УДК 519.2

**П.А. ЕНДОВИЦКИЙ**

# УТОЧНЕНИЕ АСИМПТОТИЧЕСКОЙ АППРОКСИМАЦИИ РАЗМЕРА ГРУППЫ В ПАРАДОКСЕ ДНЕЙ РОЖДЕНИЙ

**Ключевые слова:** *случайное размещение, парадокс дней рождений, асимптотические неравенства, формула Стирлинга для гамма-функции, метод коллизий для хэш-функций*.

Парадокс дней рождений состоит в следующем [1, 2]. Пусть существует группа из $n$ людей ($n \le 365$). Будем считать, что день рождения каждого человека из группы приходится с равной вероятностью на один из 365 дней. Тогда вероятность того, что в группе найдутся хотя бы два человека, дни рождений которых совпадают, равна

$$P(n) = 1 - \frac{(365)_n}{365^n} = 1 - \frac{365 \cdot 364 \cdot \ldots \cdot (365 - n + 1)}{365^n}.$$

Размер группы, где с вероятностью $p = 1$ найдутся по крайней мере два человека с совпадающими днями рождения, равен 366, а для $p < 1$ он гораздо меньше: $P(23) = 0{,}507\ldots$, $P(57) = 0{,}990\ldots$ и т.д. (т.е. в группе из 23 человек с вероятностью, большей $\frac{1}{2}$, найдутся два человека с совпадающими днями рождения). Парадокс состоит в кажущемся противоречии с малым размером группы при заданном $p \in (0; 1)$. Подробнее см. в [2].

Оценки вероятностей и размера группы в задаче о днях рождения имеют важные применения при аутентификации, построении хэш-функций, в кодировании, в криптоанализе, при решении систем уравнений над конечными алгебраическими структурами.

**1.** Сформулируем задачу о размере группы в парадоксе дней рождений в общем случае (в терминах размещения частиц по ячейкам). Пусть заданы числа $m \in N$, $p \in (0; 1)$, $m$ — число ячеек, в парадоксе $m = 365$. Пусть также $P_m(n)$ — вероятность, что при размещении $n$ частиц по $m$ ячейкам по крайней мере две частицы попадут в одну ячейку (считаем, что каждая частица размещается независимо от других и равновероятно по $m$ ячейкам):

$$P_m(n) = 1 - \frac{(m)_n}{m^n} = 1 - \frac{m \cdot (m-1) \cdot \ldots \cdot (m - n + 1)}{m^n} \equiv 1 - Q_m(n), \quad n = \overline{0, m+1}, \tag{1}$$

где $Q_m(n) = 1 - P_m(n)$ — вероятность, что все частицы попадут в разные ячейки, тогда $P(n)\uparrow$ возрастает по $n$ при фиксированном $m$, $P(0) = P(1) = 0$, $P(m+1) = 1$ (для упрощения записи вместо $P_m(n)$ будем писать $P(n)$).

Определим натуральное число $n = n(m) \in [1; m+1]$, которое зависит от $m$ и $p$, из следующего условия:

$$P(n-1) < p \le P(n), \tag{2}$$

т.е. $n(m)$ — минимальное число $n$ частиц такое, что при размещении $n$ частиц в $m$ ячейках вероятность попадания по крайней мере двух частиц в одну ячейку не меньше $p$. Например, если $p = \frac{1}{2}$, $m = 365$, то $n(365) = 23$, так как $P(22) < \frac{1}{2} < P(23)$.

Рассмотрим вопрос об асимптотическом поведении $n(m)$. Хорошо известен следующий результат (см., например, [2]):

$$n(m) = \sqrt{2am} + o(\sqrt{m}), \quad m \to \infty,$$

где $a = -\ln(1-p) > 0$.

Цель данной статьи — доказать теорему 1, которая существенно усиливает этот результат.

**Теорема 1.** Пусть $p \in (0; 1)$ фиксировано, $a = -\ln(1-p)$, тогда $n(m) = \sqrt{2am} + \gamma(m)$, где последовательность $\{\gamma(m), m \ge 1\}$ ограничена и $\underline{\lim}_{m\to\infty} \gamma(m) = \frac{1}{2} - \frac{a}{3}$, $\overline{\lim}_{m\to\infty} \gamma(m) = \frac{3}{2} - \frac{a}{3}$.

**Доказательство.** Для того чтобы воспользоваться известными результатами о свойствах гамма-функции $\Gamma(x)$ [3], будем считать, что числа $n$, $m$ из (1) могут меняться непрерывно. Пусть

$$Q_y(x) = \frac{\Gamma(y+1)}{\Gamma(y-x+1)\,y^x}, \quad y > 0, \ 0 \le x < y+1. \tag{3}$$

Очевидно, что при натуральных $x$, $y$ в (3) получаем $Q_m(n)$ из формулы (1).

Определим характер монотонности $Q_y(x)$ при фиксированном $y$

$$Q_y(x) = \frac{\Gamma(y+1)}{\Gamma(y-x+1)\,y^x} = \left| \begin{matrix} t = y-x+1 \\ 0 < t \le y+1 \end{matrix} \right| = \frac{\Gamma(y+1)}{\Gamma(t)\,y^{y+1-t}} = \frac{\Gamma(y+1)}{f(t)\,y^{y+1}},$$

где

$$f(t) = \frac{\Gamma(t)}{y^t} = \frac{1}{y^t}\int_0^\infty u^{t-1}e^{-u}\,du = \int_0^\infty v^{t-1}e^{-vy}\,dv.$$

При этом $f'(t) = \int_0^\infty v^{t-1}e^{-vy}\ln v\,dv$, $f'(t)\uparrow$, так как $f''(t) = \int_0^\infty v^{t-1}e^{-vy}\ln^2 v\,dv > 0$, дифференцирование под знаком интеграла возможно по признаку Лейбница [3], также $f(y) = \frac{\Gamma(y)}{y^y} = \frac{\Gamma(y+1)}{y^{y+1}} = f(y+1)$, т.е. по теореме Ролля $\exists\, y_0 \in (y,\, y+1) : f'(y_0) = 0$.

Отсюда и из возрастания $f'(t)$ следует, что $f(t)$ при $t \in (0;\, y_0]$ убывает и при $t \in (y_0;\, y+1]$ возрастает. Отсюда $\exists z_0 \in (0,1)$ такое, что $Q_y(\cdot)$ при $x \in [0;\, z_0]$ возрастает и при $x \in [z_0;\, y+1]$ убывает, т.е. $Q_y(\cdot)$ убывает при $x \in [1;\, y+1]$ (так как $z_0 < 1$), но $Q_y(1) = 1$, $Q_y(y+1-0) = 0$. Значит, ввиду непрерывности $Q_y(\cdot)$ для заданного $p \in (0;1)$ $\exists!\, x(y) \in (1;\, y+1)$ такое, что $Q_y(x(y)) = 1-p$.

Функция $x(y)$ есть непрерывное распространение последовательности $\{n(m),\, m \ge 1\}$ из (2) для действительных $x$, $y > 0$.

Исследуем асимптотическое поведение $x(y)$ при $y \to +\infty$ (это в дальнейшем даст возможность определить асимптотическое поведение $n(m)$, $m \ge 1$). Для этого докажем три леммы и введем обозначение $x_1(y) = x(y) - 1$ для упрощения записи некоторых формул.

**Лемма 1.** Функция $x(y) \to +\infty$ при $y \to +\infty$.

**Доказательство.** Покажем, что $x(y)$ возрастает. Для этого достаточно убедиться, что $\forall b > 0$ функция $g(y) = \frac{\Gamma(y)}{\Gamma(y-b)\,y^b}$ возрастает при $y > b$. Покажем, что $[\ln g(y)]' > 0$, т.е. $\ln g(y)$ возрастает

$$[\ln g(y)]' = [\ln\Gamma(y) - \ln\Gamma(y-b) - b\ln y]' = \frac{\Gamma'(y)}{\Gamma(y)} - \frac{\Gamma'(y-b)}{\Gamma(y-b)} - \frac{b}{y},$$

но $\frac{\Gamma'(a)}{\Gamma(a)} = -\frac{1}{a} - C + \sum_{n=1}^{\infty}\left(\frac{1}{n} - \frac{1}{n+a}\right)$, где $C$ — константа Эйлера [3].

Отсюда

$$[\ln g(y)]' = -\frac{1}{y} + \frac{1}{y-b} - \frac{b}{y} + \sum_{n=1}^{\infty}\left(\frac{1}{n} - \frac{1}{n+y}\right) - \sum_{n=1}^{\infty}\left(\frac{1}{n} - \frac{1}{n+y-b}\right) =$$

$$= \sum_{n=0}^{\infty}\frac{b}{(n+c)(n+b+c)} - \frac{b}{b+c},$$

где $c = y - b > 0$. Но

$$\sum_{n=0}^{\infty}\frac{b}{(n+c)(n+b+c)} > b\int_0^\infty \frac{dx}{(x+c)(x+b+c)} = \ln\left(1+\frac{b}{c}\right) > \frac{\frac{b}{c}}{1+\frac{b}{c}} = \frac{b}{b+c},$$

поскольку $\ln(1+t) > \frac{t}{1+t}$ при $t > 0$. Отсюда $[\ln g(y)]' > 0$, т.е. $g(y)$ и $x(y)$ возрастают.

Если теперь $\lim\limits_{y\to+\infty} x(y) = M \in R$, то для $N = [M] + 1$ имеем $x(y) < N$ $\forall y > 0$, т.е.

$$1 - p = Q_y(x) = \frac{\Gamma(y)}{\Gamma(y-x_1)\,y^{x_1}} \ge \frac{\Gamma(y)}{\Gamma(y-N)\,y^N} = \frac{(y-1)\ldots(y-N)}{y^N} \to 1, \quad y \to +\infty.$$

Получили противоречие, т.е. $\lim\limits_{y\to+\infty} x(y) = +\infty$.

Лемма доказана.

**Лемма 2.** Функция $\frac{x(y)}{y} \to 0$ при $y \to +\infty$.

**Доказательство.** Если $\overline{\lim\limits_{y\to+\infty}} \frac{x(y)}{y} > 0$, то $\exists \beta > 0 \ \ \forall M > 0 \ \ \exists y > M : x_1(y) > \beta y$.

В этом случае

$$Q_y(x) = \frac{\Gamma(y)}{\Gamma(y-x_1)y^{x_1}} < \frac{\Gamma(y)}{\Gamma(\alpha y)y^{\beta y}} = Q_y(\beta y), \ \alpha = 1-\beta.$$

Покажем, что $Q_y(\beta y) \to 0$ при $y \to +\infty$. По формуле Стирлинга для Г-функции [3] имеем:

$$-\ln Q_y(\beta y) = -\ln \frac{\Gamma(y)}{\Gamma(\alpha y)y^{\beta y}} = -\ln\Gamma(y) + \ln\Gamma(\alpha y) + \beta y \ln y =$$

$$= -\left(y - \frac{1}{2}\right)\ln y + y + \left(\alpha y - \frac{1}{2}\right)\ln \alpha y - \alpha y + \beta y \ln y + o(1) =$$

$$= y(1 - \alpha + \alpha \ln \alpha) - \frac{1}{2}\ln \alpha + o(1) \to +\infty, \ y \to +\infty,$$

так как $1 - \alpha + \alpha \ln \alpha > 0$ при $\alpha \in (0; 1)$. Отсюда $1 - p = Q_y(x) < Q_y(\beta y) \to 0, \ y \to +\infty$.

Получили противоречие, т.е. $\lim\limits_{y\to+\infty} \frac{x(y)}{y} = 0$.

Лемма доказана.

**Лемма 3.** Функция $x(y) = \sqrt{2ay} + \frac{1}{2} - \frac{a}{3} + o(1)$ при $y \to +\infty$.

**Доказательство.** Пусть $a = -\ln(1-p) > 0$, тогда по формуле Стирлинга для Г-функции имеем

$$a = -\ln Q_y(x) = -\ln\Gamma(y) + \ln\Gamma(y - x_1) + x_1 \ln y =$$

$$= -(y - x_1)\ln \frac{y}{y - x_1} + x_1 + \frac{1}{2}\ln \frac{y}{y - x_1} - \frac{\theta_1}{12y} + \frac{\theta_2}{12(y - x_1)} =$$

$$= x_1\left(1 + \frac{\ln(1-u)}{u} - \ln(1-u)\right) - \frac{1}{2}\ln(1-u) - \frac{\theta_1 u}{12x_1} + \frac{\theta_2 u}{12x_1(1-u)},$$

где $u = u(y) = \frac{x_1(y)}{y} \to 0, \ y \to +\infty, \ \theta_1, \theta_2 \in (0; 1)$.

Отсюда

$$x(y) = \frac{2a}{u(y)} + 2\left(\frac{1}{2} - \frac{a}{3}\right) + o(1) \quad y \to +\infty,$$

и, следовательно,

$$x(y) = \sqrt{2ay} + \frac{1}{2} - \frac{a}{3} + o(1), \ y \to +\infty.$$

Лемма доказана.

Теперь закончим доказательство теоремы 1. Из леммы 3 имеем, что

$$x(m) \to +\infty, \ m \to +\infty,$$

$$x(m+1) - x(m) \to 0, \ m \to +\infty,$$

отсюда следует, что дробные части $\{x(m)\}$, $m \in N$, всюду плотны в $[0; 1]$.

Из условия (2) также следует, что $n(m) - 1 < x(m) \le n(m)$. Отсюда и из леммы 3 получим, что для $\gamma(m) = n(m) - \sqrt{2am}$ будет выполняться $\overline{\{\gamma(m), m \ge 1\}} = \left[\frac{1}{2} - \frac{a}{3}; \frac{3}{2} - \frac{a}{3}\right]$ (черта над множеством означает замыкание).

В частности, $\underline{\lim\limits_{m\to\infty}} \gamma(m) = \frac{1}{2} - \frac{a}{3}$, $\overline{\lim\limits_{m\to\infty}} \gamma(m) = \frac{3}{2} - \frac{a}{3}$.

Теорема доказана.

Теорема 1 дает асимптотически точное выражение для $n(m)$, так как $\overline{\lim}\,\gamma(m) - \underline{\lim}\,\gamma(m) = 1$, т.е. отрезок $\left(\sqrt{2am} \underline{\lim\limits_{m\to\infty}} \gamma(m); \sqrt{2am} + \overline{\lim\limits_{m\to\infty}} \gamma(m)\right)$ содержит не более чем одну целую точку $n_0(m)$, которую можно рассматривать как приближение для $n(m)$ при больших $m$.

**Пример.** Если $m = 2^{64}$, $p = \frac{1}{2}$, то $n_0(m) = 5056937541$, при этом

$$P(n_0) = 0{,}500000000006094 \ldots ,$$

$$P(n_0 - 1) = 0{,}499999999869026 \ldots .$$

**2.** Рассмотрим следующий вариант парадокса дней рождений [1]: в $m$ ячейках независимо размещается две группы по $n$ частиц, причем в каждой группе частицы попадают в различные ячейки и все $C_m^n$ возможных размещений одной группы равновероятны. Пусть $R_m(n)$ — вероятность того, что при размещении двух групп по $n$ частиц по $m$ ячейкам хотя бы две частицы из разных групп попадут в одну ячейку:

$$R_m(n) = 1 - \frac{C_{m-n}^n}{C_m^n} = 1 - \frac{(m-n)!^2}{m!(m-2n)!} \equiv 1 - S_m(n), \ n = \overline{0, k}, \ k = \left[\frac{m}{2}\right],$$

где $S_m(n) = 1 - R_m(n)$ — вероятность, что все частицы попадут в разные ячейки. При $n > \left[\frac{m}{2}\right]$ имеем $R_m(n) = 1$.

Пусть задано число $p \in (0; 1)$, определим натуральное число $n = n(m) \in [1; k+1]$, которое зависит от $m$ и $p$, из условия $R(n-1) < p \le R(n)$, т.е. $n(m)$ — минимальное число $n$ частиц такое, что при размещении двух групп по $n$ частиц в $m$ ячейках вероятность попадания двух частиц в одну ячейку не меньше $p$.

Задача состоит в определении асимптотического поведения для $n(m)$. Справедлива следующая теорема.

**Теорема 2.** Пусть $p \in (0; 1)$ фиксировано, $a = -\ln(1-p)$, тогда $n(m) = \sqrt{am} + \gamma(m)$, где последовательность $\{\gamma(m), m \ge 1\}$ ограничена и $\underline{\lim\limits_{m\to\infty}} \gamma(m) = -\frac{a}{2}$, $\overline{\lim\limits_{m\to\infty}} \gamma(m) = -\frac{a}{2} + 1$.

**Доказательство**. Доказательство теоремы 2 аналогично доказательству теоремы 1. Введем функцию

$$S_y(x) = \frac{\Gamma(y-x+1)^2}{\Gamma(y+1)\Gamma(y-2x+1)} = \frac{\Gamma(y_1-x)^2}{\Gamma(y_1)\Gamma(y_1-2x)}, \ y_1 = y+1 > 1, \ 0 \le x < \frac{y}{2}.$$

Определим функцию $x(y)\colon (0; +\infty) \to (0; +\infty)$ из соотношения $S_y(x(y)) = 1 - p$. Тогда

$$n(m) - 1 < x(m) \le n(m), \ m \in N, \qquad x(y) = \sqrt{ay} - \frac{a}{2} + o(1) \text{ при } y \to +\infty .$$

Отсюда следует утверждение теоремы 2.

Эти результаты можно применить, например, для оценивания вероятности коллизий хэш-функций и трудоемкости построения коллизий [4], в криптоанализе, в теории случайных размещений и случайных отображений, при оценке совпадения редких событий [5, 6].

СПИСОК ЛИТЕРАТУРЫ


1. Ширяев А.Н. Вероятность. — М.: Наука, 1980. — 576 с.
2. Секей Г. Парадоксы в теории вероятностей и мат. статистике. — М.: Мир, 1990. — 240 с.
3. Фихтенгольц Г.М. Курс дифференциального и интегрального исчисления. Т.2. — М.: Наука, 1966. — 800 с.
4. Алферов А.П., Зубов А.Ю., Кузьмин А.С., Черемушкин А.В. Основы криптографии. — М.: Гелиос АРВ, 2001. — 480 с.
5. Cooper C., Gilchrist R., Kovalenko I.N., Novacovic D. Deriving the number of good permutstions with applications to cryptography // Кибернетика и системный анализ. — 1999. — № 5. — С. 10–16.
6. Гилкрист Р., Коваленко И.Н. Об оценке вероятности отсутствия коллизий некоторых случайных отображений // Там же. — 2000. — № 1. — С. 132–138.


*Поступила 17.11.2009*