

© 2001 г.

МИХАЙЛОВ В. Г.*

ОЦЕНКА ТОЧНОСТИ СЛОЖНОЙ ПУАССОНОВСКОЙ АППРОКСИМАЦИИ ДЛЯ РАСПРЕДЕЛЕНИЯ ЧИСЛА СОВПАДАЮЩИХ ЦЕПОЧЕК¹⁾

Пусть X_1, \dots, X_m и Y_1, \dots, Y_n — две последовательности независимых одинаково распределенных случайных величин, принимающих значения $1, 2, \dots$. С помощью специального варианта метода Стейна строится оценка точности аппроксимации распределения числа совпадений цепочек исходов X_i, \dots, X_{i+s-1} заданной длины s в первой последовательности с цепочками исходов Y_j, \dots, Y_{j+s-1} во второй последовательности. В качестве аппроксимирующего выступает распределение суммы пуассоновского числа независимых случайных величин с геометрическим распределением.

Ключевые слова и фразы: длинные повторения, совпадения слов, оценки точности пуассоновской аппроксимации, сложное пуассоновское распределение, методы Стейна и Чена–Стейна.

Рассмотрим следующую задачу. Пусть имеются две последовательности X_1, \dots, X_m и Y_1, \dots, Y_n независимых одинаково распределенных случайных величин, принимающих значения $1, 2, \dots$. При этом считаем, что за случайной величиной X_m следуют случайные величины X_1, X_2, \dots , а за случайной величиной Y_n следуют случайные величины Y_1, Y_2, \dots . Циклическая структура, заданная таким образом на множестве индексов, слабо отражается на исследуемых свойствах последовательностей и использована главным образом для некоторого упрощения изложения.

Будем использовать обозначения

$$p_k = \mathbf{P}\{X_i = k\}, \quad q_k = \mathbf{P}\{Y_i = k\}, \quad k = 1, 2, \dots,$$
$$p = \max_k p_k I\{q_k > 0\}, \quad q = \max_k q_k I\{p_k > 0\}$$

*Математический институт им. В. А. Стеклова РАН, ул. Губкина, 8, 117966 Москва, ГСП-1, Россия.

¹⁾ Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект 96-01-00531) и Совета по грантам Президента РФ и государственной поддержки ведущих научных школ (проект 96-15-96092).

(здесь и далее $I\{\cdot\}$ обозначает индикатор случайного события или неслучайного множества),

$$r = \max_k p_k q_k, \quad R = \sum_{k=1}^{\infty} p_k q_k.$$

Исследуется случайная величина

$$\xi(m, n, s) = \sum_{i=1}^m \sum_{j=1}^n I\{(X_i, \dots, X_{i+s-1}) = (Y_j, \dots, Y_{j+s-1})\}, \quad (1)$$

выражающая число совпадений цепочек исходов X_i, \dots, X_{i+s-1} заданной длины s в первой последовательности с цепочками исходов Y_j, \dots, Y_{j+s-1} во второй последовательности. Величина $\xi(m, n, s)$ вместе с группой связанных с ней случайных величин и характеристик (длина наиболее длинного совпадения, число совпадающих участков и т.п.) играет существенную роль в математических моделях молекулярной биологии и генетики (см., например, библиографию в [1]).

Задача о совпадающих участках в паре последовательностей рассматривалась во многих статьях. Предельная теорема Пуассона для числа повторений участков длины не меньше заданной (в случае, когда элементы обеих последовательностей распределены одинаково) была доказана в работе [2], оценка скорости сходимости в этой теореме была получена в работе [3]. В последнем случае был использован метод Чена–Стейна. Как следствия в этих работах были получены предельные теоремы и оценки в них для максимальной длины совпадающих участков. В литературе имеются также результаты о совпадениях участков в независимых реализациях последовательностей слабо зависимых случайных величин (см., например, [4], [5]). Довольно детально изучались вопросы, касающиеся длины максимального участка с заданной долей совпадений букв и максимального числа совпадений букв в словах заданной длины (см. [2], [6]–[8]). Здесь мы не касаемся этих вопросов.

Основной результат настоящей статьи формулируется следующим образом. Пусть $d(L(U), L(V))$ обозначает расстояние по вариации между распределениями случайных величин U и V . Пусть λ — среднее число пар повторившихся участков длины не меньше s , $\lambda = nm(1 - R)R^s$, а $CP(\Lambda)$, где $\Lambda = (\lambda_1, \lambda_2, \dots)$, обозначает распределение случайной величины $\sum_{i=1}^{\infty} i\pi(\lambda_i)$, где $\pi(\lambda_i)$ — независимые случайные величины, распределенные по закону Пуассона с параметрами λ_i . Это обычное сложное пуассоновское распределение. Будем использовать обозначения $\Theta = (\theta_1, \theta_2, \dots)$, $\theta_i = \lambda(1 - R)R^{i-1}$,

$$d(\Theta) = d(L(\xi(m, n, s)), CP(\Theta)), \quad S(R) = (1 - R)(1 - 2R). \quad (2)$$

Заметим, что распределение $CP(\Theta)$ отвечает сумме пуассоновского (с параметром λ) числа независимых случайных слагаемых, имеющих геометрическое распределение (с параметром R).

Теорема 1. Пусть $2 \leq s < \min\{n, m\}$, $0 < R < 1$. Тогда

$$d(\Theta) = d\left(L(\xi(m, n, s)), CP(\Theta)\right) < C(\lambda, R) (A\lambda + B\lambda^2) + \frac{4\lambda^2}{nmR(1-R)}, \tag{3}$$

где

$$C(\lambda, R) \leq \min\left\{1, (\lambda(1-R))^{-1}\right\} e^\lambda, \tag{4}$$

$$A = \frac{2s}{1-R} \left(2s \left(\frac{r}{R}\right)^s + mp^s + nq^s\right), \quad B = \frac{2(4s-3)}{(1-R)^2} \left(\frac{1}{n} + \frac{1}{m}\right).$$

Если при этом $0 < R < \frac{1}{2}$, то можно взять

$$C(\lambda, R) = \min\left\{1, \frac{1}{\lambda S(R)} \left[\frac{1}{4\lambda S(R)} + \left(\ln(2\lambda S(R))\right)_+\right]\right\}. \tag{5}$$

Следствие 1. Пусть $n, m \rightarrow \infty$ и $2 \leq s < \min\{n, m\}$. Тогда, если $0 < R \leq \text{const} < 1$ и $\lambda = O(1)$ или если $0 < R \leq \text{const} < \frac{1}{2}$ и $\lambda \rightarrow \infty$, то

$$d(\Theta) = O\left(\left(s \left(\frac{r}{R}\right)^s + mp^s + nq^s + \frac{\lambda}{n} + \frac{\lambda}{m}\right) s \ln \lambda\right). \tag{6}$$

Следствие 2. Пусть $n, m \rightarrow \infty$, а остальные параметры схемы меняются так, что $2 \leq s < \min\{n, m\}$, $0 < R \leq \text{const} < \frac{1}{2}$, $\lambda \rightarrow \infty$ и выполнено условие

$$\left(s \left(\frac{r}{R}\right)^s + mp^s + nq^s + \frac{\lambda}{n} + \frac{\lambda}{m}\right) s \ln \lambda \rightarrow 0.$$

Тогда распределение случайной величины

$$(\xi(m, n, s) - \lambda(1-R)^{-1})(1-R)(\lambda(1+R))^{-1/2}$$

сходится к стандартному нормальному распределению.

З а м е ч а н и е 1. Асимптотическое распределение случайной величины $\xi(m, n, s)$ представляет собой сложное пуассоновское распределение с производящей функцией $E z^{\xi(m, n, s)} = \exp\{\lambda(z-1)/(1-zR)\}$ и имеет простой вероятностный смысл. В последовательностях X_1, \dots, X_m и Y_1, \dots, Y_n имеется асимптотически пуассоновское (с параметром λ) число $N(m, n, s)$ пар совпадающих между собой участков длины не

меньше s . Длины этих участков асимптотически независимы и превосходят s на случайные добавки с асимптотически геометрическим распределением (с параметром R) длины добавка.

В связи с этим приведем (без доказательства) оценку точности пуассоновской аппроксимации для распределения случайной величины $N(m, n, s)$. Обозначим через $Po(\lambda)$ пуассоновское распределение с параметром λ .

Теорема 2. Пусть $R < 1$ и $s < \min\{n, m\}$. Тогда выполнено неравенство

$$d(L(N(m, n, s)), Po(\lambda)) < (2s + 1)(1 - e^{-\lambda})B(m, n, s), \quad (7)$$

где

$$B(m, n, s) = \frac{2s}{1 - R} \left(\frac{r}{R}\right)^s + mp^s + nq^s + (n + m)(1 - R)R^s.$$

Исследованию свойств случайной величины $N(m, n, s)$ будет посвящена отдельная статья.

З а м е ч а н и е 2. В случае, когда все величины X_1, \dots, X_m и Y_1, \dots, Y_n имеют одинаковые распределения, оценка точности пуассоновской аппроксимации для распределения случайной величины $N(m, n, s)$ была получена в работе [3].

Поскольку в теореме 1 использовано ограничение $s \geq 2$, рассмотрим отдельно случай $s = 1$. Он заметно проще, и это позволяет нам несколько расширить постановку, допустив, в частности, неодинаковые распределения у случайных величин в каждой из последовательностей X_1, \dots, X_m и Y_1, \dots, Y_n .

Введем обозначения

$$P_k = \sum_{i=1}^m \mathbf{P}\{X_i = k\}, \quad Q_k = \sum_{j=1}^n \mathbf{P}\{Y_j = k\},$$

$$p_k = \max_i \mathbf{P}\{X_i = k\}, \quad q_k = \max_j \mathbf{P}\{Y_j = k\}.$$

Пусть задано некоторое множество $A \subseteq \{1, 2, \dots\}$ и $\lambda(A) = \sum_{k \in A} P_k Q_k$.

Определим случайную величину

$$\xi(m, n; A) = \sum_{i=1}^m \sum_{j=1}^n I\{X_i = Y_j \in A\}.$$

Очевидно, что $\xi(m, n; \{1, 2, \dots\}) = \xi(m, n, 1)$ и $\mathbf{E}\xi(m, n; A) = \lambda(A)$.

Теорема 3. Выполняется неравенство

$$d(L(\xi(m, n; A)), Po(\lambda(A))) < \frac{1 - e^{-\lambda(A)}}{\lambda(A)} \left(\sum_{k \in A} P_k Q_k ((1 + q_k) P_k + (1 + p_k) Q_k) \right). \quad (8)$$

З а м е ч а н и е 3. Из теоремы 3 вытекает, что достаточными для сближения распределения величины $\xi(m, n; A)$ с законом Пуассона с параметром λ являются условия

$$\lambda(A) \rightarrow \lambda, \quad \sum_{k \in A} P_k Q_k ((1 + q_k) P_k + (1 + p_k) Q_k) \rightarrow 0. \quad (9)$$

Нетрудно убедиться также, что

$$\left\{ \sum_{k \in A} P_k \rightarrow 0 \right\} \text{ или } \left\{ \sum_{k \in A} Q_k \rightarrow 0 \right\} \implies \left\{ \mathbf{P}\{\xi(m, n; A) = 0\} \rightarrow 1 \right\}. \quad (10)$$

Следующая теорема показывает, что в случае равновероятных распределений указанные в (9) и (10) достаточные условия являются также и необходимыми.

Пусть распределения случайных величин X_i и Y_j равномерны:

$$\begin{aligned} \mathbf{P}\{X_i = k\} &= M^{-1}, \quad i = 1, \dots, m, \quad k = 1, \dots, M, \\ \mathbf{P}\{Y_j = k\} &= N^{-1}, \quad j = 1, \dots, n, \quad k = 1, \dots, N, \end{aligned}$$

и $A = \{1, \dots, K\}$, $K \leq M, N$. Тогда

$$P_k = m M^{-1}, \quad Q_k = n N^{-1}, \quad \lambda(A) = K m M^{-1} n N^{-1}. \quad (11)$$

Теорема 4. Пусть в данной схеме $m, n \rightarrow \infty$ и $\lambda(A) \rightarrow \lambda$, где $0 < \lambda < \infty$.

а) Пусть выполнено условие $K m M^{-1} \rightarrow 0$ или $K n N^{-1} \rightarrow 0$. Тогда

$$\mathbf{P}\{\xi(m, n; A) = 0\} \rightarrow 1.$$

б) Пусть выполнено условие $m M^{-1} + n N^{-1} \rightarrow 0$. Тогда

$$L(\xi(m, n; A)) \rightarrow \text{Po}(\lambda).$$

в) Пусть при некотором μ , $0 < \mu < \infty$, выполнено соотношение $\min\{|K m M^{-1} - \mu|, |K n N^{-1} - \mu|\} \rightarrow 0$ и $K \rightarrow \infty$. Тогда предельным для случайной величины $\xi(m, n; A)$ является сложное пуассоновское распределение с производящей функцией

$$\exp\left\{\mu(\exp\{\lambda\mu^{-1}(z-1)\} - 1)\right\}. \quad (12)$$

г) Пусть $K = \text{const}$, $m M^{-1} \rightarrow \mu$, $n N^{-1} \rightarrow \nu$, $0 < \mu, \nu < \infty$, $\lambda = K \mu \nu$. Тогда предельным для случайной величины $\xi(m, n; A)$ является распределение выражения $\sum_{k=1}^K \pi_k(\mu) \pi_k(\nu)$, составленного из независимых случайных величин $\pi_1(\mu), \dots, \pi_K(\mu), \pi_1(\nu), \dots, \pi_K(\nu)$, распределенных по закону Пуассона с указанным в скобках параметром.

З а м е ч а н и е 4. Условие п. а) теоремы 4 эквивалентно соотношению $mM^{-1} + nN^{-1} \rightarrow \infty$. В п. б) изучена ситуация, когда эта сумма стремится к нулю. Условия пункта в) включают в себя два частных случая: $mM^{-1} \rightarrow 0$, $nN^{-1} \rightarrow \lambda\mu^{-1}$ и $mM^{-1} \rightarrow \lambda\mu^{-1}$, $nN^{-1} \rightarrow 0$ и всевозможные комбинации этих условий на подпоследовательностях. Случаю, когда $mM^{-1} \rightarrow \mu > 0$ и $nN^{-1} \rightarrow \nu > 0$, отвечает п. г) теоремы. Таким образом, в теореме 4 приведен полный спектр предельных распределений для величины $\xi(m, n; A)$ в рассмотренном случае. Из этого факта заключаем, что в данном случае (см. (11))

$$\left\{ \mathbf{P}\{\xi(m, n; A) = 0\} \rightarrow 1 \right\} \implies \left\{ \sum_{k \in A} P_k \rightarrow 0 \right\} \text{ или } \left\{ \sum_{k \in A} Q_k \rightarrow 0 \right\},$$

а из условия $\{\lambda(A) \rightarrow \lambda \ \& \ L(\xi(m, n; A)) \rightarrow \text{Po}(\lambda)\}$ следует второе из условий (9).

Перейдем к доказательствам. Нам понадобится следующая теорема о суммах случайных индикаторов (см. [9] или [10]).

Пусть Γ — произвольный конечный набор индексов, I_a , $a \in \Gamma$, — случайные индикаторы, $W = \sum_{a \in \Gamma} I_a$. Для каждого I_a разделим некоторым образом множество Γ на четыре непересекающихся множества: $\{a\}$, $\Gamma_a^{\nu s}$, Γ_a^b , $\Gamma_a^{\nu w}$, и положим

$$U_a = \sum_{b \in \Gamma_a^{\nu s}} I_b, \quad V_a = \sum_{b \in \Gamma_a^b} I_b.$$

Определим набор $\Lambda = (\lambda_1, \dots, \lambda_{D+1}, 0, \dots)$, где

$$\lambda_i = i^{-1} \sum_{a \in \Gamma} \mathbf{E}\{I_a I\{I_a + U_a = i\}\}, \quad D = \max_a |\Gamma_a^{\nu s}|.$$

Введем обозначение $\varphi = \sum_{a \in \Gamma} \sum_{i=1}^{|\Gamma_a^{\nu s}|+1} \varphi_{ai}$, где

$$\varphi_{ai} = \mathbf{E}\left| \mathbf{E}\{I_a I\{I_a + U_a = i\} \mid (I_b: b \in \Gamma_a^{\nu w})\} - \mathbf{E}\{I_a I\{I_a + U_a = i\}\} \right|.$$

Теорема 5. При любом выборе указанных выше множеств

$$\begin{aligned} & d(L(W), \text{CP}(\Lambda)) \\ & \leq c_1(\Lambda) \varphi + c_2(\Lambda) \sum_{a \in \Gamma} \left((\mathbf{E}I_a)^2 + \mathbf{E}I_a \mathbf{E}(U_a + V_a) + \mathbf{E}I_a V_a \right), \end{aligned} \quad (13)$$

где

$$\max\{c_1(\Lambda), c_2(\Lambda)\} \leq \min\left\{1, \frac{1}{\lambda_1}\right\} \exp\left\{\sum_{s=1}^{\infty} \lambda_s\right\}.$$

Если $\nu\lambda_\nu \downarrow 0$ при $\nu \uparrow \infty$, то

$$\begin{aligned} c_1(\Lambda) &= \begin{cases} 1, & \lambda_1 - 2\lambda_2 \leq 1, \\ (\lambda_1 - 2\lambda_2)^{-1/2} [2 - (\lambda_1 - 2\lambda_2)^{-1/2}], & \lambda_1 - 2\lambda_2 > 1, \end{cases} \\ c_2(\Lambda) &= \min\left\{1, (\lambda_1 - 2\lambda_2)^{-1} [(4(\lambda_1 - 2\lambda_2))^{-1} + (\ln 2(\lambda_1 - 2\lambda_2))_+]\right\}. \end{aligned}$$

Оценка (13) доказана в работе [9] с помощью результатов работы [11], неравенства и выражения для коэффициентов получены в [11] (подробности можно найти в обзорной статье [10]).

Доказательство теоремы 1. Воспользуемся результатом теоремы 5. В нашем случае $\Gamma = \{a = (i, j): 1 \leq i \leq n, 1 \leq j \leq m\}$, $I_a = I\{(X_i, \dots, X_{i+s-1}) = (Y_j, \dots, Y_{j+s-1})\}$. Возьмем

$$\begin{aligned} \Gamma_a^{\nu s} &= \{b = (i', j') \in \Gamma \setminus \{a = (i, j)\}: -s < i' - i = j' - j < s\}, \\ \Gamma_a^{\nu w} &= \{b = (i', j') \in \Gamma: \min\{|i' - i|, |j' - j|\} > 2s - 1\}, \\ \Gamma_a^b &= \Gamma \setminus (\{a\} \cup \Gamma_a^{\nu s} \cup \Gamma_a^{\nu w}). \end{aligned}$$

В силу этих определений $D+1 = 2s-1$, т.е. в описании сопровождающего сложного пуассоновского распределения участвуют $2s - 1$ величин λ_ν , а именно $\lambda_1, \dots, \lambda_{2s-1}$. Событие $\{I_a = 1, I_a + U_a = \nu\}$ при $\nu = 1, \dots, s - 1$ является объединением ν несовместных событий

$$\{I_{a-k} \neq 1, I_{a-k+1} = \dots = I_{a-k+\nu} = 1, I_{a-k+\nu+1} \neq 1\},$$

где $k = 1, \dots, \nu$ и считается, что $a - k = (i - k, j - k)$. Вероятность каждого из этих событий равна $(1 - R)^2 R^{s+\nu-1}$. Поэтому при $\nu = 1, \dots, s - 1$

$$\begin{aligned} \lambda_\nu &= \frac{1}{\nu} \sum_{a \in \Gamma} \mathbf{E}\{I_a I\{I_a + U_a = \nu\}\} \\ &= nm(1 - R)^2 R^{s+\nu-1} = \lambda(1 - R) R^{\nu-1}, \end{aligned} \tag{14}$$

где $\lambda = nm(1 - R) R^s$ — среднее число отрезков повторений длины не меньше s .

При $\nu = s, \dots, 2s - 2$ событие $\{I_a = 1, I_a + U_a = \nu\}$ является объединением $2s - \nu - 2$ несовместных событий

$$\{I_{a-k} \neq 1, I_{a-k+1} = \dots = I_{a-k+\nu} = 1, I_{a-k+\nu+1} \neq 1\},$$

где $k = \nu - s + 1, \dots, s - 2$, и двух событий

$$\begin{aligned} \{I_{a-s+1} = \dots = I_{a-s+\nu} = 1, I_{a-s+\nu+1} \neq 1\}, \\ \{I_{a+s-\nu-1} \neq 1, I_{a+s-\nu} = \dots = I_{a+s-1} = 1\}. \end{aligned}$$

Вероятность каждого из этих двух событий равна $(1 - R) R^{s+\nu-1}$. Поэтому при $\nu = s, \dots, 2s - 2$

$$\lambda_\nu = \nu^{-1} \lambda(1 - R) R^{\nu-1} (2s - \nu + 2R). \tag{15}$$

При $\nu = 2s - 1$ выполняется равенство $\{I_a = 1, I_a + U_a = \nu\} = \{I_{a-s+1} = \dots = I_{a+s-1} = 1\}$. Вероятность этого события равна R^{3s-2} . Поэтому

$$\lambda_{2s-1} = \frac{\lambda R^{2s-2}}{(2s - 1)(1 - R)}. \tag{16}$$

Выведем теперь оценки для слагаемых в правой части (13). Очевидно, в нашем случае

$$\begin{aligned} \varphi &= 0, \quad \mathbf{E}I_a = R^s, \\ \mathbf{E}I_a(\mathbf{E}I_a + \mathbf{E}U_a + \mathbf{E}V_a) &= (nm - (n - 4s + 3)(m - 4s + 3))R^{2s}. \end{aligned} \quad (17)$$

Самое главное теперь — вычислить или оценить $\mathbf{E}I_a V_a$. Выделим в Γ_a^b подмножества

$$\begin{aligned} \Gamma_a^* &= \left\{ b = (i', j') \in \Gamma_a^b: \min\{|i' - i|, |j' - j|\} < s \right\} \setminus \Gamma_a^{\nu s}, \\ \Gamma_a^{**} &= \left\{ b = (i', j') \in \Gamma_a^*: \max\{|i' - i|, |j' - j|\} < s \right\}. \end{aligned}$$

При $(i', j') \in \Gamma_a^{**}$, как нетрудно проверить, следуя рассуждениям доказательства леммы 2 из [12], $\mathbf{E}I_a I_b \leq r^s$ (напомним, что $r = \max_k p_k q_k$). При $(i', j') \in \Gamma_a^* \setminus \Gamma_a^{**}$, если $|i' - i| < s$, то выполняется неравенство $\mathbf{E}I_a I_b \leq \mathbf{E}I_a q^s$, а если $|j' - j| < s$, то выполняется неравенство $\mathbf{E}I_a I_b \leq \mathbf{E}I_a p^s$. Поэтому $\sum_{b \in \Gamma_a^*} \mathbf{E}I_a I_b < 4s^2 r^2 + \mathbf{E}I_a(2smp^s + 2snq^s)$. Кроме того, $\mathbf{E}I_a I_b = \mathbf{E}I_a \mathbf{E}I_b$ при $b \in \Gamma_a^b \setminus \Gamma_a^*$. Поэтому

$$\begin{aligned} \mathbf{E}I_a V_a &< \sum_{b \in \Gamma_a^*} \mathbf{E}I_a I_b + \mathbf{E}I_a \mathbf{E}V_a < 4s^2 r^2 + 2sR^2(mp^s + nq^s) \\ &+ (nm - (n - 4s + 3)(m - 4s + 3))R^{2s}. \end{aligned} \quad (18)$$

Подставив (17) и (18) в оценку теоремы 5, получим после некоторых упрощений неравенство

$$d(L(\xi(m, n, s)), \text{CP}(\Lambda)) < c_2(\Lambda)(A\lambda + B\lambda^2). \quad (19)$$

Теперь о множителе $c_2(\Lambda)$. Оценка (4) следует непосредственно из теоремы 5 и наших определений. Воспользуемся условием $R < \frac{1}{2}$. В силу (14) $\lambda_1 - 2\lambda_2 = \lambda S(R) = \lambda(1 - R)(1 - 2R)$. Поэтому в этом случае

$$c_2(\Lambda) \leq C(\lambda, R) \quad (20)$$

с указанным в (5) значением $C(\lambda, R)$.

Оценим теперь расстояние между сложными пуассоновскими распределениями $\text{CP}(\Lambda)$ с $\Lambda = (\lambda_1, \dots, \lambda_{2s-1}, 0, \dots)$ и $\text{CP}(\Theta)$ с $\Theta = (\theta_1, \theta_2, \dots)$, $\theta_i = \lambda(1 - R)R^{i-1}$. Последнее, как уже отмечалось, отвечает сумме пуассоновского числа (с параметром λ) независимых слагаемых, имеющих геометрическое распределение с параметром R . Воспользуемся тем, что

$$d(\text{CP}(\Lambda), \text{CP}(\Theta)) \leq \sum_{i=1}^{2s-1} |\lambda_i - \theta_i| + \sum_{i=2s}^{\infty} \theta_i.$$

Поэтому (используем (14)–(16), а также неравенства $s \geq 2$ и $R < \frac{1}{2}$)

$$\begin{aligned} d(\text{CP}(\Lambda), \text{CP}(\Theta)) &\leq |\lambda_s - \theta_s| + \sum_{i=s+1}^{2s-1} |\lambda_i - \theta_i| + \sum_{i=2s}^{\infty} \theta_i \\ &\leq \frac{2R}{s} \lambda(1-R) R^{s-1} + \lambda R^s + \lambda \frac{R^{2s-1}}{1-R} \\ &\leq 4\lambda R^s = \frac{4\lambda^2}{nm(1-R)}. \end{aligned} \tag{21}$$

Используя (19), (21) и неравенство треугольника, получаем оценку (3). Теорема 1 доказана.

Утверждение следствия 1 вытекает непосредственно из оценки (3).

З а м е ч а н и е 5. Дополнительное слагаемое $d(\text{CP}(\Lambda), \text{CP}(\Theta))$, возникающее в оценке (3) из-за перехода к аппроксимации исследуемого распределения посредством распределения суммы пуассоновского числа геометрически распределенных слагаемых, не вносит существенного вклада в (5) и оценку (6) следствия 1.

Доказательство следствия 2 проводится по схеме доказательства следствия из теоремы 3 в работе [13]. Согласно утверждению следствия 1, в данном случае распределение случайной величины $\xi(m, n, s)$ неограниченно сближается со сложным пуассоновским распределением с производящей функцией $\exp\{\lambda(z-1)/(1-zR)\}$. Последнее в свою очередь асимптотически нормально с параметрами $\lambda(1-R)^{-1}$ (среднее) и $\lambda(1+R)(1-R)^{-2}$ (дисперсия). Используя эти свойства, получаем утверждение следствия 2.

Д о к а з а т е л ь с т в о т е о р е м ы 3. Воспользуемся следующей оценкой метода Чена–Стейна (см., например, [14]).

Пусть Γ — произвольный конечный набор индексов, $\tilde{I}_a, a \in \Gamma$, — случайные индикаторы, $\tilde{W} = \sum_{a \in \Gamma} \tilde{I}_a$. Для каждого \tilde{I}_a разделим некоторым образом множество Γ на три непересекающихся множества: $\{a\}$, Γ_a^s , Γ_a^w , и положим

$$\tilde{U}_a = \sum_{b \in \Gamma_a^s} \tilde{I}_b \quad \text{и} \quad \tilde{\varphi}_a = \mathbf{E} \left| \mathbf{E} \{ \tilde{I}_a \mid (\tilde{I}_b : b \in \Gamma_a^w) \} - \mathbf{E} \tilde{I}_a \right|.$$

Теорема 6 ([14]). Пусть $\tilde{\lambda} = \mathbf{E} \tilde{W}$. При любом выборе указанных выше множеств

$$\begin{aligned} d(L(\tilde{W}), \text{Po}(\tilde{\lambda})) &\leq \min \left\{ 1, \frac{1}{\sqrt{\tilde{\lambda}}} \right\} \sum_a \tilde{\varphi}_a \\ &\quad + \frac{1 - e^{-\tilde{\lambda}}}{\tilde{\lambda}} \left(\sum_a \mathbf{E} \tilde{I}_a (\mathbf{E} \tilde{I}_a + \mathbf{E} \tilde{U}_a) + \sum_a \mathbf{E} \tilde{I}_a \tilde{U}_a \right). \end{aligned} \tag{22}$$

В нашем случае $\tilde{\lambda} = \lambda(A)$, $\tilde{I}_a = I\{X_i = Y_j \in A\}$,

$$\Gamma = \{a = (i, j): 1 \leq i \leq n, 1 \leq j \leq m\}.$$

Возьмем

$$\Gamma_a^s = \{b = (i', j') \in \Gamma: i' \neq i, j' = j \text{ или } i' = i, j' \neq j\},$$

$$\Gamma_a^w = \{b = (i', j') \in \Gamma: i' \neq i, j' \neq j\}.$$

Выведем оценки для слагаемых в правой части (22) в нашем случае. Очевидно, что в нашем случае

$$\sum_a \tilde{\varphi}_a = 0, \quad \mathbf{E}\tilde{I}_a = \sum_{k \in A} \mathbf{P}\{X_i = k\} \mathbf{P}\{Y_j = k\}, \quad (23)$$

$$\begin{aligned} & \mathbf{E}\tilde{I}_a(\mathbf{E}\tilde{I}_a + \mathbf{E}\tilde{U}_a) \\ & \leq \sum_{k \in A} \left\{ \mathbf{P}\{X_i = k\} \mathbf{P}\{Y_j = k\} \right. \\ & \quad \times \left. \left(\sum_{i'=1}^m \mathbf{P}\{X_{i'} = k\} \mathbf{P}\{Y_j = k\} + \sum_{j'=1}^n \mathbf{P}\{X_i = k\} \mathbf{P}\{Y_{j'} = k\} \right) \right\}. \end{aligned}$$

Поэтому

$$\sum_a \mathbf{E}\tilde{I}_a(\mathbf{E}\tilde{I}_a + \mathbf{E}\tilde{U}_a) \leq \sum_{k \in A} P_k Q_k (q_k P_k + p_k Q_k). \quad (24)$$

Аналогично

$$\begin{aligned} \mathbf{E}\tilde{I}_a \tilde{U}_a &= \sum_{k \in A} \left(\sum_{\substack{i'=1 \\ i' \neq i}}^m \mathbf{P}\{X_i = X_{i'} = Y_j = k\} + \sum_{\substack{j'=1 \\ j' \neq j}}^n \mathbf{P}\{X_i = Y_j = Y_{j'} = k\} \right) \\ &\leq \sum_{k \in A} \mathbf{P}\{X_i = k\} \mathbf{P}\{Y_j = k\} \left(\sum_{i'=1}^m \mathbf{P}\{X_{i'} = k\} + \sum_{j'=1}^n \mathbf{P}\{Y_{j'} = k\} \right). \end{aligned}$$

Поэтому

$$\sum_{a \in \Gamma} \mathbf{E}\tilde{I}_a \tilde{U}_a \leq \sum_{k \in A} P_k Q_k (P_k + Q_k). \quad (25)$$

Из наших определений и оценок (22)–(25) следует (8). Теорема 3 доказана.

Доказательство теоремы 4. По условиям п. а) теоремы в одной из последовательностей среднее число встретившихся в ней знаков из множества A , т.е. тех знаков, для которых может быть зафиксировано совпадение, стремится к нулю. Значит, к нулю стремятся вероятность наличия хотя бы одного такого знака и вероятность фиксации совпадения знаков.

Утверждение п. б) является прямым следствием теоремы 3.

При доказательстве п. в) ограничимся случаем $KmM^{-1} \rightarrow \mu$. Заметим, что при этом условии число появившихся в первой последовательности знаков из множества A имеет в пределе распределение Пуассона с параметром μ . С вероятностью, стремящейся к единице, все эти элементы различны. В тех же условиях для любого фиксированного набора элементов из алфавита второй последовательности их абсолютные частоты асимптотически независимы и имеют в пределе распределение Пуассона с параметром $\lambda\mu^{-1} = \lim nN^{-1}$. Используя эти свойства, получаем сходимость к указанному распределению. Аналогично рассматриваются остальные случаи, описанные в условиях п. в). Подобным образом доказываются и п. г). Теорема 4 доказана.

Автор признателен А. М. Зубкову за ряд полезных замечаний.

СПИСОК ЛИТЕРАТУРЫ

1. *Waterman M., Vingron M.* Sequence comparison significance and Poisson approximation. — *Statist. Sci.*, 1994, v. 9, № 3, p. 367–381.
2. *Arratia R., Gordon L., Waterman M. S.* An extreme value theory for sequence matching. — *Ann. Statist.*, 1986, v. 14, № 3, p. 971–993.
3. *Новик С. Ю.* Пуассонова аппроксимация числа длинных «повторов» в случайных последовательностях. — *Теория вероятн. и ее примен.*, 1994, т. 39, в. 4, с. 731–742.
4. *Karlin S., Ost F.* Counts of long aligned word matches among random letter sequences. — *Adv. Appl. Probab.*, 1987, v. 19, № 3, p. 293–351.
5. *Karlin S., Ost F.* Maximal length of common words among random letter sequences. — *Ann. Probab.*, 1988, v. 16, № 3, p. 535–563.
6. *Arratia R., Waterman M. S.* Critical phenomena in sequence matching. — *Ann. Probab.*, 1985, v. 13, № 4, p. 1236–1249.
7. *Arratia R., Waterman M. S.* The Erdos–Renyi strong law for pattern matching with a given proportion of mismatches. — *Ann. Probab.*, 1989, v. 17, № 3, p. 1152–1169.
8. *Arratia R., Gordon L., Waterman M. S.* The Erdos–Renyi law in distribution, for coin tossing and sequence matching. — *Ann. Statist.*, 1990, v. 18, № 2, p. 539–570.
9. *Roos M.* Stein's method for compound Poisson approximation: The local approach. — *Ann. Appl. Probab.*, 1994, v. 4, № 4, p. 1177–1187.
10. *Михайлов В. Г.* Оценки точности сложной пуассоновской аппроксимации по методу Стейна–Чена. — *Обзорные прикл. промышл. матем., сер. дискретн. матем.*, 1994, т. 3, в. 4, с. 530–548.
11. *Barbour A. D., Chen L. H. Y., Loh W.-L.* Compound Poisson approximation for non-negative random variables via Stein's method. — *Ann. Probab.*, 1992, v. 20, № 4, p. 1843–1866.
12. *Зубков А. М., Михайлов В. Г.* Предельные распределения случайных величин, связанных с длинными повторениями в последовательности независимых испытаний. — *Теория вероятн. и ее примен.*, 1974, т. 19, в. 1, с. 173–181.
13. *Зубков А. М., Михайлов В. Г.* О повторениях s -цепочек в последовательности независимых величин. — *Теория вероятн. и ее примен.*, 1979, т. 24, в. 2, с. 267–279.
14. *Barbour A. D., Holst L., Janson S.* Poisson Approximation. Oxford: Clarendon Press, 1992, 277 p.
15. *Smith R. L.* Extreme value theory for dependent sequences via the Stein–Chen method of Poisson approximation. — *Stochastic Process. Appl.*, 1988, v. 30, № 2, p. 317–327.

Поступила в редакцию
29.XII.1998

Исправленный вариант
5.VII.1999