Труды по дискретной математике
Том 3, с. 147–154
М.: ФИЗМАТЛИТ, 2000

НЕРАВЕНСТВА ДЛЯ СРЕДНЕГО ЧИСЛА ПОВТОРЕНИЙ m-ЦЕПОЧЕК И ДЛЯ СРЕДНЕГО ЧИСЛА НЕПОЯВИВШИХСЯ m-ЦЕПОЧЕК ИЗ ЗАДАННОГО КЛАССА

В. Г. МИХАЙЛОВ

В работе доказываются оценки сверху и снизу для среднего числа r-кратных повторений m-цепочек из заданного класса (т.е. принадлежащих некоторому множеству S цепочек длины m над рассматриваемым алфавитом). С помощью этих оценок получены аналогичные оценки для среднего числа непоявившихся m-цепочек из заданного класса.

Рассмотрим последовательность независимых случайных величин Y_1, $Y_2, \ldots$, принимающих значения из множества $\mathbf{N} = \{1, 2, \ldots\}$ с вероятностями

$$p_{ki} = \mathbf{P}\{Y_i = k\}, \qquad \sum_{k \in \mathbf{N}} p_{ki} = 1 \qquad (i = 1, 2, \ldots). \tag{1}$$

Пусть

$$X_i = (Y_1, \ldots, Y_{i+m-1}), \qquad i = 1, 2, \ldots,$$

— последовательность m-цепочек в последовательности $\{Y_i\}$. Для произвольного подмножества S множества $\mathbf{N}^m$ определим случайные величины

$$\xi_r(S) = \sum_{1 \le i_1 < \ldots < i_r \le n} I\{X_{i_1} = \ldots = X_{i_r} \in S\}, \tag{2}$$

где $I\{B\}$ обозначает индикатор события B. Величина $\xi_r(S)$ называется числом r-кратных повторений m-цепочек из класса S в последовательности (1). В данной работе получены оценки сверху и снизу для величин $\mathbf{E}\,\xi_r(S)$.

Систематическое изучение числа r-кратных повторений m-цепочек $\xi_r = \xi_r(\mathbf{N}^m)$ началось с работ [1] и [2], в которых были получены условия сходимости к сложному пуассоновскому распределению. Скорость такой сходимости была оценена в работе [3]. Условия асимптотической нормальности ξ_r изучались в работе [4]. Суммы случайных величин ξ_r можно использовать при изучении асимптотического поведения разделимых статистик. Данный подход, разработанный в [5], позволил получить простое по форме достаточное условие асимптотической нормальности разделимой статистики, определенной на частотах m-цепочек в неоднородной полиномиальной схеме. Метод работы [5] может быть использован и для получения явных оценок математических ожиданий разделимых статистик. В настоящей работе этот подход применяется для оценивания среднего числа непоявившихся m-цепочек из заданного класса.

Пусть $N(S)$ — множество чисел из $\mathbf{N}$, встречающихся в записи цепочек из множества S. Положим

$$p = \max_i \max_{k \in N(S)} \mathbf{P}\{Y_i = k\}, \qquad P\{S\} = \max_i \mathbf{P}\{X_i \in S\},$$

$$a(\mathbf{k}) = \sum_{i=1}^{n} \mathbf{P}\{X_i = \mathbf{k}\}, \qquad Q_r = \sum_{\mathbf{k} \in S} (a(\mathbf{k}))^r.$$

Теорема 1. *Пусть $p < 1/2$. Тогда при всех r, удовлетворяющих условию $2 \le r \le n$, выполнены неравенства*

$$\mathbf{E}\,\xi_r(S) \le \frac{1}{r!} Q_r(S) + (m-1)\, np^m \,(1 + np^m)^{r-1} M_r P(S), \tag{3}$$

где

$$M_r = \max\left\{ \frac{1}{(r-1)!}, \left(\frac{ep}{1-2p}\right)^{r-1} \right\},$$

и

$$\mathbf{E}\,\xi_r(S) \ge \frac{1}{r!} Q_r(S) - \frac{2m-1}{2((r-2)!)} P(S)\,(np^m)^{r-1}. \tag{4}$$

Замечание 1. Верхняя оценка типа (3) для $\mathbf{E}\,\xi_r(S)$ была получена ранее в работе [5] в случае $S = \mathbf{N}^m$. Сформулированная выше теорема упрощает ее и распространяет на случай произвольных множеств S. Кроме этого, мы получаем нижние оценки для $\mathbf{E}\,\xi_r(S)$.

Оценки теоремы 1 можно использовать при выводе аналогичных оценок для математических ожиданий разделимых статистик от частот m-цепочек в неоднородной схеме. Это можно продемонстрировать на примере статистики $\mu_0(S)$ — числа непоявившихся цепочек из заданного множества S.

Теорема 2. *Пусть $p(2 + e + enp^m) < 1$. Тогда*

$$\begin{aligned} \left| \mathbf{E}\,\mu_0(S) - \sum_{\mathbf{k} \in S} e^{-a(\mathbf{k})} \right| &\le (m-1)\, np^m \left(e^{1+np^m} - 1\right) P(S) + \\ &+ \frac{(m-1)\, enp^m (1 + np^m)}{1 - p(2 + e + enp^m)} P(S) + \frac{1}{2}(2m-1)\, np^m e^{np^m} P(S). \end{aligned} \tag{5}$$

Замечание 2. Оценка (5) является новой даже для наиболее изученной статистики $\mu_0 = \mu_0(\mathbf{N}^m)$ в однородной схеме (когда вероятности p_{ki} не зависят от индекса i). Этот случай изучался в работе [6], где имеется асимптотическая оценка разности

$$\mathbf{E}\,\mu_0\,(N^m) - \sum_{k_1,\dots,k_m \in \mathbf{N}} e^{-np_{k_1}\cdots p_{k_m}}. \tag{6}$$

Позднее (в [7]) данный результат работы [6] был распространен на случай множеств S достаточно общего вида при сохранении предположения об однородности схемы. В отличие от результатов обеих упомянутых работ, теорема 2 дает явную оценку величины погрешности (6), которая пригодна для численных расчетов.

Нетрудно заметить, что при $m = 1$ оценки (3) и (5) значительно упрощаются. Аналогичное упрощение имеет место и при $m \geq 2$, но для специально выбранных множеств S. Рассмотрим один интересный случай.

Пусть $m \geq 2$. Обозначим через $N(m)$ множество всех m-цепочек $\mathbf{k} = (k_1, \dots, k_m)$, удовлетворяющих условиям

$$k_1 \neq k_m, \qquad (k_1, k_2) \neq (k_{m-1}, k_m), \dots, (k_1, \dots, k_{m-1}) \neq (k_2, \dots, k_m).$$

Цепочки из $N(m)$ обладают тем свойством, что начало такой цепочки не совпадает с ее концом.

Теорема 3. *Пусть $m \geq 2$, $p < 1/2$ и $S \subseteq N(m)$. Тогда при всех r, удовлетворяющих условию $2 \leq r \leq n$, выполнено неравенство*

$$\mathbf{E}\,\xi_r(S) \leq \frac{1}{r!}\,Q_r(S). \tag{7}$$

Теорема 4. *Пусть $m \geq 2$, $S \subseteq N(m)$ и $p(2 + e + enp^m) < 1$. Тогда*

$$\left|\mathbf{E}\,\mu_0(S) - \sum_{\mathbf{k} \in S} e^{-a(\mathbf{k})}\right| \leq \frac{1}{2}\,(2m-1)\,np^m e^{np^m} P(S). \tag{8}$$

Замечание 3. Интересно выяснить, сколь содержательны оценки (3)–(5) при $S \not\subseteq N(m)$. Наиболее показательным представляется следующий пример. Рассмотрим равновероятную схему с N исходами ($p_{ki} = 1/N$ при $k \leq N$ и $p_{ki} = 0$ при $k > N$). Пусть

$$S = \bigl\{(k, \dots, k) \,:\, k \in \{1, \dots, N\}\bigr\}.$$

В этом случае неравенство (5) принимает вид

$$\left|\mathbf{E}\,\mu_0(S) - Ne^{-a}\right| \leq (m-1)\,\alpha\left(e^{1+\alpha} - 1\right) N^{1-m} +$$

$$+\,\frac{(m-1)\,e\alpha(1+\alpha)}{1 - N^{-1}(2 + e + e\alpha)}\,N^{1-m} + \frac{1}{2}\,(2m-1)\,\alpha e^{\alpha} N^{1-m},$$

где $\alpha = n/N^m$. Эта оценка содержательна даже при $\alpha \to \infty$. Например, при $Ne^{-\alpha} \to \lambda$, $0 < \lambda < \infty$, величина в правой части имеет порядок

$$O\left(m\alpha e^{\alpha} N^{1-m}\right) = O\left(m \ln N \cdot N^{2-m}\right).$$

З а м е ч а н и е 4. Задача о числе непоявившихся m-цепочек имеет довольно длинную историю, начинающуюся с работы П.Ф. Беляева [8]. Из последующих работ следует отметить статью [9], в которой, в частности, были описаны условия сходимости распределения этой величины в однородной схеме к распределению Пуассона. Вопрос об условиях асимптотической нормальности этой величины изучался в работе [6].

Д о к а з а т е л ь с т в о т е о р е м ы 1. Ясно, что

$$\mathbf{E}\,\xi_r(S) = \sum_J \mathbf{P}\left\{A(i_1,\ldots,i_r)\right\},$$

где

$$A(i_1,\ldots,i_r) = \left\{X_{i_1} = \ldots = X_{i_r} \in S\right\}, \qquad J = \left\{1 \le i_1 < \ldots < i_r \le n\right\}.$$

Разобьем область суммирования J на сумму непересекающихся множеств $J(l)$:

$$J(l) = \Big\{(i_1,\ldots,i_r) \in J\colon \text{множество } V(i_1,\ldots,i_r) = \bigcup_{t=1}^{r}[i_t,\, i_t+m-1] \\ \text{состоит ровно из } l \text{ связных компонент}\Big\}.$$

Тогда

$$\mathbf{E}\,\xi_r(S) = \sum_{l=1}^{r} E_l, \qquad E_l = \sum_{J(l)} \mathbf{P}\left\{A(i_1,\ldots,i_r)\right\}.$$

Следующие утверждения дают оценки для величин E_l.

Лемма 1. *Выполняются неравенства*

$$\frac{1}{r!}\,Q_r(S) - \frac{(2m-1)\,p^m}{2(r-2)!}\,Q_{r-1}(S) \le E_r \le \frac{1}{r!}\,Q_r(S). \tag{9}$$

Следствие 1. *Выполняется неравенство*

$$E_r \ge \frac{1}{r!}\,Q_r(S) - \frac{2m-1}{2((r-2)!)}\,P(S)\,(np^m)^{r-1}. \tag{10}$$

Оценка (10) вытекает из (9) и того, что в силу неравенств

$$a(\mathbf{k}) \le np^m \qquad \text{и} \qquad \sum_{\mathbf{k}\in S} a(\mathbf{k}) \le nP(S)$$

имеем

$$Q_{r-1}(S) \le P(S)\,n^{r-1}p^{m(r-2)}.$$

Заметим, что из (10) следует оценка (4) теоремы 1.

Д о к а з а т е л ь с т в о л е м м ы 1. Пусть

$$J_*(r) = \Big\{(i_1,\ldots,i_r) \in \{1,\ldots,n\}^r\colon \text{все элементы } i_1,\ldots,i_r \text{ различны и при} \\ \text{их упорядочении по возрастанию получается набор из } J(r)\Big\}.$$

Воспользуемся тем, что случайные величины X_i конечно-зависимы. Получаем

$$\begin{aligned}E_r = \sum_{\mathbf{k}\in S}\sum_{J(r)} \mathbf{P}\left\{X_{i_1} = \ldots = X_{i_r} = \mathbf{k}\right\} =\\ = \frac{1}{r!}\sum_{\mathbf{k}\in S}\sum_{J_*(r)} \mathbf{P}\left\{X_{i_1} = \ldots = X_{i_r} = \mathbf{k}\right\} <\\ < \frac{1}{r!}\sum_{\mathbf{k}\in S}\sum_{i_1=1}^{n}\cdots\sum_{i_r=1}^{n}\prod_{s=1}^{r}\mathbf{P}\left\{X_{i_s} = \mathbf{k}\right\} = \frac{1}{r!}Q_r(S).\end{aligned} \tag{11}$$

Теперь заметим, что

$$\{1,\ldots,n\}^r \backslash J_*(r) \subseteq \bigcup_{1\le s<t\le r} J_{st},$$

где $J_{st} = \left\{(i_1,\ldots,i_r)\in\{1,\ldots,n\}^r : |i_s - i_t| < m\right\}$. Кроме этого,

$$\begin{aligned}&\sum_{J_{st}}\sum_{\mathbf{k}\in S}\prod_{u=1}^{r}\mathbf{P}\{X_{i_u} = \mathbf{k}\} \le\\ &\le \sum_{\mathbf{k}\in S}\left(\sum_{i=1}^{n}\mathbf{P}\{X_i = \mathbf{k}\}\right)^{r-2}\sum_{i_s=1}^{n}\mathbf{P}\{X_{i_s} = \mathbf{k}\}\sum_{\max\{1,i_s-m+1\}}^{\min\{n,i_s+m-1\}}\mathbf{P}\{X_{i_t} = \mathbf{k}\} \le\\ &\le (2m-1)\,p^m Q_{r-1}(S).\end{aligned}$$

Следовательно,

$$\begin{aligned}E_r &= \sum_{\mathbf{k}\in S}\sum_{J(r)}\mathbf{P}\left\{X_{i_1} = \ldots = X_{i_r} = \mathbf{k}\right\} =\\ &= \frac{1}{r!}\sum_{\mathbf{k}\in S}\sum_{J_*(r)}\prod_{u=1}^{r}\mathbf{P}\left\{X_{i_u} = \mathbf{k}\right\} \ge \frac{1}{r!}Q_r(S) - \frac{(2m-1)}{2((r-2)!)}Q_{r-1}(S).\end{aligned}$$

Лемма 1 доказана.

Лемма 2. *При* $1 \le l \le r-1$ *выполняется неравенство*

$$E_l \le (m-1)\binom{r-1}{l-1}\binom{n}{l}p^{lm}\left(\frac{p}{1-p}\right)^{r-l-1}P(S).$$

Доказательство этой леммы мы не приводим. Достаточно повторить (с незначительными уточнениями) рассуждения, проведенные на стр. 95–96 работы [5].

Лемма 3. *Пусть* $p < 1/2$. *Тогда*

$$\sum_{l=1}^{r-1}E_l \le (m-1)\,np^m\,(1+np^m)^{r-1}\,M_r P(S).$$

Доказательство леммы 3. Воспользуемся аналогией с соответствующими частями доказательства леммы 3 работы [5].

Используя неравенство леммы 2, можно записать

$$\sum_{l=1}^{r-1} E_l \le (m-1)\,P(S) \sum_{l=1}^{r-1} \binom{r-1}{l-1} \frac{(np^m)^l}{l!} \left(\frac{p}{1-p}\right)^{r-l-1} < \tag{12}$$

$$< (m-1)\,np^m\,(1+np^m)^{r-1} \max\{a_1,\ldots,a_{r-1}\}\,P(S),$$

где

$$a_l = \frac{1}{l!}\left(\frac{p}{1-p}\right)^{r-l-1}.$$

Так как $a_l/a_{l-1} = (1-p)/(lp)$, то при $r-1 \le (1-p)/p$ максимальным значением a_l является $a_{r-1} = 1-(r-1)!$, и имеет место неравенство

$$\sum_{l=1}^{r-1} E_l < (m-1)\,P(S)\,np^m\,\frac{(1+np^m)^{r-1}}{(r-1)!}. \tag{13}$$

Рассмотрим теперь случай $r-1 > (1-p)/p$. Тогда

$$\max\{a_1,\ldots,a_{r-1}\} = \max\{a_u, a_{u+1}\},$$

где $u = [(1-p)/p]$. Используя оценку $n! > (n/e)^n$, получаем

$$a_u < \left(\frac{e}{u}\right)^u \left(\frac{p}{1-p}\right)^{r-u-1} < \left(\frac{e}{(1-p)/p-1}\right)^u \left(\frac{p}{1-p}\right)^{r-u-1} < \left(\frac{ep}{1-2p}\right)^{r-1}.$$

Аналогично получается оценка

$$a_{u+1} < \left(\frac{ep}{1-2p}\right)^{r-1}.$$

Таким образом, в рассматриваемом случае

$$\max\{a_1,\ldots,a_{r-1}\} < \left(\frac{ep}{1-2p}\right)^{r-1}.$$

Подставив эту оценку в (12), получим, что при $r-1 > (1-p)/p$ выполнено неравенство

$$\sum_{l=1}^{r-1} E_l < (m-1)\,P(S)\,np^m\,(1+np^m)^{r-1}\left(\frac{ep}{1-2p}\right)^{r-1}. \tag{14}$$

Объединяя оценки (13) и (14), получаем утверждение леммы. Лемма 3 доказана.

Используя леммы 1 и 3, получаем (3). Теорема 1 доказана.

Доказательство теоремы 2. Из формулы включения-исключения следует, что

$$\mu_0(S) = |S| - \nu(S) + \sum_{r=2}^{n} (-1)^r \xi_r(S), \tag{15}$$

где $\nu(S)$ — число m-цепочек среди $X_1, \ldots, X_n$, принадлежащих множеству S. Поэтому

$$\mathbf{E}\,\mu_0(S) = |S| - \mathbf{E}\,\nu(S) + \sum_{r=2}^{n} (-1)^r \mathbf{E}\,\xi_r(S), \tag{16}$$

Воспользуемся этой формулой для оценивания $\mathbf{E}\,\mu_0(s)$. Как нетрудно заметить,

$$|S| = Q_0(S) = \sum_{\mathbf{k}\in S} (a(\mathbf{k}))^0, \qquad \mathbf{E}\,\nu(S) = Q_1(S) = \sum_{\mathbf{k}\in S} a(\mathbf{k}). \tag{17}$$

Заметим еще, что при $r > n$ верхняя оценка теоремы 1 для $\mathbf{E}\,\xi_r(S)$ положительна, а нижняя оценка — отрицательна. Поскольку $\mathbf{E}\,\xi_r(S) = 0$ при $r > n$, мы можем использовать оценки теоремы 1 при всех целых $r \geq 2$.

Начнем с верхней оценки для $\mathbf{E}\,\mu_0(S)$. Применив расширенную указанным выше образом теорему 1 и равенства (17) к тождеству (16), получим

$$\mathbf{E}\,\mu_0(S) \leq \sum_{\mathbf{k}\in S} e^{-a(\mathbf{k})} + G_1 + G_2 + G_3,$$

где

$$G_1 = \sum_{r=2}^{\infty} (m-1)\, P(S)\, np^m \,\frac{(1+np^m)^{r-1}}{(r-1)!} = (m-1)\, np^m\, (1+np^m)^{r-1}\, P(S),$$

$$G_2 = \sum_{r=2}^{\infty} (m-1)\, np^m\, (1+np^m)^{r-1}\, P(S) \left(\frac{ep}{1-2p}\right)^{r-1} =$$

$$= \frac{(m-1)\, enp^m (1+np^m)^{r-1}}{1 - p(2+e+enp^m)}\, P(S),$$

$$G_3 = \sum_{r=2}^{\infty} \frac{2m-1}{2((r-2)!)}\, (np^m)^{r-1} = \frac{1}{2}\,(2m-1)\, np^m e^{np^m}\, P(S).$$

Аналогичным образом доказывается нижняя оценка

$$\mathbf{E}\,\mu_0(S) \geq \sum_{\mathbf{k}\in S} e^{-a(\mathbf{k})} - G_1 - G_2 - G_3.$$

Теорема 2 доказана.

Доказательство теоремы 3. Применим схему доказательства оценки (3) теоремы 1. Отличие состоит в том, что в данном случае $\mathbf{E}\,\xi_r(S) = E_r$. Поэтому достаточно воспользоваться неравенствами (9) и получить (7). Теорема 3 доказана.

Доказательство теоремы 4 аналогично доказательству теоремы 2, и мы его не приводим.

Автор признателен А. М. Зубкову и В. П. Чистякову за обсуждение затронутых в статье вопросов и полезные замечания.

СПИСОК ЛИТЕРАТУРЫ

1. *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, вып. 1, с. 173–181.
2. *Михайлов В. Г.* Предельные распределения случайных величин, связанных с многократными длинными повторениями в последовательности независимых испытаний. — Теория вероятн. и ее примен., 1974, т. 19, вып. 1, с. 182–187.
3. *Зубков А. М., Михайлов В. Г.* О повторениях s-цепочек в последовательности независимых величин. — Теория вероятн. и ее примен., 1979, т. 24, вып. 1, с. 267–279.
4. *Михайлов В. Г.* Асимптотическая нормальность в схеме конечно-зависимого размещения частиц по ячейкам. — Матем. сб., 1982, т. 119, № 4, с. 509–520.
5. *Михайлов В. Г.* Асимптотическая нормальность разделимых статистик от частот m-цепочек. — Дискретная математика, 1989, т. 1, вып. 4, с. 92–103.
6. *Тихомирова М. И.* Об асимптотической нормальности числа непоявившихся s-цепочек. — Дискретная математика, 1992, т. 4, вып. 2, с. 122–129.
7. *Чистяков В. П., Тихомирова М. И.* Об асимптотике моментов числа непоявившихся s-цепочек. — Дискретная математика, 1997, т. 9, вып. 1, с. 12–29.
8. *Беляев П. Ф.* О вероятности непоявления заданного числа s-цепочек в сложных цепях Маркова. — Теория вероятн. и ее примен., 1965, т. 10, вып. 3, с. 547–551.
9. *Колчин В. Ф., Чистяков В. П.* Предельные распределения числа непоявившихся s-цепочек в полиномиальной схеме. — Теория вероятн. и ее примен., 1974, т. 19, вып. 4, с. 855–864.