Катерина Наконечна
к.е.н., доцент кафедри економічної кібернетики
НУБіП України
ORCID: orcid.org/0000-0002-1537-7201
kln273125@gmail.com
Метелиця А.М
студентка

УПРАВЛІННЯ ДАНИМИ

Використання статистики зазвичай має на увазі аналіз даних, а надійність статистичних результатів багато в чому залежить від надійності проаналізованих даних, таким чином при використанні статистики, потрібно бути обізнаним щодо управління даними. Аналіз публікацій засвідчує що цьому питанню присвячено досить мало уваги.

Використовувати коректні дані та вибрати правильний метод їх обробки - це завдання аналітика, оскільки статистична програма тільки виконує задані вами операції і не може оцінити ні якість даних, ні адекватність застосовуваної процедури. За будь-яких обставин вам може здатися корисним розуміти на базовому рівні, що відбувається при управлінні даними, і знати, що може статися, якщо це не зроблено правильно.

Збір даних та їх введення в комп'ютер проводиться людьми, які раз у раз роблять помилки. Більша частина процесу управління даними полягає у виявленні цих помилок або їх виправленні, або винаході способу обійти їх, щоб дані можна було нормально обробити.

Необхідно відзначити, що для ефективного управління даними в ході великого проекту необхідно визначити структуру, або ієрархію, людей, які відповідають за різні частини процесу. Важливо, щоб кожен учасник проекту знав, хто уповноважений приймати певні рішення, тому, коли проблема з'являється, її можна вирішити швидко і розумно.

Кодифікатор повинен містити інформацію як мінімум на наступні теми: інформація про проект та методи збору даних; методи введення даних у комп'ютер; рішення, ухвалені щодо даних; процедури кодування.

Унікальний ідентифікатор корисний для того, щоб підтвердити відсутність записів, що повторюються, виявити загальні записи для однієї одиниці дослідження (наприклад, всі відвідування лікарні однією людиною) і запобігти перемішування записів для різних людей. У великому файлі можуть фігурувати, наприклад, кілька Олександрів Петренків, і ви не хотіли б, щоб записи про них змішалися. Так Олександр Петренко міг приходити до лікарні п'ять разів на рік; переглядаючи його історію хвороби, ви хочете легко виявити всі записи, що відносяться до нього.

Перевірка того, що всі значення, назви змінних та підписи є вірними, - це наступний етап обстеження файлу. Збереження правильних значень найважливіше, оскільки назви та підписи можна створити заново, однак самі дані мають бути правильними, а в процесі перетворення файлу може статися багато несподіванок.

Наступний етап - проаналізувати самі значення набору даних і зрозуміти, чи вони правдоподібні. Деякі прості статистичні процедури (такі як обчислення середнього та дисперсії числових змінних) допомагають переконатися, що значення не змінилися при перетвореннях (за умови, що ви знаєте значення середнього та дисперсії даних до їхнього перетворення). Дати потрібно перевіряти особливо акуратно, оскільки вони є особливо частим джерелом проблем через те, що в різних програмах дати представлені в різних форматах.

Оцінка кількості пропущених даних та їх закономірності. Ваше перше завдання - це виявити поширеність пропущених даних, це можна зробити за допомогою аналізу

частот значень. Друге завдання - вивчення закономірностей пропуску даних у багатьох змінних. Наприклад, чи є такі змінні, значення яких відсутні найчастіше?

Пропущені дані створюють дві основні проблеми. Вони зменшують кількість випадків, придатних для аналізу, знижуючи таким чином статистичну потужність, а також вони можуть бути джерелом систематичної помилки.

Наведемо шляхи вирішення проблеми: докласти додаткових зусиль для заповнення пропущених даних, з'ясувавши причину їх відсутності; застосувати інший спосіб аналізу даних, такий як багаторівнева модель, замість класичної моделі повторних вимірів; відновити пропущені значення за допомогою методів найбільшої правдоподібності на кшталт тих, що доступні в модулі MVA програми SPSS, або використовувати методи множинного заміщення пропущених значень, реалізовані в таких програмах, як SAS пропущених даних вони заміщаються на значення, засновані на існуючих даних, в результаті чого ми отримуємо повний набір даних; створити додаткову змінну (0, 1) для позначення пропущених даних поряд із заміщенням пропущених даних; видалити рядки або стовпці з великою кількістю пропущених даних. (Це припустимо, тільки якщо проблема полягає у невеликому числі рядків та/або стовпців, які не дуже важливі для вашого аналізу, і це може стати джерелом систематичної помилки, якщо дані пропущені не цілком випадково.);використовувати заміщення за умови, замінюючи пропущені значення на наявні (не рекомендується, оскільки може призвести до заниження дисперсії); використовувати просте заміщення, замінивши пропущені значення, наприклад, середнім значенням (не рекомендується, оскільки майже завжди призводить до сильної недооцінки дисперсії).

Література

1. Жерліцин Д.М., Наконечна К.В., Галаєва Л.В. Статистичний аналіз і візуалізація даних. Навчальний посібник. Київ. Компринт 2022, 267 с.
2. Жерліцин Д.М., Наконечна К.В. Прикладна статистика для економічного обгрунтуваннння інженерних рішень. Київ.- Компринт 2023. 232 с.

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE

NATIONAL UNIVERSITY OF LIFE AND ENVIRONMENTAL SCIENCES OF UKRAINE

FACULTY OF INFORMATION TECHNOLOGY

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

PROCEEDINGS

МАТЕРІАЛИ

XI International scientific conference

XI Міжнародної науково-практичної конференції

GLOBAL AND REGIONAL PROBLEMS OF INFORMATIZATION IN SOCIETY AND NATURE USING '2023

ГЛОБАЛЬНІ ТА РЕГІОНАЛЬНІ ПРОБЛЕМИ ІНФОРМАТИЗАЦІЇ В СУСПІЛЬСТВІ І ПРИРОДОКОРИСТУВАННІ '2023

15-16 November 2023

Kyiv, NULES of Ukraine

15-16 листопада 2023 року

Київ, НУБіП України

Kyiv 2023

Київ 2023

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ
І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

МАТЕРІАЛИ

XI Міжнародної науково-практичної конференції

ГЛОБАЛЬНІ ТА РЕГІОНАЛЬНІ ПРОБЛЕМИ ІНФОРМАТИЗАЦІЇ В СУСПІЛЬСТВІ І ПРИРОДОКОРИСТУВАННІ '2023

15-16 листопада 2023 року

Київ, НУБіП України

Київ 2023

УДК 004

Рекомендовано до друку вченою радою факультету інформаційних технологій Національного університету біоресурсів і природокористування України (протокол № 4 від 20.11.2023)

Укладач: к.е.н., доцент Харченко В.В.

Збірник матеріалів XI Міжнародної науково-практичної конференції "Глобальні та регіональні проблеми інформатизації в суспільстві і природокористуванні '2023", 15-16 листопада 2023 року, НУБіП України, К. НУБіП України, 2023. 117 с.

Відповідальність за зміст публікацій несуть автори.