Дискретная математика

том 17 выпуск 2 * 2005

УДК 519.2

Пуассоновское приближение для числа повторений значений дискретной функции от цепочек

© 2005 г. А. М. Шойтов

Установлены достаточные условия сходимости к распределению Пуассона распределения числа повторений значений функции от цепочек последовательности независимых одинаково распределенных случайных величин, получены оценки скорости сближения распределений. Выведен ряд следствий из этого результата, в частности, в равновероятной полиномиальной схеме получены пуассоновские предельные теоремы для числа пар неперекрывающихся цепочек с совпадающими частотами встречаемости символов и для числа пар цепочек с одинаковой структурой.

Работа выполнена при поддержке программой Президента Российской Федерации поддержки молодых российских ученых, грант МК-2831.2003.09.

1. Введение

Пусть $\mathbf{X} = \{X_1, \dots, X_n, \dots\}$ есть последовательность независимых одинаково распределенных случайных величин. Для натурального s будем рассматривать s-цепочки $Y_i(s) = (X_i, \dots, X_{i+s-1})$, $i = 1, 2, \dots, n$. Будем считать, что на множестве $\mathbf{R}^s$ задана борелевская функция f, принимающая значения из множества натуральных чисел.

Положим $\mathbf{F} = \{\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_n, \dots\}$, $\mathbf{F}_i = f(Y_i(s))$, $i = 1, 2, \dots$, и определим случайную величину

$$\mathbf{Z}_n(\mathbf{F}) = \sum_{1 \le i_1 < i_2 \le n} \mathbf{I}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2}\},$$

равную числу повторений символов на отрезке длины n последовательности $\mathbf{F}$ (здесь и далее выражение $\mathbf{I}\{\cdot\}$ обозначает индикатор случайного события или неслучайного множества).

Случайная величина $\mathbf{Z}_n(\mathbf{F})$ может быть рассмотрена как индикаторная U-статистика от последовательности $(s-1)$-зависимых случайных величин (см. [5]) или как сумма индикаторов с малым числом зависимостей (см. [8]). Общие условия асимптотической нормальности таких величин получены в работах [6, 7]. В работе [15] доказана многомерная (по различным функциям f) центральная предельная теорема для случайных величин $\mathbf{Z}_n(\mathbf{F})$ в ситуации, когда растет лишь длина последовательности $\mathbf{X}$. Исследованию условий сходимости этой случайной величины к распределению Пуассона для некоторых конкретных функций f посвящены, например, работы [1, 2, 3, 10].

В пункте 3.1 настоящей работы методом Чена–Стейна (см., например, [8, 16]) установлены достаточные условия асимптотической пуассоновости случайной величины $\mathbf{Z}_n(\mathbf{F})$, а также случайной величины $\mathbf{Z}'_n(\mathbf{F})$, равной числу повторений значений функции f от неперекрывающихся s-цепочек отрезка последовательности $\mathbf{X}$ длины $n+s-1$

$$\mathbf{Z}'_n(\mathbf{F}) = \sum_{1 \leqslant i_1 < i_1+s \leqslant i_2 \leqslant n} \mathbf{I}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2}\},$$

в общем случае. Получены оценки скорости сближения распределений.

В качестве следствий этих результатов выведены предельные теоремы Пуассона для числа повторений значений функции от цепочек полиномиальной схемы для ряда конкретных типов функций f. В пункте 3.2 установлены новые достаточные условия сходимости к закону Пуассона распределения числа точных повторений цепочек полиномиальной схемы, которые являются несколько более слабыми, чем известные условия, полученные в работах А. М. Зубкова и В. Г. Михайлова (см., например, [2, 3]). В пункте 3.3 установлена сходимость к закону Пуассона распределения числа пар неперекрывающихся цепочек равновероятной полиномиальной схемы с одинаковыми векторами частот встречаемости символов. В пункте 3.4 приведен известный результат о предельном распределении числа пар цепочек равновероятной полиномиальной схемы с одинаковыми структурами (см. [1, 10]), новое доказательство которого (см. раздел 4) проводится с помощью результатов пункта 3.1.

2. Обозначения

Условимся обозначать $\rho(\Psi, \Gamma)$ расстояние по вариации между распределениями Ψ и Γ. Для распределений Ψ и Γ, сосредоточенных на множестве $\{0, 1, \dots\}$, справедливо равенство

$$\rho(\Psi, \Gamma) = \frac{1}{2} \sum_{m=0}^{\infty} |\Psi\{m\} - \Gamma\{m\}|.$$

Обозначим через $\mathrm{Po}(\lambda)$ распределение Пуассона с параметром λ, а функцию распределения случайной величины γ будем обозначать $\mathcal{L}(\gamma)$.

Пусть $\nu_x(Y_i(s))$ — частота встречаемости символа x, $x \in \mathbf{R}$, в цепочке $Y_i(s)$, то есть

$$\nu_x(Y_i(s)) = \sum_{r=1}^{s} \mathbf{I}\{X_{i+r-1} = x\}.$$

Для $r = 1, \dots, s$ положим

$$\pi_r(f) = \mathbf{P}\{\mathbf{F}_1 = \mathbf{F}_{1+r}\},$$
$$\varkappa_r(f) = \max_{a,b \in \mathbf{N}} \mathbf{P}\{\mathbf{F}_1 = a \mid \mathbf{F}_{1+r} = b\}.$$

Нетрудно видеть, что

$$\pi_s(f) = \sum_{a \in \mathbf{N}} \mathbf{P}^2\{\mathbf{F}_1 = a\},$$
$$\varkappa_s(f) = \max_{a \in \mathbf{N}} \mathbf{P}\{\mathbf{F}_1 = a\}.$$

Кроме того, если случайные величины $X_1, \dots, X_n, \dots$ имеют дискретное распределение и принимают значения из множества $\Omega = \{\omega_1, \omega_2, \dots\}$ с вероятностями $p_{\omega_1}, p_{\omega_2}, \dots$, причем $p_{\omega_i} \neq 0$, $i = 1, 2, \dots$, то величины $\varkappa_r(f)$, $r = 1, \dots, s$, допускают оценки

$$\begin{aligned}\varkappa_r(f) &\leqslant \max_{a \in \mathbf{N}} \sup_{w \in W} \mathbf{P}\{\mathbf{F}_1 = a \mid Y_{1+r}(s-r) = w\} \\ &\leqslant \max_{a \in \mathbf{N}} \mathbf{P}\{\exists w \in W \colon f(X_1, :, X_r, w) = a\},\end{aligned}$$

где

$$W = \{(\omega_{i_1}, \dots, \omega_{i_{s-r}}) \colon \omega_{i_q} \in \Omega,\ q = 1, \dots, s-r\}.$$

Причем, если носитель этого распределения конечен, то в первой оценке вместо точной верхней грани можно брать максимум.

3. Основные результаты

3.1. Общие теоремы

Следующая теорема содержит оценку близости распределения случайной величины $\mathbf{Z}_n(\mathbf{F})$ и распределения Пуассона.

Теорема 1. *Пусть $n \geqslant 2s + 1$. Тогда*

$$\rho\left(\mathcal{L}(\mathbf{Z}'_n(\mathbf{F})),\ \mathrm{Po}\left(\binom{n-2s+1}{2}\pi_s(f)\right)\right) \leqslant sn(5\pi_s(f) + 4\varkappa_s(f)) + 4\sum_{r=1}^{s-1} r\varkappa_r(f),$$

$$\rho(\mathcal{L}(\mathbf{Z}_n(\mathbf{F})), \mathcal{L}(\mathbf{Z}'_n(\mathbf{F}))) \leqslant n\sum_{r=1}^{s-1} \pi_r(f).$$

Следующая теорема описывает достаточные условия сходимости распределений случайных величин $\mathbf{Z}_n(\mathbf{F})$ и $\mathbf{Z}'_n(\mathbf{F})$ к распределению Пуассона.

Теорема 2. *Если при $n \to \infty$ параметры s, N, функция f и параметры случайной последовательности $\mathbf{X}$ меняются так, что $\mathbf{E}\mathbf{Z}'_n(\mathbf{F}) \to \lambda$, $0 < \lambda < \infty$, и выполнено условие*

$$s\varkappa_s(f)\pi_s(f)^{-1/2} + \sum_{r=1}^{s-1} r\varkappa_r(f) \to 0, \tag{1}$$

то при любом фиксированном $m = 0, 1, \dots$

$$\lim_{n\to\infty} \mathbf{P}\{\mathbf{Z}'_n(\mathbf{F}) = m\} = \frac{\lambda^m}{m!}e^{-\lambda},$$

причем

$$\rho(\mathcal{L}(\mathbf{Z}'_n(\mathbf{F})), \mathrm{Po}(\mathbf{E}\mathbf{Z}'_n(\mathbf{F}))) = O\left(s\varkappa_s(f)\pi_s(f)^{-1/2} + \sum_{r=1}^{s-1} r\varkappa_r(f)\right).$$

Если при этом выполнено условие

$$\pi_s(f)^{-1/2}\sum_{r=1}^{s-1} \pi_r(f) \to 0, \tag{2}$$

то при любом фиксированном $m = 0, 1, \ldots$

$$\lim_{n\to\infty} \mathbf{P}\{\mathbf{Z}_n(\mathbf{F}) = m\} = \frac{\lambda^m}{m!}e^{-\lambda},$$

причем

$$\rho(\mathcal{L}(\mathbf{Z}_n(\mathbf{F})), \mathrm{Po}(\mathbf{E}\mathbf{Z}_n(\mathbf{F}))) = O\left(s\varkappa_s(f)\pi_s(f)^{-1/2} + \sum_{r=1}^{s-1}(\pi_r(f)\pi_s(f)^{-1/2} + r\varkappa_r(f))\right).$$

Замечание 1. Из доказательства теоремы 1 следует, что условие (1) можно заменить условием

$$ns\varkappa_s(f) + \sum_{r=1}^{s-1} r\varkappa_r(f) \to 0.$$

Замечание 2. Предельные распределения случайных величин $\mathbf{Z}_n(\mathbf{F})$ и $\mathbf{Z}'_n(\mathbf{F})$ могут не совпадать, если не выполнено условие (2) теоремы 2. В этом случае вероятности совпадения значений функции f от неперекрывающихся цепочек убывают по порядку быстрее, чем значения f от перекрывающихся цепочек. Содержательный пример такой ситуации приведен в разделе 3.3.

Замечание 3. Рассмотрим отношение эквивалентности $\tilde{f}$ на множестве s-цепочек, заданное следующим образом: две цепочки принадлежат одному классу $\tilde{f}$-эквивалентности, если функция f принимает на них одинаковое значение. Нетрудно видеть, что функции f и f', для которых отношения $\tilde{f}$ и $\tilde{f}'$ совпадают, определяют одинаковые случайные величины $\mathbf{Z}_n(\mathbf{F})$ и $\mathbf{Z}'_n(\mathbf{F})$.

Следующая теорема описывает достаточные условия сходимости случайных величин $\mathbf{Z}_n(\mathbf{F})$ к нормальному распределению.

Теорема 3. *Если при* $n \to \infty$ *параметры* s, N, *функция* f *и параметры вероятностной схемы* $\mathbf{X}$ *меняются так, что* $\mathbf{E}\mathbf{Z}_n(\mathbf{F}) \to \infty$ *и выполнено условие*

$$sn\varkappa_s(f) + \sum_{r=1}^{s-1}(r\varkappa_r(f) + \pi_r(f)) \to 0,$$

то распределение случайной величины

$$\frac{\mathbf{Z}_n(\mathbf{F}) - \mathbf{E}\mathbf{Z}_n(\mathbf{F})}{\sqrt{\mathbf{D}\mathbf{Z}_n(\mathbf{F})}}$$

сходится к стандартному нормальному распределению.

Приведем примеры использования теоремы 2 для описания достаточных условий сходимости случайных величин $\mathbf{Z}_n(\mathbf{F})$ и $\mathbf{Z}'_n(\mathbf{F})$ к распределению Пуассона для некоторых частных случаев функций f.

3.2. Совпадающие цепочки

Будем предполагать, что последовательность $\mathbf{X}$ получена по полиномиальной схеме испытаний с вероятностями исходов $p_m = \mathbf{P}\{X_1 = m\}$, $m = 1, 2, \ldots$ Рассмотрим случайную величину

$$\delta_n = \sum_{1 \leq i_1 < i_2 \leq n} \mathbf{I}\{Y_{i_1}(s) = Y_{i_2}(s)\},$$

равную числу точных повторений s-цепочек на отрезке последовательности $\mathbf{X}$ длины $n + s - 1$. Подобные случайные величины исследовались в работах А. М. Зубкова и В. Г. Михайлова (см., например, [2, 3]). Одним из частных случаев теоремы 2 является утверждение о сходимости распределения δ_n к распределению Пуассона.

Следствие 1. *Если при $n \to \infty$ параметр s и вероятности $p_1, p_2, \ldots$ меняются так, что*

$$\max_{m \in \mathbf{N}} p_m \to 0, \quad \mathbf{E}\delta_n \to \lambda, \quad 0 < \lambda < \infty,$$

и выполнено условие

$$ns(\max_{m \in \mathbf{N}} p_m)^s \to 0, \tag{3}$$

то при любом фиксированном $m = 0, 1, \ldots$

$$\lim_{n \to \infty} \mathbf{P}\{\delta_n = m\} = \frac{\lambda^m}{m!} e^{-\lambda},$$

причем

$$\rho(\mathcal{L}(\delta_n(\mathbf{F})), \mathrm{Po}(\mathbf{E}\delta_n(\mathbf{F}))) = O(ns(\max_{m \in \mathbf{N}} p_m)^s + \max_{m \in \mathbf{N}} p_m).$$

Замечание 4. В условиях следствия 1

$$\mathbf{E}\delta_n = \binom{n}{2} P_2^s (1 + o(1)), \qquad P_2 = \sum_{m=1}^{\infty} p_m^2.$$

Замечание 5. Условия следствия 1 несколько слабее известных условий работ [2] и [3] применительно к пуассоновской зоне сходимости распределения δ_n. В [2] для сходимости распределения δ_n к закону Пуассона вместо условия (3) следствия 1 требуется выполнение условия

$$ns^t(\max_{m \in \mathbf{N}} p_m)^s \to 0, \qquad t = 1, 2, \ldots$$

Аналогичное условие из [3] выглядят следующим образом:

$$ns^2(\max_{m \in \mathbf{N}} p_m)^s (P_2^{-1} \max_{m \in \mathbf{N}} p_m)^s \to 0.$$

В [12] получены близкие по форме к условиям следствия 1 достаточные условия сходимости к закону Пуассона распределения числа пар повторений цепочек полиномиальной схемы, которые не могут быть продолжены влево.

Замечание 6. Близкой к задаче о повторениях цепочек в случайной последовательности является задача о совпадении цепочек в двух случайных последовательностях. Асимптотическое поведение числа пар совпадающих цепочек двух полиномиальных схем исследовалось, например, в [9].

3.3. Цепочки с совпадающими частотами встречаемости символов

Будем предполагать, что последовательность $\mathbf{X}$ получена по равновероятной полиномиальной схеме с исходами из множества $A = \{1, \dots, N\}$. Рассмотрим случайную величину

$$\zeta_n' = \sum_{1 \leqslant i_1 < i_1 + s \leqslant i_2 \leqslant n} \mathbf{I}\{\nu_r(Y_{i_1}(s)) = \nu_r(Y_{i_2}(s)),\ r = 1, \dots, s\},$$

равную числу пар неперекрывающихся s-цепочек с совпадающими частотами встречаемости символов на отрезке последовательности $\mathbf{X}$ длины $n + s - 1$.

Замечание 7. Совпадение частот встречаемости символов s-цепочек задает на множестве A^s отношение перестановочной эквивалентности (см. обзор [11]), которое является частным случаем $\mathbf{G}$-эквивалентности (см. монографию [13]). Другими словами, совпадение частот встречаемости символов в двух цепочках эквивалентно совпадению этих цепочек с точностью до перестановки их символов.

Следствие 2. *Если при $n \to \infty$ параметры N, s меняются так, что $N \to \infty$, $\mathbf{E}\zeta_n' \to \lambda$, $0 < \lambda < \infty$, то при любом фиксированном $m = 0, 1, \dots$*

$$\lim_{n \to \infty} \mathbf{P}\{\zeta_n' = m\} = \frac{\lambda^m}{m!} e^{-\lambda}.$$

Замечание 8. В условиях следствия 2 при $s \geqslant 2$ асимптотическое поведение случайной величины ζ_n' и случайной величины ζ_n, равной числу пар s-цепочек с совпадающими частотами встречаемости символов на отрезке последовательности $\mathbf{X}$ длины $n + s - 1$, различается. Можно доказать, что асимптотическое поведение ζ_n определяется парами соседних перекрывающихся цепочек, и в условиях следствия 2 величина $\mathbf{E}\zeta_n$ стремится к бесконечности (см. [11]).

Замечание 9. В условиях следствия 2

$$\mathbf{E}\zeta_n' = \binom{n - s + 1}{2} \pi_s, \qquad \pi_s = \sum_{k_1 + \dots + k_N = s,\, k_1, \dots, k_N \geqslant 0} \left(\frac{s!}{k_1! \dots k_N!} \right)^2.$$

Асимптотическое поведение величин π_s для различных вариантов изменения параметров s и N исследовано в [14].

3.4. Цепочки с одинаковой структурой

Как и разделе 3.3, будем предполагать, что последовательность $\mathbf{X}$ получена по равновероятной полиномиальной схеме с исходами из множества $A = \{1, \dots, N\}$. Будем говорить, что две цепочки $Y_{i_1}(s)$ и $Y_{j_1}(s)$ обладают одинаковой структурой (структурно эквивалентны, см. обзор [11]), если

$$X_{i_1 + r - 1} = X_{i_1 + q - 1} \iff X_{j_1 + r - 1} = X_{j_1 + q - 1}, \qquad r, q = 1, \dots, s.$$

Замечание 10. Отношение эквивалентности, для которого к одинаковым классам отнесены цепочки с одинаковой структурой называется (см. обзор [11]) отношением структурной эквивалентности. Нетрудно видеть, что структурная эквивалентность пары цепочек равносильна их совпадению с точностью до покоординатной замены символов с помощью некоторой подстановки из симметрической группы подстановок, заданной на множестве A, поэтому отношение структурной эквивалентности является частным случаем $\mathbf{H}$-эквивалентности (см. монографию [13]).

Обозначим через $S(Y_{i_1}(s))$ класс структурной эквивалентности, содержащий цепочку $Y_{i_1}(s)$, и определим случайную величину

$$\xi_n = \sum_{1 \leqslant i_1 < i_2 \leqslant n} \mathbf{I}\{S(Y_{i_1}(s)) = S(Y_{j_1}(s))\},$$

равную числу пар s-цепочек с одинаковой структурой в отрезке последовательности $\mathbf{X}$ длины $n + s - 1$.

Следствие 3. *Если при $n \to \infty$ параметры N, s меняются так, что $N \to \infty$, $\mathbf{E}\xi_n \to \lambda$, $0 < \lambda < \infty$, и выполнено условие*

$$s/N - \ln N \to \infty, \tag{4}$$

то при любом фиксированном $m = 0, 1, \ldots$

$$\lim_{n \to \infty} \mathbf{P}\{\xi_n = m\} = \frac{\lambda^m}{m!} e^{-\lambda}.$$

Замечание 11. Более тонкими рассуждениями условие (4) может быть несколько ослаблено (см., например, [1, 10]), однако ответ на вопрос о необходимости подобного рода условия в пуассоновской предельной теореме для ξ_n в настоящее время неясен.

Замечание 12. В условиях следствия 3

$$\mathbf{E}\xi_n = \frac{\binom{n}{2}}{N^{2s}} \sum_{m=1}^{N} \sigma(s, m)((N)_m)^2 = \binom{n}{2} \frac{N!}{N^s}(1 + o(1)).$$

4. Доказательства

Доказательство теоремы 1. Вторая оценка очевидным образом следует из неравенства

$$\rho(\mathcal{L}(\mathbf{Z}_n(\mathbf{F})), \mathcal{L}(\mathbf{Z}'_n(\mathbf{F}))) \leqslant \mathbf{E}\mathbf{Z}_n(\mathbf{F}) - \mathbf{E}\mathbf{Z}'_n(\mathbf{F}).$$

Докажем первую оценку. Положим

$$J_r = \{i = (i_1, i_2) \colon 1 \leqslant i_1 < i_1 + rn \leqslant i_2 \leqslant n\}, \qquad r = 1, 2,$$

и определим случайную величину

$$\mathbf{Z}''_n(\mathbf{F}) = \sum_{(i_1, i_2) \in J_2} \mathbf{I}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2}\}.$$

Заметим, что

$$\begin{aligned}\mathbf{P}\{\mathbf{Z}'_n(\mathbf{F}) \neq \mathbf{Z}''_n(\mathbf{F})\} &\leqslant \mathbf{E} \sum_{(i_1, i_2) \in J_1} \mathbf{I}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2}\} - \mathbf{E} \sum_{(i_1, i_2) \in J_2} \mathbf{I}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2}\} \\ &= \sum_{(i_1, i_2) \in J_1 \setminus J_2} \mathbf{P}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2}\} = |J_1 \setminus J_2| \pi(f) \leqslant sn\pi_s(f),\end{aligned}$$

так как

$$|J_1| = \binom{n-s+1}{2}, \quad |J_2| = \binom{n-2s+1}{2}, \quad |J_1 \setminus J_2| = s(n-(3s-1)/2).$$

Пусть $\lambda_n = \binom{n-2s+1}{2}\pi_s(\mathbf{F})$. Тогда по свойствам метрики полной вариации

$$\begin{aligned} d(\mathcal{L}(\mathbf{Z}'_n(\mathbf{F}), \mathrm{Po}(\lambda_n)) &\leqslant d(\mathcal{L}(\mathbf{Z}''_n(\mathbf{F}), \mathrm{Po}(\lambda_n)) + d(\mathcal{L}(\mathbf{Z}''_n(\mathbf{F}), \mathcal{L}(\mathbf{Z}'_n(\mathbf{F})) \\ &\leqslant d(\mathcal{L}(\mathbf{Z}''_n(\mathbf{F}), \mathrm{Po}(\lambda_n)) + \mathbf{P}\{\mathbf{Z}'_n(\mathbf{F}) \neq \mathbf{Z}''_n(\mathbf{F})\} \\ &\leqslant d(\mathcal{L}(\mathbf{Z}''_n(\mathbf{F}), \mathrm{Po}(\lambda_n)) + sn\pi_s(f). \end{aligned}$$

Далее оценим расстояние по вариации между распределением случайной величины $\mathcal{L}(\mathbf{Z}''_n(\mathbf{F})$ и распределением Пуассона с параметром λ_n. Будем использовать метод Чена–Стейна (см. [8, 16]) оценки точности пуассоновской аппроксимации.

Для любой пары $i = (i_1, i_2)$ из J_2 определим множество $J_2(i)$ пар $j = (j_1, j_2)$ из J_2, $j \neq i$, таких, что хотя бы один из отрезков $[i_1, i_1 + s - 1]$, $[i_2, i_2 + n - 1]$ пересекается хотя бы с одним из отрезков $[j_1, j_1 + n - 1]$, $[j_2, j_2 + n - 1]$.

Согласно теореме 2.N работы [16] для расстояния по вариации между распределением случайной величины $\mathbf{Z}''_n(\mathbf{F})$ и распределением Пуассона с параметром λ_n (нетрудно видеть, что $\mathbf{E}\mathbf{Z}''_n(\mathbf{F}) = \lambda_n$) справедлива оценка сверху

$$\begin{aligned} d(\mathcal{L}(\mathbf{Z}''_n(\mathbf{F}), \mathrm{Po}(\lambda_n))) \leqslant \frac{1-e^{-\lambda_n}}{\lambda_n} \Bigg(&\sum_{i \in J_2} (\mathbf{E}\eta_{\mathbf{F}}(i)(\mathbf{E}\eta_{\mathbf{F}}(i) + \sum_{j \in J_2(i)} \mathbf{E}\eta_{\mathbf{F}}(j))) \\ &+ \sum_{i \in J_2} \sum_{j \in J_2(i)} \mathbf{E}(\eta_{\mathbf{F}}(\bar{j})\eta_{\mathbf{F}}(i)) \Bigg), \end{aligned} \tag{5}$$

где

$$\eta_{\mathbf{F}}(k) = \mathbf{I}(\mathbf{F}_{k_1} = \mathbf{F}_{k_2}), \qquad k = (k_1, k_2) \in J_2.$$

Тогда, учитывая, что

$$|J_2(i)| \leqslant 2(2s-1)(n-2s) - 2$$

для всех $i \in J_2$, получаем, что

$$\sum_{i \in J_2} \left(\mathbf{E}\eta_{\mathbf{F}}(i) \left(\mathbf{E}\eta_{\mathbf{F}}(i) + \sum_{j \in J_2(i)} \mathbf{E}\eta_{\mathbf{F}}(j) \right) \right) \leqslant \lambda_n 4sn\pi_s(f).$$

Далее, для любых $i, j \in J_2$ определим множества

$$U_2(i) = [i_1, i_1 + s - 1] \cup [i_2, i_2 + s - 1], \qquad U_2(i, j) = U_2(i) \cup U_2(j).$$

Если $i \in J_2$ и $j \in J_2(i)$, то множество $U_2(i, j)$ состоит из двух или из трех компонент связности. Обозначим через I_m множество таких пар (i, j), $i \in J_2$, $j \in J_2(i)$, что $U_2(i, j)$ состоит из m компонент связности, $m = 2, 3$.

Пусть $(i, j) \in I_3$. В этом случае ровно одна из цепочек $Y_{i_1}(s)$, $Y_{i_2}(s)$ пересекается ровно с одной из цепочек $Y_{j_1}(s)$, $Y_{j_2}(s)$, пусть, не ограничивая общности, это будут цепочки $Y_{i_1}(s)$ и $Y_{j_1}(s)$, причем $i_1 \geqslant j_1$. Тогда

$$\begin{aligned}
\mathbf{P}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2}, \mathbf{F}_{j_1} = \mathbf{F}_{j_2}\} &= \sum_{b\in\mathbf{N}}(\mathbf{P}\{\mathbf{F}_{j_1} = \mathbf{F}_{j_2} \mid \mathbf{F}_{i_1} = \mathbf{F}_{i_2} = b\}\mathbf{P}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2} = b\}) \\
&\leqslant \sum_{b\in\mathbf{N}}(\max_{a\in\mathbf{N}}\mathbf{P}\{\mathbf{F}_{j_2} = a \mid \mathbf{F}_{i_1} = \mathbf{F}_{i_2} = b\}\mathbf{P}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2} = b\}) \\
&\leqslant \sum_{b\in\mathbf{N}}(\max_{a\in\mathbf{N}}\mathbf{P}\{\mathbf{F}_{j_2} = a\}\mathbf{P}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2} = b\}) \\
&= \varkappa_s(f)\sum_{b\in\mathbf{N}}\mathbf{P}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2} = b\} = \varkappa_s(f)\pi_s(f).
\end{aligned}$$

Тогда

$$\sum_{(i,j)\in I_3}\mathbf{E}(\eta_{\mathbf{F}}(j)\eta_{\mathbf{F}}(i)) \leqslant \sum_{(i,j)\in I_3}\pi_s(f)\varkappa_s(f) \leqslant 4\lambda_n sn\varkappa_s(f).$$

Если $(i, j) \in I_2$, то имеются две пары пересекающихся цепочек $Y_{i_1}(s)$, $Y_{j_1}(s)$ и $Y_{i_2}(s)$, $Y_{j_2}(s)$. Пусть, не ограничивая общности, $i_1 - j_1 \geqslant 1$, $i_2 - j_2 \geqslant 0$ и $i_1 - j_1 \geqslant i_2 - j_2$. Тогда

$$\begin{aligned}
\mathbf{P}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2}, \mathbf{F}_{j_1} = \mathbf{F}_{j_2}\} &= \sum_{b\in\mathbf{N}}(\mathbf{P}\{\mathbf{F}_{j_1} = \mathbf{F}_{j_2} \mid \mathbf{F}_{i_1} = \mathbf{F}_{i_2} = b\}\mathbf{P}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2} = b\}) \\
&\leqslant \max_{b\in\mathbf{N}}\mathbf{P}\{\mathbf{F}_{j_1} = \mathbf{F}_{j_2} \mid \mathbf{F}_{i_1} = \mathbf{F}_{i_2} = b\}\sum_{b\in\mathbf{N}}\mathbf{P}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2} = b\} \\
&\leqslant \max_{b\in\mathbf{N}}\mathbf{P}\{\mathbf{F}_{j_1} = \mathbf{F}_{j_2} \mid \mathbf{F}_{i_1} = \mathbf{F}_{i_2} = b\}\sum_{b\in\mathbf{N}}\mathbf{P}\{\mathbf{F}_{i_1} = \mathbf{F}_{i_2} = b\} \\
&\leqslant \max_{b\in\mathbf{N}}\max_{a\in\mathbf{N}}\mathbf{P}\{\mathbf{F}_{j_1} = a \mid \mathbf{F}_{i_1} = \mathbf{F}_{i_2} = b\}\pi_s(f) \\
&= \pi_s(f)\max_{a,b\in\mathbf{N}}\mathbf{P}\{\mathbf{F}_{j_1} = a \mid \mathbf{F}_{i_1} = b\}.
\end{aligned}$$

Следовательно,

$$\sum_{(i,j)\in I_2}\mathbf{E}(\eta_{\mathbf{F}}(j)\eta_{\mathbf{F}}(i)) \leqslant 4\pi_s(f)|J_2|\sum_{r=1}^{s-1}r\varkappa_r(f) = 4\lambda_n\sum_{r=1}^{s-1}r\varkappa_r(f).$$

Подставляя полученные оценки в неравенство (5), получаем первую оценку теоремы 1. Теорема 1 доказана.

Доказательство теоремы 2. Так как $\varkappa_s(f) \geqslant \pi_s(f)$, из условия (1) следует, что $s\pi_s(f)^{1/2} \to 0$, что с учетом условия $\mathbf{E}\mathbf{Z}'_n(\mathbf{F}) \to \lambda$, $0 < \lambda < \infty$, теоремы 2, дает соотношение $s = o(n)$.

Применяя первую оценку теоремы 1, получаем, что

$$\rho\left(\mathcal{L}(\mathbf{Z}'_n(\mathbf{F})), \mathrm{Po}\left(\binom{n-2s+1}{2}\pi_s(f)\right)\right) = O\left(s\varkappa_s(f)\pi_s(f)^{-1/2} + \sum_{r=1}^{s-1}r\varkappa_r(f)\right).$$

Для завершения доказательства остается заметить, что

$$\rho\left(\mathrm{Po}(\mathbf{E}\mathbf{Z}'_n(\mathbf{F})), \mathrm{Po}\left(\binom{n-2s+1}{2}\pi_s(f)\right)\right) \leqslant \left|\mathbf{E}\mathbf{Z}'_n(\mathbf{F}) - \binom{n-2s+1}{2}\pi_s(f)\right| = O(s\varkappa_s(f)\pi_s(f)^{-1/2}).$$

Теорема 2 доказана.

Доказательство теоремы 3. Доказательство теоремы 3 проводится с использованием теоремы 1 аналогично доказательству теоремы 2 с учетом сходимости распределения Пуассона с параметром λ к нормальному распределению при $\lambda \to \infty$.

Доказательство следствия 1. Применим теорему 2 с учетом замечания 1. Зададим функцию f, зависящую от цепочек полиномиальной схемы, так, чтобы она принимала одинаковые натуральные значения только для одинаковых цепочек. Заметим, что для так выбранной функции f

$$\varkappa_s(f) = (\max_{m\in\mathbf{N}} p_m)^s, \quad \pi_s(f) = \left(\sum_{m=1}^{\infty} p_m^2\right)^s, \quad \varkappa_r(f) = (\max_{m\in\mathbf{N}} p_m)^r.$$

Кроме того, нетрудно доказывается (см. [2]), что

$$\pi_r(f) \leqslant (\max_{m\in\mathbf{N}} p_m)^s.$$

Тогда, учитывая замечание 1, условие (1) теоремы 2 можно заменить условием

$$ns(\max_{m\in\mathbf{N}} p_m)^s + \sum_{r=1}^{s-1} r(\max_{m\in\mathbf{N}} p_m)^r \to 0.$$

Так как при $\max_{m\in\mathbf{N}} p_m \to 0$

$$\sum_{r=1}^{s-1} r(\max_{m\in\mathbf{N}} p_m)^r \leqslant \sum_{r=1}^{\infty} r(\max_{m\in\mathbf{N}} p_m)^r = \frac{\max_{m\in\mathbf{N}} p_m}{(1-\max_{m\in\mathbf{N}} p_m)^2} = O(\max_{m\in\mathbf{N}} p_m),$$

для завершения доказательства остается заметить, что в рассматриваемом случае условие (2) теоремы 2 следует из условия (1).

Следствие 1 доказано.

Доказательство следствия 2. Зададим функцию f, зависящую от цепочек полиномиальной схемы, так, чтобы она принимала одинаковые натуральные значения только для цепочек с совпадающими частотами встречаемости символов. Докажем, что для такой функции f выполнено условие (1) теоремы 2.

Заметим, что для равновероятной последовательности $\mathbf{X}$ максимальное значение вероятности

$$\mathbf{P}\{f(Y_1(s)) = a \mid f(Y_{1+r}(s)) = b\}$$

достигается, если в векторе частот $(\nu_1(Y_1(r)), \dots, \nu_N(Y_r(r)))$ величины $\nu_q(Y_r(r))$ и $\nu_t(Y_1(r))$ отличаются не более чем на единицу для всех $q, t = 1, \dots, N$. Поэтому

$$\varkappa_r(f) = N^{-r}\frac{r!}{(\lfloor r/N\rfloor!)^N(\lfloor r/N\rfloor+1)^{r-N\lfloor r/N\rfloor}}, \tag{6}$$

где $\lfloor x \rfloor$ — ближайшее целое снизу для x. Нетрудно видеть, что для величины $\varkappa_s(f)$ при $s = r$ справедливо такое же представление, поэтому далее эти величины будем отождествлять.

Оценим сверху величины $s^2\varkappa_s^2(f)\pi_s^{-1}(f)$, $s = 1, 2, \ldots$

Нетрудно видеть, что

$$\begin{aligned} \varkappa_1^2(f)\pi_1^{-1}(f) &= 1/N, \\ 2^2\varkappa_2^2(f)\pi_2^{-1}(f) &\leqslant 16/N^2. \end{aligned} \tag{7}$$

Если $3 \leqslant s \leqslant N - 3$, то существует по крайней мере $\binom{N}{s}$ различных классов перестановочной эквивалентности, имеющих вероятность $\varkappa_s(f)$, и следовательно, $\pi_s(f) \geqslant \binom{N}{s}\varkappa_s^2(f)$ и

$$s^2\varkappa_s^2(f)\pi_s^{-1}(f) \leqslant s^2/\binom{N}{s}. \tag{8}$$

Пусть $N - 2 \leqslant s \leqslant N^7$. Тогда цепочки, содержащиеся в каком-нибудь наиболее вероятном классе K_0, будут содержать по крайней мере $N - 2$ различных символа из множества A. Для $N \geqslant 18$ выберем из этих символов два неупорядоченных набора по 8 различных символов каждый. Существует по крайней мере $\binom{N-2}{8}\binom{N-10}{8}$ способов выбора таких наборов. Рассмотрим класс перестановочной эквивалентности, полученный из класса K_0 заменой в цепочках этого класса символов первого набора на символы из второго набора. Из формулы (6) следует, что вероятность попадания цепочки в любой полученный после такой замены класс будет меньше, чем вероятность попадания в класс K_0, не более чем в 2^8 раза (не более 2 раз на каждую замену). Тогда

$$\begin{aligned} s^2\varkappa_s^2(f)\pi_s^{-1}(f) &\leqslant s^2\varkappa_s^2(f)\left(\binom{N-2}{16}\binom{N-10}{8}2^{-16}\varkappa_s^2(f)\right)^{-1} \\ &\leqslant \frac{2^{16}s^2}{\binom{N-2}{16}\binom{N-10}{8}}. \end{aligned} \tag{9}$$

Пусть теперь $s > N^7$. Для достаточно большого N выберем из множества A пять пар различных символов, например, (1, 2), (3, 4), (5, 6) (7, 8) и (9, 10). Будем независимо уменьшать в векторе частот наиболее вероятного класса на единицу количество символов 1, 3, 5, 7 и 9, соответственно увеличивая количество символов 2, 4, 6, 8 и 10. Этот процесс будем продолжать по каждой паре символов независимо до момента, когда будет удалено $\lfloor\sqrt{\lfloor s/N \rfloor}\rfloor$ символов. Все полученные при этом векторы частот различны, а общее их число равно $\lfloor\sqrt{\lfloor s/N \rfloor}\rfloor^5$. Оценим снизу вероятности классов эквивалентности, соответствующих этим векторам частот. Вероятности классов больше всего уменьшатся, если удаляется по $\lfloor\sqrt{\lfloor s/N \rfloor}\rfloor$ символов 1, 3, 5, 7 и 9 (с соответствующим увеличением чисел символов 2, 4, 6, 8 и 10). Для такого наименее вероятного класса K_1 справедливо равенство

$$\begin{aligned} &\varkappa_s(f)\mathbf{P}^{-1}\{(X_1, \ldots, X_s) \in K_1\} \\ &\qquad = \left(\frac{(\lfloor s/N \rfloor - \lfloor\sqrt{\lfloor s/N \rfloor}\rfloor)!(\lfloor s/N \rfloor + \lfloor\sqrt{\lfloor s/N \rfloor}\rfloor + 1)!}{\lfloor s/N \rfloor!(\lfloor s/N \rfloor + 1)!}\right)^5. \end{aligned}$$

Заметим, что выражение, стоящее в правой части последнего равенства, не возрастает с ростом s. Кроме того, применяя при $\lfloor s/N \rfloor \to \infty$ формулу Стирлинга для факториалов в этом выражении, получаем, что

$$\lim_{N, \lfloor s/N \rfloor \to \infty} \varkappa_s(f) \mathbf{P}^{-1}\{(X_1, \dots, X_s) \in K_0\} = e^{10}.$$

Поэтому для достаточно большого N и любого $s > N^7$ справедлива оценка

$$\varkappa_s(f) \mathbf{P}^{-1}\{(X_1, \dots, X_s) \in K_0\} \leqslant 2e^{10},$$

следовательно,

$$\begin{aligned} s^2 \varkappa_s^2(f) \pi_s^{-1}(f) &\leqslant s^2 \varkappa_s^2(f)((1/4)e^{-20} \varkappa_s^2(f) \lfloor \sqrt{\lfloor s/N \rfloor} \rfloor^5)^{-1} \\ &= \frac{4e^{20} s^2}{\lfloor \sqrt{\lfloor s/N \rfloor} \rfloor^{5/2}}. \end{aligned} \tag{10}$$

Из оценок (7)–(10) с учетом очевидных неравенств

$$\sum_{r=1}^{s-1} r \varkappa_r(f) \leqslant \sum_{r=1}^{\infty} r \varkappa_r(f),$$
$$\varkappa_s(f) \geqslant \pi_s(f)$$

следует справедливость (1).

Следствие 2 доказано.

Доказательство следствия 3. Пусть $\mu(Y_i(r))$ — число различных символов встретившихся в цепочке $Y_i(r)$, $r = 1, 2, \dots$ Определим случайную величину

$$\tilde{\xi}_n = \sum_{1 \leqslant i_1 < i_2 \leqslant n} \mathbf{I}\{S(Y_{i_1}(s)) = S(Y_{i_2}(s)),\ \mu(Y_{i_2}(s - [N^{1/2} \ln N])) = N\}.$$

Заметим, что для неперекрывающихся пар цепочек $Y_{i_1}(s)$ и $Y_{i_2}(s)$

$$\begin{aligned} \mathbf{P}\{S(Y_{i_1}(s)) = S(Y_{i_2}(s)),\ \mu(Y_{i_1}(s - [N^{1/2} \ln N])) = N\} \\ = \frac{N!}{N^s} \mathbf{P}\{\mu(Y_{i_2}(s - [N^{1/2} \ln N])) = N\} \end{aligned}$$

и

$$\begin{aligned} \mathbf{P}\{S(Y_{i_1}(s)) = S(Y_{i_2}(s)),\ \mu(Y_{i_2}(s - [N^{1/2} \ln N])) < N\} \\ \leqslant \frac{N!}{N^s} \mathbf{P}\{\mu(Y_{i_2}(s - [N^{1/2} \ln N])) < N\}. \end{aligned}$$

Используя теорему 1 из книги [4] и условие (4) следствия 3, получаем, что

$$\begin{aligned} \mathbf{P}\{\mu(Y_{i_2}(s - [N^{1/2} \ln N])) < N\} &= \mathbf{P}\{\mu_0(Y_{i_2}(s - [N^{1/2} \ln N])) > 0\} \\ &\leqslant \mathbf{E}\mu_0(Y_{i_2}(s - [N^{1/2} \ln N])) \\ &\leqslant N \exp(-(s - [N^{1/2} \ln N])/N) = o(1), \end{aligned}$$

здесь $\mu_0(y)$ — случайная величина, равная числу не встретившихся в векторе y символов.

Следовательно,

$$\mathbf{P}\{\xi'_n \neq \tilde{\xi}'_n\} \leqslant \mathbf{E}\xi'_n - \mathbf{E}\tilde{\xi}'_n = o(1),$$

где

$$\tilde{\xi}'_n = \sum_{1 \leqslant i_1 < i_1 + s \leqslant i_2 \leqslant n} \mathbf{I}\{S(Y_{i_1}(s)) = S(Y_{i_2}(s)),\ \mu(Y_{i_2}(s - [N^{1/2} \ln N])) = N\}.$$

Далее проверим выполнение условия (1) теоремы 2 для случайной величины $\tilde{\xi}'_n$, определив функцию f так, чтобы она принимала одинаковые значения только для цепочек с одинаковой структурой, в первых $s - [N^{1/2} \ln N]$ координатах которых содержатся все символы алфавита A.

Заметим, что для такой функции f

$$\varkappa_s(f) = \frac{N!}{N^s}(1 + o(1)),$$
$$\pi_s(f) = \frac{N!}{N^s}(1 + o(1)),$$

поэтому

$$s\varkappa_s(f)\pi_s(f)^{-1/2} = s\left(\frac{N!}{N^s}\right)^{1/2}(1 + o(1)) = o(1).$$

Оценим величины $\varkappa_r(f)$, $r = 1, \dots, s-1$. Заметим, что если $r \leqslant [N^{1/2} \ln N]$, то $\varkappa_r(f) = N^{-r}$, а если $r > N$, то $\varkappa_r(\mathbf{F}) \leqslant N!N^{-r}$. Пусть $[N^{1/2} \ln N] < r < N + 1$, при этом условии

$$\varkappa_r(f) \leqslant \frac{N(N-1)\dots(N-r+1)}{N^{-r}} = \frac{N!}{(N-r)!N^{-r}}. \tag{11}$$

Оценивая значение выражения, стоящего в правой части (11), при $r = [N^{1/2} \ln N] + 1$ с использованием формулы Стирлинга получаем, что

$$\frac{N!}{(N - [N^{1/2} \ln N] - 1)!N^{-[N^{1/2} \ln N]-1}} = O(\exp(-\ln^2 N)).$$

Замечая, что величины, стоящие в правой части (11), убывают с ростом r, получаем, что

$$\sum_{r=1}^{s-1} r\varkappa_r(f) \leqslant \sum_{r=1}^{[N^{1/2} \ln N]} rN^{-r} + \sum_{r=[N^{1/2} \ln N]+1}^{N} rO(N^{-\ln N}) + \sum_{r=N+1}^{s-1} rN!N^{-r} = o(1).$$

Для завершения доказательства следствия 3 остается заметить (см. [1]), что $\pi_r(f) \leqslant N!\,N^{-s}$, и следовательно,

$$\pi_s(f)^{-1/2} \sum_{r=1}^{s-1} \pi_r(f) \leqslant (N!N^{-s}(1 + o(1)))^{-1/2} \sum_{r=1}^{s-1} N!N^{-s} = o(1).$$

Следствие 3 доказано.

Автор признателен В. Г. Михайлову за полезные замечания.

Список литературы

1. Буравлев С. М., Повторения с точностью до перестановок в последовательности независимых испытаний. *Дискретная математика* (1999) **11**, №1, 53–75.
2. Зубков А. М., Михайлов В. Г., Предельные распределения случайных величин, связанных с длинными повторениями в последовательности независимых испытаний. *Теория вероятностей и ее применения* (1974) **19**, №1, 173–181.
3. Зубков А. М., Михайлов В. Г., О повторениях s-цепочек в последовательности независимых величин. *Теория вероятностей и ее применения* (1979) **24**, №2, 267–279.
4. Колчин В. Ф., Севастьянов Б. А., Чистяков В. П., *Случайные размещения.* Наука, Москва, 1976.
5. Королюк В. С., Боровских Ю. В., *Теория U-статистик.* Наукова думка, Киев, 1989.
6. Михайлов В. Г., Асимптотическая нормальность в схеме конечно-зависимого размещения частиц по ячейкам. *Матем. сб.* (1982) **119**, №4, 509–520.
7. Михайлов В. Г., Асимптотическая нормальность разделимых статистик от частот m-цепочек. *Дискретная математика* (1989) **1**, №4, 92–103.
8. Михайлов В. Г., Явные оценки в предельных теоремах для сумм случайных индикаторов. *Обозрение прикладной и промышленной математики* (1994) **1**, №4, 580–617.
9. Михайлов В. Г., Оценка точности сложной пуассоновской аппроксимации для распределения числа совпадающих цепочек. *Теория вероятностей и ее применения* (2003) **46**, №4, 713–723.
10. Михайлов В. Г., Предельные теоремы Пуассона для числа пар H-связанных цепочек. *Труды по дискретной математике* (2003) **7**, 138–155.
11. Михайлов В. Г., Шойтов А. М., Структурная эквивалентность s-цепочек в случайных дискретных последовательностях. *Дискретная математика* (2003) **15**, №4, 7–34.
12. Новак С. Ю., Пуассонова аппроксимация числа длинных повторов в случайных последовательностях. *Теория вероятностей и ее применения* (1994) **39**, №4, 731–742.
13. Сачков В. Н., *Введение в комбинаторные методы дискретной математики.* Наука, Москва, 1982.
14. Тимашёв А. Н., О вероятности совпадения векторов частот исходов независимых полиномиальных схем. В сб.: *Тезисы докл. Третьей Всероссийской школа-коллоквиума по стохастическим методам.* ТВП, Москва, 1996, с. 156–157.
15. Шойтов А. М., Повторения значений функции от отрезков последовательности независимых испытаний. *Дискретная математика* (2000) **12**, №3, 49–59.
16. Barbour A. D., Holst L., Janson S., *Poisson approximation.* Oxford Univ. Press, Oxford, 1992.

Статья поступила 21.09.2004.