Math-Net.Ru

Общероссийский математический портал

В. А. Ватутин, В. Г. Михайлов, О числе чтений случайных неравновероятных файлов при устойчивой сортировке, *Дискрет. матем.*, 1996, том 8, выпуск 2, 14–30

DOI: <https://doi.org/10.4213/dm526>

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением
<http://www.mathnet.ru/rus/agreement>

Параметры загрузки:

IP: 3.93.64.190

26 апреля 2023 г., 12:33:48

О числе чтений случайных неравновероятных файлов при устойчивой сортировке

© 1996 г. В. А. Ватутин, В. Г. Михайлов

Пусть случайный файл $\mathcal{F}$ образован метками $F_1, \dots, F_n$, являющимися независимыми случайными элементами алфавита $A = \{A_1, \dots, A_N\}$, имеющими на нем не обязательно равномерное распределение. Исследуется число чтений ζ случайного файла $\mathcal{F}$ или, что то же самое, число отрезков возрастания в перестановке, осуществляющей устойчивую сортировку файла $\mathcal{F}$. Указаны достаточные условия асимптотической нормальности случайной величины ζ при $n, N \rightarrow \infty$. Случай сортировки файлов, распределенных равномерно на множестве всех слов длины n над алфавитом A , был рассмотрен ранее.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 93-011-1443.

Файл $\mathcal{F}$ представляет собой некоторый набор записей данных, причем каждой записи файла отвечает некоторое кодовое обозначение (метка или ключ). В качестве ключей обычно используются элементы некоторого линейно упорядоченного множества, скажем,

$$A = \{A_1, A_2, \dots, A_N, A_1 < A_2 < \dots < A_N\}.$$

Под сортировкой файла $\mathcal{F}$ подразумевается такая перестановка записей $\sigma[\mathcal{F}]$, в результате применения которой записи располагаются в порядке неубывания ключевых обозначений. Конкретное содержание записей при такой сортировке никакого значения не имеет. Поэтому мы будем считать далее, что файл состоит непосредственно из ключевых обозначений записей, т. е.

$$\mathcal{F} = (F_1, \dots, F_n), \quad F_k \in A, \quad k = 1, \dots, n.$$

Перестановка $\sigma[\mathcal{F}] = \sigma(1)\sigma(2) \dots \sigma(n)$ определяется условием

$$F_{\sigma(1)} \leq F_{\sigma(2)} \leq \dots \leq F_{\sigma(n)}.$$

Сортировка называется устойчивой, если дополнительно при любом $i = 1, \dots, n-1$ выполнено условие

$$F_{\sigma(i)} = F_{\sigma(i+1)} \Rightarrow \sigma(i) < \sigma(i+1).$$

Ясно, что устойчивая сортировка сохраняет порядок записей с одинаковыми ключами. Устойчивая сортировка является важнейшим инструментом построения эффективных алгоритмов (см. [2]).

Легко понять, что перестановка $\sigma[\mathcal{F}]$, реализующая устойчивую сортировку, определяется однозначно. Однако одна и та же перестановка может устойчиво сортировать, вообще говоря, несколько файлов.

Говорят, что файл $\mathcal{F} = (F_1, \dots, F_n)$ требует k чтений, если его необходимо просмотреть k раз слева направо, чтобы прочитать все его элементы в неубывающем порядке (см. [2, с. 64]).

Так, файл (39551282) требует три чтения: при первом чтении получается последовательность 122, при втором чтении 3558, а при третьем чтении 9.

Нетрудно видеть, что если файл $\mathcal{F}$ требует k чтений, то перестановка $\sigma[\mathcal{F}]$, реализующая устойчивую сортировку этого файла, имеет k отрезков возрастания. Так, в нашем примере

$$\sigma[\mathcal{F}] = (568 \mid 1374 \mid 2),$$

т. е. имеется как раз три отрезка возрастания.

В настоящей работе мы рассмотрим случай, когда ключи $F_1, \dots, F_n$ файла $\mathcal{F}$ выбираются независимо и случайно из алфавита A в соответствии с полиномиальной схемой, в которой

$$P\{F_i = A_j\} = p_j, \quad j = 1, \dots, N, \quad i = 1, \dots, n. \quad (1)$$

Надо признать, что ситуация с неодинаковыми вероятностями ключевых обозначений достаточно типична. Она, например, возникает при кодировании сообщений в узле связи именами корреспондентов.

В силу (1) на множестве A^n слов длины n над алфавитом A индуцируется распределение вероятностей

$$P(\mathcal{F} = (A_{j(1)}, \dots, A_{j(n)})) = p_{j(1)} \dots p_{j(n)}.$$

Последнее определяет распределение случайной перестановки на множестве S_n всех перестановок порядка n по правилу

$$P\{\sigma[\mathcal{F}] = \sigma\} = \sum p_{j(1)} \dots p_{j(n)}, \quad (2)$$

где суммирование распространяется на все файлы $(A_{j(1)}, \dots, A_{j(n)})$, устойчиво сортирующиеся перестановкой σ .

Главным объектом исследования данной заметки является величина $\zeta = \zeta_{n,N} = \zeta_{n,N}(\sigma[\mathcal{F}])$ — число чтений случайного файла $\mathcal{F}$. Заметим, что количество отрезков возрастания в различного рода случайных перестановках используется, например, при построении статистических тестов, проверяющих качество псевдослучайных последовательностей и приводимая ниже теорема 1 может быть применена при построении соответствующих критериев. Читатель может найти обширную библиографию, относящуюся к данной тематике в обзоре [3].

Воспользуемся обозначениями

$$p = \max_i p_i, \quad P_r = \sum_{j=1}^N p_j^r, \quad r = 1, 2, 3, \dots,$$

и предположим, что

$$N, n \rightarrow \infty, \quad nP_2 \rightarrow 0, \quad np \leq a, \quad (3)$$

где a — некоторая постоянная.

Основной результат работы формулируется следующим образом.

Теорема 1. Если при $n, N \rightarrow \infty$ параметры полиномиальной схемы (1) меняются так, что выполнены условия (3), то для любого фиксированного x

$$P\{((n+1)/12)^{-1/2}(\zeta - E\zeta) \leq x\} \rightarrow \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2} du.$$

Замечание 1. Можно показать, что в условиях теоремы 1

$$E\zeta = \frac{n+1}{n} - \frac{n^2 P_2}{12}(1 + O(np)) + O(1).$$

Замечание 2. Равновероятный случай $p_1 = \dots = p_N = 1/N$ был детально изучен в работе [1]. Там были доказаны локальные и интегральные предельные теоремы для ζ , а также выписаны асимптотические формулы для $E\zeta$ и $D\zeta$. Наша теорема 1 распространяет на неравновероятный случай интегральную нормальную предельную теорему работы [1]. Это достигается лишь ценой введения условия (3). Аналогичное условие для равновероятного случая (теорема 2 работы [1]) существенно проще: $D\zeta \rightarrow \infty$.

1. Комбинаторные свойства числа чтений

В этом разделе мы интерпретируем случайную величину ζ именно как число чтений. Зафиксируем число ключей каждого вида в файле $\mathcal{F}$. А именно, для

$$\nu_s = |\{i \in \{1, \dots, n\} : F_i = A_s\}|, \quad s = 1, \dots, N,$$

введем событие

$$\{\nu_1 = m_1, \dots, \nu_N = m_N\}, \quad m_1 + \dots + m_N = n. \quad (4)$$

При этом появившиеся ключи имеет смысл исключать из рассмотрения. Тогда алфавит A редуцируется до алфавита

$$A' = \{A_j : m_j \geq 1\} = \{A'_1, \dots, A'_k\},$$

где k — число букв нового алфавита. Обозначим через $(n_1, \dots, n_k)$ набор, полученный из $(m_1, \dots, m_N)$ вычеркиванием нулевых элементов. При условии (4) распределение файла $\mathcal{F}$ превращается в равномерное на множестве всех слов длины n , содержащих ровно по $n_j \geq 1$ букв A'_j , $j = 1, \dots, k$, алфавита A' . Пусть случайная величина I_s есть индикатор события, состоящего в том, что буквы

A'_s и A'_{s+1} алфавита A' прочитываются в файле $\mathcal{F}$ за одно чтение. Тогда, как нетрудно убедиться,

$$\zeta = 1 + \sum_{s=1}^{k-1} (1 - I_s) = k - \sum_{s=1}^{k-1} I_s. \quad (5)$$

Лемма 1. Пусть файл $\mathcal{F}$ распределен равномерно на множестве всех слов длины n , содержащих ровно по $n_j \geq 1$ букв A'_j , $j = 1, \dots, k$. Тогда случайные индикаторы $I_1, \dots, I_{k-1}$ образуют систему 1-зависимых случайных величин с параметрами

$$\mathbf{E}I_s = \frac{n_s! n_{s+1}!}{(n_s + n_{s+1})!}, \quad s = 1, \dots, k-1, \quad (6)$$

$$\mathbf{E}I_s I_{s+1} = \frac{n_s! n_{s+1}! n_{s+2}!}{(n_s + n_{s+1} + n_{s+2})!}, \quad s = 1, \dots, k-2. \quad (7)$$

Доказательство. Лемма доказывается достаточно просто. Зафиксируем некоторое натуральное число l , $1 \leq l \leq k-2$, и те места $t_1, \dots, t_L$, $L = n_1 + \dots + n_l$, которые соответствуют меткам $A'_1, \dots, A'_l$ в файле $\mathcal{F}$. Легко понять, что при такой фиксации системы индикаторов $\{I_1, \dots, I_l\}$ и $\{I_{l+1}, \dots, I_{k-1}\}$ не зависят друг от друга, а их распределения не зависят от конкретного вида множества $\{t_1, \dots, t_L\}$. Этим доказывается 1-зависимость. Аналогично, с помощью фиксации множеств мест для A'_s и A'_{s+1} или A'_s, A'_{s+1} и A'_{s+2} доказываются формулы (6) и (7). Лемма 1 доказана.

Теорема 2. Пусть ζ — число чтений случайного файла $\mathcal{F}$, распределенного равномерно на множестве всех слов длины n над алфавитом A' , содержащих ровно по $n_j \geq 1$ букв A'_j , $j = 1, \dots, k$. Тогда

$$\mathbf{E}\zeta = k - \sum_{s=1}^{k-1} \frac{n_s! n_{s+1}!}{(n_s + n_{s+1})!}, \quad (8)$$

$$\begin{aligned} \mathbf{D}\zeta = & \sum_{s=1}^{k-1} \frac{n_s! n_{s+1}!}{(n_s + n_{s+1})!} \left(1 - \frac{n_s! n_{s+1}!}{(n_s + n_{s+1})!} \right) \\ & + 2 \sum_{s=1}^{k-2} \left(\frac{n_s! n_{s+1}! n_{s+2}!}{(n_s + n_{s+1} + n_{s+2})!} - \frac{n_s! (n_{s+1}!)^2 n_{s+2}!}{(n_s + n_{s+1})! (n_{s+1} + n_{s+2})!} \right). \end{aligned} \quad (9)$$

Кроме того,

$$\sup_x |\mathbf{P}\{(\zeta - \mathbf{E}\zeta)/\sqrt{\mathbf{D}\zeta} \leq x\} - \Phi(x)| \leq \frac{cH}{(\mathbf{D}\zeta)^{3/2}}, \quad (10)$$

где c — некоторая абсолютная константа и

$$H = \sum_{s=1}^{k-1} \frac{n_s! n_{s+1}!}{(n_s + n_{s+1})!} \left(1 - \frac{n_s! n_{s+1}!}{(n_s + n_{s+1})!} \right) \left(1 - 2 \frac{n_s! n_{s+1}!}{(n_s + n_{s+1})!} + 2 \left(\frac{n_s! n_{s+1}!}{(n_s + n_{s+1})!} \right)^2 \right).$$

Замечание 3. Свойство 1-зависимости, отмеченное в лемме 1, и формулы (8) и (9), по-видимому являются достаточно известными. Существенно новым результатом в теореме 2 является неравенство (10).

Доказательство теоремы 2. Формулы (8) и (9) следуют непосредственно из леммы 1. Докажем неравенство (10). По формуле (5) и лемме 1 величина $k - \zeta$ является суммой 1-зависимых случайных индикаторов $I_1, \dots, I_{k-1}$. Следовательно, мы можем воспользоваться теоремой 1 работы [4], согласно которой

$$\begin{aligned} \sup_x |\mathbf{P}\{(\zeta - \mathbf{E}\zeta)/\sqrt{\mathbf{D}\zeta} \leq x\} - \Phi(x)| \\ = \sup_x |\mathbf{P}\{(k - \zeta - \mathbf{E}(k - \zeta))/\sqrt{\mathbf{D}(k - \zeta)} \leq x\} - \Phi(x)| \\ \leq \frac{c}{(\mathbf{D}(k - \zeta))^{3/2}} \sum_{s=1}^{k-1} \mathbf{E}|I_s - \mathbf{E}I_s|^3, \end{aligned}$$

где c — некоторая постоянная. Учитывая снова, что $\mathbf{D}(k - \zeta) = \mathbf{D}\zeta$, а также, что

$$\sum_{s=1}^{k-1} \mathbf{E}|I_s - \mathbf{E}I_s|^3 = H,$$

получаем (10). Теорема 2 доказана.

В дальнейшем случайную величину $\zeta = \zeta_{n,N}(\sigma[\mathcal{F}])$, где $\mathcal{F}$ — случайный файл, распределенный равномерно на множестве всех последовательностей длины n , содержащих n_j знаков j_s , $s = 1, \dots, k$, будем обозначать $\zeta(n_1, \dots, n_k)$.

2. Доказательство теоремы 1

В этом параграфе $\mathcal{F} = (F_1, \dots, F_n)$ — последовательность независимых исходов полиномиальной схемы с параметрами $n, N, p_1, \dots, p_N$, удовлетворяющими условию (3), $\zeta = \zeta_{n,N}(\sigma[\mathcal{F}])$.

Пусть, как и ранее, $\nu_1, \dots, \nu_N$ — частоты исходов в полиномиальной схеме.

Лемма 2. Пусть $m_1, \dots, m_N$ — неотрицательные целые числа, $m_1 + \dots + m_N = n$ и набор $(n_1, \dots, n_k)$ получен из набора $(m_1, \dots, m_N)$ отбрасыванием нулевых членов. Тогда при любом x

$$\mathbf{P}\{\zeta \leq x \mid \nu_i = m_i, i = 1, \dots, N\} = \mathbf{P}\{\zeta(n_1, \dots, n_k) \leq x\}. \quad (11)$$

Утверждение леммы следует из определений случайных величин $\zeta(n_1, \dots, n_k)$, ζ и равенства

$$\begin{aligned} \mathbf{P}\{(F_1, \dots, F_n) = (A_{r(1)}, \dots, A_{r(n)}) \mid \nu_i = m_i, i = 1, \dots, N\} \\ = \frac{m_1! \dots m_N!}{n!} = \frac{n_1! \dots n_k!}{n!}, \end{aligned}$$

справедливого для всех наборов $A_{r(1)}, \dots, A_{r(n)}$, в которых каждое число A_r встречается m_r раз, $r = 1, \dots, N$.

Из леммы 2, в частности, следует, что

$$\mathbf{E}(\zeta \mid \nu_i = m_i, i = 1, \dots, N) = \mathbf{E}\zeta(n_1, \dots, n_k). \quad (12)$$

Сопоставим каждому набору $(m_1, \dots, m_N)$ величины

$\mu_0(m_1, \dots, m_N)$ — число нулей среди $m_1, \dots, m_N$;

$\mu_{rm}(m_1, \dots, m_N)$ — число таких пар (i, j) , $1 \leq i < j \leq N$, что $m_i = r$, $m_j = m$, $m_s = 0$, $s = i + 1, \dots, j - 1$;

$\mu_{rml}(m_1, \dots, m_N)$ — число таких троек (i, j, k) , $1 \leq i < j < k \leq N$, что $m_i = r$, $m_j = m$, $m_k = l$, $m_s = 0$, $s \in \{i + 1, \dots, k - 1\} \setminus \{j\}$.

Пусть

$$a_{rm} = \frac{r!m!}{(r+m)!},$$

$$a_{rml} = \frac{r!m!}{(r+m)!} \left(1 - \frac{r!m!}{(r+m)!} - 2 \frac{m!l!}{(m+l)!} + 2 \frac{l!(r+m)!}{(r+m+l)!} \right).$$

Нетрудно проверить, что

$$0 < a_{rm} \leq 1, \quad 0 < a_{rml} \leq 1. \quad (13)$$

Пусть

$$a(m_1, \dots, m_N) = N - \mu_0(m_1, \dots, m_N) - \sum_{r,m \geq 1} a_{rm} \mu_{rm}(m_1, \dots, m_N),$$

$$\sigma^2(m_1, \dots, m_N) = a_{n_{k-1}n_k}(1 - a_{n_{k-1}n_k}) + \sum_{r,m,l \geq 1} a_{rml} \mu_{rml}(m_1, \dots, m_N).$$

Из формул (8) и (9) следует, что

$$a(m_1, \dots, m_N) = \mathbf{E}\zeta(n_1, \dots, n_k),$$

$$\sigma^2(m_1, \dots, m_N) = \mathbf{D}\zeta(n_1, \dots, n_k), \quad (14)$$

где набор $(n_1, \dots, n_k)$ получен вычеркиванием нулевых элементов из набора $(m_1, \dots, m_N)$.

Лемма 3. Пусть набор неотрицательных чисел $m_1, \dots, m_N$, $m_1 + \dots + m_N = n$, удовлетворяет условию

$$\sigma^2(m_1, \dots, m_N) > bn, \quad b > 0. \quad (15)$$

Тогда при любом x

$$|\mathbf{P}\{(\zeta - a(m_1, \dots, m_N))/\sigma(m_1, \dots, m_N) \leq x \mid \nu_i = m_i, i = 1, \dots, N\} - \Phi(x)| \leq \frac{c}{b^{3/2}n^{1/2}}. \quad (16)$$

Лемма 3 вытекает из леммы 2, (14), (10) и неравенства $H \leq k \leq n$.

Лемма 4. Пусть на некотором вероятностном пространстве $(\Omega, \mathfrak{A}, \mathbf{P})$ задана случайная величина ξ и выбраны такие $A_1, \dots, A_M, C \in \mathfrak{A}$, что $A_i \cap A_j = \emptyset$, $i \neq j$, $A_1 \cup \dots \cup A_M \cup C = \Omega$. Пусть заданы числа $a, a_1, \dots, a_M$ и положительные числа $\sigma, \sigma_1, \dots, \sigma_M$ и

$$\begin{aligned}\delta &= \max_{1 \leq k \leq M} \sup_x |\mathbf{P}\{(\xi - a_k)/\sigma_k \leq x \mid A_k\} - \Phi(x)|, \\ h_1 &= \max_k |\sigma/\sigma_k - 1| < 1/2, \\ h_2 &= \max_k |(a_k - a)/\sigma_k|.\end{aligned}$$

Тогда

$$\Delta = |\mathbf{P}\{(\xi - a)/\sigma \leq x\} - \Phi(x)| \leq \delta + ch_1 + h_2 + \mathbf{P}(C), \quad (17)$$

где c — абсолютная константа.

Доказательство. Согласно формуле полной вероятности

$$\Delta \leq \sum_{k=1}^M \Delta_k \mathbf{P}\{A_k\} + \mathbf{P}(C), \quad (18)$$

где

$$\begin{aligned}\Delta_k &= \sup_x |\mathbf{P}\{(\xi - a)/\sigma \leq x \mid A_k\} - \Phi(x)| \\ &\leq \delta + \sup_x \left| \Phi\left(\frac{\sigma}{\sigma_k} \left(x - \frac{a_k - a}{\sigma}\right)\right) - \Phi(x) \right|.\end{aligned} \quad (19)$$

Используя неравенства $h_1 < 1/2$, $|\Phi'(x)| < 1$, получаем, что

$$\begin{aligned}\left| \Phi\left(\frac{\sigma}{\sigma_k} \left(x - \frac{a_k - a}{\sigma}\right)\right) - \Phi(x) \right| \\ \leq \left| \Phi\left(\frac{\sigma}{\sigma_k} \left(x - \frac{a_k - a}{\sigma}\right)\right) - \Phi\left(\frac{\sigma}{\sigma_k} x\right) \right| + \left| \Phi\left(\frac{\sigma}{\sigma_k} x\right) - \Phi(x) \right| \\ \leq \frac{|a_k - a|}{\sigma_k} + x \left| \frac{\sigma}{\sigma_k} - 1 \right|,\end{aligned}$$

где

$$c = \sup_x |x| \Phi'(|x|/2) < \infty,$$

что в свою очередь приводит к неравенству

$$\Delta_k \leq \delta + \frac{|a_k - a|}{\sigma_k} + c \left| \frac{\sigma}{\sigma_k} - 1 \right|.$$

Осталось подставить эту оценку в (18). Лемма доказана.

Чтобы воспользоваться леммой 4, выберем некоторое δ , $0 < \delta < 1/2$, и разобьем множество всех исходов полиномиальной схемы на непересекающиеся множества (события)

$$A_{m_1, \dots, m_N} = \{\nu_1 = m_1, \dots, \nu_N = m_N\}.$$

Часть этих множеств объединим в множество

$$C_\delta = C_\delta^{(1)} \cup C_\delta^{(2)},$$

где

$$\begin{aligned} C_\delta^{(1)} &= \{|a(\nu_1, \dots, \nu_N) - \mathbf{E}\zeta| > \delta\sqrt{n}\}, \\ C_\delta^{(2)} &= \{|((n+1)/12)^{1/2} - \sigma(\nu_1, \dots, \nu_N)| > \delta\sigma(\nu_1, \dots, \nu_N)\}. \end{aligned} \quad (20)$$

Ясно, что множество C_δ и те множества $A_{m_1, \dots, m_N}$, которые не вошли в C_δ , образуют указанное в лемме 4 разбиение.

В качестве a, a_k, σ, σ_k возьмем числа $\mathbf{E}\zeta$, $a(m_1, \dots, m_N)$, $((n+1)/12)^{1/2}$, $\sigma(m_1, \dots, m_N)$ соответственно.

В силу лемм 3 и 4

$$|\mathbf{P}\{((n+1)/12)^{-1/2}(\zeta - \mathbf{E}\zeta) \leq x\} - \Phi(x)| \leq c(1/\sqrt{n} + \delta) + P(C_\delta), \quad (21)$$

где c — некоторая константа.

Лемма 5. Пусть выполнены условия (3). Тогда при любом δ , $0 < \delta < 1/2$,

$$P(C_\delta) \rightarrow 0.$$

Эта лемма будет доказана в следующем параграфе.

Используя (3), лемму 5 и (21), получаем утверждение теоремы. Теорема 1 доказана.

3. Доказательство леммы 5

Рассмотрим снова полиномиальную схему, в которой n частиц независимо размещаются по N ячейкам, причем вероятность попадания отдельной частицы в i -ю ячейку равна p_i , $i = 1, \dots, N$. Пусть ν_i — число частиц, попавших в i -ю ячейку, $i = 1, \dots, N$. Введем случайные величины

$$\begin{aligned} \gamma_i^{(r)} &= \mathbf{I}\{\nu_i = r\}, & i &= 1, \dots, N, \\ \gamma_{i,j}^{(r,m)} &= \mathbf{I}\{\nu_i = r, \nu_j = m, \nu_k = 0, i < k < j\}, & i < j, \end{aligned} \quad (22)$$

(в частности, $\gamma_{i,i+1}^{(r,m)} = \mathbf{I}\{\nu_i = r, \nu_{i+1} = m\}$), и случайные величины

$$\mu_r = \sum_{i=1}^N \gamma_i^{(r)}, \quad \mu_{rm} = \sum_{1 \leq i < j \leq N} \gamma_{i,j}^{(r,m)}, \quad \mu_{rml} = \sum_{1 \leq i < j < k \leq N} \gamma_{i,j}^{(r,m)} \gamma_{j,k}^{(m,l)}. \quad (23)$$

Величина μ_r — это число ячеек, содержащих ровно по r частиц каждая.

Отметим сразу, что в обозначениях предыдущего параграфа

$$\mu_0 = \mu_0(\nu_1, \dots, \nu_N), \quad \mu_{rm} = \mu_{rm}(\nu_1, \dots, \nu_N), \quad \mu_{rml} = \mu_{rml}(\nu_1, \dots, \nu_N).$$

Поэтому

$$\begin{aligned} a(\nu_1, \dots, \nu_N) &= N - \mu_0 - \sum_{r, m \geq 1} a_{rm} \mu_{rm}, \\ \sigma^2(\nu_1, \dots, \nu_N) &= \sum_{r, m, l \geq 1} a_{rml} \mu_{rml} + \omega, \quad 0 < \omega < 1. \end{aligned} \quad (24)$$

Перейдем к изучению свойств случайных величин μ_r, μ_{rm} и μ_{rml} . Нетрудно заметить, что

$$N - n \leq \mathbf{E}\mu_0 \leq N, \quad \mathbf{E}\mu_1 \leq n, \quad (25)$$

причем эти неравенства справедливы и для самих случайных величин.

Из определений (22) и (23) следует, что

$$\mathbf{E}\mu_r = \binom{n}{r} \sum_{i=1}^N p_i^r (1 - p_i)^{n-r}, \quad (26)$$

(см. также [5], с. 111),

$$\mathbf{E}\mu_{rm} = \frac{(n)_{r+m}}{r! m!} \sum_{1 \leq i < j \leq N} p_i^r p_j^m (1 - p_i - p_{i+1} - \dots - p_j)^{n-r-m}, \quad (27)$$

$$\mathbf{E}\mu_{rml} = \frac{(n)_{r+m+l}}{r! m! l!} \sum_{1 \leq i < j < k \leq N} p_i^r p_j^m p_k^l (1 - p_i - p_{i+1} - \dots - p_k)^{n-r-m-l}. \quad (28)$$

Здесь $(n)_r = n(n-1)\dots(n-r+1)$.

Используя (26), нетрудно показать, что при $r \geq 2$

$$\mathbf{E}\mu_r \leq \binom{n}{r} P_r \leq \frac{1}{r!} n^2 P_2 (np)^{r-2}, \quad (29)$$

(напомним, что $p = \max_i p_i$, $P_r = p_1^r + \dots + p_N^r$).

Нам понадобятся аналогичные оценки для $\mathbf{E}\mu_{rm}$ и $\mathbf{E}\mu_{rml}$. Из формул (27) и (28) их вывести сложно, однако величины μ_{rm} и μ_{rml} тесно связаны с μ_r , и это упрощает нашу задачу.

Пусть

$$\begin{aligned} \delta_{*r} &= \mathbf{I}\{\exists i \in \{1, \dots, N\}: \nu_k = 0 \forall k \in \{1, \dots, i-1\}, \nu_i = r\}, \\ \delta_{r*} &= \mathbf{I}\{\exists i \in \{1, \dots, N\}: \nu_i = r, \nu_k = 0 \forall k \in \{i+1, \dots, N\}\}. \end{aligned}$$

Из определений вытекает, что

$$\mu_r = \mu_{r1} + \dots + \mu_{r, n-r} + \delta_{r*} = \mu_{1r} + \dots + \mu_{n-r, r} + \delta_{*r}, \quad (30)$$

а также, что

$$\begin{aligned} 0 &\leq \mu_{rm} - \mu_{rm1} - \dots - \mu_{r, m, n-r-m} \leq \delta_{m*}, \\ 0 &\leq \mu_{rm} - \mu_{1rm} - \dots - \mu_{n-r-m, r, m} \leq \delta_{*r}. \end{aligned} \quad (31)$$

Отсюда следует цепочка неравенств

$$E\mu_{rml} \leq E\mu_{r'm'} \leq E\mu_{r''}, \quad (32)$$

где в качестве r', m' можно брать любые два числа из r, m, l , а в качестве r'' — любое из чисел r' и m' . В частности,

$$\begin{aligned} E\mu_{rml} &\leq E\mu_{\max\{r, m, l\}}, \\ E\mu_{rm} &\leq E\mu_{\max\{r, m\}}. \end{aligned} \quad (33)$$

Так как $E\mu_1 \leq n$, то из (32) следует, что

$$E\mu_{111} \leq E\mu_{11} \leq E\mu_1 \leq n. \quad (34)$$

Оценим теперь $E\mu_{111}$ снизу.

Лемма 6. *Справедливо неравенство*

$$E\mu_{111} > n - 3n^2 P_2 - 2. \quad (35)$$

Доказательство. Ввиду (30) и (31)

$$\begin{aligned} \mu_{111} &\geq \mu_1 - \sum_{r=2}^{n-1} \mu_{1r} - \sum_{r=2}^{n-2} \mu_{11r} - 2 \geq \mu_1 - 2 \sum_{r=2}^n \mu_r - 2 \\ &\geq \mu_1 - 2(n - \mu_1) - 2 = 3\mu_1 - 2n - 2, \end{aligned} \quad (36)$$

(здесь мы воспользовались также формулой $n \geq \mu_1 + \mu_2 + \dots + \mu_n$). Нетрудно показать, используя неравенство

$$(1 - x)^n \geq 1 - nx, \quad 0 \leq x \leq 2,$$

что

$$E\mu_1 = n \sum_{i=1}^N p_i (1 - p_i)^{n-1} \geq n - n^2 P_2.$$

Подставляя эту оценку в (36), получаем (35). Лемма доказана.

Перейдем к оценкам дисперсий случайных величин $\mu_r, \mu_{rm}, \mu_{rml}$.

Лемма 7. *Пусть $np \leq a$, где a — постоянная. Тогда*

$$D\mu_i \leq c_i(a) n^2 P_2, \quad i = 0, 1, \quad (37)$$

где $c_0(a) = 3$, $c_1(a) = 2 + 3a$.

Доказательство. Из определений следует (см. также [5], стр. 112), что

$$D\mu_0 = \sum_{i=1}^N ((1 - p_i)^n - (1 - 2p_i)^n) + \sum_{i=1}^N \sum_{j=1}^N ((1 - p_i - p_j)^n - (1 - p_i - p_j + p_i p_j)^n).$$

Теперь неравенства

$$1 - nx \leq (1 - x)^n \leq 1 - nx + \binom{n}{2} x^2, \quad 0 \leq x \leq 2,$$

$$(b + x)^n - b^n \geq nxb^{n-1}, \quad x, b \geq 0,$$

дают оценки

$$\begin{aligned} D\mu_0 &\leq n + \binom{n}{2} P_2 - n \sum_{i,j=1}^N p_i p_j (1 - p_i - p_j)^{n-1} \\ &\leq n + \binom{n}{2} P_2 - n \sum_{i,j=1}^N p_i p_j (1 - (n-1)(p_i + p_j)) \\ &< 3n^2 P_2. \end{aligned}$$

Аналогично получается оценка для $D\mu_1$, если заметить, что

$$\begin{aligned} D\mu_1 &= n \sum_{i=1}^N p_i (1 - p_i)^{n-1} - n(n-1) \sum_{i=1}^N p_i^2 (1 - 2p_i)^{n-2} \\ &\quad + n(n-1) \sum_{i,j=1}^N p_i p_j ((1 - p_i - p_j)^{n-2} - (1 - p_i - p_j + p_i p_j)^{n-2}) \\ &\quad - n \sum_{i,j=1}^N p_i p_j (1 - p_i)^{n-1} (1 - p_j)^{n-1}. \end{aligned}$$

Лемма доказана.

Лемма 8. Пусть $q = r + m \leq n$, $r, m \geq 1$, $np \leq a$, где a — постоянная. Найдется такая величина $c_2(a)$, зависящая только от a , что

$$D\mu_{rm} \leq c_2(a) E\mu_{rm} + \frac{e^2 n^4 a^{2q-4}}{(r!)^2 (m!)^2} e^{-n/q}. \quad (38)$$

Лемма 9. Пусть $q = r + m + l \leq n$, $r, m, l \geq 1$, $np \leq a$, где a — постоянная. Найдется такая величина $c_3(a)$, зависящая только от a , что

$$D\mu_{rml} \leq c_3(a) E\mu_{rml} + \frac{e^2 n^6 a^{2q-6}}{(r!)^2 (m!)^2 (l!)^2} e^{-n/q}. \quad (39)$$

Поскольку доказательства этих утверждений проводятся по одной и той же схеме, мы приводим лишь доказательство леммы 9.

Доказательство леммы 9. Пусть B — множество всех векторов $\beta = (i_1, j_1, k_1, i_2, j_2, k_2)$ с целочисленными координатами, удовлетворяющими условиям $1 \leq i_t < k_t \leq N$, $t = 1, 2$. Положим

$$\begin{aligned} P_1(\beta) &= E\gamma_{i_1 j_1}^{(r,m)} \gamma_{j_1 k_1}^{(m,l)} \gamma_{i_2 j_2}^{(r,m)} \gamma_{j_2 k_2}^{(m,l)}, \\ P_2(\beta) &= P_1(\beta) - E\gamma_{i_1 j_1}^{(r,m)} \gamma_{j_1 k_1}^{(m,l)} E\gamma_{i_2 j_2}^{(r,m)} \gamma_{j_2 k_2}^{(m,l)}. \end{aligned} \quad (40)$$

В этих обозначениях

$$D_{rml} = \sum_{\beta \in B} P_2(\beta). \quad (41)$$

Пусть

$$\begin{aligned} B_1 &= \{\beta \in B: i_1 = i_2, j_1 = j_2, k_1 = k_2\}, \\ B_2^{(1)} &= \{\beta \in B: i_1 = j_2, j_1 = k_2\}, \\ B_2^{(2)} &= \{\beta \in B: i_2 = j_1, j_2 = k_1\}, \\ B_3^{(1)} &= \{\beta \in B: i_1 = k_2\}, \\ B_3^{(2)} &= \{\beta \in B: i_2 = k_1\}, \\ B_4 &= \{\beta \in B: \{i_1, j_1, k_1\} \cap \{i_2, j_2, k_2\} = \emptyset\}, \\ B_5 &= B \setminus (B_1 \cup B_2^{(1)} \cap B_2^{(2)} \cap B_3^{(1)} \cap B_3^{(2)} \cap B_4). \end{aligned}$$

Заметим, что перечисленные множества попарно не пересекаются и образуют разбиение множества B . Отметим также, что $P_1(\beta) = 0$, $\beta \in B_5$, как вероятность произведения несовместных событий. Поэтому из (40) и (41) следует, что

$$D\mu_{rml} \leq D_1 + D_2^{(1)} + D_2^{(2)} + D_3^{(1)} + D_3^{(2)} + D_4, \quad (42)$$

где первые пять слагаемых правой части образованы суммированием величин $P_1(\beta)$ по областям B с соответствующими индексами, а

$$D_4 = \sum_{\beta \in B_4} P_2(\beta). \quad (43)$$

Из (22) и (23) следует, что $D_1 = \mathbf{E}\mu_{rml}$, $D_j^{(i)} \leq \mu_{rml}$, $i = 1, 2$, $j = 2, 3$, т.е. (см. (42))

$$D\mu_{rml} \leq 5\mathbf{E}\mu_{rml} + D_4. \quad (44)$$

Оценим величину D_4 . Очевидно, что при $\beta \in B_4$

$$P_2(\beta) \leq \frac{(n)_q^2 R(\beta)}{(r!)^2 (m!)^2 (l!)^2} p_{i_1}^r p_{j_1}^m p_{k_1}^l p_{i_2}^r p_{j_2}^m p_{k_2}^l, \quad (45)$$

где

$$R(\beta) = (1 - \Sigma_1 - \Sigma_2)^{n-2q} - (1 - \Sigma_1 - \Sigma_2 + \Sigma_1 \Sigma_2)^{n-q},$$

а

$$\Sigma_t = p_{i_t} + p_{i_t+1} + \dots + p_{k_t}, \quad t = 1, 2.$$

Если $\Sigma_1 + \Sigma_2 \geq q^{-1}$, то $R(\beta) \leq e^2 e^{-n/q}$.

Пусть теперь $\Sigma_1 + \Sigma_2 < q^{-1}$. Тогда

$$(1 - \Sigma_1 - \Sigma_2)^{-q} \leq 1 - eq(\Sigma_1 + \Sigma_2),$$

поскольку

$$(1 - x)^{-q} - 1 \leq xq(1 - x)^{-q} \leq eqx, \quad 0 \leq x \leq q^{-1}.$$

Следовательно, в обоих случаях

$$R(\beta) \leq e^2 e^{-n/q} + eq(1 - \Sigma_1)^{n-q}(1 - \Sigma_2)^{n-q}(\Sigma_1 + \Sigma_2). \quad (46)$$

Объединяя (45) и (46) с (43), получаем, что

$$D_4 \leq D_4^{(1)} + D_4^{(2)},$$

где

$$D_4^{(1)} \leq e^2 \frac{(n)_q^2 P_r^2 P_m^2 P_l^2}{(r!)^2 (m!)^2 (l!)^2} e^{-n/q}, \quad (47)$$

(здесь мы расширили область суммирования до множества всех целочисленных векторов вида $(i_1, j_1, k_1, i_2, j_2, k_2) \in \{1, \dots, N\}^6$),

$$D_4^{(2)} \leq 2eqSE_{\mu_{rml}} \frac{(n-q+3)_3 (np)^{q-3}}{r! m! l!}, \quad (48)$$

а

$$S = \sum_{1 \leq i < j < k \leq N} p_i p_j p_k (p_1 + \dots + p_k)(1 - p_i - \dots - p_k)^{n-q},$$

(здесь мы расширили область суммирования до B и воспользовались оценкой $p_i \leq p$).

Оценим величину $D_4^{(2)}$. Суммируя по j , получаем, что

$$S \leq 2pS_1^{(q)} + S_2^{(q)}, \quad (49)$$

где

$$S_t^{(q)} = \sum_{1 \leq i < k \leq N} p_i p_k (p_{i+1} + \dots + p_{k-1})^t (1 - p_i - \dots - p_k)^{n-q}, \quad t = 1, 2.$$

Нетрудно заметить, что

$$S_t^{(q)} \leq \int_0^1 \int_x^1 (y-x)^t (1+x-y)^{n-q} dx dy = \int_0^1 z^t (1-z)^{n-q+1} dz.$$

Вычисляя эти интегралы, приходим к оценкам

$$S_1^{(q)} \leq (n-q+3)_2^{-1}, \quad S_2^{(q)} \leq 2(n-q+4)_3^{-1},$$

которые вместе с (48) и (49) дают неравенство

$$D_4^{(2)} \leq \frac{4eq(np+1)^{q-2}}{r! m! l!} E_{\mu_{rml}}. \quad (50)$$

Нетрудно проверить, что при $np \leq a$ множитель при $E_{\mu_{rml}}$ не превосходит некоторой константы, не зависящей от r, m, l . Поэтому из (44), (47), (50) и неравенства $P_r \leq p^{r-1}$ следует (39). Лемма 9 доказана.

Рассмотрим величины

$$T_2 = \sum_{r+m \geq 3, r, m \geq 1} (D\mu_{rm})^{1/2},$$

$$T_3 = \sum_{r+m+l \geq 4, r, m, l \geq 1} (D\mu_{rml})^{1/2}.$$

Лемма 10. Пусть $np \leq a < \infty$. Найдутся такие величины $c'_2(a)$ и $c'_3(a)$, зависящие только от a , что

$$T_s \leq c'_s(a)(nP_2^{1/2} + n^s e^{-\sqrt{n}/2s}), \quad s = 2, 3. \quad (51)$$

Доказательство. Так как при $s = 2$ и $s = 3$ рассуждения аналогичны, мы ограничимся рассмотрением случая, когда $s = 3$.

Воспользуемся оценкой (37) и неравенством

$$(x + y)^{1/2} \leq x^{1/2} + y^{1/2}, \quad x, y \geq 0,$$

из которых следует, что

$$T_3 \leq T_{31} + T_{32} + T_{33}, \quad (52)$$

где

$$T_{31} = c_3^{1/2}(a) \sum_{r, m, l \geq 1, r+m+l \geq 4} (E\mu_{rml})^{1/2},$$

$$T_{32} = \sum_{\substack{\max\{r, m, l\} \leq \sqrt{n} \\ q=r+m+l \geq 4, r, m, l \geq 1}} \frac{en^3 a^{q-3}}{r! m! l!} e^{-n/2q},$$

$$T_{33} = \sum_{\max\{r, m, l\} > \sqrt{n}, r, m, l \geq 1} \frac{en^3 a^{q-3}}{r! m! l!} e^{-n/2q}.$$

Используя неравенства (33) и (29), для T_{31} получим оценку

$$T_{31} \leq c_3^{1/2}(a) n P_2^{1/2} \sum_{r, m, l \geq 1, r+m+l \geq 4} (s')^{-1/2} (np)^{(s-2)/2},$$

где $s = \max\{r, m, l\} \geq 2$. Проводя очевидные преобразования, находим, что

$$T_{31} \leq c_3^{1/2}(a) n P_2^{1/2} \sum_{s=2}^{\infty} \sum_{r, m, l=1}^s (s')^{-1/2} (np)^{(s-2)/2}$$

$$\leq c_3^{1/2}(a) n P_2^{1/2} \sum_{s=2}^{\infty} s^3 (s')^{-1/2} (1+a)^{(s-2)/2}.$$

Ряд в правой части сходится при любом a , следовательно, найдется такое $c_{31}(a)$, что

$$T_{31} \leq c_{31}(a) n P_2^{1/2}. \quad (53)$$

Чтобы оценить T_{32} , достаточно заметить, что

$$T_{32} < en^3 e^{-\sqrt{n}/6} \sum_{r+m+l \geq 4, r, m, l \geq 1} \frac{a^{r+m+l-3}}{r! m! l!} \leq e^{3a+1} n^3 e^{-\sqrt{n}/6}. \quad (54)$$

Наконец, оценим T_{33} . Нетрудно проверить, что

$$T_{33} \leq en^3 \sum_{\max\{r, m, l\} > \sqrt{n}} \frac{a^{r+m+l-3}}{r! m! l!} \leq 3e^{2a+1} n^3 \sum_{r \geq \sqrt{n}} \frac{a^{r-1}}{r!} \leq 3e^{3a+1} n^3 \frac{a^{\sqrt{n}} - 1}{(\sqrt{n} - 2)!}.$$

Поэтому найдется такая величина $c_{33}(a)$, что

$$T_{33} \leq c_{33}(a) n^3 e^{-\sqrt{n}/6}. \quad (55)$$

Из (52)–(55) следует (51). Лемма 10 доказана.

Лемма 11. Пусть $np \leq a$, где a — постоянная. Найдутся такие величины $c_{11}(a), c_{111}(a)$, зависящие только от a , что

$$D\mu_{11} \leq c_{11}(a)(n^2 P_2 + 1), \quad (56)$$

$$D\mu_{111} \leq c_{111}(a)(n^2 P_2 + 1). \quad (57)$$

Доказательство. В силу (30) и неравенства Минковского (см. например, [6, с. 30])

$$(D\mu_{11})^{1/2} \leq (D\mu_1)^{1/2} + \sum_{r=2}^{n-1} (D\mu_{1r})^{1/2} + 1.$$

Заметим, что сумма по r в правой части не превосходит T_2 , поэтому в силу лемм 6 и 10

$$(D\mu_{11})^{1/2} \leq (c_1^{1/2}(a) + c_2'(a)) n P_2^{1/2} + c_2'(a) n^2 e^{-\sqrt{n}/4} + 1.$$

Отсюда следует, что найдется такая величина $c_{11}(a)$, при которой будет выполнено неравенство (56).

Неравенство (57) доказывается аналогично с использованием (31) и (56) вместо (30) и (37). Лемма 11 доказана.

Перейдем к доказательству леммы 5.

Доказательство. Оценим сначала $P(C_\delta^{(1)})$. Из (24), (14) и (12) следует, что

$$E\zeta = N - E\mu_0 - \sum_{r, m \geq 1} a_{rm} E\mu_{rm}. \quad (58)$$

Используя (20), (24), (58) и (13), получаем оценки

$$\begin{aligned} P(C_\delta^{(1)}) &\leq P\{|\mu_0 - E\mu_0| + \sum_{r, m \geq 1} |\mu_{rm} - E\mu_{rm}| > \delta\sqrt{n}\} \\ &\leq \frac{1}{\delta\sqrt{n}} \left(E|\mu_0 - E\mu_0| + \sum_{r, m \geq 1} E|\mu_{rm} - E\mu_{rm}| \right) \\ &\leq \frac{1}{\delta\sqrt{n}} \left((D\mu_0)^{1/2} + \sum_{r, m \geq 1} (D\mu_{rm})^{1/2} \right). \end{aligned}$$

Применив к правой части этих неравенств леммы 6, 11, 10 и воспользовавшись условиями (3), находим, что при любом $\delta > 0$

$$P(C_\delta^{(1)}) \rightarrow 0. \quad (59)$$

Оценим $P(C_\delta^{(2)})$. Нетрудно проверить, что

$$P(C_\delta^{(2)}) \leq P'_\delta + P''_\delta,$$

где

$$\begin{aligned} P'_\delta &= \mathbf{P}\{\sigma^2(\nu_1, \dots, \nu_N) \leq (n+1)/(12(1+\delta)^2)\}, \\ P''_\delta &= \mathbf{P}\{\sigma^2(\nu_1, \dots, \nu_N) \geq (n+1)/(12(1-\delta)^2)\}. \end{aligned} \quad (60)$$

В силу (24) и (25)

$$|\sigma^2(\nu_1, \dots, \nu_N) - a_{111} \mathbf{E}\mu_{111} - E| \leq \kappa, \quad (61)$$

где

$$E = \omega + \sum_{r+m \geq 3} \mathbf{E}\mu_{rm}, \quad \kappa = \sum_{r,m,l \geq 1} |\mu_{rml} - \mathbf{E}\mu_{rml}|.$$

Так как $E \geq 0$, то из (60) и (61) следует, что

$$P'_\delta \leq \mathbf{P}\{\kappa \geq a_{111} \mathbf{E}\mu_{111} - (n+1)/(12(1+\delta)^2)\}.$$

Воспользуемся теперь оценкой (35), равенством $a_{111} = 1/12$ и неравенствами

$$\mathbf{P}\{X > \varepsilon\} \leq \varepsilon^{-1} \mathbf{E}X, \quad (\mathbf{E}X)^2 \leq \mathbf{E}X^2,$$

где X — случайная величина. Получаем, что

$$P'_\delta \leq \left(\frac{n - 3n^2 P_2 - 2}{12} - \frac{n+1}{12(1+\delta)^2} \right)^{-1} \sum_{r,m,l} (\mathbf{D}\mu_{rml})^{1/2}.$$

Из этого неравенства с помощью (3) и лемм 10 и 11 выводим, что $P'_\delta \rightarrow 0$ при любом $\delta \in (0, 1/2)$.

Аналогично доказывается, что $P''_\delta \rightarrow 0$ при любом $\delta \in (0, 1/2)$. В этом случае

$$P''_\delta \leq \mathbf{P}\{\kappa \geq (n+1)/(12(1+\delta)^2) - a_{111} \mathbf{E}\mu_{111} - E\}.$$

Величина $\mathbf{E}\mu_{111} \leq n$, а величина E оценивается как $O(n^2 P_2 + 1)$ с помощью (29).

Таким образом, доказано, что $P(C_\delta^{(2)}) \rightarrow 0$ при любом $\delta \in (0, 1/2)$. Вместе с (59) это означает, что $P(C_\delta) \rightarrow 0$ при любом $\delta \in (0, 1/2)$. Лемма 5 доказана.

Несколько слов о формуле для $\mathbf{E}\zeta$, указанной в замечании 1. Она выводится из (58) с помощью оценок (29), (32) и тождеств

$$\begin{aligned} \mu_0 + \mu_1 + \dots + \mu_n &= N, \\ \mu_1 + 2\mu_2 + \dots + n\mu_n &= n. \end{aligned}$$

Используется также равенство

$$\mathbf{E}\mu_2 = \binom{n}{2} P_2 (1 + O(np)),$$

доказываемое непосредственными вычислениями.

Список литературы

1. Ватутин В. А. Предельные теоремы для числа отрезков возрастания в случайных перестановках, порождаемых алгоритмами сортировки. *Дискретная математика* (1994) **6**, №2, 83–99.
2. Кнут Д. *Искусство программирования для ЭВМ*. Мир, Москва, 1978.
3. Barton D. E., Mallows C. L. Some aspects of the random sequences. *Ann. Math. Statist.* (1965) **36**, №1, 236–260.
4. Шергин В. В. О скорости сходимости в центральной предельной теореме для m -зависимых случайных величин. *Теория вероятностей и ее применения* (1979) **24**, №4, 781–794.
5. Колчин В. Ф., Севастьянов Б. А., Чистяков В. П. *Случайные размещения*. Наука, Москва, 1976.
6. Петров В. В. *Суммы независимых случайных величин*. Наука, Москва, 1972.

Статья поступила 1.12.94.