# ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

УДК 519.226, 519.244.3, 519.244.8
doi: 10.18101/2304-5728-2017-4-9-20

## О ПРОВЕРКЕ ГИПОТЕЗЫ О ПЛОТНОМ ВЛОЖЕНИИ ДЛЯ ДИСКРЕТНЫХ СЛУЧАЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

**© Меженная Наталья Михайловна**
кандидат физико-математических наук, доцент,
Московский государственный технический университет им. Н. Э. Баумана
Россия, 105005, г. Москва, ул. 2-я Бауманская, 5
E-mail: natalia.mezhennaya@gmail.com

Гипотеза о плотном вложении состоит в том, что одна дискретная последовательность может быть вложена в другую таким образом, что знаки вкладываемой последовательности разделены в результирующей последовательности не более, чем одним знаком. В работе предложен последовательный критерий проверки гипотезы о плотном вложении для дискретных равновероятных случайных последовательностей над конечным алфавитом и изучены его свойства. Вероятность ошибки первого рода (вероятность отклонения верной гипотезы о плотном вложении) построенного критерия равна нулю. Получено выражение для вероятности ошибки второго рода при альтернативной гипотезе, которая состоит в том, что рассматриваемые дискретные последовательности независимы. Рассмотрен также класс подобных критериев. Оказывается, что небольшое изменение процедуры проверки сильно меняет вероятности ошибок. Приведена численная иллюстрация и обсуждение полученных результатов.

***Ключевые слова**:* плотное вложение; последовательный критерий; гипотеза о независимости; вероятности ошибок первого и второго рода; дискретная случайная последовательность.

### Введение

Пусть $X_n = (x_1,...,x_n)$ и $Y_m = (y_1,\ldots,y_m)$ — последовательности элементов множества $A_N = \{0,...,N-1\}, N \geq 2,$ длин $n$ и $m$ соответственно. Будем говорить, что $X_n = (x_1,...,x_n)$ является плотной подпоследовательностью $Y_m = (y_1,\ldots,y_m)$, если существуют такие натуральные числа

$$1 = j_1 < j_2 < \ldots < j_n \leq m,\ j_{k+1} - j_k \in \{1,2\},\ k = 1,\ldots,n-1, \qquad (1)$$

что $x_k = y_{j_k}$, $k = 1,\ldots,n.$

Впервые задача о плотном вложении одной дискретной последовательности в другую рассмотрена в [1]. Получена верхняя оценка для вероятности того, что заданная двоичная случайная последовательность может быть плотно вложена в последовательность независимых двоичных случайных величин с равномерными распределениями. В работе [2] полу-

чено обобщение этого результата на последовательности со значениями в алфавите с любым конечным числом элементов, а также показано, что эта оценка неулучшаема. Также в [2] получена нижняя оценка для вероятности плотного вложения. Обобщение понятия плотного вложения на случай, когда знаки вкладываемой последовательности могут отстоять друг от друга более, чем на один знак, проведено в [3]. Там же получена нижняя оценка для вероятности вложения с произвольным допуском для дискретных случайных последовательностей. Подробно задача об ограниченных двоичных вложениях и ее значимость для задач криптоанализа рассмотрена в [4], [5]. Вопрос об исследовании свойств дискретных последовательностей общего вида и способов их перечисления приведен в [6]. В настоящей работе мы рассмотрим одну задачу о статистической проверке свойств дискретной случайной последовательности.

## 1. Построение критерия и его свойства

Рассмотрим задачу о проверке гипотезы $H_{0n}$ о том, что $X_n$ извлечена из начала последовательности независимых равномерно распределенных на множестве $A_N$ случайных величин $Y_m$ как ее плотная подпоследовательность. Ясно, что извлеченная по правилу (1) из начала последовательности $Y_m$ последовательность $X_n$ всегда может быть плотно вложена в начало последовательности $Y_m$.

Самый простой способ проверки гипотезы $H_{0n}$ состоит в том, чтобы опробовать все $2^{n-1}$ вариантов плотного вложения последовательности $X_n$ в начало последовательности $Y_m$. Вероятность отклонить гипотезу $H_{0n}$, если она верна, равна нулю. Согласно теореме 1 работы [2] вероятность ошибочного принятия гипотезы $H_{0n}$ убывает экспоненциально быстро. Неравенство (3) работы [2] дает верхнюю оценку $p_n$ для вероятности того, что последовательность $X_n$ может быть плотно вложена в начало независимой от нее последовательности $Y_m$, которая при $n \le m$ имеет вид

$$p_n = \frac{1}{2N^{2n}}\left(\left(N-\sqrt{N^2-N}\right)^n + \left(N+\sqrt{N^2-N}\right)^n\right), \qquad (2)$$

не зависит от последовательности $X_n$ и достигается на последовательностях, в которых нет совпадений соседних знаков.

Критерий $\mathcal{T}$ согласия с гипотезой $H_{0n}$, не использующий опробования всех вариантов вложения, был предложен в [7]. В настоящей работе проведем подробный анализ его свойств, а также приведем детальные доказательства сформулированных в [7] утверждений. Критерий $\mathcal{T}$ использует следующий алгоритм:

1) если $x_1 \neq y_1$, то гипотеза $H_{0n}$ отклоняется;

2) если $x_1 = y_1$, то ищем в $(y_2,\ldots,y_m)$ первый знак, равный $x_2$. Обозначим его $y_{j_2}$. Если $j_2 > 3$, то $H_{0n}$ отклоняется, в противном случае продолжаем проверку;

3) далее ищем в $(y_{j_2+1},\ldots,y_m)$ первый знак, равный $x_3$. Обозначим его $y_{j_3}$. Если $j_3 > 5$, то $H_{0n}$ отклоняется, в противном случае продолжаем проверку;

……

*k*) ищем в $(y_{j_{k-1}+1},\ldots,y_m)$ первый знак, равный $x_k$. Обозначим его $y_{j_k}$. Если $j_k > 1+2(k-1)$, то $H_{0n}$ отклоняется, в противном случае продолжаем проверку и так далее до $k = n$.

Пусть

$$L'_1 = 1, \quad L'_k = L'_k(X_n) = \min\{t \ge L'_1 + \ldots + L'_{k-1} : y_t = x_k\}, \quad k = 2,\ldots,n,$$

$$T_k = L'_2 + \ldots + L'_k. \tag{3}$$

Таким образом, критерий $\mathcal{T}$ состоит в следующем. Если $x_1 = y_1$ и на $k$-м шаге выполнено неравенство

$$T_k \le 2(k-1), \tag{4}$$

то решение не принимается. В противном случае гипотеза $H_{0n}$ отклоняется. Если $x_1 = y_1$ и при всех $k = 2,\ldots,n$ выполнено неравенство (4), то считаем, что гипотеза $H_{0n}$ не противоречит результатам наблюдений.

Заметим, если $H_{0n}$ верна, то существует набор чисел $j_1,\ldots,j_n$, удовлетворяющих (1), и $x_k = y_{j_k}$, $k = 1,\ldots,n$. Значит, $L'_k = j_k - j_{k-1} \le 2$, $k = 2,\ldots,n$, и $T_k = L'_2 + \ldots + L'_k \le 2(k-1)$. Таким образом, при описанной процедуре вероятность ошибки первого рода (вероятность отклонить верную гипотезу $H_{0n}$) равна нулю.

Изучим вероятность ошибки критерия $\mathcal{T}$ при альтернативной гипотезе $H_{1n}$ о том, что последовательность $X_n$ не зависит от последовательности $Y_m$ и состоит из независимых равномерно распределенных на множестве $A_N$ случайных величин, а также величину среднего числа знаков, используемых критерием до принятия решения. Пусть $\vartheta_{jn}$ —число проверенных знаков последовательности $X_n$ до принятия решения в критерии $\mathcal{T}$, когда верна гипотеза $H_{jn}$, $j = 0,1$. Ясно, что $\vartheta_{0n} = n$.

Обозначим через $[x]$ целую часть числа $x$, $g^{(m)}(x)$ — $m$-ю производную функции $g$ по $x$.

**Теорема 1.** *Вероятность ошибки второго рода критерия* $\mathcal{T}$ *при* $n \ge 2$ *равна*

$$\mathbf{P}\{H_{0n} \mid H_{1n}\} = \frac{1}{N}\left(1 - \sum_{k=1}^{n-1} \sigma_k\right), \tag{5}$$

*где последовательность чисел* $\sigma_k$ *имеет производящую функцию*

$$\sigma(s)=1-\frac{1-s}{1-sp}\exp\left\{\sum_{m=1}^{\infty}\frac{(1-N^{-1})^m}{m!}\left(x^{2m-1}(1-x)^{-1}\right)^{(m-1)}\Big|_{x=s/N}\right\}. \qquad (6)$$

*Величина среднего числа шагов до принятия решения при верной гипотезе* $H_{1n}$ *равна*

$$\mathbf{E}\vartheta_{1n}=\frac{1}{N}\left(N-1+\sum_{k=1}^{n-2}(k+1)\sigma_k+n\left(1-\sum_{k=1}^{n-2}\sigma_k\right)\right). \qquad (7)$$

**Замечание 1.** Формулу (6) можно также записать в виде (см. (13))

$$\sigma(s)=1-N\exp\left\{N^{-1}-1-\sum_{m=2}^{\infty}\frac{(1-N^{-1})^m}{m!}\left(\sum_{j=0}^{m-1}x^{m-1+j}\right)^{(m-1)}\Big|_{x=s/N}\right\}.$$

## 2. Доказательства

**Доказательство теоремы 1.** Вероятность ошибки второго рода критерия $\mathcal{T}$ равна

$$\mathbf{P}\{H_{0n}\mid H_{1n}\}=\mathbf{P}\{x_1=y_1,T_i\le 2(i-1),i=2,\ldots,n\}=$$
$$=\mathbf{P}\{x_1=y_1\}\left(1-\sum_{k=2}^{n}\mathbf{P}\{T_i\le 2(i-1),i=2,\ldots,k-1,T_k>2(k-1)\}\right)=$$
$$=\frac{1}{N}\left(1-\sum_{k=2}^{n}\mathbf{P}\{T_i\le 2(i-1),i=2,\ldots,k-1,T_k>2(k-1)\}\right). \qquad (8)$$

Для вычисления вероятностей в правой части (8) рассмотрим вспомогательную задачу.

Пусть $L_1,L_2,\ldots,L_n,\ldots$ — последовательность независимых одинаково распределенных случайных величин с геометрическими распределениями (см. [8], с. 238) :

$$\mathbf{P}\{L_i=k\}=pq^{k-1},k=1,2,\ldots,\ i=1,2,\ldots,q=1-p. \qquad (9)$$

Пусть $S_n=L_1+\ldots+L_n-2n$. Найдем

$$\tau_n=\mathbf{P}\{S_1\le 0,S_2\le 0,\ldots,S_{n-1}\le 0,S_n>0\} \qquad (10)$$

вероятность того, что на $n$-м шаге впервые выполнено неравенство $\sum_{i=1}^{n}L_i>2n$ при $n\ge 1$.

**Лемма 1.** *Производящая функция* $\tau(s)=\sum_{n=1}^{\infty}\tau_n s^n$, $s\in[0,1)$, *последовательности* (10) *имеет вид*

$$\tau(s)=1-\frac{1-s}{1-sp}\exp\left\{\sum_{m=1}^{\infty}\frac{q^m}{m!}\left(x^{2m-1}(1-x)^{-1}\right)^{(m-1)}\Big|_{x=sp}\right\}. \qquad (11)$$

**Замечание 2.** Ряд (11) сходится в точке $s=1$ при $p\le 1/2$. Поэтому $\tau(1)=1$ при $p\le 1/2$ (см. теорему 2 §2 главы 12 (с. 448) книги [9]).

**Замечание 3.** Формулу (11) можно преобразовать. При $m \geq 2$

$$\left(x^{2m-1}(1-x)^{-1}\right)^{(m-1)} = \left(\frac{x^{m-1}}{1-x}\right)^{(m-1)} - \left(\sum_{j=0}^{m-1} x^{m-1+j}\right)^{(m-1)} = \frac{(m-1)!}{(1-x)^m} - \left(\sum_{j=0}^{m-1} x^{m-1+j}\right)^{(m-1)}$$

и

$$\tau(s) = 1 - \frac{1-s}{1-sp}\exp\left\{q\frac{sp}{1-sp} + \sum_{m=2}^{\infty}\frac{q^m}{m!}\left(\frac{(m-1)!}{(1-x)^m} - \left(\sum_{j=0}^{m-1} x^{m-1+j}\right)^{(m-1)}\Bigg|_{x=sp}\right)\right\} =$$

$$= 1 - \frac{1-s}{1-ps}\exp\left\{q\frac{sp}{1-sp} + \sum_{m=2}^{\infty}\frac{q^m}{m}\frac{1}{(1-sp)^m} - \sum_{m=2}^{\infty}\frac{q^m}{m!}\left(\sum_{j=0}^{m-1} x^{m-1+j}\right)^{(m-1)}\Bigg|_{x=sp}\right\}.$$

Теперь воспользуемся разложением логарифма в ряд Тейлора при $|s|<1$

$$\sum_{n=1}^{\infty}\frac{s^n}{n} = -\ln(1-s). \tag{12}$$

Получим

$$\tau(s) = 1 - \frac{1-s}{1-ps}\exp\left\{q\frac{sp}{1-sp} - \frac{q}{1-sp} - \right.$$

$$\left. - \ln\left(1-\frac{q}{1-sp}\right) - \sum_{m=2}^{\infty}\frac{q^m}{m!}\left(\sum_{j=0}^{m-1} x^{m-1+j}\right)^{(m-1)}\Bigg|_{x=sp}\right\}.$$

Так как $\ln\left(1-\frac{q}{1-sp}\right) = \ln\left(1-\frac{1-p}{1-sp}\right) = \ln\frac{p(1-s)}{1-sp}$, то

$$\tau(s) = 1 - \frac{1-s}{1-ps}\frac{1-sp}{p(1-s)}\exp\left\{-q - \sum_{m=2}^{\infty}\frac{q^m}{m!}\left(\sum_{j=0}^{m-1} x^{m-1+j}\right)^{(m-1)}\Bigg|_{x=sp}\right\} =$$

$$= 1 - p^{-1}\exp\left\{-q - \sum_{m=2}^{\infty}\frac{q^m}{m!}\left(\sum_{j=0}^{m-1} x^{m-1+j}\right)^{(m-1)}\Bigg|_{x=sp}\right\}. \tag{13}$$

Вернемся к нашему критерию. Так как знаки последовательности $Y_m$ независимы и распределены на множестве $A_N$ равномерно, то распределения случайных величин $L'_k(X_n)$ одинаковы при всех $X_n$. Известно, что если $X_n$ состоит из всех нулей, то случайные величины $L'_2,\ldots,L'_n$ независимы в совокупности и для них выполнены равенства

$$\mathbf{P}\{L'_k = l\} = N^{-1}(1-N^{-1})^{l-1},\ l \geq 1,\ k = 2,\ldots,n. \tag{14}$$

(см., например, [8], с. 327-328). Значит, эти свойства выполнены для любой последовательности $X_n$. Закон распределения случайный величин

$L'_2,...,L'_n$ — это тот же геометрический закон распределения (9) с $p=1/N$ и $q=1-1/N$.

Обозначим $\sigma_n=\tau_n\big|_{p=1/N,q=1-1/N}$ и $\sigma(s)=\sum\limits_{n=1}^{\infty}\sigma_n s^n$. Очевидно, что $\sigma(s)=\tau(s)\big|_{p=1/N,q=1-1/N}$, где $\tau(s)$ определена формулой (11). В этих обозначениях равенство (8) можно записать в виде

$$\mathbf{P}\{H_{0n}\mid H_{1n}\}=\begin{cases}\dfrac{1}{N}\left(1-\sum\limits_{k=1}^{n-1}\sigma_k\right), & n\geq 2,\\ \dfrac{1}{N}, & n=1.\end{cases} \tag{15}$$

Теперь перейдем к вычислению среднего числа $\mathbf{E}\vartheta_{1n}$ знаков последовательности $X_n$, используемых критерием. Так как

$$\mathbf{P}\{\vartheta_{1n}=1\}=1-N^{-1},\quad \mathbf{P}\{\vartheta_{1n}=k+1\}=\sigma_k N^{-1},\ k=1,\dots,n-2,$$

$$\mathbf{P}\{\vartheta_{1n}=n\}=\frac{1}{N}\sum_{k=n-1}^{\infty}\sigma_k=\frac{1}{N}\left(1-\sum_{k=1}^{n-2}\sigma_k\right),$$

то $\mathbf{E}\vartheta_{1n}$ задается формулой (7). Теорема 1 доказана.

**Доказательство леммы 1.** Так как случайная величина $L_i$ равна номеру испытания Бернулли, в котором впервые произошел успех, то сумма $\sum\limits_{i=1}^{n}L_i$ — это номер опыта, в котором произошел $n$-й успех в испытаниях Бернулли. Поэтому

$$\mathbf{P}\left\{\sum_{i=1}^{n}L_i=n+k\right\}=C_{n+k-1}^{n-1}p^nq^k,\ k=0,1,2,\dots. \tag{16}$$

Значит,

$$\mathbf{P}\{S_n=m\}=\mathbf{P}\left\{\sum_{i=1}^{n}L_i=2n+m\right\}=C_{2n+m-1}^{n-1}p^nq^{n+m},m=-n,-n+1,\dots \tag{17}$$

Известно (см. [9], с. 466), что

$$\ln(1-\tau(s))^{-1}=\sum_{n=1}^{\infty}\frac{s^n}{n}\mathbf{P}\{S_n>0\}. \tag{18}$$

Найдем производящую функцию $\tau(s)$, вычислив правую часть (18). Из (17) получаем, что

$$\mathbf{P}\{S_n>0\}=1-\sum_{m=-n}^{0}\mathbf{P}\{S_n=m\}=1-\sum_{m=-n}^{0}C_{2n+m-1}^{n-1}p^nq^{n+m}=1-p^n\sum_{m=0}^{n}C_{n+m-1}^{n-1}q^m.$$

Подставив полученное выражение в правую часть (18), получим

$$\ln(1-\tau(s))^{-1}=\sum_{n=1}^{\infty}\frac{s^n}{n}\left(1-p^n\sum_{m=0}^{n}C_{n+m-1}^{n-1}q^m\right)=\sum_{n=1}^{\infty}\frac{s^n}{n}\left(1-p^n-p^n\sum_{m=1}^{n}C_{n+m-1}^{n-1}q^m\right).$$

Теперь используем разложение для натурального логарифма (12). Имеем

$$\ln(1-\tau(s))^{-1} = -\ln(1-s) + \ln(1-sp) - \sum_{n=1}^{\infty}\frac{(sp)^n}{n}\sum_{m=1}^{n} C_{n+m-1}^{n-1} q^m. \qquad (19)$$

Рассмотрим отдельно последнее слагаемое в правой части (19). Сначала изменим порядок суммирования:

$$\sum_{n=1}^{\infty}\frac{(sp)^n}{n}\sum_{m=1}^{n} C_{n+m-1}^{n-1} q^m = \sum_{m=1}^{\infty}\frac{q^m}{m!}\sum_{n=m}^{\infty}\frac{(n+m-1)!}{(n-1)!}\frac{(sp)^n}{n} =$$

$$= \sum_{m=1}^{\infty}\frac{q^m}{m!}\sum_{n=m}^{\infty}(n+m-1)^{[m-1]}(sp)^n . \qquad (20)$$

Здесь и далее через $a^{[k]}$ обозначена $k$ -я факториальная степень числа $a$ .

При $|x|<1$ выполнено равенство $\sum_{n=m}^{\infty} x^{n+m-1} = x^{2m-1}(1-x)^{-1}$, которое можно продифференцировать $m-1$ раз:

$$\sum_{n=m}^{\infty}(n+m-1)^{[m-1]}x^{n+m-1} = \left(x^{2m-1}(1-x)^{-1}\right)^{(m-1)} .$$

Подставив последнее выражение в (20), получим

$$\sum_{n=1}^{\infty}\frac{(sp)^n}{n}\sum_{m=1}^{n} C_{n+m-1}^{n-1} q^m = \sum_{m=1}^{\infty}\frac{q^m}{m!}\left(x^{2m-1}(1-x)^{-1}\right)^{(m-1)}\Big|_{x=sp} .$$

Тогда из (19)

$$\tau(s) = 1 - \frac{1-s}{1-sp}\exp\left\{\sum_{m=1}^{\infty}\frac{q^m}{m!}\left(x^{2m-1}(1-x)^{-1}\right)^{(m-1)}\Big|_{x=sp}\right\}.$$

Лемма 1 доказана.

### 3. Численная иллюстрация

Так как $\sum_{n=1}^{\infty}\sigma_n = 1$ при всех $N \ge 2$ (см. замечание 2), то вероятность ошибки второго рода критерия $\tau$ стремится к нулю при $n \to \infty$. Так как $T_n = S_{n-1}$, то среднее число знаков последовательности $X_n$, достаточных для принятия верной гипотезы $H_{1n}$, равно $\mathbf{E}\vartheta_{1n}$, задаваемому формулой (7). В таблице 1 приведены значения вероятности ошибки второго рода (ошибочного принятия гипотезы $H_{0n}$). Они вычислены по формуле (15), исходя из разложения функции $\sigma(s)$ в ряд Тейлора в точке $s = 0$. Данное разложение получено с помощью системы Wolfram Mathematica 10. При произвольном $N$ первые 15 членов разложения имеют вид

$$\sigma(s) = \frac{(N-1)^2 s}{N^2} + \frac{2(N-1)^3 s^2}{N^4} + \frac{5(N-1)^4 s^3}{N^6} + \frac{14(N-1)^5 s^4}{N^8} +$$

$$+ \frac{42(N-1)^6 s^5}{N^{10}} + \frac{132(N-1)^7 s^6}{N^{12}} + \frac{429(N-1)^8 s^7}{N^{14}} +$$

$$+\frac{1430(N-1)^9 s^8}{N^{16}}+\frac{4862(N-1)^{10} s^9}{N^{18}}+\frac{16796(N-1)^{11} s^{10}}{N^{20}}+$$
$$+\frac{58786(N-1)^{12} s^{11}}{N^{22}}+\frac{208012(N-1)^{13} s^{12}}{N^{24}}+\frac{742900(N-1)^{14} s^{13}}{N^{26}}+$$
$$\frac{2674440(N-1)^{15} s^{14}}{N^{28}}+\frac{9694845(N-1)^{16} s^{15}}{N^{30}}+\ldots.$$

Таблица 1. Значения вероятности ошибки второго рода критерия $\mathcal{T}$.

| | $n$=2 | $n$=3 | $n$=4 | $n$=5 | $n$=6 | $n$=7 | $n$=8 |
|---|---|---|---|---|---|---|---|
| $N$=3 | 0,1852 | 0,1193 | 0,0828 | 0,0600 | 0,0448 | 0,0342 | 0,026576 |
| $N$=4 | 0,1094 | 0,0566 | 0,0319 | 0,0189 | 0,0116 | 0,0073 | 0,004721 |
| $N$=5 | 0,0720 | 0,0310 | 0,0147 | 0,0073 | 0,0038 | 0,0020 | 0,001100 |
| $N$=6 | 0,0509 | 0,0188 | 0,0076 | 0,0033 | 0,0015 | 0,0007 | 0,000316 |
| $N$=7 | 0,0379 | 0,0122 | 0,0043 | 0,0016 | 0,0006 | 0,0003 | 0,000107 |
| $N$=8 | 0,0293 | 0,0084 | 0,0026 | 0,0009 | 0,0003 | 0,0001 | $41\cdot10^{-6}$ |
| $N$=9 | 0,0233 | 0,0060 | 0,0017 | 0,0005 | 0,0002 | $52\cdot10^{-6}$ | $17\cdot10^{-6}$ |
| $N$=10 | 0,0190 | 0,0044 | 0,0011 | 0,0003 | 0,0001 | $26\cdot10^{-6}$ | $8\cdot10^{-6}$ |

В таблице 2 приведены значения среднего числа знаков $\mathbf{E}\vartheta_{1n}$ при разных $N$ и $n$. При $N=2$ наблюдается наибольший рост $\mathbf{E}\vartheta_{1n}$ с ростом $n$. Это вызвано тем, что в этом случае $\sigma'(1)=\infty$.

Таблица 2. Значение среднего числа знаков $\mathbf{E}\vartheta_{1n}$, используемых критерием $\mathcal{T}$ при гипотезе $H_{1n}$.

| | $n$=2 | $n$=3 | $n$=4 | $n$=5 | $n$=6 | $n$=7 | $n$=8 | $n$=9 | $n$=10 |
|---|---|---|---|---|---|---|---|---|---|
| $N$=2 | 1,50 | 1,88 | 2,19 | 2,46 | 2,71 | 2,93 | 3,14 | 3,34 | 3,52 |
| $N$=3 | 1,33 | 1,52 | 1,64 | 1,72 | 1,78 | 1,83 | 1,86 | 1,89 | 1,91 |
| $N$=4 | 1,25 | 1,36 | 1,42 | 1,45 | 1,47 | 1,48 | 1,49 | 1,49 | 1,49 |
| $N$=5 | 1,20 | 1,27 | 1,30 | 1,32 | 1,33 | 1,33 | 1,33 | 1,33 | 1,33 |
| $N$=6 | 1,17 | 1,22 | 1,24 | 1,24 | 1,25 | 1,25 | 1,25 | 1,25 | 1,25 |
| $N$=7 | 1,14 | 1,18 | 1,19 | 1,20 | 1,20 | 1,20 | 1,20 | 1,20 | 1,20 |
| $N$=8 | 1,13 | 1,15 | 1,16 | 1,17 | 1,17 | 1,17 | 1,17 | 1,17 | 1,17 |
| $N$=9 | 1,11 | 1,13 | 1,14 | 1,14 | 1,14 | 1,14 | 1,14 | 1,14 | 1,14 |
| $N$=10 | 1,10 | 1,12 | 1,12 | 1,12 | 1,12 | 1,12 | 1,12 | 1,12 | 1,12 |

Ниже (рис. 1) представлен график зависимости вероятности ошибки второго рода $\beta_n=\mathbf{P}\{H_{0n}\,|\,H_{1n}\}$ от $n$ при $N=2$, а также верхняя оценка для вероятности плотного вложения $p_n$ по формуле (2). Из графиков видно, что вероятность ошибки второго рода значительно больше, чем $p_n$. Это обусловлено тем, что при проверке гипотезы $H_{0n}$ по критерию $\mathcal{T}$ мы

вкладываем каждый следующий знак последовательности $X_n$ на ближайшее возможное место, при этом некоторые из величин $L'_2,\ldots,L'_k$ могут быть больше 2. Например, если $n=5$ $L'_2=1, L'_3=1, L'_4=1, L'_5=3,$ то $T_2=1\le 2(2-1), T_3=2\le 2(3-1), T_4=3\le 2(4-1), T_5=6\le 2(5-1)$ и гипотеза $H_{0n}$ принимается, хотя места расположения знаков $X_n$ в $Y_m$ не удовлетворяют условию (1).

Из приведенных расчетов видно, что было бы хорошо подкорректировать свойства критерия $\mathcal{T}$ выбором подходящего множителя в правой части неравенства (4). В частности, это позволит сократить среднее количество проверяемых знаков при всех $N\ge 2$. Естественно, в этом случае вероятность ошибки первого рода будет положительна. Оказывается, что выбор одной и той же константы $c\in(1,2)$ при всех $n$ в (4) ведет к резкому увеличению ошибки первого рода.

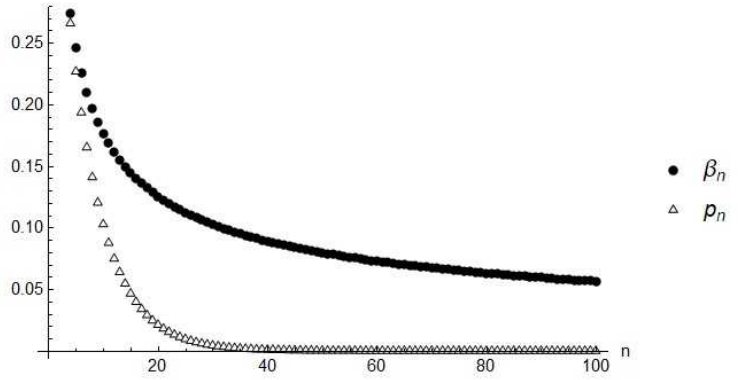


Рис. 1. График зависимости вероятности ошибки второго рода $\beta_n=\mathbf{P}\{H_{0n}\,|\,H_{1n}\}$ критерия $\mathcal{T}$ от $n$ при $N=2$. Для сравнения приведена оценка вероятности плотного вложения по формуле (2).

Например, пусть $1<c<2$ и вместо (4) используем $T_k<c(k-1)$. Оценим вероятность ошибки второго рода при $N=2.$ При верной гипотезе $H_{0n}$ мы всегда будем ее отклонять, если первый и второй знаки последовательности $X_n$ вкладываются через один, т.е.

$$\mathbf{P}\{H_{1n}\,|\,H_{0n}\}\ge\mathbf{P}\{x_1=y_1, x_2\ne y_2, x_2=y_3\}=1/8.$$

На самом деле приведенная оценка является грубой. Экспериментальная оценка вероятности ошибки первого рода по 1000 наблюдений при $N=2$ и различных значениях $c$ представлена на рисунке 2.

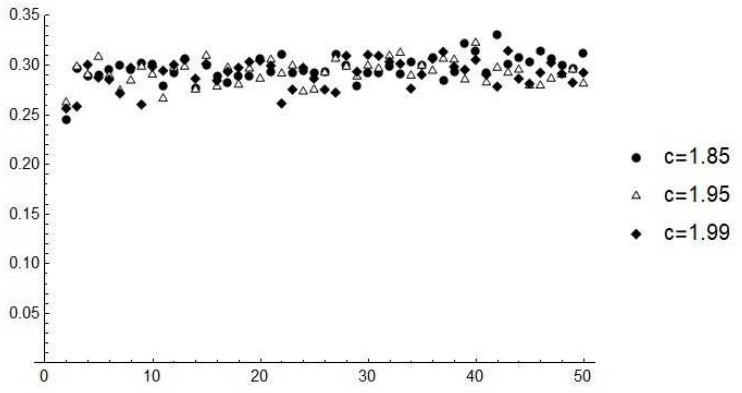


Рис. 2. Экспериментальная оценка вероятности ошибки первого рода по 1000 наблюдений при $N = 2$ и различных значениях $c$.

Вероятность ошибки второго рода при разных $c$ представлена на графике ниже. Видно, что при достижении значения $c = 2$ она меняется скачком.

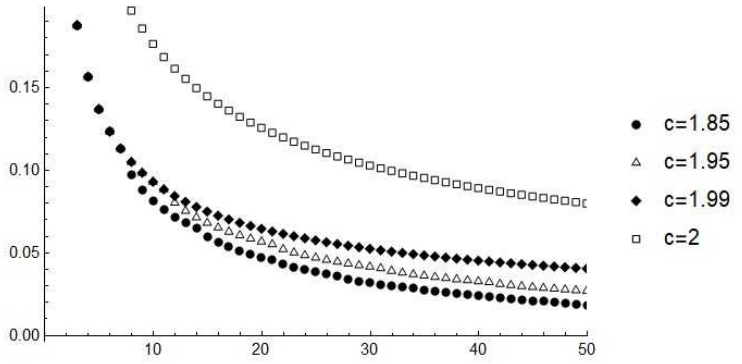


Рис. 3. Вероятность ошибки второго рода при $N = 2$ и различных значениях $c$.

Таким образом, дальнейшая модификация критерия требует определения границы $c$ как функции от длины вкладываемой последовательности $n$. При этом ясно, что для первых нескольких шагов $c$ не может быть меньше 2.

**Заключение**

В работе рассмотрен последовательный критерий проверки гипотезы о плотном вложении одной дискретной последовательности в другую. Вероятность ошибки первого рода этого критерия равна нулю. Найдено аналитическое выражение для вероятности ошибки второго рода при альтернативной гипотезе о независимости рассматриваемых последовательностей. При описанной процедуре она оказывается не слишком маленькой

при небольшом размере алфавита. Аналогичными рассуждениями можно изучить класс подобных критериев, у которых вероятность ошибки первого рода положительна. Оказалось, что даже при небольшом изменении параметра $c$ вероятность ошибки первого рода перестает быть нулевой и сразу достигает величины порядка 0,3, а вероятность ошибки второго рода уменьшается приблизительно в два раза.

**Литература**

1. Golic J. Dj. Constrained embedding probability for two binary strings // SIAM J. Discrete Math. 1996. Vol. 9, No. 3. P. 360–364.

2. Михайлов В. Г., Меженная Н. М. Оценки для вероятности плотного вложения одной дискретной последовательности в другую // Дискретная математика. 2005. Т. 17, № 3. С. 19–27.

3. Меженная Н. М., Михайлов В. Г. Нижние оценки для вероятности вложения с произвольным допуском // Вестник Московского государственного технического университета им. Н. Э. Баумана. Серия: Естественные науки. 2012. № 2. С. 3–11.

4. Donovan D. M., Lefevre J., Simpson L. A Discussion of Constrained Binary Embeddings with Applications to Cryptanalysis of Irregularly Clocked Stream Ciphers // Balakrishnan R. Veni Madhavan C. (Eds.) Discrete mathematics. Proceedings of the international conference on discrete mathematics, Indian Institute of Science, Bangalore, December 2006. P. 73–86.

5. Kholosha A. Clock-Controlled Shift Registers for Key-Stream Generation. IACR Cryptology ePrint Archive 2001: 61 (2001). URL: eprint.iacr.org/2001/061.pdf.

6. Кошевой Н. Д., Костенко Е. М., Доценко Н. В., Павлик А. В. Метод перечисления символьных последовательностей // Радіоелектронні і комп'ютерні системи. 2012. № 3 (55). С. 45–49.

7. Меженная Н. М. Предельные теоремы в задачах о плотном вложении и плотных сериях в дискретных случайных последовательностях. дис… канд. физ.-мат. наук / Московский государственный институт электроники и математики. М., 2009.

8. Феллер В. Введение в теорию вероятностей и ее приложения: в 2 т. М.: Мир, 1984. Т. 1. 528 с

9. Феллер В. Введение в теорию вероятностей и ее приложения: в 2 т. М.: Мир, 1984. Т. 2. 751 с.

# ABOUT TESTING THE DENSE EMBEDDING HYPOTHESIS FOR DISCRETE RANDOM SEQUENCES

*Natalya M. Mezhennaya*
Cand. Sci. (Phys. and Math.), A/Prof.,
Bauman Moscow State Technical University
5 2nd Baumanskaya St., Moscow 105005, Russia

The dense embedding hypothesis says that one discrete sequence can be embedded in the other in such a way that the characters of the inserted sequence are separated in the resulting sequence by at most one character. We propose a sequential test for the dense imbedding hypothesis for discrete equiprobable random sequences over a finite alphabet and study its properties. The probability of type I error (the probability of rejection of the dense embedding hypothesis when it's true) of the constructed test equals zero. We derive an expression for the probability of type II error under the alternative hypothesis that the discrete sequences under consideration are independent. A class of similar test is also considered. It turns out that a small change in the testing procedure greatly changes the error probabilities. A numerical illustration and discussion of the results are given.

*Keywords*: dense embedding; sequential test; hypothesis of independence; probabilities of type I and type II errors; discrete random sequence.