ДНІПРОВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ОЛЕСЯ ГОНЧАРА

Механіко-математичний факультет

Кафедра статистики й теорії ймовірностей

**КУРСОВА РОБОТА**

другий (магістерський) рівень вищої освіти

спеціальність 112 Статистика

**РАНДОМІЗОВАНИЙ КРИТЕРІЙ ПОРІВНЯННЯ СЕРЕДНІХ ДВОХ ГРУП**

Виконавець

студент групи МС-23м-1

Роман ЩЕРБАК

Керівник

доцент каф. МСТ, к. ф.-м. н.,

Євген КАРНАУХ

Завідувач випускової кафедри

к.ф.-м.н. Валерій ТУРЧИН

Дніпро

2024

**ЗМІСТ**

[ВСТУП 3](#_Toc166846778)

[**Актуальність теми дослідження** 3](#_Toc166846779)

[**Завдання роботи** 3](#_Toc166846780)

[РОЗДІЛ 1. ТЕОРЕТИЧНА ЧАСТИНА 5](#_Toc166846781)

[**1.1** **Постановка задачі** 5](#_Toc166846782)

[**1.2** **Лейкемії та їх типи** 5](#_Toc166846783)

[**1.3** **Гени та рівень їх експресії** 5](#_Toc166846784)

[**1.4** **Статистичні критерії порівняння середніх** 7](#_Toc166846785)

[РОЗДІЛ 2. Практична частина 12](#_Toc166846786)

[**2.1** **Аналіз даних** 12](#_Toc166846787)

[**2.2** **Опис методу** 12](#_Toc166846788)

[**2.3** **Результат роботи програми** 13](#_Toc166846789)

[ВИСНОВКИ 14](#_Toc166846790)

[СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 15](#_Toc166846791)

[ДОДАТОК А. Код програми мовою R 16](#_Toc166846792)

# ВСТУП

**Актуальність теми дослідження**

Гострий лімфобластний лейкоз (Acute lymphoblastic leukemia, ГЛЛ, ALL) – онкологічне захворювання клітин крови, за тип і перебіг якого відповідає генетична інформація організму. Серед типів захворювання виділяють два основні: BCR/ABL та NEG Їхня відмінність полягає в тому, що перший має наявну генетичну абнормалію на 22 хромосомі, тоді як для другого жодної генетичної абнормалії не зафіксовано.

У роботі розглянуто

Об'єкт дослідження –.

Предмет дослідження –.

Мета роботи – .

**Огляд літератури**

**Завдання роботи**

1. .
2. .
3. .
4. .
5. .

Курсова робота складається з таких частин:

* вступу, який обґрунтовує актуальність роботи; визначає цілі проведення наукового дослідження;
* аналітичного огляду літературних джерел, який містить постановку задачі та …;
* програмної реалізації;
* висновків;
* списку використаної літератури;
* додатку А. Код програми мовою R.

# ТЕОРЕТИЧНА ЧАСТИНА

* 1. **Постановка задачі**

Задано набір даних із бібліотеки ALL для мови R, що містить інформацію про 128 пацієнтів із T-лейкемією та B-лейкемією та рівень експресії в них 12 625 генів. Назви цих лейкемій походять від типу лімфоцитів, які вражає цей вид захворювання: відповідно, T-лімфоцити та B-лімфоцити.

Необхідно порівняти середній рівень експресії релевантних генів між пацієнтами з захворюванням B-лейкемії типу BCR/ABL та NEG та визначити, відсутня (гіпотеза ) чи наявна (гіпотеза ) в них відмінність. Усього в наборі даних наявно 37 пацієнтів із типом лейкемії BCR/ABL та 42 – з NEG.

* 1. **Лейкемії та їх типи**

B-лейкемії типу BCR/ABL та NEG відрізняються біологічною природою цього типу раку. Для типу BCR/ABL його появу зумовлено генетичною абнормалією 22 хромосоми (так званою філадельфійською хромосомою), що спричинено реципрокною транслокацією (взаємним обміном) ділянок із хромосомою 9, що призводить до того, що ген 9 хромосоми ABL1 опиняється та вступає в контрастну дію з геном BCR хромосоми 22, що призводить до появи гібридного сигнального протеїну для тирозинкінази. Оскільки цей протеїн постійно ввімкнено, то клітина з цією мутацією, починає ділитися безконтрольно, що є однією з передумов ракового захворювання. Для типу NEG неможливо виявити жодну схожу генетичну абнормалію, тому аналіз рівнів експресії генів може допомогти у виявленні та класифікації хвороби за типом.

* 1. **Гени та рівень їх експресії**

Для відбору релевантних генів було застосовано процедуру, запропоновану Gentleman et al. [1], що вилучає гени, рівень експресії яких у кров’яних тільцях низький що у хворих на лейкемію типу BCR/ABL або NEG, що у здорових. Також було вилучено гени з низькою мінливістю, з використанням міжквартильного розмаху (IQR).

За Kellis et al. [2] існує декілька різних методів вимірювання експресії генів, серед них описано, наприклад: ДНК-біочипи та РНК-послідовності. Для отримання експресії генів за методом біочипів беруться короткі ділянки ДНК, що звуться пробами. Їх прикріпляють до твердої поверхні, відомої як ДНК-біочип. Тоді отримана з клітини РНК-популяція інтересу зворотно транскрибується у кДНК (компліментарну ДНК). Тоді біочип промивають кДНК, що запускає процес гібридизації, який призводить до флуоресцентного світіння проб. Детекція цього світіння дозволяє визначити відносну кількість мРНК у пробі. РНК-послідовність є більш сучасною технологією визначення експресії генів. Її функція подібна до ДНК-біочипів, але з більшою точністю. Відмінність полягає в тому, що для методу ДНК-біочипів необхідне використання специфічних проб, і їх створення потребує знання про геном та розмір утвореного масиву даних. Технологія РНК-послідовности не має цих обмежень, вона дозволяє секвенувати всю кДНК, отриману в експериментах із біочипами за допомогою технології секвенування нового покоління. Цю техніку широко використовують, наприклад, для вивчення раку. Отримані з обох методів дані аналізуються однаковим чином за допомогою кластерингу.

Результати аналізу експресії генів часто подають у вигляді матриць та їхніх теплових карт. Для їх отримання за допомогою вищезазначених методів проводиться вимірювання генів за різних умов: часу, стадій розвитку, фенотипів, здоров’я чи хвороба тощо. За допомогою цього отримують значення рівня експресії генів у чисельній формі. Якщо було проведено багато експериментів, то можна побудувати матрицю значень, що відображає значення , де – рівень експресії гена в тестовому зразку, – рівень експресії гена в еталонному зразку. Такі матриці можна кластерувати за ієрархією, відображаючи відносини між парами генів, парами пар тощо. Це утворює дендрограму, з колонками та рядками, упорядкованими за певним алгоритмом. Це дозволяє відкрити приховану структуру довгого сегмента геному та отримати розуміння про його функцію та, відповідно, краще розуміння причини певної хвороби. Такі теплові карти для розглянутого набору даних про ГЛЛ можна знайти, наприклад, у Chiaretti et al [3].

* 1. **Статистичні критерії порівняння середніх**

Стандартно для порівняння середніх застосовують критерій Готелінґа з фіксованим , меншим за та за , однак його сфера застосування не включає ситуацію, коли , тобто розмірність даних значно перевищує розмір вибірки. Bai and Sarandasa [4] показали також, що за його точність зменшується зі збільшенням , зокрема через те, що матриця вибіркової коваріації не збігається до популяційної коваріації коли та мають однаковий порядок. Yin, Bai та Krishnaiah [5] показали, що за максимум і мінімум власних значень вибіркової коваріації не збігаються до відповідних власних значень . За може не бути оборотною, тому статистику не визначено.

Bai and Sarandasa запропонували обчислення статистики , а Chen та Qin [6] розширили її до статистик та .

Використання цих статистик полягає в наявності двох незалежних однаково розподілених випадкових вибірок із , а саме

, (1.1)

де – розподіл із із середнім та коваріацією . Одна з цілей багатовимірного аналізу – перевірка гіпотези про рівність середніх двох багатовимірних популяцій:

*.*

Ця гіпотеза складається з маргінальних гіпотез щодо середніх кожного з виміру даних.

Тестування гіпотези про рівність середніх за умови за статистикою Bai and Sarandasa базується на:

, (1.2)

де та . Пропозиція Bai and Sarandasa полягає у вилученні з тесту , зважаючи на відсутність користі від нього за . Також для отримання віднімають . Для цього тесту припущено такі умови:

(1.3)

, (1.4)

де – найбільше власне значення .

Статистика має вигляд:

(1.5)

після вилучення для з . Можна показати, що

.

Тож є всім необхідним для тестування. За правильности гіпотези та умови

(1.6)

маємо:

,

де за зникає.

Нехай

, (1.7)

де кожне – матриця для деякого такого, що , – -вимірні незалежні однаково розподілені випадкові вектори, де – одинична матриця .

Якщо переписати , припускаємо, що та

(1.8)

для додатного цілого такого, що .

(1.9)

Якщо (1.6) не дійсна, то використовуємо

(1.10)

На ставиться умова:

(1.11)

за .

З (1.7), (1.8), (1.9), (1.11) та (1.6) або (1.10) дійсна така теорема:

*ТЕОРЕМА 1*

за . Асимптотичну нормальність витримано без певних прямих обмежень між та . Єдине обмеження на виміри дано в (1.11). Дисперсія цієї статистики має вигляд:

,

де за (1.6)

(1.12)

та за (1.10):

(1.13)

Для формулювання тестової процедури, що базується на цій теоремі необхідно оцінити із (1.12).

Для виключення з таких доданків, як Chen та Qin запропонували такі оцінки та :

та

,

де – -е вибіркове середнє після вилучення та , – -е вибіркове середнє після вилучення . Це подібне до ідеї кросвалідації, де під час утворення відхилень та від вибіркового середнього та виключаються з обчислення вибіркового середнього. Таким чином оцінки та можна подати як слід сум добутків незалежних матриць.

За припущеннями (1.6)-(1.9) та (1.11) для вище подані оцінки є спроможними за співвідношенням за такою теоремою:

*ТЕОРЕМА 2*

*.*

За оцінка , спроможна за співвідношенням:

*.* (1.14)

Це спільно з теоремою 1 дає тестову статистику:

за . Цей тест із рівнем статистичної значущости відкидає гіпотезу якщо , де – верхній квантиль розподілу .

За Manly [7] рандомізований критерій визначення дійсности нульової гіпотези полягає у виборі статистики , що дозволяє виміряти, наскільки в заданих даних наявний шуканий ефект. Значення статистики для наявних даних порівнюють із розподілом , отриманим із випадкових перестановок даних. У випадку, якщо нульова гіпотеза дійсна, то будь-які варіанти даних могли статися з однаковою ймовірністю, і спостережувані дані є лише одним із однаково правдоподібних порядків, значення має бути типовим значенням розподілу . Якщо такий результат не спостережено, то є значущою, тому більш правдоподібною є альтернативна гіпотеза. Рівень значущости визначають як пропорцію значень, що є так само чи більш екстремальними за його значення з рандомізованого розподілу. У такому разі якщо менше за 5%, то є певні підстави вважати нульову гіпотезу неправильною, якщо менше за 1%, то є досить сильні підстави вважати, що нульова гіпотеза неправильна, якщо менше за 0.1%, то є дуже сильні підстави відкидати нульову гіпотезу.

# Практична частина

* 1. **Аналіз даних**

Набір даних ALL для мови R містить інформацію про 128 пацієнтів із T-лейкемією та B-лейкемією та рівень експресії в них 12 625 генів. Після фільтрування за вищезазначеною процедурою з Gentleman et al. було знайдено 2391 ген, релевантний для розгляду в цій задачі.

Зі 128 пацієнтів у 79 наявна лейкемія типів BCR/ABL або NEG. Для обчислення статистик за процедурою, запропонованою Gentleman et al. з оригінального набору даних вибрано цих пацієнтів та розподілено на дві матриці, що містять, відповідно, лише пацієнтів із лейкемію типу BCR/ABL (позначено як X1) та лейкемією типу NEG (позначено як X2).

* 1. **Опис методу**

Робота програми починається з задання змінної рівня значущости sign\_level як 0.05. З використанням функції sum\_fot\_T з модуля sum\_for\_T\_stat проводиться обчислення за формулою для статистики (1.5) для вектору X1, для вектору X2 та для обох векторів, після чого три отримані значення підсумовуються для отримання статистики .

Наступним кроком проводиться обчислення оцінки сліду матриці за заданими формулами для статистики (1.14) для вектору X1, для вектору X2 та для обох векторів, після чого три отримані значення множаться на відповіді підсумовуються для отримання оцінки .

Тоді значення отриманої статистики ділиться на значення оцінки , що дає значення статистики . Для проведення тесту використовується значення , отримане за допомогою функції qnorm(1-sign\_level/2).

* 1. **Результат роботи програми**

У результаті обчислення було отримано значення – 0.4584315, тоді як значення – 1.959964, що не дозволяє відкидати гіпотезу про рівність середніх.

# ВИСНОВКИ

# СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Robert Gentleman et al. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer Science+Business Media, 2005. 473 p.
2. Manolis Kellis et al. Computational Biology - Genomes, Networks, and Evolution. URL: https://bio.libretexts.org/Bookshelves/Computational\_Biology/Book%3A\_Computational\_Biology\_-\_Genomes\_Networks\_and\_Evolution\_(Kellis\_et\_al.)
3. Sabina Chiaretti et al. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival
4. Z. Bai, H. Saranadasa. Effect of high dimension: By an example of a two sample problem. Statist. Sinica 6, 1996. 311–329 p.
5. Y. Yin, Z. Bai, P. R. Krishnaiah. On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. Probab. Theory Related Fields 78, 1988. 509–521 p.
6. Song Xi Chen, Ying-Li Qin. A two-sample test for high-dimensional data with applications to gene-set testing. The Annals of Statistics Vol. 38, No. 2, 2010, 808-835 p. URL: 10.1214/09-AOS716
7. Bryan F. J. Manly. Randomization Bootstrap and Monte Carlo Methods in Biology. Boca Raton: Chapman & Hall/CRC, 2007. 450 p.

# ДОДАТОК А. Код програми мовою R