УДК 519.226

А. М. Семенюк[1]

МОВА "R" І СТАТИСТИКА В МЕДИЦИНІ

[1] Донецький національний університет імені Василя Стуса, м. Вінниця;

Анотація

Тема роботи пов'язана з використанням пакету аналізу та візуалізації мовою R для статистичної обробки даних у медичній сфері. Цей інструмент дозволить групувати та ранжувати вихідні дані, конвертувати якісні показники в кількісні та візуалізувати отримані результати. Важливими характеристиками мови R, які призвели до його швидкого поширення, є стабільне ядро і проста система розширення можливостей за рахунок підключення додаткових пакетів, доступних для скачування.

Ключові слова: мова R; аналіз статистичних даних в медицині; моделювання процесів.

Abstract

The topic of the work is related to the use of the R language analysis and visualization package for statistical data processing in the medical field. This tool will allow grouping and ranking of initial data to convert qualitative indicators into quantitative ones, and visualization of the obtained results. Important characteristics of the "R" environment, which led to its rapid spread, are a stable core and a simple system of extending capabilities by connecting additional packages available for download.

Keywords: R language; analysis of statistical data in medicine; process modeling.

Вступ

З поширенням цифровізації досліджень неймовірно збільшують обсяги даних які підлягають аналізу. Це веде до пошуку неординарних рішень для її обробки та інтерпретації результатів, подачу інформації у зручному вигляді. Одним із таких варіантів є використання методів математичної статистики, які останнім часом почали використовувати в різних "нетрадиційно математичних" галузях:

- в соціально-економічних та інших дослідженнях;
- прийнятті різноманітних управлінських рішень в умовах певної невизначеності;
- для перевірки висунутих наукових гіпотез;
- побудови математичних моделей різноманітних об’єктів та явищ, що відбуваються в природі та суспільстві;
- для аналізу і визначенню експертних висновків на основі статистичного дослідження медичних даних [1].

Застосування подібних математичних засобів дозволяє об’єктивно оцінювати кількісні результати досліджень і експериментів, а візуалізація даних полегшує сприйняття інформації.

Статистичний аналіз інформації

Дослідження медичної інформації із застосуванням статистичних методів вимагає умілого підходу до вибору об'єкту розгляду, елементарної одиниці контролю та її ознак. Для вирішення цих задач пропонується використовувати мову "R" – пакет створений для аналізу та візуалізації наукових розрахунків.

Отримання при медичних обстеженнях значних об’ємів інформації та формування при цьому великої кількості таблиць потребує наступних процедур для їхньої обробки і аналізу:

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

Якщо предметом статистичного вивчення стають якісно різнорідні показники, то їх розуміння, отримані без попереднього групування за якісними ознаками, не відповідають об’єктивній дійсності. Наприклад не розділення осіб за віковими критеріями, по місцю проживання, робочим професіям, тощо, тобто на групи соціальної неоднорідності здоров’я, веде до спотворення висновків.

У мові програмування "R" для попередньої підготовки даних існують методи якісного аналізу. Це дозволяє виконувати моделювання, з метою прогнозування майбутніх подій, і розроблення ефективних лікувально-профілактичних процедур [2]. Серед таких методів найбільш поширені наступні:

- регресійний аналіз – виявлення та дослідження функціональних залежностей між різними показниками, прогнозування майбутніх тенденцій, виконання складних обчислень;
- кластерний аналіз – виокремлювання групи пацієнтів за їхніми поведінковими характеристиками.

В результаті буде отримане формування послідовних груп даних, а також ранжування в середині кожної групи.

Математичний апарат подібного статистичного аналізу має наступний вигляд:

1) якщо в групі значень декілька з них потрапляють до однієї градації, то таким параметрам 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");

2) для перевірка правильності розміщення знаходимо суму всіх рангів і порівнюємо з перевірочною сумою, яку визначаємо за формулою:

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

Таке перетворення дасть можливість "замінити" якісні ознак параметра на кількісні ознаки, тобто отримуємо набір статистичних даних.

Більшість процедур в середовищі програмування "R" реалізовані з використання підпрограм-функцій. Вони ґрунтуються на трьох складових:

- переліку формальних аргументів;
- тілом функції;
- оточенням.

Перелік аргументів складається з назв змінних, при цьому їм можуть бути присвоєні значення за замовченням у вигляді "аргумент = значення".

Модель даного процесу у вигляді "умовної чорної скриньки" – пакету досліджень на мові "R" представлено на рис. 1.

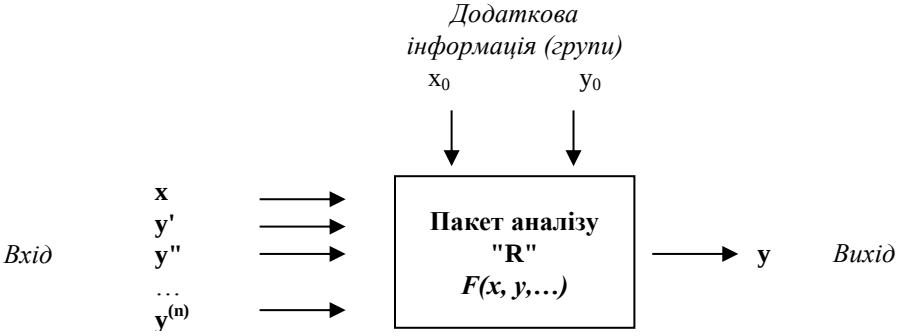

Рис. 1 – Модель процесу формування статистичних даних

Подібна кластеризація складних об’єктів створює технологію попередньої обробки великих масивів даних та проведенню їх аналізу та представленню сучасними методами.

Використання мови "R" надає широкі можливості для здійснення статистичних аналізів [3], які включають: лінійну і нелінійну регресію, класичні статистичні тести, аналіз часових рядів (серій), кластерний аналіз і т. д. Мова R завдяки використанню додаткових функцій і пакетів легко перебудовується на різні типи задач.

Важливими характеристиками середовища "R", які зумовили його стрімке розповсюдження, є стабільне ядро та проста система розширення можливостей за допомогою підключення додаткових пакетів, доступних для скачування. Базові пакети середовища включають цілий спектр функцій для визначення узагальнюючих характеристик масиву даних, зокрема, характеристики центральної тен-

денції та варіації.

Статистичні данні для зручності зображають за допомогою статистичних таблиць та статистичних графіків. Графічне зображення даних можна реалізовувати досить ефективно, для цього наявна значна кількість різних процедур.

З появою мови програмування "R", дешевого та легкого доступу до аналізу даних, відбулася зміна парадигми обробки інформації. Замість попередньої установки всіх параметрів аналізу, процес став значною мірою інтерактивним. Водночас результати кожного етапу аналізу слугує вхідними даними для подальшого етапу.

Висновки

З появою мови програмування "R", простого та легкого доступу до аналізу даних, відбулася зміна парадигми обробки інформації. Замість попередньої установки всіх параметрів аналізу, процес став значною мірою інтерактивним. Водночас результати кожного етапу аналізу слугує вхідними даними для подальшого етапу.

У мові "R" наявні інструменти наступні методи:

- дослідницького факторного аналізу EFA – дослідження прихованої факторної структури без попередніх відомостей щодо кількості факторів та навантажень (кореляцій між початковою змінною та фактором);
- підтверджуючого факторного аналізу CFA – перевірки гіпотез щодо факторів та навантажень.

Ще одним фактором який сприяє використанню мови "R" – відкритість коду та вільнодоступність поширення вихідного коду за ліцензією GNU General Public License. А розроблені різні графічні інтерфейси користувача дозволяють використовувати продукт не підготовленим в галузі програмування спеціалістам.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

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. Семенюк А. М., Хмелівський Ю. С., Статистичний аналіз медичних даних на мові R, Прикладні аспекти сучасних міждисциплінарних досліджень: матеріали II Міжнародної науково-практичної конференції (м. Вінниця, 24 листопада 2023 р.). Вінниця: ДонНУ імені Василя Стуса. 2023. 282 с.
3. Методи програмування в R. [Електронний ресурс] – URL: https://tvimc.jimdofree.com

Семенюк Андрій Михайлович — студент групи КН-21-Б2, Факультет інформаційних і прикладних технологій, Донецький національний університет імені Василя Стуса, Вінниця, e-mail: sam12122003@gmail.com

Науковий керівник: ***Хмелівський Юрій Сергійович*** — асистент, кафедра інформаційних технологій, Донецький національний університет імені Василя Стуса, м. Вінниця

Semeniuk Andriy M. — student of KN-21-B2, Faculty of Information and Applied Technologies, Vasyl' Stus Donetsk National University, email : sam12122003@gmail.com

Scientific supervisor: ***Khmelivskyi Yuryi S.*** — assistant, Department of Information Technologies, Vasyl Stus Donetsk National University, Vinnytsia