Записки научных
семинаров ПОМИ
Том 446, 2017 г.

**И. В. Бельков, В. Б. Невзоров**

# ОБ ОДНОЙ ЗАДАЧЕ ОПТИМАЛЬНОГО ВЫБОРА РЕКОРДНЫХ ВЕЛИЧИН

Пусть $X_1, X_2, \ldots$ – последовательность независимых случайных величин (с.в.), имеющих непрерывную функцию распределения (ф.р.) $F(x)$. Определим для этой последовательности верхние рекордные моменты $L(n)$ и верхние рекордные величины $X(n)$, $n = 1, 2, \ldots$. Задаем эти случайные величины следующими соотношениями:

$$L(1) = 1, \quad X(1) = X_1$$

и

$$L(n) = \min\{j : X_j > X(n-1)\}, \quad X(n) = X_{L(n)}, \quad n = 2, 3, \ldots.$$

Будем также использовать специальные обозначения $U_1, U_2, \ldots$ для с.в., имеющих U([0, 1])-равномерное распределение на интервале $[0, 1]$, и $U_1 = U(1) < U(2) < \ldots$ – для соответствующих верхних рекордных величин. Различные результаты для рекордных моментов и рекордных величин можно найти, например, в монографиях [1–3].

С рекордными величинами связана классическая задача оптимального выбора, также называемая "задачей о разборчивой невесте" или "проблемой секретаря", которую можно сформулировать следующим образом.

Имеется набор из $n$ случайных величин $X_1, X_2, \ldots, X_n$. Выделим из них случайное число $M = M(n)$ рекордных величин

$$X(1) = X_1 < X(2) < \cdots < X(M) = \max\{X_1, X_2, \ldots, X_n\}.$$

Последовательно получаем наблюдаемые значения $x_1, x_2, ..., x_n$ этих с.в. Получив очередное наблюдение $x_k$, которое больше всех предыдущих величин $x_1, x_2, \ldots, x_{k-1}$, нужно решить, останавливаться ли на этом наблюдении, предполагая, что оно и является значением с.в. $X(M)$, или, расставшись с этим наблюдением (и уже не имея возможности в этом случае к нему вернуться), продолжить процесс выбора в надежде выйти позже на максимальное из всех $n$ наблюдений.

*Ключевые слова*: рекордные моменты, рекордные величины, суммы рекордных величин, среднее число рекордов, равномерное распределение, задача оптимального выбора.

В классической постановке "задачи о разборчивой невесте", или "проблемы секретаря"( [4–6]), последовательно должны появиться $n$ (известное заранее число) претендентов. Невесту или работодателя интересует какой-то определенный, присущий претенденту, фактор, выраженный количественно, например, зарплата жениха или стаж работы по специальности кандидата в секретари. Опираясь на последовательно получаемые при знакомстве с претендентами соответствующие данные $x_1, x_2, \ldots$, невеста или работодатель надеются не пропустить самого достойного из кандидатов. По сути дела, это и означает, что нужно угадать момент появления последнего рекорда среди величин $X_1, X_2, \ldots, X_n$.

Оптимальная стратегия (см., например, [5, 6]) заключается в том, что нужно при фиксированном числе $n$ элементов выборки пропустить какое-то число $r = r(n)$ наблюдений $x_1, x_2, \ldots, x_r$, а затем выбрать первое же значение $x_s$, $s > r$, которое превышает все предыдущие. Имеются таблицы, позволяющие для значений $n = 2, 3, \ldots$ находить величины $r = r(n)$ и вероятности $p_n$ угадать при этом момент появления последнего рекордного значения в наборе $X_1, X_2, \ldots, X_n$. Отметим лишь, что при $n \to \infty$ получаем, что $r(n) \sim n/e$ и $p_n \to 1/e = 0,368\ldots$.

В статье [7] была рассмотрена задача оптимального выбора, в которой интерес представлял вопрос о возможности увеличить математическое ожидание числа рекордов среди с.в. $X_1, X_2, \ldots, X_n$, если начинать их отсчет не с первого элемента последовательности, а лишь дождавшись появления некоторого достаточно малого по величине наблюдения $x_r$.

Мы рассмотрим несколько иной вариант задачи оптимального выбора, также связанный с рекордными величинами в наборе $X_1, X_2, \ldots, X_n$. Пусть $M$ – число рекордов $X(1) < X(2) < \cdots < X(M)$ среди этих величин, $S_n$ – сумма этих рекордов и $T(n) = \mathbf{E}S_n$ – сумма их математических ожиданий.

Приведем следующий результат для общего случая (для $X$-ов с непрерывной ф.р. $F(x)$).

**Лемма 1.** *Для любого $n = 1, 2, \ldots$ справедливо равенство*

$$T(n) = \int_{-\infty}^{\infty} x \frac{1 - F^n(x)}{1 - F(x)}\, dF(x). \tag{1}$$

**Доказательство.**

$$T(n) = \sum_{k=1}^{n} \int_{-\infty}^{\infty} xF^{k-1}(x)\,dF(x)$$
$$= \int_{-\infty}^{\infty} \sum_{k=0}^{n-1} F^k(x)\,dF(x) = \int_{-\infty}^{\infty} \frac{1-F^n(x)}{1-F(x)}\,dF(x). \qquad \square$$

**Следствие.** *Если $X_1, X_2, \ldots$ имеют равномерное $U([0,1])$-распределение, то*

$$T(n) = \sum_{k=1}^{n} \int_{0}^{1} x^k dx = \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n+1}, \quad n = 1, 2, \ldots. \qquad (2)$$

Рассмотрим новую задачу оптимального выбора в ситуации, когда исходные случайные величины имеют $U([0,1])$-распределение. Зададимся целью, как и в классической задаче оптимального выбора, максимально увеличить (по сравнению с выражением в правой части равенства (2)) математическое ожидание суммы рекордных значений среди величин $U_1, U_2, \ldots, U_n$ за счет правильного выбора в этом наборе начальной точки отсчета рекордов, т.е. за счет того, что не все рекордные величины будем фиксировать. Дело в том, что если, скажем, начальная в этом наборе величина $U_1$ принимает достаточно большое (близкое к единице) значение, то оно вносит хороший вклад в интересующую нас сумму $T(n)$, но уже остальные из величин $U_2, \ldots, U_n$ имеют существенно меньше шансов стать рекордными и внести свой вклад в сумму $T(n)$. Остается надеяться, что если не будем принимать в расчет эту случайную величину, а перейдем к набору $U_2, \ldots, U_n$, то сможем получить существенно большее значение для математического ожидания суммы рекордных величин. Надо при этом учесть, что история может повториться и на следующих этапах выбора начальной точки отсчета рекордов. Каким же должно быть наблюдаемое значение $x_1$ случайной величины $U_1$, которое позволит определить, учитывать его как первое рекордное или, отвергнув это наблюдение, перейти к набору $U_2, U_3, \ldots, U_n$. Конечно, если $n = 1$ или $n = 2$, то никакие переходы не могут увеличить величины $T(n)$.

Итак, рассмотрим ситуацию, когда $U_1 = x$, $0 \leqslant x \leqslant 1$. Пусть $T_n(x)$ обозначает математическое ожидание суммы рекордов в наборе $x, U_2, \ldots, U_n$. В этом случае справедлив следующий результат.

**Лемма 2.** *Для любого* $n = 2, 3, \ldots$

$$T_n(x) = \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} + x - \frac{x^2}{2} - \frac{x^3}{3} - \ldots - \frac{x^n}{n}. \tag{3}$$

Действительно, для ф.р. $F(x) = x$, $0 \leqslant x \leqslant 1$, справедливы равенства

$$\begin{aligned} T_n(x) &= x + \int_1^x u\, dF(u) + \int_1^x uF(u)\, dF(u) + \ldots + \int_1^x uF^{n-2}(u)\, dF(u) \\ &= x + \int_1^x u\, du + \int_1^x u^2\, du + \ldots + \int_1^x u^{n-1}\, du \\ &= \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} + x - \frac{x^2}{2} - \frac{x^3}{3} - \ldots - \frac{x^n}{n}. \end{aligned}$$

Таким образом, нужно начать со сравнения при каждом фиксированном $n = 2, 3, \ldots$ и любом $0 < x < 1$ величины $T_n(x)$ и математического ожидания суммарного числа рекордов в наборе $U_2, U_3, \ldots, U_n$, которое совпадает с $T(n-1)$. В дальнейшем (возможно!) потребуются такого типа сравнения при переходе от $U_1$ к $U_2$, от $U_2$ к $U_3$ и т.д.

Рассмотрим разности

$$R_n(x) = T(n-1) - T_n(x), \quad n = 2, 3, \ldots.$$

Если $U_1 = x$ и $R_n(x) \leqslant 0$, то надо начинать отсчет рекордов с величины $U_1$. Если же $R_n(x) > 0$, то следует, отбросив $U_1$, переходить к величине $U_2$ и рассматривать уже знаки величины $R_{n-1}(x) = T(n-2) - T_{n-1}(x)$ с возможным переходом к следующему элементу $U_3$ и так далее. Поэтому появляется необходимость знать при каждом $n = 2, 3, \ldots$ значения $x_n$, представляющие корни уравнения

$$R_n(x) = 0, \quad 0 < x < 1,$$

в котором

$$R_n(x) = T(n-1) - T_n(x)$$
$$= \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} - \Big(\Big(\frac{1}{2} + \ldots + \frac{1}{n}\Big) + x - \frac{x^2}{2} - \cdots - \frac{x^n}{n}\Big)$$
$$= x\Big(\frac{x}{2} + \frac{x^2}{3} + \ldots + \frac{x^{n-1}}{n} - 1\Big). \quad (4)$$

Таким образом, при каждом $n$ нас интересуют корни уравнений

$$\frac{x}{2} + \frac{x^2}{3} + \ldots + \frac{x^{n-1}}{n} = 1, \quad (5)$$

лежащие внутри интервала $(0,\ 1)$. Отметим сразу, что таких корней нет, если $n = 2$ и $n = 3$.

Заметим также, что справедливы неравенства

$$1 > x_4 > x_5 > \cdots > x_m > x_{m+1} > \cdots . \quad (6)$$

В приводимой таблице даны значения $x_n$ (с точностью до 4 знаков после запятой) для $n = 4, 5, \ldots, 10$.

| $n$ | $x_n$ |
|---|---|
| 4 | 0,9554 |
| 5 | 0,8852 |
| 6 | 0,8509 |
| 7 | 0,8318 |
| 8 | 0,8204 |
| 9 | 0,8131 |
| 10 | 0,8084 |

Пусть $V_n$ обозначает полученное при такой процедуре математическое ожидание суммарного числа рекордов в наборе $X_1, X_2, \ldots, X_n$. Очевидно, что если $n = 1, 2, 3$, то

$$V_1 = T(1) = \frac{1}{2},$$
$$V_2 = T(2) = \frac{1}{2} + \frac{1}{3} = \frac{5}{6}, \quad (7)$$
$$V_3 = T(3) = \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = \frac{13}{12}.$$

Если $n = 4$, то принимаем во внимание значение $x_4 = 0{,}9554\ldots$. Если $X_1 \leqslant x_4$, то рассматриваем все рекорды, включая $X_1$, но если $X_1 > x_4$,

то учитываем только рекорды в наборе $X_2, X_3, X_4$. Получаем в этом случае, что

$$V_4 = (1-x_4)T(3) + \int_0^{x_4} T_4(u)\,du = (1-x_4)T(3) + \int_0^{x_4} (T(3) - R_4(u))\,du$$

$$= T(3) - \int_0^{x_4} \Big(\frac{u^2}{2} + \frac{u^3}{3} + \frac{u^4}{4} - u\Big)\,du = T(3) + \frac{(x_4)^2}{2} - \frac{(x_4)^3}{6} - \frac{(x_4)^4}{12} - \frac{(x_4)^5}{20}$$

$$= \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{(x_4)^2}{2} - \frac{(x_4)^3}{6} - \frac{(x_4)^4}{12} - \frac{(x_4)^5}{20}.$$

Если $n = 5$, то

$$V_5 = (1-x_5)V_4 + \int_0^{x_5} T_5(u)\,du = (1-x_5)V_4 + \int_0^{x_5} (T(4) - R_5(u))\,du$$

$$= (1-x_5)V_4 + x_5T(4) - \int_0^{x_5} \Big(\frac{u^2}{2} + \frac{u^3}{3} + \frac{u^4}{4} + \frac{u^5}{5} - u\Big)\,du$$

$$= (1-x_5)V_4 + x_5T(4) + \frac{(x_5)^2}{2} - \frac{(x_5)^3}{6} - \frac{(x_5)^4}{12} - \frac{(x_5)^5}{20} - \frac{(x_5)^6}{30}.$$

Для $n > 5$ справедливы аналогичные рекуррентные соотношения, задаваемые равенствами

$$V_n = (1-x_n)V_{n-1} + \int_0^{x_n} T_n(u)\,du$$

$$= (1-x_n)V_{n-1} + \int_0^{x_n} (T(n-1) - R_n(u))\,du.$$

Приведем таблицу значений $V_n$ с точностью до 4 знаков после запятой для $n = 4, 5, \ldots, 10$. В таблице даны также значения величин $T(n)$ и $d(n) = V_n - T(n)$.

| $n$ | $V_n$ | $T(n)$ | $d(n)$ |
|---|---|---|---|
| 4 | 1,2851 | 1,2833 | 0,0018 |
| 5 | 1,4653 | 1,4500 | 0,0153 |
| 6 | 1,6253 | 1,5928 | 0,0324 |
| 7 | 1,7668 | 1,7178 | 0,0490 |
| 8 | 1,8927 | 1,8289 | 0,0637 |
| 9 | 2,0054 | 1,9289 | 0,0785 |
| 10 | 2,1072 | 2,0198 | 0,0873 |

Можно отметить, возвращаясь к неравенствам (6), что имеет место соотношение $\lim x_n = x^*$, $n \to \infty$, где $x^*$ представляет собой единственный внутри интервала (0,1) корень уравнения

$$-\ln(1-x) = 2x.$$

Этот корень вычислен с точностью до 9 знаков после запятой и $x^* = 0{,}796812130\ldots$. Величины $x_n$ достаточно быстро стремятся к этому предельному значению $x^*$. В частности, $x_{30} = 0{,}796854973\ldots$, $x_{50} = 0{,}796812417\ldots$, а $x_{100}$ уже практически (с точностью до 9 знаков после запятой) совпадает с $x^*$. Уравнение $-\ln(1-x) = 2x$ появляется как предельное при $n \to \infty$ для соотношений

$$x + \frac{x^2}{2} + \frac{x^3}{3} + \ldots + \frac{x^n}{n} = 2x,$$

получаемых из равенств (5).

## Литература


1. B. C. Arnold, N. Balakrishnan, H. N. Nagaraja, *Records*, Wiley, New York, 1998.
2. M. Ahsanullah, V. B. Nevzorov, *Records via probability theory*, Atlantis Press, 2015.
3. В. Б. Невзоров, *Рекорды. Математическая теория*, Фазис, М., 2000.
4. M. Gardner, *Mathematical Games. A fifth collection of "brain-teasers.* — Sci. Amer. **202**, No. 2 (1960), 150–154.
5. Е. Б. Дынкин. *Оптимальный выбор момента остановки марковского процесса.* — ДАН СССР **150**, No. 2 (1963), 238–240.
6. С. М. Гусейн-Заде, *Разборчивая невеста*, МЦНМО, М., 2003.
7. В. Б. Невзоров, С. А. Товмасян, *О максимальном значении среднего числа рекордов.* — Вестник СПбГУ, сер. 1, вып. 2, том 1 **59** (2014), 196–200.

Bel'kov I. V., Nevzorov V. B. On one problem of the optimal choice of record values.

Independent random variables $X_1, X_2, \ldots, X_n$ having U([0,1])-uniform distribution and the upper record values in this set are considered. We study the problem how to maximize (taking into account some consecutively observed values $x_1, x_2, \ldots, x_k$ of these $X$-s) the expectation of sums of records in this sequence under the optimal choice of the corresponding variable $X_k$ (instead of $X_1$) as the initial record value.

С.-Петербургский
государственный университет,
Университетский пр., 28,
Петродворец,
198504, С.-Петербург, Россия
*E-mail*: igor.belkov@gmail.com
*E-mail*: valnev@mail.ru

Поступило 12 октября 2017 г.