УДК 519.2 Вестник СПбГУ. Сер. 1. Т. 1 (59). 2014. Вып. 2

О МАКСИМАЛЬНОМ ЗНАЧЕНИИ СРЕДНЕГО ЧИСЛА РЕКОРДОВ*

В. Б. Невзоров, С. А. Товмасян

Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7/9

Имеются n независимых случайных величин с общей непрерывной функцией распределения. Рассматривается проблема как, получая по очереди значения этих величин, выбрать одну из них в качестве начала отсчета, чтобы максимизировать при этом математическое ожидание числа рекордов среди оставшихся случайных величин, значения которых нам не известны. Библиогр. 4 назв. Табл. 1.

Ключевые слова: рекордные моменты, рекордные величины, индикаторы рекордов.

1. Рассмотрим последовательность независимых случайных величин (с.в.) $X_1, X_2, \ldots$ с общей непрерывной функцией распределения (ф.р.) $F(x)$ и порядковые статистики $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$, $n = 1, 2, \ldots$, построенные по этой последовательности. Каждой из величин X_n соответствует последовательный ранг R_n, принимающий с равными вероятностями $1/n$ значения $1, 2, \ldots, n$. Справедливы равенства

$$P\{R_n = m\} = P\{X_n = X_{m,n}\}, \quad 1 \leq m \leq n, \quad n = 1, 2, \ldots \tag{1}$$

Определим рекордные индикаторы $\xi_1, \xi_2, \ldots$ таким образом, что $\xi_n = 1$, если $R_n = n$, и $\xi_n = 0$, если $R_n < n$. Если $\xi_n = 1$, то это означает, что X_n представляет собой верхнее рекордное значение, т. е. $X_n > \max\{X_1, X_2, \ldots, X_{n-1}\}$. Заметим, что если ф.р. F непрерывна, то

$$P\{\xi_n = 1\} = 1/n, \quad n = 1, 2, \ldots$$

Верхние рекордные величины $X(1) < X(2) < \ldots$, где $X(1) = X_1$, образуют монотонно возрастающую подпоследовательность $X_{L(1)} < X_{L(2)} < \ldots$ последовательности $X_1, X_2, \ldots$

Случайные индексы $1 = L(1) < L(2) < \ldots$ называют рекордными моментами. Имеется большое число работ (см., например, [1–3]), в которых исследуются распределения и свойства рекордных величин $X(n)$ и рекордных моментов $L(n)$.

Рассмотрим следующую проблему. Пусть

$$N(n) = \xi_1 + \ldots + \xi_n, \quad n = 1, 2, \ldots \tag{2}$$

обозначает число рекордных значений среди величин $X_1, X_2, \ldots, X_n$ и $a(n) = EN(n)$ — среднее число рекордов среди первых n X-ов. Отметим, что $a(n) = 1 + 1/2 + \ldots + 1/n$ и $a(n) \sim \ln n, n \to \infty$.

Если нам становится известным значение x_1 случайной величины X_1, то среднее число рекордов в наборе $x_1, X_2, X_3, \ldots, X_n$ будет больше $a(n)$, если наблюдаемое значение x_1 достаточно мало, или меньше $a(n)$ — при больших значениях x_1. Возникает вопрос, насколько мы можем увеличить значение $EN(n)$, если начнём отсчёт рекордов не с X_1 (в случае, когда эта величина принимает слишком большие значения),

*Работа выполнена при финансовой поддержке СПбГУ (грант №6.38.672.2013) и РГНФ (грант №13-02-00338).

не с X_2 (если соответствующее значение x_2 также слишком велико), а подсчитаем, дождавшись появления некоторого достаточно малого значения x_k, среднее число рекордов в наборе $x_k, X_{k+1}, X_{k+2}, \ldots, X_n$. Каждый раз, получив очередное наблюдение x_k, нужно решать, начать ли отсчёт рекордов со случайной величины $X_k = x_k$ или, отвергнув её (и уже не имея возможности в дальнейшем к ней вернуться), продолжить процедуру, рассматривая значение x_{k+1} следующей в нашей последовательности с.в. X_{k+1}. Наша цель — получить в результате такого процесса максимально возможное среднее значение числа рекордов.

Естественно, что эта задача близка по постановке к классической «проблеме секретаря» («задаче о разборчивой невесте»), в которой с помощью подобной процедуры нужно по сути максимизировать вероятность выбора последней рекордной величины в наборе $X_1, X_2, \ldots, X_n$. Также близкой к данной задаче (но только существенно более сложной по процедуре) является другая классическая проблема, связанная с нахождением математического ожидания максимального числа элементов монотонно возрастающей подпоследовательности X-ов, которую можно зафиксировать в наборе $X_1, X_2, \ldots, X_n$. В этой проблеме представляют интерес два её варианта. В первом из них, зафиксировав в качестве очередного k-го элемента монотонной подпоследовательности некоторое наблюдаемое значение $x_{\alpha(k)}$, переходят последовательно к следующим наблюдениям $x_{\alpha(k)+1}, x_{\alpha(k)+2}, \ldots$, чтобы, дождавшись появления нужного значения $x_{\alpha(k+1)}$, отвечающего процедуре, взять его в качестве следующего элемента подпоследовательности. Ещё раз подчеркнём, что в этом варианте задачи, каждый раз выбирая (или отвергая) в качестве элемента монотонной подпоследовательности некоторое очередное наблюдаемое значение x_m, мы не знаем значений с.в. $X_{m+1}, X_{m+2}, \ldots, X_n$. Было показано [4], что при такой процедуре выбора максимальное среднее число $ET_1(n)$ элементов монотонной подпоследовательности удовлетворяет соотношению

$$ET_1(n) \sim (2n)^{1/2}, \quad n \to \infty. \tag{3}$$

Во втором варианте задачи, когда выбор нужных элементов будет производиться после знакомства с наблюдаемыми значениями всех X-ов, соответствующее математическое ожидание $ET_2(n)$ числа элементов подпоследовательности имеет порядок

$$ET_2(n) \sim 2n^{1/2}, \quad n \to \infty. \tag{4}$$

Рассматриваемая нами модель имеет более простую конструкцию, чем в первом варианте задаче о выборе максимальной монотонной подпоследовательности, что позволяет получить более точные соотношения для соответствующих математических ожиданий $a(n)$.

2. Пусть независимые случайные величины (с.в.) $X_1, X_2, \ldots$ имеют общую непрерывную функцию распределения (ф.р.) $F(x)$. Предполагается последовательно получать значения $x_1, x_2, \ldots$ этих случайных величин и, наблюдая каждый раз очередное значение, решать, что делать дальше. Можно начать отсчёт рекордов с данного наблюдаемого значения или, отвергнув его, продолжить процедуру, переходя к следующей случайной величине. Для описания соответствующей процедуры достаточно рассмотреть случай, когда исходные случайные величины (обозначим их в этой ситуации $U_1, U_2, \ldots, U_n, \ldots$) имеют равномерное $U([0,1])$ распределение с ф.р. $F(x) = x, 0 \le x \le 1$.

Пусть $N(n,x)$ обозначает число рекордов среди величин $U_1 = x, U_2, \ldots, U_n$, а $T(n,x) = E(n,x)$. Рассмотрим также индикаторы $\xi_k, k = 2,3,\ldots$, такие, что $\xi_k = 1$, если $U_k > \max\{x, U_2, \ldots, U_{k-1}\}$, и $\xi_k = 0$ иначе. Получаем, что

$$P\{\xi_k = 1\} = \int_x^1 u^{k-2}du = (1 - x^{k-1})/(k-1)$$

и

$$T(n,x) = 1 + \sum_{k=2}^{n} \frac{1}{k-1}(1 - x^{k-1}) = 1 + a(n-1) - \sum_{k=1}^{n-1}(x^k/k). \tag{5}$$

Сравнивая $T(n,x)$ со средним числом $a(n-1)$ рекордных значений в наборе $U_2, \ldots, U_n$, к которому мы можем перейти при слишком больших значениях x с.в. U_1, видим, что $T(n,x) < a(n-1)$, если

$$\sum_{k=1}^{n-1}(x^k/k) > 1. \tag{6}$$

В этом случае мы отвергаем $U_1 = x$, переходим к с.в. U_2 и так далее.

Таким образом, при фиксированном $n = 3,4,\ldots$ процедура максимизации среднего числа рекордов сводится к нахождению критических значений $x(n), x(n-1), \ldots$, являющихся корнями уравнений

$$\sum_{k=1}^{r-1}(x(r)^k/k) = 1, \quad r = 3,4,\ldots \tag{7}$$

Как только происходит первое из событий $\{U_1 < x(n)\}, \{U_1 \geq x(n), U_2 < x(n-1)\}, \ldots, \{U_1 \geq x(n), U_2 \geq x(n-1), \ldots, U_{r-1} \geq x(n-r+2), U_r < x(n-r-1)\}$, выбираем значение с.в. U_r в качестве начальной точки отсчета рекордов.

Обозначим через $e(n)$ среднее число рекордов, которое обеспечивает данная процедура начала их отсчёта. Очевидно, что $e(1) = 1, e(2) = 3/2$, поскольку при $n = 1$ и $n = 2$ невозможно увеличить эти значения за счёт переноса начала отсчёта рекордов. Если $n = 3$ и мы имеем три величины U_1, U_2 и U_3, то ищем соответствующее решение $0 < x(3) < 1$ уравнения

$$x(3)^2/2 + x(3) = 1$$

и получаем, что если $U_1 \geq x(3) = \sqrt{3} - 1 = 0,73205\ldots$, то переходим (и уже окончательно останавливаемся) к набору U_2, U_3, гарантирующему среднее число рекордов $a(2) = 3/2$. Если же $U_1 = x_1 < \sqrt{3} - 1$, то останавливаемся на наборе (x_1, U_2, U_3). В итоге получаем, что

$$e(3) = \int_0^{x(3)} (1 + a(2) - u - u^2/2)du + 3(1 - x(3))/2 = 1,898717\ldots$$

и среднее число рекордов в результате применения такой процедуры увеличивается на величину $d(3) = e(3) - a(3) = 0,065384\ldots$

Для $n > 2$ нужно использовать рекуррентное соотношение

$$e(n) = \int_0^{x(n)} T(n,u)du + (1 - x(n))e(n-1), \quad n = 3, 4, 5 \ldots,$$

которое сводится к равенству

$$e(n) = x(n)(1 + a(n-1)) - \sum_{k=1}^{n-1} \frac{x(n)^{k+1}}{k(k+1)} + (1 - x(n))e(n-1). \tag{8}$$

Учитывая, что

$$\sum_{k=1}^{n-1} (x(n)^k/k) = 1,$$

получаем, что справедливо соотношение

$$e(n) = x(n)a(n-1) + (1 - x(n))(1 + e(n-1)) + (x(n))^n/n. \tag{9}$$

Для разностей $d(n) = e(n) - a(n), n = 3, 4, \ldots$, будут тогда выполняться равенства

$$d(n) = (1 - x(n))d(n-1) + (1 - x(n)) - 1/n + (x(n))^n/n. \tag{10}$$

Соотношения (7), (9) и (10) позволили получить следующие (приведённые в таблице) значения величин $x(n), e(n)$ и $d(n)$.

n	$x(n)$	$e(n)$	$d(n)$	n	$x(n)$	$e(n)$	$d(n)$
3	0,732050808	1,898717474	0,065384141	24	0,632121207	4,290253984	0,514295807
4	0,672495811	2,233383757	0,150050424	25	0,632120953	4,333036295	0,517078117
5	0,651099053	2,50798968	0,224656347	30	0,632120592	4,523125871	0,52813874
6	0,641771969	2,733684397	0,283684397	35	0,632120562	4,682756211	0,535974792
7	0,637265996	2,921733417	0,328876274	40	0,632120559	4,820361162	0,541818123
8	0,634949572	3,081266352	0,363409209	41	0,632120559	4,845746228	0,542812946
9	0,633709492	3,219092336	0,390124082	45	0,632120559	4,941291656	0,54634354
10	0,633026867	3,340139731	0,411171477	46	0,632120559	4,963816113	0,547128867
11	0,632643531	3,447961911	0,428084566	47	0,632120559	4,985844149	0,547880307
12	0,632425057	3,545146583	0,441935905	48	0,632120559	5,007397184	0,548600009
13	0,632299128	3,6336104	0,453476645	49	0,632120559	5,028495281	0,549289943
14	0,632225894	3,714801263	0,463238937	50	0,632120559	5,049157257	0,549951918
15	0,632183	3,789835118	0,471606125	60	0,632120559	5,235215378	0,555344965
20	0,632125377	4,098013975	0,500274317	70	0,632120559	5,392019831	0,559183073
21	0,632123469	4,149658493	0,504299788	80	0,632120559	5,527533376	0,562054097
22	0,632122319	4,198758906	0,507945656	90	0,632120559	5,646853299	0,564282696
23	0,632121626	4,245554985	0,511263474	100	0,632120559	5,753440335	0,566062818

Мы рассмотрели в качестве исходной последовательность случайных величин $U_1, U_2, \ldots$, имеющих равномерное $U([0,1])$ распределение. При переходе к независимым с.в. $X_1, X_2, \ldots$ с произвольной непрерывной ф.р. $F(x)$ изменятся лишь значения $x(n), n = 3, 4, \ldots$ Вместо табличных значений $x(n)$ нужно будет брать величины $G(x(n))$, где $G(x) = F(x)^{-1}$ — функция, обратная функции распределения $F(x)$

3. Отметим, что величины $x(n)$ монотонно убывают и их предел x^* можно найти, учитывая, что при $n \to \infty$ из соотношения (7) получим равенство

$$-\ln(1 - x^*) = 1, \tag{11}$$

из которого следует, что $x^* = 1 - 1/e = 0,63212\ldots$ Теперь уже нетрудно найти и предельное значение d^* монотонно возрастающей последовательности $0 = d(2) < d(3) < \ldots$ Из равенства (10) получаем, что для этого предела справедливо соотношение

$$d^* = (1 - x^*)d^* + (1 - x^*). \tag{12}$$

Таким образом получаем, что

$$d^* = (1 - x^*)x^* = 1/(e - 1) = 0,58197\ldots \tag{13}$$

Литература

1. *Ahsanullah M., Nevzorov V. B.* Ordered random variables. New York: Nova Science Publishers, 2001. 412 p.

2. *Arnold C., Balakrishnan N., Nagaraja H. N.* Records. New York: John Wiley and Sons, 1998. 312 p.

3. *Невзоров В. Б.* Рекорды. Математическая теория. М.: ФАЗИС, 2000. 244 с.

4. *Samuels S. M., Steele J. M.* Optimal sequential selection of a monotone sequence from a random sample // Ann. Probab., 1981. Vol. 9, N 6. P. 937–947.

Статья поступила в редакцию 26 декабря 2013 г.

Сведения об авторах

Невзоров Валерий Борисович — доктор физико-математических наук, профессор; probabil@pisem.net

Товмасян Сергей Ашотович — аспирант; sergei_tovmasyan@mail.ru

ON THE MAXIMAL VALUE OF EXPECTATION OF RECORD NUMBERS

Valery B. Nevzorov, Sergey A. Tovmasyan

St.Petersburg State University, Universitetskaya nab., 7/9, St.Petersburg, 199034, Russian Federation; probabil@pisem.net, sergei_tovmasyan@mail.ru

There are n independent random variables with some continuous distribution function. We consider the problem, how getting sequentially values of this variables and selecting one of them as the initial point to maximize (without knowledge of future values) the expected number of records among the rest in this sequence random variables. Refs 4. Tables 1.

Keywords: record times, record values, record indicators.