УДК 616:316:311.21

МЕТОДИЧНИЙ ПІДХІД ЩОДО ВИБОРУ МЕТОДУ СТАТИСТИЧНОЇ ОБРОБКИ ДАНИХ ДЛЯ МЕДИКО-СОЦІОЛОГІЧНИХ ДОСЛІДЖЕНЬ

О. В. Гойко

Національна медична академія післядипломної освіти імені П. Л. Шупика

Запропоновано методичний підхід, в основу якого покладено алгоритм, що вказує на послідовність дій, які слід виконувати досліднику при обробленні й аналізі одержаних результатів наукового дослідження.

Ключові слова: комп'ютерний аналіз, сучасні технології аналізу даних, математичні методи обробки.

МЕТОДИЧЕСКИЙ ПОДХОД К ВЫБОРУ МЕТОДА СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ ДЛЯ МЕДИКО-СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

О. В. Гойко

Национальная медицинская академия последипломного образования имени П. Л. Шупика

Предложен методический подход, в основе которого лежит алгоритм, указывающий на последовательность действий, которые следует выполнять исследователю при обработке и анализе полученных результатов научного исследования.

Ключевые слова: компьютерный анализ, современные технологии анализа данных, математические методы обработки.

METHODICAL APPROACH TO CHOOSING A METHOD OF STATISTICAL DATA FOR MEDICAL AND SOCIOLOGICAL RESEARCH

O. V. Hoyko

Shupyk National Medical Academy of Postgraduate Education

There is proposed a methodical approach, which is based on an algorithm that indicates the sequence of actions that must be performed by a researcher for processing and analysis ofthe results of scientific research.

Key words: computer analysis, advanced data mining technology, mathematical methods of processing.

Вступ. Статистичні методи оброблення даних уже давно застосовуються в найрізноманітніших сферах людської діяльності й, насамперед, там, де досліджуються закономірності, властиві великим обсягам об'єктів. Це стосується і медичної галузі, де завдяки використанню більш потужної сучасної апаратури та застосуванню нових методів обстеження невпинно зростають інформаційні масиви та кількісні дані про стан здоров'я пацієнта. Все це змушує лікарів замислюватись над тим, яким чином проаналізувати й коректно обробити такі великі обсяги медичної інформації з метою прийняття правильних рішень щодо діагностики, вибору адекватного лікування, прогнозування перебігу лікування тощо. Аналіз й оброблення такої інформації без використання статистичних методів і сучасних комп'ютерних технологій став просто неможливим, оскільки буде стримувати не лише подальший розвиток медичної науки, а й практичне надання невідкладної допомоги.

Насамперед з'ясуємо сутність статистичного підходу, що полягає у заміні дослідження великої множини об'єктів дослідженням значно меншої її частини та подальшому «поширенні» результатів дослідження на всю множину, зробивши відповідні висновки щодо властивостей так званої генеральної сукупності в цілому. Звичайно ж, ці висновки повинні бути обґрунтованими і достовірними, що, в першу чергу, залежить від якості даних, які використовуються, від їх достовірності та точності, адже саме від первинного матеріалу залежить достовірність аналізованих

результатів проведених наукових досліджень та зроблених на їх основі практичних рекомендацій щодо використання останніх в практичній медицині. Оскільки збір даних лежить в основі всього дослідження, то на сьогодні перед статистичною наукою постають актуальні проблеми щодо подальшого вдосконалення системи показників, прийомів і методів збору, оброблення, зберігання та аналізу статистичної інформації. Від дослідника вимагається провести коректне статистичне оброблення вихідного матеріалу, суть якого полягає в тому, щоб зібрати необхідний масив валідних даних про масові явища, представити їх у формі, зручній для аналізу з допомогою комп'ютерних програм, обробити їх, використовуючи адекватні математичні методи. В зв'язку з цим, на сьогоднішній день значно підвищуються вимоги до вивчення прикладної статистики, яка є невід'ємною частиною формування наукового підходу у наукових співробітників, студентів, аспірантів, лікарів-фахівців, які планують і проводять дослідження у сфері медицини.

Не дивлячись нате, що в останні роки широко використовуються програмні засоби для статистичного аналізу даних у різних прикладних галузях, включаючи медичні додатки, необхідність володіння хоча б основами статистики та математичного апарату стає все актуальнішою. Користувач пакетів прикладних програм з аналізу й оброблення статистичних даних повинен вміти грамотно вибирати відповідні статистичні процедури, знати їх можливості та обмеження, коректно й осмислено інтерпретувати одержані результати, оскільки довільне застосування статистичних методів може призвести до помилкових висновків [3]. Недостатня увага до планування досліджень тягне за собою нестачу даних для формування статистично значущого висновку після закінчення етапу збору інформації. У цьому випадку навіть найскладніші математичні методи аналізу отриманих результатів не зможуть дати необхідноїдослідникуінформації [2].

Мета роботи. Розробити методичний підхід, в основу якого покладено алгоритм, що вказує на послідовність дій, які слід виконувати дослідникупри обробленні, аналізі та інтерпретації одержаних результатів наукового дослідження.

Для реалізації окресленої мети вирішували такі завдання:

1. Проаналізувати критерії, які найчастіше використовуються в медико-соціологічних дослідженнях для отримання значимої інформації з позиції доказової медицини.

2. Розробити алгоритм медико-соціологічного дослідження на підставі обраних методів оброблення й аналізу одержаної інформації, який допоможе дослідникусистематизувати свої знання й вибрати оптимальне поєднання методів оброблення одержаної інформації.

Результати й обговорення. Запропонований алгоритм (рис. 1) передбачає, насамперед, опис даних науково-практичного медичного дослідження, що включає в себе ряд одиниць спостереження (хворих, лабораторних піддослідних тварин тощо), які характеризуються певними ознаками. Аналізуючи набір даних, слід чітко визначити їх типи. Кількісні ознаки виражаються числовими значеннями, наприклад, вік, зріст, вага, тиск. Порядкові ознаки можуть бути виміряні в шкалах (наприклад, шкільні оцінки, ступінь тяжкості захворювання: легкий, середній, тяжкий). Якісні ознаки характеризують деякий стан об'єкта, але не можуть бути виміряні кількісно (наприклад, стать, професія, діагноз). Включаючи ознаку в опис даних, дослідник повинен досить чітко уявити, для чого ця ознака знадобиться йому в подальшому. Це необхідно, щоб уникнути перевантаженості інформації, проте база даних повинна бути достатньо повною й інформативною.

Залежно від типу даних вибираємо метод описових статистик. Тут надзвичайно важливо знати, чи підпорядковуються кількісні дані нормальному закону розподілу, адже знання щодо нормального розподілу важливе з багатьох причин.

Розподіл багатьох статистик є нормальним або може бути отриманий із нормальних за допомогою деяких перетворень. Міркуючи філософськи, можна сказати, що нормальний розподіл являє собою одну з емпірично перевірених істин щодо загальної природи дійсності і його положення може розглядатися як один із фундаментальних законів природи. Точна форма нормального розподілу (характерна «дзвоноподібна крива») визначається тільки двома параметрами: середнім значенням (M) і стандартним відхиленням (±д). Характерна властивість нормального розподілу полягає у тому, що 68 % усіх його спостережень лежать у діапазоні ±1 стандартне відхилення від середнього, а діапазон ±2 стандартних відхилення містить 95 % значень. Іншими словами, при нормальному розподілі, стандартизовані спостереження, менші -2 або більші +2, мають відносну частоту менше 5 %. Якщо ж значення ознаки розподілені несиметрично щодо середнього, то сукупність краще описати за допомогою медіани і квартилів (процентилів). Медіана є середньою

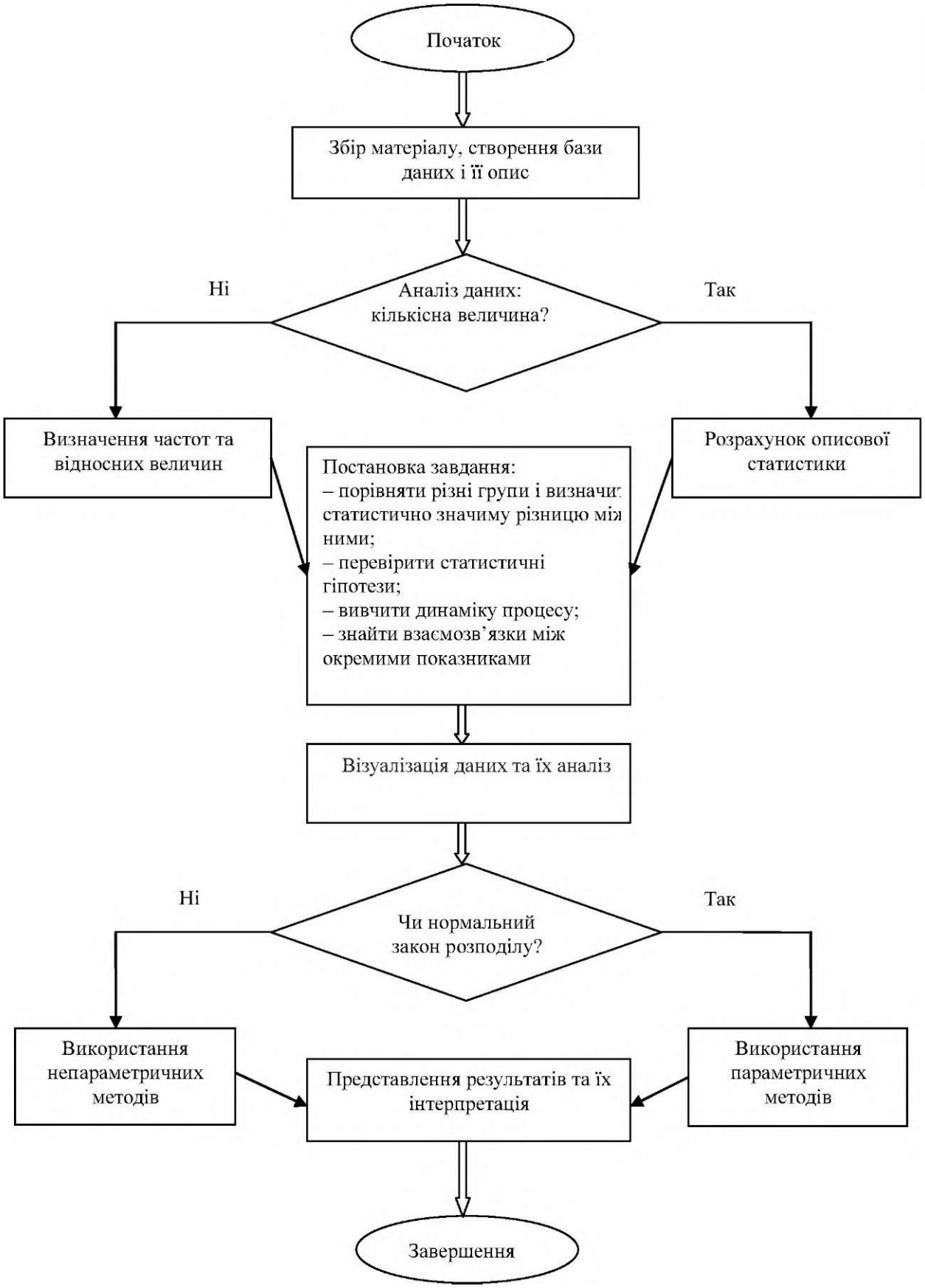

Рис. 1. Алгоритм аналізу й оброблення даних наукового дослідження.

точкою варіаційного ряду, тому вона не так чутлива до «викидів». Медіана вибірки розбиває її на дві рівні частини. Половина спостережень лежить нижче медіани, а половина - вище. Якщо ж розподіл несиметричний (зсунутий вліво або вправо), то медіана і міжквартильний розмах можуть дати більше інформації про те, в якій галузі концентруються спостереження. Якщо медіана менша середнього, то розподіл зсунутий вправо, якщо медіана більша середнього, то розподіл зсунутий вліво. Звичайно, така схема вибору приймається за умови, що розподіл одномодальний, тобто має одну моду. Якщо дані катетеризовані, то краще використовувати моду, тобто значення змінної, яка найчастіше зустрічається. Якщо не всі наявні значення змінної становлять інтерес, розподіл несиметричний і є «викиди», то краще використовувати медіану. У протилежному випадку слід працювати із середнім значенням.

Отже, для кількісних даних, які підкоряються нормальному закону розподілу, найчастіше розраховують середнє значення величини (M), стандартну похибку (±т) і стандартне відхилення (±д). Розраховані описові статистики представляють у вигляді M±т або M±д. Для кількісних даних, що не підкоряються нормальному закону розподілу, розраховують медіану і квартилі. Такі описові характеристики представляються у вигляді Me (Q1—Q3) (наприклад, вік учасників дослідження становив 20 (15-22) років). Для якісних даних розраховуються абсолютні частоти або відносні величини (відсотки) та їх похибки. Представляти їх треба або тільки у відносних величинах (наприклад, захворювання «Х» у досліджуваній сукупності склало (27±2,1) %), або в абсолютних та відносних разом (наприклад, виявлено 16 випадків захворювань «Х», що склало (27±2,1) %).

Після описових статистик дослідник формулює завдання, які необхідно вирішити в процесі проведення наукового дослідження, тобто уточнити статистичні гіпотези та перевірити їх, конкретизувати, які групи підлягають порівнянню і визначити статистично значиму різницю між ними, визначитись з показниками, між якими слід знайти кореляційні взаємозв'язки тощо.

Наступний етап запропонованого алгоритму передбачає вибір статистичного критерію, що являє собою суворе математичне правило, за яким приймається чи відхиляється та або інша статистична гіпотеза. Вибір методу критеріального аналізу залежить від багатьох чинників і, насамперед, від типу досліджуваної ознаки, її розподілу, виду дослідження тощо [1,3,4]. Для того, щоб орієнтуватися у виборі необхідного критерію, існує декілька алгоритмів [1,5]. Алгоритми вибору критеріїв, що залежать від типу даних, представлені на рисунках 2,3.

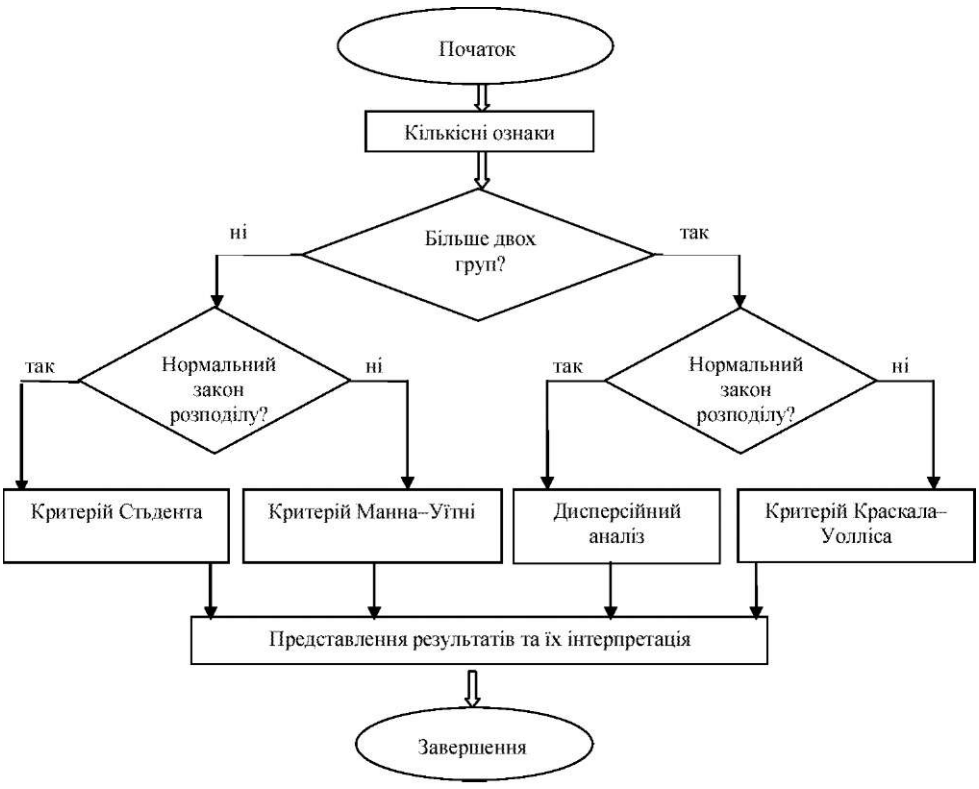

Рис. 2. Алгоритм порівняння двох і більше груп за кількісною ознакою.

Рис. 3. Алгоритм порівняння груп за якісною ознакою.

При виборі критерію для перевірки статистичної гіпотези щодо виявлення статистично значимої різниці між окремими ознаками, що виражені кількісними значеннями, необхідно перевірити їх розподіл (нормальний чи ні) і визначити кількість порівнюваних груп (дві чи більше), оскільки ці факти впливають на проведення критеріального аналізу (рис. 2). У випадку, коли дані в порівнюваних групах мають нормальний розподіл, застосовується параметричний критерій Стьюдента, при значному відхиленні розподілу від нормального слід використовувати непараметричні критерії: для залежних груп - критерій Манна-Уїтні (Вілкоксона), для незалежних - критерій Уайта.

Досить часто у дослідників виникає питання: чи можна використовувати непараметричні критерії при нормальному розподілі? Так, можна, але при цьому слід пам'ятати, що параметричні критерії мають більшу статистичну потужність, аніж непараметричні при нормальному розподілі.

Інше питання: чи можна використовувати параметричні критерії при відхиленні розподілу від нормального? Не рекомендується, оскільки при наявності великих вибірок критерій Стьюдента досить стійкий до невеликих відхилень розподілу від нормального, то при малих вибірках його застосування для скошених розподілів може призвести до спотворення результатів.

При виборі статистичного критерію для якісних ознак необхідно, насамперед, визначити кількість порівнюваних показників. Якщо визначається статистично значима різниця між двома показниками у двох групах, які можна представити так званою таблицею 2х2 (рис. 4), тоді можна використати кілька критеріїв: хі-квадрат, точний критерій Фішера, Макнемара (рис. 5).

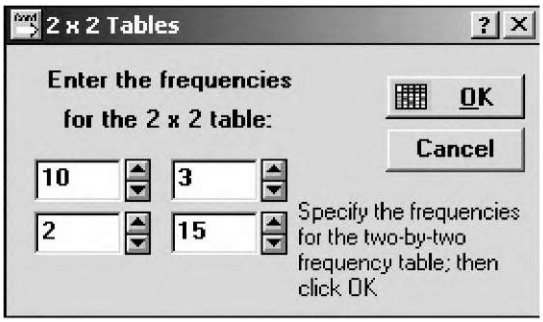

Рис. 4. Представлення якісних ознак в пакеті STATISTIC A таблицею 2х2.

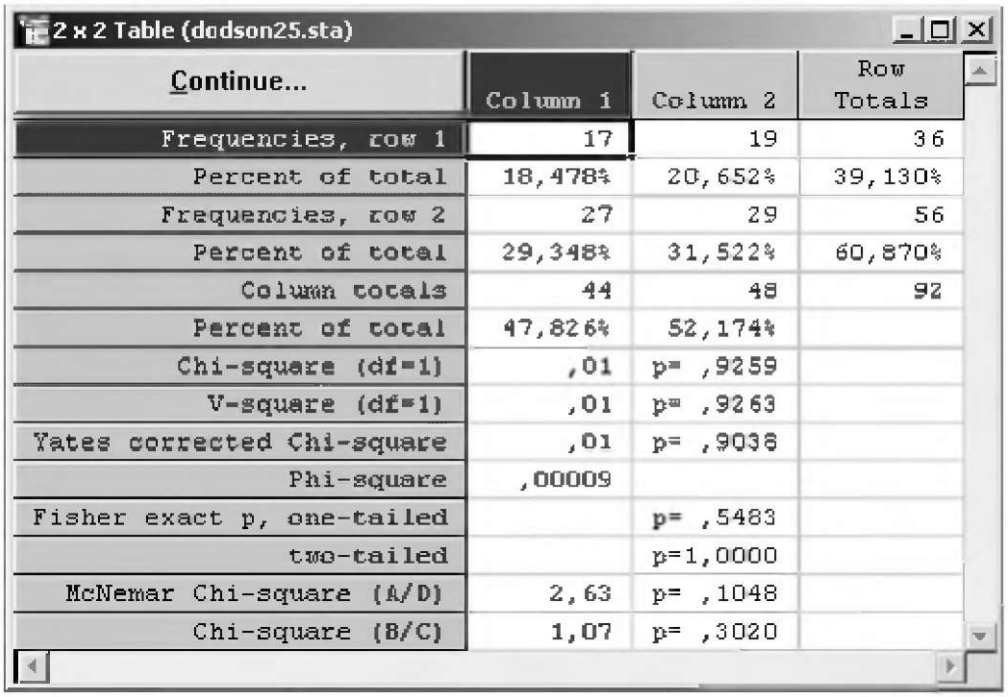

Рис. 5. Результат аналізу якісних ознак, представлених таблицею 2х2, у пакеті STATISTICA (модуль «Непараметрична статистика»).

Для якісних ознак, що не можуть бути представлені таблицею 2х2, можна використовувати цілу низку статистичних критеріїв непараметричної статистики (рис. 6).

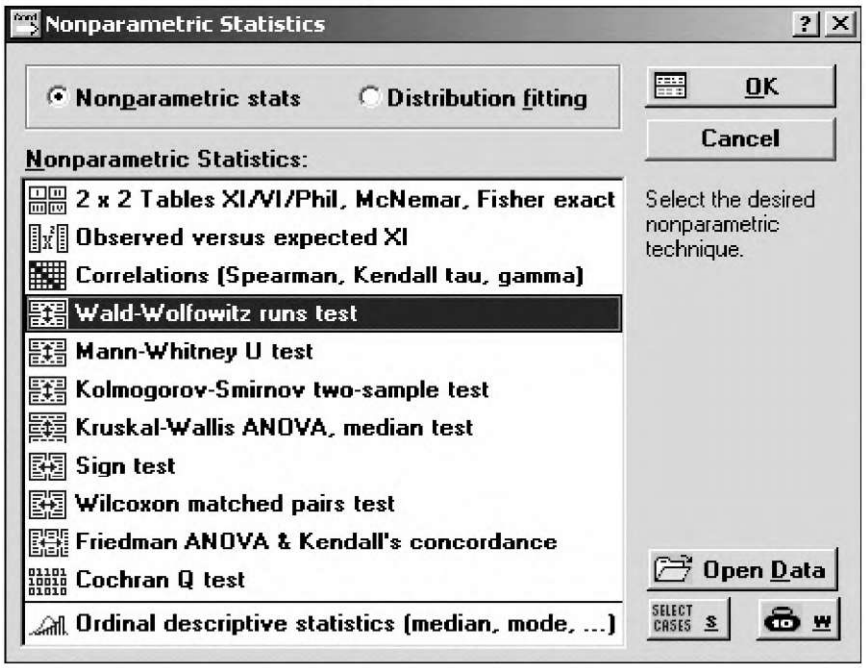

Рис. 6. Непараметричні статистичні критерії в пакеті STATISTICA (модуль «Непараметрична статистика»).

Безліч критеріїв, наведених у підручниках з математичної статистики, багато з яких використовуються досить рідко, часто бентежать дослідника. Кожен дослідник вибирає статистичні критерії виходячи зі своїх знань, досвіду, типу завдання і виду даних, які підлягають обробці. При виборі математико-статистичного критерію потрібно орієнтуватися також на тип розподілу даних дослідження, оскільки при нормальному розподілі отриманих даних використовуються параметричні критерії, а при використанні непараметричних критеріїв тип розподілу даних не має значення. При визначенні статистично значимої різниці між групами в медико-соціологічних дослідженнях слід звернути увагу на те, які ці групи, а саме: незалежні (наприклад, контрольна й експериментальна група) чи залежні (група до проведення експерименту і після), оскільки це також має значення при виборі критеріїв порівняння.

Здійснюючи пошук взаємозв'язку між окремими показниками слід використовувати коефіцієнти кореляції. Зв'язок між двома кількісними показниками характеризується парним коефіцієнтом Пірсона, якщо вони мають нормальний закон розподілу, в іншому випадку розраховується коефіцієнт рангової кореляції Спірмена чи Кендалла.

Слід зауважити, що для дослідження впливу, а тим більше взаємовпливу декількох факторів на досліджуваний параметр доцільніше використовувати дисперсійний аналіз. Дослідник виходить з припущення, що одні змінні можуть розглядатися як причини, а інші - як наслідок. Змінні першого роду вважаються факторами, а змінні другого роду - результативними ознаками. У цьому відмінність дисперсійного аналізу від кореляційного, в якому передбачається, що зміни однієї ознаки просто пов'язані з певними змінами іншої [2, 3,4].

Важливо звернути увагу на обмеження, які має кожен критерій. Якщо один критерій не підходить для аналізу наявних даних, завжди можна знайти якийсь інший. Вибір і реалізація методу аналізу у зв'язку з їх різноманіттям може виявитися завданням нетривіальним. Проте використання комп'ютерних технологій для обробки й аналізу даних дозволяє вирішити це завдання кількома подібними методами і вибрати той, котрий дає найкращий результат. І дійсно, у сучасних пакетах прикладних програм внесені дані досить просто обробити з використанням різних критеріїв і процедур, а потім можна вибрати ті, що дають найкращі результати.

Заключним етапом технології аналізу даних є інтерпретація і подання результатів аналізу. Дуже важливе значення мають повнота й рівень опису, як самого аналізу, так і його результатів та їхньої інтерпретації. Так, при інтерпретації результатів статистичної обробки даних завжди необхідно пам'ятати про їхній імовірнісний зміст, суть якого полягає в тому, що не завжди отримані результати є точними, а лише статистичними оцінками істотних значень.

Висновок. Правильний методичний підхід щодо вибору методу статистичної обробки й аналізу даних медико-соціологічних досліджень, використання комп'ютерів і широке впровадження сучасних інформаційних технологій дає можливість складні процедури обчислення досить великих обсягів медичної інформації зробити набагато простішими і провести цей аналіз на високому науковому рівні з метою прийняття правильних рішень щодо діагностики, вибору адекватного лікування, прогнозування перебігу лікування тощо.

Література

1. Гланц С. Медико-биологическая статистика / С. Гланц. - М. : Практика, 1999. - 461 с.

2 Гойко О. В. Практичне використання пакета STATISTICA для аналізу медико-біологічних даних: навч. посібник / О. В. Гойко. - К. : КМАПО імені П. Л. Шупика, 2004. - 76 с.

3. Жилина Н. М. Приложения математической статистики к медицинским научным исследованиям : учебное пособие/Н. М. Жилина. - Новокузнецк: МОУ ДПО ИПК, 2005.-41 с.

4. Мінцер О. П. Оброблення клінічних і експериментальних даних у медицині : навч. посібник / Ю. В. Вороненко, О. П. Мінцер, В. В. Власов. - К.: Вища школа, 2003. - 350 с.

5. Сергиенко В. И. Математическая статистика в клинических исследованиях / В. И. Сергиенко, И. Б. Бондарева - М. : ГЭОТАР МЕДИЦИНА, 2000. - 256 с.