**Український Радіологічний Журнал**

**УРЖ 2002; 10:** 426 – 429

Є.Б.Радзішевська,
М.І. Пилипенко,
В.Г. Кнігавко

*Харківський державний медичний університет*

*Інститут медичної радіології ім. С.П. Григор'єва АМН України, м. Харків*

# Елементи медичної статистики. Лекція 9. Багатовимірний статистичний аналіз

## Elements of medical statistics. Lecture 9. Multidimensional statistical analysis

Досі розглядалися ситуації, коли випадкова мінливість торкалась однієї змінної (див. Лекції 1–8, УРЖ. — 2000. — Т. VIII, вип. 1–4; 2001. — Т. IX, вип. 1, 3, 4; 2002. — Т. X, вип. 1). Вочевидь, такі ситуації не характерні для світу явищ, а є лише максимально спрощеними його моделями, зручними для аналізу зв'язків саме однієї змінної величини.

Більш складні моделі розглядають одночасні зміни кількох параметрів, між якими, можливо, існують складні, не завжди очевидні, зв'язки. Для опису таких ситуацій створені методи багатовимірного статистичного аналізу (БСА). За їх допомогою розв'язують задачі класифікації, визначення прихованих закономірностей у складному явищі, встановлення зв'язків і залежностей між факторами або характеру й спрямованості зв'язків.

До головних методів БСА належать багатовимірна кореляція й регресія, факторний та кластерний аналізи, дискримінантний аналіз тощо.

На жаль, побудова теорії для багатовимірних статистичних даних виявилася справою вельми важкою. Добре розроблено лише теорію для гауссових (що мають багатовимірний нормальний розподіл) даних. Тут майже для кожного одновимірного гауссового статистичного методу існує відповідний багатовимірний варіант.

Побудова багатовимірних версій для інших статистичних методів здійснюється не так гладко. Зокрема, непараметричні методи, такі важливі й ефективні в одновимірному випадку, все ще не мають свого завершеного аналога (відповідна теорія перебуває на стадії розробки). Тому для акуратного статистичного аналізу існуючих даних часто не знаходиться адекватних статистичних засобів. Через це, зокрема, розраховані на гауссові дані правила доводиться застосовувати й там, де для цього немає достатніх підстав.

Остаточні висновки в таких випадках буває нелегко інтерпретувати.

Попри все це, необхідність методів БСА така очевидна, що в усьому світі кількість спеціалістів, які використовують ці методи на практиці, стрімко зростає. Популярність багатовимірних технологій охоплює методи багатовимірного аналізу даних. Критерієм істинності побудованих моделей у цьому випадку є виключно практика: якщо результати не суперечать здоровому глуздові та фундаментальним поняттям, допускають однозначне медичне тлумачення й виявляють усталеність (тобто повторюваність при зміні обсягу вибірки), то немає підстав вважати їх нереальними.

Перш ніж перейти до розгляду окремих методів БСА, конкретизуємо сенс категорії «багатовимірність» методів статистичного аналізу.

Неодмінною умовою встановлення діагнозу пацієнта, оцінки тяжкості захворювання, стану хворого та іншого є аналіз цілого комплексу різноманітних кількісних і якісних показників. Можна сказати, що стан пацієнтів у цілому є багатоознаковою системою. Для статистичного аналізу будь-якої багатоознакової системи розглядають таблицю (матрицю), рядками якої є дані щодо кожного пацієнта досліджуваної сукупності, а графами — аналізовані показники. В подальшому викладі матриці такого виду ми називатимемо матрицями «об'єкти — властивості».

Розглянемо призначення основних методів БСА, застосовуваних для аналізу багатоознакових систем.

**1. Факторний аналіз**

Основна мета такого аналізу полягає в тому, щоб виявити приховані спільні фактори, що пояснюють зв'язки між спостережуваними ознаками (параметрами) об'єкта. Кількість досліджуваних ознак може бути великою, а взаємозв'язки надзвичайно складними.

Факторний аналіз слугує для перевірки гіпотези про те, що існує невелика кількість чинників, які впливають на вимірювані параметри.

Ці чинники (нові ознаки) не можна виміряти безпосередньо, вони є лінійними комбінаціями попередніх, «вбирають» у себе більшу частину загальної мінливості спостережуваних ознак і тому передають основну частину інформації, отриманої в початкових спостереженнях. Виявлені таким чином фактори називаються загальними, бо вони впливають на всі ознаки, а не на одну ознаку чи їх групу.

Якщо вплив нового фактора проявляється в кількох вимірюваних ознаках, останні можуть виявляти тісний взаємозв'язок між собою (наприклад, корельованість), тому загальна кількість факторів може бути значно меншою, ніж кількість вимірюваних ознак, яку дослідник зазвичай обирає тією чи іншою мірою довільно.

Таким чином, головними цілями факторного аналізу є:

а) скорочення кількості ознак (редукція даних);

б) визначення структури взаємозв'язків між змінними, тобто класифікація змінних.

Зверніть увагу на те, що факторний аналіз — це метод зовсім іншого призначення, ніж одно-, дво- та багатофакторний аналізи, розглянуті раніше. В одно-, двофакторному і подібних аналізах (англійською one-way, two-way і таке інше Analysis of Variance) фактори, що впливають на результат, вважаються відомими, і йдеться тільки про визначення суттєвості цього впливу. А в факторному аналізі (Factor analysis) розглядають виокремлення з множини вимірюваних характеристик об'єкта нових чинників, які адекватніше відображують його властивості.

Наприклад, методами факторного аналізу можна вивчити причини виникнення того чи іншого захворювання. Для цього на кожного хворого заводять анкету, в якій висвітлюють питання, що стосуються найрізноманітніших сторін його життя — спадковість, рід діяльності, харчування, звички, характеристики району проживання тощо — ту інформацію, яка, на погляд дослідника, має відношення до проблеми. Отримані результати піддають факторному аналізу, що дозволяє виділити кілька нових факторів як комбінацію вихідних. При цьому якісні відмінності між факторами визначаються кількісними особливостями ознак, що входять до них. Чим менший внесок ознаки у формування фактора, тим менше він значущий для вивчення спостережуваного явища.

**2. Кластерний аналіз**

Методи кластерного аналізу дозволяють розбити сукупність об'єктів, яку вивчають, на групи «подібних» об'єктів, званих кластерами. Іншими словами, процедури кластерного аналізу дозволяють упорядкувати об'єкти за однорідними групами.

Так, наприклад, у галузі психіатрії правильне виділення кластерів симптомів параної, шизофренії тощо є вирішальним для успішної терапії.

Розвиток різних баз даних у медичних установах дозволяє накопичувати великий обсяг різнопрофільної інформації про хворих, починаючи з антропологічних і соціальних відомостей до інформації про захворювання та лікування в динаміці. Для того, щоб уся інформація була дійсно корисною і застосованою на практиці, необхідні засоби її обробки. Одним із першочергових завдань обробки є зведення даних в однорідні групи, всередині яких слід виконувати пошук закономірностей. Необхідність використання для цього процедур кластерного аналізу пов'язана з тим, що велика кількість факторів і різнорідність самої інформації роблять вибір ознак кластеризації неочевидним.

Традиційно задачу згрупування первинних даних розв'язують таким чином. Із множини ознак, що описують об'єкт чи явище, відбирають одну, найбільш інформативну з погляду дослідника, та проводять згрупування відповідно до значень цієї ознаки. Якщо є потреба класифікації за кількома ознаками, рангованими між собою за ступенем важливості, то

спочатку проводять класифікацію за першою ознакою, потім кожен з одержаних класів розбивають на підкласи за другою ознакою і т.д. Методи багатовимірного аналізу дозволяють зробити цей досить складний алгоритм, що потребує багато часу, більш компактним та швидким. Наявність ПЕОМ та спеціалізованого програмного забезпечення взагалі дозволяє миттєво виконувати всі розрахунки. За таких умов на перший план висувається спроможність дослідника коректно сформулювати завдання дослідження, вибрати відповідну статистичну процедуру й програмне забезпечення та знайти відповідне медичне тлумачення одержаних результатів.

Як вже було зазначено, кластерний аналіз є одним із методів багатовимірного статистичного аналізу, коли кожне спостереження подають не одним числом, а сукупністю чисел-ознак. Фактично кластерний аналіз є не стільки звичайним статистичним методом, скільки «набором» різноманітних алгоритмів розподілу об'єктів за кластерами. Одним із таких алгоритмів є так званий метод *k*-середніх. Цей метод припускає відомою кількість кластерів *k* та будує рівно *k* таксономій (груп, кластерів), розташованих на якомога більших відстанях одна від одної.

З обчислювального погляду це є дисперсійний аналіз «навпаки». Спочатку об'єкти поділяють на *k* кластерів випадковим способом, а потім переміщують із кластеру в кластер так, щоб мінімізувати мінливість усередині кластеру та максимізувати мінливість між кластерами.

Після того, як кластери утворено, розраховують середні значення для кожного з них за кожним із вимірювань, щоб оцінити, наскільки кластери відрізняються один від одного.

**3. Дискримінантний аналіз**

Цей аналіз як розділ БСА включає в себе статистичні методи кластеризації багатовимірних спостережень у ситуації, коли дослідник володіє так званими вибірками, що навчають («кластеризація з учителем»). Сутність аналізу розглянемо нижче.

Припустимо, існує сукупність об'єктів, розбита на кілька груп (тобто для кожного об'єкта ми можемо сказати, до якої групи він належить). Нехай для кожного об'єкта існують значення кількох кількісних характеристик. Ми хочемо знайти спосіб, як на підставі значень цих характеристик можна визначити групу, до якої належить розглядуваний об'єкт. Це дозволить нам для нових об'єктів із тієї ж сукупності передбачати групи, до яких вони належать.

Для медицини таке завдання дуже типове. Наприклад, якщо як об'єкти розглядають пацієнтів (здорових чи хворих на ту чи іншу недугу), а як характеристики — результати медичних аналізів. На підставі накопичених раніше даних (вибірки-вчителі) методами дискримінантного аналізу будують так звані «класифікуючі функції». Значення аналізів нового пацієнта, який надійшов, підставляють у класифікуючі функції та на підставі отриманих значень (конкретних чисел) видають діагноз.

Іншою не менш типовою задачею, що розв'язується методами дискримінантного аналізу, є задача диференційної діагностики. Як вибірку-вчитель тут розглядають набір аналізів хворих на важко розрізнювані захворювання (наприклад, гепатит, цироз печінки, її метастатичні ураження), для яких вже існує референтний діагноз. Отримані на підставі вибірки, що навчає, класифікуючі функції (своя функція для кожної нозології) дозволяють віднести новообстеженого пацієнта з невідомим діагнозом до тієї чи іншої групи.

Звернімо увагу читача на те, що кластерний і дискримінантний аналізи можна застосовувати як до вихідних, так і до узагальнених ознак (факторів), отриманих методами факторного аналізу.

**4. Багатовимірний регресійний аналіз**

Крім завдань зниження розмірності та класифікації, методи БСА дозволяють дослідити також багатовимірні залежності. Для цього існують методи множинного регресійного аналізу й багатовимірні кореляції.

Загальне призначення множинної лінійної регресії полягає в аналізі зв'язку між кількома незалежними змінними $X_1, X_2, X_3, \ldots, X_p$ (які називають також предикторами і в подальшому позначають просто X) та залежної змінної Y. При цьому кожну змінну подають у вигляді *n* спостережень.

| Змінна | Спостереження | | | | |
|---|---|---|---|---|---|
| $Y$ | $Y_1$, | $Y_2$, | $Y_3$, | …, | $Y_n$ |
| $X_1$ | $X_{11}$, | $X_{12}$, | $X_{13}$, | …, | $X_{1n}$ |
| $X_2$ | $X_{21}$, | $X_{22}$, | $X_{23}$, | …, | $X_{2n}$ |
| $X_3$ | $X_{31}$, | $X_{32}$, | $X_{33}$, | …, | $X_{3n}$ |
| … | … | … | … | … | … |
| $X_p$ | $X_{p1}$, | $X_{p2}$, | $X_{p3}$, | …, | $X_{pn}$ |

Тоді, в загальному випадку, задачею процедур множинної регресії є оцінка параметрів лінійного рівняння виду:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p .$$

Регресійні коефіцієнти (або β-коефіцієнти) характеризують внески кожної незалежної змінної. Кожний із коефіцієнтів вимірює середнє за сукупністю відхилення результативної ознаки від її середньої величини при відхиленні даного фактора $X_i$ від своєї середньої величини на одиницю за умови, що всі інші фактори, які входять у рівняння регресії, закріплені на середніх значеннях (тобто не змінюються). Якщо б ми знали всі фактори, що впливають на результативну ознаку Y, і включили б їх у рівняння регресії, то величини $\beta_i$ можна було б вважати мірою чистого впливу факторів.

Наприклад, цілком імовірно виявити значущий негативний зв'язок між довжиною волосся і зростом (невисокі люди мають довше волосся). На перший погляд це може здатися дивним, однак, якщо додати змінну *Стать* у рівняння множинної регресії, цей зв'язок, скоріше за все, зникне. Це станеться, через те що жінки в середньому мають довше волосся, ніж чоловіки, при цьому вони також у середньому менші на зріст. Таким чином, після видалення різниці за статтю шляхом уведення змінної *Стать* у рівняння зв'язок між довжиною волосся та зростом зникає, оскільки довжина волосся не дає якого-небудь самостійного внеску в передбачення зросту, крім того, який вона поділяє із змінною *Стать*.

Лінія регресії виражає найліпше передбачення залежної змінної (Y) за незалежними змінними (X). Однак природа рідко (якщо й узагалі коли-небудь) буває повністю передбачуваною, і зазвичай існує істотний розкид спостережуваних точок відносно підігнаної прямої. Відхилення спостереження від лінії регресії (від передбаченого значення) називають залишком.

Чим менший розкид значень залишків коло лінії регресії відносно загального розкиду значень, тим кращий прогноз. Наприклад, якщо зв'язок між змінними X і Y відсутній, то відношення залишкової мінливості змінної Y до вихідної дисперсії дорівнює 1. Якщо ж X та Y жорстко зв'язані, то залишкова мінливість відсутня, а відношення дисперсій дорівнює 0. В більшості випадків відношення лежатиме десь між цими екстремальними значеннями, тобто між 0 і 1. Різницю між одиницею і цим відношенням називають R-квадратом, або коефіцієнтом детермінації. Якщо, наприклад, R-квадрат становить 0,4, то це означає, що 40 % від вихідної мінливості можуть бути пояснені побудованою моделлю, а 60 % від залишкової мінливості залишаються непоясненими.

Зазвичай ступінь залежності двох чи більше предикторів (незалежних змінних або змінних X) від залежної змінної (Y) виражають за допомогою коефіцієнта множинної кореляції R. За визначенням він дорівнює квадратному кореню з коефіцієнта детермінації. Це позитивна величина, що набирає значень між 0 та 1. Для інтерпретації напрямку зв'язку між змінними дивляться на знаки (плюс чи мінус) регресійних коефіцієнтів або β-коефіцієнтів. Якщо β-коефіцієнт позитивний, то зв'язок даної змінної із залежною позитивний, якщо негативний, то зв'язок теж має негативний характер.

Примітка. Ми розглянули лише окремий випадок множинної регресії — лінійну регресію. Часто регресійний зв'язок має складніший вигляд і описується іншими видами рівнянь (поліноміальним, ступеневим тощо). В більшості сучасних статистичних пакетів передбачається вибір відповідної моделі залежно від фізичної сутності явища, яке вивчають.

Дата надходження: 26.09.2002.

Адреса для листування:
Пилипенко Микола Іванович,
ІМР ім. С.П. Григор'єва АМНУ,
вул. Пушкінська, 82, Харків, 61024, Україна