Math-Net.Ru

Общероссийский математический портал

А. М. Зубков, Новые оценки расстояния по вариации между двумя распределениями выборки, *Матем. вопр. криптогр.*, 2018, том 9, выпуск 3, 45–60

DOI: <https://doi.org/10.4213/mvk262>

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., 16:11:02

Новые оценки расстояния по вариации между двумя распределениями выборки

А. М. Зубков

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

Получено 11.V.2017

Аннотация. Для пары выборок, состоящих из независимых одинаково распределенных случайных величин, получены новые двусторонние неравенства для расстояний по вариации между их распределениями. В случаях когда эти расстояния малы, оценки пропорциональны расстоянию по вариации между распределениями элементов выборок и квадратному корню из объема выборки; коэффициенты пропорциональности зависят от распределений элементов выборок. Результаты представляют интерес для получения оценок объема выборки, необходимого для различения двух близких статистических гипотез.

Ключевые слова: расстояние по вариации, однородные выборки, вероятностные неравенства, двусторонние оценки

New estimates for the variational distance between two distributions of a sample

A. M. Zubkov

Steklov Mathematical Institute of Russian Academy of Sciences, Moscow

Abstract. For a pair of samples consisting of i.i.d. random variables we obtain new upper and lower bounds for the total variation distance between their distributions. In cases of small distances these estimates are proportional to the total variation distance between distributions of elements of samples and the square root of the sample volume with coefficients depending on the distributions of elements of samples. The results are useful for estimates of the sample volume necessary for testing close hypotheses.

Key words: total variation distance, homogeneous samples, probabilistic inequalities, two-side estimates

Citation: *Mathematical Aspects of Cryptography*, 2018, v. 9, № 3, pp. 45–60 (Russian)

© Академия криптографии Российской Федерации, 2018 г.

Статистические методы криптографического анализа в большинстве случаев связаны с задачами проверки статистических гипотез о распределениях наблюдаемых последовательностей. Как правило, при этом предполагается, что наблюдения независимы и одинаково распределены. Для исследования таких ситуаций актуальной является задача получения теоретических оценок объема выборки, при котором возможно построение статистических критериев различения близких статистических гипотез с заданными вероятностями ошибок.

Для конкретных статистических критериев оценка необходимого объема выборки сводится к получению асимптотических формул для распределений статистик при разных гипотезах, но при этом остается нерешенным вопрос о возможности построения более эффективных критериев. Мерой принципиальной возможности различения гипотез с заданными вероятностями ошибок является расстояние по вариации. С ростом объема выборки n расстояние по вариации между возможными распределениями выборки увеличивается и стремится к 1 при $n \rightarrow \infty$. Поэтому нетривиальные теоретические оценки расстояния по вариации позволяют делать объективные выводы об объеме выборки, необходимом или достаточном для различения близких гипотез.

В последнее время верхние и нижние оценки скорости роста расстояния по вариации при увеличении объема n выборки были получены в [3–5]; порядком роста этих оценок при некоторых условиях является $\sqrt{n}$, что согласуется с результатами, полученными для конкретных критериев. Однако ситуация не вполне простая, так как универсальная верхняя оценка расстояния по вариации между распределениями выборки объема n линейна по n (это следует из неравенства треугольника) и в схемах серий эта оценка неулучшаема.

Основные результаты препринтов [3, 4] приведены ниже в теоремах 1 и 3. Они послужили исходными точками для поиска новых оценок расстояния по вариации, содержащихся в теоремах 2 и 4. Верхние оценки расстояния по вариации в [5] охватывают случай, когда элементы выборок могут быть зависимыми, и имеют более сложный вид.

Основные свойства расстояния по вариации

Для случайных величин X и Y , принимающих значения в одном и том же пространстве B с σ -алгеброй $\mathcal{B}$ измеримых множеств, расстояние по вариации между их распределениями P_X и P_Y определяется равенством

$$\rho(X, Y) = \rho(P_X, P_Y) = \sup_{A \in \mathcal{B}} |P_X(A) - P_Y(A)|.$$

Лемма 1 (см. [1, 6, 7]). Если X и Y — случайные величины с не более чем счетным множеством значений B , то

$$\rho(X, Y) = \frac{1}{2} \sum_{b \in B} |\mathbf{P}\{X=b\} - \mathbf{P}\{Y=b\}|.$$

Лемма 2 (см. [1, 6, 7]). Пусть $X_1, \dots, X_n, Y_1, \dots, Y_n$ — независимые случайные величины, принимающие значения в одном и том же пространстве B , и $g(x_1, \dots, x_n)$ — измеримая функция, определенная на B^n . Тогда

$$\rho(g(X_1, \dots, X_n), g(Y_1, \dots, Y_n)) \leq \rho(X_1, Y_1) + \dots + \rho(X_n, Y_n).$$

Лемма 3 (см. [6, 7]). Пусть $X, X_1, \dots, X_n$ — независимые одинаково распределенные случайные величины, $Y, Y_1, \dots, Y_n$ — тоже независимые одинаково распределенные случайные величины и $\rho(X, Y) = \varepsilon$. Тогда

$$n\varepsilon \geq \rho((X_1, \dots, X_n), (Y_1, \dots, Y_n)) \geq 1 - 2e^{-n\varepsilon^2/2}.$$

Обе оценки в лемме 3 следуют из леммы 2, при доказательстве нижней оценки используется также неравенство Чернова. Нижняя оценка содержательна только при $n\varepsilon^2 > 2 \ln 2 \approx 1,3863$.

Расстояние по вариации является естественной мерой возможности различения простых статистических гипотез. Пусть по наблюдению случайной величины Z , принимающей значения в измеримом пространстве $(B, \mathcal{B})$, нужно различить две простые гипотезы:

H_0 : Z имеет распределение P_0 ,

H_1 : Z имеет распределение P_1 ,

с ошибками первого и второго родов, не превосходящими α и β соответственно. (Случай $B = \mathbb{R}$ соответствует проверке статистических гипотез по одному наблюдению, случай $B = \mathbb{R}^n$ можно интерпретировать как проверку гипотез по выборке объема n .)

Допустим, что решение определяется детерминированной функцией от Z : выбрано некоторое множество $C \subseteq B$ и при $Z \in C$ принимается гипотеза H_0 , а при $Z \notin C$ принимается гипотеза H_1 . Тогда вероятности ошибок первого и второго родов равны соответственно

$$\alpha = \mathbf{P}_0\{Z \notin C\} \quad \text{и} \quad \beta = \mathbf{P}_1\{Z \in C\}.$$

Из определений вероятностей ошибок и расстояния по вариации следует, что

$$P_0\{Z \in C\} - P_1\{Z \in C\} = 1 - \alpha - \beta \leq \rho(P_0, P_1).$$

Значит,

$$\alpha + \beta \geq 1 - \rho(P_0, P_1).$$

Таким образом, если расстояние по вариации между P_0 и P_1 мало, то сумма вероятностей ошибок при различении двух простых гипотез не может быть малой, т. е. не существует статистических критериев, которые различают гипотезы, близкие в смысле расстояния по вариации.

Обратно, если расстояние по вариации между мерами P_0 и P_1 равно Δ , то по теореме Хана существует такое измеримое множество $C \subseteq B$, что $P_0(D) > P_1(D)$ для любого измеримого $D \subseteq C$ и $P_0(D) \leq P_1(D)$ для любого измеримого $D \subseteq B \setminus C$. Тогда

$$P_0(C) - P_1(C) = P_1(B \setminus C) - P_0(B \setminus C) = \Delta.$$

Поэтому если принимать гипотезу H_0 тогда и только тогда, когда $Z \in C$, то сумма вероятностей ошибок первого и второго родов будет равна

$$P_0\{Z \notin C\} + P_1\{Z \in C\} = 1 - P_0\{Z \in C\} + P_1\{Z \in C\} = 1 - \Delta.$$

Неравенства для расстояния по вариации между двумя распределениями выборки

Пусть $X_1, \dots, X_n$ — независимые случайные величины с одним и тем же распределением P , а $Y_1, \dots, Y_n$ — независимые случайные величины с одним и тем же распределением Q . Совместные распределения $X_1, \dots, X_n$ и $Y_1, \dots, Y_n$ будем обозначать соответственно P^n и Q^n . Из леммы 2 следует, что для любой измеримой функции $g(\cdot)$

$$\rho(g(X_1, \dots, X_n), g(Y_1, \dots, Y_n)) \leq \rho(P^n, Q^n) \leq n\rho(X_1, Y_1).$$

В схеме серий эта тривиальная оценка может быть близка к точной. Например, пусть $P(0) = 1$, $Q(1) = \varepsilon > 0$, $Q(0) = 1 - \varepsilon$. Тогда $\rho(P, Q) = \varepsilon$,

$$P^n((0, \dots, 0)) = 1, \quad P^n(x) = 0 \quad \text{для всех } x \in \{0, 1\}^n \setminus \{(0, \dots, 0)\}$$

и

$$Q^n((0, \dots, 0)) = (1 - \varepsilon)^n, \quad Q^n(x) > P^n(x) \quad \text{для всех } x \in \{0, 1\}^n \setminus \{(0, \dots, 0)\}.$$

По лемме 1

$$\begin{aligned}\rho(P^n, Q^n) &= \frac{1}{2} \sum_{(z_1, \dots, z_n) \in \{0,1\}^n} |P^n(z_1, \dots, z_n) - Q^n(z_1, \dots, z_n)| = \\ &= \frac{1}{2} \left(P^n((0, \dots, 0)) - Q^n(0, \dots, 0) + \sum_{(z_1, \dots, z_n) \neq (0, \dots, 0)} Q^n(z_1, \dots, z_n) \right) = \\ &= 1 - (1 - \varepsilon)^n > \\ &> 1 - e^{-n\varepsilon} = n\varepsilon (1 + O(n\varepsilon)) \quad \text{при } n\varepsilon \rightarrow 0.\end{aligned}$$

В [3] приводится следующая нижняя оценка для $\rho(P^n, Q^n)$.

Теорема 1. Пусть $X, X_1, X_2, \dots$ — независимые целочисленные одинаково распределенные случайные величины и $Y, Y_1, Y_2, \dots$ — тоже независимые целочисленные одинаково распределенные случайные величины, $\rho(X, Y) = \varepsilon > 0$ и

$$\max_x \mathbf{P}\{X = x\} = m_X, \quad \max_x \mathbf{P}\{Y = x\} = m_Y.$$

Если $\max\{m_X, m_Y\} \leq \varepsilon/2$, то

$$\rho((X_1, \dots, X_n), (Y_1, \dots, Y_n)) > \left(\sqrt{\frac{n}{2\pi}} - \frac{1}{\sqrt{8n\pi}} \right) \varepsilon - \frac{3}{8} n\varepsilon^2 - \frac{3}{32} n^2 \varepsilon^3.$$

Если $n^{3/4}\varepsilon \rightarrow 0$, то

$$\rho((X_1, \dots, X_n), (Y_1, \dots, Y_n)) > C\varepsilon\sqrt{n}$$

при некотором $C \in (0, \infty)$.

Замечание 1. Из условия $\max\{m_X, m_Y\} \leq \varepsilon/2$ следует, что хотя бы для одного из распределений число исходов с положительными вероятностями должно быть не меньше $2/\varepsilon$. В случае когда условие $\max\{m_X, m_Y\} \leq \varepsilon/2$ не выполняется, в [3] предлагается объединять случайные величины X_i и Y_i в группы вида

$$\begin{aligned}X^{(t)} &= (X_{(t-1)k+1}, X_{(t-1)k+2}, \dots, X_{tk}), \\ Y^{(t)} &= (Y_{(t-1)k+1}, Y_{(t-1)k+2}, \dots, Y_{tk}), \quad t = 1, 2, \dots,\end{aligned}$$

выбирая k таким образом, чтобы для случайных векторов $X^{(t)}, Y^{(t)}$ условие $\max\{m_{X^{(t)}}, m_{Y^{(t)}}\} \leq \varepsilon/2$ выполнялось; при этом, очевидно, $\rho(X, Y) \leq \rho(X^{(1)}, Y^{(1)})$, но в окончательной оценке придется заменить n на n/k .

Замечание 2. Нижняя оценка из теоремы 1 заведомо становится отрицательной, если $n > 3\varepsilon^{-4/3}$, так как тогда

$$\frac{3}{32} n^2 \varepsilon^3 > \frac{3}{32} 3^{3/2} \sqrt{n} \varepsilon > \frac{15}{32} \sqrt{n} \varepsilon > \sqrt{\frac{n}{2\pi}} \varepsilon,$$

но даже при $n = 3\varepsilon^{-4/3}$ единственное положительное слагаемое в нижней оценке невелико:

$$\sqrt{\frac{n}{2\pi}} \varepsilon = \sqrt{\frac{3\varepsilon^{-4/3}}{2\pi}} \varepsilon = \sqrt{\frac{3}{2\pi}} \sqrt[3]{\varepsilon},$$

т. е. при малых ε нижняя оценка теоремы 1 не может существенно отличаться от 0 ни при каких n .

Как уже отмечалось, для вероятностных мер P и Q на одном и том же измеримом множестве Ω с $\rho(P, Q) = \varepsilon$ по определению расстояния по вариации существует такое множество $C \subset \Omega$, что

$$P(C) - Q(C) = Q(\Omega \setminus C) - P(\Omega \setminus C) = \varepsilon \quad (1)$$

и $P(B) \geq Q(B)$ для любого измеримого $B \subseteq C$, $P(B) \leq Q(B)$ для любого измеримого $B \subseteq \Omega \setminus C$. (В случае дискретного пространства Ω можно положить $C = \{\omega \in \Omega: P(\omega) > Q(\omega)\}$.) Используя это множество C , введем величину

$$v = v(P, Q) = \min \{P(C), Q(C), P(\Omega \setminus C), Q(\Omega \setminus C)\},$$

тогда $v \leq (1-\varepsilon)/2$, так как если, например, $v = P(C) > (1-\varepsilon)/2$, то

$$Q(C) = v + \varepsilon, \quad Q(\Omega \setminus C) = 1 - v - \varepsilon < \frac{1}{2} (1 - \varepsilon).$$

Если обе величины $v = v(P, Q)$ и $\varepsilon = \rho(P, Q)$ малы, то одна из мер P, Q мажорирует другую на множестве меры, близкой к 1. Действительно, если, например, $\rho(P, Q) = Q(C) - P(C) = \varepsilon$ и $v = Q(C)$, то $P(C) = v + \varepsilon$, $P(\Omega \setminus C) = 1 - v - \varepsilon$, $Q(\Omega \setminus C) = 1 - v$, т. е. мера Q мажорирует меру P на множестве $\Omega \setminus C$, у которого P -мера и Q -мера близки к 1.

Теорема 2 содержит нижнюю оценку расстояния по вариации, отличную от оценки теоремы 1.

Теорема 2. Пусть $X_1, X_2, \dots$ — независимые целочисленные случайные величины с одним и тем же распределением P и $Y_1, X_2, \dots$ — независимые целочисленные случайные величины с одним и тем же распределением Q , $\rho(P, Q) = \varepsilon > 0$. Справедливо неравенство

$$\begin{aligned} \rho(P^n, Q^n) &= \rho((X_1, \dots, X_n), (Y_1, \dots, Y_n)) \geq \\ &\geq \frac{e^{-4/(nv(1-v))}}{2(1+2/n)} \sqrt{\frac{v}{1-v}} \left(\Phi(2\varepsilon\sqrt{n}) - \frac{1}{2} \right), \end{aligned}$$

где $v = v(P, Q)$, $2 \leq nv < (n-1)/2$ и

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2} du$$

— функция стандартного нормального распределения.

Замечание 3. Разность $\Phi(2\varepsilon\sqrt{n}) - 1/2$ не превышает $1/2$ и эквивалентна $\varepsilon\sqrt{2n/\pi}$, если $\varepsilon\sqrt{n} \rightarrow 0$, т.е. нижняя оценка при $n \rightarrow \infty$ в области своих малых значений растет по порядку как $\sqrt{n}$ и принимает значения порядка $\sqrt{v}$ при $n \sim 1/(4\varepsilon^2)$. Согласно лемме 3 дальнейшее увеличение n в соответствующее число раз сделает нижнюю оценку сколь угодно близкой к 1.

Доказательство. Пусть множество C — то же, что в (1). Не ограничивая общности, будем считать, что

$$\mathbf{P}\{X \in C\} = v, \quad \mathbf{P}\{Y \in C\} = v + \varepsilon$$

при некотором $v \in [0, (1-\varepsilon)/2]$.

Для фиксированного $n \geq 1$ рассмотрим функцию

$$f(x_1, \dots, x_n) = \sum_{i=1}^n \mathbb{I}\{x_i \in C\}.$$

Тогда $f(X_1, \dots, X_n)$ имеет биномиальное распределение $\text{Bin}(n, v)$, а $f(Y_1, \dots, Y_n)$ — биномиальное распределение $\text{Bin}(n, v + \varepsilon)$.

По лемме 2

$$\begin{aligned}
 \rho(P^n, Q^n) &= \rho((X_1, \dots, X_n), (Y_1, \dots, Y_n)) \geq \\
 &\geq \rho(f(X_1, \dots, X_n), f(Y_1, \dots, Y_n)) = \\
 &= \rho(\text{Bin}(n, v), \text{Bin}(n, v + \varepsilon)) = \\
 &= \frac{1}{2} \sum_{k=0}^n \left| C_n^k v^k (1-v)^{n-k} - C_n^k (v + \varepsilon)^k (1-v-\varepsilon)^{n-k} \right|.
 \end{aligned}$$

Так как отношение

$$\frac{C_n^k v^k (1-v)^{n-k}}{C_n^k (v + \varepsilon)^k (1-v-\varepsilon)^{n-k}} = \left(1 - \frac{\varepsilon}{(1-v)(v + \varepsilon)} \right)^k \left(\frac{1-v}{1-v-\varepsilon} \right)^n$$

монотонно убывает с ростом k и

$$\begin{aligned}
 &\left(1 - \frac{\varepsilon}{(1-v)(v + \varepsilon)} \right)^{nv} \left(\frac{1-v}{1-v-\varepsilon} \right)^n = \left(\frac{v^v (1-v)^{1-v}}{(v + \varepsilon)^v (1-v-\varepsilon)^{1-v}} \right)^n = \\
 &= \exp \left\{ n \left(-v \ln \left(1 + \frac{\varepsilon}{v} \right) - (1-v) \ln \left(1 - \frac{\varepsilon}{1-v} \right) \right) \right\} > \\
 &> \exp \left\{ n \left(-v \frac{\varepsilon}{v} - (1-v) \left(-\frac{\varepsilon}{1-v} \right) \right) \right\} = e^0 = 1,
 \end{aligned}$$

то

$$C_n^k v^k (1-v)^{n-k} > C_n^k (v + \varepsilon)^k (1-v-\varepsilon)^{n-k} \quad \text{при } 0 \leq k \leq nv.$$

Поэтому, используя обозначение $\beta_{n,q}$ для случайной величины с распределением $\text{Bin}(n, q)$, $q \in (0, 1)$, находим:

$$\begin{aligned}
 \rho(P^n, Q^n) &\geq \rho(\text{Bin}(n, v), \text{Bin}(n, v + \varepsilon)) = \\
 &= \frac{1}{2} \sum_{0 \leq k \leq n} \left| C_n^k v^k (1-v)^{n-k} - C_n^k (v + \varepsilon)^k (1-v-\varepsilon)^{n-k} \right| \geq \\
 &\geq \sum_{0 \leq k \leq nv} \left(C_n^k v^k (1-v)^{n-k} - C_n^k (v + \varepsilon)^k (1-v-\varepsilon)^{n-k} \right) = \\
 &= \mathbf{P}\{\beta_{n,v} \leq nv\} - \mathbf{P}\{\beta_{n,v+\varepsilon} \leq nv\}.
 \end{aligned} \tag{2}$$

Из известного тождества

$$\mathbf{P}\{\beta_{n,v} \leq k\} = (n-k) C_n^k \int_v^1 z^k (1-z)^{n-k-1} dz, \quad k = 0, \dots, n-1,$$

и (2) следует, что

$$\begin{aligned} \rho(P^n, Q^n) &\geq (n - \lfloor nv \rfloor) C_n^{\lfloor nv \rfloor} \left(\int_v^1 - \int_{v+\varepsilon}^1 \right) z^{\lfloor nv \rfloor} (1-z)^{n-\lfloor nv \rfloor-1} dz = \\ &= (n - \lfloor nv \rfloor) C_n^{\lfloor nv \rfloor} \int_v^{v+\varepsilon} z^{\lfloor nv \rfloor} (1-z)^{n-\lfloor nv \rfloor-1} dz \geq \\ &\geq n(1-v) C_n^{\lfloor nv \rfloor} \int_v^{v+\varepsilon} z^{nv} (1-z)^{n(1-v)} dz = \\ &= n(1-v) C_n^{\lfloor nv \rfloor} v^{nv} (1-v)^{n(1-v)} \int_v^{v+\varepsilon} \left(\frac{z}{v}\right)^{nv} \left(\frac{1-z}{1-v}\right)^{n(1-v)} dz. \end{aligned} \quad (3)$$

Покажем, что при $z \in [v, v+\varepsilon]$

$$\left(\frac{z}{v}\right)^{nv} \left(\frac{1-z}{1-v}\right)^{n(1-v)} \geq e^{-2n(z-v)^2}. \quad (4)$$

Действительно, при $z = v$ левая и правая части (4) равны 1, а при $z \in [v, v+\varepsilon]$ производная логарифма левой части не меньше производной логарифма правой части:

$$\begin{aligned} &\frac{d}{dz} \ln \left(\left(\frac{z}{v}\right)^{nv} \left(\frac{1-z}{1-v}\right)^{n(1-v)} \right) = \\ &= n \frac{d}{dz} \left(v \ln \frac{z}{v} + (1-v) \ln \frac{1-z}{1-v} \right) = \\ &= n \left(\frac{v}{z} - \frac{1-v}{1-z} \right) = n \frac{v-z}{z(1-z)} \geq \\ &\geq 4n(v-z), \\ &\frac{d}{dz} \ln \left(e^{-2n(z-v)^2} \right) = n \frac{d}{dz} (-2(z-v)^2) = \\ &= 4n(v-z). \end{aligned}$$

Используя замену $x = 2(z - v)\sqrt{n}$, получаем: при $v \leq (1 - \varepsilon)/2$

$$\begin{aligned} & \int_v^{v+\varepsilon} \left(\frac{z}{v}\right)^{nv} \left(\frac{1-z}{1-v}\right)^{n(1-v)} dz \geq \\ & \geq \int_v^{v+\varepsilon} e^{-2n(z-v)^2} dz = \frac{\sqrt{2\pi}}{2\sqrt{2\pi n}} \int_0^{2\varepsilon\sqrt{n}} e^{-x^2/2} dx = \\ & = \sqrt{\frac{\pi}{2n}} \left(\Phi(2\varepsilon\sqrt{n}) - \frac{1}{2} \right). \end{aligned}$$

Отсюда и из (3) следует, что

$$\rho(P^n, Q^n) \geq \sqrt{\frac{\pi n}{2}} (1-v) C_n^{\lfloor nv \rfloor} v^{nv} (1-v)^{n(1-v)} \left(\Phi(2\varepsilon\sqrt{n}) - \frac{1}{2} \right). \quad (5)$$

В [8, теорема 3.1] получены нижние оценки биномиальных вероятностей в виде приращений функции распределения стандартного нормального закона: если $2 \leq nv \leq k \leq n(1-v)$, то

$$\begin{aligned} C_n^k v^k (1-v)^{n-k} & \geq \Phi\left(\frac{k-nv+1}{\sqrt{nv(1-v)}}\right) - \Phi\left(\frac{k-nv}{\sqrt{nv(1-v)}}\right) = \\ & = \frac{1}{\sqrt{2\pi}} \int_{\frac{k-nv}{\sqrt{nv(1-v)}}}^{\frac{k-nv+1}{\sqrt{nv(1-v)}}} e^{-u^2/2} du \geq \\ & \geq \frac{1}{\sqrt{2\pi nv(1-v)}} \exp\left\{-\frac{(k-nv+1)^2}{nv(1-v)}\right\}. \end{aligned} \quad (6)$$

Поэтому при

$$2 \leq nv < \frac{n-1}{2} \quad \text{и} \quad k = \lceil nv \rceil \in [nv, n(1-v)]$$

из (6) следует, что

$$C_n^{\lceil nv \rceil} v^{\lceil nv \rceil} (1-v)^{n-\lceil nv \rceil} \geq \frac{e^{-4/(nv(1-v))}}{\sqrt{2\pi nv(1-v)}}.$$

Так как

$$C_n^{\lfloor nv \rfloor} v^{nv} (1-v)^{n(1-v)} = C_n^{\lceil nv \rceil} v^{\lceil nv \rceil} (1-v)^{n-\lceil nv \rceil}$$

при целых значениях nv , а при

$$2 \leq nv = \lfloor nv \rfloor + x, \quad x \in (0, 1), \quad v < \frac{1}{2},$$

имеем

$$\begin{aligned} \frac{C_n^{\lfloor nv \rfloor} v^{nv} (1-v)^{n(1-v)}}{C_n^{\lceil nv \rceil} v^{\lceil nv \rceil} (1-v)^{n-\lceil nv \rceil}} &= \frac{(\lfloor nv \rfloor + 1)(1-v)^{1-x}}{(n - \lfloor nv \rfloor) v^{1-x}} > \\ &> \frac{nv}{n(1-v) + x} = \frac{v}{1-v + x/n} > \\ &> \frac{v}{(1-v)(1 + \frac{1}{n(1-v)})} > \\ &> \frac{v}{1-v} \frac{1}{1 + 2/n} = \frac{v}{1-v} \frac{n}{n+2}, \end{aligned}$$

то

$$C_n^{\lfloor nv \rfloor} v^{nv} (1-v)^{n(1-v)} > \frac{v}{1-v} \frac{n}{n+2} \frac{e^{-4/(nv(1-v))}}{\sqrt{2\pi nv(1-v)}}.$$

Подставляя эту оценку в (5), завершаем доказательство теоремы:

$$\begin{aligned} \rho(P^n, Q^n) &\geq \sqrt{\frac{\pi n}{2}} (1-v) \frac{v}{1-v} \frac{n}{n+2} \frac{e^{-4/(nv(1-v))}}{\sqrt{2\pi nv(1-v)}} \left(\Phi(2\varepsilon\sqrt{n}) - \frac{1}{2} \right) = \\ &= \frac{e^{-4/(nv(1-v))}}{2(1 + 2/n)} \sqrt{\frac{v}{1-v}} \left(\Phi(2\varepsilon\sqrt{n}) - \frac{1}{2} \right). \end{aligned}$$

Теорема 2 доказана. $\square$

В [4] было доказано следующее утверждение, показывающее, что расстояние по вариации между n -мерными векторами с независимыми одинаково распределенными целочисленными координатами растет по порядку не быстрее, чем $\sqrt{n}$, если все вероятности отделены от 0.

Теорема 3. Пусть P и Q — два вероятностных распределения на $\mathbb{Z}$, которые различаются на множестве $\mathcal{D}(P, Q) = \{z \in \mathbb{Z} : P(z) \neq Q(z)\}$, и $p_* = \inf_{z \in \mathcal{D}(P, Q)} \min\{P(z), Q(z)\}$. Если $p_* > 0$, то при любом $n \in \mathbb{N}$

$$\rho(P^n, Q^n) \leq \sqrt{\frac{1}{\pi p_*}} \sqrt{n + \frac{1}{p_*}} \rho(P, Q)$$

и

$$\rho(P^n, Q^n) \leq \sqrt{\frac{n}{2p_*}} \rho(P, Q).$$

Основным недостатком оценок теоремы 3 является наличие числа p_* в знаменателях. Это не только делает оценки бессодержательными в случаях распределений с нулевыми вероятностями, но даже для почти равновероятных распределений на конечных множествах большой мощности существенно увеличивает числовой коэффициент в правых частях оценок. В теореме 4 приводится более точная оценка.

Теорема 4. Пусть $X_1, X_2, \dots$ и $Y_1, Y_2, \dots$ — независимые случайные величины, принимающие значения $1, \dots, N$:

$$\mathbf{P}\{X_t = k\} = p_k, \quad \mathbf{P}\{Y_t = k\} = r_k, \quad k \in \{1, \dots, N\}, \quad t = 1, 2, \dots,$$

и $\rho(X_1, Y_1) = \frac{1}{2} \sum_{j=1}^N |p_j - r_j|$. Тогда

$$\rho((X_1, \dots, X_n), (Y_1, \dots, Y_n)) \leq \rho \sqrt{n} \left(\frac{1}{\sqrt{P} + \sqrt{P+\rho}} + \frac{1}{\sqrt{R} + \sqrt{R+\rho}} \right),$$

где $P = \sum_{j=1}^N p_j \mathbb{I}\{p_j < r_j\}$, $R = \sum_{j=1}^N r_j \mathbb{I}\{p_j > r_j\}$.

Замечание 4. Набор чисел $P, P + \rho, R, R + \rho$ аналогичен набору чисел $P(C), Q(C), P(\Omega \setminus C), Q(\Omega \setminus C)$ в определении величины $v(P, Q)$, использованной в теореме 2.

Замечание 5. Одним из промежуточных результатов было неравенство [9]

$$\rho((X_1, \dots, X_n), (Y_1, \dots, Y_n)) \leq \sqrt{n} \sum_{k=1}^N |\sqrt{p_k} - \sqrt{r_k}|,$$

которое доказывается с помощью того же подхода, что в теореме 4, но короче и проще. Его недостатком является то, что даже при малой величине $\rho(X_1, Y_1)$ сумма в правой части может принимать сколь угодно большие значения при достаточно больших N .

Доказательство. По определению

$$\begin{aligned} \rho((X_1, \dots, X_n), (Y_1, \dots, Y_n)) &= \frac{1}{2} \sum_{k_1, \dots, k_n=1}^N \left| \prod_{j=1}^n p_{k_j} - \prod_{j=1}^n r_{k_j} \right| = \\ &= \frac{1}{2} \sum_{\substack{m_1, \dots, m_N \geq 0 \\ m_1 + \dots + m_N = n}} \frac{n!}{m_1! \dots m_N!} \left| \prod_{k=1}^N p_k^{m_k} - \prod_{k=1}^N r_k^{m_k} \right|. \end{aligned}$$

Далее,

$$\frac{d}{dt} \prod_{k=1}^N (x_k + \theta_k t)^{m_k} = \sum_{j=1}^N \frac{m_j \theta_j}{x_j + \theta_j t} \prod_{k=1}^N (x_k + \theta_k t)^{m_k}.$$

Поэтому при $\theta_j = r_j - p_j$, $j = 1, \dots, N$,

$$\begin{aligned} \left| \prod_{k=1}^N p_k^{m_k} - \prod_{k=1}^N r_k^{m_k} \right| &= \left| \int_0^1 \sum_{j=1}^N \frac{m_j \theta_j}{p_j + \theta_j t} \prod_{k=1}^N (p_k + \theta_k t)^{m_k} dt \right| \leq \\ &\leq \int_0^1 \left| \sum_{j=1}^N \frac{m_j \theta_j}{p_j + \theta_j t} \right| \prod_{k=1}^N (p_k + \theta_k t)^{m_k} dt. \end{aligned}$$

Значит,

$$\begin{aligned} \rho((X_1, \dots, X_n), (Y_1, \dots, Y_n)) &\leq \\ &\leq \frac{1}{2} \int_0^1 \sum_{\substack{m_1, \dots, m_N \geq 0 \\ m_1 + \dots + m_N = n}} \left| \sum_{j=1}^N \frac{m_j \theta_j}{p_j + \theta_j t} \right| \frac{n!}{m_1! \dots m_N!} \prod_{k=1}^N (p_k + \theta_k t)^{m_k} dt. \end{aligned} \quad (7)$$

Выражение под знаком интеграла представляет собой математическое ожидание случайной величины

$$\left| \sum_{j=1}^N \frac{\nu_j \theta_j}{p_j + \theta_j t} \right|,$$

здесь вектор $(\nu_1, \dots, \nu_N)$, составленный имеющими биномиальные распределения случайными величинами $\nu_j \sim \text{Bin}(n, p_j + \theta_j t)$, $j = 1, \dots, N$, с $\mathbf{M} \nu_j = n(p_j + \theta_j t)$, $\mathbf{D} \nu_j = n(p_j + \theta_j t)(1 - p_j - \theta_j t)$, имеет полиномиальное распределение $\text{Poly}(n; p_1 + \theta_1 t, \dots, p_N + \theta_N t)$.

Так как по определению $\theta_j = r_j - p_j$, $j = 1, \dots, N$, то

$$\mathbf{M} \sum_{j=1}^N \frac{\nu_j \theta_j}{p_j + \theta_j t} = \sum_{j=1}^N \frac{\mathbf{M} \nu_j \theta_j}{p_j + \theta_j t} = \sum_{j=1}^N n \theta_j = 0.$$

Далее, $\text{cov}(\nu_i, \nu_j) = -n(p_i + \theta_i t)(p_j + \theta_j t) \leq 0$ при $i \neq j$, поэтому

$$\begin{aligned} \mathbf{M} \left| \sum_{j=1}^N \frac{\nu_j \theta_j}{p_j + \theta_j t} \right| &\leq \sqrt{\mathbf{M} \left(\sum_{j=1}^N \frac{\nu_j \theta_j}{p_j + \theta_j t} \right)^2} \leq \\ &\leq \sqrt{\sum_{j=1}^N \mathbf{D} \frac{\nu_j \theta_j}{p_j + \theta_j t}} = \sqrt{\sum_{j=1}^N \frac{n(p_j + \theta_j t)(1 - p_j - \theta_j t) \theta_j^2}{(p_j + \theta_j t)^2}} \leq \\ &\leq \sqrt{n \sum_{j=1}^N \frac{\theta_j^2}{p_j + \theta_j t}}. \end{aligned} \quad (8)$$

Чтобы оценить сверху сумму в правой части, будем считать, не ограничивая общности, что при некоторых $1 \leq w < M \leq N$

$$\theta_j > 0 \quad (1 \leq j \leq w), \quad \theta_j < 0 \quad (w < j \leq M), \quad \theta_j = 0 \quad (M < j \leq N);$$

тогда

$$\sum_{j=1}^w \theta_j = - \sum_{j=w+1}^M \theta_j = \frac{1}{2} \sum_{j=1}^N |\theta_j| = \rho(\mathbf{p}, \mathbf{r}) = \rho$$

и

$$\sum_{j=1}^N \frac{\theta_j^2}{p_j + \theta_j t} = \sum_{j=1}^w \frac{\theta_j^2}{p_j + \theta_j t} + \sum_{j=w+1}^M \frac{\theta_j^2}{p_j + \theta_j t} = \sum_{j=1}^w \frac{\theta_j^2}{p_j + \theta_j t} + \sum_{j=w+1}^M \frac{(-\theta_j)^2}{r_j + (-\theta_j)t}. \quad (9)$$

Считая $p_1, \dots, p_N$ и $t \in [0, 1]$ фиксированными, найдем максимальное значение функции

$$\mathcal{L}(\theta_1, \dots, \theta_w) = \sum_{j=1}^w \frac{\theta_j^2}{p_j + \theta_j t}$$

при условии, что $\sum_{j=1}^w \theta_j = \rho < 1$ и $\theta_1, \dots, \theta_w \geq 0$. Система уравнений Лагранжа для точки экстремума функции $\mathcal{L}(\theta_1, \dots, \theta_w)$ при условии $\sum_{j=1}^w \theta_j = \rho$ имеет вид

$$\frac{\partial}{\partial \theta_j} \left(\sum_{j=1}^w \frac{\theta_j^2}{p_j + \theta_j t} - \lambda \sum_{j=1}^w \theta_j \right) = \frac{\theta_j(2p_j + \theta_j t)}{(p_j + \theta_j t)^2} - \lambda = 0, \quad j = 1, \dots, w.$$

Переписывая эти уравнения в виде $t + 2p_j/\theta_j = \lambda(t + p_j/\theta_j)^2$, находим, что отношение $p_j/\theta_j \geq 0$ при любом $j = 1, \dots, w$ является корнем квадратного уравнения $t + 2x = \lambda(t + x)^2$ и поэтому может принимать не более двух значений, т. е. либо $\theta_j = cp_j$ для некоторого $c > 0$ при всех $j = 1, \dots, w$, либо $\theta_j = c_1 p_j$ или $\theta_j = c_2 p_j$ для некоторых $c_1, c_2 > 0$ при всех $j = 1, \dots, w$. Покажем, что последний вариант не соответствует точке экстремума.

Не ограничивая общности, можно считать, что $0 < c_1 < c_2$ и $\theta_j = c_1 p_j$ ($j = 1, \dots, v$), $\theta_j = c_2 p_j$ ($j = v+1, \dots, w$). Тогда если $P_1 = \sum_{j=1}^v p_j$, $P_2 = \sum_{j=v+1}^w p_j$, то $c_1 P_1 + c_2 P_2 = \rho$ и

$$\sum_{j=1}^w \frac{\theta_j^2}{p_j + \theta_j t} = \sum_{j=1}^v \frac{c_1^2 p_j}{1 + c_1 t} + \sum_{j=v+1}^w \frac{c_2^2 p_j}{1 + c_2 t} = \frac{c_1^2}{1 + c_1 t} P_1 + \frac{c_2^2}{1 + c_2 t} P_2.$$

Чтобы найти максимальное по c_1, c_2 значение правой части при $c_1 P_1 + c_2 P_2 = \rho$ и $0 < c_1 < c_2$, опять составим систему уравнений Лагранжа:

$$\frac{\partial}{\partial c_j} \left(\sum_{i=1}^2 \frac{c_i^2}{1 + c_i t} P_i - \lambda \sum_{i=1}^2 c_i P_i \right) = \frac{c_j P_j (2 + c_j t)}{(1 + c_j t)^2} - \lambda P_j = 0, \quad j = 1, 2.$$

Таким образом, c_1 и c_2 являются положительными корнями квадратного уравнения $x(2 + xt) = \lambda(1 + xt)^2 \Leftrightarrow t(1 - t\lambda)x^2 + 2(1 - t\lambda)x - \lambda = 0$. При $t \in [0, 1]$ и $\lambda \leq 0$ это уравнение не имеет положительных корней. Приводя уравнение к виду

$$\left(x + \frac{1}{t}\right)^2 = \frac{1}{t^2} + \frac{\lambda}{t(1 - t\lambda)} = \frac{1}{t^2(1 - t\lambda)},$$

находим, что уравнение имеет действительные корни только при $0 \leq t\lambda < 1$, но при этом условии и $t > 0$ оба корня не могут быть положительными. Следовательно, экстремальные значения функции $\mathcal{L}(\theta_1, \dots, \theta_w)$ могут достигаться только на границе области $0 < c_1 < c_2$, т. е. только при $c_1 = c_2 = c$.

Пусть $\sum_{j=1}^w p_j = P$. Если $\theta_j = c p_j$ ($j = 1, \dots, w$), то $\sum_{j=1}^w \theta_j = cP = \rho$, т. е. $c = \rho/P$, и при фиксированных $p_1, \dots, p_w$ максимальное значение первой суммы в правой части (9) равно

$$\sum_{j=1}^w \frac{\theta_j^2}{p_j + \theta_j t} = \sum_{j=1}^w \frac{\rho^2 p_j^2}{P^2 (p_j + \rho p_j t / P)} = \sum_{j=1}^w \frac{\rho^2 p_j}{P(P + t\rho)} = \frac{\rho^2}{P + t\rho}.$$

Применяя этот результат ко второй сумме в правой части (9), находим, что ее максимальное значение равно $\rho^2 / (R + t\rho)$, $R = \sum_{j=w+1}^M r_j$.

Таким образом,

$$\sum_{j=1}^N \frac{\theta_j^2}{p_j + \theta_j t} \leq \rho^2 \left(\frac{1}{P + t\rho} + \frac{1}{R + t\rho} \right),$$

и из (8) следует, что при любом $t \in [0, 1]$

$$\mathbf{M} \left| \sum_{j=1}^N \frac{\nu_j \theta_j}{p_j + \theta_j t} \right| \leq \rho \sqrt{n \left(\frac{1}{P + t\rho} + \frac{1}{R + t\rho} \right)} \leq \rho \sqrt{n} \left(\sqrt{\frac{1}{P + t\rho}} + \sqrt{\frac{1}{R + t\rho}} \right).$$

Тогда в силу (7)

$$\begin{aligned}\rho((X_1, \dots, X_n), (Y_1, \dots, Y_n)) &\leq \frac{1}{2} \int_0^1 \mathbf{M} \left| \sum_{j=1}^N \frac{\nu_j \theta_j}{p_j + \theta_j t} \right| dt \leq \\ &\leq \frac{1}{2} \rho \sqrt{n} \int_0^1 \left(\sqrt{\frac{1}{P+t\rho}} + \sqrt{\frac{1}{R+t\rho}} \right) dt.\end{aligned}$$

Так как

$$\int_0^1 \frac{1}{\sqrt{a+t\rho}} dt = \frac{2}{\rho} \sqrt{a+t\rho} \Big|_{t=0}^1 = \frac{2}{\rho} (\sqrt{a+\rho} - \sqrt{a}) = \frac{2}{\sqrt{a+\rho} + \sqrt{a}},$$

то

$$\rho((X_1, \dots, X_n), (Y_1, \dots, Y_n)) \leq \rho \sqrt{n} \left(\frac{1}{\sqrt{P} + \sqrt{P+\rho}} + \frac{1}{\sqrt{R} + \sqrt{R+\rho}} \right).$$

Теорема доказана. $\square$

Список литературы

- [1] Золотарев В. М., *Современная теория суммирования независимых случайных величин*, М.: Наука, 1986, 416 с.
- [2] Феллер В., *Введение в теорию вероятностей и ее приложения*. Т. 1, М.: Мир, 1984, 528 с.
- [3] Reyzin L., *A note on the statistical difference of small direct products*, Boston Univ. Computer Science, Techn. Rep. BUCS-TR-2004-032.
- [4] Renner R., *On the variational distance of independently repeated experiments*, arXiv: cs.IT/0509013.
- [5] Bellare M., Impagliazzo R., *A tool for obtaining tighter security analyses of pseudorandom function based constructions, with applications to PRP $\rightarrow$ PRF conversion*, Cryptology ePrint Archive, Report 1999/024, <http://eprint.iacr.org/1999/024>.
- [6] Sahai A., Vadhan S., "Manipulating statistical difference". In: *"Randomization Methods in Algorithm Design"*, DIMACS Workshop, December 1997, DIMACS Ser. in Discr. Math. and Theor. Comput. Sci., **43**, Providence, R.I.: Amer. Math. Soc., 1999, 251–270.
- [7] Sahai A., Vadhan S., "A complete problem for statistical zero knowledge", *J. ACM*, **50**:2 (2003), 196–249.
- [8] Slud E. V., "Distributional inequalities for the binomial law", *Ann. Probab.*, **5**:3 (1977), 404–412.
- [9] Zubkov A. M., "New inequalities for the binomial law and for the total variation distance between iid samples", Proc. 10th Int. Conf. «Computer Data Analysis and Modeling», v. 2 (Minsk, 2013), Minsk: Publ. center of BSU, 48–50.