

МОДЕЛІ КОНФІГУРАЦІЙНО-ЧАСТОТНОГО АНАЛІЗУ: МАТЕМАТИЧНИЙ АНОНС

Конфігураційно-частотний аналіз (КЧА) є статистичним методом, що дає змогу відповісти на питання про типову чи атипову поведінку. На підставі порівняння спостережуваних та очікуваних частот КЧА визначає патерни змінних як тип або антитип. В КЧА непропорційно велике (або мале) значення частот у багатовимірних таблицях сполучення виявляється на підставі критерію Пірсона χ^2 -квадрат або біноміального критерію.

У представлених матеріалах подані такі моделі КЧА: КЧА першого порядку, предикторний КЧА, КЧА вісьової симетрії та модель лонгітюдного КЧА. Більшість моделей шансів КЧА можуть бути представлені у термінах логлінійних моделей.

Вітчизняні фахівці з аналізу даних мали нагоду ознайомитися зі змістом, перевагами та модифікаціями конфігураційно-частотного аналізу (КЧА), застосуванням його для розв'язання конкретних наукових проблем [1—3]. Основна ідея цього багатовимірного методу статистики запропонована Й. Краутом (Krauth) та Г. А. Лінертом (Linert) [4]. Надалі розвиток КЧА отримав у роботах Т. Хавранека (Navranek), В. Лемахера (Lehmacher), В. Малі (Maľu), А. фон Айя (von Eye) та ін.

Багатовимірний статистичний метод конфігураційно-частотного аналізу активно використовується західними соціологами, займає належне місце у сучасних словниках у рубриці статистичних методів, що використовуються в соціальних

© Бова А., 2001

науках, більше того — входить у навчальні програми курсу вивчення статистичних методів для студентів психологічних факультетів. Поряд з тим вітчизняні науковці недостатньо знайомі з широкими можливостями КЧА. Для ґрунтовного розуміння цього методу слід викласти в загальному вигляді основні його моделі й сучасні різновиди. Завдяки поданому нижче матеріалу можливе самостійне обчислення відповідних статистик та подальша алгоритмізація окремих моделей.

Власне метод базується на схемах дисперсійного аналізу, за допомогою яких можна поглиблено досліджувати таблиці сполучення та перевіряти різні гіпотези щодо незалежності та квазі-незалежності ознак, порядку взаємодії змінних,

впливу предикторів на відгук тощо (аналог ANOVA) [5, 6]. Моделі, що дають змогу досліджувати як зв'язок у цілому, так і зв'язок, що проявляється на окремих підтаблицях, умовно назвемо *локально-глобальними* (у цих моделях пріоритетним є локальний зв'язок). На наш погляд, саме розглянутий у даній публікації метод багатовимірної статистики орієнтований на виявлення локального та глобального зв'язку в таблицях сполучення.

Статистичні критерії в КЧА

КЧА призначається для обробки даних, які представлені в табличній формі у вигляді перехресної класифікації об'єктів за різними ознаками. Ця форма наведення даних є поширеною передусім у соціологічних, психологічних та біологічних дослідженнях. У термінах КЧА патерни змінних можуть трактуватися як конфігурації. Метою цього методу є виявлення конфігурації, яка конституює тип або антитип. Тип і антитип визначаються в ході порівняння емпіричних і теоретичних частот в окремих клітинках перехресної класифікації.

Якщо o — спостережувані частоти, а e — очікувані, то КЧА можна записати статистично як

$$o \neq e. \quad (1)$$

При $o > e$, конфігурація утворює *тип* та, навпаки, якщо $o < e$ — конфігурація є *антитипом*.

Для виявлення (анти)типів застосовують різноманітні статистичні критерії [7, 8]. Найбільш поширеними є критерій χ^2 -квадрат (χ^2) та біноміальний критерій (Z). Конфігурація становить певну закономірність, якщо значення χ^2 -квадрату для одного ступеня волі (або іншого критерію), що обчислено в клітинці таблиці сполучення, більше граничного. Найбільший інтерес для дослідника становлять типи і антитипи, що утворюються в результаті сполучення відповідних категорій ознак. Наведемо відповідні формули для обчислення зазначених вище статистичних критеріїв для випадку тривимірної таблиці:

$$\chi^2_{ijk} = \frac{(o_{ijk} - e_{ijk})^2}{e_{ijk}}. \quad (2)$$

Теоретичні частоти обчислюються, як:

$$e = \frac{N_{i..} N_{.j.} N_{..k}}{N^2}, \quad (3)$$

де $N_{i..} N_{.j.} N_{..k}$ — суми за маргіналами, а N — загальна кількість об'єктів.

$$Z_{ijk} = \frac{o_{ijk} - e_{ijk}}{\sqrt{e_{ijk} q_{ijk}}}, \quad (4)$$

де $q_{ijk} = 1 - p$. Даний тест використовується при $Np_{ijk} \geq 10$, де $p_{ijk} = e_{ijk}/N$. Таким чином, для запропонованого критерію спочатку обчислюється значення нормальної апроксимації конфігурації,

яке порівнюється з табличним критичним значенням. Зазвичай використовують рівень значущості $\alpha = 0,05$, який ділиться на кількість усіх можливих конфігурацій (r). Критичне значення міститься в таблиці нормального розподілу.

Деякі статистичні критерії можуть бути модифіковані. Наприклад, при невеликій кількості наповнення клітинок об'єктами для обчислення може застосовуватись біноміальний критерій з корекцією на неперервність,

$$Z^T = \left(o_{ijk} - e_{ijk} - \frac{1}{2} \right) / \sqrt{e_{ijk} (N - e_{ijk}) / N}. \quad (5)$$

Якщо очікування є менше одиниці, обчислюють таку статистику:

$$T = \frac{o_{ijk}}{N - o_{ijk} + 1} \cdot \frac{N - e_{ijk}}{e_{ijk}}, \quad (6)$$

яку порівнюють з критичною F -межею для $df_1 = 2(N - o_{ijk} + 1)$, $df_2 = 2o_{ijk}$, та $\alpha^* = 0,05/r$, де r — кількість клітинок (конфігурацій) у таблиці контингенції. При великому значенні df_2 ($df_2 = \infty$) критична F -межа може бути отримана з таблиці χ^2 з $df = df_1$ як $F = \chi^2 / df$.

Після виявлення типів конфігурацій можна оцінити їх виразність за допомогою спеціального коефіцієнту кореляції [9]. Спочатку встановлюється значущість типів, оскільки відомо, що велика кореляція може бути незначущою при невеликому обсязі N . Коефіцієнт виразності змінюється від 0 до 1, де 0 означає слабку, а 1 — добру виразність типу. Отже, якщо відома спостережувана частота o та обчислена відповідна очікувана частота e , то:

$$Q = \frac{o - e}{\text{Max}(e, N - e)} = \frac{2|o - e|}{N + |2e - N|} = \frac{2\sqrt{e}}{N + |2e - N|} \times \sqrt{\frac{(o - e)^2}{e}}. \quad (7)$$

Коефіцієнт дає змогу визначити і порівняти виразність різних типів і має такі властивості. Q знаходиться завжди в інтервалі від 0 до 1. $Q = 0$, якщо $o = 0$. Максимального значення коефіцієнт досягає у випадку: $e < N/2$ та $o = N$, $e > N/2$ і $o = 0$, $e = N/2$ та $o = N$. При застосуванні асимптотичного критерію до Q^2 отримуємо

$$Q^2 > \frac{4e}{(N + |2e - N|)^2} \chi^2_{\alpha}(1df) = \frac{e}{(\text{Max}(e, N - e))^2} \chi^2_{\alpha}(1df). \quad (8)$$

Конфігураційно-частотний аналіз можна подати у термінах логлінійних моделей, які мають властивість адитивності. Важливо зауважити, що більшість моделей КЧА з точки зору перспективи обчислювальних процедур є класом простих

логічних моделей, коли очікувані частоти обчислюються, виходячи з маргіналів. Представимо логічні моделі для пояснення моделей шансів КЧА, а також наведемо формули для очікуваних частот деяких моделей [8, 10]. Перша являє собою загальну та класичну модель *першого порядку*. Додатково наводяться формули й окремі пояснення до моделей *нульового* та *другого порядку*. Друга є регіональною моделлю *предикторного* КЧА (П-КЧА). Дана форма КЧА є проблематикою *інтераційно-структурного* аналізу, основна мета якого — виявлення порядку взаємодії змінних [11]. Третя є моделлю *вісьової симетрії* та остання модель — *лонгitudного* КЧА.

Модель першого порядку КЧА

У моделі нульового порядку, т. з. конфігураційно-кластерному аналізі, перевіряється нульова гіпотеза про відсутність ефектів. Для моделі нульового порядку чотирьох змінних маємо:

$$\log e_{ijkl} = u, \quad (9)$$

де очікувана частота клітинки x_{ijkl} обчислюється як

$$e_{ijkl} = \frac{N}{IJKL}, \quad (10)$$

де I, J, K, L — кількість категорій ознак A, B, C та D відповідно, а N — означає обсяг вибірки. Подібність до кластерного аналізу моделі нульового порядку КЧА та ідентифікації типів полягає в агломерації об'єктів.

Модель КЧА першого порядку передбачає відсутність взаємодії між змінними. При оцінці теоретичних частот розглядають лише головні ефекти. Типи виникають у результаті застосування КЧА першого порядку, якщо взаємодія між змінними, наприклад локальні зв'язки, можуть бути пояснені, виходячи з головних ефектів.

Логічна модель головних ефектів для чотирьох категоризованих змінних має такий вигляд:

$$\log e_{ijkl} = u + u_{i(A)} + u_{j(B)} + u_{k(C)} + u_{l(D)}, \quad (11)$$

де u — константа та $u_{i(A)}, u_{j(B)}, u_{k(C)}, u_{l(D)}$ — параметри ефектів. Як ілюстрацію застосування моделі розглянемо формулу для оцінки очікуваних частот, що використовується в КЧА першого порядку. Припустимо, що досліджується зв'язок ознак A, B, C та D . Змінна A містить i категорій, змінна B містить j категорій, змінна C містить k категорій, а змінна D містить l категорій, тобто ми маємо коефіцієнти a_i, b_j, c_k та d_l , де $i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$ та $l = 1, \dots, L$. Отже, очікувана частота для елемента $ijkl$ в перехресному розташуванні ознак A, B, C та D в таблиці оцінюється за допомогою:

$$e_{ijkl} = \frac{a_i \cdot b_j \cdot c_k \cdot d_l}{N^3}, \quad (12)$$

де крапками в підрядковому індексі позначені маргінали, які отримують підсумком відповідних емпіричних частот. Іншими словами $a_{i..}$ означає частоту i -ї категорії ознаки A , $b_{.j.}$ — частоту j -ї категорії ознаки B , $c_{..k}$ — частоту k -ї категорії ознаки C та $d_{...l}$ — частоту l -ї категорії ознаки D .

Зазначимо, що модель головних ефектів може бути застосована й при перевірці гіпотез стосовно квазінезалежності ознак, коли спеціальним чином елімінуються окремі частоти в комірцях таблиці сполучення.

Таким чином, модель другого порядку КЧА виявляє головні ефекти та парну асоціацію. Отже, типи й антитипи в цій моделі проявляються лише при наявності зв'язку вищого порядку. В термінах логічного аналізу відповідна формула для оцінки очікуваних частот КЧА другого порядку така:

$$\log e_{ijkl} = u + u_{ij(AB)} + u_{ik(AC)} + u_{il(AD)} + u_{jk(BC)} + u_{jl(BD)} + u_{kl(CD)}. \quad (13)$$

Існують приклади відповідності логічних моделей та моделей КЧА з числовою оцінкою параметрів останніх [10, р. 62—65].

Предикторна модель КЧА

Якщо в моделі першого порядку передбачається однаковий статус усіх змінних, то в предикторному КЧА ознаки поділяються на незалежні та критеріальні. Логічна модель для оцінки очікуваних частот для П-КЧА, може бути визначена таким чином: предиктори та критерії є незалежними, але може існувати зв'язок будь-якого типу у групі залежних і незалежних ознак.

Порушення моделі можливе у випадку, якщо існує зв'язок між незалежними ознаками та критеріями. Ці зв'язки проявляються у вигляді типів і антитипів. Нехай P означає низку незалежних ознак, а C — набір критеріїв. Припустимо, що існують два предиктори (A та B) та критерії (D та E). Ознака A має I категорій, B — J , D — K , E — L категорій відповідно (при $I, J, K, L \geq 2$). Тоді модель матиме такий вигляд:

$$\log e_{ijkl} = u + u_{i(A)} + u_{j(B)} + u_{k(C)} + u_{l(D)} + u_{ij(AB)} + u_{kl(CD)}, \quad (14)$$

де параметри з одним підрядковим індексом позначають головні ефекти, а з двома — взаємозв'язки. Очікувані частоти П-КЧА також обчислюються виходячи з маргіналів. Для цього утворюється одна незалежна ознака, яка складається з низки незалежних ознак та одного критерію, що складається із залежних ознак. Для нашого випадку з чотирма ознаками композитна ознака P має $M = I \times J$ категорій. Ці категорії формуються в результаті перехресного сполучення ознак A та B . Складовий критерій C має $O = K \times L$ категорій,

котрі є елементами перехресної класифікації категорій ознак D та C . Тобто, утворені дві комбіновані ознаки трансформують чотиривимірну таблицю контингенції у двовимірну. Очікувані частоти e_{mo} обчислюються за аналогією з (12) таким чином:

$$e_{mo} = \frac{p_m \cdot c_o}{N}, \quad (15)$$

де p_m — маргінальна частота P , а c_o — маргінальна частота C . Різновидом П-КЧА є стратифікована предикторна модель [12].

Модель вісьової симетрії для КЧА

Завдяки моделям вісьової симетрії можна перевіряти гіпотези стосовно характеристик об'єктів при повторних замірах. При цьому використовуються квадратні таблиці сполучення ознак, а зразки перевіряються на стабільність і змінювальність. Квадратні, тобто $I \times I$, перехресні таблиці є симетричними щодо їх головної діагоналі, якщо

$$o_{ij} = o_{ji}. \quad (16)$$

У цілому вісьова симетрія оцінюється за допомогою спеціальних тестів або логлінійних моделей, які можуть бути представлені у вигляді:

$$\log e_{ij} = u + u_{i(A)} + u_{j(B)} + u_{ij(AB)}, \quad (17)$$

з такими крайніми обмеженнями:

$$u_{ij(AB)} = u_{ji(AB)}, \quad (18)$$

та

$$\sum_i u_{i(A)} = \sum_i u_{ij(AB)}, \quad (19)$$

де (18) характеристика взаємодії, що подає оцінювані частоти, котрі задовольняють визначенню вісьової симетрії.

З визначення вісьової симетрії можна отримати просту формулу оцінки очікуваних частот для кожної пари елементів ij та ji таким чином:

$$e_{ij} = e_{ji} = \frac{o_{ij} + o_{ji}}{2}. \quad (20)$$

КЧА вісьової симетрії перевіряє, чи є переміщення від i до j таким же як переміщення від j до i . Ця перевірка проводиться для кожної пари елементів ij та ji . Якщо один з елементів, i або j являє собою тип (відповідно, інший елемент завжди являтиме антитип), можна зробити висновок, що переміщення від i до j буде, імовірно, відрізнятися від j до i .

Лонгitudний КЧА

Складним завданням при аналізі даних у динаміці є повторні спостереження. Загальною формою є спостереження одних і тих самих змінних кілька разів. У деяких дослідницьких проектах один набір ознак спостерігається перший раз, інший набір ознак — другий раз, тобто враховуються як ознаки, так і моменти спостереження. Переважно такі завдання постають перед психо-

метристами і соціологами, що здійснюють лонгitudні дослідження.

КЧА не має обмежень, які притаманні іншим статистичним методам аналізу стосовно залежності структури від часу. Таким чином, існування залежних від часу структур не є причиною того, аби не довіряти результатам КЧА, який визначає і описує такі структури.

Головні передбачення лонгitudного КЧА такі. Кожен об'єкт (область, район) повинен мати однакову кількість моментів спостереження. Тому при обчисленні в лонгitudному КЧА передбачається, що ознаки спостерігаються неодноразово та їх кількість визначена. Розглянемо випадок, де одна змінна спостерігається $T = 4$ рази ($t = 1, \dots, 4$). Логлінійна модель шансів лонгitudного КЧА для цих даних:

$$\log e_{ijkl} = u + \sum_{t=1}^T u_{ij(AB)t}, \quad (21)$$

де i, j, k, l — показник спостережуваної категорії в моменти спостереження.

Остаточне рівняння для оцінки очікуваних частот може бути отримано за аналогією з (12). Модель лонгitudного КЧА, запропонована як (21), передбачає, що значення для неодноразово спостережуваної змінної X є повністю незалежними один від одного в часі.

Модель, яка показана в (21), може бути змінена для того, аби розмістити додаткові змінні, наприклад коваріації, та висунути передбачення стосовно структури автокореляції. Якщо розглядаються автокореляції, то типи й антитипи будуть проявлятися лише у випадку існування взаємодії між змінними, поза яких передбачена модель КЧА.

Модифікації КЧА

Модифікаціями цього методу є ієрархічний та аглютинований КЧА [2]. Ієрархічний варіант КЧА дозволяє скоротити деякі ознаки у випадку, коли кількість об'єктів є досить малою, або, незалежно від кількості об'єктів, елімінувати ознаки, що мають малий внесок у конституювання типів при максимальному збереженні корисної інформації. Ієрархічний КЧА полягає в послідовному вилученні ознак та аналізу всіх таблиць, які утворюються залишеними змінними. Порядок моделі, тобто виявлення наявності зв'язку певного порядку, в КЧА та логлінійному аналізі обчислюється однаково.

Охоплення окремих конфігурацій називається аглютинацією (об'єднанням), відповідно така форма КЧА — аглютинованим (об'єднуючим) КЧА. Завдяки аглютинованому КЧА можна розв'язати завдання щодо:

1) побудови інтегрованих типів при недостатньо наповнених об'єктами або слабо виражених конфігураціях;

2) формування виразних аглютинованих типів зі слабо виразних окремих типів;

- 3) редукції кількості типів до мінімуму;
- 4) простої та переконливої інтерпретації;
- 5) уникнення низького значення очікуваної частоти для оцінки конфігурації.

Застосовуючи конфігураційно-частотний аналіз, також можна отримати відповідь щодо внеску окремих ефектів та коваріацій у загальний зв'язок.

Слід зазначити, що КЧА виявляє лише клітини в таблиці контингенції, в яких міститься непропорційно багато або непропорційно мало спостережень, тобто в певному сенсі ті спостереження, які статистично суперечать логлінійним моделям. Тобто КЧА є певним антиподом логлінійного аналізу. Отже параметри, їх числові значення і підгонка моделей в цілому не відіграють важливої ролі при такому аналізі. Не є метою КЧА також знаходження параметрів логлінійної моделі, які добре описують дані.

Завдяки наведеним формулам можливо виявляти типи й антитипи в емпіричній інформації, оцінювати їх значущість, навіть при невеликій кількості частот, та обчислювати відповідний коефіцієнт виразності. В кінцевому підсумку дані показники повніше характеризують закономірності, дають змогу здійснювати обґрунтовану інтерпретацію. Слід наголосити, що КЧА має значні переваги порівняно з іншими методами аналізу даних, а його основні моделі потребують нескладних обчислень. Більше того, результати, досягнуті КЧА, не можуть бути виявлені шляхом проведення факторного, кластерного або латентно-структурного аналізу. Викладений вище матеріал стане в пригоді соціологам та психологам, які беруть участь у прикладних дослідженнях та цікавляться сучасними методами багатовимірної статистики.

1. Краус Ж., Линерт Г. А. Основы конфигурационно-частотного анализа (КЧА) и его модификации // Психологический журнал.— 1984.— Том 5.— № 1.— С. 26—34.

2. Бова А. А. Конфігураційно-частотний аналіз у психологічних дослідженнях // Практична психологія та соціальна робота.— 1988.— № 8.— С. 42—45.

3. Бова А. А. Моделирование межпоколенной профессиональной мобильности // Подвижность структуры. Современные процессы социальной мобильности / С. А. Макеев, И. М. Прибыткова, Е. В. Симончук и др.— К.: Институт социологии НАН Украины, 1999.— С. 97—114.

4. Krauth J., Linert G. A. Nichtparametrischer Nachweis von Syndromen durch simultane Binominaltests // Biometrische Zeitschrift.— 1973.— Bd. 15.— S. 13—20.

5. Havranek T., Kohlen R., Lienert G. A. Nonparametric evaluation of ANOVA designs by local, regional, and global contingency testing // Biometrical Journal.— 1986.— Vol. 28.— P. 11—21.

6. Linert G. A., Longo N. Nonparametric evaluation of learning curves by hierarchical contingency analysis // Biometrical Journal.— 1988.— Vol. 30.— P. 17—24.

7. Lehman W. A more powerful simultaneous test procedure in in configural frequency analysis // Biometrical Journal.— 1981.— Vol. 23.— P. 429 — 432.

8. Eke von A., Spiel C., Wood P. K. Configural frequency analysis in applied psychological research // Applied Psychology: an International Review.— 1996.— Vol. 45.— P. 305—309.

9. Krauth J., Lienert G. A. Die Konfigurationsfrequenzanalyse (KFA) und ihre Anwendung in Psychologie und Medizin: ein multivariates nichtparametrisches Verfahren zur Aufdeckung von Typen und Syndromen. Weinheim: Beltz, Psychologie Verlags Union, 1995.— S. 33—34.

10. Eke von A. The general model as framework for models in configural frequency analysis // Biometrical Journal.— 1988.— Vol. 30.— P. 59—63.

11. Krauth J., Lienert G. A. Zum Nachweis syndromgenerierender Symptominteraktionen in mehrdimensionalen Kontingenztafeln (Interaktionsstrukturanalyse) // Biometrische Zeitschrift.— 1974.— Bd. 16.— S. 203—211.

12. Helmann W. -R., Lienert G. A., Maly V. Prediction models in in configural frequency analysis // Biometrical Journal.— 1979.— Vol. 21.— P. 79—86.

Bova A.

MODELS OF CONFIGURAL FREQUENCY ANALYSIS: MATHEMATICAL ANNOUNCE

Configural Frequency Analysis (CFA) is a statistical method that allows one to answer questions concerning typical or atypical behaviors. Specifically, CFA allows one to identify variable patterns that are, relative to some expectation, more frequent (typical) or less frequent (atypical). In the CFA overfrequent (or underfrequent) cells in multivariate contingency tables identified by Pearson χ^2 test or simultaneous binomial test.

The present paper discusses model of CFA: the First Order CFA, the Prediction CFA, Axial Symmetry CFA, and model Longitudinal CFA. Most CFA chance model can be cast in terms of log-linear models.