Київський національний університет імені Тараса Шевченка
Кафедра теорії ймовірностей, статистики та актуарної математики

Р.Є. Майборода, О.В. Сугакова

ВСТУП ДО СТАТИСТИКИ

Завдання самостійних робіт
та рекомендації по виконанню

Для студентів освітньо-професійної програми "Статистика"
освітнього рівня "бакалавр"

Рекомендовано до друку вченою радою механко-математичного факультету Київського національного університету імені Тараса Шевченка, протокол № 10 від 06.03.2026.

Київ — 2026

УДК 519.22.35
ББК 22.172я73

Майборода Р.Є., Сугакова О.В. ВСТУП ДО СТАТИСТИКИ (Завдання самостійних робіт та рекомендації по виконанню). Навчально-методичний посібник.

Посібник містить 10 варіантів завдань для самостійного виконання, що охоплюють курс “Вступ до статистики” для студентів механіко-математичного факультету спеціальності “статистика”. Розглядаються базові методи лінійної регресії, візуального відображення даних, кореляційного аналізу, перевірки статистичних гіпотез, застосування статистики та імітаційного моделювання у актуарних розрахунках. Для кожного завдання наведено рекомендації по виконанню та приклади виконання аналогічних завдань. Основна увага приділена змістовній інтерпретації результатів статистичного аналізу.

Вступ

Цей посібник призначений для використання при вивченні навчальної дисципліни "Вступ до статистики", який читають студентам другого курсу механіко-математичного факультету що спеціалізуються у галузі теорії ймовірності та математичної статистики. Метою курсу є початкове знайомство з предметом і методами статистики. Виконуючи самостійну роботу, студенти можуть познайомитись з тим, як теоретичні знання, отримані ними на лекціях, використовуються у прикладній статистиці. В усіх завданнях самостійної роботи передбачається статистичний аналіз деяких даних: модельованих, реальних наданих викладачем, або таких, які студент сам збирає у цікавій йому області.

Для виконання завдань потрібно застосовувати систему статистичного програмування `R`. Останню версію `R` для операційної системи Windows можна отримати за адресою:

`http://cran.r-project.org/bin/windows/base/`

На цьому ж сайті є версії `R` і для інших операційних систем. Мінімальні необхідні знання з `R` можна знайти у підручнику [4]. Зокрема, у першому розділі цього підручника міститься інформація про встановлення `R` і початок роботи з ним. Для роботи з `R` зручно використовувати систему `RStudio`, яка дозволяє працювати одночасно з програмним кодом, результатами роботи, графікою, системою підказки, тощо. Рекомендації по встановленню `RStudio` також можна подивитись у [4].

Встановивши `R` на своєму комп'ютері, студенти можуть виконувати завдання цілком самостійно і оформлювати результати у вигляді звітів, які надсилаються викладачу у форматі pdf на адресу:

`rostmaiboroda@knu.ua`.

У заголовку звіту має бути вказано: назва дисципліни, прізвище та ім'я студента, курс, група, номер роботи, номер варіанту. У звітах потрібно вказувати постановку задачі, результати статистичної обробки комп'ютерною програмою та змістовну інтерпретацію отриманих результатів.

Приклади, наведені у посібнику, показують, як це можна робити.

Основні теоретичні відомості, потрібні для виконання завдань, можна знайти у навчальному посібнику [5]. Додаткові теоретичні відомості з математичної статистики можна дивитись у підручниках [1] і [6]. У книгах [7], [8] можна знайти більше інформації про практичне застосування статистичних алгоритмів.

Всі робочі матеріали по курсу, включаючи даний текст, навчальний посібник [5], файли з даними для виконання самостійних завдань та ін. можна знайти у папці на гугл-диску за адресою:
`https://drive.google.com/drive/folders/1iBpFTBCUWzrs4f0pBKoytqaiUYSSRoup?usp=sharing`

Теоретичні питання

1. Метод найменших квадратів для простої лінійної регресії. Оцінки коефіцієнтів за цим методом. ([5], п. 1.2)

2. Лінеаризація регресійних залежностей. ([5], п. 1.3)

3. Характеристики центрального положення у вибірці: вибіркове середнє та вибіркова медіана. Їх стійкість по відношенню до забруднень — робастність. ([5], п. 2.1.1)

4. Характеристики розкиду вибірки: дисперсія, стандартне відхилення, IQR. Їх робастність. ([5], п. 2.1.1)

5. Алгебраїчні властивості статистик центрального положення і розкиду: еквіваріантність та інваріантність відносно додавання та множення. ([5], п. 2.1.2)

6. Гістограми абсолютних та відносних частот. Їх використання для аналізу та порівняння розподілів даних. ([5], п. 2.2)

7. Скриньки з вусами. ([5], п. 2.3)

8. Коефіцієнт кореляції Пірсона. Його інтерпретація та властивості. ([5], п. 2.4)

9. Ранги спостережень. Рангові коефіцієнти кореляції Спірмена і Кендалла. Їх інтерпретація та властивості.([5], п. 2.5)

10. Значущість кореляцій. Тест для перевірки значущості кореляції Спірмена.([5], п. 2.7)

11. Візуалізація кореляційних зв'язків між змінними. Карти кореляцій та кореляційні мережі. ([5], п. 2.6)

12. Ймовірність події з погляду статистики. Умовні ймовірності. Поняття незалежності двох подій.([5], п. 3.1, 3.2)

13. Основні властивості ймовірностей (правило додавання і правило множення). ([5], п. 3.1, 3.2)

14. Поняття математичного сподівання дискретної випадкової величини. Основні властивості математичного сподівання. ([5], п. 3.6)

15. Поняття дисперсії випадкової величини. Основні властивості дисперсії випадкової величини. Поняття незалежності випадкових величин. ([5], п. 3.6, 3.7)

16. Статистичний тест для перевірки гіпотез. Помилки першого і другого роду. Стандартний рівень значущості тесту. Досягнутий рівень значущості тесту. ([5], п. 5.1)

17. Довірчі інтервали і їх застосування для побудови статистичних тестів. ([5], п. 5.3.1)

18. Односторонній тест для перевірки гіпотези про ймовірність події за частотою. ([5], п. 5.2)

19. Двосторонній тест для перевірки гіпотези про ймовірність події за частотою. ([5], п. 5.2)

20. Довірчий інтервал Вільсона для ймовірності події. ([5], п. 5.3.1)

21. Використання довірчих інтервалів для перевірки однорідності ймовірностей за різними вибірками. Одночасні довірчі інтервали.([5], п. 5.3.2)

22. Функція розподілу і функція виживання випадкової величини. Таблиця смертності у страхуванні життя. ([5], п. 3.3, 6.1, 6.2)

23. Емпірична функція розподілу як оцінка для теоретичної функції розподілу. Її незміщеність та консистентність. ([5], п. 6.1)

24. Розрахунок вартості полісу тимчасового страхування життя за принципом еквівалентності. ([5], п. 6.3)

25. Розрахунок резерву за полісами страхування, що забезпечує гарантований ризик перевищення виплат понад резерв. Плата за ризик. ([5], п. 6.4)

26. Застосування імітаційного моделювання для аналізу ризиків по страхових полісах. ([5], п. 6.5)

27. Центральна гранична теорема (без доведення). Приклади застосування. ([5], п. 3.8)

Робота 1. Проста лінійна регресія.

Завдання роботи.

Метою роботи є проведення попереднього аналізу залежності між цінами на два різних товари (послуги) у різних містах Європи. Для цього виконайте наступні завдання.

1. На сайті `numbeo.com` знайдіть поточні значення цін на два товари/послуги у десяти містах Європи (конкретні товари і міста виберіть відповідно до вашого варіанту за таблицею індивідуальних завдань, поданою нижче). Запишіть отримані дані у текстову таблицю і збережіть її у вигляді файлу на диску.

2. Прочитайте файл з вашими даними у `R` у вигляді фрейму даних.

3. Виведіть діаграму розсіювання для досліджуваних цін разом з назвами міст.

4. Підрахуйте оцінки методу найменших квадратів для коефіцієнтів простої лінійної регресії між досліджуваними цінами за теоретичними формулами, використовуючи арифметичні функції `R`.

5. Підрахуйте ті самі оцінки коефіцієнтів, використовуючи функцію `lm()`. Порівняйте з результатом п. 4.

6. На діаграмі розсіювання нарисуйте лінію регресії за методом найменших квадратів.

7. Опишіть особливості даних, помітні на діаграмі розсіювання. Чи є викиди, нелінійність, кластеризація (розбиття даних на групи)?

8. Якщо виявлені певні особливості, спробуйте проаналізувати дані з урахуванням цих особливостей. Наприклад, спробуйте видалити викид з набору даних і подивіться, наскільки зміниться лінія регресії. Або спробуйте лінеаризувати залежність, логарифмуючи одну чи обидві змінні (або іншим перетворенням). Опишіть результати ваших експериментів.

9. Зробіть висновки: як на вашу думку, чи виявилась якась залежність? Який вигляд вона має, якщо виявилась.

10. (Для бажаючих) Зберіть дані про іще 10 міст з тих країн, з яких була взята попередня вибірка. Додайте значення цін у цих містах до ваших даних і повторіть дослідження. Наскільки отримані результати відрізняються від тих, що були отримані раніше?

Індивідуальні варіанти завдань.

У варіантах 1-5 зберіть дані по таких містах Європи: Paris, Orlean, London, Edinburgh, Rome, Milan, Berlin, Munich, Madrid, Barcelona.

У варіантах 6-10 зберіть дані по таких містах Європи: Warsaw, Krakow, Prague, Brno, Budapest, Debrecen, Bucharest, Constanta, Sofia, Gabrovo.

Табл. 1: Ціни у французьких містах

City	Water	Fitness
Paris	0.92	38.28
Orleans	0.5	23
Marseille	1.56	36.4
Toulouse	0.68	36
Nantes	0.5	28.5
Lyon	0.83	36.85
Amiens	0.66	31.67
Nice	0.81	33.62
Strasbourg	0.6	35
Lille	0.68	37.86

Для аналізу використовуються ціни на такі товари/послуги:
`Meal` — Meal, Inexpensive Restaurant;
`Milk` — Milk (regular), (1 liter);
`Water` — Water (1.5 liter bottle);
`Gasoline` — Gasoline (1 liter);
`MobPhone` — Mobile Phone Monthly Plan with Calls and 10GB+ Data;
`Fitness` — Fitness Club, Monthly Fee for 1 Adult.
Пари товарів для різних варіантів:
Варіанти 1 і 6: `MobPhone`, `Fitness`
Варіанти 2 і 7: `Fitness`, `Milk`
Варіанти 3 і 8: `Meal`, `Milk`
Варіанти 4 і 9: `Water`, `Meal`
Варіанти 5 і 10: `Milk`, `Gasoline`.

Рекомендації по виконанню роботи

Ми розглянемо приклад даних по десяти містах Франції і двох змінних (цінах): `Milk` і `Fitness`. Зберемо дані на сайті `numbeo.com` і запишемо їх у текстову таблицю 1.

На моєму комп'ютері повне ім'я файлу з цією таблицею:
`c:/rem_d/online/vstup/data/cost_france.txt`

Прочитаємо дані в `R` і відобразимо діаграму розсіювання даних (рис. 1).

```
> cost <- read.table("c:/rem_d/online/vstup/data/cost_france.txt",
+                    header=TRUE)
```

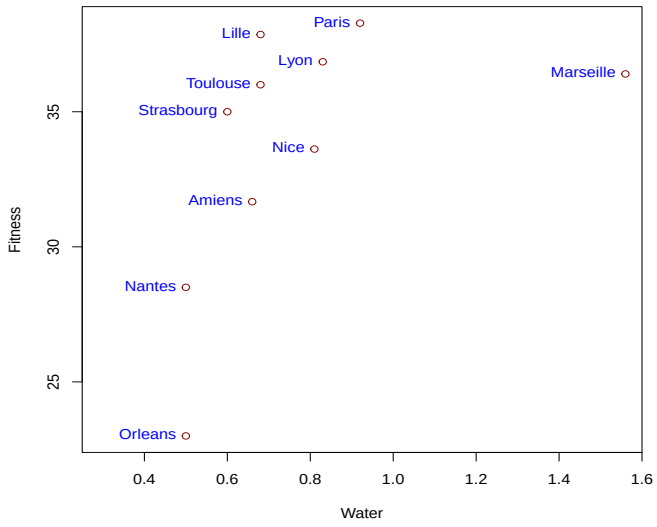

Рис. 1: Діаграма розсіювання цін

```
> plot(cost$Water,cost$Fitness,col="darkred",
+      xlim=c(0.3,1.55),xlab="Water",ylab="Fitness")
> text(cost$Water,cost$Fitness,labels=cost$City,col="blue",pos=2)
```

Підрахуємо оцінки МНК для коефіцієнтів лінійної регресії безпосередньо за формулами (1.4) і (1.8) підручника [5].

```
> n <- 10
> MW <- sum(cost$Water)/n
> MF <- sum(cost$Fitness)/n
> b1 <- sum((cost$Water-MW)*(cost$Fitness-MF))/sum((cost$Water-MW)^2)
> b0 <- MF - b1*MW
> b1

[1] 7.861766

> b0

[1] 27.63299
```

І порівнюємо з результатом функції `lm()`:

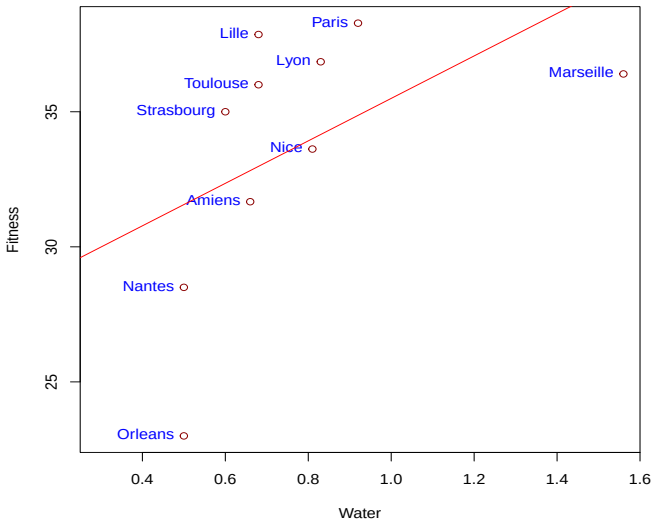

Рис. 2: Діаграма розсіювання з лінією регресії

```
> lm(Fitness~Water,data=cost)


Call:
lm(formula = Fitness ~ Water, data = cost)

Coefficients:
(Intercept)        Water
     27.633        7.862
```

Результати вийшли однакові, отже можна сподіватись, що помилки немає.

Відобразимо на діаграмі розсіювання лінію регресії (рис. 2):

```
> plot(cost$Water,cost$Fitness,col="darkred",
+      xlim=c(0.3,1.55),xlab="Water",ylab="Fitness")
> text(cost$Water,cost$Fitness,labels=cost$City,col="blue",pos=2)
> abline(b0,b1,col="red")
```

Лінія регресії на цьому рисунку відображена червоним кольором. Ми бачимо, що вона проходить нижче ніж можна було б прикинути на око.

Можливо, це тому, що вона притягається до даних по Марселю, у якому аномально високою є вартість пляшки води. Спробуємо розглянути Марсель (третій рядок нашої таблиці — фрейму даних) як викид і вилучити його з розгляду. При підгонці за домопогою МНК отримуємо

```
> B9 <- lm(Fitness~Water,data=cost[-3,])
> B9


Call:
lm(formula = Fitness ~ Water, data = cost[-3, ])

Coefficients:
(Intercept)        Water
      15.55        26.02

> plot(cost$Water,cost$Fitness,col="darkred",
+      xlim=c(0.3,1.55),xlab="Water",ylab="Fitness")
> text(cost$Water,cost$Fitness,labels=cost$City,col="blue",pos=2)
> abline(b0,b1,col="red")
> abline(B9$coefficients,col="darkgreen")
```

Бачимо, що коефіцієнти суттєво змінились. Порівнюючи лінії регресії на рис. 3, бачимо, що підгонка залежності між цінами у дев'яти містах за даними без Марселя вийшла значно кращою, ніж з урахуванням цін Марселя.

(Далі можна було б додати дані іще по 10 французьким містам і подивитись, яка лінія краще описує залежність на отриманій діаграмі розсіювання).

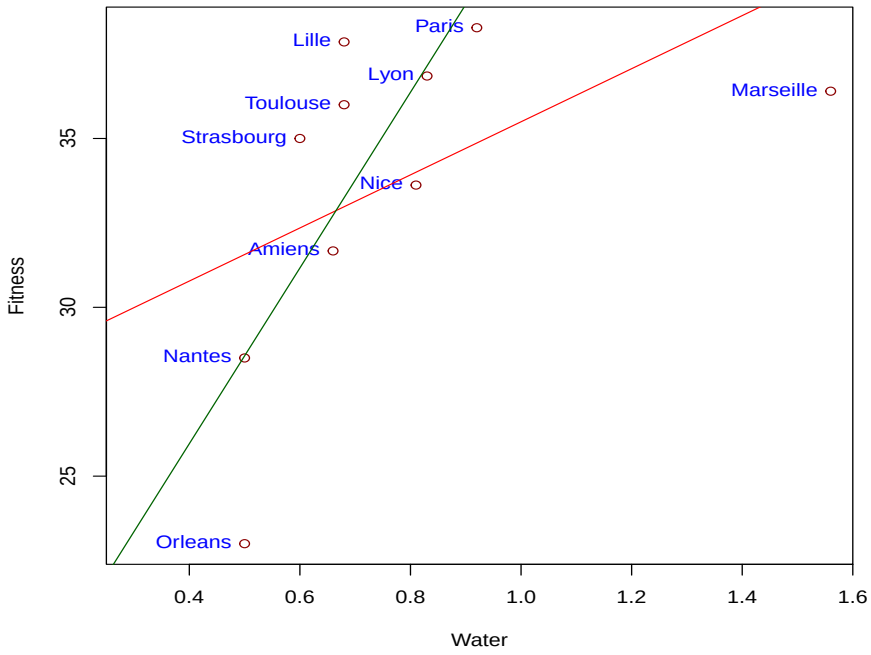

Рис. 3: Червона лінія — регресія по всіх спостереженнях, зелена — без Марселя

Робота 2. Візуалізація розподілу даних.

Метою роботи є дослідження розподілу одновимірних даних, порівняння розподілів.

Завдання роботи.

1. Завантажте дані про ціни на акції двох фірм, які ви вибрали відповідно до вашого варіанту.

2. Для кожної фірми на окремих рисунках відобразіть гістограми абсолютних частот цін закриття на Нью-йоркськй біржі на акції цих фірм по всіх наявних даних.

3. Відобразіть також гістограми rel — відношення максимальної та мінімальної цін для кожного дня торгів. Опишіть особливості отриманих розподілів, які можна помітити на цих гістограмах.

4. Для кожної з обраних фірм відобразіть скриньки з вусами для розподілу цін закриття по кожному року, по якому є інформація у даних (за всі роки для даної фірми на одному рисунку). Опишіть зміни, які відбувались з розподілом даних на цих рисунках.

5. Виділіть 2-3 часових інтервали, на яких розподіл цін закриття був однорідним. По цих інтервалах побудуйте гістограми відносних частот на одному рисунку. Опишіть відмінності розподілів, які ви помітили.

6. Для часових інтервалів, які ви вибрали у п. 5, нарисуйте гістограми відносних частот змінної rel, визначеної у п. 3 на одному рисунку. Опишіть результати.

Індивідуальні варіанти завдань.

Дані про ціни на акції знаходяться у файлі `daily_sp500.zip`. Цей архів потрібно розпакувати у яку-небуть папку на вашому диску. Кожній фірмі відповідає один csv-файл, назва якого має вигляд

`table_`⟨мнемоніка⟩`.csv`,

де ⟨мнемоніка⟩ — скорочена назва фірми. Інформація про мнемоніки фірм та сферу їхньої діяльності — у файлі

`List_of_S&P_500_companies.pdf`.

Виберіть дві компанії, які вас цікавлять серед тих, назви яких починаються з літери, що відповідає вашому варіанту.

Варіант 1. Літери A або L.

Варіант 2. Літери B або M.

Варіант 3. Літери C або N.

Варіант 4. Літери D,O або X.

Варіант 5. Літери E або K.

Варіант 6. Літери F,Q або Y.

Варіант 7. Літери G або R.
Варіант 8. Літери H або S.
Варіант 9. Літери I або T.
Варіант 10. Літери E,U або V.

Рекомендації по виконанню

Розпакуйте архів `daily_sp500.zip`, виберіть компанію за вашим варіантом. Завантажте дані в `R`. Наприклад, тут папка з диску, де знаходяться розпаковані файли встановлена як робоча, а потім завантажується таблиця даних для компанії nflx:

```
> setwd("C:\\rem_d\\online\\vstup\\data\\stoks\\daily_intro")
> nflx <- read.csv(file="table_nflx.csv",header=F)
```

Присвоюємо імена змінним у нашому фреймі даних про ціни акцій компанії:

```
> colnames(nflx) <- c("dat","z","opn","mx","mn","clo","vol")
```

Тут використані такі назви змінних:
dat — дата торгів;
z — (не використовується);
opn — ціна відкриття;
mx — максимальна ціна;
mn — мінімальна ціна;
clo — ціна закриття;
vol — обсяг продаж.
Робимо перетворення змінної dat у формат дати:

```
> nflx$dat <- as.Date(as.character(nflx$dat),format="%Y%m%d")
```

Рисуємо гістограму абсолютних частот для ціни закриття (рис. 4):

```
> hist(nflx$clo,main="Netflix 2002-2013",xlab="closure price")
```

На цій гістограмі помітно, що дані являють собою суміш, як мінімум, двох наборів з різними розподілами. Викидів не помітно.

Нарисуємо гістограму відношень максимальних цін до мінімальних (рис. 5):

```
> rel <- nflx$mx/nflx$mn
> hist(rel, breaks=30,main="Netflix 2002-2013",xlab="max to min")
```

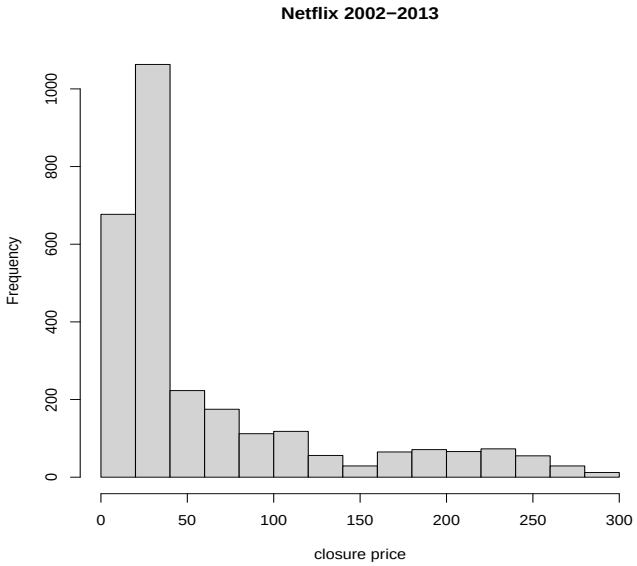

Рис. 4: Гістограма цін закриття для Netflix за всіма даними

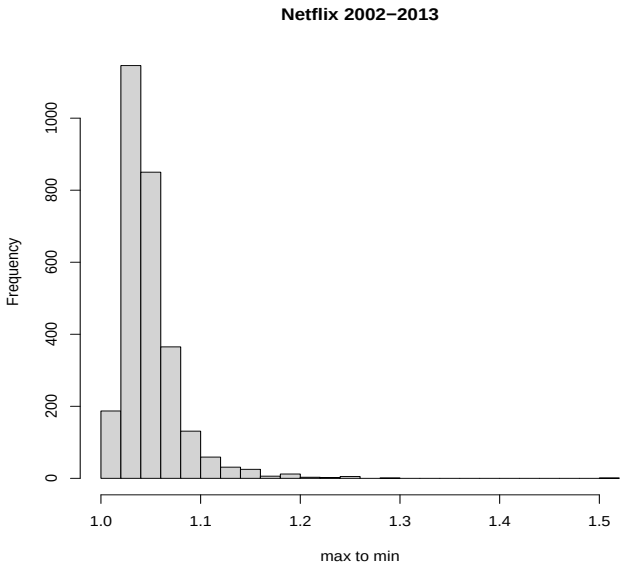

Рис. 5: Гістограма відношень максимальної ціни до мінімальої для Netflix за всіма даними

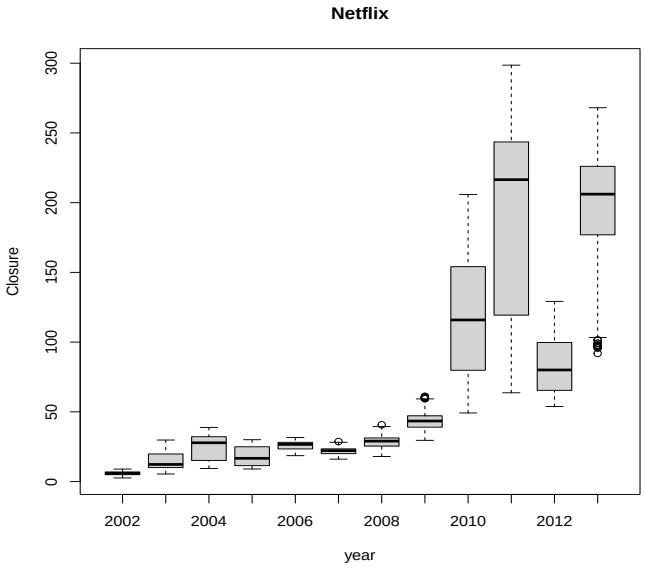

Рис. 6: Скриньки з вусами цін закриття по роках для Netflix за всіма даними

Бачимо, що розподіл цього відношення асиметричний, має важкий правий хвіст, наявні викиди.

Нарисуємо скриньки з вусами для цін закриття по роках

```
> library(lubridate)
> boxplot(nflx$clo ~year(nflx$dat),xlab="year",ylab="Closure",main="Netflix")
```

Помітна різка зміна розподілу цін закриття при переході від 2009 до 2010 року. Розіб'ємо дані на дві групи — перша група 2002-2009 роки (синій колір), друга — 2010-2013 (червоний колір). Нарисуємо гістограми відносних частот для обох груп на одному рисунку (рис. 7)

```
> hist(nflx$clo[year(nflx$dat) <= 2009],col=rgb(0,0,1,1/3),
+      breaks=seq(0,310,10),main="",xlab="",probability=TRUE)
> hist(nflx$clo[year(nflx$dat) > 2009],col=rgb(1,0,0,1/3),
+      breaks=seq(0,310,30),main="",xlab="",add=TRUE,probability=TRUE)
```

(Цю картинку також треба описати).

Аналогічно виводяться гістограми для відношення максимальної та мінімальної ціни. Не забудьте зробити загальні висновки.

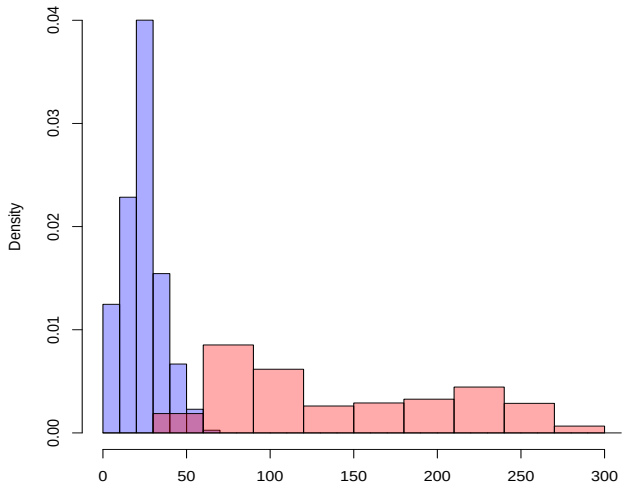

Рис. 7: Гістограми цін закриття для Netflix: синім — 2002-2009 роки, червоним 2010-2013 роки

Робота 3. Кореляційний аналіз

Мета роботи — дослідити залежності між цінами на акції різних компаній на Нью-йоркській біржі за даними 2002-2013 років.

1. З фірм, що відповідають вашому варіанту у завданні 2, виберіть десять таких, які вам найбільше цікаві. Зберіть у один фрейм даних дані по цінах закриття на акції цих фірм у такий спосіб, щоб кожен рядочок фрейму відповідав одному дню торгів на біржі, а кожен стовпчик — одній фірмі.

2. Підрахуйте коефіцієнти кореляції Пірсона між цінами для всіх фірм. Виведіть отримані значення у числовому вигляді та у вигляді карти кореляцій.

3. Спробуйте виділити за кореляціями групи (кластери) близьких за поведінкою цін фірм.

4. Нарисуйте матричні діаграми розсіювання цін для наборів 3-4х фірм: (а) для фірм з одного кластера; (б) коли кожна фірма належить іншому кластеру. (Якщо кластери не виділяються, виберіть 3-4 фірми, зв'язки між якими мають бути найбільш цікавими на вашу думку). Зробіть висновок про те, наскільки доцільним є використання саме кореляції

Пірсона для опису залежностей, що спостерігаються на цих діаграмах.

5. Нарисуйте кореляційні мережі для набору всіх фірм, які ви досліджуєте. Спробуйте вибрати варіант мережі, який найбільш адекватно передає залежність між цінами на акції різних фірм.

6. Виконайте п. 2, 3 і 5 використовуючи замість кореляцій Пірсона кореляції Спірмена і Кенделла. Зробіть висновок про те, які кореляції на вашу думку найбільш адекватні для опису залежностей між цінами акцій.

7. (Для бажаючих) Зберіть самі, або пошукайте в інтернеті статистичні дані про дві (або більше) змінні, які могли б становити для вас інтерес. Підрахуйте кореляцію між цими змінними, перевірте, чи є вона значущою. Результати дослідження опишіть у звіті.

Рекомендації по виконанню завдання.

Для виконання завдання потрібно об'єднати дані, описані у завданні 1 в один фрейм даних. Це робиться наступним чином. Складіть файли даних для фірм, які ви будете досліджувати, у одну папку на диску. Виконайте наступний скрипт, вказавши цю папку як робочу у команді `setwd` у першому рядочку скрипту.

```
> # Встановлюємо робочий каталог:
> setwd("C:\\rem_d\\online\\vstup\\data\\stoks\\daily_intro")
> # у filenames --- повні імена всіх файлів, що лежать у каталозі
> filenames=list.files(path="C:\\rem_d\\online\\vstup\\data\\stoks\\daily_intro",
+                      full.names=FALSE)
> # читаємо файли, виймаємо 1-ший і 6-й стовпчики і кладемо в окремий фрейм
> datalist = lapply(filenames,
+ function(x){x0<-read.csv(file=x,header=F)[,c(1,6)];
+ colnames(x0)<-c("data",
+ unlist(strsplit(x,"[_.]"))[2]);# назва компанії стає назвою стовчика
+ x0})
> # зливаємо фрейми в один:
> y<-Reduce(function(x,y) {merge(x,y,by="data")}, datalist)
```

— тут з таблиць даних, що вміщені у каталозі (папці)
`C:\\rem_d\\online\\vstup\\data\\stoks\\daily_intro`
збираються значення змінної `clo` (6-тий стовпчик кожної таблиці) і записуються у один фрейм даних `y`. У цьому фреймі перша змінна (стовпчик) відповідає даті біржової сесії, а всі наступні змінні мають назви

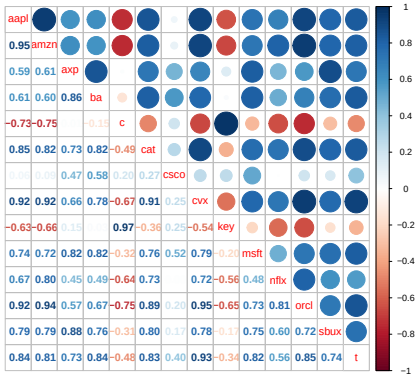

Рис. 8: Карта кореляцій

що є скороченими назвами фірм з вашого набору даних. У цих змінних знаходяться ціни закриття на акції відповідних фірм у даній сесії.

Для підрахунку коефіцієнтів кореляції можна використати функцію `cor()`:

```
> M <- cor(y[,-1])
```

(ми вилучили перший стовпчик з даних, оскільки кореляції з датою біржової сесії рахувати непотрібно). Тепер `M` містить матрицю кореляцій. Відобразити її у вигляді карти кореляцій і/або у числовому вигляді можна, використовуючи функцію `corrplot()` та інші функції з бібліотеки `corrplot`:

```
> library(corrplot)
> corrplot.mixed(M)
```

(див. рис. 8).

На карті помітно, що ціни на акції компаній `key` і `c` від'ємно корельовані з усіма іншими цінами крім цін на компанію `csco`. Ці дві компанії виділяються у окремий кластер. Розбити інші компанії на кластери не так просто. Можна застосувати опцію `order="hclust"` у функції `corrplot()`

```
>  corrplot(M,order="hclust",addrect=5)
```

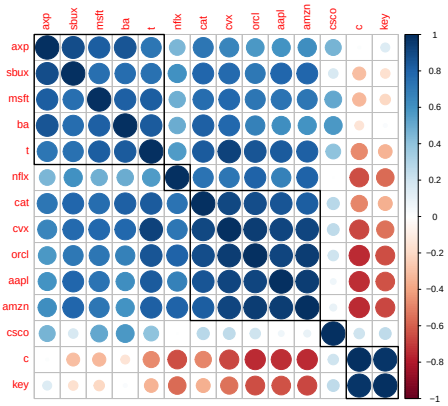

Рис. 9: Карта кореляцій з кластерами

(результат — на рис. 9). Змінні переставлені у порядку, за яким зручно відслідковувати їх залежність, квадрати виділяють змінні, які програма віднесла до одного кластера.

Для відображення матричної діаграми розсіювання виберемо змінні `aapl`, `amzn`, `csco` і `key`:

```
> pairs(y[,c("aapl","amzn","csco","key")])
```

(див. рис. 10). Залежність між змінними `aapl` і `amzn` загалом можна інтерпретувати як приблизно лінійну (з відхиленнями). Інші залежності на цій діаграмі мало схожі на лінійні, тому використання коефіцієнта кореляції Пірсона може не бути доцільним для їх опису.

Кореляційну мережу можна будувати, використовуючи функцію `qgraph()` з бібліотеки `qgraph`:

```
> library("qgraph")
> qgraph(M,layout="spring",threshold=0.5,labels=colnames(M))
```

(див. рис. 11).

Для підрахунку кореляцій Кендалла і Спірмена потрібно встановити відповідні значення опції `method` у функції `cor`, див. п. 5.4 у [4].

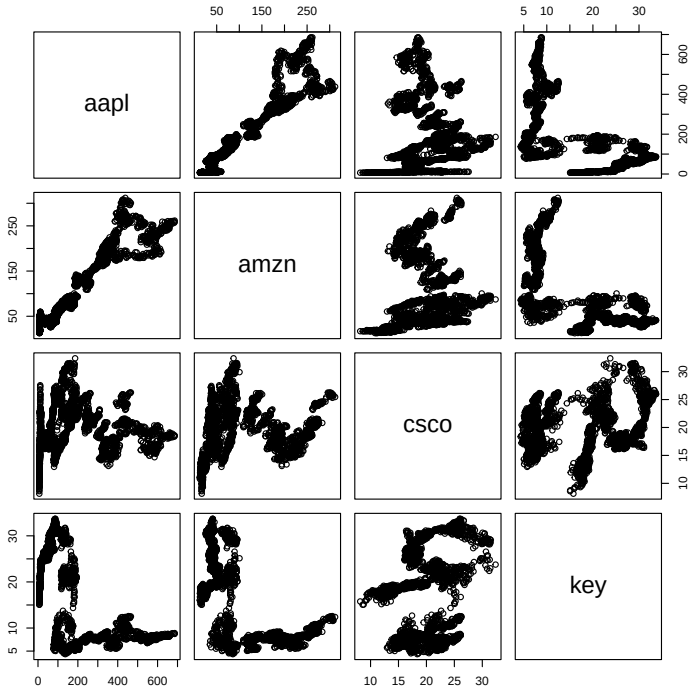

Рис. 10: Карта кореляцій з кластерами

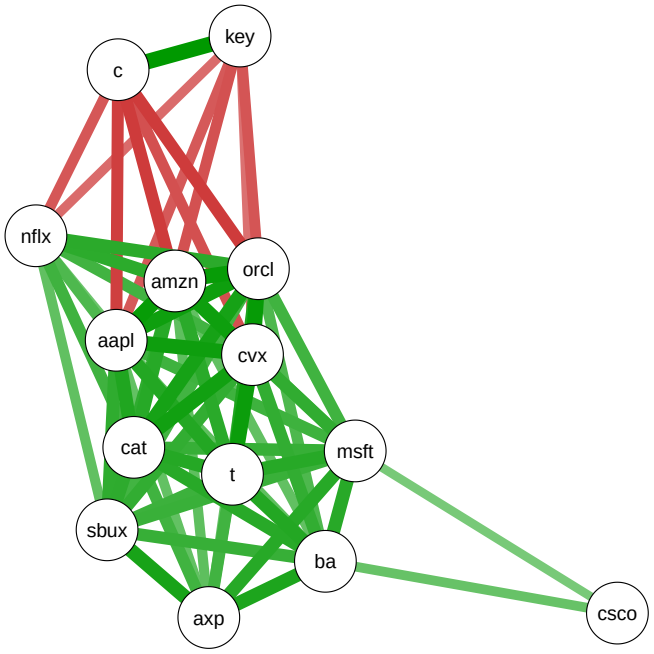

Рис. 11: Кореляційна мережа

Робота 4. Довірчі інтервали і перевірка гіпотез

Мета роботи — дослідити дані про гендерний склад дітей у різних штатах США, перевірити, чи є значущі відмінності співвідношення статей у різних штатах. Дані для аналізу містяться у файлі

`ChGendUS2023.csv`

Кожен рядок у цьому файлі відповідає одному штату США (+ всі США в цілому, Пуерто Ріко та столичний округ Колумбія). Стовпчик Location містить назву штату, Male — кількість дітей до 18 років чоловічої статі у цьому штаті на 2023 рік, Female — відповідно, кількість дітей жінокої статі.

1. Завантажте файл у `R` у вигляді фрейму даних. Виділіть набір даних, який відповідає штатам, вказаним у індивідуальному завданні.

2. Підрахуйте частки дітей чоловічої статі серед всіх дітей у вибраних штатах, подайте результат у вигляді таблиці.

3. Відобразіть на рисунку систему одночасних довірчих інтервалів для математичних сподівань часток дітей чоловічої статі у вибраних штатах з рівнем значущості $\alpha = 0.05$.

4. За отриманими довірчими інтервалами перевірте основну гіпотезу про те, що в усіх обраних штатах математичне сподівання цієї частки є однаковим. Альтернативна гіпотеза — є математичні сподівання, що відрізняються від інших.

5. Опишіть результати, вкажіть, для яких саме штатів є значуща відмінність математичних сподівань.

6. Для одного зі штатів підрахуйте межі довірчого інтервалу Вільсона за формулами (5.1) з [5]. Перевірте, чи збігаються вони з межами, які підрахувала функція `BinomCI` в `R`.

Індивідуальні варіанти завдань

Перелік штатів, які потрібно аналізувати у відповідному варіанті:

Варіант 1. Alabama,Hawaii, Massachusetts,New Mexico, Puerto Rico

Варіант 2. Alaska, Idaho, Michigan, New York, Vermont

Варіант 3. Arizona, Illinois,North Carolina, South Carolina, Wyoming

Варіант 4. Arkansas, Indiana, Ohio, South Dakota,Wisconsin

Варіант 5. California, Iowa, Oklahoma,Tennessee, West Virginia

Варіант 6. Colorado, Kansas, Oregon, Texas, Washington

Варіант 7. Connecticut, Alaska, Louisiana, Maine,Mississippi

Варіант 8. Delaware, District of Columbia, Alaska, Wyoming, Vermont

Варіант 9. Maryland, Arizona, Illinois, Pennsylvania, Puerto Rico

Варіант 10. Rhode Island, District of Columbia, Idaho, Hawaii, Georgia

Рекомендації по виконанню завдання.
Завантажуємо допоміжні бібліотеки:

```
>  library(plotrix)
> library(DescTools)
```

Читаємо дані з файлу:

```
>  x <- read.csv("C:/rem_d/online/vstup/data/ChGendUS2023.csv",header=TRUE)
>  x$Male <- as.numeric(x$Male)
> x$Female <- as.numeric(x$Female)
```

Виділяємо рядочки, що відповідають обраним штатам:

```
> rownames(x) <- x$Location
> Stat <- c("California","Colorado","Minnesota","Delaware","Vermont")
> y <- x[Stat,]
```

Підраховуємо і виводимо частку дітей чоловічої статі:

```
>  data.frame(Stat=y$Location,fraction = y$Male/(y$Male+y$Female))
```

```
        Stat  fraction
1 California 0.5118178
2   Colorado 0.5119183
3  Minnesota 0.5110828
4   Delaware 0.5084789
5    Vermont 0.5158240
```

Як бачимо з цієї таблички, частоти коливаються дуже мало навколо значення 0.51. Чи є ці коливання значущими? Перевіримо гіпотезу про рівність їх математичних сподівань, використовуючи довірчі інтервали:

```
> alpha <- 0.05
> ci <- BinomCI(y$Male,y$Male+y$Female,conf.level=(1-alpha)^(1/5),
+               sides="two.sided",method="wilson")
> plotCI(1:5,y=ci[,1],ui=ci[,3],li=ci[,2],lab=c(3,5,4),xlab="",ylab="",xaxt="n")
> axis(1,at=(1:5),labels=Stat)
```

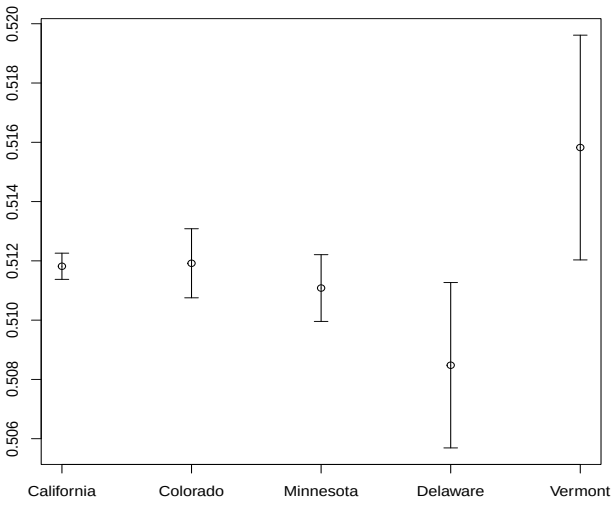

Рис. 12: Довірчі інтервали для частот

Ми бачимо, що довірчі інтервали для штатів Delaware і Vermont не перекриваються, отже для них математичні сподівання досліджуваної частоти є різними. Основну гіпотезу про однаковість математичних сподівань у всіх штатах потрібно відхилити.

Висновки: за побудованими довірчими інтервалами можна припустити, що коливання частот у штатах California, Colorado, Minesota є не значущими, їхні математичні сподівання можуть бути однаковими. У штатах Delaware і Vermont математичні сподівання різні. Немає можливості за даними встановити, чи дорівнює математичне сподівання для штату Delaware спільному математичному сподіванню штатів California, Colorado, Minesota, чи відрізняється від нього. Те саме справедливо і для штату Vermont.

Пункт 6 завдання спробуйте виконати самостійно.

Робота 5. Розрахунок вартості страхового полісу

У цій роботі метою є визначення вартості страхового полісу з використанням принципу еквівалентності та додаткової плати за ризик. При цьому використовуються як теоретичні формули, так і імітаційне моделювання.

Завдання роботи: Для полісу тимчасового страхування для особи віку x на n років розрахуйте нетто-вартість полісу зі страховою сумою S при відсотковій ставці i. Визначте плату за ризик якщо розглядається N незалежних полісів а допустима ймовірність перевищення виплат компанії понад резерв становить $\beta = 0.01$. Значення параметрів x, n, S, i, N виберіть з таблиці відповідно до номера вашого варіанту. Параметри смертності візьміть з таблиці 6.1 [5].

Розрахунок потрібних характеристик проведіть за теоретичними формулами. Після цього перевірте результат, використовуючи імітаційне моделювання. Наведіть у звіті теоретичні формули та скрипти, використані вами для всіх розрахунків, результати теоретичних розрахунків та імітаційного експерименту. Зробіть висновки.

Таблиця індивідуальних варіантів

Варіант	x	n	S	i	N
1	35	7	10000	0.05	1000
2	40	6	15000	0.04	2000
3	45	7	20000	0.03	1000
4	32	6	25000	0.05	2000
5	47	8	18000	0.04	1000
6	51	5	24000	0.03	2000
7	30	8	14000	0.05	1000
8	37	5	30000	0.04	2000
9	35	7	10000	0.05	1000
10	35	7	10000	0.05	1000

Рекомендації по виконанню роботи

У даному прикладі використовуються параметри з прикладів 5.3.1 та 5.4.1 [5].

Читаю таблицю смертності з файлу на своєму комп'ютері:

```
> tbl <- read.table("c:\\rem_d\\online\\vstup\\data\\Canada0.txt",header=TRUE)
```

(У своїй роботі ви можете використати лише ту частину таблиці смертності з табл. 6.1 [5], яка вам потрібна для розрахунків)

```
> lx <- tbl$lx
> dx <- tbl$dx
```

Тут `lx` — l_x, `dx` — d_k.

Задаємо індивідуальні значення параметрів:

```
> n <- 5
> S <- 1000000
> x <- 30
> irate <- 0.03
> nu <- 1/(1+irate)
> N <- 1000
> beta <- 0.001
```

Тут `nu` — ν — дисконтний множник.

```
> V <- (S/lx[x+1])*sum(nu^(1:n)*dx[(x+1):(x+n)])
> V

[1] 17787.58
```

Нетто-вартість полісу за правилом еквівалентності — 17 787.58.

Дисперсія виплат за полісом і стандартне відхилення σ:

```
> V2 <- (S^2/lx[x+1])*sum(nu^(2*(1:n))*dx[(x+1):(x+n)])-V^2
> sigma <-sqrt(V2)
```

Значення λ_β:

```
> lambda <- qnorm(1-beta)
```

Резерв з урахуванням ризику для N однотипних полісів:

```
> Rezerve <- V*N +lambda*sigma*sqrt(N)
```

Вартість одного полісу з надбавкою за ризик:

```
> Vrisk <- Rezerve/N
> Vrisk

[1] 30150.01
```

З урахуванням ризику випадкових коливань, вартість полісу зростає до 30 150.01.

На скільки відсотків зросла вартість?

```
> Vrisk/V-1
```

```
[1] 0.6950034
```

— на 69.5%.

Робимо імітаційне моделювання з кількістю повторних випробувань B= 1 000 000.

```
> set.seed(123456)
> B <- 1000000
```

Спочатку розраховуємо ймовірності смерті за кожен рік дії полісу на основі таблиці смертності:

```
> prob <- dx[(x+1):(x+n)]/lx[x+1]
> prob <- c(prob,1-sum(prob))
```

Генеруємо дані і підраховуємо випадкові вартості виплат по всіх полісах з урахуванням дисконту:

```
> discont <- c(nu^(1:n),0)
> ca <- replicate(B,sum(sample(discont*S,size=N,replace=TRUE,prob=prob)))
```

Середнє значення виплат ділене на кількість полісів — оцінка нетто-вартості одного полісу:

```
> Vsim <- mean(ca)/N
> Vsim
```

```
[1] 17783.72
```

— майже те саме, що за теоретичною формулою.

```
> RezerveSim <- quantile(ca,1-beta)
> Vsim <- RezerveSim/N
> Vsim
```

```
   99.9%
31249.47
```

Оцінка вартості з урахуванням ризику — 31249.47. Це досить близько до теоретичної вартості, отриманої вище.

Бібліоґрафія

[1] Карташов М.В. "Імовірність, процеси, статистика".— Київ, Видавничо-поліграфічний центр “Київський університет”, 2007, 494 с.

[2] Майборода Р.Є. Регресія: Лінійні моделі.— К. ВПЦ "Київський університет 2007, 296с.

[3] Майборода Р.Є., Сугакова О.В. "Аналіз даних за допомогою пакета R". , 2015 65 с.

[4] Майборода Р. Комп’ютерна статистика — професійний старт.— Київ, 2024.

[5] Майборода Р. Вступ до статистики.— Київ, 2024.

[6] Турчин В.М. Теорія ймовірностей і математична статистика.— Дніпропетровськ, ІМА-пресс, 2014 - 566 с.

[7] James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R.— Springer NY 2013.— 440p.

[8] Ramsey F., Schafer D.W. The Statistical Sleuth A Course in Methods of Data Analysis.—Brooks Cole. 2013.