УДК 519.226

Семенюк А. М., здобувач вищої освіти;
Хмелівський Ю. С., асистент кафедри інформаційних технологій,
Донецький національний університет імені Василя Стуса

СТАТИСТИЧНИЙ АНАЛІЗ МЕДИЧНИХ ДАНИХ НА МОВІ R

Ключові слова: BIG DATA, мова R, статистична інформація, медицина.

Вступ. У роботі досліджуються основні етапи попереднього аналізу великих масивів даних, пов'язаних із роботою медико-соціальних експертних комісій (МСЕК). Аналіз звітної медичної інформації із застосуванням статистичних методів вимагає умілого підходу до вибору об'єкта аналізу, елементарної одиниці контролю та її ознак. Для реалізації методів пропонується використовувати пакети аналізу та візуалізації мови наукових розрахунків **R.**

Актуальність. Мова наукових розрахунків **R** широко застосовується для проведення різноманітних досліджень і містить засоби реалізації технології кластерного аналізу у вигляді пакетів. У медичній практиці, а особливо у медичних дослідженнях, часто застосовують різноманітні методи аналізу й обробки даних математичної статистики, оскільки математичні методи дають змогу об'єктивно оцінювати кількісні результати досліджень.

Проблема статистичного аналізу медичних даних у наш час є надзвичайно актуальною. Неймовірне збільшення обсягів інформації, пов'язане з інформатизацією медичної галузі, потребує неординарних рішень для її обробки та інтерпретації результатів. Останніми роками методи й засоби аналізу даних зазнали принципових змін, а представлення інформації у зручному вигляді взагалі перетворилось на мистецтво.

Статистична інформація, яка подана у вигляді таблиць, пов'язаних між собою, вимагає таких процедур їх обробки й аналізу:

- встановлення потенційно можливих закономірностей та зв'язків між окремими компонентами;
- наявність можливостей передбачення нових фактів.

Реалізація цих вимог неможлива без представлення даних у комп'ютерному форматі з подальшою обробкою їх у цифровому вигляді.

Якщо предметом статистичного вивчення стають якісно різнорідні показники, то розуміння їх, отримані без попереднього групування за якісними ознаками, не відповідають об'єктивній дійсності. Наприклад, нерозділення осіб за віковими критеріями, за місцем проживання, робочими професіями тощо, тобто на групи соціальної неоднорідності здоров'я, призводить до спотворення висновків.

Використання мови програмування **R** дасть змогу дослідити отримані результати роботи МСЕК. Методи регресійного аналізу дадуть змогу виявити й дослідити залежності між різними показниками, спрогнозувати майбутні тенденції, виконати складні обчислення в галузі медичної експертизи. Кластерний аналіз, який можливо виконати на мові **R,** дасть змогу виокремити

групи пацієнтів за їхніми поведінковими характеристиками. Також цей ПЗ дасть змогу виконувати моделювання з метою прогнозування майбутніх подій і розроблення ефективних лікувально-профілактичних процедур. Паралельно ми отримаємо ранжування груп даних і даних у середині груп.

Якщо серед ранжованих значень декілька потрапляють до однієї градації, то тоді всім їм приписують однаковий ранг, який розраховують за формулою:

$$R_n(x) = \sum_{i=0}^{n-1} y_i + \frac{y_n + 1}{2}, \quad (1)$$

де n – номер градації;

R_n – ранг кожного значення ознаки, що потрапив до градації i;

y_n – кількість значень, що потрапили до градації n (y_0 приймається таким, що дорівнює 0).

Перевірка правильності ранжування здійснюється таким шляхом: знаходимо суму всіх рангів і порівнюємо з перевірочною сумою, яку визначаємо за формулою:

$$S_R^T = \frac{N(N+1)}{2}. \quad (2)$$

Отже, операція ранжування дасть змогу перейти від якісних ознак до кількісних ознак. Якщо провести аналіз за різними обліковими ознаками (вік, стать, місце проживання, професія, захворювання тощо), це дасть змогу дослідити не тільки кожен елемент сукупності, але і всю сукупність загалом.

Пакет візуалізації мови **R** дає змогу видавати результат у вигляді графіків кореляції показників та різноманітних гістограм.

Висновки

У роботі представлено технологію попередньої обробки великих масивів даних та проведено аналіз сучасних методів кластеризації складних об’єктів. Використання мови R надає широкі можливості для здійснення статистичних аналізів, які включають: лінійну і нелінійну регресію, класичні статистичні тести, аналіз часових рядів (серій), кластерний аналіз та ін. Мова R завдяки використанню додаткових функцій і пакетів легко перебудовується на різні типи задач.

Список використаних джерел

1. Основні показники медико-соціальної реабілітації осіб з інвалідністю в Україні за 2022 рік / В. І. Шевчук, Р. Я. Перепелична, Л. О. Сторожук, І. В. Куриленко, Л. Г. Семененко, М. В. Семенюк, А. М. Семенюк. *Аналітико-інформаційний довідник*. Вінниця: ФОП Данилюк В. Г., 2023. 119 с.

2. Основні показники інвалідності та діяльності медико-соціальних експертних комісій України за 2020 рік / А. В. Іпатов, О. М. Мороз, І. Я. Ханюкова, Н. О. Гондуленко, Н. А. Саніна, А. М. Ульянова. *Аналітико-інформаційний довідник*. Дніпро, Акцент ПП, 2021. 188 с.