EESTI NSV TEADUSTE AKADEEMIA TOIMETISED. XIV KÖIDE
FÜÜSIKA-MATEMAATIKA- JA TEHNIKATEADUSTE SEERIA. 1965, NR. 1

ИЗВЕСТИЯ АКАДЕМИИ НАУК ЭСТОНСКОЙ ССР. ТОМ XIV
СЕРИЯ ФИЗИКО-МАТЕМАТИЧЕСКИХ И ТЕХНИЧЕСКИХ НАУК. 1965, № 1

https://doi.org/10.3176/phys.math.tech.1965.1.03

А. НИЛЬСОН

НЕКОТОРЫЕ СВОЙСТВА СУММ КВАДРАТОВ ВЕРОЯТНОСТЕЙ И ИХ МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ ПРИЛОЖЕНИЯ

В работе вводится класс статистик, для построения которых используются свойства сумм квадратов вероятностей случайных событий. В приложениях математико-статистического анализа установленные статистики могут заменить многие статистики, характеризующие энтропию, сопряженность, рассеивание и зависимость в системах случайных событий или величин. Введенные статистики имеют ясное вероятностное содержание и исключительно легко вычисляются как аналитически, так и методом статистических испытаний. Свойства сумм квадратов вероятностей могут быть применены для анализа целесообразности кодов в информационно-поисковых системах. Указывается на некоторые возможности упрощения вычисления статистик нелинейной связи.

1. Некоторые свойства сумм квадратов вероятностей, k-разбросанность и ее связь с энтропией и дисперсией

Пусть имеется полная система событий $A_i = \begin{Bmatrix} A_1, \ldots, A_h \\ p_1, \ldots, p_h \end{Bmatrix}$ и такая последовательность независимых опытов, что $P(A_i^1) = P(A_i^2) = \ldots = P(A_i^n) = p_i$, где верхние индексы обозначают порядковый номер опыта. Тогда при $1 \leqslant l,\ m \leqslant n$ и $l \neq m$ имеем $P(A_i^l A_i^m) = p_i^2$ и в силу несовместимости событий A_i

$$P\left(\sum_{i=1}^{h} A_i^l A_i^m\right) = \sum_{i=1}^{h} p_i^2. \tag{1}$$

Итак, сумма квадратов вероятностей всех событий полной системы событий выражает вероятность одинаковых результатов двух независимых опытов, вероятности разных исходов которых совпадают с вероятностями событий в упомянутой системе. Совершенно очевидно, что такая вероятность тесно связана с неопределенностью (энтропией) системы.

Ввиду сказанного, а также исключительной простоты вероятностного смысла и вычисления сумм квадратов вероятностей событий в полной системе событий, эта сумма во многих случаях может успешно заменить энтропию или даже дисперсию распределения. Связь с энтропией (информацией) второго порядка $I_2(P)$, вводимой Рени [10], очень тесна

$$\left(\sum_{i=1}^{h} p_i^2\right)^{-1} = 2^{I_2(P)}, \text{ где } I_2(P) = -\log_2 \sum_{i=1}^{h} p_i^2. \tag{2}$$

Связь с энтропией $H(x)$, вычисленной по формуле Шеннона, будет установлена ниже [см. выражения (13), (14)].

Укажем на некоторые простые свойства суммы квадратов вероятностей полной системы событий. Имеет место равенство

$$\sum_{i=1}^{h} p_i^2 = \frac{1}{h} + \sum_{i=1}^{h} \left(p_i - \frac{1}{h}\right)^2, \tag{3}$$

откуда легко получим следующие свойства *а*, *б* и *в*:

а) $\frac{1}{h} \leqslant \sum_{i=1}^{h} p_i^2 \leqslant 1$, при этом $\sum_{i=1}^{h} p_i^2 = \frac{1}{h}$ только при

$$p_i = \frac{1}{h},\ i = 1, 2, \ldots, h \text{ и } \sum_{i=1}^{h} p_i^2 = 1 \text{ только } p_i = 0,$$

$$i = 1, 2, \ldots, h,\ i \neq k, \text{ а } p_i = 1,\ i = k\ (1 \leqslant k \leqslant h);$$

б) $\sum_{i=1}^{h} p_i^2 \to 0$, если $\lim \max_i p_i = 0$;

в) если часть вероятностей $p_i > 0$, $i = 1, 2, \ldots, h$, $i \neq k, \ldots, l$, а остальные $p_i \to 0$, $i = k, \ldots, l$ $(1 \leqslant k < l \leqslant h)$, то $\sum_{i=1,\ i \neq k,\ldots,l}^{} p_i^2 \to \sum_{i=1}^{h} p_i^2$;

г) если в данной полной системе $A_i = \begin{Bmatrix} A_1, \ldots, A_h \\ p_1, \ldots, p_h \end{Bmatrix}$, исключив некоторые события, перейти к новой полной системе $C_j = \begin{Bmatrix} C_r, \ldots, C_s \\ q_r, \ldots, q_s \end{Bmatrix}$ где $1 \leqslant r < \ldots < s \leqslant h$ и при $i = j$, $C_j = A_i$, $q_j = p_i \left(\sum_{i=r}^{s} p_i\right)^{-1}$, то для новой полной системы получим

$$\sum_{j=r}^{s} q_j^2 = \sum_{i=r}^{s} p_i^2 \left(\sum_{i=r}^{s} p_i\right)^{-2}. \tag{4}$$

Свойством *г* удобно пользоваться для анализа систем по несовместным частям.

Обозначим в дальнейшем

$$\left(\sum_{i=1}^{h} p_i^2\right)^{-1} = k, \text{ где } k \leqslant h. \tag{5}$$

Если k — натуральное число, то по свойству *а* мы можем сопоставить с данной полной системой h неравновероятных событий систему k равновероятных событий $B_j = \begin{Bmatrix} B_1, \ldots, B_k \\ q_1, \ldots, q_k \end{Bmatrix}$, где $q_1 = q_2 = \ldots = q_k = \frac{1}{k}$

и $\sum_{i=1}^{h} p_i^2 = \sum_{j=1}^{k} q_j^2 = \frac{1}{k}$. Если же k — не целое число, то оно и тогда позволяет характеризовать разбросанность вероятностей (распределение вероятностей по отдельным событиям) в системе событий. В дальнейшем будем величину k называть квадратичной разбросанностью вероятностей (сокращенно k-разбросанностью).

Статистика χ^2 данного распределения вероятностей p_i, $i = 1, 2, \ldots, h$, относительно равномерного распределения вероятностей $p_i' = 1/h$, $i = 1, 2, \ldots, h$, выражается через k и h:

$$\chi^2_{\nu,\lambda^2} = \sum_{i=1}^{h} \frac{(Np_i - Np_i')^2}{Np_i'} = N\left(\frac{h}{k} - 1\right), \tag{6}$$

где число степеней свободы $\nu = h - 1$ и параметр нецентральности $\lambda^2 = N\left(\frac{h}{k} - 1\right)$ (см. [6], стр. 418).

Рассмотрим далее связь k-разбросанности и среднего квадратического отклонения распределения случайной величины. Пусть случайная величина $x/\sigma = t$ (где σ — среднее квадратическое отклонение случайной величины x) имеет непрерывное распределение с плотностью вероятностей $f(t)$. При усеченном распределении в промежутке $(a/\sigma, b/\sigma)$ для каждого натурального числа h определяется полная система событий $A^{(h)} = \{A_1^{(h)}, \ldots, A_h^{(h)}\}$, где $A_i^{(h)}$ — событие $\frac{x}{\sigma} \in \left[\frac{a}{\sigma} + \frac{(b-a)(i-1)}{\sigma h}, \frac{a}{\sigma} + \frac{(b-a)i}{\sigma h}\right]$. Обозначая $(b-a)/h = \Delta_h$, мы можем вероятности событий вычислять по приближенной формуле

$$p(A_i^{(h)}) = \frac{\frac{\Delta_h}{\sigma} f\left(\frac{a + i\Delta_h}{\sigma}\right)}{\sum_{i=1}^{h} \frac{\Delta_h}{\sigma} f\left(\frac{a + i\Delta_h}{\sigma}\right)}.$$

Для каждой системы $A^{(h)}$ вычисляем следующее отношение:

$$\frac{h}{k_h} = h \sum_{i=1}^{h} p^2(A_i^{(h)}).$$

В пределе $h \to \infty$ имеем

$$\lim \frac{h}{k_h} = \lim \frac{h\left(\frac{\Delta_h}{\sigma}\right)^2 \sum_{i=1}^{h} f^2\left(\frac{a + i\Delta_h}{\sigma}\right)}{\left[\sum_{i=1}^{h} f\left(\frac{a + i\Delta_h}{\sigma}\right)\frac{\Delta_h}{\sigma}\right]^2} = \frac{h\Delta_h \int_{a'}^{b'} f^2(t)\,dt}{\sigma\left(\int_{a'}^{b'} f(t)\,dt\right)^2}, \tag{7}$$

где $a' = a/\sigma$ и $b' = b/\sigma$.

В частности для симметрически усеченного нормального распределения $N(0, \sigma)$, где σ — среднее квадратическое отклонение неусеченного распределения, имеем

$$\lim \frac{h}{k_h} = \frac{h\Delta_h \int_0^{b'} e^{-t^2} dt}{2\sigma \left(\int_0^{b'} e^{-\frac{t^2}{2}} dt \right)^2}, \tag{8}$$

или по табулированным функциям $\Phi(x) = \sqrt{\frac{2}{\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$ и

$\operatorname{erf} x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ получим

$$\lim \frac{h}{k_h} = \frac{h\Delta_h \operatorname{erf} b'}{2\sigma\sqrt{\pi}\,\Phi^2(b')} = \frac{h\Delta_h \Phi(b'\sqrt{2})}{2\sigma\sqrt{\pi}\,\Phi^2(b')}. \tag{8а}$$

В целях выяснения влияния величины Δ_h (грубости группировки) на отношение h/k_h при нормальном распределении по данным таблиц $\operatorname{erf} x$ нами вычислялись $\lim(h/k_h)$ и значения $h/k_h = h \sum_{i=1}^{h} \{\operatorname{erf}(\Delta_h i/\sigma) - \operatorname{erf}[\Delta_h(i-1)/\sigma]\}^2 \operatorname{erf}^{-2}(b')$ для разных величин Δ_h при $b_1' = 3{,}2\sqrt{2}$ и $b'_2 = 4{,}8\sqrt{2}$. Таким образом, мы смогли составить эмпирическую формулу $\lim \frac{h}{k_h} \approx \frac{2k_h h}{2k_h^2 - 1} = \frac{h}{k}$, откуда

$$k = k_h - \frac{1}{2k_h}, \tag{9}$$

где k — исправленная оценка k_h. При $\Delta_h \leqslant \sigma$ имеем

$$100 \left| \frac{h}{k} - \lim \frac{h}{k_h} \right| \left(\lim \frac{h}{k_h} \right)^{-1} < 0{,}5\%. \tag{10}$$

При $\Delta_h > \sigma$ ошибка быстро увеличивается. Из выражений (9) и (10) получим $\frac{h}{k} \approx \lim \frac{h}{k_h}$ и, обозначая $\frac{\Phi(b'\sqrt{2})}{\Phi^2(b')} = G(b')$, по (8а) имеем

$$\frac{h}{k} \approx \frac{h\Delta_h}{2\sigma\sqrt{\pi}} G(b') = \frac{b-a}{2\sigma\sqrt{\pi}} \cdot G(b') \tag{11}$$

или

$$\sigma \approx \frac{k \cdot \Delta_h}{2\sqrt{\pi}} G(b'). \tag{11а}$$

Если промежуток (a', b') выбрать так, что $G(b') \to 1$, то имеем

$$\sigma \approx \frac{k\Delta_h}{2\sqrt{\pi}} \approx 0{,}282\, k\Delta_h \tag{11б}$$

или без поправки на грубость группировки

$$\sigma \approx 0{,}282\, k_h \Delta_h, \tag{11в}$$

а с поправкой Шеппарда приблизительно

$$\sigma \approx 0{,}282\, \Delta_h \sqrt{k_h^2 - 1}. \tag{11г}$$

Подобные же отношения между k-разбросанностью и средним квадратическим отклонением можно получить из выражения (7) и для ряда других типов распределений.

Если $\frac{h}{k} - 1 \to 0$, т. е. распределение по критерию χ^2 (6) близко к равномерному, то, зная, что дисперсия равномерного распределения $\sigma^2 = \frac{(b-a)^2}{12}$, по свойству a можем установить оценку

$$\sigma \approx \frac{k\Delta_h}{2\sqrt{3}} \approx 0{,}289\, k\Delta_h. \tag{12}$$

При равномерном распределении из свойства a и из свойств энтропии $H(X)$, вычисленной по формуле Шеннона, легко получаем

$$H(X) = \log k = \log h, \tag{13}$$

а в случае нормального распределения из выражения (11в) и из выражения энтропии нормального распределения $H(X) = \log\left(\frac{\sqrt{2\pi e}\sigma}{\Delta_h}\right)$ (см. [3], стр. 487) имеем

$$H(X) = \log k \sqrt{\frac{2}{e}} \text{ или } \log k = H(X) \sqrt{\frac{e}{2}}. \tag{14}$$

Применение и вычисление k-разбросанности мыслимо, если общее число испытаний $N \gg h \geqslant 2$. При выборе h можно воспользоваться практической рекомендацией $h \approx \log_2 N + 1$, которая принята в информационных методах корреляционного анализа [4]. Исходя из условия $N \gg h$, результатов (9) и (10) и из вычислительных соображений, в большинстве случаев оказываются полезными значения $20 \geqslant h \geqslant 5$, но применимы и значения $5 > h > 20$.

Если непрерывное распределение представлено в дискретной форме и каждый интервал представлен своим средним значением x_i, то имеем $a = x_1 - \Delta_h/2$, $b = x_h + \Delta_h/2$, откуда $\Delta_h = (x_h - x_1)/(h-1)$. Если исходить из окончательно дискретной модели распределения, то построенные по свойству a равномерно распределенные аналоги исходных распределений, очевидно, также должны быть окончательно дискретными и поскольку размах их в таком случае вместо $k\Delta_h$ равен $(k-1)\Delta_h$, мы можем в формулах (11) — (12) вместо k поставить $(k-1)$.

2. Некоторые обсуждения k-разбросанности сложных систем

Пусть имеется сложная система $AB = \begin{Bmatrix} A_1 B_1, \ldots, A_h B_r \\ p_{11}, \ldots, p_{hr} \end{Bmatrix}$, которая состоит из полных систем случайных событий.

$$A = \begin{Bmatrix} A_1, \ldots, A_h \\ p_1, \ldots, p_h \end{Bmatrix} \quad \text{и} \quad B = \begin{Bmatrix} B_1, \ldots, B_r \\ q_1, \ldots, q_r \end{Bmatrix}.$$

Вводим обозначения

$$k_A = \left(\sum_{i=1}^{h} p_i^2\right)^{-1}, \quad k_B = \left(\sum_{j=1}^{r} q_j^2\right)^{-1} \quad \text{и} \quad k_{AB} = \left(\sum_{i=1}^{h}\sum_{j=1}^{r} p_{ij}^2\right)^{-1}.$$

Тогда в случае полной независимости систем A и B по теореме умножения вероятностей получаем

$$k_{AB} = k_A k_B, \tag{15}$$

а в случае взаимно однозначной зависимости

$$k_{AB} = k_A = k_B = \sqrt{k_A k_B}. \tag{15а}$$

Из определения условной вероятности $q_{j/i} = p_{ij}/p_i$ получаем $p_{ij}^2 = p_i^2\, q_{j/i}^2$ и $\sum_i \sum_j p_{ij}^2 = \sum_i \sum_j p_i^2 q_{j/i}^2$. Если $\sum_j q_{j/i}^2 = \text{const}$ при $i = 1, 2, \ldots, h$, то из последнего получаем $\sum_i \sum_j p_{ij}^2 = \sum_i p_i^2 \sum_j q_{j/i}^2$ или, обозначая $\left(\sum_j q_{j/i}^2\right)^{-1} = k_{B/A}$, имеем

$$k_{AB} = k_A\, k_{B/A} = k_B\, k_{A/B}. \tag{15б}$$

Выражение (15б) имеет силу всегда, когда выполняется условие $\sum_j q_{j/i}^2 = \text{const}$ (например, при двухмерном нормальном распределении). Если $\sum_j q_{j/i}^2 \neq \text{const}$, то вместо k_{B/A_i} можем использовать математическое ожидание условной k-разбросанности $k_{B/A} = M(k_{B/A}) = \sum_i p_i k_{B/A_i}$. Однако аналитическое вычисление k_{AB} значительно проще, чем $\sum_i p_i k_{B/A_i}$, и $k_{B/A}$, вычисленное по формуле (15б), является обратной величиной от $(k_{B/A})^{-1}$, усредненной по весам $p_i^2 / \sum_i p_i^2$, поэтому в приближенных расчетах для вычисления оценки $k_{B/A}$ полезно применить (15б).

Те же результаты (15), (15а) и (15б) можем легко получить из свойств энтропии, используя взаимоотношения k-разбросанности и энтропии [см. выражения (2), (13) и (14)].

Таким образом каждой оценке энтропии мы можем сопоставить соответствующие оценки k-разбросанности. Ниже приводим сводную таблицу таких соотношений в сложной системе $\{AB\}$ с соответствующими оценками k-разбросанностей по вероятностям и по частотам (табл. 1).

Обозначения частот в табл. 1: N — общее число наблюдений; n_{ij} — число наблюдений с признаком $A_i B_j$; $n_{i\cdot} = \sum_j n_{ij}$ — число наблюдений с признаком A_i; $n_{\cdot j} = \sum_i n_{ij}$ — число наблюдений с признаком B_j.

Оценки k-разбросанностей по вероятностям в табл. 1 являются дефинициями, остальное вытекает из результатов п. 3 или может быть получено из соотношений k-разбросанности и энтропии. Результаты п. 2 при помощи математической индукции легко распространяются на случай сложных систем, состоящих более чем из двух составляющих систем событий.

Приведенные в табл. 1 примеры показывают исключительную простоту вычислений оценок k-разбросанностей непосредственно по наблюденным частотам, а также возможность применения k-разбросанностей для построения статистик, характеризующих зависимость составляющих систем событий в сложной системе. Результаты, полученные в п. 1 относительно отношений k-разбросанности и среднего квадратического отклонения распределений случайных величин, создают предпосылки для построения по k-разбросанности статистик, характеризующих зависимость случайных величин. Вопросы построения таких статистик будут рассмотрены ниже.

Таблица 1

Часть сложной системы	Энтропия	Соответствующие k-разбросанности		
		Обозначение	Оценки по вероятностям	Оценки по частотам
A	$H(A)$	k_A	$\left(\sum_i p_i^2\right)^{-1}$	$N^2\left(\sum_i n_{i.}^2\right)^{-1}$
B	$H(B)$	k_B	$\left(\sum_j q_j^2\right)^{-1}$	$N^2\left(\sum_j n_{.j}^2\right)^{-1}$
$A \cup B$	$H(AB)$	k_{AB}	$\left(\sum_i \sum_j p_{ij}^2\right)^{-1}$	$N^2\left(\sum_i \sum_j n_{ij}^2\right)^{-1}$
$B - A \cap B$	$H(AB) - H(A)$	$k_{B\|A} = \frac{k_{AB}}{k_A}$	$\sum_i p_i^2 \left(\sum_i \sum_j p_{ij}^2\right)^{-1}$	$\sum_i n_{i.}^2 \left(\sum_i \sum_j n_{ij}^2\right)^{-1}$
$A - A \cap B$	$H(AB) - H(B)$	$k_{A\|B} = \frac{k_{AB}}{k_B}$	$\sum_j q_j^2 \left(\sum_i \sum_j p_{ij}^2\right)^{-1}$	$\sum_j n_{.j}^2 \left(\sum_i \sum_j n_{ij}^2\right)^{-1}$
$A \cup B - A \cap B$	$2H(AB) - H(A) - H(B)$	$\frac{k_{AB}^2}{k_A k_B}$	$\sum_i p_i^2 \sum_j q_j^2 \left(\sum_i \sum_j p_{ij}^2\right)^{-2}$	$\sum_i n_{i.}^2 \cdot \sum_j n_{.j}^2 \left(\sum_i \sum_j n_{ij}^2\right)^{-2}$
$A \cap B$	$H(A) + H(B) - H(AB) = I_{A \leftrightarrow B}$	$\frac{k_A k_B}{k_{AB}}$	$\sum_i \sum_j p_{ij}^2 \left(\sum_i p_i^2 \sum_j q_j^2\right)^{-1}$	$N^2 \sum_i \sum_j n_{ij}^2 \left(\sum_i n_{i.}^2 \sum_j n_{.j}^2\right)^{-1}$
$A \cup B + A \cap B$	$H(A) + H(B)$	$k_A k_B$	$\left(\sum_i p_i^2 \sum_j q_j^2\right)^{-1}$	$N^4\left(\sum_i n_{i.}^2 \sum_j n_{.j}^2\right)^{-1}$

3. О некоторых вопросах корреляционного анализа и применении в этих целях k-разбросанности

Наиболее известными статистиками, характеризующими зависимость систем случайных событий или случайных величин, являются:

а) показатель взаимной сопряженности Пирсона [9]

$$C = \sqrt{\frac{\chi^2}{N+\chi^2}} = \sqrt{\frac{\varphi^2}{1+\varphi^2}}, \text{ где } \chi^2 = N \sum_i \sum_j \frac{(p_{ij} - p_i q_j)^2}{p_i q_j}; \quad \varphi^2 = \frac{\chi^2}{N};$$

б) коэффициент взаимной сопряженности Чупрова [8]

$$T^2 = \frac{\varphi^2}{\sqrt{(h-1)(l-1)}};$$

в) информационно-теоретический коэффициент корреляции Линфута $L = \sqrt{1 - e^{-2I_{x \leftrightarrow y}}}$, где количество информации $I_{x \leftrightarrow y} = H(X) + H(Y) - H(X, Y)$, а энтропии $H(X)$, $H(Y)$ и $H(X, Y)$ вычисляются по формуле Шеннона [10];

г) коэффициент ранговой корреляции Спирмена [2]

$$\varrho = 1 - \frac{6\sum d^2}{N(N-1)(N+1)},$$

где d — разность порядковых номеров признаков в одном наблюдении (признаки упорядочены по убыванию значений или качества);

д) коэффициент корреляций

$$r = \frac{M\{(x-\bar{x})(y-\bar{y})\}}{s_x s_y},$$

где в числителе математическое ожидание произведения отклонений случайных величин X и Y и в знаменателе произведение их средних квадратических отклонений;

е) корреляционное отношение Y к X Пирсона [8]

$$\eta_{yx} = \sqrt{1 - \left(\frac{s_{y/x}}{s_y}\right)^2},$$

где $s_{y/x}$ условное среднее квадратическое отклонение Y по X, $s^2_{y/x} = \sum_i p_i s^2_{y/x_i}$.

Известно, что в случае двухмерного нормального распределения при $\Delta_h \to 0$ $\eta_{yx} = \eta_{xy} = r = L = C = 2 \sin \frac{\pi}{6} \varrho$ (см. [2, 4, 9]).

Используя нормальность условных распределений такого распределения и выражение (11) из формул η или r, легко получим их оценки $\varkappa$ по k

$$\varkappa_{yx} = r = \eta_{yx} = \sqrt{1 - \left(\frac{k_{y/x}}{k_y}\right)^2},$$

или по (15б) имеем $\varkappa = \varkappa_{yx} = \varkappa_{xy} = \eta_{yx} = \eta_{xy} = r = L = C = 2 \sin \frac{\pi}{6} \varrho$ и

$$\varkappa = \sqrt{1 - \left(\frac{k_{y/x}}{k_y}\right)^2} = \sqrt{1 - \left(\frac{k_{x/y}}{k_x}\right)^2} = \sqrt{1 - \left(\frac{k_{xy}}{k_x k_y}\right)^2}. \qquad (16)$$

К результату (16) приводит и замена $I_{x \leftrightarrow y}$ в формуле L на $I_{2x \leftrightarrow y} = I_2(P) + I_2(Q) - I_2(P, Q)$, где I_2 — информация второго порядка по

Рени [10], $I_2(P) = -\log_2 \sum_i p_i^2$, $I_2(Q) = -\log_2 \sum_j q_j^2$ и $I_2(P, Q) = -\log_2 \sum_i \sum_j p_{ij}^2$. Тогда имеем $L_2 = \sqrt{1 - 2^{-2I_2, x \leftrightarrow y}}$, откуда получаем $\varkappa = L_2$.

По свойству *а* п. 1 имеем $1 \leqslant k \leqslant h$ и, следовательно, при $\Delta_h^{-1} \to 0$, т. е. при $k^{-1} \to \infty$

$$0 \leqslant \varkappa \leqslant \sqrt{\frac{k_y^2 - 1}{k_y^2}}, \quad \sqrt{\frac{k_x^2 - 1}{k_x^2}}, \quad \sqrt{\frac{k_x k_y - 1}{k_x k_y}} < 1. \tag{17}$$

Для нормирования $\varkappa$ так, чтобы оценка его $\widehat{\varkappa}$ удовлетворяла $0 \leqslant \widehat{\varkappa} \leqslant 1$, вместо выражения (11) используем (11г). При этом имеем

$$\widehat{\varkappa} = \sqrt{\frac{k_y^2 - k_{y/x}^2}{k_y^2 - 1}} = \sqrt{\frac{k_x^2 - k_{x/y}^2}{k_x^2 - 1}} = \sqrt{\frac{k_x^2 k_y^2 - k_{xy}^2}{k_x^2 k_y^2 - k_x k_y}}. \tag{16а}$$

К тем же результатам (16а) ведет умножение $\varkappa$ на обратную величину своего максимального значения по (17). Но, если значение $k_x k_y$ достаточно большое, то разность $1 - \sqrt{\frac{k_x k_y - 1}{k_x k_y}}$ приближается к точности вычисления оценки $\varkappa$ и вместо $\widehat{\varkappa}$ можно взять $\varkappa$.

В случае нелинейности связи, если $k_{y/x_i} \approx \text{const}$ при $i = 1, 2, \ldots, h$ и законы условных распределений не особенно резко различаются (от безусловного распределения), то приближенно имеет силу (15б) и ввиду того, что (11) и (12) имеют силу только тогда, если тип условных распределений близок к типу безусловного распределения (особенно в смысле компактности и числа экстремумов), то обычно имеем

$$\frac{k_{y/x}}{k_y} \approx \frac{k_{x/y}}{k_x} \approx \frac{k_{xy}}{k_x k_y} \approx \min\left(\frac{s_{y/x}}{s_y}, \frac{s_{x/y}}{s_x}\right),$$

откуда вытекает

$$\varkappa \approx \max(\eta_{xy}, \eta_{yx}). \tag{18}$$

Наиболее неблагоприятными для выражения (18) являются двухмерные распределения, где $\eta_{xy} \to 0$ и $\eta_{yx} \gg 0$ (немонотонные зависимости) и где условные распределения резко различного типа. Но даже тогда согласие $\widehat{\varkappa}$ и η оказывается хорошим; обычно $\max \eta - \varkappa < 3s_{\tau_i}$ (например, из табл. 74 [7], где $N = 314$, $h = 11$; $l = 11$; $r = 0{,}00$; $\eta_{xy} = 0{,}00$; $\eta_{yx} = 0{,}51$ и условные распределения резко различны, мы получили $\widehat{\varkappa} = 0{,}42$, т. е. $\eta_{yx} - \widehat{\varkappa} \approx 2s_{\tau_{i/yx}}$).

Результат $\varkappa = L_2$ указан выше. Сходство L и L_2 ($L_2 \approx L$) можно доказать по результатам Рени [10], а в таком случае $\varkappa \approx L$ для любых распределений и приближенно обладает свойствами L, указанными в [5]. Следовательно, $\varkappa$ имеет много общего как с математико-статистическими, так и информационно-теоретическими, параметрическими и непараметрическими статистиками зависимости.

Вычисление оценок C, T^2, φ, r и η довольно трудоемко и при этом r характеризует только линейную зависимость.

Намного легче вычисление оценки L, если $I_{x\leftrightarrow y}$ вычисляется по формуле (см. [4])

$$I_{x\leftrightarrow y}=\frac{1}{N}\left(N\ln N+\sum_i\sum_j n_{ij}\ln n_{ij}-\sum_i n_{i.}\ln n_{i.}-\sum_j n_{.j}\ln n_{.j}\right).$$

Если для вычислений используются таблицы* значений $-p\log_2 p$, то запятую в значениях наблюденных частот n следует передвинуть на α разрядов влево, где α выбирается так, чтобы $10^{\alpha-1}<N\leqslant 10^{\alpha}$. Обозначая $N'=10^{-\alpha}N$ и $n'=10^{-\alpha}n$, имеем

$$I_{x\leftrightarrow y}=\frac{1}{N'}\left(N'\log_2 N'+\sum_i\sum_j n'_{ij}\log_2 n'_{ij}-\sum_i n'_{i.}\log_2 n'_{i.}-\sum_j n'_{.j}\log_2 n'_{.j}\right)$$

и

$$L=\sqrt{1-2^{-2I_{x\leftrightarrow y}}}.$$

Для облегчения вычислений полезно протабулировать $L=f(I_{x\leftrightarrow y})$, но поскольку в табл. 2 обратной функции $I_{x\leftrightarrow y}=f^{-1}(L)$ можно придать более компактную форму, мы ниже приведем табличку значений $I_{x\leftrightarrow y}=f^{-1}(L)=-\frac{1}{2}\log_2(1-L^2)$. Оценка L практически не достигает единицы, так как $\Delta_h\gg 0$.

Еще проще вычисление оценок $\varkappa$ и $\widehat{\varkappa}$. По формулам (15) и (17) и по табл. 1 имеем

$$\varkappa=\sqrt{1-\left(\frac{\sum_i n_{i.}^2\sum_j n_{.j}^2}{N^2\sum_i\sum_j n_{ij}^2}\right)^2}\quad\text{и}\quad\widehat{\varkappa}=\varkappa\gamma,\tag{19}$$

Таблица 2

Величины информации $I_{x\leftrightarrow y}$ по значениям информационно-теоретического коэффициента корреляции L

L	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0001	0003	0006	0012	0018	0026	0035	0046	0059
0,1	0072	0088	0105	0123	0143	0164	0187	0212	0238	0265
0,2	0294	0325	0358	0392	0428	0466	0505	0546	0589	0634
0,3	0680	0729	0779	0832	0886	0943	1001	1062	1125	1190
0,4	1258	1328	1400	1475	1552	1632	1715	1801	1889	1981
0,5	2075	2173	2274	2379	2487	2599	2714	2834	2958	3086
0,6	3219	3357	3500	3648	3801	3961	4126	4298	4477	4663
0,7	4857	5059	5270	5491	5722	5963	6217	6483	6793	7058
0,8	7370	7700	8050	8423	8821	9247	9706	1,0202	1,0741	1,1330
0,9	1,1980	1,2702	1,3514	1,4439	1,5514	1,6792	1,8365	2,0403	2,3292	2,8255

где значение $\gamma=\sqrt{\frac{k_xk_y}{k_xk_y-1}}=\sqrt{\frac{N^4}{N^4-\sum_i n_{i.}^2\sum_j n_{.j}^2}}$ можно протабулировать по значениям k_xk_y. Однако в большинстве случаев достаточно

* Таблицы значений $-p\log_2 p$ приведены в [2], в книге А. М. Яглома и И. М. Яглома «Вероятность и информация», М., 1960, и в других источниках.

вычисления $\varkappa$. Для вычисления оценки $\varkappa$ требуются только таблицы квадратов целых чисел, счеты и счетная линейка, доступные каждому. Полезной может быть таблица значений функции $\sqrt{1-\varkappa^2}$, которая наряду с облегчением вычислений $\varkappa$ позволяет по найденным значениям $\varkappa$ (или η) найти значения условной k-разбросанности или условного среднего квадратического отклонения относительно соответствующих безусловных показателей, что дает возможность легко построить наглядную геометрическую модель двухмерного распределения, играющую немаловажную роль при интерпретации результатов вычислений.*

Одним из важнейших преимуществ статистик, вычисляемых по оценкам k-разбросанности, является то, что содержание выражения (1) создает хорошие возможности вычисления оценок k-разбросанностей методом статистических испытаний (методом Монте-Карло). Идея метода состоит в образовании случайных пар независимых событий, в сравнении их в парах и в подсчете числа сравнений и пар одинаковых событий. Для вычисления оценок условных и безусловных k-разбросанностей не потребуются большие быстродействующие электронно-вычислительные машины, вычисления можно успешно провести на счетно-перфорационных машинах (на сортировке и табуляторе). Случайные пары событий образуются отсортировкой массива перфокарт исходных данных по пробитым на них случайным или псевдослучайным числам. При этом на табуляторе Т5-М можно рациональной коммутацией достичь скорости работы до 55 сравнений в секунду при параллельном анализе 22 показателей. Это, с одной стороны, несравнимо повышает эксплуатационные возможности счетно-перфорационных машин и, с другой стороны, позволяет расширить многие приемы статистического анализа на довольно широкий круг массовых данных науки и народного хозяйства, обрабатываемых на этих машинах.

Наметим далее еще некоторые возможности применения k-разбросанности к анализу структуры сложных систем. Частная и множественная $\varkappa$ в случае нормальности трехмерного распределения могут быть вычислены по формулам частного и множественного коэффициентов корреляции, заменяя в них r на соответствующие $\varkappa$. Если распределение не нормально, то в качестве оценки частной $\varkappa_{yx\cdot z}$, характеризующей зависимость Y от X без влияния Z, можем применить

$$\varkappa_{yx\cdot z}=\sqrt{1-\left(\frac{k_{y/xz}}{k_{y/z}}\right)^2}, \tag{20}$$

а в качестве оценки множественной $\varkappa_{y,\,xz}$, характеризующей зависимость Y от случайного вектора $\vec{XZ}$

$$\varkappa_{y,\,xz}=\sqrt{1-\left(\frac{k_{y/xz}}{k_y}\right)^2}. \tag{21}$$

---

* В связи с вопросом геометрического представления многомерных распределений возникает вопрос — не целесообразно ли представить статистики зависимости для интерпретаций в виде $\psi=1-\frac{s_{y/x}}{s_y}\approx 1-\frac{k_{y/x}}{k_y}$. Статистики типа ψ легко вычислять непосредственно по данным наблюдений (например, $\psi_\varkappa=1-\frac{k_{y/x}}{k_y}$) и переход от них к обычным статистикам и обратно очень прост (например, $\psi_{\tau_i}=1-\sqrt{1-\eta^2}$ и $\eta=\sqrt{\psi_{\tau_i}(2-\psi_{\tau_i})}$).

Формулам (20) и (21) можно придать вид

$$\varkappa_{xy\cdot z} = \varkappa_{yx\cdot z} = \sqrt{1 - \left(\frac{k_{xyz}\, k_z}{k_{xz}\, k_{yz}}\right)^2}, \tag{20a}$$

(однако, при этом обычно $\varkappa_{xy\cdot z} \neq \varkappa_{x\cdot yz}$)

$$\varkappa_{y,xz} = \varkappa_{xz,y} = \sqrt{1 - \left(\frac{k_{xyz}}{k_y\, k_{xz}}\right)^2}. \tag{21a}$$

Аналогично можем сконструировать оценки $\varkappa_{xy\,\cdot\, zq}$, $\varkappa_{x,\,yzq}$ и т. д.

Из связи k-разбросанности с дисперсией (11), (12) и из возможности разложения k-разбросанности многомерных систем на компоненты (табл. 1 и другие результаты пп. 2 и 3) вытекает возможность применения таких разложений в некоторых задачах вместо приемов дисперсионного анализа.

Но имеется возможность еще одного интересного разложения разбросанности сложной системы. Из выражения (15) следует

$$k_{xy} = h_x \frac{k_x}{h_x} h_y \frac{k_y}{h_y} \sqrt{1 - \varkappa^2},$$

индуктивным обобщением чего на сложную систему более чем двух составляющих систем является

$$k_{1\,2\,3\,\ldots\,z} = \prod_{i=1}^{z} h_i \frac{k_i}{h_i} \sqrt{1 - \varkappa_i^2}, \tag{22}$$

где

$$\varkappa_i = \sqrt{1 - \left(\frac{k_{1\,2\,3\,\ldots\,i}}{k_{1\,2\,3\,\ldots\,(i-1)}\, k_i}\right)^2}. \tag{22a}$$

В качестве примера приводим некоторые из возможных толкований содержания выражения (22):

1) $k_{1\,2\,3\,\ldots\,z}$ — k-разбросанность сложной системы, состоящей из z систем событий;

2) $\prod\limits_{i=1}^{z} h_i$ — максимальная емкость k-разбросанности сложной системы, определенная в предположении, что составляющие системы независимы и распределения равномерны;

3) $1 - \prod\limits_{i=1}^{z} \frac{k_i}{h_i}$ — ограничение, наложенное на использование $\prod\limits_{i=1}^{z} h_i$ законами распределения составляющихся систем в предположении их независимости;

4) $1 - \prod\limits_{i=1}^{z} \sqrt{1 - \varkappa_i^2} = 1 - \left(\prod\limits_{i=1}^{z} k_i\right)^{-1} k_{1\,2\,3\,\ldots\,z}$ — ограничение, наложенное на использование $\prod\limits_{i=1}^{z} h_i$ зависимостью составляющих систем;

5) $1 - \prod\limits_{i=1}^{z} \frac{k_i}{h_i} \sqrt{1 - \varkappa_i^2} = 1 - \left(\prod\limits_{i=1}^{z} h_i\right)^{-1} k_{1\,2\,3\,\ldots\,z}$ — ограничение, наложенное на использование $\prod\limits_{i=1}^{z} h_i$ законом z-мерного распределения,

6) $h_i \frac{k_i}{h_i} \sqrt{1-\varkappa_i^2} = \frac{k_{1\,2\,3\ldots i}}{k_{1\,2\,3\ldots i-1}} = k_{i/1\,2\,3\ldots i-1}$

и т. д.

Вышеизложенное еще раз потверждает целесообразность включения k-разбросанности и показателя связи $\varkappa$ в арсенал вычислительной статистики. Из других приложений свойств сумм квадратов вероятностей можно еще указать на возможность применения выражения (1) в качестве теста независимости случайных чисел. Важным преимуществом показателя связи $\varkappa$ является свойство $\varkappa \to \max(\eta_{xy}, \eta_{yx})$, позволяющее применить $\varkappa$ в целях анализа нелинейных связей.

Из других возможностей упрощения вычисления статистик, характеризующих нелинейную связь, одной из простейших является применение в качестве такой статистики корреляционного отношения «по закону связи» (см. [8], стр. 63)

$$\eta_{yx}^2 = 1 - \frac{1}{s_y^2} \sum_i p_i s_{y/x_i}^2,$$

где $s_{y/x_i}^2 = \sum_j p_{ij}(y - \bar{y}_{\cdot i})$ и $\bar{y}_{\cdot i}$ — значение y на месте x_i по закону связи (по линии регрессии).

Корреляционное отношение по закону связи сравнительно легко вычисляется на электронных вычислительных машинах, не требует предварительной группировки исходных данных и менее чувствительно к числу исходных данных, чем обычное корреляционное отношение.

В приложениях многие нелинейные связи часто сравнительно хорошо аппроксимируются полиномом второй степени (параболой второго порядка). Если в качестве закона связи использовать параболу второго порядка, то по [8] имеем

$$\eta_{yx}^2 = r_{1/1}^2 + \frac{(r_{2/1} - r_{3/0}\, r_{1/1})^2}{r_{4/0} - r_{3/0}^2 - 1}, \tag{23}$$

где $r_{i/k}$ — смешанный основной момент (i-й от X и k-й от Y).

Используя известные соотношения между моментами, дисперсией и коэффициентом корреляции, можем формуле (23) после некоторых преобразований придать вид

$$\eta_{yx}^2 = r_{xy}^2 + \frac{(r_{x^2y} - r_{x^2x}\, r_{xy})^2}{1 - r_{x^2x}^2} \quad \text{и} \quad \eta_{xy}^2 = r_{xy}^2 + \frac{(r_{xy^2} - r_{y^2y}\, r_{xy})^2}{1 - r_{y^2y}^2}, \tag{24}$$

где r_{xy} — коэффициент корреляций между X и Y, r_{x^2y} — коэффициент корреляций между X^2 и Y, r_{x^2x} — коэффициент корреляций между X^2 и X и т. д.

В формулах (24) вместо квадратов случайных величин можем использовать и другие функции от них (кубы, логарифмы и т. д.). При применении формул (24) вычисление матрицы корреляционных отношений по параболе второго порядка в основном превращается в вычисление матрицы коэффициентов корреляций случайных величин и их квадратов. В целях вычисления таких корреляционных отношений после незначительных дополнений применимы все программы множественного статистического анализа, в которых вычисляется матрица коэф-

фициентов корреляции. Важнейшими дополнениями являются: 1) ввод в машину (или образование в машине) наряду со случайными величинами квадратов их и 2) применение формул (24). Время вычислений увеличивается при этом обычно незначительно.

О полезности применения корреляционного отношения по параболе второго порядка можно судить по следующему примеру: нами анализировались 10 лесотаксационных показателей, причем вычислялись матрицы коэффициентов корреляций (r) и корреляционных отношений по параболе второго порядка (η). Критерием нелинейности служил

$$\zeta_1 = \max(\eta^2_{xy},\ \eta^2_{yx}) - r^2_{xy}$$

и в качестве существенности критерия ζ_1 применялся критерий Стюдента $t = \sqrt{N\zeta_1}$. По пятипроцентному критерию t существенными оказались 28 из 45 возможных критериев нелинейности. Применение в этих случаях корреляционного отношения по параболе второго порядка вместо коэффициента корреляции существенно улучшает анализ.

Из приведенных в статье статистик, характеризующих нелинейную связь, наиболее выгодными в вычислительном отношении являются показатель связи $\varkappa$, информационно-теоретический коэффициент корреляций L и корреляционное отношение по закону связи η (в случае нормальной корреляции они совпадают с коэффициентом корреляции r). Существенно упрощаются и многие другие математико-статистические вычисления, если в этих целях применяется k-разбросанность.

В заключение выражаю благодарность Л. Выханду и Р. Тамместе за их ценные замечания, которые были учтены при оформлении статьи.

ЛИТЕРАТУРА

1. Бусленко Н. П., Шрейдер Ю. Я., Метод статистических испытаний, М., 1961.
2. Варден (ван дер) Б. Л., Математическая статистика, М., 1960.
3. Венцель Е. С., Теория вероятностей, М.—Л., 1962.
4. Занина Е. Н., Калинин О. М., Фалева Т. А., Сб.: Применение математических методов в биологии, Изд. Лен. гос. ун-та, 1963, стр. 107—109.
5. Занина Е. Н., Фалева Т. А., Тезисы докладов Третьего совещания по применению математических методов в биологии, Изд. Лен. гос. ун-та, 1961, стр. 26—27.
6. Леман Э., Проверка статистических гипотез, М., 1964.
7. Плохинский Н. А., Биометрия, Новосибирск, 1961.
8. Чупров А. А., Основные проблемы теории корреляции, М., 1960.
9. Юл Дж. Э., Кендэл М. Дж., Теория статистики, М., 1960.
10. Renyi A., Wahrscheinlichkeitsrechnung mit einem Anhang über Informationstheorie, Berlin, 1962.

Институт физики и астрономии
Академии наук Эстонской ССР

Поступила в редакцию
22/VII 1964

A. NILSON

TÕENÄOSUSTE RUUTUDE SUMMADE MÕNEDEST OMADUSTEST JA NENDE MATEMAATILIS-STATISTILISEST RAKENDUSEST

Tõenäosuste ruutude summade omadused täielikes sündmuste süsteemides võimaldavad juhuslike sündmuste ja suuruste täielike liht- ja liitsüsteemide analüüsimiseks kasutada nn. k-hajuvust (5), mille väärtuste leidmine on lihtne nii analüütiliselt kui ka Monte-Carlo meetodiga. Statistikud, mis on tuletatud k-hajuvusest, on tihedasti seotud nn. «klassikaliste» ja informatsiooniteoreetiliste statistikutega ja võimaldavad oluliselt lihtsustada paljusid mitmemõõtmelise statistilise analüüsi menetlusi (korrelatsioon- ning dispersioonanalüüs jt.).

Mittelineaarsete sõltuvuste analüüsimisel võib osutuda otstarbekaks ka nn. «sõltuvuse seaduste» järgi leitud korrelatsioonisuhete kasutamine (24).

A. NILSON

SOME PROPERTIES AND STATISTICAL APPLICATIONS OF SUMS OF SQUARES OF PROBABILITIES

If $P(A_i)=p_i$ *and* $\sum_i p_i=1$, some properties of the $\sum_i p_i^2$ are proved and the $\left(\sum_i p_i^2\right)^{-1}=k$ as the k-variance is defined. The random variables and vectors as well as random attributes and systems can be analysed by the k-variance. The appropriate analytical and Monte-Carlo methods for computing the estimations of the k-variance are discussed. A correction for grouping of the k-variance of normal distribution (9) is proposed. The schemes for computing the variance (7), (8), (11), (12), entropy (2), (13), (14), components of variance or entropy (table 1), (22) and measures of dependence (16), (18), (19), (20), (21) by k-variance are constructed.

A new formula (24) to estimate the correlation ratio by some curvilinear curves of regression is proposed.