УДК 519.2
DOI: 10.20535/1810-0546.2016.4.76267

П.О. Єндовицький
Національний технічний університет України “КПІ”, Київ, Україна

ДВІ МОДИФІКАЦІЇ ЗАДАЧІ ПРО ДНІ НАРОДЖЕННЯ: НЕРІВНОМІРНИЙ ВИПАДОК

Background. The scheme of random allocation of particles in cells is studied both in probability theory and mathematical statistics. In probability theory usual study is concerning limit theorems, in mathematical statistics – construction statistical criteria's. Birthday problem is one of main questions in this theory.
Objective. Two modifications of the birthday problem are considered in the paper. One was formulated in Fermi statistic scheme, another – in non-uniform and independent random allocation scheme. In both cases the objective was to solve a birthday problem.
Methods. Standard asymptotical methods were used. At first we needed to prove one limit theorem and to estimate rapidity of convergence in it. Using these results numerical calculation of probabilities from birthday problem was made. Also formulas for the group size from birthday problem were obtained.
Results. As a result numerical estimates for birthday problem probability and group size were obtained.
Conclusion. For both modifications main asymptotic values coincide, as in the formula for probability calculation, as in the formula for the group size. But second terms from their asymptotic series are already different.
Keywords: birthday problem; birthday paradox; random allocations; Fermi statistic; Uval attack.

Вступ

Задача про дні народження є класичною в теорії ймовірностей [1]. У цій задачі потрібно знайти ймовірність того, що принаймні двоє осіб з групи, що складається з n осіб, народилися в один день. Будемо називати таку задачу прямою задачею про дні народження – у прямій задачі потрібно знайти ймовірність. Цю задачу можна розв'язувати як у припущенні, що всі можливі дні народження є рівноймовірними, так і без цього припущення [2].

Також задачею про дні народження називають [3, 4] задачу знаходження мінімального n такого, що в групі з n осіб знайдуться з ймовірністю не менше ніж задане число $p \in (0,1)$ принаймні двоє осіб, які народилися в один день. Цю задачу будемо називати оберненою задачею про дні народження.

У цій статі буде розглянуто дві модифікації класичної задачі (КЗ) про дні народження. Обидві модифіковані задачі будуть формулюватися в термінах розміщення частинок по комірках. Наприклад, подія “двоє осіб народилися в один день” буде формулюватися так: “дві частинки потрапили в одну комірку”.

В обох модифікаціях КЗ, які розглядатимуться далі, в комірках розміщуються частинки двох типів (а не одного, як у класичній задачі) і шукається ймовірність події, що існує комірка, яка містить частинки обох типів.

Різниця між цими двома модифікаціями КЗ полягає в тому, що у першій модифікації, яку будемо називати статистикою Фермі, частинки одного типу розміщуються в комірках по одній – тобто тут відсутні комірки, які містять дві і більше частинок одного типу. При цьому потрапляння частинок різних типів у одну комірку дозволяється. У другій модифікації, яку будемо називати статистикою Ювала, частинки одного типу можуть потрапляти в одну комірку.

Природними видаються такі міркування. При розміщенні частинок двох типів за статистикою Фермі ці частинки будуть займати “більше” комірок, ніж відповідні частинки зі статистики Ювала. Отже, для статистики Фермі ймовірність F існування комірки, що містить частинки обох типів, буде вищою за відповідну ймовірність U зі статистики Ювала:

$$F > U.$$

А з цієї нерівності випливатиме, що для досягнення певної заданої ймовірності існування комірки з частинками обох типів (обернена задача) статистика Фермі потребуватиме меншої кількості частинок n_F, що розміщуються, ніж статистика Ювала:

$$n_F \le n_U.$$

Ці міркування будуть правильними у випадку рівномірного розміщення частинок по комірках [5]. Чи будуть вони правильними і для випадку нерівномірного – поліноміального – розміщення частинок?

Історичний огляд питання, сучасний стан та перспективи досліджень у задачі про дні народження можна знайти, наприклад, у [6].

Постановка задачі

Метою роботи є розв'язання оберненої задачі про дні народження для двох модифікацій нерівномірного класичного випадку цієї задачі.

Основний результат

У статі розглядатимуться дві ймовірнісні схеми, які ми будемо називати статистикою Фермі та статистикою Ювала. Обидві схеми формулюються у термінах розміщення частинок по комірках.

Нехай задано число m – кількість комірок та невід'ємні числа $p_1, p_2, ..., p_m$ такі, що $\sum_{i=1}^{m} p_i = 1$. Також маємо частинки двох типів, по n частинок кожного типу. Ці $2n$ частинок випадковим чином розміщуються в m комірках.

Статистика Фермі. Тут n частинок одного типу розміщуються в m комірках по одній, і при цьому ймовірність $P = P(i_1, ..., i_n)$ того, що зайняті комірки з номерами $i_1, ..., i_n$, $1 \le i_1 < ... < i_n \le m$, дорівнює

$$P = \frac{p_{i_1} \cdot \ldots \cdot p_{i_n}}{\sum\limits_{1 \le j_1 < \ldots < j_n \le m} p_{j_1} \cdot \ldots \cdot p_{j_n}}. \qquad (1)$$

Ця формула визначає нерівноймовірну статистику Фермі [7]. При незалежному поліноміальному розміщенні n частинок по m комірках, де ймовірності потрапляння однієї частинки в комірки дорівнюють $p_1, ..., p_m$, ймовірність $P(i_1, ..., i_n)$ з (1) дорівнює умовній ймовірності того, що частинки потрапили в комірки з номерами $i_1, ..., i_n$, $1 \le i_1 < ... < i_n \le m$, за умови, що всі n частинок потрапили в різні комірки.

Якщо $p_1 = ... = p_m = m^{-1}$, то отримуємо рівномірну статистику Фермі з $P = (C_m^n)^{-1}$. У рівномірному випадку всі можливі C_m^n розміщень n частинок по m комірках будуть рівноймовірними.

Статистика Ювала. Тут n частинок одного типу незалежно розміщуються у m комірках. Ймовірності потрапляння однієї частинки в комірки дорівнюють $p_1, ..., p_m$.

На відміну від статистики Фермі, у статистиці Ювала частинки одного типу можуть потрапити в одну комірку.

Частинки другого типу в обох статистиках розміщуються незалежно від частинок першого типу і з таким самим розподілом, як і частинки першого типу.

Позначимо $F_m(n)$ та $U_m(n)$ ймовірності того, що існує комірка, яка містить частинки обох типів для статистики Фермі та для статистики Ювала відповідно. Для обох статистик будемо розв'язувати задачу, яку вище назвали оберненою задачею про дні народження.

***Задача* 1**. Нехай задано параметр m – кількість комірок, ймовірності $p_1, ..., p_m$ та число $p \in (0,1)$ і потрібно знайти кількість частинок $n_F(m)$ та $n_U(m)$, які задовольняють рівності:

$$n_F(m) = \min\{n \ge 1 : F_m(n) \ge p\},$$
$$n_U(m) = \min\{n \ge 1 : U_m(n) \ge p\}.$$

У рівномірному випадку, $p_1 = ... = p_m = m^{-1}$, виконується нерівність

$$n_F(m) \le n_U(m).$$

Основним результатом нашого дослідження є наведена нижче теорема, в якій задача 1 розв'язується для певного класу поліноміальних схем, зокрема і для рівномірної схеми.

Теорема 1. Нехай задано деяку невід'ємну двічі диференційовану на відрізку [0,1] функцію $p(x)$ таку, що $\int_0^1 p(x)\,dx = 1$. Розглянемо поліноміальну схему $\{p_i(m), m \ge 1, i = \overline{1,m}\}$, яка задається через функцію $p(x)$ за формулою: $\forall m \ge 1$ $\forall i = \overline{1,m}$:

$$p_i = p_i(m) = \int_{\Delta_i} p(x)\,dx, \ \Delta_i = \left(\frac{i-1}{m}, \frac{i}{m}\right).$$

Тоді кількість частинок $n_F(m)$ та $n_U(m)$ із задачі 1 буде дорівнювати

$$n_F(m) = \sqrt{\frac{am}{M_2} + a\left(\frac{M_3}{M_2^2} - \frac{3}{2}\right)} + \alpha_m, \qquad (2)$$

$$n_U(m) = \sqrt{\frac{am}{M_2} + a\left(\frac{M_3}{2M_2^2} - \frac{1}{2}\right)} + \beta_m, \qquad (3)$$

де $a = -\ln(1-p)$, $M_2 = \int_0^1 p(x)^2 dx$, $M_3 = \int_0^1 p(x)^3 dx$, а для послідовностей $\{\alpha_m, m \ge 1\}$, $\{\beta_m, m \ge 1\}$ виконуються співвідношення

$$0 = \underline{\lim}\,\alpha_m = \underline{\lim}\,\beta_m,\ \overline{\lim}\,\beta_m = \overline{\lim}\,\alpha_m = 1.$$

Доведення теореми 1 буде наведено нижче, а спочатку зробимо кілька зауважень.

***Зауваження* 1.** У лівій частині формули (2) стоїть ціле число $n_F(m)$, тому в правій частині цієї формули з'являється величина α_m, яка робить цілим і значення правої частини формули (2).

При цьому виконуються рівності: $0 = \underline{\lim}\,\alpha_m$, $\overline{\lim}\,\alpha_m = 1$, але нерівність $0 \le \alpha_m \le 1$ може не виконуватися для деяких значень *m*.

***Зауваження* 2.** Якщо $p(x) \equiv 1,\ x \in [0,1]$, то поліноміальна схема в цьому випадку є рівномірною: $\forall m \ge 1 : p_1 = ... = p_m = m^{-1}$. Тоді $M_2 = M_3 = 1$ і формули (2), (3) набувають вигляду

$$n_F(m) = \sqrt{am} - \frac{a}{2} + \alpha_m, \qquad (4)$$

$$n_U(m) = \sqrt{am} + \beta_m, \qquad (5)$$

де $a = -\ln(1-p) > 0$.

Формули (4), (5) доведені в [5]. Вираз $\sqrt{am} - \frac{a}{2}$ із правої частини (4) є меншим за аналогічний вираз $\sqrt{am}$ із правої частини (5). Це узгоджується з нерівністю $n_F(m) \le n_U(m)$, про яку йшла мова у вступі.

***Зауваження* 3.** Головний асимптотичний член $\sqrt{\frac{am}{M_2}}$ є однаковим в обох формулах (2) і (3). Він виражає "парадокс днів народжень": необхідна кількість частинок $n_F(m)$ і $n_U(m)$ є незначною порівняно з числом комірок *m*:

$$n_F(m) << m,\ n_U(m) << m.$$

***Зауваження* 4.** Другий асимптотичний член в обох формулах (2) і (3) є константою, і ці константи різні. Нерівність між ними

$$a\left(\frac{M_3}{M_2^2} - \frac{3}{2}\right) < a\left(\frac{M_3}{2M_2^2} - \frac{1}{2}\right)$$

приводить до нерівності

$$\frac{M_3}{M_2^2} = \frac{M_1\ M_3}{M_2^2} < 2. \qquad (6)$$

Із формул (2) і (3) випливає, що умова (6) є достатньою для того, щоб нерівність $n_F(m) \le$ $\le n_U(m)$ виконувалася для всіх достатньо великих *m*, *n* (асимптотична нерівність).

Справді, з умови

$$0 = \underline{\lim}\,\alpha_m = \underline{\lim}\,\beta_m,\ \overline{\lim}\,\beta_m = \overline{\lim}\,\alpha_m = 1$$

випливає, що $\forall \varepsilon > 0\ \ \exists m_0\ \forall m \ge m_0$:

$$-\varepsilon < \alpha_m < 1 + \varepsilon,$$

$$-\varepsilon < \beta_m < 1 + \varepsilon.$$

Якщо позначити

$$c = a\left(\frac{M_3}{2M_2^2} - \frac{1}{2}\right) - a\left(\frac{M_3}{M_2^2} - \frac{3}{2}\right) > 0,$$

то при $\varepsilon < \frac{c}{2}$ буде виконуватися нерівність $n_F(m) \le n_U(m)$, бо тут

$$n_F(m) - n_U(m) = \alpha_m - \beta_m - c < 1 + \varepsilon + \varepsilon - c = 1 + 2\varepsilon - c < 1.$$

Отже, значення виразу $\frac{M_1\ M_3}{M_2^2}$ є визначальним для асимптотичного порівняння величин $n_F(m)$ та $n_U(m)$:

якщо $\frac{M_1\ M_3}{M_2^2} < 2$, то $n_F(m) \le n_U(m)$,

якщо $\frac{M_1\ M_3}{M_2^2} > 2$, то $n_F(m) \ge n_U(m)$

(в обох випадках нерівності між $n_F(m)$ та $n_U(m)$ є асимптотичними).

Зауважимо, що вираз

$$\frac{M_1\ M_3}{M_2^2} = \frac{\int_0^1 p(x)\,dx \int_0^1 p(x)^3 dx}{\left(\int_0^1 p(x)^2 dx\right)^2}$$

може набувати як завгодно великих значень, наприклад, для

$$p(x) = \begin{cases} N,\ 0 \le x \le \frac{1}{N^2}, \\ 1,\ \frac{1}{N^2} < x \le 1. \end{cases}$$

Але з нерівності Коші–Буняковського випливає, що завжди $\frac{M_1\ M_3}{M_2^2} \ge 1$.

Значить, у нерівномірному випадку може (!) справджуватися нерівність

$$n_F(m) > n_U(m).$$

Доведення теореми 1

Позначимо $G_m(n) = 1 - F_m(n)$ та $V_m(n) = 1 - U_m(n)$ ймовірності того, що комірки, які містять частинки обох типів, відсутні.

Статистика Фермі. Доведемо спочатку формулу (2) для статистики Фермі. Для цього розглянемо поведінку ймовірності $G(n)$ (іноді, для спрощення запису, будемо писати $G(n)$ замість $G_m(n)$). Нехай A означає подію, що немає комірок, які містять частинки обох типів, $P(A) = G(n)$, а $B(j_1,...,j_n)$ — що частинки другого типу зайняли комірки з номерами $j_1,...,j_n$, $1 \le j_1 < ... < j_n \le m$. Тоді з формули повної ймовірності маємо:

$$G(n) = P\{A\} =$$

$$= \sum_{1 \le j_1 < ... < j_n \le m} P\{A | B(j_1,...,j_n)\} \, P\{B(j_1,...,j_n)\} =$$

$$= \sum_{1 \le j_1 < \ldots < j_n \le m} \left(\sum_{1 \le i_1 < \ldots < i_n \le m} \frac{p_{i_1} \cdot \ldots \cdot p_{i_n}}{T} \right) \frac{p_{j_1} \cdot \ldots \cdot p_{j_n}}{T} =$$

$$= \frac{1}{T^2} \sum_{\substack{1 \le j_1 < ... < j_n \le m \\ 1 \le i_1 < ... < i_n \le m}} (p_{i_1} \cdot \ldots \cdot p_{i_n})(p_{j_1} \cdot \ldots \cdot p_{j_n}),$$

де позначено $T = \sum_{1 \le k_1 < ... < k_n \le m} p_{k_1} \cdot \ldots \cdot p_{k_n}$, а сумування у внутрішній сумі відбувається за всіма наборами $(i_1,...,i_n)$, які не мають спільних елементів з набором $(j_1,...,j_n) : \{i_1,...,i_n\} \cap \{j_1,...,j_n\} = \varnothing$.

Позначимо також $Q_m(n)$ ймовірність того, що при поліноміальному розміщенні n частинок у m комірках всі частинки потраплять у різні комірки:

$$Q(n) = n! \sum_{1 \le l_1 < ... < l_n \le m} p_{l_1} \cdot \ldots \cdot p_{l_n},$$

тоді $T = \frac{1}{n!} Q(n)$ і

$$\sum_{\substack{1 \le j_1 < ... < j_n \le m \\ 1 \le i_1 < ... < i_n \le m}} (p_{i_1} \cdot \ldots \cdot p_{i_n})(p_{j_1} \cdot \ldots \cdot p_{j_n}) =$$

$$= C_{2n}^n \sum_{1 \le k_1 < ... < k_{2n} \le m} p_{k_1} \cdot \ldots \cdot p_{k_{2n}} = \frac{Q(2n)}{n!\,n!}.$$

Підставимо це значення у формулу для $G(n)$:

$$G(n) = \frac{1}{T^2} \frac{Q(2n)}{n!\,n!} = \left(\frac{n!}{Q(n)} \right)^2 \frac{Q(2n)}{n!\,n!} = \frac{Q(2n)}{Q(n)^2}$$

або

$$G(n) = \frac{Q(2n)}{Q(n)^2}. \qquad (7)$$

Отже, ймовірність $G(n)$ зі статистики Фермі виражається через ймовірність $Q(n)$ з поліноміального розміщення.

Із формули (7) випливає, що розв'язок $n_F(m)$ оберненої задачі буде задовольняти співвідношення $\frac{n^2}{m} = O(1)$, $n, m \to \infty$. Справді, відомо [8], що при $\frac{n^2}{m} \to \lambda > 0 : Q(n) \to e^{-\frac{\lambda M_2}{2}}$, $n, m \to \infty$. Отже, якщо $\frac{n^2}{m} \to \lambda > 0$, то

$$G(n) = \frac{Q(2n)}{Q(n)^2} \to e^{-\frac{4\lambda M_2}{2}} \cdot (e^{-\frac{\lambda M_2}{2}})^{-2} = e^{-\lambda M_2}, \; n, m \to \infty.$$

Звідси можна показати, що

$$n_F(m) = \sqrt{\frac{a\,m}{M_2}} + o(\sqrt{m}), \; m \to \infty, \qquad (8)$$

де $a = -\ln(1-p)$.

(Справді, рівність (8) означає, що

$$\frac{n(m)^2}{m} \to \frac{a}{M_2}, \; m \to \infty.$$

Для величини $n = n(m)$ виконані співвідношення:

$$F(n-1) < p \le F(n)$$

$$G(n) \le 1 - p < G(n-1).$$

Отже, якщо $\underline{\lim} \frac{n(m)^2}{m} = \frac{a}{M_2} - \varepsilon$, $\varepsilon > 0$, то при $\frac{n(m_k)^2}{m_k} \to \frac{a}{M_2} - \varepsilon$, $k \to \infty$, матимемо протиріччя:

$$1 - p \ge G_m(n) \to e^{-\left(\frac{a}{M_2} - \varepsilon\right) M_2} =$$

$$= e^{-a} e^{\varepsilon M_2} = (1-p)\, e^{\varepsilon M_2} > 1 - p.$$

Аналогічно розглядається і випадок $\overline{\lim} \frac{n(m)^2}{m} = \frac{a}{M_2} + \varepsilon$, $\varepsilon > 0$.

Також можна показати, що $G(n) \to 1$, $\frac{n^2}{m} \to 0$, та $G(n) \to 0$, $\frac{n^2}{m} \to \infty$.)

Далі вважаємо, що співвідношення $\frac{n^2}{m} = O(1)$, $\frac{n^2}{m} = O(1)$, $n, m \to \infty$, виконується.

Щоб отримати більш точну, ніж (8), формулу для $n_F(m)$, прологарифмуємо (7):

$$\ln G(n) = \ln Q(2n) - 2\ln Q(n),$$

та скористаємося розкладом [9] для ймовірності $Q(n)$:

$$\ln Q(n) = -\frac{n^2}{2m} M_2 + \frac{n}{2m} M_2 + \frac{n^3}{3m^2}\left(M_3 - \frac{3}{2} M_2^2 \right) + O\left(\frac{n^2}{m^2} \right),\ m \to \infty.$$

Після перетворень отримаємо

$$\ln G(n) = -\frac{n^2}{m} M_2 + \frac{n^3}{m^2}(2M_3 - 3M_2^2) + O\left(\frac{n^2}{m^2} \right),\ m \to \infty. \quad (9)$$

Ця формула виражає величину $\ln G(n)$ з точністю $O\left(\frac{n^2}{m^2} \right)$, з її допомогою можна буде знайти другий доданок у формулі (2).

Отже, нехай задане число $p \in (0,1)$, тоді кількість частинок $n(m)$ з оберненої задачі задовольняє співвідношення

$$F(n-1) < p \le F(n),$$

звідси

$$G(n) \le 1 - p < G(n-1),$$

$$-\ln G(n-1) < -\ln(1-p) \le -\ln G(n).$$

Позначимо $a = -\ln(1-p) > 0$, тоді

$$a = -\ln G(n) - \theta_m(\ln G(n-1) - \ln G(n)),\ \theta_m \in [0,1). \quad (10)$$

Із формули (9) можемо отримати оцінку для виразу в дужках у формулі (10):

$$-\ln G(n) + \ln G(n-1) =$$

$$= \frac{n^2}{m} M_2 - \frac{n^3}{m^2}(2M_3 - 3M_2^2) - \frac{(n-1)^2}{m} M_2 + \frac{(n-1)^3}{m^2}(2M_3 - 3M_2^2) + O\left(\frac{n^2}{m^2} \right) =$$

$$= 2\frac{n}{m} M_2 + O\left(\frac{n^2}{m^2} \right),\ n, m \to \infty.$$

Тепер, підставивши це значення в (10), отримаємо

$$a = \frac{n^2}{m} M_2 - \frac{n^3}{m^2}(2M_3 - 3M_2^2) - 2\theta_m M_2 \frac{n}{m} + O\left(\frac{n^2}{m^2} \right).$$

З цієї рівності, що пов'язує неявним чином змінні m та $n(m)$, можна знайти і явний вираз для $n(m)$. Із формули (8) випливає, що

$$n(m) = \sqrt{\frac{a\,m}{M_2}} + z_m,$$

де $z_m = o(\sqrt{m}),\ m \to \infty$.

Підставивши цей вираз у попередню формулу, можна показати, що

$$z_m = a\left(\frac{M_3}{M_2^2} - \frac{3}{2} \right) + \theta_m + o(1),\ m \to \infty.$$

Отже,

$$n(m) = \sqrt{\frac{a\,m}{M_2}} + a\left(\frac{M_3}{M_2^2} - \frac{3}{2} \right) + \theta_m + \gamma_m, \quad (11)$$

де $\theta_m \in [0,1)$, $\gamma_m = o(1)$, $m \to \infty$.

Зробивши заміну $\alpha_m = \theta_m + \gamma_m$, приведемо формулу (11) до вигляду

$$n(m) = \sqrt{\frac{a\,m}{M_2}} + a\left(\frac{M_3}{M_2^2} - \frac{3}{2} \right) + \alpha_m, \quad (12)$$

де $0 \le \underline{\lim}\, \alpha_m$, $\overline{\lim}\, \alpha_m \le 1$.

Насправді для послідовності $\{\alpha_m, m \ge 1\}$ виконуються рівності $\underline{\lim}\, \alpha_m = 0$, $\overline{\lim}\, \alpha_m = 1$, бо дробові частини $\left\{ \sqrt{\frac{am}{M_2}} \right\}$, $m \ge 1$, всюди щільно заповнюють відрізок (0,1), а у лівій частині формули (12) стоїть ціле число.

Формулу (2) доведено.

Статистика Ювала. Далі доведемо формулу (3). Тут розглядаємо незалежне поліноміальне розміщення частинок двох типів, по n частинок кожного типу, в m комірках.

Дослідимо поведінку величини $V_m(n) = 1 - U_m(n)$ — ймовірності того, що комірки, які містять частинки обох типів, відсутні. Нехай A означає подію, що відсутні комірки, які містять частинки обох типів, $P(A) = V(n)$, а $B(i_1, ..., i_{n-k})$ —

що частинки першого типу зайняли комірки з номерами $i_1,\dots,i_{n-k}$, $1\le i_1<\dots<i_{n-k}\le m,\ 0\le k\le n-1$. Тоді з формули повної ймовірності маємо

$$V(n)=P(A)=$$

$$=\sum_{k=0}^{n-1}\sum_{1\le i_1<\dots<i_{n-k}\le m}P\{B(i_1,\dots,i_{n-k})\}\cdot P\{A|B(i_1,\dots,i_{n-k})\}=$$

$$=\sum_{k=0}^{n-1}\sum_{1\le i_1<\dots<i_{n-k}\le m}r(i_1,\dots,i_{n-k})(1-p_{i_1}-\dots-p_{i_{n-k}})^n,$$

де позначено

$$r(i_1,\dots,i_{n-k})=P\{B(i_1,\dots,i_{n-k})\}$$

і

$$r(i_1,\dots,i_{n-k})=\sum_{\substack{l_1+\dots+l_{n-k}=n\\ l_1\ge1,\dots,l_{n-k}\ge1}}\frac{n!}{l_1!\dots l_{n-k}!}p_{i_1}^{l_1}\cdot\dots\cdot p_{i_{n-k}}^{l_{n-k}}.$$

Для подальшого доведення нам знадобляться дві леми.

Лема 1. Нехай $a_1,\dots,a_m\in R$, тоді

$$\sum_{k=0}^{n-1}\sum_{1\le i_1<\dots<i_{n-k}\le m}r(i_1,\dots,i_{n-k})(a_{i_1}+\dots+a_{i_{n-k}})=$$

$$=\sum_{i=1}^{m}a_i\,(1-(1-p_i)^n).$$

Доведення. Нехай

$$f(a_1,\dots,a_m)=$$

$$=\sum_{k=0}^{n-1}\sum_{1\le i_1<\dots<i_{n-k}\le m}r(i_1,\dots,i_{n-k})(a_{i_1}+\dots+a_{i_{n-k}}),$$

тоді очевидно, що

$$f(a_1,\dots,a_m)=\lambda_1\,a_1+\dots+\lambda_m\,a_m,$$

де, наприклад,

$$\lambda_1=\frac{\partial f}{\partial a_1}=\sum_{k=0}^{n-2}\sum_{2\le i_2<\dots<i_{n-k}\le m}r(1,i_2,\dots,i_{n-k}).$$

Але подвійна сума у правій частині виражає ймовірність того, що перша комірка зайнята, отже, $\lambda_1=1-(1-p_1)^n$ та $f(a_1,\dots,a_m)=\sum_{i=1}^{m}a_i\,(1-(1-p_i)^n)$.

Лему 1 доведено.

Лема 2. Нехай $x\in(0,1)$, тоді справджується нерівність

$$0\le e^{-nx}-(1-x)^n\le n\,x^2e^{-nx}.$$

Доведення. З нерівності Бернуллі маємо $(1-x^2)^n\ge1-n\,x^2$, отже,

$$e^{nx}(1-x)^n\ge(1+x)^n(1-x)^n=(1-x^2)^n\ge1-n\,x^2$$

або

$$(1-x)^n\ge e^{-nx}-n\,x^2e^{-nx}.$$

Лему 2 доведено.

Далі припустимо, що $\frac{n^2}{m}\to\lambda>0,\ n,m\to\infty$, і покажемо, що тоді $V(n)\to e^{-\lambda M_2},\ n,m\to\infty$.

Для цього, використовуючи доведені леми, спочатку оцінимо різницю $R(n)-V(n)$, де

$$R(n)=\sum_{k=0}^{n-1}\sum_{1\le i_1<\dots<i_{n-k}\le m}r(i_1,\dots,i_{n-k})\,e^{-n(p_{i_1}+\dots+p_{i_{n-k}})},$$

$$V(n)=\sum_{k=0}^{n-1}\sum_{1\le i_1<\dots<i_{n-k}\le m}r(i_1,\dots,i_{n-k})(1-p_{i_1}-\dots-p_{i_{n-k}})^n.$$

Позначимо $x=p_{i_1}+\dots+p_{i_{n-k}}$, тоді з доведених лем маємо

$$0\le R(n)-V(n)=$$

$$=\sum_{k=0}^{n-1}\sum_{1\le i_1<\dots<i_{n-k}\le m}r(i_1,\dots,i_{n-k})(e^{-nx}-(1-x)^n)\le$$

$$\le\sum_{k=0}^{n-1}\sum_{1\le i_1<\dots<i_{n-k}\le m}r(i_1,\dots,i_{n-k})\,n\,x^2e^{-nx}\le$$

$$\le\sum_{k=0}^{n-1}\sum_{1\le i_1<\dots<i_{n-k}\le m}r(i_1,\dots,i_{n-k})\,n\,(p_{i_1}+\dots+p_{i_{n-k}})^2\le$$

$$\le\sum_{k=0}^{n-1}\sum_{1\le i_1<\dots<i_{n-k}\le m}r(i_1,\dots,i_{n-k})\,n(n-k)\,(p_{i_1}^2+\dots+p_{i_{n-k}}^2)\le$$

$$\le n^2\sum_{k=0}^{n-1}\sum_{1\le i_1<\dots<i_{n-k}\le m}r(i_1,\dots,i_{n-k})\,(p_{i_1}^2+\dots+p_{i_{n-k}}^2)=$$

$$=n^2\sum_{i=1}^{m}p_i^2(1-(1-p_i)^n)\le n^2\sum_{i=1}^{m}p_i^2\cdot np_i=$$

$$=n^3\sum_{i=1}^{m}p_i^3=n^3S_3.$$

Отже,

$$V(n)=R(n)-\theta\,n^3S_3,\qquad(13)$$

де $\theta\in(0,1)$, $S_3=S_3(m)=\sum_{i=1}^{m}p_i^3$.

Можна показати, що якщо $p(x)\in C^{(2)}([0,1])$, то

$$M_2-mS_2(m)=M_2-m\sum_{i=1}^{m}p_i^2=O(m^{-2}),\ m\to\infty,\qquad(14)$$

$$M_3 - m^2 S_3(m) = M_3 - m^2 \sum_{i=1}^{m} p_i^3 = O(m^{-2}),\ m \to \infty. \quad (15)$$

Отже, $n^3 S_3 = O\left(\frac{n^3}{m^2}\right) = O\left(\frac{n}{m}\right),\ m \to \infty$, або, після підстановки цього значення в (13):

$$V(n) = R(n) + O\left(\frac{n}{m}\right),\ m \to \infty. \quad (16)$$

Далі оцінимо величину $R(n)$. Для цього зауважимо, що

$$r(i_1,\dots,i_{n-k}) = \sum_{\substack{l_1+\dots+l_{n-k}=n \\ l_1\ge 1,\dots,l_{n-k}\ge 1}} \frac{n!}{l_1!\dots l_{n-k}!} p_{i_1}^{l_1} \cdot \dots \cdot p_{i_{n-k}}^{l_{n-k}} =$$

$$= n!\ \underset{x^n}{\operatorname{coef}} \left\{ \left(p_{i_1} x + \frac{p_{i_1}^2 x^2}{2!} + \dots \right) \dots \left(p_{i_1} x + \frac{p_{i_1}^2 x^2}{2!} + \dots \right) \right\} =$$

$$= n! \underset{x^n}{\operatorname{coef}} \{ (e^{p_{i_1} x} - 1) \dots (e^{p_{i_{n-k}} x} - 1) \}.$$

Звідси для величини $R(n)$ маємо представлення:

$$R(n) = \sum_{k=0}^{n-1} \sum_{1 \le i_1 < \dots < i_{n-k} \le m} r(i_1,\dots,i_{n-k}) e^{-n(p_{i_1}+\dots+p_{i_{n-k}})} =$$

$$= \sum_{k=0}^{n-1} \sum_{1 \le i_1 < \dots < i_{n-k} \le m} n! \underset{x^n}{\operatorname{coef}} \{ (e^{p_{i_1} x} - 1) \dots (e^{p_{i_{n-k}} x} - 1) \} \times$$

$$\times e^{-n(p_{i_1}+\dots+p_{i_{n-k}})} = n! \underset{x^n}{\operatorname{coef}} \left\{ \sum_{k=0}^{n-1} \sum_{1 \le i_1 < \dots < i_{n-k} \le m} (e^{p_{i_1} x} - 1) \times \right.$$

$$\left. \times e^{-np_{i_1}} \cdot \dots \cdot (e^{p_{i_{n-k}} x} - 1) e^{-np_{i_{n-k}}} \right\} = n! \underset{x^n}{\operatorname{coef}} \{ (1 + (e^{p_1 x} - 1) \times$$

$$\times e^{-np_1}) \cdot \dots \cdot (1 + (e^{p_m x} - 1) e^{-np_m}) \}.$$

Отже, якщо

$$F(x) = \prod_{i=1}^{m} (1 + (e^{p_i x} - 1) e^{-np_i}),$$

то значення $R(n)$ можна знаходити за формулою $R(n) = F^{(n)}(0)$.

Позначимо $\xi_1,\dots,\xi_m$ незалежні випадкові величини, які набувають двох значень: 0 та 1, і

$$P\{\xi_i = 0\} = 1 - e^{-np_i},\ P\{\xi_i = 1\} = e^{-np_i}, \quad i = \overline{1,m}.$$

Тоді

$$F(x) = \prod_{i=1}^{m} (1 + (e^{p_i x} - 1) e^{-np_i}) =$$

$$= \prod_{i=1}^{m} M\, e^{p_i \xi_i x} = M\, e^{x \sum_{i=1}^{m} p_i \xi_i}.$$

Тут математичне сподівання є скінченною сумою, отже,

$$F^{(n)}(x) = M \left(\sum_{i=1}^{m} p_i \xi_i \right)^n e^{x \sum_{i=1}^{m} p_i \xi_i}$$

та

$$R(n) = F^{(n)}(0) = M \left(\sum_{i=1}^{m} p_i \xi_i \right)^n = M \left(1 - \sum_{i=1}^{m} p_i \eta_i \right)^n,$$

де $\eta_i = 1 - \xi_i,\ i = \overline{1,m}$.

Позначимо $\zeta = \sum_{i=1}^{m} p_i \eta_i$, тоді з леми 2 матимемо:

$$R(n) = M(1-\zeta)^n =$$

$$= Me^{-n\zeta} - \theta\, n\, M \zeta^2 e^{-n\zeta},\ \theta \in (0,1), \quad (17)$$

де

$$Me^{-n\zeta} = Me^{-n\sum_{i=1}^{m} p_i \eta_i} = \prod_{i=1}^{m} Me^{-np_i \eta_i} =$$

$$= \prod_{i=1}^{m} \{ e^{-np_i} + (1 - e^{-np_i}) e^{-np_i} \} = \prod_{i=1}^{m} \{ 2e^{-np_i} - e^{-2np_i} \} =$$

$$= \prod_{i=1}^{m} (1 - (1 - e^{-np_i})^2).$$

Оцінимо другий доданок у (17):

$$M \zeta^2 e^{-n\zeta} \le M \zeta^2 = M \left(\sum_{i=1}^{m} p_i \eta_i \right)^2 =$$

$$= M \sum_{i,j=1}^{m} p_i p_j \eta_i \eta_j = \sum_{i,j=1}^{m} p_i p_j M \eta_i \eta_j =$$

$$= \sum_{i \ne j} p_i p_j\, M \eta_i \eta_j + \sum_{i=1}^{m} p_i^2 M \eta_i \le \left(\sum_{i=1}^{m} p_i M \eta_i \right)^2 +$$

$$+ \sum_{i=1}^{m} p_i^2 M \eta_i \le |\, M \eta_i = 1 - e^{-np_i} \le np_i,$$

$$i = \overline{1,m}\,| \le \left(\sum_{i=1}^{m} np_i^2 \right)^2 + n \sum_{i=1}^{m} p_i^3 = n^2 S_2^2 + n S_3.$$

Зі співвідношень (14) і (15) випливає, що $n^2 S_2^2 = O\left(\frac{n^2}{m^2}\right),\ n S_3 = O\left(\frac{n}{m^2}\right)$, отже,

$$n\, M\zeta^2 e^{-n\zeta} = n\, O\left(\frac{n^2}{m^2}\right) = O\left(\frac{n^3}{m^2}\right) = O\left(\frac{n}{m}\right),\ n,m \to \infty.$$

Отримали оцінку для $R(n)$:

$$R(n) = A(n) + O\left(\frac{n}{m}\right),\ n,m \to \infty, \qquad (18)$$

де $A(n) = \prod_{i=1}^{m}(1-(1-e^{-np_i})^2)$.

Далі отримаємо оцінку для величини $A(n)$. Спочатку зауважимо, що

$$np_i = n\int_{\Delta_i} p(x)\,dx \le \frac{n}{m}L,\ i = \overline{1,m},$$

де $L = \max_{[0,1]} p(x)$.

Отже, можемо вважати, що $np_i \le \delta$, де $\delta > 0$ — певне додатне число.

Із формули Тейлора маємо

$$\ln(1-(1-e^{-x})^2) = -x^2 + O(x^3),\ x \to 0.$$

або

$$\ln(1-(1-e^{-x})^2) = -x^2 + x^3 C(x),\ x \in [0,\delta].$$

Функція $C(x)$ є аналітичною в нулі, а отже, є обмеженою в певному околі цієї точки. Нехай $C = \max_{[0,\delta]} C(x)$, тоді

$$\ln A(n) =$$

$$= \ln\prod_{i=1}^{m}(1-(1-e^{-np_i})^2) = \sum_{i=1}^{m}\ln(1-(1-e^{-np_i})^2) =$$

$$= \sum_{i=1}^{m}\{-(np_i)^2 + C(np_i)(np_i)^3\} = -n^2 S_2 + \theta C n^3 S_3 =$$

$$= -n^2 S_2 + O\left(\frac{n}{m}\right),\ \ n,m \to \infty.$$

Отже

$$A(n) = e^{-n^2 S_2 + O\left(\frac{n}{m}\right)} = e^{-n^2 S_2}\left(1 + O\left(\frac{n}{m}\right)\right) =$$

$$= e^{-n^2 S_2} + O\left(\frac{n}{m}\right),\ n,m \to \infty.$$

Далі з формули (14) маємо

$$A(n) = e^{-n^2 S_2} + O\left(\frac{n}{m}\right) = e^{-n^2\left(\frac{M_2}{m} + O\left(\frac{1}{m^2}\right)\right)} +$$

$$+ O\left(\frac{n}{m}\right) = e^{-\frac{n^2}{m}M_2} + O\left(\frac{n}{m}\right),\ n,m \to \infty.$$

Таким чином, для величини $A(n)$ можна отримати оцінку:

$$A(n) = e^{-\frac{n^2}{m}M_2} + O\left(\frac{n}{m}\right),\ n,m \to \infty. \qquad (19)$$

Із формул (16), (18) і (19) отримуємо

$$V(n) = e^{-\frac{n^2}{m}M_2} + O\left(\frac{n}{m}\right),\ n,m \to \infty.$$

Звідси, якщо $\frac{n^2}{m} \to \lambda > 0$, то $V(n) \to e^{-\lambda M_2}$, $n,\ m \to \infty$. З цього граничного співвідношення, таким самим способом, як і для статистики Фермі (див. вище), можна отримати перший доданок із формули (3):

$$n_U(m) = \sqrt{\frac{a\,m}{M_2}} + o(\sqrt{m}),\ m \to \infty.$$

Щоб отримати другий доданок у (3), треба повторити всі обчислення для статистики Ювала, наведені тут, але з точністю $O\left(\frac{n^2}{m^2}\right)$. Вище всі формули наводилися лише з точністю $O\left(\frac{n}{m}\right)$, що дало змогу знайти тільки перший доданок із (3). Отже, схема отримання другого доданка така ж, як і для першого доданка. Ця схема складається з трьох кроків.

К р о к 1. Перехід від $V(n)$ до $R(n)$. Цей перехід відбувається за допомогою співвідношення [5], що узагальнює лему 2:

$$(1-x)^n = e^{-nx} - \frac{n}{2}x^2 e^{-nx} -$$

$$-\theta_1\frac{n}{2}x^3 e^{-nx} + \theta_2\frac{n^2}{2}x^4 e^{-nx}, \qquad (20)$$

де $x \in (0,1)$, $\theta_1, \theta_2 \in (0,1)$.

З допомогою рівності (20) отримуємо оцінку

$$V(n) = R(n) - \frac{n}{2}S(n) + O\left(\frac{n^2}{m^2}\right),\ m \to \infty,$$

де

$$S(n) = \sum_{k=0}^{n-1} \sum_{1 \le i_1 < \ldots < i_{n-k} \le m} r(i_1, \ldots, i_{n-k})(p_{i_1} + \ldots + p_{i_{n-k}})^2 \times$$
$$\times e^{-n(p_{i_1} + \ldots + p_{i_{n-k}})}$$

Крок 2. Перехід від $R(n), S(n)$ до $A(n)$. Цей перехід відбувається з допомогою представлення $R(n) = M(1-\varsigma)^n$ та рівності (20). Можна отримати такі співвідношення:

$$R(n) = A(n)\left(1 - \frac{n^3}{2} S_2^2\right) + O\left(\frac{n^2}{m^2}\right), \; m \to \infty,$$

$$S(n) = n^2 S_2^2 R(n) + O\left(\frac{n^2}{m^2}\right), \; m \to \infty.$$

Звідси

$$V(n) = A(n)(1 - n^3 S_2^2) + O\left(\frac{n^2}{m^2}\right), \; m \to \infty.$$

Крок 3. Розклад для величини $A(n)$. Цей розклад можна отримати так само, як це було зроблено для першого доданка:

$$A(n) = e^{-n^2 S_2 + n^3 S_3} + O\left(\frac{n^2}{m^2}\right), \; m \to \infty.$$

Тепер можна отримати формулу для ймовірності $V(n)$ з точністю $O\left(\frac{n^2}{m^2}\right)$:

$$\ln V(n) = -n^2 S_2 + n^3(S_3 - S_2^2) + O\left(\frac{n^2}{m^2}\right) =$$
$$= -\frac{n^2}{m} M_2 + \frac{n^3}{m^2}(M_3 - M_2^2) + O\left(\frac{n^2}{m^2}\right), \; n, m \to \infty.$$

Подальші перетворення аналогічні тим, що були зроблені вище для статистики Фермі.

Отже, другий доданок із (3) знаходиться аналогічно. Формулу (3) доведено. Теорему 1 доведено.

Висновки

У статі наведено формули (2) і (3), які дають асимптотичний розв'язок оберненої задачі про дні народження для статистик Фермі та Ювала відповідно. У нерівномірному випадку, що розглядався у статті, нерівність між необхідним числом частинок

$$n_F \le n_U$$

може не виконуватися (на відміну від рівномірного випадку). Це є своєрідним "парадоксом" у "парадоксі днів народжень". Умова (6) визначає, коли нерівність $n_F \le n_U$ все ж справджується. Видно, що, зокрема, рівномірна поліноміальна схема цю умову задовольняє.

Формули (2) і (3) дають асимптотичний розв'язок оберненої задачі. Метою подальших досліджень може бути отримання аналогічних формул для розв'язання прямої задачі.

Список літератури

1. *Секей Г.* Парадоксы в теории вероятностей и математической статистике. – М.: Мир, 1990. – 240 с.
2. *Nunnikhoven T.* A birthday problem solution for nonuniform birth frequencies // The American Statistician. – 1992. – **46**. – P. 601–606.
3. *Mathis F.* A generalized birthday problem // SIAM Rev. – 1991. – **33**, № 2. – P. 265–270.
4. *Heuer G.A.* Estimation of a certain probability problem // Am. Math. Monthly. – 1959. – **66**. – P. 704–706.
5. *Єндовицький П.О.* Дві модифікації задачі про дні народження // Наукові вісті НТУУ "КПІ". – 2015. – № 4. – С. 47–55.
6. *DasGupta A.* The matching, birthday and strong birthday problem: a contemporary review // J. Statist. Planning Inference. – 2005. – **130**. – P. 377–389.
7. *Иванов В.А., Теребулин С.Ю.* Некоторые предельные теоремы в неравновероятной схеме размещения частиц комплектами // Теория вероятностей и ее применения. – 1978. – **3**. – С. 643–650.
8. *Колчин В.Ф., Севастьянов Б.А., Чистяков В.П.* Случайные размещения. – М.: Наука, 1976. – 224 с.
9. *Єндовицький П.О.* Точна асимптотична оцінка розміру групи в узагальненні парадоксу днів народжень // Наукові вісті НТУУ "КПІ". – 2010. – № 4. – С. 55–60.

References

1. G. Szekely, Paradoxes in Probability Theory and Mathematical Statistics. Moscow, USSR: Mir, 1990 (in Russian).
2. T. Nunnikhoven, "A birthday problem solution for nonuniform birth frequencies", *The American Statistician*, vol. 46, pp. 601–606, 1992.

3. F. Mathis, "A generalized birthday problem", *SIAM Rev.*, vol. 33, pp. 265–270, 1991.
4. G.A. Heuer, "Estimation of a certain probability problem", *Am. Math. Monthly*, vol. 66, pp. 704–706, 1959.
5. P. Yendovytskij, "Two birthday problem modifications", *Naukovi Visti NTUU "KPI"*, vol. 4, pp. 47–55, 2015 (in Ukrainian).
6. A. DasGupta, "The matching, birthday and strong birthday problem: a contemporary review", *J. Statist. Planning Inference*, vol. 130, pp. 377–389, 2005.
7. V. Ivanov *et al.*, "Some limit theorems in the non-uniform allocation scheme", *Teorija Verojatnostej i ee Primenenija*, vol. 3, pp. 643–650, 1978 (in Russian).
8. V.F. Colchin *et al.*, Random Allocations. Moscow, USSR: Nauka, 1976 (in Russian).
9. P. Yendovytskij, "Exact asymptotic approximation of the group size in generalization of birthday paradox", *Naukovi Visti NTUU "KPI"*, vol. 4, pp. 55–60, 2010 (in Ukrainian).

П.О. Єндовицький

ДВІ МОДИФІКАЦІЇ ЗАДАЧІ ПРО ДНІ НАРОДЖЕННЯ: НЕРІВНОМІРНИЙ ВИПАДОК

Проблематика. Схема розміщення частинок по комірках досліджується як у теорії ймовірностей, так і в математичній статистиці. В теорії ймовірностей мова йде про доведення граничних теорем для цієї схеми, в математичній статистиці – про побудову статистичних критеріїв. Одним із важливих питань у цій теорії є задача про дні народження.

Мета дослідження. У статі розглядаються дві модифікації класичної задачі про дні народження. Одна модифікація формулюється у схемі статистики Фермі, інша – у схемі поліноміального та незалежного розміщення частинок по комірках. В обох випадках метою дослідження був розв'язок задачі про дні народження.

Методика реалізації. Використовувалися стандартні асимптотичні методи. При цьому спочатку було доведено певну граничну теорему та знайдено швидкість збіжності в ній. З допомогою цих результатів було проведено числовий підрахунок ймовірностей у задачі про дні народження та отримано формули для розміру групи в цій задачі.

Результати дослідження. Отримано числові оцінки для ймовірності та розміру групи із задачі про дні народження.

Висновки. Для обох модифікацій збігаються головні члени асимптотики, як у формулі для підрахунку ймовірності, так і у формулі для розміру групи, але вже другі доданки в отриманих асимптотичних формулах різняться.

Ключові слова: задача про дні народження; парадокс днів народжень; випадкові розміщення; статистика Фермі; атака Ювала.

П.А. Ендовицкий

ДВЕ МОДИФИКАЦИИ ЗАДАЧИ ПРО ДНИ РОЖДЕНИЯ: НЕРАВНОМЕРНЫЙ СЛУЧАЙ

Проблематика. Схема размещения частиц по ячейкам исследуется как в теории вероятностей, так и в математической статистике. В теории вероятностей речь идет о предельных теоремах для этой схемы, в математической статистике – о построении статистических критериев. Одним из важных вопросов в этой теории является задача про дни рождения.

Цель исследования. В статье рассмотрены две модификации классической задачи про дни рождения. Одна модификация формулировалась в схеме статистики Ферми, вторая – в терминах неравновероятного и независимого размещения частиц по ячейкам. В обоих случаях целью исследования было решение задачи про дни рождения.

Методика реализации. Использовались стандартные асимптотические методы. При этом сначала была доказана определенная предельная теорема и найдена скорость сходимости в ней. С помощью этих результатов был проведен численный подсчет вероятностей в задаче про дни рождения и получены формулы для размера группы в этой задаче.

Результаты исследования. Получены числовые оценки для вероятностей и размера группы из задачи про дни рождения.

Вывод. Для обеих модификаций совпадают главные члены асимптотики, как в формуле для подсчета вероятностей, так и в формуле для размера группы, но уже вторые слагаемые в полученных асимптотических формулах отличаются.

Ключевые слова: задача про дни рождения; парадокс дней рождений; случайные размещения; статистика Ферми; атака Ювала.

Рекомендована Радою
Фізико-технічного інституту
НТУУ "КПІ"

Надійшла до редакції
07 червня 2016 року