УДК 003.26+519.212.2

Миронкин В.О., Чухно А.Б.

Национальный исследовательский университет «Высшая школа экономики», Москва

ОБ ОДНОМ ОБОБЩЕНИИ ПАРАДОКСА «ДНЕЙ РОЖДЕНИЯ»

В работе рассматривается обобщение классического парадокса «дней рождения» на случай нескольких независимых в совокупности упорядоченных выборок произвольной мощности. Получены точные и асимптотические выражения, описывающие вероятность пересечения соответствующих выборок.

ПАРАДОКС «ДНЕЙ РОЖДЕНИЯ», УПОРЯДОЧЕННАЯ ВЫБОРКА, КОЛЛИЗИЯ

Mironkin V.O, Chukhno A.B.

National Research University Higher School of Economics, Moscow

ON ONE GENERALIZATION OF THE BIRTHDAY PROBLEM

In this paper a generalization of the classical birthday problem for the case of several independent samples of arbitrary power is considered. Exact and asymptotic expressions describing the probability of the intersection of these samples are obtained.

THE BIRTHDAY PROBLEM, ORDERED SAMPLE, COLLISION

Введение

Парадокс «дней рождения» [2], описывающий совпадение элементов одной выборки некоторой дискретной случайной величины, имеющей равновероятное распределение, в его классическом виде [4] был сформулирован Р. Фон Мизесом в 1939 году [1].

В настоящее время в литературе все чаще появляются новые модификации парадокса «дней рождения» [11], что объясняется его широким применением при решении большого класса прикладных задач теории вероятностей, математической статистики и криптографии [7-10], в частности, задач, связанных с анализом хэш-функций и описанием ряда методов дискретного логарифмирования.

Рассмотрим обобщение парадокса «дней рождения» на случай нескольких независимых в совокупности упорядоченных выборок.

Пусть ξ - дискретная случайная величина, имеющая равновероятное распределение на множестве $\{1,2,...,n\}$, $n \in \mathbb{N}$. Пусть далее $B_{n,1}^{(m_1)},...,B_{n,k}^{(m_k)}$ - упорядоченные независимые в совокупности выборки [6] случайной величины ξ объемов $m_1,...,m_k \in \mathbb{N}$, $k \in \mathbb{N}$.

Замечание 1. *В рамках классической постановки парадокса «дней рождения»* [4] *случайная величина ξ имеет равновероятное распределение. В общем случае для произвольного дискретного распределения ξ известные результаты* [3, 11]*, а также результаты настоящей статьи не верны.*

Определим событие $Z_{m_1,...,m_k}^{(n)} = \left\{ \bigcap_{i=1}^{k} B_{n,i}^{(m_i)} \neq \varnothing \right\}$. Таким образом, в рамках решаемой задачи требуется найти вероятность $\mathbf{P}\left\{ Z_{m_1,...,m_k}^{(n)} \right\}$ для произвольных значений n, k и $m_1,...,m_k$.

На рис. 1 изображен пример реализации четырех независимых в совокупности упорядоченных выборок $B_{n,1}^{(m_1)}, B_{n,2}^{(m_2)}, B_{n,3}^{(m_3)}, B_{n,4}^{(m_4)}$ случайной величины ξ объемов $m_1,...,m_4$, для которой справедливо включение множеств

$\{\alpha,\beta\}\subseteq\bigcap_{i=1}^{4}B_{n,i}^{(m_i)}$, где α,β - некоторые элементы множества $\{1,2,...,n\}$. Заметим, что в силу упорядоченности выборок $B_{n,i}^{(m_i)}$, $i=\overline{1,4}$, различные позиции элементов α и β в каждой отдельной выборке соответствуют различным исходам, принадлежащим событию $\left\{\{\alpha,\beta\}\subseteq\bigcap_{i=1}^{4}B_{n,i}^{(m_i)}\right\}$.

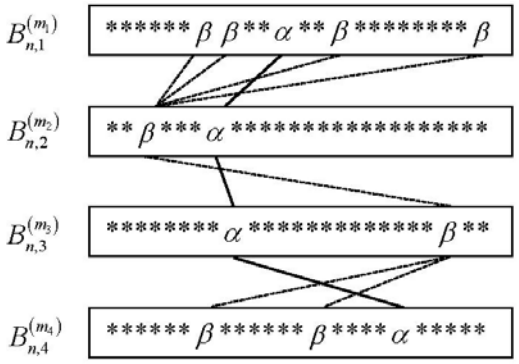

Рис. 1. Появление элементов α,β в упорядоченных выборках $B_{n,1}^{(m_1)},...,B_{n,4}^{(m_4)}$

1. Основные результаты

Через $A_n^{(z)}$, $z\in\overline{1,n}$, обозначим произвольное z-элементное подмножество множества $\{1,2,...,n\}$. Тогда справедлива следующая теорема.

Теорема 1. *Пусть* $n\in\mathbb{N}$, $z\in\overline{1,n}$. *Тогда для произвольного* $A_n^{(z)}$ *и произвольных независимых в совокупности упорядоченных выборок* $B_{n,1}^{(m_1)},...,B_{n,k}^{(m_k)}$, $k\in\mathbb{N}$, *справедливо равенство*

$$\mathbf{P}\left\{A_n^{(z)}\subseteq\bigcap_{i=1}^{k}B_i^{(m_i)}\right\}=\prod_{i=1}^{k}\sum_{s=0}^{z}(-1)^s C_z^s\left(1-\tfrac{s}{n}\right)^{m_i}. \qquad (1)$$

Доказательство. Обозначим через $\alpha_j^{(i)}$ число вхождений произвольного элемента $j\in\{1,2,...,n\}$ в упорядоченную выборку $B_i^{(m_i)}$, $i\in\overline{1,k}$. Тогда для произвольного $A_n^{(z)}$ с использованием формулы «включения-исключения» получаем цепочку соотношений:

$$\mathbf{P}\left\{A_n^{(z)} \subseteq B_i^{(m_i)}\right\} = 1 - \mathbf{P}\left\{A_n^{(z)} \not\subseteq B_i^{(m_i)}\right\} =$$

$$= 1 - \mathbf{P}\left\{B_i^{(m_i)} : \exists \alpha_j^{(i)} = 0, \text{ где } j \in A_n^{(z)}\right\} = 1 - \mathbf{P}\left\{\bigcup_{s=1}^{z}\left\{\alpha_{j_1}^{(i)} = \ldots = \alpha_{j_s}^{(i)} = 0 : \{j_1, \ldots, j_s\} \in A_n^{(z)}\right\}\right\} =$$

$$= 1 + \sum_{s=1}^{z}(-1)^s C_z^s \sum_{\substack{c_1+\ldots+c_{n-s}=m_i \\ c_j \geq 0,\ j=\overline{1,n-s}}} \mathbf{P}\left\{\alpha_1^{(i)} = c_1, \ldots, \alpha_n^{(i)} = c_n\right\} = 1 + \frac{1}{n^{m_i}} \sum_{s=1}^{z}(-1)^s C_z^s \sum_{\substack{c_1+\ldots+c_{n-s}=m_i \\ c_j \geq 0,\ j=\overline{1,n-s}}} \frac{m_i!}{c_1! \cdot \ldots \cdot c_{n-s}!}.$$

Далее, учитывая равенство $\sum_{\substack{c_1+\ldots+c_{n-s}=m_i \\ c_j \geq 0,\ j=\overline{1,n-s}}} \frac{m_i!}{c_1! \cdot \ldots \cdot c_{n-s}!} = (n-s)^{m_i}$ для суммы полиномиальных коэффициентов [6], получаем:

$$\mathbf{P}\left\{A_n^{(z)} \subseteq B_i^{(m_i)}\right\} = 1 + \frac{1}{n^{m_i}} \sum_{s=1}^{z}(-1)^s C_z^s (n-s)^{m_i} = \sum_{s=0}^{z}(-1)^s C_z^s \left(1 - \frac{s}{n}\right)^{m_i}.$$

Из условия независимости в совокупности выборок $B_1^{(m_1)}, \ldots, B_k^{(m_k)}$ следует, что для произвольного $k \in \mathbb{N}$

$$\mathbf{P}\left\{A_n^{(z)} \subseteq \bigcap_{i=1}^{k} B_i^{(m_i)}\right\} = \prod_{i=1}^{k} \mathbf{P}\left\{A_n^{(z)} \subseteq B^{(m_i)}\right\} = \prod_{i=1}^{k} \sum_{s=0}^{z}(-1)^s C_z^s \left(1 - \frac{s}{n}\right)^{m_i}. \ \square$$

Следствие 1. *Если в условиях теоремы 1 выполняются соотношения* $m_i = \alpha n$, $i = \overline{1,k}$, *то справедливо равенство*

$$\lim_{n\to\infty} \mathbf{P}\left\{A_n^{(z)} \subseteq \bigcap_{i=1}^{k} B_{n,i}^{(m_i)}\right\} = \left(1 - e^{-\alpha}\right)^{zk}.$$

Доказательство. Рассмотрим при фиксированном $z \in \overline{1,n}$ выражение, стоящее под знаком произведения в (1), которое в случае $m_i = \alpha n$ при $n \to \infty$ представляет собой бином Ньютона. Действительно, используя равенство $\lim_{n\to\infty}\left(1 - \frac{1}{x}\right)^x = e^{-1}$ [5], получаем:

$$\sum_{s=0}^{z}(-1)^s C_z^s \left(1 - \frac{s}{n}\right)^{m_i} = \sum_{s=0}^{z}(-1)^s C_z^s \left(1 - \frac{s}{n}\right)^{\alpha n} \underset{n\to\infty}{=} \sum_{s=0}^{z}(-1)^s C_z^s e^{-\alpha s} = \left(1 - e^{-\alpha}\right)^z.$$

Тогда из соотношения (1) следует искомое равенство. $\square$

Теорема 2. *Для произвольных* $n,k \in \mathbb{N}$ *и* $m_1,...,m_k \geq 1$ *справедливо равенство*

$$\mathbf{P}\left\{Z^{(n)}_{m_1,...,m_k}\right\} = \frac{m_1!}{n^{m_1}} \sum_{s=1}^{n} \sum_{\substack{c_1+...+c_s=m_1 \\ c_i \geq 1,\ i=\overline{1,s}}} \sum_{j=1}^{s} \prod_{i=2}^{k} \sum_{u=0}^{j} (-1)^{u+j-1} \frac{C_n^s \cdot C_s^j \cdot C_j^u}{c_1! \cdot ... \cdot c_s!} \left(1-\frac{u}{n}\right)^{m_i}. \quad (2)$$

Доказательство. Не ограничивая общности, рассмотрим упорядоченную выборку $B^{(m_1)}_{n,1}$. Разделим множество всех упорядоченных разбиений $c_1 + ... + c_n = m_1$, где $c_i \geq 0$, $i = \overline{1,n}$, на непересекающиеся подмножества U_s, $s = \overline{1,n}$, по количеству ненулевых элементов в указанном разбиении:

$$U_s = \left\{ (c_1,...,c_n) \left| \begin{array}{c} c_1 + ... + c_n = m_1,\ c_{j_1} \geq 1,...,c_{j_s} \geq 1, \\ \{j_1,...,j_s\} \in \{1,2,...,n\},\ c_i = 0,\ i \notin \{j_1,...,j_s\} \end{array} \right. \right\}.$$

Определим далее события $\widetilde{U}_s$, $s = \overline{1,n}$:

$$\widetilde{U}_s = \left\{ (c_1,...,c_n) \left| \begin{array}{c} c_1 + ... + c_n = m_1,\ c_1 \geq 1,...,c_s \geq 1, \\ c_i = 0,\ i = \overline{s+1,n} \end{array} \right. \right\}.$$

Тогда по формуле полной вероятности получаем цепочку равенств:

$$\mathbf{P}\left\{Z^{(n)}_{m_1,...,m_k}\right\} = \sum_{s=1}^{n} \mathbf{P}\left\{Z^{(n)}_{m_1,...,m_k}, U_s\right\} = \sum_{s=1}^{n} C_n^s \mathbf{P}\left\{Z^{(n)}_{m_1,...,m_k}, \widetilde{U}_s\right\} = \sum_{s=1}^{n} C_n^s \mathbf{P}\left\{Z^{(n)}_{m_1,...,m_k} \middle/ \widetilde{U}_s\right\} \mathbf{P}\left\{\widetilde{U}_s\right\}. \quad (3)$$

Рассмотрим отдельно величину $\mathbf{P}\left\{\widetilde{U}_s\right\}$ в выражении (3).

$$\mathbf{P}\left\{\widetilde{U}_s\right\} = \frac{1}{n^{m_1}} \sum_{\substack{c_1+...+c_s=m_1 \\ c_i \geq 1,\ i=\overline{1,s}}} C_{m_1}^{c_1} C_{m_1-c_1}^{c_2} \cdot ... \cdot C_{m_1-c_1-...-c_{s-1}}^{c_s} = \frac{m_1!}{n^{m_1}} \sum_{\substack{c_1+...+c_s=m_1 \\ c_i \geq 1,\ i=\overline{1,s}}} \frac{1}{c_1! \cdot ... \cdot c_s!}. \quad (4)$$

Далее рассмотрим условную вероятность $\mathbf{P}\left\{Z^{(n)}_{m_1,...,m_k} \middle/ \widetilde{U}_s\right\}$ в выражении (3). Событие $\widetilde{U}_s$ означает, что в выборке $B^{(m_1)}_{n,1}$ реализованы элементы $1,2,...,s$. Таким образом, при реализации $\widetilde{U}_s$ выполняется равенство:

$$\mathbf{P}\left\{Z^{(n)}_{m_1,...,m_k} \middle/ \widetilde{U}_s\right\} = \mathbf{P}\left\{\bigcup_{j=1}^{s} \left\{ j \subseteq \bigcap_{i=2}^{k} B^{(m_i)}_{n,i} \right\}\right\},$$

Тогда с учетом результата теоремы 1 по формуле «включения-исключения» имеем:

$$\mathbf{P}\left\{Z^{(n)}_{m_1,\dots,m_k}\middle/\widetilde{U}_s\right\}=\sum_{j_1=1}^{s}\mathbf{P}\left\{j_1\subseteq\bigcap_{i=2}^{k}B^{(m_i)}_{n,i}\right\}-C_s^2\sum_{1\le j_1<j_2\le s}\mathbf{P}\left\{\{j_1,j_2\}\subseteq\bigcap_{i=2}^{k}B^{(m_i)}_{n,i}\right\}+\dots+$$

$$+(-1)^{s-1}C_s^s\mathbf{P}\left\{\{1,\dots,s\}\subseteq\bigcap_{i=2}^{k}B^{(m_i)}_{n,i}\right\}=\sum_{j=1}^{s}(-1)^{j-1}C_s^j\mathbf{P}\left\{A^{(j)}_n\subseteq\bigcap_{i=2}^{k}B^{(m_i)}_{n,i}\right\}= \qquad (5)$$

$$=\sum_{j=1}^{s}(-1)^{j-1}C_s^j\prod_{i=2}^{k}\sum_{u=0}^{j}(-1)^u C_j^u\left(1-\tfrac{u}{n}\right)^{m_i}.$$

В итоге, подставляя выражения (4), (5) в (3), получаем искомое равенство. □

Как и в случае классического парадокса «дней рождения» [4] интерес представляет собой задача определения минимального значения m, при котором вероятность $\mathbf{P}\left\{Z^{(n)}_{m,\dots,m}\right\}>\frac{1}{2}$. В таблице 1 представлены соответствующие значения m при некоторых $k\in\mathbb{N}$ и $n=365$.

k	**1**	**2**	**3**	**4**	**5**	**6**
m	23	16	48	86	123	158
$\mathbf{P}\left\{Z^{(n)}_{m,\dots,m}\right\}$	0.507	0.504	0.503	0.514	0.509	0.502

Таб. 1. Минимальное значение m, при котором $\mathbf{P}\left\{Z^{(n)}_{m,\dots,m}\right\}>\frac{1}{2}$

Следствие 2. *Если в условиях теоремы 2 выполняются соотношения* $m_i=\alpha n$, $i=\overline{1,k}$, *то при* $n\to\infty$ *справедливо равенство*

$$\mathbf{P}\left\{Z^{(n)}_{\underbrace{m,\dots,m}_{k}}\right\}=1-\tfrac{m!}{n^m}\sum_{s=1}^{n}\sum_{\substack{c_1+\dots+c_s=m\\ c_i\ge1,\ i=\overline{1,s}}}\tfrac{C_n^s}{c_1!\cdot\dots\cdot c_s!}\left(1-\left(1-e^{-\alpha}\right)^{(k-1)}\right)^s. \qquad (5)$$

Доказательство. С учетом результата следствия 1 имеем цепочку равенств при $n\to\infty$:

$$\mathbf{P}\left\{Z^{(n)}_{\underbrace{m,\dots,m}_{k}}\right\}=\tfrac{m!}{n^m}\sum_{s=1}^{n}\sum_{\substack{c_1+\dots+c_s=m\\ c_i\ge1,\ i=\overline{1,s}}}\tfrac{C_n^s}{c_1!\cdot\dots\cdot c_s!}\sum_{j=1}^{s}(-1)^{j-1}C_s^j\prod_{i=2}^{k}\sum_{u=0}^{j}(-1)^u C_j^u\left(1-\tfrac{u}{n}\right)^{m_i}=$$

$$=\tfrac{m!}{n^m}\sum_{s=1}^{n}\sum_{\substack{c_1+\dots+c_s=m\\ c_i\ge1,\ i=\overline{1,s}}}\tfrac{C_n^s}{c_1!\cdot\dots\cdot c_s!}\sum_{j=1}^{s}(-1)^{j-1}C_s^j\left(1-e^{-\alpha}\right)^{j(k-1)}=$$

$$=\tfrac{m!}{n^m}\sum_{s=1}^{n}\sum_{\substack{c_1+\dots+c_s=m\\ c_i\ge1,\ i=\overline{1,s}}}\tfrac{C_n^s}{c_1!\cdot\dots\cdot c_s!}(-1)\left(\left(1-\left(1-e^{-\alpha}\right)^{(k-1)}\right)^s-1\right)=$$

$$= \frac{m!}{n^m} \sum_{s=1}^{n} \sum_{\substack{c_1+\ldots+c_s=m \\ c_i \geq 1,\, i=\overline{1,s}}} \frac{C_n^s}{c_1!\cdot\ldots\cdot c_s!} \left(1 - \left(1 - \left(1 - e^{-\alpha}\right)^{(k-1)}\right)^s\right) =$$

$$= 1 - \frac{m!}{n^m} \sum_{s=1}^{n} \sum_{\substack{c_1+\ldots+c_s=m \\ c_i \geq 1,\, i=\overline{1,s}}} \frac{C_n^s}{c_1!\cdot\ldots\cdot c_s!} \left(1 - \left(1 - e^{-\alpha}\right)^{(k-1)}\right)^s . \ \square$$

В случаях, когда число независимых в совокупности упорядоченных выборок случайной величины ξ не велико, а объем каждой из них существенно превосходит n, справедлив следующий результат.

Следствие 3. *Если в условиях теоремы 2 выполняются соотношения* $m_i = \alpha n$, $i = \overline{1,k}$, *и* $k \leq 1 + e^{\alpha}$, *то при* $n \to \infty$ *справедлива оценка снизу:*

$$\mathbf{P}\left\{ Z^{(n)}_{\underbrace{m,\ldots,m}_{k}} \right\} \geq 1 - (k-1)e^{-\alpha} .$$

Доказательство. Для оценки выражения, стоящего под знаком суммы в (5) воспользуемся двусторонним неравенством [5]:

$$1 - rx \leq (1-x)^r \leq 1 - rx + C_r^2 x^2 , \tag{6}$$

справедливом при $-1 < x < 1$, где $r \in \mathbb{N}$.

Левое неравенство в (6) является нетривиальным при $1 - rx \geq 0$, что, в свою очередь, выполняется в условиях настоящего следствия при $r = k-1$ и $x = e^{-\alpha}$. Тогда

$$\left(1 - \left(1 - e^{-\alpha}\right)^{k-1}\right)^s \leq \left(1 - \left(1 - (k-1)e^{-\alpha}\right)\right)^s = \left((k-1)e^{-\alpha}\right)^s = (k-1)^s e^{-\alpha s} .$$

Учитывая, что с ростом s величина $(k-1)^s e^{-\alpha s}$ убывает, получаем цепочку неравенств при $n \to \infty$:

$$\mathbf{P}\left\{ Z^{(n)}_{\underbrace{m,\ldots,m}_{k}} \right\} \geq 1 - \frac{m!}{n^m} \sum_{s=1}^{n} \sum_{\substack{c_1+\ldots+c_s=m \\ c_i \geq 1,\, i=\overline{1,s}}} \frac{C_n^s}{c_1!\cdot\ldots\cdot c_s!} (k-1)^s e^{-\alpha s} \geq 1 - \frac{1}{n^m} \sum_{s=1}^{n} \sum_{\substack{c_1+\ldots+c_s=m \\ c_i \geq 1,\, i=\overline{1,s}}} \binom{n}{s} \frac{m!}{\alpha_1!\cdot\ldots\cdot\alpha_s!} (k-1)e^{-\alpha} =$$

$$= 1 - (k-1)e^{-\alpha} \frac{1}{n^m} \sum_{s=1}^{n} \sum_{\substack{c_1+\ldots+c_s=m \\ c_i \geq 1,\, i=\overline{1,s}}} \binom{n}{s} \frac{m!}{\alpha_1!\cdot\ldots\cdot\alpha_s!} = 1 - (k-1)e^{-\alpha} . \ \square$$

Заключение

Полученные в статье результаты могут найти практическое применение при решении ряда задач современной криптографии, например, при оценке эффективности параллельной реализации методов определения мультиколлизий хэш-функций, а также методов дискретного логарифмирования, подобных ρ-методу Полларда.

Список литературы

1. *Крилли Т.* 50 идей, о которых нужно знать. Математика. - М.: Фантом Пресс, 2014, 208 с.
2. *Погорелов Б.А., Сачков В.Н.* Словарь криптографических терминов. – М.:МЦНМО, 2006. – 94 с.
3. *Сачков В. Н.* Курс комбинаторного анализа. - Ижевск: НИЦ «Регулярная и хаотическая динамика», 2013, 336 с.
4. *Секей Г.* Парадоксы в теории вероятностей и математической статистике. - М.: МИР, 1990, 240 с.
5. *Фихтенгольц Г.М.* Курс дифференциального и интегрального исчисления. Том 1 (5-е изд.). – М.: ФИЗМАТЛИТ, 1962.
6. *Ширяев А.Н.* Вероятность-1 (6-е изд.). – М.: МЦНМО, 2017.
7. *Bellare M., Micciancio D.* A new paradigm for collision-free hashing: incrementality at reduced cost. - EUROCRYPT'97, Lect. Notes Comp. Sci., v. 1233 (1997), p. 163-192.
8. *Kelsey J., Schneier B., Wagner D., Hall C.* Cryptanalytic attacks on pseudorandom number generators. - FSE'98, Lect. Notes Comp. Sci., v. 1372 (1998), p. 168-188.
9. *Neuenschwander D.* Probabilistic and Statistical Methods in Cryptology (An Introduction by Selected Topics). - Berlin, Heidelberg: Springer-Verlag, 2004, LNCS 3028, 158 p.
10. *Preenel B.* Analysis and Design of Cryptographic Hash Functions. PhD, 2003, 324 p.
11. *Wagner D.* A generalized birthday problem. CRYPTO 2002, Lect. Notes Comp. Sci., v. 2442 (2002), p. 288-303.