ИЗВЕСТИЯ ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЙ

2000 МАТЕМАТИКА № 5 (456)

УДК 519.216

Л.А. ЗОЛОТУХИНА

АСИМПТОТИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ ЧИСЛА СОВПАДЕНИЙ ДВУМЕРНОЙ ВЫБОРКИ ПРИ НАТУРАЛЬНОМ СОВМЕЩЕНИИ

1. Введение

Классическая задача о совпадениях сформулирована Монмортом в 1708 г. Пусть мы имеем, например, две колоды карт. Перетасуем колоды каждую в отдельности, а затем разложим на столе одну под другой. Если сверху и снизу оказалась одна и та же карта, то будем говорить о совпадении. Известно, что среднее число совпадений при таком сопоставлении при любом числе карт в колоде равно 1. Если же будем увеличивать число карт в колоде, распределение числа совпадений будет стремиться к распределению Пуассона с параметром 1.

В работах [1]–[3] было рассмотрено следующее обобщение этой задачи. Пусть выборка (u_i, t_i) $(i = 1, \ldots, n)$ объема n извлечена из двумерной генеральной совокупности с плотностью распределения $f(x, y) > 0$ при $a < x < b$, $c < y < d$. Возможно a, c равны $-\infty$, а b, d равны $+\infty$. Но компоненты каждого выборочного вектора (u_i, t_i) наблюдаются отдельно, так что имеем два набора наблюдений:

$\vec{t} = (t_1, \ldots, t_n)$ — наблюдения первой компоненты,

$\vec{u} = (u_1, \ldots, u_n)$ — наблюдения второй компоненты.

Затем элементы каждой из выборок располагаются в случайном порядке (любая перестановка $(t_1, \ldots t_n)$ и $(u_1, \ldots u_n)$ является равновероятной). Необходимо найти такой способ соединения элементов выборок $\vec{t}$ и $\vec{u}$ в пары, чтобы наилучшим образом репродуцировать пары исходной выборки. В [1] было предложено несколько критериев оптимальности, в частности, а) максимизация вероятности репродукции пар и б) максимизация среднего числа совпадений. Способ соединения в пары, удовлетворяющий критерию б), был назван “оптимальным” в отличие от удовлетворяющего критерию а) “максимального правдоподобия” (ML). Способ, при котором в пары объединяются соответствующие элементы вариационных рядов, будем называть натуральным совмещением.

В совокупности в [1]–[3] были получены следующие результаты. Пусть плотность распределения генеральной совокупности $f(x, y) > 0$, $x \in R$, $y \in R$, и обладает монотонным отношением правдоподобия

$$f(x_1, y_1)f(x_2, y_2) \geq f(x_1, y_2)f(x_2, y_1) \quad \text{для} \quad x_1 < x_2; \quad y_1 < y_2.$$

Тогда а) ML-правило есть натуральное совмещение; б) оптимальное правило предполагает объединение в пары наибольших и наименьших элементов $\vec{t}$ и $\vec{u}$; в) при натуральном совмещении среднее число совпадений не меньше 1, т. е. натуральное совмещение всегда не хуже случайного объединения в пары.

Все эти выводы были сделаны авторами для любых объемов выборки n. Позже в [4], [5] были получены интегральные представления математического ожидания и дисперсии числа совпадений, а также вычислена асимптотика при $n \to +\infty$ среднего числа совпадений при натуральном совмещении.

Целью данной статьи является получение интегрального представления всех моментов числа совпадений (п.2), а также доказательство того факта, что асимптотическое распределение числа

совпадений — распределение Пуассона с параметром $\lambda = \int_a^b \frac{f(x,y^*(x))}{f_2(y^*(x))}dx < \infty$, где $f_2(y) = \int_a^b f(x,y)dx > 0$ — маргинальная плотность распределения второй компоненты,

$$F_1(x) = \int_a^x \int_c^d f(x,y)dx\,dy, \quad F_2(x) = \int_a^b \int_c^y f(x,y)dx\,dy$$

— маргинальные функции распределения u_i и t_i соответственно, а $y^*(x) = F_2^{-1}(F_1(x))$.

Если же $\lambda = \infty$, то распределение числа совпадений асимптотически нормально, точнее

$$P\left(\frac{N - MN}{\sqrt{DN}} < x\right) \to \phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(\frac{-u^2}{2}\right)du.$$

Кроме того, п. 5 содержит доказательство следующего утверждения: для любого n число совпадений распределено по рандомизированному биномиальному закону.

2. Интегральные представления моментов

В дальнейшем для определенности положим $f(x,y) > 0$, $x \in R$, $y \in R$. Итак, пусть из двумерной генеральной совокупности с плотностью $f(x,y)$ извлечена выборка $((t_1,\ldots,t_n),(u_1,\ldots,u_n))$ отдельно по компонентам. Пусть, кроме того, $(x_1,x_2,\ldots,x_n)$ — упорядоченные $(t_1,t_2,\ldots,t_n)$, а $(y_1,y_2,\ldots,y_n)$ — упорядоченные $(u_1,u_2,\ldots,u_n)$. Будем объединять в пары x_i с y_i — в этом состоит натуральное совмещение.

Случайные величины χ_i введем следующим образом: $\chi_i = 1$, если среди образовавшихся пар присутствует пара (t_i,u_i), и $\chi_i = 0$ в противном случае.

Пусть N — число совпадений. Очевидно, $N = \sum\limits_{i=1}^n \chi_i$ и χ_i зависимы симметрично. Тогда

$$MN^k = M\left(\sum_{i=1}^n \chi_i\right)^k = M \sum_{(k_1,\ldots,k_n)} \frac{\chi_1^{k_1}\chi_2^{k_2}\cdots\chi_n^{k_n}}{k_1!k_2!\cdots k_n!}k! = \sum_{j=1}^k \frac{c_{jk}\,n!}{(n-j)!}M\Big(\prod_{m=1}^j \chi_m\Big), \tag{2.1}$$

т. к. $\chi_i^k = \chi_i$ для любых целых i и k, $M\chi_{i_1}\chi_{i_2}\cdots\chi_{i_j} = M\chi_1\chi_2\cdots\chi_j$ в силу симметричной зависимости χ_i. Необходимо вычислить моменты вида $M(\chi_1\cdots\chi_k)$.

Нетрудно видеть, что $M(\chi_1\ldots\chi_k) = P(\chi_1 = 1,\ldots,\chi_k = 1)$. Зафиксируем первые k пар $(t_1,u_1),\ldots,(t_k,u_k)$. Пусть все эти пары истинны при натуральном совмещении, $(x_1,\ldots,x_k)$ упорядоченны по возрастанию $(t_1,\ldots,t_k)$, а $(y_1,\ldots,y_k)$ упорядоченны $(u_1,\ldots,u_k)$. Тогда множества пар $\{(x_1,y_1),\ldots,(x_k,y_k)\}$ и $\{(t_1,u_1),\ldots,(t_k,u_k)\}$ совпадают. Посмотрим, как могут распределиться остальные $n-k$ пар. Ясно, что, сложив вероятности возможных исходов, будем иметь искомую вероятность.

Чтобы облегчить описание процедуры подсчета, введем несколько понятий.

Пусть пары (x_i,y_i), $i \in 1,\ldots,k$, истинны при натуральном совмещении. Точки с координатами (x_i,y_i), $i \in 1,\ldots,k$, назовем узлами. Каждый узел прямыми $y = y_i$ и $x = x_i$ делит плоскость на четыре части. Очевидно, для каждого узла число остальных пар, попавших в левую верхнюю четверть, и в нижнюю правую должны совпадать. Это число для i-й истинной пары будем называть i-м ящиком и обозначать S_i. Далее, левую верхнюю четверть назовем для удобства i-м верхним ящиком, правую нижнюю — i-м нижним ящиком.

Каждой клетке сопоставим вероятность попадания в нее и число пар, в нее попавших. Клетка — прямоугольник, ограниченный прямыми $y = y_i$, $y = y_{i+1}$, $x = x_i$, $x = x_{i+1}$, где $x_0 = y_0 = -\infty$, $x_{k+1} = y_{k+1} = +\infty$. Нумерацию строк будем вести снизу вверх, столбцов — слева направо.

Для каждого ящика число пар, попавших в клетки, прилегающие к главной диагонали (соединяющей левую нижнюю клетку с правой верхней), будем вычислять как S_i минус число пар, попавших в остальные клетки ящика. Эти клетки назовем соответственно результирующими верхнего (нижнего) ящика. Клетки ящика, не прилегающие к главной диагонали, назовем

элементами ящика. По аналогии с понятием ящика будем числу пар, попавших в клетку, давать то же название, что и самой клетке.

Итак, пусть k_{ij} — элемент ящика. Символом $\sum_i^{(1)}$ будем обозначать суммы элементов i-го верхнего ящика, символом $\sum_i^{(2)}$ — i-го нижнего. Пусть, наконец, p_{ij} — вероятность попадания в клетку (i,j), $g = 1 - \sum\limits_{i \neq j} p_{ij}$, а $\sum^*$ — число пар, не попавших на главную диагональ. Искомая вероятность равна

$$\sum_{k_{ij}, S_i} \frac{(n-k)! \prod\limits_{|i-j|\geq 2} p_{ij}^{k_{ij}} \prod\limits_i \left(p_{i,i+1}^{S_i - \sum_i^{(2)}}\right)\left(p_{i+1,i}^{S_i - \sum_i^{(1)}}\right)\left(g^{n-k-\sum^*}\right)}{\prod\limits_{|i-j|\geq 2}(k_{ij})! \prod\limits_i \left(S_i - \sum_i^{(1)}\right)!\left(S_i - \sum_i^{(2)}\right)!\left(n-k-\sum^*\right)!},$$

где суммирование проводится при неотрицательности всех чисел, стоящих в знаменателе под знаком факториала; k_{ij} исчезли при суммировании (см. [4]).

Легко видеть, что каждый член суммы равен

$$\frac{1}{(2\pi i)^{k^2+k}} \int\limits_{\Gamma \subset C^{k^2+k}} \frac{\left(g + \sum\limits_{i\neq j} p_{ij} u_{ij}\right)^{n-k} du}{\prod\limits_{i\neq j} u_{ij} \prod\limits_{|i-j|\geq 2} u_{ij}^{k_{ij}} \prod\limits_i \left(u_{i,i+1}^{S_i - \sum_i^{(1)}} u_{i+1,i}^{S_i - \sum_i^{(2)}}\right)},$$

где Γ — декартово произведение k^2+k замкнутых контуров на комплексных плоскостях, содержащих точку $(0,0)$ внутри себя. Если какая-то из степеней в знаменателе будет больше $n-k$, то соответствующий интеграл обратится в нуль; поэтому суммирование по всем переменным можно вести до бесконечности. Теперь все суммы по k_{ij} — бесконечные геометрические прогрессии. Выберем контур интегрирования так, чтобы эти прогрессии бесконечно убывали.

Обозначив искомую вероятность при фиксированных $(x_1,y_1),(x_2,y_2),\ldots,(x_k,y_k)$ через P, будем иметь

$$P = \frac{1}{(2\pi i)^{k^2+k}} \sum_{S_i} \int\limits_{\Gamma \subset C^{k^2+k}} \frac{\left(g + \sum\limits_{i\neq j} p_{ij} u_{ij}\right)^{n-k}}{\prod\limits_i (u_{i,i+1}u_{i+1,i})^{S_i} \prod\limits_{i\neq j} u_{ij} \prod\limits_{i-j>1} \left(1 - \sum\limits_{m=j}^{i-1} u_{m+1,m}/u_{ij}\right)} \times$$

$$\times \frac{du}{\prod\limits_{i-j<-1} \left(1 - \sum\limits_{m=i}^{j-1} u_{m,m+1}/u_{ij}\right)}.$$

Проинтегрируем по u_{ij}, где $|i-j|>1$, затем просуммируем по ящикам и проинтегрируем по $u_{i+1,i}$, т. е. по соответствующим результирующим верхних ящиков. В итоге получим

$$P = \frac{1}{(2\pi i)^k} \int\limits_{\Gamma \subset C^k} f(u)du,$$

где

$$f(u) = \frac{\left(g + \sum\limits_{i<j}\left(p_{ij}\prod\limits_{m=i}^{j-1} u_{m,m+1} + p_{ji}\prod\limits_{m=i}^{j-1} u_{m,m+1}^{-1}\right)\right)^{n-k}}{\prod\limits_{i=1}^{k} u_{i,j+1}}.$$

Осталось осреднить P по возможным значениям $(x_1,y_1),(x_2,y_2),\ldots,(x_k,y_k)$ и учесть, что всего существует $k!$ перестановок $(x_1,y_1),(x_2,y_2),\ldots,(x_k,y_k)$. Тогда для момента $M(\chi_1\ldots\chi_k)$ имеем

интегральное представление

$$\frac{k!}{(2\pi i)^k}\int\limits_{x_1<\cdots<x_k}\cdots\int\cdot\int\limits_{y_1<\cdots<y_k}\cdots\int\cdot\int\limits_{\Gamma\subset C^k} f(u)du\times\prod_{i=1}^{k}(f(x_i,y_i)dx_idy_i). \tag{2.2}$$

Для нахождения MN^k осталось применить (2.1).

3. Асимптотика MN^k при $n\to\infty$

Как и в предыдущем пункте, займемся сначала моментами вида $M(\chi_1\cdots\chi_k)$, а затем воспользуемся формулой (2.1).

Для нахождения асимптотики $M(\chi_1\cdots\chi_k)$ применим метод перевала по переменным $u_{m,m+1}$ и метод Лапласа по переменным y_i. Обозначим

$$H=H(u,x,y)=g+\sum_{i<j}\left(p_{ij}\prod_{m=i}^{j-1}u_{m,m+1}+p_{ji}\prod_{m=i}^{j-1}u_{m,m+1}^{-1}\right),$$

где

$$g=1-\sum_{i<j}(p_{ij}+p_{ji}),$$
$$u=(u_{12},u_{23},\ldots,u_{k,k+1}),\quad x=(x_1,\ldots,x_k),\quad y=(y_1,\ldots,y_k),$$

причем $y_1<y_2<\cdots<y_k$, $x_1<x_2<\cdots<x_k$.

Для нахождения асимптотики интеграла по замкнутому контуру Γ необходимо найти точку, в которой достигается минимакс

$$\min_{\Gamma\in G}\max_{u\in\Gamma}\operatorname{Re}H(u,x,y), \tag{3.1}$$

где G — множество всех контуров, сохраняющих значение интеграла. В качестве контуров G будем рассматривать окружности $u_{m,m+1}=r_m\exp(i\psi_m)$, $m=1,\ldots,k$. Тогда для нахождения асимптотики интеграла (2.2) нужно найти точку (y^*,u^*), в которой достигается минимакс

$$L(x)=\max_{y}\min_{r_m}\max_{\psi_m}\operatorname{Re}\,H(u,x,y). \tag{3.2}$$

Теорема 1. *Минимакс в* (3.2) $L(x)\equiv 1$ $\forall x$ *и достигается при*

$$u^*_{m,m+1}=1,\quad m=1,\ldots,k;\quad y_i^*(x_i)=F_2^{-1}(F_1(x_i)),\quad i=1,\ldots,k,$$

где

$$F_1(x)=\int_{-\infty}^{x}\int_{-\infty}^{+\infty}f(x,y)dx\,dy,\qquad F_2(y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{y}f(x,y)dx\,dy.$$

Доказательство. Найдем сначала $\max\limits_{\psi_m}\operatorname{Re}H(u,x,y)$. Имеем

$$\operatorname{Re}H(u,x,y)=\operatorname{Re}\Big(g+\sum_{i<j}(p_{ij}r_i\ldots r_{j-1}\exp\left(i(\psi_1+\cdots+\psi_{j-1})\right)+$$
$$+p_{ji}r_i^{-1}\ldots r_{j-1}^{-1}\exp(-i(\psi_1+\cdots+\psi_{j-1}))\Big)=$$
$$=g+\sum_{i<j}(p_{ij}r_i\ldots r_{j-1}\cos(\psi_1+\cdots+\psi_{j-1})+p_{ji}r_i^{-1}\ldots r_{j-1}^{-1}\cos(\psi_1+\cdots+\psi_{j-1})),$$

и максимум достигается при $\psi_1=\psi_2=\cdots=\psi_k=0$.

В точке минимакса $\frac{\partial H}{\partial y_1} = 0$, поэтому

$$\frac{\partial H}{\partial y_1} = -(1-r_1)\int_{-\infty}^{x_1} f(x,y_1)dx + \left(1-\frac{1}{r_1}\right)\int_{x_1}^{x_2} f(x,y_1)dx +$$
$$+\left(1-\frac{1}{r_1 r_2}\right)\int_{x_2}^{x_3} f(x,y_1)dx + \cdots + \left(1-\frac{1}{r_1\ldots r_{k-1}}\right)\int_{x_k}^{+\infty} f(x,y_1)dx -$$
$$-\left(1-\frac{1}{r_2}\right)\int_{x_2}^{x_3} f(x,y_1)dx - \left(1-\frac{1}{r_2 r_3}\right)\int_{x_3}^{x_4} f(x,y_1)dx - \cdots - \left(1-\frac{1}{r_2\ldots r_{k-1}}\right)\int_{x_k}^{+\infty} f(x,y_1)dx =$$
$$= (r_1-1)\Bigg(\int_{-\infty}^{x_1} f(x,y_1)dx + \frac{1}{r_1}\int_{x_1}^{x_2} f(x,y_1)dx + \frac{1}{r_1 r_2}\int_{x_2}^{x_3} f(x,y_1)dx + \cdots +$$
$$+\frac{1}{r_1\ldots r_{k-1}}\int_{x_k}^{+\infty} f(x,y_1)dx\Bigg) = 0.$$

Отсюда $r_1 = 1$. Аналогично, из условия $\frac{\partial H}{\partial y_i} = 0$ следует $r_i = 1$, $i = 2,\ldots,k$.

Итак, минимакс достигается при $u^*_{m,m+1} = 1$, $m = 1,\ldots,k$, поэтому $\frac{\partial H}{\partial u_{m,m+1}}\big|_{u^*=1} = 0$. Значит,

$$\frac{\partial H}{\partial u_{m,m+1}}\bigg|_{u^*=1} = \sum_{i<j}(p_{ij}-p_{ji}) = \int_{-\infty}^{x_m}\int_{y_m}^{+\infty} f(x,y)dx\,dy - \int_{x_m}^{+\infty}\int_{-\infty}^{y_m} f(x,y)dx\,dy = 0.$$

Поскольку $F_1(x)$ и $F_2(y)$ — возрастающие непрерывные функции, то $\forall x_m \in R\ \exists z \in (0,1)$: $F_1(x_m) = z$, а для $z\ \ \exists y_m : F_2(y_m) = z$. Следовательно, $F_1(x_m) = F_2(y_m)$, а $L(x) = 1$. □

Теперь можно получить асимптотику интеграла (2.2) (см. [6], с. 418, предложение 4.1, с. 122, теорема 4.1) в виде

$$k!\int\limits_{x_1<\cdots<x_k}\cdots\int \frac{\prod\limits_i f(x_i,y_i^*)dx_i}{\sqrt{\det H(z,x,y)''_{zz}\big|_{\substack{z=1\\y=y^*}}\det H(r^*,x,y)''_{xy}\big|_{\substack{r^*=1\\y=y^*}}}}, \tag{3.3}$$

где $r^*(x,y)$ — радиус перевального контура, а $y^* : F_1(x_m) = F_2(y_m^*)\ \forall m$. Заметим, что в точке $(r^*=1,\ y=y^*)$ $p_{ij} = p_{ji}$ для любых i и j.

4. Выражение для DN

Получим выражение $DN = MN^2 - (MN)^2 = nM\chi_1 + n(n-1)M\chi_1\chi_2 - n^2(M\chi_1)^2$. Найдем $\lim\limits_{n\to+\infty} n(n-1)M\chi_1\chi_2$. В соответствии с (3.3) имеем

$$n^2 M\chi_1\chi_2 \sim 2!\iint\limits_{x_1<x_2}\frac{f(x_1,y_1^*(x_1))f(x_2,y_2^*(x_2))dx_1dx_2}{\sqrt{\det H(z,x,y)''_{zz}\big|_{\substack{z=1\\y=y^*}}\det H(r^*,x,y)''_{xy}\big|_{\substack{r^*=1\\y=y^*}}}}.$$

Теорема 2. *Имеет место равенство*

$$\sqrt{\det H(z,x,y)''_{zz}\big|_{\substack{z=1\\y=y^*}}\det H(r^*,x,y)''_{xy}\big|_{\substack{r^*=1\\y=y^*}}} = f_2(y_1^*)f_2(y_2^*) > 0. \tag{4.1}$$

Доказательство. Для $k = 2$ имеем

$$H = g + p_{12}z_1 + p_{21}\frac{1}{z_1} + p_{23}z_2 + p_{32}\frac{1}{z_2} + p_{13}z_1z_2 + p_{31}\frac{1}{z_1z_2}$$

и

$$\det H(z,x,y)''_{zz}\big|_{\substack{z=1\\y=y^*}} = 4(p_{12}p_{23} + p_{12}p_{13} + p_{23}p_{13}).$$

Для получения второго определителя найдем производные вида $(r_i^*)'_{y_j}$ в точке $(r^*=1,\ y=y^*)$. Так как r^* — радиус перевального контура, то $H(r^*,x,y)'_{r_i^*}=0$, т. е.

$$p_{12}-\frac{p_{21}}{(r_1^*)^2}+p_{13}r_2^*-\frac{p_{31}}{(r_1^*)^2r_2^*}=0,\quad p_{23}-\frac{p_{32}}{(r_2^*)^2}+p_{13}r_2^*-\frac{p_{31}}{(r_2^*)^2r_1^*}=0.$$

Продифференцируем каждое из этих уравнений по y_1

$$f_2(y_1)+2r^{*'}_{1\,y_1}(p_{12}+p_{13})+2p_{13}r^{*'}_{2y_1}=0;\quad 2r^{*'}_{1\,y_1}(p_{23}+p_{13})+2p_{13}r^{*'}_{1\,y_1}=0.$$

Отсюда

$$r^{*'}_{1y_1}=-\frac{f_2(y_1)}{2}\frac{p_{23}+p_{13}}{p_{23}p_{13}+p_{12}p_{13}+p_{23}+p_{12}};\quad r^{*'}_{2y_1}=\frac{f_2(y_1)}{2}\frac{p_{13}}{p_{23}p_{13}+p_{12}p_{13}+p_{23}+p_{12}}. \tag{4.2}$$

Аналогично,

$$r^{*'}_{1y_2}=\frac{f_2(y_2)}{2}\frac{p_{13}}{p_{23}p_{13}+p_{12}p_{13}+p_{23}p_{12}},\quad r^{*'}_{2y_2}=-\frac{f_2(y_2)}{2}\frac{p_{13}+p_{12}}{p_{23}p_{13}+p_{12}p_{13}+p_{23}p_{12}}. \tag{4.2$'$}$$

Введем обозначения $r^{*'}_{iy_j}=a_{ij}$; $(p_{12}-p_{21})'_{y_j}=b_{1j}$; $(p_{13}-p_{31})'_{y_j}=b_{2j}$; $(p_{23}-p_{32})'_{y_j}=b_{3j}$. Учитывая, что $b_{21}=-b_{31}=f_2(y_1)-b_{11}$; $b_{22}=-b_{12}=f_2(y_2)-b_{32}$, имеем

$$\begin{aligned}
H''_{y_1y_1}&=2a_{11}b_{11}+2a_{21}b_{31}+2a_{11}^2p_{12}+2a_{21}^2p_{23}+2b_{21}(a_{11}+a_{21})+2p_{13}(a_{11}+a_{21})^2=\\
&\qquad=2a_{11}f_2(y_1)+2a_{11}^2p_{12}+2a_{21}^2p_{23}+2p_{13}(a_{11}+a_{21})^2,\\
H''_{y_2y_2}&=2a_{22}f_2(y_2)+2a_{12}^2p_{12}+2a_{22}^2p_{23}+2p_{13}(a_{12}+a_{22})^2,\\
H''_{y_1y_2}&=a_{12}f_2(y_1)+a_{12}^2f_2(y_2)+2a_{11}a_{12}p_{12}+2a_{21}a_{22}p_{23}+2p_{13}(a_{12}+a_{22})(a_{11}+a_{21}).
\end{aligned} \tag{4.3}$$

Подставив (4.2), (4.2′) в (4.3), получим

$$H''_{y_1y_1}H''_{y_2y_2}-(H''_{y_1y_2})^2=f_2^2(y_1)f_2^2(y_2)\left(a_{11}a_{22}\frac{1}{f_2(y_1)f_2(y_2)}-\left(\frac{a_{12}}{f_2(y_2)}+\frac{a_{21}}{f_2(y_1)}\right)^2\right)=$$
$$=f_2^2(y_1)f_2^2(y_2)\frac{1}{4(p_{12}p_{13}+p_{12}p_{23}+p_{23}p_{13})},$$

отсюда вытекает (4.1). □

Теперь асимптотика $M\chi_1\chi_2$ значительно упрощается. При $n\to+\infty$ имеем

$$n^2M\chi_1\chi_2\sim 2\iint\limits_{x_1<x_2}\frac{f(x_1,y_1^*)f(x_2,y_2^*)}{f_2(y_1^*)f_2(y_2^*)}dx_1dx_2=$$
$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{f(x_1,y_1^*)f(x_2,y_2^*)}{f_2(y_1^*)f_2(y_2^*)}dx_1dx_2=\left(\int_{-\infty}^{+\infty}\frac{f(x_1,y_1^*)}{f_2(y_1^*)}dx_1\right)^2=(nM\chi_1)^2.$$

В итоге

$$DN\sim MN\sim\int_{-\infty}^{+\infty}\frac{f(x_1,y_1^*)}{f_2(y_1^*)}dx_1=\lambda>0.$$

5. Вид распределения числа совпадений

Предложение ([7], с. 265). *Пусть случайные величины $\chi_1, \ldots, \chi_n$ принимают только значения 0 и 1 и симметрично зависимы. Тогда*

1. *моменты $C_{nk} = M(\chi_1 \cdots \chi_k)$ образуют вполне монотонную последовательность (т. е. знаки конечных разностей элементов последовательности чередуются) для каждого n;*
2. *существует случайная величина λ_n, распределение которой сосредоточено на решетке j/n, $j = 0, \ldots, n$, такая, что $M\lambda_n^k = C_{nk}$;*
3. *случайные величины N_n распределены по рандомизированному биномиальному закону*

$$P(N_n = k) = \sum_{l=0}^{n} C_n^k \left(\frac{l}{k}\right)^k \left(1 - \frac{l}{k}\right)^{n_k} P\left(\lambda_n = \frac{l}{k}\right).$$

Теорема 3. *Пусть N_n — число совпадений для объема выборки n,*

$$C_{nk} = M(\chi_1 \cdots \chi_k), \quad S_k = \lim_{n \to +\infty} n^k C_{nk}.$$

Тогда

а) *существует случайная величина λ, распределение которой сосредоточено на решетке j/n, $j = 0, \ldots, n$, такая, что $M\lambda_n^k = C_{nk}$, и $p(N_n = k) = \sum\limits_{l=0}^{n} C_n^k l^k (1-l)^{n-k} P(\lambda_n = \frac{l}{k})$;*

б) *для любого частичного слабого предела N последовательности N_n существует $\mu \stackrel{\mathrm{D}}{=} \lim\limits_{n \to +\infty} n\lambda_n : M\mu^k = S_k$ и распределение N — рандомизированное по параметру распределение Пуассона*

$$P(N = k) = \sum_{l=0}^{+\infty} \frac{e^{-l} l^k}{k!} P(\mu = l).$$

Доказательство. Легко видеть, что утверждение а) прямо следует из пп. 2, 3 предложения, т. к. $\chi_1, \chi_2, \ldots, \chi_k$ симметрично зависимы.

б) Пусть $\mu_n = n\lambda_n$, F_n — функция распределения μ_n. Легко показать, что последовательность μ_n стохастически ограничена. Действительно, из неравенства Чебышева

$$P(\mu_n \geq t) \leq \frac{1}{t^2} M(\mu_n^2).$$

Для фиксированного $\varepsilon^*(n > n^*) \to P(\mu_n \geq t) \leq \frac{1}{t^2}(S_2 + \varepsilon^*)$. Тогда $\forall(\varepsilon > 0,\ n > n^*)$ существует $a = \sqrt{S_2 + \varepsilon^* / \varepsilon}$: $P(\mu_n \geq a) \leq \varepsilon$, что и означает стохастическую ограниченность последовательности μ_n. Теперь из теоремы Хелли о выборе ([7], с. 307) ясно, что если последовательность F_n и не имеет предела, то все ее частичные пределы будут функциями распределения и будут обладать тем свойством, что их k-й момент равен S_k. Выберем любой из этих пределов. Пусть μ – случайная величина, ему соответствующая. Далее утверждение б) становится очевидным. □

6. Асимптотическое распределение числа совпадений

Для нахождения предельного распределения числа совпадений воспользуемся методом моментов. Прежде всего методом математической индукции выведем рекуррентные соотношения для коэффициентов c_{jk} в (2.1).

При $k = 1$, очевидно, $c_{11} = 1$. Предположим, что известны c_{jk}, $j = 1, \ldots, k$, и

$$MN_n^{k+1} = M\left(\sum_{j=1}^{k} c_{jk} \sum_{i_1 \neq i_2 \neq \cdots \neq i_j} \chi_{i_1} \cdots \chi_{i_j}\right)\left(\sum_{i=1}^{n} \chi_i\right) =$$

$$= M\left(\sum_{j=1}^{k} c_{jk} \sum_{i_1 \neq \cdots \neq i_{j+1}} \chi_{i_1} \cdots \chi_{i_{j+1}} + \sum_{j=1}^{k} j c_{jk} \sum_{i_1 \neq \cdots \neq i_j} \chi_{i_1} \cdots \chi_{i_j}\right).$$

Таким образом, $c_{1,k+1} = 1$, $c_{j,k+1} = c_{j-1,k} + jc_{jk}$.

Предположим, что c_{jk} суть коэффициенты полинома $\Phi_k(\lambda) = \sum\limits_{j=1}^{k} c_{jk}\lambda^k$, тогда

$$\begin{gathered}\Phi_1(\lambda) = 1,\\ \Phi_{k+1}(\lambda) = \lambda(\Phi_k(\lambda) + \Phi_k'(\lambda)), \quad k = 2, 3, \ldots\end{gathered} \tag{6.1}$$

Из (2.1) и (3.3) имеем $\lim\limits_{n\to\infty} MN_n^k = \sum\limits_{j=1}^{k} c_{jk}\lambda^j = \Phi_k(\lambda)$, где $\lambda = \lim\limits_{n\to\infty} nM\chi_1$.

Начальные моменты случайной величины W, подчиненной распределению Пуассона, удовлетворяют (6.1). Действительно, $MW = \lambda = \Phi_1(\lambda)$. Предположим $MW^k = \Phi_k(\lambda)$. Это означает, что $\phi^{(k)}(t) = (Me^{itw})^{(k)} = e^{\lambda(e^{it}-1)}(\Phi_k(\lambda e^{it}))i^k = e^{\lambda(e^{it}-1)} \sum\limits_{j=1}^{k} c_{jk}\lambda^j e^{ijt} i^k$. Дифференцируя последнее равенство, имеем $\phi^{(k+1)}(t) = e^{\lambda(e^{it}-1)}\Big(\sum\limits_{j=1}^{k} \lambda e^{it} i c_{jk}\lambda^j e^{ijt} + ij c_{jk}\lambda^j e^{ijt}\Big) i^k = e^{\lambda(e^{it}-1)}\Phi_{k+1}(\lambda e^{it}) i^{k+1}$. Следовательно, $MW^{k+1} = \Phi_{k+1}(\lambda)$.

Так как распределение Пуассона однозначно определяется своими моментами, предельное распределение N_n пуассоновское.

Заметим, что при $\lambda < \infty$ этот результат следует из теоремы 3, поскольку

$$M\mu = \lim_{n\to\infty} M\mu_n = \lambda < +\infty,$$
$$M\mu_n^2 = (M\mu_n)^2 + \mathrm{o}((M\mu_n)^2) = (M\mu_n)^0 + \mathrm{o}(1),$$

т. к. $n^2 M\chi_1\chi_2 \sim (nM\chi_1)^2$ и $D\mu = \lim\limits_{n\to\infty} D\mu_n = 0$.

Покажем, что при $\lambda \to \infty$ распределение числа совпадений N асимптотически нормально. Пусть $\{W_n\}$ — последовательность случайных величин, подчиненных закону Пуассона с параметрами $\lambda_n = nM\chi_1$. Случайные величины N_n и W_n определены на одном и том же вероятностном пространстве $\mathbf{Z}$ и сходятся по распределению к одной и той же случайной величине W, поэтому $N_n - W_n \xrightarrow[n\to 0]{p} 0$. Имеем $N_n = W_n + (N_n - W_n)$, поскольку при $\lambda \to \infty$ распределение W_n асимптотически нормально, таким же является и предельное распределение N_n, т.е.

$$P\left(\frac{N_n - MN_n}{\sqrt{DN_n}} < x\right) \to \Phi(x).$$

Литература

1. De Groot M.H., Feder P.I., Goel P.K. *Matchmaking* // Ann. Math. Stat. – 1971. – V. 42. – №. 2. – P. 578–593.
2. Chew M.C. *On pairing observations from a distribution with monotone likehood ratio* // Ann. Math. Stat. – 1973. – V. 1. – №3. – P. 433–445.
3. De Groot M.H., Goel P.K. *The matching problem for multivariate normal data* // Sankhya. – 1976. – V. 38. – № 1. – P. 14–29.
4. Золотухина Л.А., Латышев К.П. *Асимптотическое представление среднего числа совпадений элементов вариационных рядов двумерной выборки* // Зап. научн. семин. ЛОМИ. – 1978. – Т. 79. – С. 4–10.
5. Латышев К.П., Золотухина Л.А. *Интегральное представление дисперсии числа правильно угаданных пар в задаче о совпадении двумерной выборки* // Зап. научн. семин. ЛОМИ. – 1982. – Т. 136. – С. 113–120.
6. Федорюк М.В. *Асимптотика. Интегралы и ряды.* – М.: Наука, 1987. – 361 с.
7. Феллер В.И. *Введение в теорию вероятностей и ее применения.* Т. 2. – М.: Мир, 1984. – 751 с.

Санкт-Петербургский государственный морской технический университет

Поступили
первый вариант 11.04.1997
окончательный вариант 29.10.1998