathbf{i} \neq \mathbf{j}}} p_{\mathbf{i}} p_{\mathbf{j}} \\ &= 2\sqrt{d} \sum_{\mathbf{j} \in B_d} p_{\mathbf{j}}^2 + \frac{2}{2^d - 1} \sum_{\mathbf{k} \neq \mathbf{m} \in B_d} (\sqrt{d - |\mathbf{k} \oplus \mathbf{m}|} - \sqrt{|\mathbf{k} \oplus \mathbf{m}|}) \left(1 - \sum_{\mathbf{j} \in B_d} p_{\mathbf{j}}^2\right). \end{aligned}$$

Заметим, что

$$\sum_{\mathbf{k} \in B_d} \sum_{\mathbf{m} \in B_d} (\sqrt{d - |\mathbf{k} \oplus \mathbf{m}|} - \sqrt{|\mathbf{k} \oplus \mathbf{m}|}) = 0.$$

Поэтому

$$\begin{aligned} & \sum_{\mathbf{k} \neq \mathbf{m} \in B_d} (\sqrt{d - |\mathbf{k} \oplus \mathbf{m}|} - \sqrt{|\mathbf{k} \oplus \mathbf{m}|}) \\ &= - \sum_{\mathbf{k} = \mathbf{m} \in B_d} (\sqrt{d - |\mathbf{k} \oplus \mathbf{m}|} - \sqrt{|\mathbf{k} \oplus \mathbf{m}|}) = -2\sqrt{d}. \end{aligned}$$

Таким образом

$$\frac{1}{2^d!} \sum_{\sigma \in \Sigma_d} \mathbb{E}_{\sigma}(\|X + Y\| - \|X - Y\|) = 2\sqrt{d} \sum_{\mathbf{j} \in B_d} p_{\mathbf{j}}^2 - \frac{2\sqrt{d}}{2^d - 1} \left(1 - \sum_{\mathbf{j} \in B_d} p_{\mathbf{j}}^2\right),$$

и теорема 4 доказана.

СПИСОК ЦИТИРОВАННОЙ ЛИТЕРАТУРЫ

- [1] G. J. Székely, T. F. Móri, “A characteristic measure of asymmetry and its application for testing diagonal symmetry”, *Comm. Statist. Theory Methods*, **30**:8-9 (2001), 1633–1639.
- [2] G. G. Gregory, “Large sample theory for U -statistics and tests of fit”, *Ann. Statist.*, **5**:1 (1977), 110–123.
- [3] В. С. Королук, Ю. В. Боровских, *Теория U -статистик*, Наукова Думка, Киев, 1989.
- [4] G. J. Székely, N. K. Bakirov, “Extremal probabilities for Gaussian quadratic forms”, *Probab. Theory Related Fields*, **126**:2 (2003), 184–202.
- [5] W. Hoeffding, “A class of statistics with asymptotically normal distribution”, *Ann. Math. Statist.*, **19**:3 (1948), 293–325.
- [6] D. O. Menshenin, A. M. Zubkov, “On the Székely–Móri asymmetry criterion statistics for binary vectors with independent components”, *Austrian J. Statist.*, **37**:1 (2008), 137–144.

А. М. Зубков

Математический институт им. В. А. Стеклова РАН, г. Москва

E-mail: zubkov@mi.ras.ru

Поступило

17.11.2010

Д. О. Меньшенин

Московский государственный университет

им. М. В. Ломоносова