МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

Р. Є. Майборода

КОМП'ЮТЕРНА СТАТИСТИКА

Підручник

КИЇВСЬКИЙ ВІТЛ УНІВЕРСИТЕТ 2019

УДК 519.22/.25:004.42(075.8)
М14

Рецензенти:

д-р фіз.-мат. наук, проф. О. І. Клесов
(Національний технічний університет України,
канд. фіз.-мат. наук, доц. О. О. Диховичний
"Київський політехнічний інститут імені Ігоря Сікорського"),
д-р фіз.-мат. наук, провід. наук. співроб. А. Ю. Пилипенко
(Інститут математики НАН України),
канд. фіз.-мат. наук, асист. В. В. Голомозий
(Київський національний університет імені Тараса Шевченка)

Рекомендовано до друку вченою радою
механіко-математичного факультету
(протокол № 8 від 16 квітня 2018 року)

Ухвалено науково-методичною радою
Київського національного університету імені Тараса Шевченка
(протокол № 5-17/18 н. р. від 21 червня 2018 року)

Майборода Р. Є.

М14 Комп'ютерна статистика : підручник / Р. Є. Майборода. – К. : ВПЦ "Київський університет", 2019. – 589 с.

ISBN 978-966-933-010-9

Розглянуто основні питання, що стосуються використання комп'ютерних методів статистичного аналізу для дослідження прикладних даних. Описано основні властивості мови статистичного програмування R, розглянуто методи дескриптивного та візуального аналізу статистичних даних. Досліджено методи статистичного оцінювання та перевірки гіпотез. Розглянуто основні засоби лінійного та нелінійного регресійного аналізу. Основну увагу приділено змістовній інтерпретації результатів статистичного аналізу.

Для студентів механіко-математичного факультету спеціальностей "Статистика" і "Математика".

УДК 519.22/.25:004.42(075.8)

ISBN 978-966-933-010-9

© Майборода Р. Є., 2019
© Київський національний університет імені Тараса Шевченка,
ВПЦ "Київський університет", 2019

Зміст

Список позначень та скорочень 8

Передмова **10**

1 Початок роботи з R **14**
1.1 Що таке R і де його взяти 14
1.2 Система R-Studio 18
1.3 Пакети, Help та інше 20

2 Мова статистичного програмування R **25**
2.1 Базові поняття 25
2.2 Типи даних та елементарні функції 30
2.2.1 Вектори. Арифметичні та логічні операції . 30
2.2.2 Індексація векторів 35
2.2.3 Фактори 38
2.2.4 Матриці, масиви та фрейми даних 41
2.2.5 Векторні і матричні функції. Функція apply. Пропущені значення 52
2.3 Деякі корисні функції 57
2.3.1 Функції покрокового обчислення 58
2.3.2 Впорядкування та пошук у масивах 61
2.3.3 Злиття наборів даних — merge 63
2.4 Експорт та імпорт даних у R 65
2.4.1 Експорт та імпорт даних у внутрішньому форматі 65

2.4.2 Експорт та імпорт текстових таблиць 66
2.5 Переформатування фреймів даних 70
2.6 Підстановки (pipes) 73
2.7 Програмування у R 76
2.7.1 Створення власних функцій 76
2.7.2 Векторизація функцій 82
2.7.3 Структури управління виконанням програм у мові R 84
2.7.4 Вибір з кількох умов: switch 86
2.7.5 Цикли while та repeat 87
2.7.6 Цикл for 89
2.8 Формули: y ~ x 91

3 Базова графіка в R **94**
3.1 Стовпцеві та кругові діаграми 95
3.2 Точки та лінії на площині 100
3.3 Елементи тривимірної графіки 110
3.4 Географічні карти 115

4 Одновимірна описова статистика **128**
4.1 Статистики середнього положення 130
4.2 Статистики розкиду 141
4.3 Алгебраїчні властивості описових статистик 146
4.4 Статистики форми розподілу 150
4.5 Групування та навантаження 152
4.6 Обчислення описових статистик у R 158

5 Опис залежностей **165**
5.1 Діаграми розсіювання 165
5.2 Коефіцієнт кореляції Пірсона 172
5.3 Візуалізація кореляцій 179
5.4 Ранги та рангові кореляції 188
5.5 Сила і значущість кореляції 197

6 Основні ймовірнісні розподіли 206
6.1 Загальні поняття 207
6.2 Неперервні розподіли на прямій 210
6.2.1 Одновимірний гауссів (нормальний) розподіл 210
6.2.2 Півнормальний розподіл 214
6.2.3 Логнормальний розподіл 215
6.2.4 Експоненційний і гамма розподіли та розподіл Лапласа 217
6.2.5 Розподіли екстремальних типів: Вейбулла, Фреше і Гумбеля 221
6.2.6 Рівномірний розподіл 222
6.2.7 Розподіли, пов'язані з гауссовим: χ^2, T і F . 225
6.3 Дискретні розподіли на прямій 227
6.3.1 Біноміальний розподіл 227
6.3.2 Розподіл Пуассона 229
6.3.3 Геометричний розподіл 230
6.4 Комбінації та перетворення розподілів 231
6.4.1 Зрізані розподіли 231
6.4.2 Суми незалежних випадкових величин . . . 235
6.4.3 Суміші кількох розподілів 240
6.5 Генерація псевдовипадкових чисел 242
6.5.1 Генератори рівномірних псевдовипадкових чисел 244
6.5.2 Генерація псевдовипадкових чисел із заданим розподілом 253
6.5.3 Випадкові числа в R 259

7 Одновимірний графічний аналіз 262
7.1 Гістограми 262
7.2 P-P- та Q-Q- діаграми 270
7.3 Q-Q-діаграма з прогнозними інтервалами 278

7.4 Порівняння розподілів кількох наборів даних 281
7.5 Скриньки з вусами 285

8 Оцінювання параметрів 291
8.1 Оцінки методу моментів 292
8.2 Оцінки методу квантилів 306
8.3 Оцінки методу найбільшої вірогідності 310
8.4 Асимптотична нормальність 320
8.5 Довірчі інтервали та еліпсоїди 343
8.5.1 Побудова асимптотичних довірчих інтервалів 343
8.5.2 Побудова довірчих еліпсоїдів 350
8.6 Оцінювання для стандартних розподілів 360

9 Перевірка гіпотез 368
9.1 Загальні відомості 368
9.2 Тест для простих гіпотез 374
9.3 Тест відношення вірогідності для складних гіпотез 391
9.3.1 Загальна схема тестів відношення вірогідності 391
9.3.2 Асимптотика тесту відношення вірогідності для вкладених гіпотез 396
9.3.3 Багатовибіркові задачі 400
9.4 Тести за довірчими інтервалами 402
9.5 Тести для нормального розподілу 411
9.5.1 T-тест. Перевірка гіпотез про середні. Дисперсія — заважаючий параметр 413
9.5.2 F-тест. Перевірка гіпотез про дисперсії. Заважаючий параметр — математичне сподівання 423
9.5.3 Z-тест для гіпотез про середнє без заважаючих параметрів 425
9.5.4 Знову про тести для дисперсії 429

9.5.5 Знову про тести для математичних сподівань 434
9.6 Тести χ^2 437
9.6.1 Тест χ^2 для простих основних гіпотез 437
9.6.2 Тест χ^2 для складної основної гіпотези . . . 443
9.6.3 Тест χ^2 з групуванням для перевірки узгодженості 450
9.6.4 Перевірка незалежності двох змінних тестом χ^2 453
9.7 Перевірка залежності двох змінних 459
9.7.1 Однофакторний дисперсійний аналіз 460
9.7.2 Тести кореляцій 469
9.7.3 Порівняння різних підходів до пошуку залежності 472

10 Регресія **484**
10.1 Проста лінійна регресія 487
10.2 Лінійний метод найменших квадратів 507
10.3 Позначення у формулах для функції `lm()` 533
10.4 Перевірка лінійних гіпотез. Тест Фішера 538
10.5 Нелінійний МНК 557

Література **566**

Додатки **571**

А Векторна і матрична алгебра **571**

Б Відомості з теорії ймовірностей **576**
Б.1 Випадкові величини та вектори 576
Б.2 Умовні ймовірності. Незалежність 580
Б.3 Багатовимірний нормальний розподіл 583
Б.4 Збіжність та граничні теореми 584

Покажчик засобів `R` **586**

Список позначень та скорочень

в.в. — випадкова величина;

м.н. — майже напевне;

ф.р. — функція розподілу;

$\mathrm{Binom}(n, p)$ — біноміальний розподіл для n випробувань з ймовірністю успіху p;

cov — коваріація або коваріаційна матриця;

cor — кореляція (теоретична);

$\mathsf{D}\,\xi$ — дисперсія сподівання в.в. ξ;

$\mathrm{Exp}(\lambda)$ — еспоненційний розподіл з інтенсивністю λ;

$\mathsf{E}\,\xi$ — математичне сподівання в.в. ξ;

$F_\xi(x)$ — функція розподілу випадкової велични ξ у точці x.

$\mathrm{GM}(X)$ — середнє геометричне вибірки X;

$\mathrm{HalfM}(\sigma^2)$ — півнормальний розподіл з параметром σ^2;

$\mathrm{HM}(X)$ — середнє гармонічне вибірки X;

$\mathrm{IQ}(X)$ — інтерквартильний розмах вибірки X;

$\mathrm{LN}(\mu, \sigma^2)$ — логнормальний розподіл з параметрами μ, σ^2;

$\mathrm{lr}(X)$ — логарифмічне відношення вірогідності;

$\mathrm{LR}(X)$ — відношення вірогідності;

$\mathrm{Poiss}(\lambda)$ — розподіл Пуассона з параметром λ;

$\mathsf{P}(A)$ — ймовірність події A;

$r(X, Y)$ — коефіцієнт кореляції Пірсона між змінними X і Y;

$S^2(X)$ — вибіркова дисперсія вибірки X;

$S_0^2(X)$ — виправлена вибіркова дисперсія вибірки X;

$\mathrm{med}(X)$ — медіана вибірки X;

$N(\mu, \sigma^2)$ — нормальний розподіл з математичним сподіванням μ і дисперсією σ^2;

$Q^F(\alpha)$ — квантиль рівня α для розподілу F;

$Q^X(\alpha)$ — вибірковий квантиль рівня α для вибірки X;

$\bar{X}$ — вибіркове середнє вибірки X;

$X_{[j]}$ — j-ий елемент вибірки X, впорядкованої за зростанням;

$\mathrm{Unif}(a, b)$ — рівномірний розподіл на інтервалі $[a, b]$;

$\rho(X, Y)$ — коефіцієнт кореляції Спірмена між змінними X і Y;

$\tau(X, Y)$ — коефіцієнт кореляції Кендалла між змінними X і Y;

$\Phi(x)$ — функція розподілу стандартного нормального розподілу;

$\varphi(x)$ — щільність стандартного нормального розподілу;

$\mathbb{1}\{A\}$ — індикатор події A;

$\#A$ — кількість елементів множини A;

◀ — закінчення тексту прикладу.

Передмова

Всім, хто хоч трохи знайомий із статистикою або комп'ютерами, зрозуміло, що в одній книзі розповісти про всю комп'ютерну статистику неможливо. Ця книга створювалась як підручник для студентів механіко-математичного факультету, які вивчають курси комп'ютерної, дескриптивної та математичної статистики. Проблема полягає в тому, що такі курси читаються і на третьому, і на четвертому, і на п'ятому роках навчання. При цьому студенти можуть спеціалізуватись і з математики, і з статистики, і з актуарної справи. Значною мірою, саме цим обумовлена структура книги: поєднання розділів, орієнтованих на читачів з різним науковим вантажем, але об'єднаних наскрізними спільними мотивами та сюжетами. Сподіваюсь, що вона може стати у пригоді не тільки нашим студентам, а і багатьом іншим читачам, що цікавляться застосуванням статистики до дослідження реальних даних та можливостями розробки своїх власних технологій статистичного аналізу.

Отже, про що йдеться?

Перші три розділи присвячені системі статистичного програмування R. Тут розповідається, що таке R, звідки його треба брати і як встановлювати, які є основні засоби програмування мови R і як працює базова графіка у цій системі. Цю частину книги можна використовувати для першого знайомства з R. Тут скажу тільки, що R це не лише гнучка і зручна мова програмування та набір функцій, які реалізують практично всі сучасні технології статистичної обробки даних. Це також перепустка в R-спільноту, братство, де ви

завжди зможете знайти підтримку та відповідь на найнесподіваніші питання. R — це свого роду сучасна латина, міжнародна мова, яка об'єднує статистиків-практиків і розробників нових ідей.

Далі у четвертому і п'ятому розділах розповідається про дескриптивні статистики, тобто про числові характеристики, які використовуються, коли великі набори статистичних даних треба коротко описати одним або кількома числами. Тут читач може дізнатись, коли середнє положення у вибірці доцільно характеризувати медіаною, а коли — геометричними середнім. Що таке робастність. Чому треба брати корінь з вибіркової дисперсії. Як відобразити структуру кореляцій у вигляді графа. І багато іншого.

Для читання перших п'яти розділів достатньо тільки знання елементарної шкільної математики. Аби лише не боятись символів $\sum$, $\prod$ і $\sqrt[n]{x}$.

Із шостого розділу починається застосування математичних, ймовірнісних моделей у статистиці. Тут описані основні ймовірністні розподіли, які використовуються для побудови теоретичних моделей даних. Крім того, пояснюється техніка генерації псевдовипадкових послідовностей із заданим розподілом. Такі послідовності широко використовуються у сучасній статистиці для аналізу реальних даних і для перевірки якості алгоритмів, про це розповідається далі. Для читання цього і наступних розділів треба мати хоча б мінімальне уявлення про похідні та інтеграли.

У сьомому розділі йдеться про графічні засоби аналізу розподілу даних, вибору теоретичної моделі і порівняння розподілів різних наборів.

Що можна робити, коли ви вибрали теоретичну модель? Оцінювати її невідомі параметри і перевіряти гіпотези про них. Цьому присвячені, відповідно, восьмий і дев'ятий розділи книги. Вони побудовані трохи незвично: від складного — до простого. Спочатку розповідь йде про загальні підходи до розв'язання задач оцінки і тестування гіпотез, показано, як їх можна самостійно реалізувати в R для дослідження даних, що описуються порівняно складними, нестандартними моделями. І лише потім сказано про стандартні

функції R, що реалізують ті ж методи у стандартних, широко застосовних моделях. По-справжньому розібратись у складному легше, ніж у простому: складне можна розкласти на простіші складові, а простого так не зрозумієш! Крім того, майстру, навченому робити складні речі, непотрібно буде вганяти нетипову справу у стандартні межі — як це часом буває з тими, хто починав з простого.

Нарешті, десятий розділ оповідає про техніку регресійного аналізу — розділ статистики, який на сьогодні має найбільше застосувань. Тут застосовуються всі ті підходи, які описані у книзі раніше. Але значну частину розділу можна читати майже не звертаючись до попереднього матеріалу. Наприклад, за бажанням читач може почати знайомство з книгою з підрозд. 10.1. І вже склавши певне уявлення про те, що буде наприкінці, звернутись до початку книги, якщо захоче розплутати всю інтригу.

У додатках вміщені ті відомості з алгебри та теорії ймовірностей, які студенти-математики зазвичай отримують у інших навчальних курсах, але деяким читачам вони можуть бути не зовсім відомі, або відомі в дещо іншому освітленні.

Трохи про те, чого немає у цій книзі — я одразу зазначу, де про це можна прочитати. Немає аналізу часових рядів [24, 47]. Немає того, що зазвичай називають машинним навчанням: нейронних мереж, техніки кросс-валідації, зниження вимірності, кластерного аналізу і т.д. [31, 33]. Майже немає непараметричної статистики [52]. Немає технік вибіркових обстежень [1]. Немає послідовного викладу математичних основ статистики і теорії ймовірностей [3, 9, 46, 51]. І хотілося б про все це написати, але десь треба ставити крапку.

Іще три рекомендації для тих, кому, можливо, варто відкласти цю книгу і пошукати більш корисної. Якщо вам потрібний швидкий огляд того, як реалізувати стандартні статистичні техніки у R, можна почати з [12]. Якщо ви взагалі не хочете мати справу навіть з мінімальним програмуванням, а волієте обмежитись такою обробкою, яку можна реалізувати, обираючи пункти у стандартному меню, варто скористатись, пакетом Statistica, див. [15]. Якщо вас

цікавить не конкретна комп'ютерна платформа, а загальна логіка комп'ютерного статистичного аналізу у застосуванні до конкретних прикладних задач, хорошим підручником для вас може стати [44].

Я прошу читачів за все цікаве і хороше, що знайдеться у цій книзі, разом зі мною подякувати моїм колегам, чиї поради, коментарі і критика значною мірою сформували її. Спеціально хочу подякувати тим, хто читав і вичитував цю книгу в процесі її підготовки — О. Сугаковій, А. Оленку, А. Пилипенку. А також своїм студентам, від яких я також багато чому навчився, особливо у програмуванні на R. Ви можете приєднатись до цих людей, написавши мені свої відгуки на адресу mre@univ.kiev.ua

Успіхів вам у статистиці і не тільки.

Р. Майборода

Розділ 1

Початок роботи з R

У цьому підрозділі міститься переважно технічна інформація, корисна для тих, хто вперше вирішив поекспериментувати з R. Тут розподідається, що таке R, R-Studio, пакети (бібліотеки), чим вони можуть бути корисні, де їх взяти і як встановити на своєму комп'ютері. Пояснюється також, як вести діалог з R і R-Studio, користуватись help-системою і де в інтернеті можна шукати порад щодо роботи з R, коли вони вам потрібні.

Якщо R вже встановлений на вашому комп'ютері і перші кроки були успішними, пропускайте цей розділ і переходьте до наступного.

1.1 Що таке R і де його взяти

R — це середовище програмування для статистичного аналізу даних. Воно складається з базової програми R, що працює як інтерпретатор мови статистичного програмування S, та окремих пакетів, які реалізують спеціальні методи та технології статистичної обробки даних. Базова програма створена як складова проекту GNU. Вона є альтернативною програмною реалізацією мови S (ця мова та комерційний пакет S+ для її реалізації були розроблені

у Bell Laboratories під керівництвом Дж. Чемберса). На відміну від S+, програма R є некомерційною і вільно розповсюджується за умови дотримання вимог GNU General Public License. Комерційний проект S+ сьогодні практично не активний, остання версія програми випущена у 2007 р. Подальший розвиток ідей, закладених у мові S, та їхня реалізація продовжується в системі R. Тому сучасна версія мови також має назву R. Але ряд книг, написаних з орієнтацією на S та S+, зберігає свою актуальність, оскільки у них питання прикладного застосування часто пояснюються детальніше і зрозуміліше, ніж у документації до R, розробленій ентузіастами.

Офіційна сторінка проекту R розміщена за адресою:

`http://www.r-project.org/` .

Отримати останню версію інсталятора базової програми R для операційної системи Windows можна за адресою[1]:

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

(На 12 січня 2019 р. це була версія R-3.5.2). Інсталятор завантажується у вигляді exe-файла. Для інсталяції програми досить запустити цей файл і відповідати на запитання, які задаватиме інсталятор. При першій спробі роботи з R рекомендовано погоджуватись з усіма пропозиціями, які робить інсталятор.

Проблеми можуть виникнути, якщо на вашому комп'ютері встановлені різні права доступу для різних користувачів. Справа в тому, що R наприкінці кожної сесії роботи зберігає на диску "робочий простір" (workspace) — сукупність даних та програм, які були завантажені під час сесії. На початку наступної сесії робочий простір завантажується з диска. Якщо під час інсталяції для зберігання робочого простору буде обрано директорію, недоступну певному користувачеві, то при роботі з R можуть виникати повідомлення про неможливість завантаження або зберігання workspace. Для усунення таких повідомлень потрібно або вибрати директорію вільного доступу при інсталяції, або змінити директорію, використовуючи пункт File->Change Dir. . . у головному меню

[1] У цій книзі я орієнтуюсь на використання R у системі Windows, але на сайті `r-project` можна знайти версії і для інших операційних систем.

головного вікна програми R.

Після інсталяції R його можна запустити й отримати приблизно таке вікно, як зображено на рис. 1.1. Тут зверху розміщено головне меню, а нижче відкрито вікно "консолі R", де можна давати команди програмі та отримувати її відповіді. Синім кольором у цьому вікні виведено початкову інформацію про вашу версію базової програми R. Потім іде інформація про те, що R поширюється вільно за умови виконання вимог ліцензії. Далі червоним кольором можуть бути вказані команди, які R виконав автоматично при завантаженні. Нарешті, червоний символ > є запрошенням користувачу вводити власні команди.

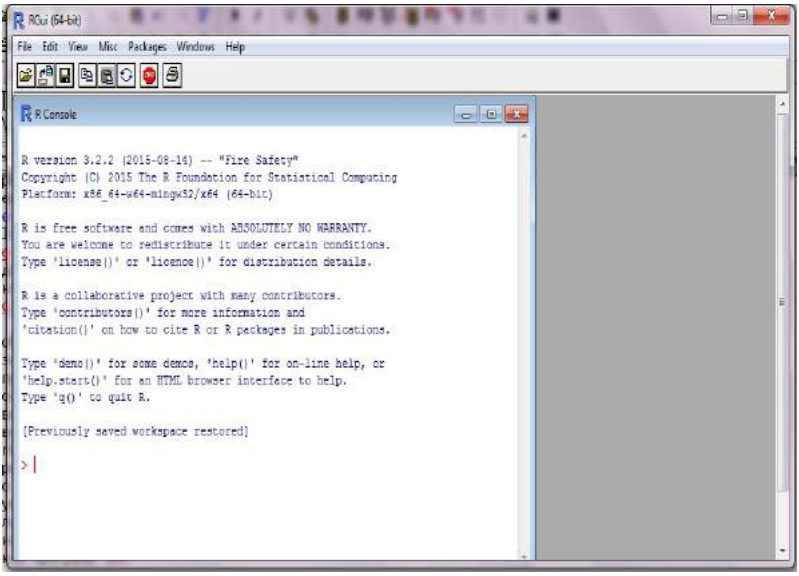

Рис. 1.1. Початок роботи з R

Для перевірки роботи системи можна після того, як на екрані з'явиться символ > ввести з клавіатури 2+2 і натиснути Enter. Результат виконання буде виведено на консоль буде виведено на консоль. Діалог з комп'ютером буде мати такий вигляд:

```
> 2+2

[1] 4
```

R виводить результати виконання безпосередньо після команди синім кольором, після чого переходить у режим очікування наступної команди, про що повідомляє червоним знаком >. При роботі з R можна виконувати одразу багато команд, що записані в окремому файлі. Найпростіший спосіб зробити це — в якому-небудь текстовому редакторі завантажити такий файл, зробити там сору, а потім — paste на консолі. При цьому, якщо команди в файлі розміщені в окремих рядках, розділових знаків між ними не потрібно. Команди, вміщені в одному рядку, розділяють символом ; .

Якщо довга команда не вміщується у одному рядку, її можна розбити на декілька рядків, причому, при переході до наступного рядку R автоматично виводить символ продовження +. R сам здогадується, що команда не закінчена — за її синтаксисом. Тому деякі синтаксичні помилки (як от — забуті дужки) можуть сприйматись як незакінчені команди. У цьому випадку R виставить + на початку наступного рядка і перейде у режим очікування. Натисніть esc, щоб перейти у режим введення нової команди без продовження аналізу попередньої.

Програми, що складаються з команд R, називають *скриптами* (script). Вони мають стандартне розширення .r. У базовій програмі є можливість відкрити вікно редактора для створення нового скрипта або завантажити файл зі скриптом, використовуючи пункти головного меню File->New script, або File->Open script. Виконати завантажений у вікні редактора скрипт повністю можна, використовуючи Edit->Run all. Можна також виконати виділену частину скрипта використовуючи кнопку “Run line or selection”. Закінчивши роботу зі скриптом, його можна зберегти, використовуючи File->Save.

1.2 Система R-Studio

За потреби, всі технології статистичної обробки можна реалізовувати, використовуючи лише базовий пакет R. Але він спеціально розроблений так, щоб забезпечувати лише мінімально необхідні засоби реалізації. Для більш зручного користування R-технологіями можна використовувати спеціальні надбудови-оболонки над R, які дають більше можливостей для програмування, перегляду використаних змінних, користування графікою та роботи з Help-системою.

Система R-Studio є однією з таких оболонок-інтеграторів. Вона також розповсюджується безкоштовно. Інсталятор R-Studio можна отримати на офіційному сайті www.rstudio.com. Перш ніж інсталювати цю програму, треба встановити на комп'ютері базовий R. Після цього можна запустити інсталятор R-Studio і погоджуватись з усіма його запитаннями.

При роботі вікно R-Studio може мати приблизно такий вигляд, як зображено на рис. 1.2.

Основне вікно розділене на чотири дочірні вікна. У лівому верхньому вікні виведено script — програму, яка редагується. (У цього вікна багато закладинок, які дозволяють працювати з кількома файлами-скриптами одразу). У лівому нижньому вікні — консоль, у якій виконуються команди. Тут також можна запускати скрипти або їхні частини. У правому верхньому вікні можна переглядати активні змінні, з якими працює програма. Тут також можна побачити історію роботи — що ви робили у R не тільки підчас даної сесії (це відображено на консолі), а і раніше, у попередніх сесіях.

Найбільш навантажене вікно внизу праворуч. Сюди виводять рисунки, які робить програма, тут можна проглянути Help, подивитись, які додаткові пакети завантажені, а також працювати з різними файлами з вашого комп'ютера.

Звичайно, користувач може поміняти ці вікна місцями, змінювати їх розміри та користуватись іншими можливостями системи.

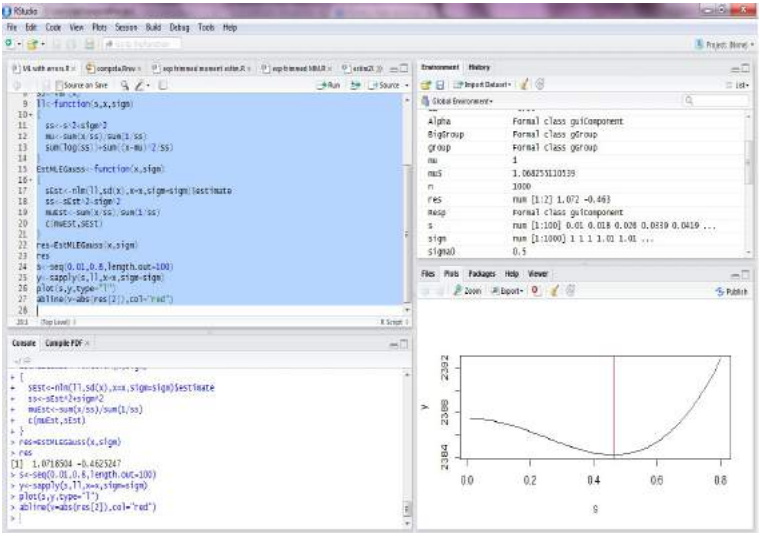

Рис. 1.2. Початок роботи з R

Зокрема, користуючись головним меню, можна перезавантажувати R, вибирати новий робочий каталог (тобто каталог, з якого R завантажує файли стандартно), зберігати та завантажувати у пам'ять workspace, отримувати з інтернету нові бібліотеки програм/даних (packages). Усе це можна робити і безпосередньо з R, але в R-Studio такі речі організовані зручніше.

R-Studio корисний також підказками, які він робить під час набору команд на консолі та у вікні скриптів.

Ще одна додаткова зручність R-Studio — можливість генерації текстових звітів, які виконуються з поєднанням системи програмування R та системи форматування текстів LaTeX. При цьому R-Studio спирається на R пакет `knitr`. За допомогою цієї технології підготовлена ця книга. На жаль, для повного опису `knitr` необхідно пояснювати не тільки роботу R, а і принципи організації LaTeX, що виходить за межі цієї книги.

1.3 Пакети, Help та інше

Базовий R має великий набір функцій для реалізації математичних та статистичних алгоритмів. Але користувачі весь час розробляють свої власні функції, що доповнюють базові. Коли деякий набір функцій, що реалізують певну технологію статистичної обробки даних, буде відпрацьований настільки, що у розробника виникає бажання поділитись ним також з іншими можливими користувачами, він оформлює такий набір у вигляді пакета (package). Пакет повинен мати також help-документацію, яка дозволить можливим користувачам зрозуміти його призначення. До пакета часто включають і набори даних, на яких можна перевірити роботу його функцій. Бувають пакети, складені лише з даних, — це просто колекції цікавих або популярних прикладів, які хтось підібрав для власних потреб.

Правильно оформлені пакети розробники відсилають до депозитаріїв, звідки їх можна переписати на свій комп'ютер у каталог, доступний для R (інсталювати). Оскільки пакети створюються різними розробниками за власною ініціативою, незалежно один від одного, то між ними можуть існувати неузгодженості. Наприклад, функції з різних пакетів можуть мати однакові імена та типи параметрів, тоді при завантаженні у пам'ять комп'ютера обох пакетів користувач не зможе правильно їх використовувати[2]. Тому рекомендується завантажувати не всі інстальовані на комп'ютері пакети, а лише ті, які дійсно потрібні для роботи під час даної сесії.

Програмісти, що створюють у R нові пакети для загального користування, розрізняють поняття *пакет* (package) і *бібліотека* (library). *Пакетом* називають файл, або набір файлів з скриптами та їхнім описом, а *бібліотекою* — місце, тобто каталог у файловій системі, де лежить пакет. З точки зору користувача-початківця, ця відмінність не суттєва. Старожили пам'ятають, що в мові S library

[2]Функції, що мають однакові імена, але працюють з параметрами різних типів, — це нормальне явище для об'єктно-орієнтованих мов. Комп'ютер під час виклику обирає правильну функцію, виходячи зі специфікації її параметрів.

означало приблизно те ж, що у R називають package. У цій книзі ми також не будемо надавати ваги цій відмінності.

Інсталювати пакет на комп'ютері, тобто отримати його з інтернету у вигляді zip-архіву, розархівувати і покласти у зручне для R місце, можна таким чином:

1. Під час сесії роботи з R з консолі, викликавши функцію інсталяції `install.packages`. Наприклад, команда

```
install.packages('raster')
```

викличе звертання комп'ютера до стандартного депозитарію (зазвичай це `cran.us.r-project.org`), отримання від нього пакета і розміщення його у відповідному каталозі на комп'ютері. Звичайно, якщо комп'ютер не має виходу в інтернет, або депозитарій недоступний, в результаті виконання функції виникне помилка.

2. При роботі безпосередньо з базовим R інсталяцію можна робити, використовуючи пункти головного меню Packages-> Install package(s). Спочатку програма пропонує вибрати інтернет-архів, з якого робиться інсталяція. Варіант 0-cloud, що пропонується стандартно, зазвичай працює цілком задовільно. Після цього треба у списку вибрати потрібний для вас пакет. Якщо цей пакет використовує які-небудь інші, яких немає на вашому комп'ютері, вони будуть інстальовані автоматично.

3. При роботі з R-Studio можна скористатись пунктами головного меню Tools->Install packages. . . При цьому відкривається діалогове вікно, де ви можете вказати, звідки проводиться інсталяція (з інтернет-депозитарію, чи з zip-архіву на вашому комп'ютері), який пакет ви хочете інсталювати, місце, де буде розміщений пакет, і чи треба інсталювати інші пакети, які ним використовуються.

Для завантаження (підключення) пакета у робочу область пам'яті (`Workspase`) під час сесії використовують функцію `library()`. Наприклад, `library(raster)` завантажує в робочу область пакет `raster` і дає змогу використовувати всі його функції у подальшій сесії.

Функцію `detach()` відключає пакети. Наприклад, команда

```
detach("package:raster")
```

зробить пакет `raster` неактивним — його функції перестануть бути доступними у подальшій сесії. Але при цьому він залишиться в оперативній пам'яті.

Видалити непотрібний пакет з `Workspase` можна, використовуючи функцію `remove()`, наприклад:

```
remove.packages("raster").
```

Для того, щоб очистити весь `Workspase`, можна набрати

```
remove(list = ls()).
```

Для того, щоб отримати довідку про які-небудь можливості `R`, можна скористатись help-системою. Для цього призначена функція `help()`, або скорочено `?`. Набравши на консолі `R` команду

```
?sin
```

ви отримаєте довідку про тригонометричні функції в `R`, зокрема, і про функцію $\sin(x)$. Довідка, зазвичай, починається з інформації про те, у якому пакеті міститься функція (дані, об'єкти тощо) У випадку `?sin` це має вигляд: `Trig {base}`, що вказує на набір тригонометричних функцій з базового `R`.

Інколи назву функції (тему help) після знака запитання слід задавати у лапках. Так, при спробі викликати help запитом `?+` (або `?for`) ви у відповідь отримаєте запрошення продовжувати введення команди (`+`). Якщо набрати `?"+"` (відповідно — `?"for"`) можна отримати довідку про реалізацію арифметичних операцій (або про цикл `for`) у `R`.

Базовий `R` для перегляду help-документації може запускати інтернет-браузер, але це не означає, що документація шукається у інтернеті. Все, що видається за командою `?` або `??`, міститься на вашому комп'ютері і не потребує доступу до інтернету.

Команда `?` виводить основний файл, пов'язаний з темою запитання. Для прпереглядy всх файлів, де згадується ця тема, можна скористатись функцією `help.search()`, скорочено — `??`. Наприклад, за запитом

```
??"linear models"
```

у браузері буде виведена сторінка з переліком усіх документів help-у, де згадуються лінійні моделі з коротким описом їхнього

змісту. Більшість із цих сторінок буде стосуватись лінійних регресійних моделей, або узагальнених лінійних моделей. Але можливі і посилання на лінійні моделі чогось зовсім іншого. Переходячи за гіперпосиланнями, можна продивлятись ці документи.

Задаючи спеціальні параметри функцій `help.search()` або `help()`, можна отримувати довідки для окремих пакетів, або тільки за ключовими словами, або тільки за документами з певного каталогу і т.д.

Документацію для help автори пакетів розробляють самі, і поставляється вона разом з пакетами. Тому за запитом `?` ви отримуєте інформацію лише про ті функції, які містяться у пакетах, доступних під час сесії (підключених при запуску `R` або додатково командою `library()`). Якщо, скажімо, на початку сесії я запитаю

```
?ginv
```

відповідь буде:

```
No documentation for 'ginv' in specified packages
and libraries: you could try '??ginv'
```

При запиті `??ginv` виводиться вся інформація про `ginv`, що є на комп'ютері (як підключена до сесії, так і не підключена). Зокрема, на моєму комп'ютері на сторінці довідки з'являється гіперпосилання

```
MASS::ginv      Generalized Inverse of a Matrix,
```

що вказує на наявність функції `ginv` у пакеті `MASS` і коротко описує її призначення — знаходження узагальнених обернених матриць. А після підключення пакета `MASS` довідка про цю функцію стане доступною за запитом `?ginv`. При роботі в R-Studio комп'ютер стане підказувати вам параметри цієї функції при наборі і т.д.

У R-Studio у вікні help (праворуч знизу на екрані у стандартній конфігурації) є поле для пошуку (Search), яке діє аналогічно запиту `?`, але при цьому дає додаткову підказку при наборі.

Якщо ви хочете отримувати інформацію про можливості всіх функцій з усіх пакетів, що лежать у всіх доступних депозитаріях, то ви можете інсталювати на своєму комп'ютері пакет `sos`. Після підключення його до сесії (`library(sos)`) стане доступним запит

у формі ???<тема>, за яким буде видаватись результат пошуку заданої теми по всіх R-депозитаріях світу. Розібратись у таких багатосторінкових переліках буває не просто, але інколи вони дають несподівані і дуже корисні результати.

Оскільки документацію до пакетів розробляють їхні автори, то вона часто буває переобтяженою технічними подробицями, не дуже зрозумілими початківцю. Логіка застосування програми (очевидна авторам) при цьому втрачається. Тому дуже корисним буває ознайомлення з думками користувачів. Найпростіше знайти такі думки, скориставшись якою-небудь пошуковою інтернет-машиною (я віддаю перевагу Google). Набравши, скажімо, запит

`"inverse matrix in r"`

— ви отримаєте посилання на багато різних рекомендацій щодо знаходження обернених матриць за допомогою R. Не всі вони будуть адекватними! Я рекомендую звертати увагу на поради сайтів:

`stackoverflow.com`

— це сайт програмістів та математиків, тут можна знайти поради спеціалістів та обговорення проблем на серйозному рівні.

`www.statmethods.net`

— тут можна шукати швидкі і прості поради.

`cran.r-project.org/doc/FAQ/`

— це офіційний сайт R, місце де зібрані відповіді на запитання, що виникають особливо часто.

На ютубі можна також побачити лекції з багатьох окремих питань використання R у стилі "зрозуміло навіть немовлятам". Вони можуть бути корисними на перших етапах вивчення R, щоб не почуватися зовсім безпорадним. Потім їх зрозумілість починає дратувати. Але, коли ви набуваєте певного досвіду у роботі з R, і виникає потреба поділитись ним з іншими, перегляд таких лекцій знову може стати у пригоді.

Розділ 2

Мова статистичного програмування R

Цей розділ присвячений першому знайомству з програмуванням мовою R. Тут, в першу чергу, розповідається те, без чого неможливо розуміти тексти на R, вміщені у наступних розділах. Звичайно, дещо викладено більш детально, щоб читач міг також орієнтуватись у простих R-програмах, які можуть потрапити у поле його зору, наприклад, при пошуку в інтернеті. Але неможливо стати програмістом на R прочитавши тільки цей розділ. Для подальшого знайомства з програмуванням на R, а також з особливостями комп'ютерної реалізації цього середовища, можна рекомендувати книгу [54]. Різним перетворенням наборів статистичних даних за допомогою R присвячені книги [20] і [48]. З техніками статистичного аналізу даних на основі R можна ознайомитись у [49] і [50].

2.1 Базові поняття

Мова R складається з *команд*. Кожна команда може виконуватись окремо або у складі програми. Програми у R називають *скриптами* (script). Окрема команда записується у командному рядку

системи R після *запрошення* “>” і запускається на виконання клавішею Enter:

```
> 1+1

[1] 2
```

(Запрошення комп’ютер видає автоматично).

Команда може бути *виразом* (тоді результат її виконання одразу виводиться на екран, як у попередньому прикладі) або привласненням:

```
> x<-1+1
> x

[1] 2
```

Тут “<-” це символ привласнення, праворуч від нього іде вираз, значення якого обчислюється, а ліворуч — ім’я змінної, якій привласнено обчислене значення. Саме значення не виводиться на екран, щоб побачити, чому тепер дорівнює змінна x, ми ввели її назву у наступному рядку після запрошення.

Команди виконуються після натискання на клавішу Enter. Коли при цьому комп’ютер за синтаксисом помічає, що команда не закінчена, він у наступному рядку замість запрошення “>” виводить символ продовження введення “+” і ви можете закінчити введення команди:

```
> x<-2*
+ 3
> x

[1] 6
```

Команди працюють з *об'єктами*[1]. Прикладами об'єктів є вектори, матриці, фрейми (набори) даних, функції. Часто функція, що виконує певну процедуру статистичного аналізу даних (напр. — регресійний аналіз) всі свої результати записує в окремий об'єкт, який потім можна надрукувати на екрані у вигляді звіту або відобразити у вигляді серії рисунків, або використати для подальшого аналізу іншими функціями.

Грубо кажучи, об'єкт — це поіменована ділянка пам'яті, де зберігається все, що стосується цього об'єкта: значення координат вектора, опис роботи функції тощо. У R, як і в інших об'єктно-орієнтованих мовах, в одному об'єкті можуть об'єднуватись дані і функції, що їх обробляють. Дані, з яких складається об'єкт, називають його *атрибутами*, а функції, які з ним пов'язані — *методами*. Об'єкти належать різним *класам*. Клас визначає, які атрибути може/повинен мати об'єкт і які методи можна до нього застосовувати.

До атрибута `atr` об'єкта `x` можна звертатись використовуючи символ `$`:

```
x$atr
```

Наприклад:

```
> x<-NULL
> x$name<-"Коваленко"
> x$age<-24
> x$married<-F
> x

$name
[1] "Коваленко"

$age
```

[1] Якщо ви не зовсім відчуваєте, про що йдется у цьому та наступному абзацах, не переймайтесь. Все стане зрозуміліше у подальших прикладах. При першому читанні можна одразу перейти до наступного підрозділу.

```
[1] 24

$married
[1] FALSE
```

△

— ми спочатку створили порожній (NULL) об'єкт x, а потім надали йому атрибути name, age і married. Щоб подивитись, з чого тепер складений об'єкт x, ми просто набрали його назву у скрипті, і R відобразив всі x атрибути на екрані.

Тепер з x можна працювати як з єдиним цілим, наприклад, присвоїти його значення новому об'єкту y:

▽

```
> y<-x
> y

$name
[1] "Коваленко"

$age
[1] 24

$married
[1] FALSE
```

△

— тепер y також став 24-літнім неодруженим Коваленком.

Зверніть увагу, що при привласненні за допомогою y<-x створюється новий об'єкт y (якщо у y було старе значення, воно втрачається) і у цей об'єкт копіюється значення x. Тепер можна змінювати атрибути y, але атрибути x при цьому залишаться старими:

▽

```
> y$married<-T
> y$married

[1] TRUE
```

```
> x$married

[1] FALSE
```

Наш y-Коваленко одружився, але x-Коваленко залишився не одруженим.

Для того, щоб побачити значення об'єкта x, зазвичай, досить ввести команду x. Якщо необхідно більш акуратне відображення, можна скористатись функцією print(). Зокрема, якщо записати print(z,digits=3) то R намагатиметься виводити не більше трьох значущих цифр у всіх числових атрибутах z:

```
> z=1/3
> z

[1] 0.3333333

> print(z,digits=3)

[1] 0.333
```

Структуру об'єкта зручно перевіряти, використовуючи функцію str():

```
> str(x)

List of 3
 $ name   : chr "Коваленко"
 $ age    : num 24
 $ married: logi FALSE
```

Тут ми не тільки бачимо імена і значення атрибутів, але можемо дізнатись про їхній тип: символьний для name, числовий для age і логічний для married.

2.2 Типи даних та елементарні функції

2.2.1 Вектори. Арифметичні та логічні операції

Найпростішою структурою у мові R є вектор (скаляри як окремі структури не існують, а трактуються як вектори одиничної довжини).

R використовує п'ять простих векторних типів об'єктів:

- `logical`: логічний — вектор складений з елементів що набувають значень "істинно" (`TRUE` або `T`) та "хибно" (`FALSE` або `F`);
- `numeric`: числовий — вектор, складений з дійсних чисел;
- `integer`: цілий — вектор, складений з цілих чисел;
- `complex`: вектор, складений з комплексних чисел;
- `character`: символьний — вектор, елементами якого є символьні рядки.

Якщо в одному наборі даних треба об'єднати елементи різної природи, використовують об'єкт типу `list` — список[2].

Створити будь-який вектор можна використовуючи функцію `c()`, яка об'єднує різні списки в один об'єкт (конкатенація)[3]:

```
> c(1,5,-3,4)

[1]  1  5 -3  4
```

— з чотирьох одноелементних векторів створений числовий вектор, що складається з чотирьох елементів (1,5,-3,4) і результат роботи виведено на екран.

[2] Список це не вектор!

[3] Списки створюються функцією `list()`.

Зверніть увагу, що, хоча всі використані нами числа були цілими, вектор має тип `numeric`. Для того, щоб сказати комп'ютеру, що ви хочете по можливості використовувати цілочислову арифметику при роботі з даним вектором, необхідно або перетворити його на цілочисловий за допомогою функції `as.integer()`, або поставити літеру `L` після кожного числа:

```
> str(c(1,5,-3,4))

 num [1:4] 1 5 -3 4

> str(c(1L,5L,-3L,4L))

 int [1:4] 1 5 -3 4

>  str(as.integer(c(1,5,-3,4)))

 int [1:4] 1 5 -3 4
```

З числовими векторами можна виконувати звичайні дії додавання, множення і т.д. З логічними — операції `&` (логічне і), `|` (логічне або), `!` (заперечення) та ін. Операції порівняння (`<`, `>`, `<=`,`=>`,`==`,`!=`) застосовуються до числових даних і дають логічний результат.

Усі ці операції застосовуються до векторів поелементно:

```
>  x<-c(1,5,-3,4)
>  y<-c(3,-1,2,1)
>  x+y

[1]  4  4 -1  5
```

Також поелементно застосовуються до векторів елементарні функції `sin`, `log` і т.д.

Для цілочисельного ділення використовується операція `%/%` для знаходження залишку від ділення — `%%`..

Якщо у бінарній операції вектори-аргументи мають різну довжину, то коротший аргумент повторюється циклічно при виконанні операції:

```
>   x<-c(1,2)
>   y<-c(3,3,3,3,3)
>   x*y

[1] 3 6 3 6 3
```

При цьому, якщо довжина довшого вектора не кратна довжині коротшого, комп'ютер видає попередження (warning) про це. Далі у прикладах попередження не відображаються.

При виконанні арифметичних дій можуть виникати значення Inf (нескінченність) та NaN (невизначено). З ними можна виконувати різні дії, які дають осмислений результат:

```
>   x<-1/0; x

[1] Inf

>   3-x

[1] -Inf

>   x>3

[1] TRUE

>   x/x

[1] NaN
```

Окрім значення NaN, яке відповідає невизначеності, пов'язаній з арифметичними операціями, в R використовується також значення NA, що позначає *пропущені значення*, тобто значення, які невідомі статистику. Особливості обробки NA-значень обговорюються далі. Крім того, можливе іще значення NULL, яке позначає порожній список.

Вектори можуть бути іменованими (named), у такому випадку кожен елемент вектора має ім'я. Щоб зробити вектор іменованим, треба задати для нього *атрибут* names:

```
> x<-c(5,4,3,2)
> names(x)<-c("відмінно","добре","задовільно",
+             "незадовільно")
> x

    відмінно        добре   задовільно незадовільно
           5            4            3            2
```

(У цьому скрипті виклик функції names() стоїть ліворуч від знака привласнення. У R такий синтаксис дозволений лише у невеликій кількості випадків. Зазвичай вживання виклику довільної функції ліворуч від <- трактується як помилка). Використання іменованих векторів часто буває зручним саме у статистичних застосуваннях, зокрема при звертанні до того чи іншого елемента вектора або масиву складнішої структури.

Розглянемо дві зручні функції для створення векторів. Якщо потрібен вектор, елементи якого утворюють арифметичну прогресію, можна скористатись функцією seq():

```
> seq(2.5,6,0.5)

[1] 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0
```

Виклик `seq` можливий у різних форматах (це характерна особливість не лише `seq`, а всіх функцій мови R) Формальна специфікація цієї функції така:

```
 seq(from = 1, to = 1,
 by = ((to - from)/(length.out - 1)),
    length.out = NULL, along.with = NULL, ...)
```

У цьому записі `seq` — ім'я функції, `from` (перший елемент), `to` (останній), `by` (крок), `length.out` (кількість елементів) і `along.with` — імена формальних параметрів. Після знака рівності вказані значення, яких ці параметри набувають стандартно (by default), якщо вони не вказані у виклику функції (якщо, скажімо, параметр `from` не заданий у виклику, то першим елементом буде 1). Наприклад, у цьому скрипті крок прогресії не заданий явно, він обирається комп'ютером так, щоб кількість елементів дорівнювала заданому значенню `length.out`:

```
> seq(2,10,length.out=6)

[1]  2.0  3.6  5.2  6.8  8.4 10.0
```

Символ `...` у специфікації позначає, що у функції можуть бути й інші параметри.

Як працюватиме функція `seq()` при виклику з іншими наборами параметрів, можна подивитись у help, задавши команду `?seq`.

Для випадку, коли крок послідовності дорівнює ± 1, можна використовувати скорочений запис `seq` у вигляді `from:to`, наприклад:

```
> 5:10

[1]  5  6  7  8  9 10

> -5:10

 [1] -5 -4 -3 -2 -1  0  1  2  3  4  5  6  7  8  9 10
```

```
> -(5:10)

[1]  -5  -6  -7  -8  -9 -10

> -10:-5

[1] -10  -9  -8  -7  -6  -5
```

Функція `rep()` розмножує свій перший параметр задану кількість разів:

```
> x<-1:4
> rep(x,3)

 [1] 1 2 3 4 1 2 3 4 1 2 3 4

> rep(x,each=3)

 [1] 1 1 1 2 2 2 3 3 3 4 4 4
```

Числовий вектор, що складається з нулів, можна створити також функцією `numeric(n)`, де `n` — кількість елементів вектора:

```
> x<-numeric(5)
> x

[1] 0 0 0 0 0
```

2.2.2 Індексація векторів

Для того, щоб при обробці мати можливість використати певну частину вектора (матриці, багатовимірного масиву), у `R` застосовується дуже гнучка система індексації. Зараз ми обмежимось прикладами її застосування до векторів, матриці розглянемо далі. Як і у багатьох інших мовах програмування, i-й елемент вектора можна виділити, використовуючи прямі дужки:

```
> x<-5:1
> names(x)<-c("відмінно","добре","задовільно",
+             "незадовільно", "погано")
> x[2]

добре
    4
```

(Нумерація елементів векторів завжди починається з 1).

Можна звернутись до елемента за ім'ям, якщо воно є:

```
>  x["задовільно"]

задовільно
         3
```

Якщо у прямих дужках вказати вектор індексів, то виділиться підвектор відповідних елементів:

```
>  x[c(3,1,5)]

задовільно   відмінно     погано
         3          5          1
```

(елементи переставлені в тому порядку, в якому йдуть індекси). І нарешті, якщо задати від'ємні значення індексів, то відповідні елементи будуть вилучені з підвектора:

```
> x[-c(3,1,5)]

       добре незадовільно
          4            2
```

Це ще не все. Можна для індексації використовувати логічні вектори, тоді включатись у підвектор будуть тільки ті елементи, яким відповідає значення TRUE:

```
> x[c(T,T,F,F,T)]

відмінно    добре   погано
       5        4        1
```

У прямих дужках можна записувати будь-який вираз, значення якого будуть використані для індексації:

```
> x[x%%2==0]

      добре незадовільно
         4            2
```

Роботу цієї команди можна описати так: спочатку створюється логічний вектор `x%%2==0`, в якому TRUE відповідає тим елементам, які є парними числами, а потім за цим логічним вектором робиться відбір відповідних елементів у підвектор.

Вираз вигляду `x[...]` може стояти і у лівій частині команди привласнення, наприклад:

```
> alp<-c('a','b','c','d','e','f')
> alp[2]<-'bbb'
> alp

[1] "a"   "bbb" "c"   "d"   "e"   "f"

> alp[c(1,3)]<-c('u','v')
> alp

[1] "u"   "bbb" "v"   "d"   "e"   "f"
```

2.2.3 Фактори

Ще один векторний тип даних — *фактори* (`factors`) заслуговує спеціального розгляду. Елементи вектора факторів можуть набувати значень лише з фіксованого набору значень. Дані такого типу часто виникають у статистичних дослідженнях, коли досліджувані об'єкти розбиваються на кілька груп (категорій) за деякою ознакою, наприклад, люди за національністю, статтю, юридичні особи — за формою власності, слова — за частинами мови (іменник, прикметник, дієслово...) тощо[4]. Різні значення, яких може набувати фактор, прийнято називати *рівнями* (levels).

Різні рівні зручно позначати їхніми назвами, наприклад, тип валюти — USD, EUR, UAH, RUR. Скажімо, набір даних про тип валют, якими було зроблено платежі протягом дня, може мати вигляд

('USD','EUR','EUR','UAH','EUR','USD','UAH','RUR').

Якщо задати такий вектор конкатенацією:

```
> z<-c('USD','EUR','EUR','UAH','EUR','USD','UAH','RUR')
> z

[1] "USD" "EUR" "EUR" "UAH" "EUR" "USD" "UAH" "RUR"
```

то `z` буде мати тип character (символьні рядки). Щоб пояснити комп'ютеру, що йдеться про рівні деякого фактора, треба зробити перетворення типу:

```
> zf<-factor(z)
> zf

[1] USD EUR EUR UAH EUR USD UAH RUR
Levels: EUR RUR UAH USD
```

[4] Відповідно, змінні, що можуть набувати лише скінченну кількість значень, прийнято називати **категорійними** (*categorical variables*).

Тепер, хоча на екрані рівні фактора відображаються їхніми назвами, у внутрішньому представленні комп'ютера вони кодуються натуральними числами. Перелік різних рівнів виведено в рядку `Levels` у порядку зростання кодів. Якщо вам потрібен тільки цей перелік у вигляді символьного рядка, можна скористатись функцією `levels()`:

```
> zl<-levels(zf)
> zl

[1] "EUR" "RUR" "UAH" "USD"
```

Відповідні коди можна побачити, наприклад, використовуючи функцію `unclass()`:

```
> unclass(zf)

[1] 4 1 1 3 1 4 3 2
attr(,"levels")
[1] "EUR" "RUR" "UAH" "USD"
```

Зрозуміло, що використання векторів із факторів замість символьних рядків дозволяє економити місце у пам'яті комп'ютера, якщо довжина вектора велика, а кількість рівнів помірна. Крім того, задання переліку рівнів дозволяє перевірити наявність зайвих назв, що могли б утворитись внаслідок якихось помилок. У статистиці є багато алгоритмів обробки даних, що працюють саме з категорійними даними (напр., у дисперсійному аналізі та аналізі таблиць спряженості). З цим пов'язано виділення факторів у окремий тип.

Зауважимо, що у векторі факторів можуть зустрічатись не всі допустимі рівні, але для них будуть зарезервовані числові коди. Інформація про можливість їхньої появи зберігається в *атрибуті* levels:

```
> z2<-zf[c(1,2)]
> z2

[1] USD EUR
Levels: EUR RUR UAH USD
```

Якщо при виділенні підмножини вектора факторів треба вилучити рівні, що не зустрічаються у підмножині, це можна зробити, задавши *опцію* `drop`:

```
> z2d<-zf[c(1,2),drop=T]
> z2d

[1] USD EUR
Levels: EUR USD
```

У статистичних дослідженнях часто розбиття досліджуваних об'єктів на категорії проводится залежно від того, у який діапазон потрапляє певна числова характеристика цих об'єктів. Наприклад, домогосподарства можна розділити на категорії з високим (high), середнім (mid) та низьким (low) рівнем прибутку залежно від числового розміру їхніх прибутків. Для того, щоб робити це автоматично, застосовується функція `cut`:

```
> u<-c(6,5,4,3,2,1)
> ul<-cut(u,breaks=c(-Inf,2.5,3.5,Inf),
+         labels=(c('low','mid','high')))
> ul

[1] high high high mid  low  low
Levels: low mid high
```

Тут ми створили вектор зі значеннями числової характеристики `u` і розбили досліджувані об'єкти на три категорії в залежності від значень `u`. Опція `breaks` вказує межі інтервалів, що визначають ці категорії: до першої потрапляють об'єкти, для яких $u \in (-\infty, 2.5]$, до другої — з $u \in (2.5, 3.5]$, до третьої — $u \in (3.5, \infty)$. Назви цих категорій (рівнів факторів) задані в опції `labels`.

У цьому випадку (а також у багатьох інших) для рівнів фактора можна вказати природний порядок: low<mid<high. Для деяких інших факторів (як от для національності) якогось природного порядку не існує. Щоб вказати комп'ютеру на наявність порядку рівнів, вводиться тип `ordered` (впорядкований фактор):

```
> ulo<-ordered(ul)
> ulo

[1] high high high mid  low  low
Levels: low < mid < high
```

Деякі функції `R` аналізують впорядковані фактори спеціальним чином, не так, як невпорядковані.

2.2.4 Матриці, масиви та фрейми даних

Матриці в `R` обов'язково складаються з елементів одного типу (напр., тільки з чисел або тільки з логічних значень). Існує багато різних способів створити матрицю, наприклад, її можна скласти з окремих векторів-рядків функцією `rbind()` або з векторів-стовпчиків функцією `cbind()`:

```
> x1<-1:3
> x2<-5:7
> u<-rbind(x1,x2)
> u
```

```
   [,1] [,2] [,3]
x1    1    2    3
x2    5    6    7

> v<-cbind(x1,x2)
> v

     x1 x2
[1,]  1  5
[2,]  2  6
[3,]  3  7
```

(Зверніть увагу, що імена векторів перетворились на імена відповідних стовпчиків або рядків).

З цього прикладу зрозумілі правила індексації: перший індекс позначає рядок, другий — стовпчик, тобто `u[2,3]` — це елемент на перетині другого рядка і третього стовпчика матриці `u`. Використання індексів та імен дуже гнучке, як показують такі приклади:

```
> u[,1]

x1 x2
 1  5

> u[2,]

[1] 5 6 7

> v[,"x2"]

[1] 5 6 7

> v[1:2,"x2"]

[1] 5 6
```

"Вийнятий" з матриці стовпчик перетворюється на вектор-рядок. Якщо ви хочете отримати як результат матрицю, що складається з одного стовпчика, скористайтесь опцією `drop=F`:

```
> v[,"x2",drop=F]

     x2
[1,]  5
[2,]  6
[3,]  7
```

Інколи виділення частини матриці або вектора буває зручно оформляти не через індексацію, а використовуючи спеціальну функцію

`subset(x, subset, select, drop = FALSE, ...)`.

Параметри у цій функції: `x` — об'єкт, з якого виділяється частина, `subset` — умова на рядки, за якою відбувається виділення, `select` — перелік стовпчиків, що виділяються.

Інший спосіб створення матриці — функція `matrix()`, яка перетворює вектор на матрицю. Першим параметром функції є вектор, який використовується для заповнення матриці, параметри `ncol` і `nrow` задають кількість стовпчиків і рядків утвореної матриці.

Логіка роботи функції зрозуміла з таких прикладів:

```
> x<-1:10
> matrix(x,nrow=2)

     [,1] [,2] [,3] [,4] [,5]
[1,]    1    3    5    7    9
[2,]    2    4    6    8   10

> matrix(x,ncol=2)
```

```
     [,1] [,2]
[1,]    1    6
[2,]    2    7
[3,]    3    8
[4,]    4    9
[5,]    5   10
```

```
> matrix(x,ncol=2,nrow=2)
```

```
     [,1] [,2]
[1,]    1    3
[2,]    2    4
```

```
> matrix(x,ncol=3)
```

```
     [,1] [,2] [,3]
[1,]    1    5    9
[2,]    2    6   10
[3,]    3    7    1
[4,]    4    8    2
```

△

В останньому прикладі для того, щоб заповнити матрицю, довелось циклічно повторити вектор `x`.

Імена рядків та стовпчиків матриці можна задавати, використовуючи функцію `dimnames()` ліворуч від знака привласнення `<-`, як показано на прикладі:

▽

```
> x<-1:10
> X<-matrix(x,nrow=2)
> dimnames(X)<-list(c('first','second'),letters[1:5])
> X
```

```
       a b c d  e
first  1 3 5 7  9
second 2 4 6 8 10
```

△

(Тут функція `list()` створює список, що складається з двох елементів, кожний з яких є вектором. `letters` у `R` — це вектор, складений з латинських літер у алфавітному порядку.)

Кількість рядків (або стовпчиків) вже існуючої матриці `X` можна дізнатись, використовуючи `nrow(X)` (або `ncol(X)`).

Для задання назв тільки рядків (стовпчиків) можна використовувати функції `rownames()` (`colnames()`) . Ті ж функції використовуються, якщо треба дізнатись імена для існуючої матриці:

```
> rownames(X)

[1] "first"  "second"
```

Часто буває корисною функція `diag()`, яку можна застосовувати різними способами. Якщо її параметром є вектор, вона породжує діагональну матрицю:

```
> x<-1:3
> diag(x)

     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    2    0
[3,]    0    0    3
```

Якщо параметр — матриця, `diag()` виділяє її головну діагональ у вигляді вектора:

```
> X<-matrix(1:9,ncol=3)
> diag(X)

[1] 1 5 9
```

Нарешті, якщо `diag` зустрічається ліворуч від символа привласнення, вона замінює діагональ свого матричного параметра:

```
> diag(X)<-rep(0,3)
> X

     [,1] [,2] [,3]
[1,]    0    4    7
[2,]    2    0    8
[3,]    3    6    0
```

Арифметичні та логічні дії виконуються з матрицями поелементно. Для того, щоб виконати матричне множення, слід застосувати операцію `%*%`. Функція `t()` транспонує матрицю.

Функція `solve(A,b)` розв'язує рівняння $Ax = b$. Якщо викликати її без другого параметра, вона обчислює обернену матрицю: значенням `solve(A)` буде A^{-1}. (Зрозуміло, що матриця A має бути невиродженою, інакше `solve(A)` повідомить про помилку).

Обернену матрицю можна також обчислити, використовуючи функцію `ginv()`, яка не входить у ядро `R`, а міститься у пакеті (бібліотеці) `MASS`. Якщо ця бібліотека не була підключена раніше, її слід підключити перед використанням `ginv()`. Точніше, `ginv()` обчислює псевдообернену матрицю Мура — Пенроуза, яка для невироджених матриць дорівнює звичайній оберненій:

```
> X<-matrix(1:6,ncol=2)
> X

     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

```
> Y<-t(X)
> Z<-Y%*%X
> Z

     [,1] [,2]
[1,]   14   32
[2,]   32   77

> library(MASS)
> iZ<-ginv(Z) # iZ матриця обернена до Z
> Z%*%iZ      # матричне множення дає одиничну матрицю:

     [,1]         [,2]
[1,]    1 1.776357e-15
[2,]    0 1.000000e+00

> Z*Z         # тут множення поелементне:

     [,1] [,2]
[1,]  196 1024
[2,] 1024 5929
```

△

При обчисленні оберненої матриці `ginv()` буде більш стабільною, ніж `solve()`: вона дає точніші результати, коли визначник матриці близький до 0. Це добре, якщо ви використовуєте функції правильно. Але, якщо ви помилитеся і параметр функції буде виродженою або не квадратною матрицею, то `solve()` повідомить вас про помилку, а `ginv()` — ні, тому що узагальнена обернена визначена і для таких матриць.

Фрейми даних відрізняються від матриць насамперед тим, що в них стовпчики можуть мати різні типи. Такий формат особливо зручний для запису типових статистичнх даних у вигляді таблиці, в якій кожному спостережуваному об'єкту відповідає один рядок, а

змінні, що характеризують об'єкти, записуються у відповідні стовпчики. При цьому кожна змінна може бути свого типу — числового, логічного, символьного чи факторного.

Приклад 2.2.1. Набір `iris` входить у пакет `R Datasets`, що являє собою колекцію даних для популярних прикладів статистичної обробки. Зазвичай цей пакет завантажується системою автоматично, якщо це не так, його можна завантажити командою `library(Datasets))`. Після завантаження до цього набору можна звертатись як до фрейма даних `iris`. У ньому міститься інформація про квіти півники (іриси). Кожен рядок цих даних відповідає одній квітці. Для кожної дослідженої квітки у відповідному стовпчику записані змінні `Sepal.Length`, `Sepal.Width` (довжина та ширина чашолистків), `Petal.Length`, `Petal.Width` (довжина та ширина пелюсток) а також змінна `Species` — вид роду Iris, до якого належить дана квітка: setosa (ірис щетинистий), versicolor (різнокольоровий) і virginica (віргінський).

Наступний приклад показує, як можна вивести на екран значення перших чотирьох змінних, що містяться у 49–51 рядках цього набору даних:

```
> print(iris[49:51,1:4])

   Sepal.Length Sepal.Width Petal.Length Petal.Width
49          5.3         3.7          1.5         0.2
50          5.0         3.3          1.4         0.2
51          7.0         3.2          4.7         1.4
```

Відповідні значення 5-ї змінної можна вивести так:

```
> print(iris[49:51,5])

[1] setosa     setosa     versicolor
Levels: setosa versicolor virginica
```

◄

Створити фрейм даних з окремих векторів-стовпчиків змінних можна, використовуючи функцію `data.frame()`:

```
> numb<-1:5
> let<-letters[numb]
> Name<-c('Alfa','Bravo','Charlie','Delta','Echo')
> type<-factor(c('vowel','consonant','consonant',
+                'consonant','vowel'))
> L<-data.frame(numb,let,type,row.names=Name,
+               stringsAsFactors = FALSE)
> print(L)

        numb let      type
Alfa       1   a     vowel
Bravo      2   b consonant
Charlie    3   c consonant
Delta      4   d consonant
Echo       5   e     vowel
```

Параметр `row.names` вказує імена об'єктів (рядків таблиці). Іменами змінних стають імена векторів стовпчиків, з яких склали фрейм. Імена рядків і стовпчиків можна продивлятись і змінювати функціями `row.names()` і `names()`.

Параметр `stringsAsFactors` показує, чи слід при створенні фрейма перетворювати вектори символьних рядків у змінні типу фактор. Стандартно таке перетворення виконується, тому, якщо вам потрібні саме символьні змінні, задавайте значення `stringsAsFactors = FALSE`.

Для того, щоб продивлятись, а за необхідності і виправляти великі фрейми даних, можна використовувати вбудований редактор R, який викликається функцією `edit()`.

З даними, що складають фрейм, можна працювати як з елементами матриці, наприклад:

```
> L[2,]

      numb let      type
Bravo    2   b consonant

> L[,2]

[1] "a" "b" "c" "d" "e"

> L[,'let']

[1] "a" "b" "c" "d" "e"
```

Окрім того, змінні є атрибутами фрейма, тому до них можна звертатись, використовуючи формат

ім'я об'єкта$*ім'я атрибута*.

Наприклад, задавши команду

```
> L$let

[1] "a" "b" "c" "d" "e"
```

отримуємо вектор значень змінної `let` для фрейма `L`.

При роботі з фреймами інколи виникає потреба перевірити, який тип у тієї чи іншої змінної. Це можна зробити, використовуючи функції перевірки типів `is.numeric`, `is.logical`, `is.integer` та подібні їм. Ці функції видають логічне значення `TRUE` якщо їх параметр має відповідний тип і `FALSE` — якщо тип не той. Функція `is.numeric()` дає результат `TRUE` для всіх векторів з числовими елементами. Функція `is.matrix()` перевіряє, чи є її аргумент матрицею.

Наприклад, тут ми використовуємо функції групи `is.` для перевірки того, що змінна `num` у фреймі `L` є числовою, а змінна `let` — факторною або символьною:

```
> is.numeric(L$num)

[1] TRUE

> is.factor(L['let'])

[1] FALSE

> is.character(L$let)

[1] TRUE
```

Зміну типу можна робити, використовуючи відповідно функції `as.numeric`, `as.character` та ін. Наприклад:

```
> x<-c('12','3')
> x[1]+x[2]

Error in x[1] + x[2]: nonnumeric argument in binary
operator

> y<-as.numeric(x)
> y[1]+y[2]

[1] 15
```

Варто мати на увазі, що R виконує автоматичне перетворення типів у елементарних операціях, наприклад:

```
> 1+TRUE

[1] 2

> TRUE&(-0.5)

[1] TRUE
```

(`TRUE` трактується як 1, `FALSE` — як 0 у арифметичних операціях. Ненульові числа трактуються як `TRUE`, а 0 — як `FALSE` у логічних операціях).

2.2.5 Векторні і матричні функції. Функція apply. Пропущені значення

Як вже зазаначалось, елементарні операції та функції виконуються над масивами поелементно. Цю властивість мають не всі функції. Якщо необхідно явно вказати, що деяка функція має застосовуватись до кожного елемента вектора, використовують функцію `sapply()` або `lapply()`[5]. Першим параметром `sapply()` має бути вектор `x`, до якого застосовується функція, другим — функція `FUN`, яка застосовується до кожного елемента `x`. Точніше, елементи `x` підставляються у `FUN` замість першого її параметра. Якщо у `FUN` є іще параметри, їх можна вказати як додаткові параметри-опції `sapply`, і вони будуть передані у `FUN` за їхніми назвами. Наприклад, тут ми застосовуємо двійковий логарифм до вектора ступенів двійки:

```
> sapply(c(1,2,4,8),log,base=2)

[1] 0 1 2 3
```

(У цьому прикладі застосування `sapply()` не обов'язкове, такий самий ефект буде при виклику `log(c(1,2,4,8),base=2)`).

Функція `lapply()` аналогічна `sapply()`, але її можна застосовувати до списків з довільних елементів.

Інколи доводиться один і той самий вираз обчислювати багато разів. Це здається трохи дивним. Чому б не зробити обчислення один раз, а потім розмножити результат функцією `rep()`? Але справа в тому, що значення виразу може змінюватись на кожному кроці обчислень. Для цього в нього мають входити глобальні змінні, які змінюються шляхом глобального привласнення[6]. Щоб виконати такі обчислення можна використати функцію `replicate()`.

[5] Інша можливість змусити функцію обчислюватись поелементно при застосуванні до масиву — векторизація, див. п. 2.7.2.

[6] Див. п. 2.7.1.

Перший параметр цієї функції, `n` — кількість повторень, другий, `expr` — вираз, який обчислюється `n` разів. Результат обчислень — вектор з `n` послідовних значень `expr`:

```
> I<-1
> # глобальне I змінюється глобальним привласненням:
> replicate(5,I<<-I+1)

[1] 2 3 4 5 6

> # а тут змінюється лише локальне I
> # всередині виклику функції:
> replicate(5,I<-I+1)

[1] 7 7 7 7 7
```

Часто зустрічаються функції, які працюють з вектором "в цілому". Наприклад:

`length()` — функція, що повертає кількість елементів масиву;

`max()`, `min()`, `sum()`, `prod()` — функції, що знаходять відповідно найбільше, найменше значення, суму або добуток всіх елементів масиву[7];

`sort()` — функція, що переставляє (відсортовує) елементи масиву у порядку зростання (або спадання, якщо вказана опція `decreasing=T`).

[7] Ці функції (а також функція `range()`) мають цікаву особливість, що відрізняє їх від інших функцій `R`, призначених для підрахунку узагальнених характеристик даних (таких як `mean()`): ці функції об'єднують всі свої аргументи в один набір і далі працюють із цим набором. Наприклад, результатм виконання `sum(1,2,3,4)` буде 10. Якщо спробувати аналогічно обчислити середнє значення — `mean(1,2,3,4)`, отримаємо хибний результат 1: `mean()` ігнорує всі перелічені через кому аргументи, окрім першого. Ця властивість часом виявляється незручною, коли, наприклад, є два числових вектори $(x_1, \ldots, x_n)$ і $(y_1, \ldots, y_n)$ і треба знайти вектор, складений з попарних мінімумів їхніх елементів: $(\min(x_1, y_1), \ldots, \min(x_n, y_n))$. У цьому випадку можна скористатись функцією `pmin()` (`pmax()` для максимумів).

Якщо аргументом такої функції є матриці, то матриця трактується як "довгий" вектор, складений з усіх її елементів:

```
> u1<-c(3,1,2)
> u2<-c(0,-1,-2)
> z<-cbind(u1,u2)
> z

     u1 u2
[1,]  3  0
[2,]  1 -1
[3,]  2 -2

> sum(z)

[1] 3

> sort(z)

[1] -2 -1  0  1  2  3

> sort(z,decreasing=T)

[1]  3  2  1  0 -1 -2
```

Часто доводиться застосувати функцію від векторного елемента до кожного рядка або до кожного стовпчика матриці окремо. У такому випадку використовується функція `apply()`, що має специфікацію

`apply(X, MARGIN, FUN, ...)`,

де `X` — масив, до якого буде застосовуватись функція; `MARGIN=1`, якщо функція застосовується до рядків і `MARGIN=2`, — якщо до стовпчиків; `FUN` — ім'я функції, яку треба застосувати.

Символ `...` позначає, що у виклику функції `apply()` можна також задавати будь-які інші опції. Ці опції `apply()` передасть у функцію `FUN` без змін.

Першим параметром функції `FUN` має бути вектор. Замість цього вектора `apply()` підставляє послідовно рядки (або стовпчики) матриці `X` і результат також записує у список результатів. Наприклад, використовуючи матрицю `z` з попереднього прикладу, отримуємо

```
> apply(z,1,sum)

[1] 3 0 0

> apply(z,2,sum)

u1 u2
 6 -3

> apply(z,2,sort)

     u1 u2
[1,]  1 -2
[2,]  2 -1
[3,]  3  0
```

У третьому прикладі елементи матриці `z` відсортувались окремо всередині кожного стовпчика. А для того, щоб отримати матрицю з елементами, відсортованими всередині кожного рядка, результат роботи `apply` треба транспонувати:

```
> apply(z,1,sort)

   [,1] [,2] [,3]
u2    0   -1   -2
u1    3    1    2

> t(apply(z,1,sort))
```

```
     u2 u1
[1,]  0  3
[2,] -1  1
[3,] -2  2
```

Приклад передачі опції при виклику `sort()` через `apply()`:

```
> apply(z,2,sort,decreasing=T)

     u1 u2
[1,]  3  0
[2,]  2 -1
[3,]  1 -2
```

Ще один варіант поелементної обробки векторів реалізує функція `outer()`. Перші два її параметри `x, y` є векторами, третій `FUN` — функцією, у якої не менше двох параметрів. Замість цих параметрів `outer()` підставляє послідовно всі можливі пари елементів `x` і `y` (`x` замість першого параметра, `y` — замість другого). Отримані значення утворюють матрицю. Наприклад,

```
> x<-1:4
> y<-5:7
> f<-function(x,y){x^2+y^2}
> z<-outer(x,y,f)
> z

     [,1] [,2] [,3]
[1,]   26   37   50
[2,]   29   40   53
[3,]   34   45   58
[4,]   41   52   65
```

Тут ми створили нову функцію `f`, яка обчислює суму квадратів двох своїх параметрів[8] і застосували її до всіх пар елементів векторів

[8] Про створення власних функцій див. далі, у п. 2.7.1.

`x` і `y`. Як і функція `apply()`, `outer()` вміє передавати додаткові параметри всередину функції `FUN`.

Такі функції як `sum()` та `prod()` можуть обробляти пропущені значення (NA) по-різному. Можна вважати, що коли якесь значення у векторі невідоме, то і сума невідома. А можна вилучити всі пропущені значення і обчислити суму не пропущених. Вибір реалізується за допомогою опції `na.rm` ("NA remove" — видалення пропущених):

```
> x<-c(2,NA,1,4,3)
> sum(x)

[1] NA

> sum(x,na.rm=T)

[1] 10
```

При стандартному застосуванні функції `sort()` значення NA видаляються. Можна скористатись опцією `na.last=T`, щоб при сортуванні значення NA потрапляли у кінець вектора. При цьому значення NA та NaN обробляються однаково:

```
> x<-c(2,NA,1,NaN,4,NA,3,NaN)
> sort(x,na.last=T)

[1]   1   2   3   4  NA NaN  NA NaN
```

2.3 Деякі корисні функції

У цьому підрозділі розглянемо ряд простих функцій, що працюють з векторами, матрицями та фреймами даних. Значну частину того, що роблять ці функції неважко запрограмувати самому, але краще використовувати стандартні готові засоби: це робить програму зрозумілішою, а часто — й ефективнішою.

2.3.1 Функції покрокового обчислення

У статистиці часто виникає потреба обчислювати "накопичені суми/добутки" послідовних значень спостережуваних величин. Якщо задана послідовність x_1, x_2,..., x_n, то послідовністю її накопичених сум буде

$$S_1 = x_1,\ S_2 = x_1 + x_2, \ldots, S_n = x_1 + x_2 + \cdots + x_n.$$

Накопичені добутки визначаються як

$$P_1 = x_1, \ldots, P_2 = x_1 \cdot x_2 \ldots, P_n = x_1 \cdot x_2 \cdot \cdots \cdot x_n.$$

Для обчислення накопичених (кумулятивних) сум можна використати функцію `cumsum()`, а накопичених добутків — функцію `cumprod()`.

Аналогічно вводяться послідовності накопичених максимумів і мінімумів, які обчислюються функціями `cummax()` і `cummin()`:

```
> x<-c(1,-2,3,-4,5,-6,7,-8)
> names(x)<-c("a","b","c","d","e","f","g","h")
> cumsum(x)

 a  b  c  d  e  f  g  h
 1 -1  2 -2  3 -3  4 -4

> cumprod(x)

    a     b     c     d     e     f     g     h
    1    -2    -6    24   120  -720 -5040 40320

> cummax(x)

a b c d e f g h
1 1 3 3 5 5 7 7

> cummin(x)
```

```
 a  b  c  d  e  f  g  h
 1 -2 -2 -4 -4 -6 -6 -8
```

(Зверніть увагу, що всі ці функції зберігають імена елементів послідовності).

Якщо аргумент `cumsum()` є матрицею, то її стовпчики об'єднуються в один довгий вектор, для якого і обчислюються накопичені суми:

```
> X<-cbind(1:5,6:10)
> cumsum(X)

 [1]  1  3  6 10 15 21 28 36 45 55
```

Якщо треба підсумовувати окремо по кожному стовпцю (або по кожному рядку), це можна зробити, використовуючи функцію `apply()`:

```
> apply(X,2,cumsum)

     [,1] [,2]
[1,]    1    6
[2,]    3   13
[3,]    6   21
[4,]   10   30
[5,]   15   40

> apply(X,1,cumsum)

     [,1] [,2] [,3] [,4] [,5]
[1,]    1    2    3    4    5
[2,]    7    9   11   13   15
```

Дія, обернена до обчислення кумулятивних сум — обчислення приростів (скінченних різниць). Її можна виконати, використовуючи функцію `diff()`.

Скінченними приростами першого порядку з лагом k для послідовності x_i називають послідовність

$$D_1 = x_{1+k} - x_1,\ D_2 = x_{2+k} - x_2, \ldots, D_{n-k} = x_n - x_{n-k}.$$

Наприклад,

```
> x<-(1:8)^2
> names(x)<-c("a","b","c","d","e","f","g","h")
> diff(x)

 b  c  d  e  f  g  h
 3  5  7  9 11 13 15

> diff(x,lag=2)

 c  d  e  f  g  h
 8 12 16 20 24 28
```

Зазначимо, що скінченні різниці першого порядку з лагом 1 для масиву `x` можна отримати так:

`x[-1]-x[-length(x)]`.

Різниці (прирости) порядку d можна отримати повторивши d-кратно операцію взяття різниць першого порядку. Але у функції `diff()` для цього використана більш ефективна процедура:

```
> diff(x,differences = 2)

c d e f g h
2 2 2 2 2 2

> diff(diff(x))

c d e f g h
2 2 2 2 2 2
```

2.3.2 Впорядкування та пошук у масивах

Ми вже розглядали у п. 2.2.5 функцію `sort()`, яка переставляє елементи масивів у порядку зростання (спадання). Інколи буває корисно переставляти не сам масив, а послідовність номерів його елементів так, щоб отримана перестановка відповідала порядку зростання у масиві. Це робить функція `order()`. Команда `ind<-order(x)` створює масив номерів `ind` такий, що `ind[1]` це номер найменшого елемента у масиві `x`, `ind[2]` — номер наступного за порядком елемента, і.т.д.:

```
> x<-c(6,1,5,2,4,3)
> order(x)

[1] 2 4 6 5 3 1
```

(Вказавши опцію `decreasing=T` отримуємо перестановку за спаданням).

Функцію verb|order()| зручно використовувати, наприклад, тоді, коли треба переставити рядки матриці (фрейма даних) у порядку зростання елементів з певного її стовпця. У наступному прикладі — перестановка за другим стовпцем:

```
> first<-c("a","b","c")
> second<-c(3,1,2)
> y<-data.frame(first,second)
> y

  first second
1     a      3
2     b      1
3     c      2

> y[order(y$second),]
```

```
  first second
2     b      1
3     c      2
1     a      3
```

Для символьних рядків функція `order()` використовує лексикографічний порядок, тому її можна використовувати для сортування в алфавітному порядку[9].

Якщо вам потрібен номер найбільшого елемента масиву `x`, його можна отримати як `order(x)[length(x)]`. Але насправді не треба впорядковувати всі елементи масиву для того, щоб знайти найбільший. Краще скористатись стандартною функцією `which.max()`. Відповідно, для знаходження номера найменшого елемента можна використовувати функцію `which.min()`:

```
> which.min(c(2,-3,4,-2))

[1] 2
```

Якщо мінімум у масиві досягається для кількох елементів, результатом `which.min()` буде найменший з індексів, що відповідають мінімуму. (Аналогічно для `which.max()`). Для логічних значень ці функції використовують порядок `FALSE`<`TRUE`. Тому їх можна застосовувати для знаходження положення першого елемента масиву, для якого виконується деяка умова:

```
> x<-seq(0,pi,0.001)
> x0<-which.max(sin(x)>0.5)
> x0

[1] 525
```

[9] Точніше, у порядку латинського алфавіту. З кирилічними літерами будуть проблеми.

```
> x[x0]

[1] 0.524

> sin(x[x0])

[1] 0.5003474
```

— вперше $\sin(x) > 1/2$ на 525-му елементі масиву `x`.

Якщо потрібний перелік індексів всіх елементів масиву, для яких виконується певна умова, можна скористатись функцією `which()`:

```
> x<-seq(0,6*pi,0.5)
> which(sin(x)>0.95)

[1]  4 17 29
```

— серед елементів масиву `x` лише для 4-го, 17-го і 29-го виконується умова[10] $\sin(x) > 0.95$.

2.3.3 Злиття наборів даних — merge

У статистиці часто виникає потреба об'єднувати інформацію з різних джерел. Якщо у двох фреймах даних міститься різна інформація про одні і ті самі об'єкти, її можна перенести в один фрейм, використовуючи функцію `merge()` (злиття фреймів).

У наступному прикладі маємо два фрейми `weights` — з даними про вагу людей і `heights` — з даними про їхній зріст. У кожному з цих фреймів перша змінна — `name` містить ім'я людини, а друга (`weight` або `height`) — її вагу або зріст. Ми хочемо

[10] Згідно з стандартом, в Україні для відділення дробової частини десяткових дробів використовується кома. У `R` дробова частина відділяється крапкою. Тому, для зручності читачів у цій книзі з цією метою всюди використана крапка. Автор перепрошує за відхилення від стандарту.

отримати фрейм, у якому кожен рядок відповідатиме одній людині, а змінними будуть `name`, `weight` і `height`. От як це робиться:

```
> # створюємо два фрейми:
> weights<-data.frame(name=c("John","Ivan","Kate"),
+                     weight=c(80,75,48))
> heights<-data.frame(name=c("Kate","Ivan","Mary"),
+                     height=c(160,180,182))
> # робимо злиття фреймів:
> merge(weights,heights,all=T)

  name weight height
1 Ivan     75    180
2 John     80     NA
3 Kate     48    160
4 Mary     NA    182
```

Як бачимо, набір людей у першому і другому файлах дещо різний: для `John` ми знаємо `weight`, але не знаємо `height`. Для `Mary` — навпаки. У файл, що був утворений злиттям, ввійшли всі люди і з першого, і другого фрейма, а їхні змінні, по яких немає інформації, отримали значення `NA`.

Такий спосіб злиття визначений опцією `all=T`. Якщо `all=F`, то у фрейм-результат будуть внесені тільки люди, які описані і у першому, і другому фреймі[11].

Змінна, що використовується для узгодження рядків (у нашом прикладі це`name`) називають *ключем* (key). Таких змінних може бути декілька (напр., ім'я, прізвище та рік народження). Стандартно `merge` використовує як ключі всі змінні, які мають однакові імена у першому та другому фреймі. При бажанні список ключів можна задати опцією `by`, якщо вони мають однакові імена в обох

[11] Можна також використати опцію `all.x=T`, якщо треба включити у результат всіх людей, перелічених у першому фреймі, а тих, хто згаданий лише у другому, — пропустити. І навпаки — `all.y=T`.

фреймах, або опціями `by.x` (ключі першого фрейма) і `by.y` (ключі другого фрейма). В останньому випадку першій змінній зі списку `by.x` буде відповідати перша у `by.y`, другій — друга і т.д.

Опція `sort` вказує, чи сортувати отриманий злиттям фрейм у порядку зростання ключів. Стандартне значення `sort=T`, тобто сортування виконується.

2.4 Експорт та імпорт даних у R

2.4.1 Експорт та імпорт даних у внутрішньому форматі

Інколи буває треба зберігти деякі результати роботи програми у форматі R, наприклад, для використання їх іншим користувачем R у своїй програмі. Для цього можна скористатись функцією `save()`:

```
> a<-1:10
> save(a,file="c:/rem/term/example.Rdata")
```

У цьому прикладі ми створили вектор `a`, а потім записали його у файлі `example.Rdata` у каталозі `rem/term` на диску `c`. (Зверніть увагу, що при записі шляху до файла використовується символ /, прийнятий в Unix, а не \, як це прийнято у Windows).

Функція `save()` зберігає об'єкти у внутрішньому кодуванні системи R. Прочитати записаний файл можна тільки в R. Якщо продивлятись його в якому-небудь текстовому редакторі, будуть відображатись лише незрозумілі символи. Для читання можна використати функцію `load()`:

```
> a<-0
> a

[1] 0

> load(file="c:/rem/term/example.Rdata")
> a
```

```
 [1]  1  2  3  4  5  6  7  8  9 10
```

(ми спочатку надали нового значення a, а потім відновили старе, прочитавши його з файла). Об'єкти записуються разом зі своїми іменами, тому load() розуміє без додаткових пояснень, що саме слід змінити.

Якщо в одному файлі треба зберігти багато об'єктів, їх перелічують через кому у списку параметрів save().

Якщо файл для запису або читання треба вибрати під час роботи програми інтерактивно, використовують функцією file.choose(), яка відкриває стандартне вікно вибору файла. Скажімо, для завантаження з файла, який ви хочете обрати вручну, можна написати:

```
load(file=file.choose()).
```

2.4.2 Експорт та імпорт текстових таблиць

Більшість статистичних програм загального призначення має можливості створення та читання файлів даних у вигляді текстових таблиць. Зміст таких файлів легко зрозуміти, проглядаючи їх у звичайних текстових редакторах. Тому природно використовувати такі файли для обміну статистичними даними між програмами.

Нехай таблиця записана в текстовому файлі у зручному для людського сприйняття вигляді:

```
  Name       Weight Married
Ahmad        70     T
John         82     F
Victoria     60     T
Olga         54     F
```

Тут у першому рядку записані назви змінних, а у кожному наступному рядку — значення цих змінних для певної людини. Для читання таких таблиць використовують функцію read.table(). Якщо таблиця записана у файлі

```
c:/rem/term/table.txt,
```

то прочитати її можна так:

```
> tbl<-read.table(file="c:/rem/term/table.txt",
+                 header=T)
> tbl

      Name Weight Married
1    Ahmad     70    TRUE
2     John     82   FALSE
3 Victoria     60    TRUE
4     Olga     54   FALSE
```

(Результат читання записано у фрейм `tbl`. Опція `header=T` вказує на те, що у першому рядку містяться назви змінних).

Якщо один зі стовпчиків таблиці треба прочитати як імена об'єктів-рядків, це можна зробити, задавши опцію `row.names`. У ній можна вказати або номер стовпчика імен, або його назву, наприклад:

```
>  tbl<-read.table(file="c:/rem/term/table.txt",
+                  header=T,row.names="Name")
>  tbl

         Weight Married
Ahmad        70    TRUE
John         82   FALSE
Victoria     60    TRUE
Olga         54   FALSE
```

Записати таку таблицю на диск можна за допомогою функції `write.table()`:

```
write.table(tbl,file="c:/rem/term/table.txt")
```

Стандартно, у першому стовпчику таблиці будуть записані імена об'єктів-рядків фрейма, а у першому рядку — назви змінних-стовпчиків фрейма. Якщо це не потрібно, слід вказати опції `row.names=F`, `col.names=F`.

При читанні з файла `read.table()` визначає кількість змінних (стовпчиків) у таблиці за кількістю назв у першому рядку. Тип змінної визначається за форматом запису елементів у відповідному стовпчику. Скажімо, якщо всі елементи стовпчика мають вигляд `TRUE`, `FALSE` або `NA`, то відповідна змінна отримає у прочитаному фреймі даних тип `logical`. Якщо хоча б один елемент не можна трактувати як логічний, тип буде `character`, навіть, якщо всі інші елементи можна було б інтерпретувати як логічні.

Різна кількість елементів, розділених пробілами у різних рядках таблиці, призводить до помилки читання. Якщо у файлі зустрічаються символьні рядки з пробілами всередині, ці рядки треба вміщувати у лапки, як у наступному прикладі:

```
Name        Weight Married
Ahmad         70     T
"John R.C."   82     F
Victoria      60     T
"Olga V."     54     F
```

Інший текстовий формат — csv (скор. від comma separated values) в якому окремі значення змінних розділяються комами. Цей формат менш зручний для людського сприйняття, ніж табличний, але він дає більше можливостей для передачі даних різних типів.

Щоб записати (або прочитати) файл у форматі csv, можна використовувати функції `write.csv()` (`read.csv`). В основному вони влаштовані аналогічно функціям `write.table()` та `read.table()`. (По суті, відмінність між функціями, що обробляють table та csv, полягає лише в іншому виборі стандартних значень тих опцій, які регулюють вибір символів, що розділяють значення. Вибирати ці опції (вони описані у help) можна самому, якщо необхідно створити або прочитати файл з нестандартного формату.

У форматі `csv` числа розділяються комами, а крапка викори-

стана для відокремлення дробової частини десяткового числа. У форматі csv2 дробова частина відділяється комою, а як розділовий знак між числами використана крапка з комою — ";". Цей формат зручний для передачі таблиць з EXCEL, оскільки у EXCEL є можливість збереження/читання файлів у форматі csv2, а в R такі таблиці, як ви вже здогадались, читаються/записуються функціями read.csv2() і write.csv2().

Параметр file у функції read.table та подібних їй не обов'язково має бути іменем файла на вашому комп'ютері. Наприклад, вказавши опцію file="clipboard", можна прочитати таблицю, яка розміщена у буфері обміну Windows (або іншої операційної системи, яка працює на комп'ютері). Зрозуміло, що цю таблицю спочатку треба помістити у буфер обміну, виконавши "copy".

Наприклад, якщо ви відкриєте Excel-файл, виділите у ньому числову таблицю і виконаєте функцію copy, ця таблиця опиниться у буфері обміну. Тепер її можна прочитати і зберігти у змінній x в R, виконавши команду

```
x<-read.table(file="clipboard",dec=",")
```

(Опція dec="," вказує, що для відділення дробової частини числа використано кому).

Якщо вказати у file інтернет-адресу ресурсу (URL), R прочитає таблицю в інтернеті, що розміщена за цією адресою. (Звичайно, таблиця має бути у відповідному текстовому форматі. HTML-таблицю так не прочитати.) Наприклад, якщо виконати команди

```
fl<-"http://datasets.flowingdata.com/
     crimeRatesByState2005.tsv"
crime <- read.csv(fl, header=TRUE, sep="\t")
```

то R прочитає у фрейм crime дані кримінальної статистики 2005 року по штатах США, які розміщені на навчальному сайті Натана Яу Flowingdata у файлі

crimeRatesByState2005.tsv[12].

[12]Цей файл був доступний в інтернеті у серпні 2017 р.

2.5 Переформатування фреймів даних

При роботі з фреймами даних кожен рядок відповідає одному спостереженню (subject, case — це може бути спостережуваний об'єкт або день, коли проводились спостереження тощо). Кожен стовпчик трактують як змінну, певну характеристику, що описує спостережуваний об'єкт. Саме за цією схемою працює більшість функцій статистичного аналізу даних. Але визначення того, що вважати об'єктом, а що змінною, залежить від точки зору. При зміні точки зору те, що було об'єктом, може стати характеристикою об'єкта, і навпаки. При цьому виникає потреба переформатувати відповідний фрейм даних, щоб його було зручно обробляти стандартним способом. Для цього можна скористатись функціями бібліотеки `tidyr`. Ми розглянемо приклад застосування двох таких функцій.

Приклад 2.5.1. У файлі `potato1.txt` містяться дані про ціни на картоплю на різних ринках міста Києва[13]. Кожному ринку відповідає один рядок. Перший стовпчик — назва ринку, наступні містять ціну 1 кг картоплі різних типів на даному ринку. Розглядаються чотири типи картоплі — молода біла (new.white), молода рожева (new.pink), стара біла (old.white) і стара рожева. Кожному типу відповідає свій стовпчик-змінна. Завантажимо цей файл у фрейм `potato` і надрукуємо:

```
> potato<-read.table(file="c:/rem/term/potato1.txt",
+                    header=T)
> print(potato)

     Market new.white new.pink old.white old.pink
1    Odessa      15.5     16.5       4.5      5.0
2 Goloseevo      17.5     18.0       5.0      5.0
3   Central      20.0     22.0       6.0      6.5
```

[13] Дані умовні. Скажімо, можна уявити, що вони відповідають середнім цінам кінця травня 2016 р. у гривнях.

Така структура фрейма доцільна, якщо дослідник хоче описати зв'язок між цінами на різні типи картоплі. Наприклад, за ним зручно провести регресійний аналіз залежності ціни на молоду рожеву картоплю від цін на стару білу. Але можуть бути й інші варіанти статистичних досліджень цих даних. Скажімо, можна вивчати питання про те, як на ціну картоплі впливають такі фактори, як її тип і ринок, на якому вона продається. Для такого дослідження доцільно окремі рядки присвятити кожному набору обставин, що визначають ціну картоплі. Самі ціни у такому файлі мають вміщуватись в один стовпчик-змінну, а інші стовпчики повинні вказувати на обставини, що формували ціну: місце продажу та тип картоплі. Ціна картоплі у цьому випадку — те, що називають залежною (ключовою) змінною, а всі інші змінні — фактори, що впливають на залежну змінну. Це називають "перетворенням фрейма з широкого формату на довгий". Застосуємо для цього функцію

```
gather(data,key,value,...)
```

з бібліотеки `tidyr`. Першим параметром (`data`) цієї функції є фрейм, який треба переформатувати. Параметр `key` вказує нову назву стовпчика у переформатованому фреймі, у якому буде створений фактор з назв змінних у старому фреймі, `value` — назва нової залежної змінної, у якій зберуться значення об'єднаних старих змінних, замість `...` треба перелічити назви стовпчиків у старому фреймі, з яких буде формуватись нова залежна змінна.

Для фрейма `potato` відповідний скрипт має вигляд:

```
> library(tidyr)
> pt<-gather(potato,type,price,new.white:old.pink)
> print(pt)

      Market      type price
1     Odessa new.white  15.5
2  Goloseevo new.white  17.5
3    Central new.white  20.0
4     Odessa  new.pink  16.5
```

```
5  Goloseevo  new.pink  18.0
6    Central  new.pink  22.0
7     Odessa old.white   4.5
8  Goloseevo old.white   5.0
9    Central old.white   6.0
10    Odessa  old.pink   5.0
11 Goloseevo  old.pink   5.0
12   Central  old.pink   6.5
```

(двокрапка у виклику `gather()` між `new.white` і `old.pink` позначає, що треба взяти ці стовпчики і всі, що лежать між ними).

У результаті нашого перетворення назви чотирьох змінних від `new.white` до `old.pink` перетворились на значення змінної `type`, котра вказує тепер, з якого старого стовпчика була отримана відповідна ціна у новому стовпчику `price`. Значення змінної `Market` також розмножились, щоб відповідати кожній окремій ціні.

Але у змінній `type` зараз змішані дві різні характеристики картоплі: вік (`age`: молода—стара) та колір (`color`: біла—рожева). Щоб досліджувати вплив кожної з цих характеристик окремо, варто створити для них дві окремі змінні.

Для цього призначена функція

`separate(data,col,into,sep)`.

У цій функції параметр `data` позначає фрейм, що переформатовується, `col` — назва стовпчика-змінної, яка розбивається на кілька нових, `into` — перелік назв нових змінних, `sep` — символ, що розділяє у старій змінній назви нових факторів. (Параметр `sep` може також бути вектором, що складається з номерів позицій початків назв нових факторів у символьному рядку-назві старого фактора).

От як це можна зробити для нашого фрейма `pt`:

```
>   library(dplyr)
> pt2<-separate(pt,type,
+               into = c("age","color"),sep="\\.")
> print(slice(pt2,3:5))
```

```
# A tibble: 3 x 4
     Market   age color price
     <fctr> <chr> <chr> <dbl>
1   Central   new white  20.0
2    Odessa   new  pink  16.5
3 Goloseevo   new  pink  18.0
```

△

(ми поклали переформатований фрейм у змінну `pt2` і вивели лише три рядки цього фрейма. Функція `slice()` виділяє обрані рядки — з 3-го по 5-й з фрейма `pt2`. Вона визначена у бібліотеці `dplyr`. По суті, її дія еквівалентна `pt2[2:3,]`. Можливості її використання і робота бібліотеки `dplyr` пояснюються у підрозд. 2.6) ◄

2.6 Підстановки (pipes)

Розглянемо мовний засіб `R`, який називають pipe. Він використовує оператор `%>%`. Це, фактично, особливий спосіб запису виклику функцій, який можна назвати *підстановкою*. Для використання підстановок потрібно завантажити пакет `dplyr`.

Щоб зрозуміти, для чого використовуються підстановки, повернемось до розгляду прикладу 2.5.1. Там ми робили перетворення фрейма даних `potato`, розміщуючи проміжні результати у фрейми `pt` і `pt2`. Це було зроблено для наочності викладу. Взагалі, якщо проміжні результати не потрібні після закінчення перетворень, їх краще залишати безіменними. Це, по-перше, прискорить роботу комп'ютера: йому не треба буде вносити імена змінних у таблиці імен і робити відповідні привласнення. По друге, завдяки цьому людина, що розбирає роботу програми, не буде відволікатись на з'ясування призначення змінних, які миттєво втрачають зміст у процесі перетворень.

Якщо не вживати імен для поміжних змінних, а одразу підставляти результати виконання попереднього перетворення у функцію, що виконує наступну дію, скрипт прикладу 2.5.1 можна оформити так:

```
library(tidyr)
library(dplyr)
print(
  slice(
    separate(
      gather(
        read.table(file=
        "c:/rem/term/potato1.txt",header=T),
      type,price,new.white:old.pink),
    type,into = c("age","color"),sep="\\."),
  3:5)
)
```

(Переконайтесь, що ця програма дійсно видає такий самий остаточний результат, як і у прикладі 2.5.1).

Такий запис має очевидні недоліки. Назви функцій у програмі розташовані у порядку, оберненому порядку їх виконання — робота починається з `read.table`, але щоб це побачити, треба прослідкувати ланцюжок викликів від початку до кінця. Математики, які звикли мати справу з формулами, легко з цим справляються. Спеціалістам у прикладних галузях, які роблять статистичну обробку своїх даних, це незручно, оскільки суперечить їх інтуїції.

Окрім того, при такому записі важко зрозуміти, до виклику якої функції належать параметри й опції, записані після підстановки чергового проміжного результату. Скажімо, `type` і `price` — це параметри `gather`, `separate`, чи `slice`? Для точної відповіді на це питання треба акуратно підрахувати дужки, що відкриваються та закриваються. Для полегшення сприйняття таких ланцюгових викликів різних перетворень у бібліотеці `dplyr` створена структура, яку я називатиму *підстановка*[14].

[14] англійська назва pipe (pipeline) —"труба" або "трубопровід", на мою думку, не передає змісту цієї структури.

Оператор підстановки має вигляд %>%. Ліворуч від нього записується те, що треба підставити, а праворуч — функція, у яку воно підставляється як перший параметр. Всі інші параметри функції записуються, як при звичайному виклику:

```
> library(tidyr)
> library(dplyr)
> c(1:4,NA)%>%sum(na.rm=T)

[1] 10
```

(Ми підставили вектор (1,2,3,4,NA) у функцію sum(), причому вказали опцію вилучення пропущених значень. Отримали суму всіх непропущених. Фактично це те саме, що виклик sum(c(1:4,NA),na.rm=T).

От як можна записати перетворення прикладу 2.5.1, використовуючи оператор підстановки:

```
read.table(file="c:/rem/term/potato1.txt",header=T)%>%
  gather(type,price,new.white:old.pink)%>%
   separate(type,into = c("age","color"),sep="\\.")%>%
     slice(3:5)%>%
        print
```

Тепер функції записані у порядку їх виконання: програма послідовно читає фрейм, переформатовує його у довгий формат, розділяє фактор type, вирізає частину фрейма і друкує її. Додаткові параметри кожної функції записані у виклику одразу після її імені. Результат попереднього перетворення завжди підставляється у наступну функцію як перший параметр.

Запис викликів функцій з підстановками доцільний при виконанні таких ланцюгів перетворень з базами даних. Там, де програми виконують дії математичного характеру, мабуть, краще використовувати класичний запис. Зрозуміло, що це не є загальним

правилом, і програміст вільний у виборі мовного оформлення своєї програми. Але бажано не застосовувати ці два стилі упереміш всередині коротких логічних фрагментів програми.

При застосуванні підстановок техніка індексації даних за допомогою прямих дужок з п. 2.2.2 стає незручною. Тому для виділення певних частин даних для дослідження краще користуватись спеціальними функціями. Одна з них — `silce()` з бібліотеки `dplyr`, вже була використана раніше. Вона дозволяє вирізати потрібні рядки з фрейма даних за переліком їхніх номерів. Якщо треба вирізати певні стовпчики (змінні), можна скористатись функцією `select()`. Для того, щоб вирізати частину даних, використовуючи логічні умови, застосовують функцію `subset()`. Більш докладно про це можна подивитись у help `R`.

2.7 Програмування у R

2.7.1 Створення власних функцій

Функції у R є об'єктами, тому для того, щоб ввести нову функцію, треба створити об'єкт типу `function` і привласнити його значення деякій змінній. Наприклад, тут ми у першому рядку створюємо функцію `t.sum()`, а у наступному викликаємо її з параметрами `x=1:10` та `t=8`:

```
> t.sum<-function(x,t){sum(x[x>t])}
> t.sum(1:10,8)

[1] 19
```

Призначення цієї функції зрозуміле — вона розраховує суму тих елементів вектора-параметра `x`, які перевищують поріг, заданий параметром `t`.

В загальному вигляді команда створення нової функції (специфікація) має формат

`function(`*список формальних параметрів* `){`*тіло функції*`}`.

Тіло функції — це послідовність команд, які будуть виконані при виклику функції. Результат останньої виконаної у тілі функції команди є *значенням функції*. Це значення і буде результатом виразу виклику функції.

При виклику функції фактичні значення параметрів, задані у дужках після імені функції, підставляються замість формальних параметрів, вказаних у специфікації функції. Можна використовувати неіменований спосіб підстановки, коли формальні параметри заміняються фактичними в порядку їхнього переліку у специфікації. Так зроблено у попередньому прикладі. Можна застосувати іменовану підстановку, вказуючи ім'я формального параметра, який треба замінити при виклику:

```
> t.sum(t=9,x=1:10)

[1] 10
```

Виконання такого виклику нічим не відрізняється від попереднього. Можна комбінувати ці два способи:

```
> t.sum(1:10,t=9)

[1] 10
```

R дозволяє задавати при виклику функції менше параметрів, ніж вказано у специфікації. При цьому функція мусить знати, яких значень ці параметри набувають стандартно (англ. by default, тобто коли вони не вказані). У специфікації такі стандартні значення можна вказати, використовуючи знак = після імені формального параметра:

```
> t.sum<-function(x,t=0){sum(x[x>t])}
> t.sum(-5:5)
```

```
[1] 15

> t.sum(1:10,t=9)

[1] 10
```

Тут стандартно функція `t.sum` розраховує суму додатних елементів вектора, але поріг можна змінити, задавши значення параметра `t` явно. Параметри, для яких, зазвичай, використовуються стандартні значення, прийнято називати *опціями функції*. Відмінність між опціями та іншіими параметрами функції тільки у способі їх використання. Для комп'ютера опції нічим не відрізняються від інших параметрів.

При заданні стандартних значень опцій можна використовувати вирази, в які входять інші формальні параметри функції:

```
> t.sum<-function(x,t=sum(x)/length(x)){sum(x[x>t])}
> t.sum(1:10)

[1] 40
```

Тут стандартно функція обчислює суму всіх елементів `x`, які перевищують середнє `x`.

У тілі функції може бути багато команд (їх можна записувати в окремих рядках програми або розділяти крапкою з комою). Можна також використовувати змінні, що не входять до списку формальних параметрів. Зокрема, функція `t.sum()` з попередніх прикладів може бути реалізована таким чином:

```
> t.sum<-function(x,t=0)
+         {
+         z<-x>t
+         sum(z)
+         }
```

Тут z — допоміжна змінна, що використовується у функції. За правилами R всі такі змінні є *локальними*, тобто вони існують лише всередині функції і знищуються при завершенні виконання її виклику. Якщо поза тілом функції було введено змінну з тим самим іменем (*глобальну змінну*), її значення не зміниться після виклику функції:

```
> z<-0
> t.sum(1:10,t=8)

[1] 2

> z

[1] 0
```

(Значення глобального z залишилось 0, незважаючи на виклик функції, в якій локальній змінній z було зроблене привласнення).

Локальність дозволяє усунути можливість небажаних побічних ефектів (side effect), коли функція змінює значення змінних, що не мають стосунку до її виклику. Інколи буває необхідно, щоб функція мала побічний ефект, впливаючи на певну глобальну змінну. Для цього можна використати *глобальне привласнення* <<-. При використанні <<- всередині опису функції, комп'ютер шукає змінну, якій треба привласнити значення, у зовнішньому оточенні цієї функції.

Нехай, наприклад, треба підрахувати, скільки разів відбувався виклик функції t.sum() у програмі. Створимо для цього глобальну змінну n і модифікуємо функцію так:

```
> n<-0
> t.sum<-function(x,t=0){n<<-n+1;sum(x>t)}
> t.sum(1:10)

[1] 10
```

```
> n
[1] 1
> t.sum(1:10)
[1] 10
> n
[1] 2
```

Тут змінна `n` збільшується на 1 при кожному виклику функції `t.sum`.

Глобальні привласнення рекомендується використовувати дуже обережно, оскільки вони можуть зробити логіку виконання функції незрозумілою[15].

Насправді, при виклику функції, з усіх фактичних параметрів робляться копії, і саме вони підставляються у функцію замість формальних параметрів. Тому, навіть якщо в тілі функції деякому формальному параметру привласнюється нове значення, це не вплине на відповідний фактичний параметр у зовнішній програмі:

```
> my.sort<-function(x){x<-sort(x)}
> z<-c(3,5,1)
> y<-my.sort(z)
> y
[1] 1 3 5
> z
[1] 3 5 1
```

[15] Ці пояснення щодо локальних і глобальних змінних є "правильними у першому наближенні". Для того, щоб розуміти, як працюватиме глобальне привласнення у тому чи іншому випадку, потрібне більш повне уявлення про організацію пошуку змінних за їх іменами у `R`. Більш детально логіка `R` у роботі з об'єктами з різних середовищ описана у розд. 8 [54].

(Змінна z не відсортувалась, хоча вона використана у виклику функції, яка сортує свій формальний параметр у тілі). Цей спосіб передачі параметрів *за значенням* а не *за назвою*[16] також запобігає небажаним побічним ефектам. Всю інформацію, яку треба буде використовувати після виконання виклику функції, слід записувати в її значення.

У списку формальних параметрів функції можна використовувати символ ... — трикрапка. Він позначає, що функцію можна викликати з довільною кількістю параметрів. Параметри, що стоять на місці ... можна використовувати у тілі функції так само, як інші, єдина їхня відмінність полягає в тому, що вони не мають індивідуальних імен. У наступному прикладі створюється функція, що має один іменований параметр x та може викликатись із довільною кількістю інших параметрів. Параметр x не використовується, а всі параметри, що підставляються у виклику на місці ..., збираються в один список, який і є результатом виконання функції:

```
> f<-function(x,...){list(...)}
> z<-f(2,1:5,"aaa",T)
> z

[[1]]
[1] 1 2 3 4 5

[[2]]
[1] "aaa"

[[3]]
[1] TRUE

> z[1]
```

[16] При передачі параметра у функцію за назвою функція використовує безпосередньо той глобальний параметр, ім'я якого їй передається. Якщо функція його змінює, то він залишається зміненим і після виклику функції.

```
[[1]]
[1] 1 2 3 4 5
```

2.7.2 Векторизація функцій

Більшість таких елементарних функцій, як `sin()` чи `log()`, при застосуванні до масивів виконуються поелементно. Цю властивість успадковують і формули, які комбінуються з таких функцій:

```
> sin(log(1:3))+cos(1:3)

[1]  0.54030231  0.22281444 -0.09941545
```

Ця властивість може зберігатись і тоді, коли введено власну функцію. Наприклад, функція

```
> f<-function(x){sin(log(x))+cos(x)}
```

обчислюється для векторних `x` без помилки:

```
> f(1:3)

[1]  0.54030231  0.22281444 -0.09941545
```

Але, якщо у тілі функції зустрічаються умовні оператори або інші оператори управління, можливість поелементного виконання втрачається:

```
> f<-function(x){if(x>0) 1 else 0 }
> f(c(2,-1,0.5))

[1] 1
```

Для того, щоб R застосовував такі функції поелементно, можна використати техніку векторизації:

```
> fv<-Vectorize(f)
>   fv(c(2,-1,0.5))

[1] 1 0 1
```

Функція `Vectorize()` у цьому прикладі робить із функції `f` її векторизований варіант `fv`. Якщо не векторизована функція `f` не буде використовуватись сама по собі, то її можна не називати, а викликати `Vectorize()` з анонімною функцією:

```
> fv<-Vectorize(function(x){if(x>0) 1 else 0 })
```

Це дасть той самий результат, що і попередній приклад.

При використанні `Vectorize()` без додаткових опцій, функція векторизується по всіх аргументах. Наприклад:

```
> g<-function(x,t){if(x>t) 1 else 0}
> gv<-Vectorize(g)
> gv(c(1,-1,2),0)

[1] 1 0 1

> gv(c(1,-1,2),c(0,1,4))

[1] 1 0 0
```

Тут кожен елемент вектора `x` порівнюється з відповідним елементом вектора `t` (якщо ці вектори різної довжини, коротший використовується циклічно). Інколи виникає потреба мати функції, векторизовані лише за кількома своїми аргументами і не векторизовані за іншими. Наприклад, якщо ми хочемо мати функцію

ifin(x,d), яка перевіряє, чи належать елементи вектора x інтервалу d= (d_1, d_2), то вектор d треба використовувати як єдиний об'єкт, а не поелементно. Для цього можна вказати в опції vectorize.args список аргументів, по яких потрібна векторизація:

```
> ifin<-Vectorize(function(x,d)
+                 {if(x>d[1]&x<d[2]) 1 else 0},
+                 vectorize.args="x")
> ifin(1:6,c(2,5))

[1] 0 0 1 1 0 0
```

2.7.3 Структури управління виконанням програм у мові R

У R порівняно небагато мовних структур, які забезпечують зміну порядку виконання команд у програмі. Мова розроблена так, щоб мінімізувати потребу їхнього використання. Наприклад, там, де у інших мовах програміст змушений використовувати цикл for, у R можна скористатись векторними виразами. Можливість використання логічних масивів при індексації та параметрів-опцій зі стандартними значеннями помітно звужує область застосування структур умовних переходів типу if...else. Хороший стиль програмування у R полягає в тому, щоб не використовувати подібні структури там, де можна обійтись іншими. Однак у деяких випадках саме використання цих структур робить програму ефективною, а код зрозумілим. Опишемо ці структури послідовно.

Умовне виконання: if

У R є три варіанти структур, що реалізують класичний умовний перехід:

— `if(`*умова*`)`*команда* — якщо *умова* істина, *команда* виконується, інакше — не виконується (тут і далі *команда* може бути складною, тобто складатись із послідовності команд, об'єднаних фігурними дужками);

— `if(`*умова*`)` *команда1* `else` *команда2* — тут *команда1* виконується, якщо *умова* істина, *команда2* виконується, якщо *умова* — хибна;

— `ifelse(` *умова*,*команда1*,*команда2*`)` — логіка виконання така ж, як і у попередньому варіанті.

Наприклад:

```
> x<-1
> y<-2
> if(x<y) x else y

[1] 1
```

(Результатом виконання `if` тут буде менше з чисел x та y).

У третьому варіанті `ifelse()` працює як функція, зокрема, при застосуванні до векторів, вона дає векторні значення:

```
> x<-c(-4,4)
> sqrt(x)

[1] NaN   2

> sqrt(ifelse(x>0,x,NA))

[1] NA  2
```

В умовах `if` та інших структур управління можна використовувати логічні операції `&` (логічне і) та `|` (логічне або). При визначенні результату цих операцій спочатку обчислюється значення виразів ліворуч та праворуч від знака операції, а потім виконується сама

операція. Інколи результат операції можна визначити лише за значенням, що стоїть ліворуч, наприклад значення `T|x` завжди `T`, яким би не був `x`. Якщо у таких випадках вам не треба обчислювати вираз праворуч від знака операції, можна скористатись операціями `&&` та `||`:

```
> T|(sqrt(-5)>0)

Warning in sqrt(-5):  NaNs produced

[1] TRUE

> T||(sqrt(-5)>0)

[1] TRUE
```

(У другому варіанті не було спроби обчислити `sqrt(-5)`).

2.7.4 Вибір з кількох умов: switch

Функція `switch()` дозволяє обирати один з багатьох варіантів виконання програми залежно від значення деякого виразу. Її формат

`switch(`*вираз-умова*`,`*список варіантів*`)`.

Як приклад розглянемо застосування `switch()` у функції `f()`, що обчислює суму або добуток елементів вектора залежно від того, яких значень набуває параметр `type` — `"add"` чи `multiply`.

```
> f <- function(x, type)
+ {
+ switch(type, add = sum(x), multiply = prod(x), NA)
+ }
> f(1:4,type="add")

[1] 10
```

```
> f(1:4,type="multiply")

[1] 24

> f(1:4,type="x")

[1] NA
```

△

У цьому прикладі `switch()` обчислює значення виразу `type`, знаходить далі у списку параметрів такий параметр, назва якого відповідає `type`, та обчислює вираз, що стоїть після знака = для цього параметра (тобто значення цього параметра по умовчанню). Результат обчислення є значенням, яке дає `switch()`.

Останній елемент у списку параметрів `switch` у цьому прикладі (`NA`), записаний без знака =, задає дії, котрі будуть виконані, якщо значення виразу-умови не дорівнює жодному з попередніх варіантів. Якщо такого останнього елемента немає, жодні дії не виконуються, а значення `switch` дорівнює `NULL`.

2.7.5 Цикли while та repeat

`R`, як і інші мови програмування, використовує цикли для організації серій повторних обчислень. Загальний формат циклу `while`:

`while(`*умова*`)`*команда*

Спочатку перевіряється *умова*, і якщо вона дає результат `TRUE`, виконується *команда*. Цей процес повторюється циклічно і зупиняється як тільки *умова* набуде значення `FALSE`.

У наступному прикладі цей цикл використано для наближеного обчислення кореня рівняння $x = \cos(x)$ (`eps` — точність обчислень):

▽

```
> x<-1
> eps<-0.0000001
> while(abs(x-cos(x))>eps)x<-cos(x)
> x
```

```
[1] 0.7390851

> cos(x)

[1] 0.7390852
```

Інколи буває необхідно розмістити перевірку умови не на початку циклу, а в кінці, або навіть по середині. Для організації таких циклів зручно використовувати структуру `repeat` з командами `break` та `next`.

Формат команди

`repeat` *команда*

де *команда* — це "тіло циклу", тобто послідовність команд, які повинні виконуватись циклічно. Щоб комп'тер міг зупинитись, всередині тіла циклу має бути команда `break`, яка перериває виконання циклу і передає управління на команду, що йде одразу після тіла циклу:

```
> x<-NULL
> t<-100
> i<-0
> repeat
+ {
+ i<-i+1
+ if(i^2>t) break
+ x<-c(x,i^2)
+ }
> x

 [1]   1   4   9  16  25  36  49  64  81 100
```

(квадрати натуральних чисел додаються до списку доти, доки вони не перевищують поріг `t`).

Команда `next` всередині тіла циклу перериває виконання даного циклу і передає управління на першу команду тіла.

2.7.6 Цикл for

Цикл for використовується тоді, коли одну і ту саму дію треба виконати для певної послідовності (вектора) індексів. Формат відповідної структури

`for(`*індекс* in *послідовність*`)` *команда*

тут *індекс* — назва змінної, що використовується для індексації у тілі циклу; *послідовність* — вектор значень, що будуть підставлені замість індексу при виконанні циклу; *команда* — тіло циклу, тобто команда, або набір команд, вміщених у фігурні дужки, який буде виконано для всіх значень індексу.

Наприклад, якщо для вектора $x = (x_1, \ldots, x_n)$ треба обчислити вектор різниць $y = (x_2 - x_1, \ldots, x_n - x_{n-1})$, це можна зробити, використовуючи цикл:

```
> x<-(1:10)^2
> y<-rep(NA,length(x)-1)
> for(i in 1:(length(x)-1))y[i]<-x[i+1]-x[i]
> x

 [1]   1   4   9  16  25  36  49  64  81 100

> y

[1]  3  5  7  9 11 13 15 17 19
```

Вектор значень, по яких буде проводитись ітерація у `for` не обов'язково має складатись з цілих чисел. Це можуть бути будь-які об'єкти:

```
> for(i in c("a","b","c"))cat(i)

abc
```

(Функція `cat()` друкує свої аргументи на консолі).

Тим, хто звик програмувати на класичних мовах подібних до `C`, варто звернути увагу ще на одну особливість циклу `for` у `R`:

```
> for(i in 1:3){
+  cat("b: ",i)
+  i<-10
+  cat(" a: ",i,"; ")
+ }

b:  1 a:  10 ; b:  2 a:  10 ; b:  3 a:  10 ;

> cat("end",i)

end 10
```

Хоча змінна ітерації `i` змінюється всередині тіла циклу, це не впливає на значення, які їй привласнюються на наступній ітерації.

Потреба використання циклів for у `R` значно менша, ніж у більшості класичних мов програмування, завдяки можливостям застосування векторних функцій та гнучкій індексації масивів. Зокрема, у попередньому прикладі, вектор різниць можна обчислити так:

```
> x[-1]-x[-length(x)]

[1]  3  5  7  9 11 13 15 17 19
```

(Нагадаємо, що від'ємний індекс наказує вилучити відповідний елемент з вектора: `x[-1]` — вектор з усіх елементів `x`, окрім першого).

Хороший стиль програмування у `R` вимагає не використовувати цикли for, якщо без них можна обійтись.

2.8 Формули: y ~ x

У R є ще один специфічний клас об'єктів — формули. За допомогою формул користувач повідомляє статистичним функціям *специфікацію моделі*, на основі якої буде проводитись обробка даних. Кожна функція може розуміти формулу по-своєму, тому докладний опис використання формул відкладемо до вивчення відповідних функцій. Тут скажемо лише кілька слів про найбільш загальні їхні риси.

Зазвичай формули використовують там, де проводиться дослідження залежності однієї змінної від якихось інших. Змінну, залежність якої досліджується, називають *відгуком*, або залежною змінною. Змінні, від яких може залежати відгук — *регресорами*, пояснюючими або незалежними змінними.

У тексті програми формулу легко помітити завдяки наявності символу ~, який можна читати як "залежить від". Загальний формат запису формул:

відгук ~ опис моделі залежності від регресорів

Наприклад, нехай є два вектори x і y однакової довжини. Треба зобразити на координатній площині точки з відповідними координатами. Це можна зробити так:

```
> x<-seq(-1,1,0.1)
> y<-x^2
> plot(y~x)
```

Тут ми спочатку визначили x та y, а потім викликали функцію рисування[17], якій передали формулу: нарисувати залежність y від x. Результат — на рис. 2.1.

[17] Див. підрозд. 3.2.

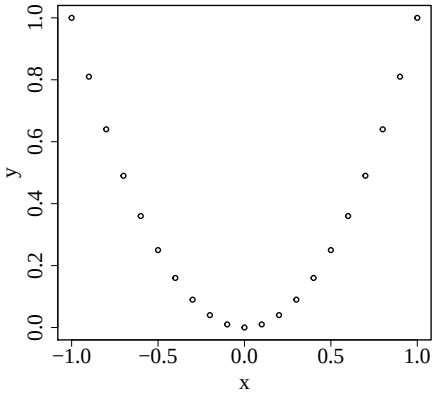

Рис. 2.1. Залежність `y` від `x`

Схожою є логіка використання формул у функції `boxplot()`, що рисує скриньки з вусами (див. підрозд. 7.5), але тут досліджується залежність відгука від деякого фактора, тобто змінна-регресор може набувати лише значень з деякого скінченного набору.

Для реалізації техніки лінійного регресійного аналізу в `R` використовується функція `lm()` (див. підрозд. 10.2). Формули дозволяють зручно задавати модель регресії для цієї функції. Наприклад, формула `y~x+u+v` відповідає регресійній моделі

$$y_j = b_0 + b_1 x_j + b_2 u_j + b_3 v_j + \varepsilon_j,$$

де b_i, $i = 0, \ldots, 3$ — невідомі коефіцієнти регресії, які функція `lm()` буде підганяти за методом найменших квадратів, ε_j — похибки регресії.

Таким чином, у розглянутих прикладах формула не є явним записом форми залежності, а лише пояснює, де взяти відгук, а де — регресори.

А от при підгонці нелінійної регресії за допомогою функції

`nls()`[18], у формулі треба безпосередньо записати ту функцію, яку підганяють:

```
y~C*exp(a*x)
```

буде формулою для підгонки нелінійної регресійної моделі

$$y_j = C \exp(a * x_j) + \varepsilon_j,$$

причому C і a будуть вважатись невідомими коефіцієнтами, які треба підігнати, якщо на момент виклику функції

```
nls(y~C*exp(a*x))
```

об'єкти `C` і `a` не були визначені. (Звичайно, `x` та `y` мають бути визначеними векторами однакової довжини для того, щоб підгонку можна було зробити).

Таким чином, записуючи формулу, треба завжди враховувати, яка саме функція буде її використовувати.

[18] Див. підрозд. 10.5.

Розділ 3

Базова графіка в R

Цей розділ присвячений основним засобам графічного відображення даних у R, які далі будуть використовуватись у даній книзі. R має надзвичайно розвинену графіку, з якої тут описана лише дуже мала частина. Ми познайомимось з базовими функціями, що дозволяють зображувати точки і лінії на площині, відображати тривимірні графіки, робити написи на рисунках.

Окрім того, розглядаються стовпцеві та кругові діаграми, що є дуже поширеним засобом відображення статистичних даних.

Окремо розглянемо питання про відображення статистичної інформації на географічній карті і на прикладі побачимо, як таке відображення дозволяє помічати особливості у даних і висувати гіпотези щодо їхньої природи.

Більш специфічні графічні засоби, які дозволяють аналізувати розподіл одновимірних даних, розглядаються далі у розд. 7. Візуалізація залежностей між різними змінними, що описують спостереження розглянута у розд. 5.

3.1 Стовпцеві та кругові діаграми

Один з найбільш популярних способів відображення не дуже великих наборів чисел — діаграми, на яких кожному числу відповідає один стовпчик. Англійською мовою такі рисунки називаються barplot або barchart. Для їхнього відображення можна використовувати функцію `barplot()`. Наприклад, у наборі даних `ldeaths` вміщені щомісячні дані про кількість смертей у Великій Британії від бронхітів, астми та емфіземи легенів. Перші 12 елементів набору відповідають 1974 р. Зобразимо їх на стовпцевій діаграмі:

```
> barplot(ldeaths[1:12],names.arg =month.abb)
> barplot(ldeaths[1:12],names.arg =1:12,
+         horiz=T,col=2:4)
```

Результати виконання відображені на рис. 3.1 — перший виклик `barplot()`, на рис. 3.2 — другий.

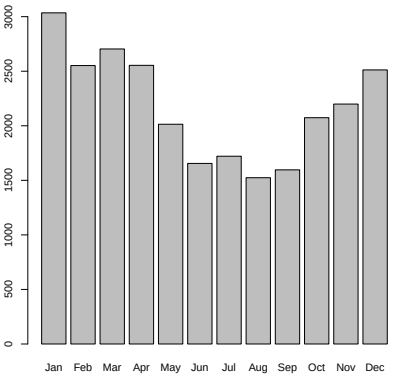

Рис. 3.1. Стовпцева діаграма вертикальна

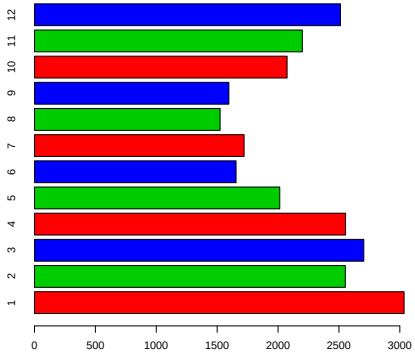

Рис. 3.2. Стовпцева діаграма горизонтальна

Функція `barplot()` має багато параметрів, вкажемо лише основні. Перший параметр `height` задає висоти стовпчиків, якщо стовпчики вертикальні, як на рисунку 3.1 або довжини — коли стовпчики горизонтальні, як на рис. 3.2. Вибір орієнтації стовпчиків задає параметр `horiz` (`T` — горизонтальні, `F` — вертикальні). Параметр `col` задає колір стовпчика, `names.arg` — назви, які будуть підписані під стовпчиками. (У першому прикладі ці назви взяті з масиву `month.abb`, який містить скорочені імена місяців).

Параметр `main` задає заголовок, що виводиться над рисунком. Можна також задати текст підпису під рисунком — використовуючи параметр `sub`.

Параметр `height` можна задати як матрицю. Це дає можливість порівнювати різні набори даних на одній діаграмі. Наприклад, у наборах `mdeaths` і `fdeaths` містяться дані про смертність окремо чоловіків та жінок. Зберемо ці дані в одну матрицю і виведемо:

```
> h<-rbind(fdeaths[1:12],mdeaths[1:12])
> rownames(h)<-c("female","male")
> colnames(h)<-month.abb
```

```
> barplot(h,sub="(a)",legend.text=T,
+       args.legend = list(x = "top",cex=2),
+       col=c("red","green"),angle=c(45,-45),
+       density=15)
> barplot(h,sub="(b)",legend.text=T,
+       args.legend = list(x = "top",cex=2,inset=0.1),
+       col=c("red","green"),angle=c(45,-45),
+       density=15,beside=T)
```

△

Тут перший рядок матриці відповідає жіночій смертності по місяцях, другий - чоловічій. Ми надали цим рядкам імена `female` і `male`, а стовпчики матриці даних назвали скороченими іменами місяців. Результати виконання відображені на рис. 3.3 — перший виклик `barplot()`, на рис. 3.4 — другий.

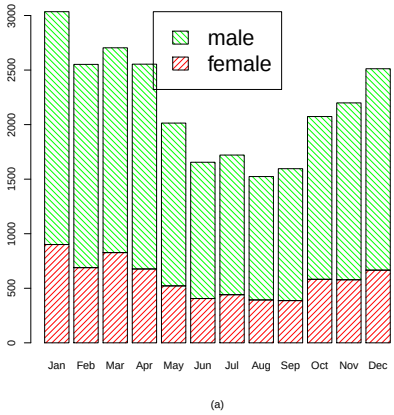

Рис. 3.3. Діаграма смертності чоловіків та жінок

У варіанті на рис. 3.3 стовпчики на діаграмі, що відповідають чоловікам, виведені як продовження стовпчиків для жінок. Це зручно тим, що одразу можна порівнювати сумарні смертності чоловіків та жінок протягом різних місяців. У варіанті на рис.

3.4 стовпчики чоловіків та жінок виводяться поруч, так зручніше порівнювати їх між собою. Вибір із цих варіантів відображення робить параметр `beside`.

Щоб пояснити читачу, де на діаграмі виводиться який рядок матриці, можна відобразити пояснення (легенду). Текст пояснення вказується у параметрі `legend.text`. Якщо просто задати `legend.text=T`, то для пояснення будуть використані назви рядків матриці `height`, як це зроблено у нашому випадку. Для рис. 3.3 ми задали положення легенди на рисунку, задавши конструкцію з опцій

```
args.legend = list(x = "top",cex=2).
```

Тут `x = "top"` вказує, що легенда має бути вгорі. Можливі варіанти: `"bottomright"`, `"bottom"`, `"topleft"`, `"center"` та аналогічні.

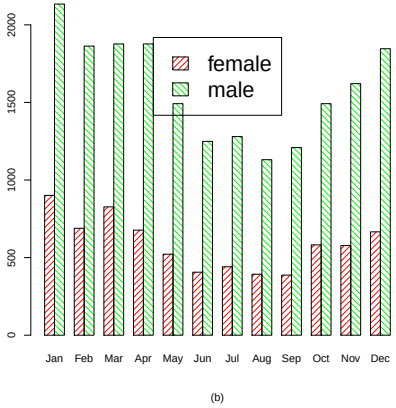

Рис. 3.4. Стовпцева діаграма смертності чоловіків та жінок

Задавши `x = "top"` у списку опцій легенди, ми збільшили розмір шрифту вдвічі порівняно із стандартним. Для рис. 3.4 відстань від середини легенди до правого поля рисунка задано опцією `inset`.

У функції `barplot()` є багато інших опцій, зокрема параметри `density`, `angle`, `border` регулюють штриховку стовпчиків та зображення контуру аналогічно тому, як це робиться у функції відображення прямокутників `rect()`. Опції, що керують відображенням осей координат: `axes`, `xlab`, `ylab` та розміром рисунка: `xlim`, `ylim` аналогічні відповідним опціям функції `plot()` (див. підрозд. 3.2).

Логічна опція `add` вказує: треба відкривати нове вікно для відображення діаграми (`add=F`), чи діаграма відображається у вже відкритому вікні доповнюючи існуючий рисунок (`add=T`).

Інколи для відображення даних використовують не стовпцеві, а кругові діаграми. Ідея полягає в тому, щоб зобразити "частки спільного пирога", які дістались різним їдокам. Відповідно англійська назва таких діаграм — pie charts. Наприклад, за даними перепису 1939 р. у Києві проживало

українців — 450 556,

євреїв — 224 236,

росіян — 139 495,

людей інших національностей — 27991.

Кругову діаграму для цих даних можна зобразити так:

```
> population<-c(450556,224236,139495,27991)
> names(population)<-c("U", "J","R","O")
> pie(population,col=c("red","green","blue","yellow"),
+ density=10,angle=c(45,90,-45,-90),
+ radius=1,cex=2)
```

Результат — на рис. 3.5.

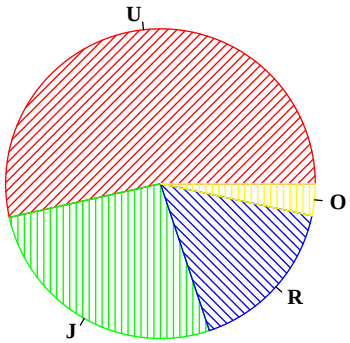

Рис. 3.5. Національний склад населення Києва у 1939 р.
U — українці, J — євреї, R — росіяни, O — інші.

Кругові діаграми вважаються менш візуально сприйнятними, ніж рисунки, складені зі стовпчиків, тому у серйозних дослідженнях як окремий засіб графічного відображення застосовуються не часто. Зазвичай, стовпцеві діаграми використовувати доцільніше. Але кругові діаграми завдяки своїй компактності можуть бути зручними для порівняння великої кількості наборів даних, наприклад, при відображенні на географічній карті складу населення різних міст країни тощо.

3.2 Точки та лінії на площині

У `R` для графічного відображення об'єктів часто використовується функція `plot()`. Ця функція є родовою (generic function), тобто вона працює по-різному для різних класів об'єктів. Зараз розглянемо простіше застосування `plot()` у випадку, коли треба відобразити які-небудь точки або лінії на площині. У такому випадку координати точок можна задати у вигляді векторних параметрів

функції:

```
> x<-sin(1:20)
> y<-cos(1:20)
> plot(x,y)
```

Результат виконання цієї функції — зображення на координатній площині набору точок, у яких горизонтальні координати взяті з вектора `x`, а вертикальні — з `y` (рис. 3.6). Зрозуміло, що для нормальної роботи функції ці вектори повинні мати однакову довжину.

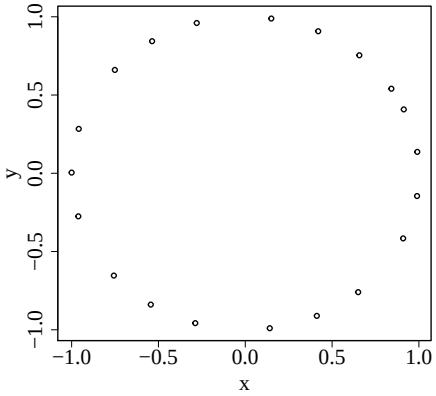

Рис. 3.6. Набір точок, виведений функцією plot()

Функція `plot()` спочатку створює нове вікно виводу, а потім виводить у нього об'єкти (у нашому випадку — точки). Якщо треба додати нові об'єкти на старому рисунку, замість `plot()` краще використати іншу функцію[1], наприклад — `points(x,y,...)` або `lines(x,y,...)`.

[1]Або скористатись функцією `plot()`, але перед цим встановити параметр відображення `new=T` за допомогою функції `par()`. Наприклад: `par(new=T);plot(x,y)`.

```
> x<--4:4
> y<-x^2
> plot(x,y,pch=2,cex=2)
> x1<-c(-1,1)
> y1<-c(10,10)
> points(x1,y1,pch=3,cex=2)
```

Опція `pch` у функціях `plot()` і `points()` задає символ, яким будуть відображатись точки на рисунку.

На рис. 3.7 — результат, що буде відображений після виконання `plot(x,y,pch=2)`, на рис. 3.8 — як зміниться рисунок після додавання точок командою `points(x1,y1,pch=3)`.

Якщо до рисунка треба додати лінії, можна скористатись функцією `lines(x,y,...)`.

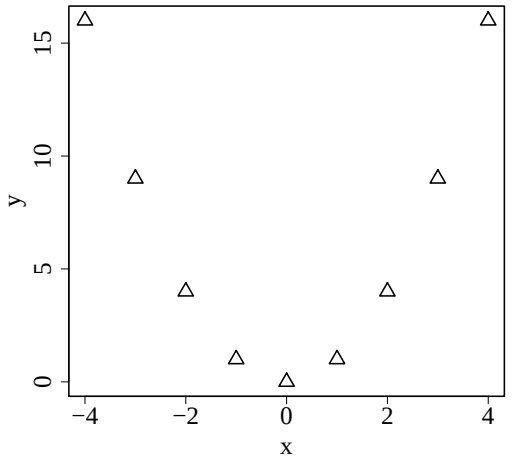

Рис. 3.7. Виведено `plot(x,y,pch=2,cex=2)`

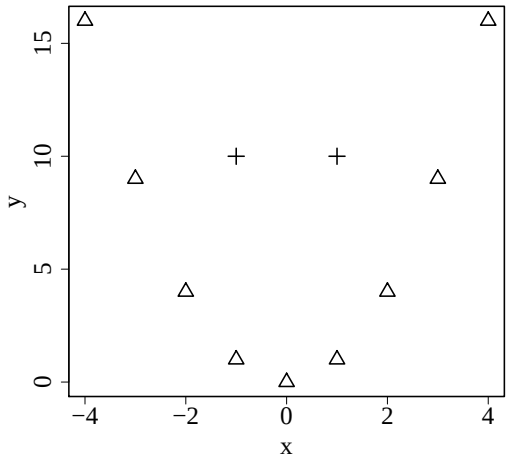

Рис. 3.8. Додали точки, використовуючи `points(x1,y1,pch=3,cex=2)`

Функція `plot()` має багато опцій, які дозволяють обирати формат відображення рисунка. Перелічимо найбільш вживані.

`axes` — логічна, якщо `T` — осі координат відображаються, якщо `F` — ні.

`xlab`, `ylab` — символьні, задають текст, яким будуть підписані координатні осі.

`sub` — задає підпис знизу під рисунком;

`main` — задає заголовок над рисунком.

`xlim`, `ylim` — задають межі для значень по горизонтальній та вертикальній осях.

`asp` — "aspect ratio" — співвідношення масштабних одиниць по вертикалі та горизонталі. Якщо не задавати цю опцію, то масштаб по вертикалі та горизонталі буде задаватись незалежно, виходячи із зручності розміщення рисунка. Якщо задати `asp=1`, масштабні одиниці по обох осях матимуть однакову довжину.

`log` — вказує, по якій осі вибрати логарифмічний масштаб, наприклад, `log="y"` задає логарифмічний масштаб по вертикалі,

`log="xy"` — по обох осях.

Наступні опції можна використовувати як у `plot()`, так і у `points()`, `lines()` та ряді інших графічних функцій.

`type` — тип точок/ліній. Ця опція може набувати значень:

`"p"` — відображати лише точки;

`"l"` — відображати лише прямі лінії, що з'єднують задані точки;

`"b"` — відображати і точки, і лінії, причому лінії торкаються точок;

`"o"` — відображати лінії, що перекривають точки;

`"s"`, `"S"` — з'єднати точки ступінчастими лініями, (`"s"` — стрибок ліворуч, `"S"` — праворуч);

`"h"` — відображаються вертикальні відрізки, що з'єднують задані точки з віссю абсцис.

`"n"` — на рисунку буде відображена координатна площина, у якій вміщуються всі задані точки, але ані самі точки, ані лінії, що їх з'єднують, не відображаються. (Ця опція використовується для того, щоб підготувати місце на якому рисунок буде відображений іншими функціями).

`col` — колір або кольори, яким будуть відображатись об'єкти (це може бути числовий вектор, або кольори можна задавати їх англійськими назвами `"red"`, `"blue"` тощо).

`pch` — символ, яким відображаються точки.

`cex` — контролює розмір символів.

`lwd` — контролює ширину ліній (стандартній ширині відповідає `lwd=1`).

`lty` — тип лінії. Може бути `"solid"`, `"dashed"`, `"dotted"`, `"dotdash"`, `"longdash"` або `"twodash"`. Можна також задати послідовністю символів від 1 до 9 та від A до F, які позначають довжину штриха та інтервалу між штрихами у шістнадцятковій системі числення: `lty="2F"` задає штрихову лінію з довжиною штриха 2 та інтервалом між штрихами 15.

Наприклад, рис. 3.9 відобразиться в результаті виконання такого скрипта:

```
> plot(1,1,xlim=c(1,16),ylim=c(-1.5,5),
+     type="n",xlab="",ylab="",
+     xaxt="n",yaxt="n")
> points(seq(1,15,2),rep(4,8),cex=1:8,col=1:8,pch=0:7)
> text(seq(1,15,2),rep(2,8),labels=paste(0:7),
+      cex=1:8,col=1:8)
> points(seq(1,15,2),rep(0,8),pch=8:15,cex=2)
> text(seq(1,15,2)+0.7,rep(0,8),
+      labels=paste(8:15),cex=1.5)
> points(seq(1,15,2),rep(-1,8),pch=16:23,cex=2)
> text(seq(1,15,2)+0.7,rep(-1,8),
+      labels=paste(16:23),cex=1.5)
```

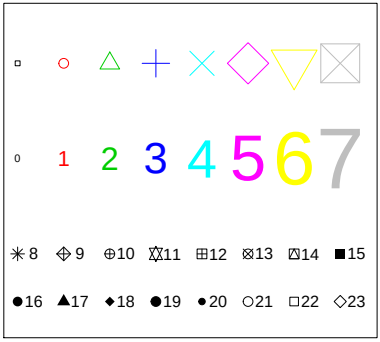

Рис. 3.9. Символи, розміри та кольори

Для того, щоб відображати на рисунках написи у заданих точках, використовується функція `text(x,y,labels,...)`. Тут вектори `x` та `y` задають координати точок, де розміщується текст,

`labels` — вектор символьних рядків, які будуть виводитись у заданих точках. Для цієї функції опція `pos` вказує, як розміщується текст щодо заданої точки (1 — під точкою, 2 — ліворуч, 3 — над, 4 — праворуч). `ofset` задає зміщення тексту щодо точки.

Опції `cex` та `col` задають розмір символів та їхній колір у `text()` так само, як у функції `plot()`.

Інколи буває необхідно відобразити одразу кілька наборів точок, заданих своїми координатами, що записані у стовпчиках деякої матриці. Це можна зробити, викликаючи функцію `points()` у циклі, окремо для кожного стовпчика. Але зручніше скористатись функцією `matplot()`, роботу якої демонструє рис. 3.10.

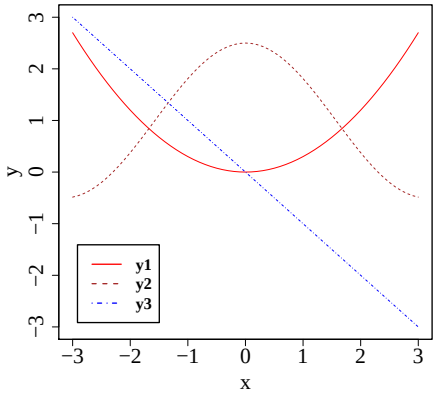

Рис. 3.10. Лінії, виведені функцією matplot()

Цей рисунок відображений таким скриптом:

```
> x <- seq(-3,3,0.1)
> y1 <- 0.3*x^2
> y2 <- 1.5*cos(x) + 1
> y3<- -x
> matplot(x, cbind(y1,y2,y3),  type="l",
```

```
+          col=c("red","brown","blue"),
+          lty=c("solid","dashed","dotdash"),ylab="y")
> legend("bottomleft", inset=.05,
+         legend=c("y1", "y2","y3"),
+         col=c("red","brown","blue"),
+         lty=c("solid","dashed","dotdash"),cex=1.5)
```

△

У цьому скрипті ми спочатку створили набір координат по горизонталі — `x` а потім підрахували значення трьох функцій `y1`, `y2`, `y3` для цих значень `x`. Далі графіки цих функцій виводяться функцією `matplot()`. Перші два параметри цієї функції — матриці, стовпчики яких містять координати точок по горизонталі і по вертикалі. Інші параметри-опції мають те ж значення, що й у функції `plot()`.

Далі ми використали функцію `legend()`, щоб відобразити табличку (легенду) з поясненням того, які кольори і типи ліній на графіку відповідають кожній з функцій `y1`, `y2`, `y3`. Перший параметр задає положення легенди (у лівому нижньому кутку), `inset` вказує, наскільки треба відсунути легенду від рамки, що оточує рисунок. Параметр `legend` задає текст легенди, інші параметри — ті ж, що у функції `matplot()`.

Функція `segments(x0,y0,x1,y1)` відображає набір відрізків. У векторах `x0`, `y0` розміщені x і y координати точок-початків відрізків, у `x1`, `y1` — точок-кінців.

Аналогічно, функція `arrows(x0,y0,x1,y1,length,angle)` рисує стрілки, `lenght` задає довжину "наконечника" стріли, `angle` — гостроту кута наконечника.

Функція

```
rect(xleft, ybottom, xright, ytop, density = NULL,
border = angle = 45,col = NA,NULL,...)
```

відображає прямокутники, координати лівих нижніх кутів беруться з `xleft`, `ybottom`, правих верхніх — з `xright`, `ytop`. Прямокутник може заповнюватись штриховкою, щільність цієї штриховки задається `density`, кут нахилу — `angle`, колір — `col`. Параметр `border` визначає колір контуру прямокутника.

Наприклад, на рис. 3.11 зображено результат виконання таких команд:

```
> plot(c(0,10),c(0,10),type="n")
> rect(1:9,rep(1,9),2:10,2:10,
+      density=5:13,col=1:9,border=10:19)
> arrows(rep(2,4),4:8,4:8,5:9,
+        angle=20,length=0.1)
> text(rep(2,4),4:8,
+      labels=c("a","b",'c',"d","e"),pos=2)
```

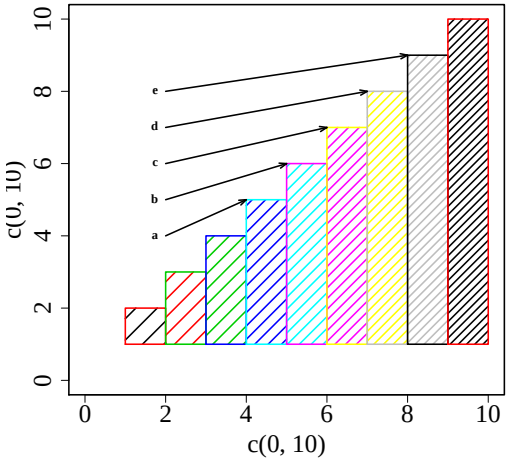

Рис. 3.11. Рисування прямокутників, стрілок та написів

Функція `abline(a,b)` рисує пряму лінію, що описується рівнянням $y = a + bx$. Якщо треба провести вертикальну лінію з горизонтальною координатою x, це можна зробити функцією `abline(h=x)`.

Для відображення графіків нелінійних функцій можна використовувати функцію

```
curve(expr,from=NULL,to=NULL,n=101,add=FALSE,
      type="l",xname="x",xlab=xname,ylab = NULL).
```

Параметри цієї функції:

`expr` — ім'я функції, що залежить від параметра `x`, або вираз, що залежить від змінної `x`;

`from`, `to` — лівий та правий кінці інтервалу зміни `x`, на якому будується графік;

`n` — кількість точок для зображення графіка;

`add` — логічний параметр, якщо він `TRUE`, графік будується на старому рисунку, якщо `FALSE` — для рисунка відкривається нове вікно;

`type, xlab, ylab` — такі ж, як у функції `plot`;

`xname` — ім'я що використовується для осі x.

Наприклад (рис. 3.12):

```
> curve(pnorm,-3,3,add=FALSE,col="blue",lty="dashed")
> curve((x/3)^2,col="red",lty="dotted",add=TRUE)
> abline(0.5,1)
```

Тут `pnorm` — функція розподілу для стандартного нормального розподілу (див. табл. 6.1).

У цьому прикладі другий виклик функції відображення кривих — `curve((x/3)^2,col="red",add=TRUE)` має опцію `add=TRUE`, тобто новий рисунок не створюється, крива відображається на старому. При цьому не задані параметри, що вказують діапазон по горизонталі (такі як `from`, `to`). У такому випадку криву рисують через весь старий рисунок — від його лівого до правого поля.

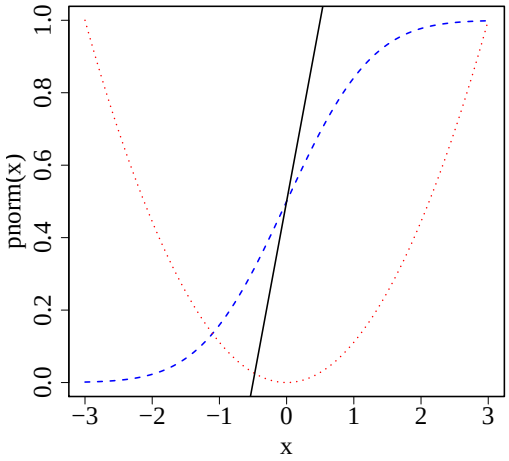

Рис. 3.12. Рисування кривих та прямих

3.3 Елементи тривимірної графіки

У статистиці досить часто спостережувані дані відображають у вигляді точок у просторі. Якщо кожній точці відповідає одне спостереження, а змінним — координати цієї точки, то такий рисунок називають *діаграмою розсіювання*. Якщо для відображення використовують лише дві змінні, утворюється двовимірна діаграма розсіювання, яку можна вивести на екран функцією `plot()`, як описано вище. Для відображення трьох змінних одразу використовують тривимірні діаграми розсіювання. Приклад такої діаграми — на рис. 3.13. У `R` діаграми розсіювання можна виводити багатьма різними способами, один з найпростіших — використання функції `scatterplot3d()` з пакета `scatterplot3d`.

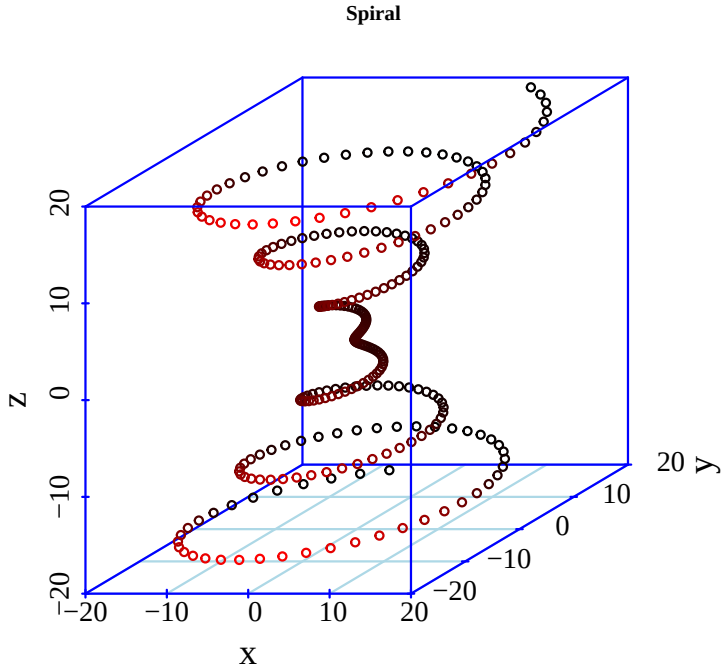

Рис. 3.13. Тривимірна діаграма розсіювання

У виклику функції `scatterplot3d()` перші три параметри `x, y, z` задають положення точок у тривимірному просторі. Опція `highlight.3d` визначає, чи треба розфарбовувати точки залежно від того, як вони розташовані по осі Ox. Два наступних параметри визначають колір осей координат та координатної сітки, зображеної у площині $x - y$. Нарешті, параметр `angle` визначає кут, який будуть утворювати на двовимірній проекції осі Ox та Oy. (Проекція завжди будується так, щоб вісь Ox на ній розташовувалась горизонтально, вісь Oz — вертикально, а напрямок Oy на рисунку визначається напрямком проектування. Змінюючи `angle` можна

розглядати дані “з різних точок зору”.

Приклад виклику цієї функції, який відображає рис. 3.13:

```
> library(scatterplot3d)
> z <- seq(-20, 20, 0.15)
> x <- z*cos(z)
> y <- z*sin(z)
> scatterplot3d(x, y, z,
+               highlight.3d=TRUE, col.axis="blue",
+               col.grid="lightblue", main="Spiral",
+               pch=1,angle=30)
```

У скрипті спочатку створена послідовність z-координат точок, потім розраховані відповідні x- та y-координати. Після цього відбувається відображення цих точок функцією `scatterplot3d()`.

Як бачимо, можливості `scatterplot3d()` при виборі напрямку проекції досить обмежені. Для того, щоб мати змогу повертати зображення даних інтерактивно, можна скористатись функцією `plot3d()` з пакета `rgl`.

Наприклад, викликавши цю функцію на даних попереднього прикладу у такий спосіб

```
library(rgl)
plot3d(x,y,z)
```

і покрутивши отриманий рисунок мишею, можна отримати зображення з рис. 3.14.

Інколи статистику треба відобразити на рисунку поведінку деякої числової функції двовимірного аргументу. Досить популярним засобом такого відображення є контурні графіки (contour plot), на яких зображають “лінії рівня” функції.

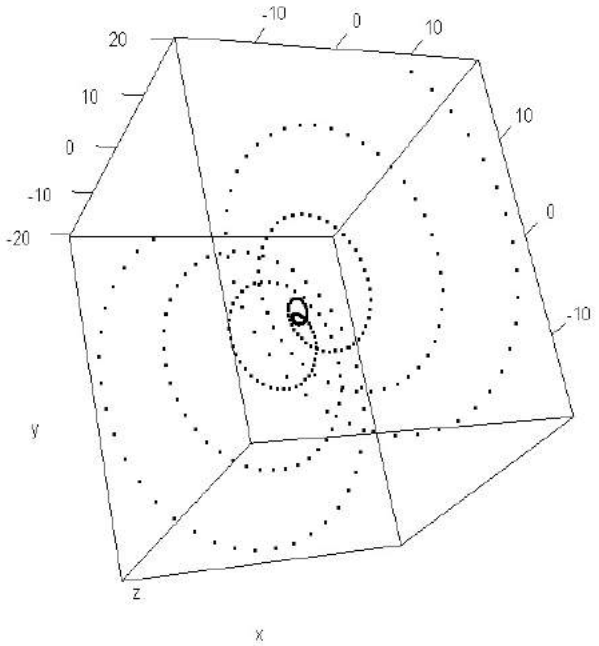

Рис. 3.14. Тривимірний рисунок за допомогою plot3d

Лінія рівня функції $f(x, y)$, що відповідає рівню c — це множина всіх точок на площині з такими координатами (x, y), що $f(x, y) = c$. На контурних графіках (рис. 3.15) відображають лінії рівня для різних рівнів, підбираючи їх так, щоб можна було побачити горби та западини функції подібно до того, як це роблять на географічних картах. У R контурні графіки зображають, використовуючи функцію `contour()`..

Інший варіант тривимірного графіка — зображення його проекції на площину аналогічно тому, як це було зроблено вище для діаграм розсіювання (рис. 3.16). Таке відображення тривимірних графіків у перспективі забезпечує функція `persp()`.

І `contour()`, і `persp()` працюють не безпосередньо з функ-

цією f, яку треба відобразити, а із матрицею `z` значень цієї функції у вузлах прямокутної сітки, визначеної векторами координат `x`, `y`, тобто `z[i,k]=f(x[i],y[k])`, `i=1,...,length(x)`, `k=1,...,length(y)`. Для того, щоб розраховувати значення `z`, зручно використовувати функцію `outer()`.

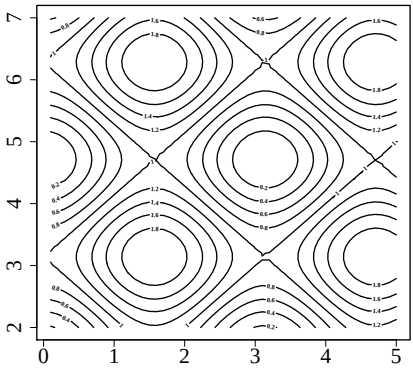

Рис. 3.15. Контурний графік

Приклад застосування `contour()` і `persp()` для відображення функції $f(x,y) = \sin^2 x + \cos^2 y$, коли $x \in (0,5)$, $y \in (2,7)$:

```
> x<-(1:50)/10
> y<-(20:70)/10
> f<-function(x,y){sin(x)^2+cos(y)^2}
> z<-outer(x,y,f)
> contour(x,y,z)
> persp(x, y, z, theta = 30, phi = 30, expand = 0.75,
+ col = "lightblue")
```

Результати виконання цього скрипта зображено на рис. 3.15 і рис. 3.16.

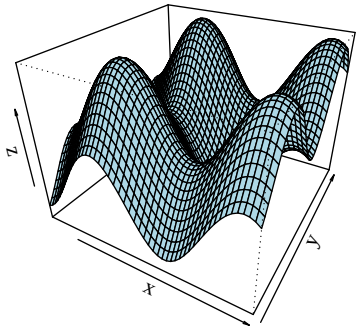

Рис. 3.16. Тривимірний графік

Функція `persp()` має більше можливостей вибору напрямку проекції, ніж `scatterplot3d()`. В ній цей напрямок задається двома кутами `theta` (азимут) та `phi` (90^o мінус широта). Крім того, параметр `expand` (зазвичай, його обирають від 0 до 1) можна використовувати для стиснення графіка по осі `z`.

3.4 Географічні карти

Статистична інформація часто має географічну прив'язку, тому для її відображення природно використовувати географічні карти. У `R` передбачений великий вибір можливостей такого відображення. У цьому підрозділі розглядаються лише два найпростіші приклади: відображення інформації фарбуванням різних областей різними кольорами та відображення кругових діаграм на географічних картах.

Спочатку розберемось, як у `R` зображати географічні карти. Це можна робити багатьма різними способами. Для нашої мети одним з найпростіших є використання пакетів (бібліотек) `sp`, `maptools`

та `raster`. Вони не входять у стандартну поставку `R`, тому їх треба інсталювати на комп'ютері звичайним способом та завантажувати перед використанням у робочу область, використовуючи функцію `library()`.

Приклад 3.4.1. Політична карта світу та її частини. У пакеті `maptools` міститься політична карта земної кулі під назвою `wrld_simpl`. Для того, щоб вивести її у вигляді рисунка можна скористатись звичайною функцією `plot()`, наприклад:

```
> library('sp')
> library('maptools')
> data(wrld_simpl)
> par(mai=c(0,0,0,0))
> par(mar=c(0,0,0,0))
> plot(wrld_simpl, xlim=c(-10,50),
+    ylim=c(-40,35), bg='white', col='khaki',
+    border='black')
```

У цьому скрипті ми спочатку функцією `data()` завантажили потрібну змінну у пам'ять. Потім задали командою `par` нульовий розмір полів рисунка.

І, нарешті, надрукували карту функцією `plot()`, використовуючи параметри:

`col` — колір, яким фарбується основна частина (суходіл),

`bg` — колір заднього плану (background) — море,

`border` — колір для границь позначених на карті країн,

`xlim`, `ylim` — межі регіону, який треба відобразити на карті по горизонталі та вертикалі.

Результат зображено на рис. 3.17. (Карта вже дещо застаріла, на ній не відмічено, наприклад, таку країну, як Південний Судан). ◄

На картах використовується географічна шкала координат, горизотнальна вісь відповідає довготі (longitude), вертикальна — широті (latitude). Як відомо, це кутові міри, вони визначаються у

градусах та хвилинах. Один градус становить 60 хвилин. У R використовується звичайне десяткове позначення для цих координат, тобто, скажімо, latitude -10.5 це 10 градусів 30 хвилин південної широти.

Рис. 3.17. Карта Африки

Карти кордонів окремих країн також містяться в об'єкті `wrld_simpl`. Їх можна використовувати, звертаючись по номерах країн у списку, розташованому приблизно за алфавітним порядком. Цей список міститься в атрибуті `wrld_simpl$NAME`. Вивівши цей список у R, можна побачити, що Демократична республіка Конго має номер 28, Нігерія — 153, Мадагаскар — 108. Нехай ми хочемо на вже існуючій карті Африки відмітити ці країни різними кольорами: Нігерію — зеленим, Мадагаскар — червоним, Конго — білим. Це можна зробити викликавши знову функцію `plot()` з параметром `add=T`, що означає — дорисувати новий рисунок поверх

попереднього:

```
plot(wrld_simpl[c(28,153,108),],
col=c('white','green','blue'),add=T)+
```

Написи на картах можна наносити, використовуючи функцію `text` так, як це було описано у п. 3.2.

Приклад 3.4.2. Кругові діаграми на карті. Покажемо, як відображати на картах кругові діаграми, подібні до тих, що описані у п. 3.1. Для цього ми використаємо функцію `floating.pie()` з пакета `plotrix` (його треба інсталювати на комп'ютері та завантажити).

Цю функцію можна викликати так:

```
floating.pie(xpos,ypos,x,col,radius)
```

де

`xpos`, `ypos` — координати центра кругової діаграми на рисунку (на карті),

`x` — вектор, координати якого відповідають розмірам секторів на круговій діаграмі,

`col` — кольори секторів,

`radius` — радіус діаграми.

Нехай, наприклад, ми хочемо для вибраних нами країн відобразити круговими діаграмами розподіл населення за релігійною ознакою. Для Нігерії, скажімо, маємо: 58% християн, 41% — прибічники ісламу, 1% — інші релігії. Програма може мати такий вигляд:

```
> par(mai=c(0,0,0,0))
> par(mar=c(0,0,0,0))
> plot(wrld_simpl, xlim=c(-10,50),
+     ylim=c(-40,35), bg='azure2', col='white',
+     border='black')
> # Розфарбовуємо Конго, Нігерію, Мадагаскар
> plot(wrld_simpl[c(28,153,108),],
+      col=c('white','green','blue'),add=T)
> library(plotrix)
```

```
> # Nigeria
> floating.pie(7,9,c(58,41,1),
+              col=c("red","yellow","magenta"),
+              radius=2.5)
> # Congo
> floating.pie(22,-4,c(79.8,11.3,8.9),
+              col=c("red","yellow","magenta"),
+              radius=2.5)
> # Madagaskar
> floating.pie(47,-18,c(41,7,52),
+              col=c("red","yellow","magenta"),
+              radius=2.5)
> legend("bottomleft",
+       legend=c("Christian","Islam","Other"),
+       fill=c("red","yellow","magenta"))
```

△

Спочатку рисуємо карту Африки як у попередньому прикладі, потім розфарбовуємо три країни і рисуємо кругові діаграми. Остання виконана функція — `legend()` створює легенду (пояснення) до карти. Перший параметр `"bottomleft"` визначає положення легенди у лівому нижньому кутку карти. Параметр `legend` — це вектор символьних рядків, кожен з яких відповідає одному рядку легенди. Параметр `fill` задає кольори, які будуть пояснюватись легендою.

Результат — на рис. 3.18. ◄

Приклад 3.4.3. Регіони країни. Карти країн, що містяться у наборі `wrld_simpl`, є досить грубими, вони не містять адміністративного поділу. Тому для того, щоб відображати статистику по регіонах якої-небудь країни, потрібні більш детальні карти. Їх можна взяти у пакеті `raster`. (Не забудьте його завантажити). У цьому пакеті є функція `getData()`, яка завантажує з інтернету карти різних країн за їхніми кодами `ISO`. Отримати перелік всіх країн з їхніми кодами можна, викликавши `getData('ISO3')` Наприклад, код України — `UKR`.

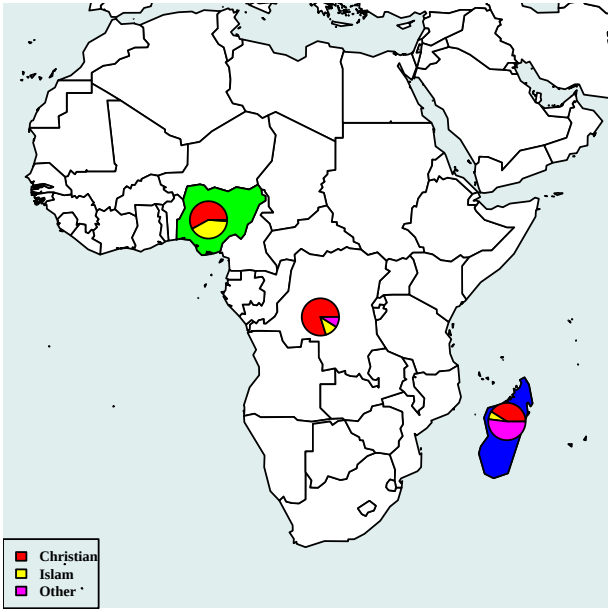

Рис. 3.18. Карта Африки з розподілом релігій

Для завантаження карти викликаємо `getData()`:

```
> library(raster)
> ukraine <- getData('GADM', country='UKR', level=1)
```

Параметр `'GADM'` показує, що карта буде завантажена з бази даних про адміністративні кордони (є іще бази кліматичних та топографічних карт). Параметр `country='UKR'` вказує, що буде завантажуватись карта України. Параметр `level` визначає деталізацію карти. Значенню 0 відповідає карта з лише державними кордонами, 1 — регіональні кордони (для України — областні, для США — штатів, для Польщі — воєводств), 2 відповідає районам для України, графствам для США, повітам для Польщі. Можливий також 3-й рівень для ще дрібніших одиниць (гміни у Польщі).

Таким чином, ми завантажили карту України з границями областей і зберегли її у вигляді змінної `ukraine`. Ця змінна розміщена у робочій області `R`. Якщо робочу область не зберігати наприкінці сеансу, карта загубиться. Доцільно зберегти її окремо у файлі для подальшого користування. Це можна зробити, використовуючи функцію `save()`:

```
save(ukraine,file="c:/rem/term/ukrmap.Rdata")
```

— зберігає карту у вигляді об'єкта `R` на диску `c` у каталозі `term` під назвою `ukrmap.Rdata`. Назва і каталог можуть бути довільними, розширення `Rdata` стандартне для `R`, втім, при бажанні можна використовувати і інші розширення, але формат файла при цьому не зміниться (тобто якщо вказати `ukrmap.pdf`, `R` збереже карту у такому файлі, але не у форматі pdf, а у своєму внутрішньому форматі).

Для того, щоб завантажити збережену карту під час нової сесії роботи з `R` тепер досить набрати

```
load(file="c:/rem/term/ukrmap.Rdata")
```

Після цього карта стане доступною у вигляді об'єкта з назвою `ukraine`. Її можна відобразити, використовуючи `plot()`.

Відображати карти окремих регіонів тепер можна, викликаючи цю функцію, наприклад, так:

```
plot(ukraine[list_reg],col=list_col)
```

де `list_reg` — список номерів регіонів, які треба відобразити, `list_col` — список кольорів, якими ці регіони будуть зафарбовані. Регіони у змінній `ukraine` розташовані в алфавітному порядку їхніх англійських назв. Щоб побачити ці назви і їхній порядок, можна вивести атрибут `NAME_1` змінної `ukraine`, тобто `ukraine$NAME_1`.

Як приклад, розглянемо відображення густоти населення України у різних областях (рис. 3.19).

▽

```
> library(raster)
> load(file="c:/rem/term/ukrmap.Rdata")
> dens<-read.csv(file="c:/rem/term/gustotan.csv")
```

```
> brk<-seq(30,170,20)
> int<-8-findInterval(dens[,2],brk)
> palette(gray(0:7/7))
> par(mai=c(0,0,0,0))
> par(mar=c(0,0,0,0))
> plot(ukraine,col=int)
> plot(ukraine[c(11,20),],col="red",add=T)
> legend("bottomleft",title="Population Density",
+     legend=c("150-170","130-150","110-130",
+              "90-110","70-90","50-70","30-50"),
+     fill=gray(0:7/7))
```

△

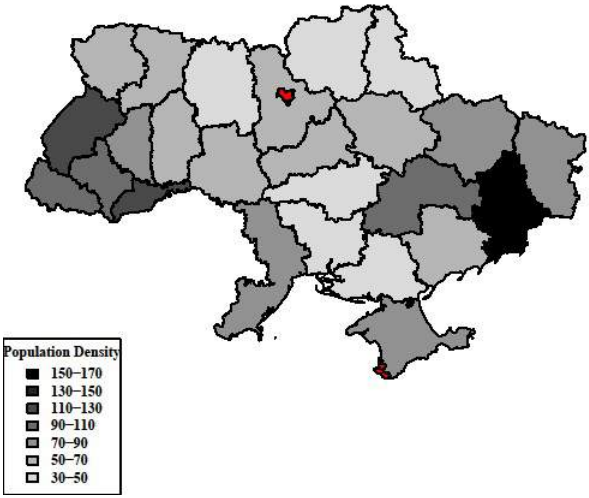

Рис. 3.19. Густота населення України

Дані щодо густоти населення містяться у файлі `gustotan.csv`. Перший стовпчик таблиці даних називається `region` і містить ан-

глійські назви областей (регіонів України) в алфавітному порядку. Другий — густоту населення (кількість осіб на 1 кв. км) у даному регіоні. Ці дані коливаються у діапазоні від 32.9 у Чернігівській області до 3442.6 у місті Києві. Зрозуміло, що міста Київ та Севастополь у цьому наборі даних різко виділяються (є викидами), тому при побудові шкали густот їх краще не враховувати. Серед областей найбільшу густоту населення має Донецька — 161.3. Тому ми вибрали інтервал від 30 до 170 і розбили його на підінтервали ширини 20. Кожному підінтервалу відповідає певна насиченість білого/сірого/чорного кольору, яким зафарбовується область. Така палітра кольорів створюється функцією `grey(0:7/7)` (0 відповідає чорний, 1 — білий колір). Функція

```
palette(grey(0:7/7))
```

встановлює такий набір кольорів, як палітру, що використовується іншими функціями, подібними до `plot()`. Щоб визначити, який номер кольору відповідає густоті населення кожної області, використовується функція `findInterval`, котра визначає, якому з інтервалів, заданих набором точок `brk`, належать густоти різних областей (ці густоти розміщені у другому стовпчику фрейма `dens`).

Далі розфарбовуємо всі регіони функцією `plot()`. Щоб виділити Київ та Севастополь (11-й і 20-й регіони), зафарбовуємо їх червоним кольором, використовуючи `plot()` з параметром `add=T`, тобто зверху попереднього рисунка. Нарешті, `legend()` додає пояснення кольорової шкали у лівому нижньому кутку карти. ◄

Приклад 3.4.4. Кому подобаються джинсові шорти. У цьому прикладі подивимось, яку користь може дати статистикові відображення інформації на карті. Ідея прикладу взята зі статті, розміщеної на сайті газети Wasington Post [32]. Автор статті аналізує дані з Google Trends[2] про кількість пошукових запитів, які надсилали жителі США до Google з двома наборами ключових слів: (1) запити зі словами “jean shorts” (джинсові шорти) і (2) запити зі словами “cargo shorts” (шорти у стилю “cargo”). Google Trends дозволяє

[2] `www.google.com/trends/` — це сайт, що показує статистику різних пошукових запитів, які різні користувачі інтернету надсилають у Google.

подивитись, скільки запитів надійшло за цими словами за певний період часу з кожного регіону США (або іншої країни, якщо вам цікаво). Отримані дані автор відобразив на карті.

Я повторив цей експримент. Мої дані дещо відрізняються від описаних у статті, оскільки зібрані за трохи інший період (у мене — з 01.01.2004 по 20.08.2016). Мої дані розміщені у файлі `shortU.txt`. В ньому кожен рядок відповідає одному штату США. П'ять стовпчиків-змінних вказують:

`n` — номер штату США у списку відповідної карти;

`state` — українська назва штату;

`cargo` — кількість пошукових запитів "cargo shorts" з даного штату;

`jean` — кількість запитів "jean shorts";

`urban` — відсоток міських жителів серед населення штату;

`stateen` — стандартне скорочене позначення штату англійською мовою.

Зафарбуємо на карті США червоним кольором (вертикальною штриховкою) штати, у яких переважають карго-шорти і синім (горизонтальною штриховкою) — ті, де більше цікавляться джинсовими шортами (рис. 3.20). (Для простоти Аляска та Гаваї не відображені. По двох штатах — Вайомінгу і Північній Дакоті даних немає, вони білі). Цей рисунок був створений таким скриптом:

```
> library(raster)
> load(file="c:/rem/term/usamap.Rdata")
> tb<-read.table("c:/rem/term/shortU.txt",header=T)
> color<-c("red","blue")
> stcol<-color[(tb$cargo<tb$jean)+1]
> angle<-c(0,90)
> stangle<-angle[(tb$cargo<tb$jean)+1]
> par(mai=c(0,0,0,0))
> par(mar=c(0,0,0,0))
> plot(usamap[tb$n,],col=stcol,ylim=c(27,43),
+   density=17,angle=stangle,
```

```
+    xlim=c(-124,-65))
> plot(usamap,lwd=2,ylim=c(27,43),add=T)
```

△

Графічні дані для рисування карти завантажені з файла `usamap.Rdata`, який можна створити аналогічно описаному у попередньому прикладі для карти України.

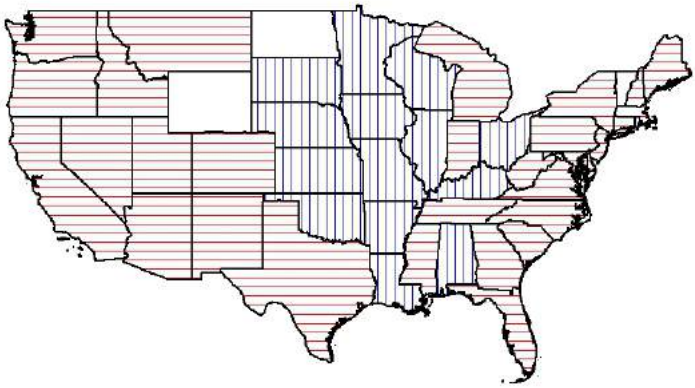

Рис. 3.20. Карго-шорти проти джинсових

Як бачимо, "сині" штати, де переважають джинси, зібрались компактною групою, що приблизно відповідає річковому басейну Міссісіпі — Міссурі (рис 3.21). Побачити цей ефект без відображення на карті навряд чи можливо. З чим він може бути пов'язаний? Автор [32] каже, що у цих штатах зосереджене фермерське землеробство США, тобто саме тут живуть "справжні американці", яким до вподоби справжній американський одяг — джинси.

Рис. 3.21. Басейн Міссісіпі — Міссурі

Щоб перевірити гіпотезу про зв'язок між інтересом до джинстових шортів та фермерством, можна поглянути на карту рівнів урбанізації (тобто відсотка міського населення) різних штатів. Для цього скористаємось наступною програмою (рис. 3.22):

```
> library(raster)
> load(file="c:/rem/term/usamap.Rdata")
> tb<-read.table("c:/rem/term/shortU.txt",header=T)
> numc<-10
> palette(gray(0:numc/numc))
> z<-tb$urban
> colorU<-floor(numc*(z-min(z))/(max(z)-min(z)))+1
> plot(usamap[tb$n,],col=colorU,ylim=c(27,43),
+   xlim=c(-124,-65))
> plot(usamap,lwd=1.5,add=T)
```

(Чим більш урбанізований штат, тим він світліший, найбільш аграрні — темні). Певна схожість цієї карти з попередньою помітна, але помітні і відмінності.

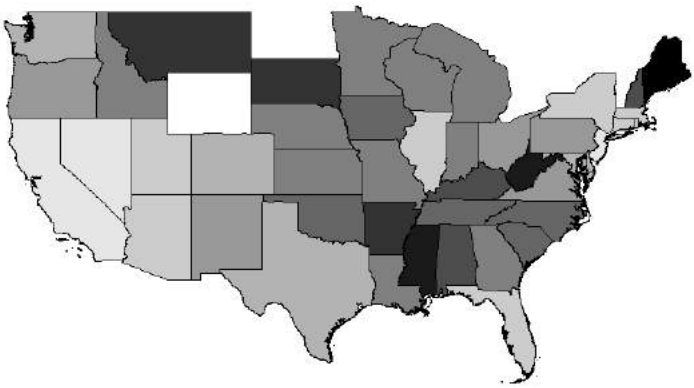

Рис. 3.22. Рівень урбанізації у США

Наскільки обґрунтована цими даними наша гіпотеза про зв'язок між інтересом до джинсових шортів і наявністю аграрного населення? Чи є дійсно зв'язок між положенням штату у басейні Міссісіпі-Міссурі і популярністю джинсових шортів? Відповідь непроста, повернемось до цього питання у прикладах 9.5.2 і 10.2.3.
◄

Розділ 4

Одновимірна описова статистика

Статистик, зазвичай, має справу з великими обсягами даних. Їх неможливо охопити оком, важко порівнювати з іншими аналогічними наборами даних. Тому виникає потреба описати основні особливості даних однією або кількома числовими характеристиками. Техніку такого опису називають *описовою* статистикою, а самі числові характеристики даних — дескриптивними[1] статистиками. При використанні та аналізі таких статистик дослідник намагається вивчати дані не на основі якоїсь наперед заданої теоретичної моделі, а виходячи зі структури самих даних. У цьому розділі ми розглядаємо техніку дескриптивної статистики саме з такої точки зору.

У розділах 7–10 розглядається інший підхід до статистичного аналізу, що спирається на використання теоретичних моделей розподілу даних. Описані у цьому розділі дескриптивні статистики можна використовувати і при такому підході, скажімо, як оцінки параметрів моделі, статистики тестів для перевірки гіпотез, про-

[1] Від лат. descriptio — опис. Інколи кажуть також “описові статистики” або просто “статистики”.

гнози для очікуваних спостережень. У цьому розділі ми не обговорюємо такі застосування, але інколи, для пояснення переваг тої чи іншої статистики, ми будемо згадувати трактування даних як **кратної вибірки** — набору незалежних, однаково розподілених випадкових величин[2]. Читачі, яким така трактовка не зовсім зрозуміла або здається недоречною при застосуванні до розглядуваних даних, можуть пропускати ці пояснення.

У цьому розділі обговоримо основні дескриптивні статистики наборів числових статистичних даних, у яких для кожного спостереження вимірюється одна числова характеристика (змінна)[3].

Наприклад, спостережуваними об'єктами можуть бути призовники до армії, а змінною, що досліджується — їх зріст. (Властивості цієї характеристики важливі для тих, хто займається забезпеченням військовослужбовців одягом). Інший приклад — вимірювання температури повітря на вулиці, які проводяться протягом року щодня о певній годині. Тут кожне спостереження відповідає дню вимірювання, а температура є змінною, що досліджується.

У випадку зросту призовників порядок, в якому розташовані об'єкти у наборі несуттєвий, він склався випадково і не пов'язаний з досліджуваним явищем. Перетасувавши виміряні значення зросту у довільному порядку, ми не втрачаємо корисної інформації. Такі набори даних прийнято називати вибірками.

Для вимірювань температури порядок суттєвий: температура на вулиці залежить від пори року, сьогоднішня температура залежить від вчорашньої і т.д. Дані, для яких важливими є такі ефекти, називають часовими рядами. Зрозуміло, що перетасувавши елементи часового ряду ми втратимо інформацію про ці залежності. Але інформація про деякі важливі особливості досліджуваної температури збережеться: якщо, наприклад, нас цікавить найбільша температура протягом року, на порядок вимірювань можна не звертати уваги. При дослідженні таких особливостей часові ряди

[2] Формальне означення кратної вибірки див. у підрозд. 8.1.

[3] Описова статистика багатовимірних даних розглядається у розд. 5, а графічні засоби аналізу одновимірних даних — у розд. 7.

можна (з певними застереженнями) розглядати як вибірки.

У цьому підрозділі ми зосередимось на аналізі вибірок, тобто таких наборів даних, для яких порядок спостережень несуттєвий.

Надалі ми будемо позначати X_j — значення досліджуваної змінної у j-му спостереженні, n — кількість елементів у вибірці, $\mathbf{X} = (X_1, \ldots, X_n)$ — вибірка.

4.1 Статистики середнього положення

Найпростіший спосіб схарактеризувати вибірку в цілому одним числом полягає в тому, щоб вказати "середнє положення", "центр вибірки", навколо якого коливаються вибіркові значення. Існує багато способів визначення такого числа і, відповідно, різні статистики середнього положення. Далі ми розглянемо найбільш поширені з них та обговоримо їх властивості.

Вибіркове середнє — статистика, що першою спадає на думку, коли треба визначити центр вибірки. Для вибірки $\mathbf{X}$ воно визначається за формулою

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j.$$

(У статистиці вибіркове середнє стандартно позначається тією ж літерою, що і усереднювана змінна з рискою над нею.)

Поширеність цієї статистики пов'язана із загальновживаним способом міркування на зразок: "середня врожайність цього сорту картоплі у наших умовах становить 20 т з гектара, отже, з наших трьох гектарів можна сподіватись приблизно 60 т врожаю". В основі такого прогнозування лежить уявлення про стабільність середніх за даними великого обсягу: врожай з одного куща картоплі змінюється унаслідок багатьох причин. Але при усередненні за всіма кущами з одного гектара поля індивідуальні особливості, що змінюють врожай різних кущів у різних напрямках, взаємно врівноважуються й отримується результат, який має бути приблизно

однаковим для всіх трьох гектарів нашого поля. У теорії ймовірностей цей ефект називають законом великих чисел і він є властивістю вибіркових середніх для даних, що описуються певними теоретичними моделями.

Окрім закону великих чисел у нашому міркуванні була використана ще "властивість адитивності" врожаю: повний врожай з усіх ділянок дорівнює сумі врожаїв кожної окремої ділянки. Цю властивість можна також назвати подібністю, або приблизною пропорційністю аналізованих явищ: при зростанні ділянки втричі, врожай також має зрости приблизно втричі.

Приклад 4.1.1. У деякому магазині протягом тижня щодня фіксували кількість покупців. Отримані дані — у табл. 4.1.

Таблиця 4.1. Кількості покупців у магазині протягом тижня.

День	пн.	вт.	ср.	чт.	пт.	сб.	нд.
Кільк. покупців	8	12	23	14	7	92	24

Потрібно оцінити, скількох покупців можна очікувати протягом кварталу (90 дн.). Зрозуміло, що найпростіша оцінка — визначити середню кількість покупців на день і помножити її на 90. Це можна зробити в R таким чином:

```
> # Середня кількість покупців на день:
> MeanBuy<-(8+12+23+14+7+92+24)/7
> MeanBuy

[1] 25.71429

> # Оцінка кількості покупців за квартал:
> 90*MeanBuy

[1] 2314.286
```

Середнє дорівнює 25.71429, а прогноз для кількості покупців за квартал[4] — 2314.286. Якщо для інших потреб ці дані використовуватись не будуть, то при такому обсязі це, мабуть, найбільш зручний варіант обчислення вибіркового середнього.

Якщо ви збираєтесь щось іще робити з даними, то їх краще запам'ятати в окремій змінній. Тоді з ними можна буде працювати використовуючи стандартні функції R. Наприклад:

```
> NumBuy<-c(8,12,23,14,7,92,24) # запам'ятали дані
> sum(NumBuy)/7                 # підрахували середнє

[1] 25.71429
```

— результат, звичайно, той самий, що й у попередньому варіанті.

У R також є спеціальна функція, що обчислює вибіркові середні —

```
> mean(NumBuy)

[1] 25.71429
```

Докладніше про роботу цієї та інших аналогічних функцій див. у підрозд. 4.6. ◄

Інше просте пояснення цієї характеристики — *по справедливості*: "якщо у всіх відібрати і розділити порівну, то кожен отримає вибіркове середнє". Це пояснення дозволяє зрозуміти, чому у деяких випадках вибіркове середнє заміняють іншими характеристиками.

[4] Зрозуміло, що цей приклад суто ілюстративний. Вибірка з семи спостережень обрана для того, щоб зручно було записати розрахунки. За даними такого обсягу прогноз не може бути точним. Більше того, він спирається на припущення про стабільність середніх, але, якщо наш магазин торгує морозивом, то середня кількість покупців у березні має відрізнятись від середнього за травень. Такого роду міркування не слід забувати, коли працюєш з реальними даними.

Середнє геометричне визначається для вибірок $\mathbf{X}$, у яких значення змінної X_j $j = 1, \ldots, n$ є додатними. Воно дорівнює

$$\mathrm{GM}(\mathbf{X}) = \left(\prod_{j=1}^{n} X_j \right)^{1/n}. \tag{4.1}$$

Приклад 4.1.2. (Застосування геометричного середнього у фінансовій математиці). Нехай укладено кредитну умову на n років, у якій боржник на початку терміну умови отримує суму S. За перший рік кредитування нараховується відсоток p_1, за другий — p_2 і т.д. Нарахування відбувається за схемою складних відсотків. Сплата боргу з усіма відсотками передбачається наприкінці терміну дії угоди. Як визначити середній відсоток по кредитуванню за цією угодою?

Що таке середній відсоток? Це таке $\hat{p}$, що, якби ми уклали угоду на n років з фіксованим відсотком $\hat{p}$, то виплата при поверненні боргу була б така сама, як і в розглянутій угоді зі змінними відсотками.

Виплата у схемі змінних відсотків дорівнює

$$S_n = S \prod_{j=1}^{n} (1 + p_j/100),$$

а виплата з фіксованим відсотком $\hat{p}$, вочевидь, була б

$$S'_n = S(1 + \hat{p}/100)^n.$$

Прирівнюючи S_n і S'_n, отримуємо

$$\hat{p} = 100 \left(\prod_{j=1}^{n} (1 + p_j/100) \right)^{1/n} - 100. \tag{4.2}$$

У цьому виразі легко побачити геометричне середнє величин $X_j = (1 + p_j/100)$ — приростів боргу протягом j-го року дії угоди. ◀

Отже, геометричне середнє природно застосовувати там, де загальний ефект виражається не як сума, а як добуток ефектів окремих спостережень.

Зауважимо також, що логарифм геометричного середнього є вибірковим середнім логарифмів спостережень:

$$\log(\mathrm{GM}(X)) = \overline{\log(X)}.$$

Можна сказати, що логарифмічне перетворення даних переводить геометричні середні у вибіркові.

Приклад 4.1.3. Нехай обсяг продаж магазину за січень збільшився порівняно з попереднім місяцем на 5%, за лютий — на 10%, а за березень — зменшився на 5%. Яким був середній темп зростання щомісячного обсягу продаж протягом першого кварталу?

Нас цікавить середній відсоток у схемі, що цілком відповідає логіці складних відсотків. Тому природно скористатись формулою (4.2):

$$p_{\text{сер}} = 100\sqrt[3]{1.05 \times 1.10 \times 0.95} - 100.$$

Обчислюється це у R так:

```
> 100*(1.05*1.1*0.95)^(1/3)-100

[1] 3.141916
```

або, якщо треба буде за цією формулою визначати середній відсоток для багатьох різних наборів даних:

```
> x<-c(5,10,-5) # вводимо дані
> # середній відсоток:
> 100*(prod((1+0.01*x))^(1/length(x))-1)

[1] 3.141916
```

Отримали 3.141916. Економісти зазвичай округлюють такі результати, наприклад, до двох цифр після коми. Це можна зробити автоматично, використовуючи функцію `round()`:

```
> round(100*(prod((1+0.01*x))^(1/length(x))-1),2 )

[1] 3.14
```

Якщо використати для усереднення відсотків звичайне вибіркове середнє, отримаємо у цьому прикладі $\bar{x} = 3.333333$. Відмінність від геометричного середнього невелика, але помітна. ◄

Середнє гармонічне — це величина, обернена до вибіркового середнього обернених величин спостережень:

$$\mathrm{HM(X)} = \frac{n}{\sum_{j=1}^{n} \frac{1}{X_j}} = \frac{1}{1/X}. \tag{4.3}$$

Гармонічні середні природно застосовувати для характеризації середніх положень змінних, які самі можна визначити, як відношення двох характеристик одного об'єкта, якщо чисельник є менш мінливим, ніж знаменник.

Приклад 4.1.4. (Середнє гармонічне у підрахунку mpg) Важливою характеристикою економічності автомобіля є шлях, що він проходить, витративши одиницю об'єму пального. У країнах з британською системою мір ця величина визначається у мілях шляху на галон пального і позначається mpg.

Для визначення mpg даного автомобіля використовуються тестові поїздки заданим маршрутом. Якщо довжина маршруту у мілях дорівнює S, а об'єм витраченого пального у галонах — V, то $\mathrm{mpg} = S/V$. Для надійності тестові поїздки повторюють декілька разів одним маршрутом, отримуючи різні об'єми витраченого пального V_1, V_2,. . . ,V_n. Відповідно, для кожного тесту можна визначити своє значення $\mathrm{mpg}_j = S/V_j$. Середнє значення mpg за

всіма тестами природно визначити як відношення загальної довжини пройденого в усіх тестах шляху до об'єму всього витраченого пального:

$$\widehat{\mathrm{mpg}} = \frac{\sum_{j=1}^{n} S}{\sum_{j=1}^{n} V_j} = \frac{n}{\sum_{j=1}^{n} 1/\mathrm{mpg}_j} = \mathrm{HM}(\mathrm{mpg}).$$

Таким чином, для усереднення mpg отриманих у серії тестових поїздок слід використовувати середнє гармонічне. ◄

На основі схожих міркувань рекомендується застосовувати середнє гармонічне для визначення середнього значення коефіцієнта ціна/прибуток (P/E, earnings multiple) при порівнянні інвестиційної привабливості акціонерних компаній [38].

Приклад 4.1.5. На склад для реалізації прибула партія харчової солі, розфасованої у пакети вагою 1 кг. Три пакети вибрали з партії і провели вимірювання насипної[5] густини солі, яка в них містилась. Отримали такі значення для кожного пакету: 1123, 1115 і 1284 (кг/м^3). Чому дорівнює середня густина солі у досліджених пакетах?

Зазвичай у таких дослідах обчислюють просто вибіркове середнє (у нас воно становить 1174 кг/м^3). Але можна помітити, що в даному випадку маса всіх зразків однакова, отже, відмінності густини пов'язані з різним об'ємом солі у різних пакетах. Оскільки густина обернено пропорційна об'єму, то природно використати гармонічне середнє:

```
> 3/(1/1123+1/1115+1/1284)

[1] 1169.067
```

[5] Нагадаємо, що густина речовини — це маса, яка припадає на одиницю об'єму. При визначенні насипної густини сипких матеріалів (піску, цукру, меленої кави, солі...) вимірюють той об'єм, який займає речовина, насипана вільно без стискання. Насипна густина солі залежить від помелу та вологості, тобто її значення важливе для перевірки якості продукту.

або так:

```
> x<-c(1123,1115,1284)
> 1/mean(1/x)

[1] 1169.067
```

Отримали 1169.067 — помітно менше ніж вибіркове середнє. ◄

Забруднення і робастність. При виборі статистики для характеризації середнього положення вибірки доцільно враховувати можливість забруднень. Забрудненою називається вибірка, у якій наявні значення, що не пов'язані з досліджуваним явищем, а потрапили до неї внаслідок помилки. Якщо таке неадекватне значення можна розпізнати і вилучити з вибірки, його називають "грубою помилкою" (напр., якщо вибірка складається зі значень ваги різних людей, всі від'ємні значення у ній будуть грубими помилками).

Але бувають забруднення, які не можна однозначно розпізнати, тому вони впливають на значення сумарних статистик, що обчислюються за вибіркою. Якщо дослідник не може з теоретичних міркувань виключити можливість забруднення, то для загальної характеризації вибірки бажано використовувати статистики, які не дуже сильно змінюються за наявності невеликої кількості забруднень. Такі статистики називають *робастними* (стійкими щодо забруднень).

Наприклад, вибіркове середнє $\bar{X}$ не є робастним: забруднення, при якому змінюється одне єдине значення у вибірці X може змінити $\bar{X}$ як завгодно сильно, якщо змінене значення обрати дуже великим.

Спостереження наздвичайно великі або надзвичайно малі порівняно з основною масою спостережень, називають **викидами**. Забруднення, які є викидами, зазвичай, є небезпечними з огляду на можливі впливи на описові статистики[6].

[6]Важливо розуміти, що викиди не обов'язково є наслідком забруднення вибірки. Вони можуть виникати природним чином, як її елементи. Наприклад, вибірки з логнормального розподілу (6.2.3) зазвичай мають викиди.

Те ж можна сказати і про середнє геометричне: збільшуючи лише один множник у добутку (4.1), можна зробити весь добуток, а отже, і середнє, як завгодно великим. А от для середнього гармонічного це неправильно. Дійсно, якщо у (4.3) одне спостереження, наприклад, X_n спрямувати до нескінченності, то середнє гармонічне прямуватиме до

$$\frac{n}{\sum_{j=1}^{n-1} 1/X_j},$$

тобто до величини, яка при великих n, приблизно дорівнює гармонічному середньому, обчисленому за даними $X_1, \ldots, X_{n-1}$. Тобто наявність одного великого викиду змінює гармонічне середнє не дуже сильно. Але, якщо спрямувати X_n до 0, то гармонічне середнє вибірки прямуватиме до 0, тобто до величини, шо може як завгодно сильно відрізнятись від гармонічного середнього початкової вибірки. Отже, для гармонічного середнього небезпечними є не великі, а малі (близькі до 0) викиди.

Зрізані середні. Розглянуті середні характеристики можна зробити стійкими до невеликої кількості забруднень, якщо застосувати техніку зрізання (truncation, trimming).

Переставимо елементи нашої вибірки у порядку зростання:

$$X_{[1]} \le X_{[2]} \le \cdots \le X_{[n-1]} \le X_{[n]}.$$

Тут $X_{[1]}$ — найменше значення у вибірці, $X_{[2]}$ — наступне за величиною, і т.д. аж до $X_{[n]}$ — найбільшого значення. $X_{[j]}$ називають j-ю **порядковою статистикою** вибірки $\mathbf{X}$, а послідовність порядкових статистик — **варіаційним рядом**.

Для того, щоб знайти зрізане середнє вибірки з рівнем зрізання α, треба відкинути[7] $\lceil n\alpha/2 \rceil$ найбільших та $\lceil n\alpha/2 \rceil$ найменших порядкових статистик і усереднити те, що залишилось:

$$\mathrm{TM}_\alpha(X) = \frac{1}{n - 2\lceil n\alpha/2 \rceil} \sum_{j=\lceil n\alpha/2 \rceil + 1}^{n - \lceil n\alpha/2 \rceil} X_{[j]}. \tag{4.4}$$

[7]Тут $\lceil x \rceil$ — найменше ціле число, що є більшим x або дорівнює x.

Аналогічно можна використовувати зрізане геометричне або гармонічне середнє.

Чим більшою вибрати частку відкинутих порядкових статистик, тим більш стійким до забруднення буде зрізане середнє. Граничний випадок досягається, коли відкидають всі спосереження крім того одного або двох, що знаходяться по середині варіаційного ряду. В результаті отримуємо характеристику середнього положення, яку називають *вибірковою медіаною*.

Вибіркова медіана — це статистика, що обчислюється за формулою

$$\operatorname{med}(X) = \begin{cases} X_{[(n+1)/2]}, & \text{якщо } n \text{ непарне,} \\ \frac{1}{2}(X_{[n/2]} + X_{[n/2+1]}), & \text{якщо } n \text{ непарне.} \end{cases} \tag{4.5}$$

Коротко можна сказати, що медіана — це середина варіаційного ряду: ліворуч від медіани міститься стільки ж значень, скільки і праворуч.

Медіана — найбільш робастна характеристика середнього положення у вибірці. Цим значною мірою пояснюється її популярність у багатьох застосуваннях.

Приклад 4.1.6. Розрахуємо вибіркову медіану на даних про кількості покупців магазину з табл. 4.1. Впорядкуємо цю вибірку у порядку зростання щоб отримати варіаційний ряд:

```
> sort(c(8,12,23,14,7,92,24))

[1]  7  8 12 14 23 24 92
```

По середині у цьому ряду стоїть число 14. Це і є медіана.

Можна обчислити медіану безпосередньо, використовуючи відповідну функцію R:

```
> x<-c(8,12,23,14,7,92,24)
> median(x)
```

```
[1] 14
```

△

Отримана нами медіана помітно менша, ніж вибіркове середнє знайдене у прикладі 4.1.1 — 25.71429. Це пов'язано з викидом: кількість покупців у суботу (92) в нас була у кілька разів більша, ніж у будь-який інший день. Такий викид не можна вважати забрудненням — природно, що по суботах в Україні люди більше займаються покупками порівняно з буднями. Тому вилучати це число з вибірки не варто. Більше того, цей ефект доцільно було враховувати при оцінці загальної кількості покупців за квартал у прикладі 4.1.1, оскільки суботи будуть регулярно повторюватись протягом кварталу.

Однак, якщо ми хочемо схарактеризувати "типову кількість" покупців у магазині протягом дня, природно використати для цього саме медіану — це кількість покупців у день, який лежить по середині між менш і більш вдалими днями нашого магазину. ◀

Можна сказати, що порядкові статистики, які розташовані поблизу середини варіаційного ряду, є найбільш робастними. І навпаки, найбільш чутливими до забруднень є "екстремальні" порядкові статистики $X_{[1]} = \min(X_1, \ldots, X_n)$ та $X_{[n]} = \max(X_1, \ldots, X_n)$. Інтервал $[X_{[1]}, X_{[n]}]$ називають *діапазоном вибірки*, а величину

$$\mathrm{MR}(X) = \frac{X_{[1]} + X_{[n]}}{2}$$

— **серединою діапазону** (midrange). $\mathrm{MR}(X)$ також є характеристикою середнього положення у вибірці, хоча і зовсім не робастною. Скажімо, якщо у вибірці є одне забруднення, воно може помітно змінити вибіркове середнє. Але при зростанні обсягу вибірки вплив цього забруднення буде зменшуватись. Для середини діапазону це не так: одне значення забруднення, яке є більшим, ніж всі спостережувані значення досліджуваної змінної, залишиться $X_{[n]}$, скільки б нових спостережень ми не зробили. Отже, використовувати середину діапазону для характеризації середнього

положення слід дуже обережно.

Приклад 4.1.7. Для даних з табл. 4.1 середину діапазону можна обчислити так (x визначено у попередньому прикладі):

```
> (min(x)+max(x))/2
[1] 49.5
```

— отримали 49.5. Це більше, ніж всі спостереження у вибірці крім одного. Навряд чи в даному випадку можна казати про вдалу характеризацію середнього положення. Зрозуміло, що це результат впливу суботнього викиду.

Але, якщо нас цікавить, скажімо, яку можливу кількість покупців доведеться обслуговувати протягом одного дня у нашому магазині, таке велике значення середини діапазону буде вказувати, що нам треба передбачити можливий наплив клієнтів у окремі пікові дні. З цієї точки зору середина діапазону може бути корисною характеристикою. ◄

Зауважимо, що у тих випадках, коли забруднень немає, середина діапазону може виявитись значно більш точною оцінкою теоретичного середнього положення, ніж вибіркове середнє (для математичного сподівання рівномірного розподілу за кратною вибіркою — див. приклад 8.4.5).

4.2 Статистики розкиду

Щоб одним числом показати, як далеко вибіркові значення можуть відхилятись від середнього положення, використовують статистики розкиду.

Найбільш популярною такою статистикою є **вибіркова дисперсія** (sample variance). Вона визначається як середнє квадратів відхилень спостережень від вибіркового середнього:

$$S^2(X) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2. \tag{4.6}$$

Часто використовується **виправлена вибіркова дисперсія**, яка відрізняється від звичайної лише нормуючим множником[8] $(n-1)/n$:

$$S_0^2(X) = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2.$$

Використання виправленої вибіркової дисперсії пов'язане з тим, що вона є незміщеною оцінкою для теоретичної дисперсії по кратній вибірці. У багатьох підручниках та комп'ютерних програмах $S_0^2(X)$ називають *вибірковою дисперсією*, а $S^2(X)$ — *популяційною дисперсією*, або дисперсією генеральної сукупності. Вибіркову дисперсію інколи позначають σ^2.

Корінь квадратний з вибіркової дисперсії називають (вибірковим) **середньоквадратичним відхиленням** (або вибірковим стандартним відхиленням):

$$S(X) = \sqrt{S^2(X)} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

і, аналогічно,

$$S_0(X) = \sqrt{S_0^2(X)} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2}.$$

Вибірковим середнім абсолютним відхиленням (mean absolute deviation, або mean deviation) називають середнє абсолютних відхилень вибіркових значень від вибіркового середнього:

$$\mathrm{MAD}(X) = \frac{1}{n}\sum_{j=1}^{n}|X_j - \bar{X}|.$$

[8] Цей множник називають *поправкою Бесселя*.

Інколи використовують також середнє абсолютне відхилення від медіани:

$$\mathrm{MAD}_{\mu}(X) = \frac{1}{n}\sum_{j=1}^{n} |X_j - \mathrm{med}(X)|.$$

Зрозуміло, що вибіркова дисперсія та середнє абсолютне відхилення не є робастними — одне забруднення, що лежить далеко від інших спостережень, може змінити ці характеристики як завгодно сильно. Тому для забруднених вибірок розроблені спеціальні робастні характеристики розкиду, серед яких найбільш поширений інтерквартильний розмах.

Квартилі та інтерквартильний розмах (quartiles and interquartile range). Вибіркова медіана розбиває варіаційний ряд на дві частини однакового розміру. Медіани кожної з цих двох частин називають *квартилями*. За тою частиною, де вибіркові значення менші або дорівнюють медіані всієї вибірки, визначають нижній (кажуть також лівий або перший) квартиль $Q_1(X)$, а за тією, у якій значення більші або дорівнюють медіані — верхній (правий або третій) квартиль $Q_3(X)$. Медіану інколи називають другим квартилем $\mathrm{med}(X) = Q_2(X)$. Таким чином, квартилі розбивають вибірку на чотири частини приблизно однакового розміру.

Інтерквартильний розмах визначається як відстань від нижнього до верхнього квартиля:

$$\mathrm{IQ}(X) = Q_3(X) - Q_1(X).$$

— це одна з найбільш популярних робастних характеристик розкиду вибірки.

Іще одна робастна характеристика розкиду — медіанне абсолютне відхилення (median absolute deviation):

$$\mathrm{MedAD}(X) = \mathrm{med}(\{|X_j - \mathrm{med}(X)|,\ j = 1\ldots, n\}).$$

Ця характеристика використовується не часто.

Шириною інтервалу, або розмахом вибірки (range), називають

$$\mathrm{Range}(X) = X_{[n]} - X_{[1]},$$

тобто відстань від найменшого до найбільшого значення у вибірці. Зрозуміло, що це найменш стійка до забруднень характеристика розкиду вибірки.

Приклад 4.2.1. Обчислимо статистики розкиду для даних про кількості покупців магазину з табл. 4.1.

Почнемо з дисперсії. Підрахуємо її двома способами — за означенням і, використовучи функцію `var()`[9]:

```
> NumBuy<-c(8,12,23,14,7,92,24) # вводимо дані
> # дисперсія за означенням:
> sum((NumBuy-mean(NumBuy))^2)/6

[1] 898.9048

> # стандартна функція:
> var(NumBuy)

[1] 898.9048
```

Це значення мало що може сказати адміністрації магазину, але для статистика дисперсія даних відіграє велику роль, наприклад, при аналізі точності оцінок, побудові довірчих інтервалів тощо (див. розд. 8).

Як ми побачимо у наступному підрозділі, для характеризації розкиду даних більш природно використовувати корінь з дисперсії — середньоквадратичне відхилення:

```
> sqrt(var(NumBuy))
```

[9] — скорочення англ. *variance* — дисперсія.

```
[1] 29.98174

> sd(NumBuy) # спеціальна функція для с.кв. відхилення

[1] 29.98174
```

Отримали значення ≈30. Можна сказати, що типовий розкид кількості відвідувачів нашого магазину навколо середнього значення 25.7 (див. приклад 4.1.1) становить 30 чоловік. Це дещо дивне твердження: виходить, що типова кількість відвідувачів коливається від $25.7 - 30 = -4.3$ до $25.7 + 30 = 55.7$. Але ж у магазину не може бути -4.3 відвідувачі. Це нетипово!

У даному випадку невдалим результатом ми завдячуємо викиду у суботу: 92 відвідувачі не тільки вплинули на середнє, але збільшили і дисперсію. Причому на дисперсію і середньоквадратичне відхилення викид мав більший вплив, ніж на середнє, тому що піднесення до квадрата виділяє особливо великі значення, робить їх іще більшими порівняно з основною масою спостережень.

Подивимось на робастну характеристику розкиду — інтерквартильний розмах. У першому варіанті розрахунків ми знаходимо медіани верхньої і нижньої половин вибірки[10]. У другому варіанті використовуємо спеціальну функцію `IQR()`:

```
> median(sort(NumBuy)[4:7])-median(sort(NumBuy)[1:4])

[1] 13.5

> IQR(NumBuy)

[1] 13.5
```

[10] У нашій вибірці 7 елементів, тому у "половини" потрапляє по 4 елементи: медіана самої вибірки входить і у "верхню половину" і у "нижню". Це як у дитинстві:

— Тобі яку половину пиріжка?

— Більшу!

— приблизно половина даних розташована в середині варіаційного ряду, лежить у смузі шириною 13.5. Це більш-менш адекватний показник розкиду для наших даних.

Нарешті, ширина діапазону становить

```
> max(NumBuy)-min(NumBuy)

[1] 85
```

— 85 покупців. Цей показник дозволяє побачити, до якої аритмії потоку покупців треба підготуватись персоналу магазину. ◄

4.3 Алгебраїчні властивості описових статистик

Чому одні статистики доцільно використовувати саме для характеризації середнього положення, а інші — для опису розкиду? Які властивості повинна мати статистика, щоб задовільно описувати середнє положення у вибірці? У цьому підрозділі ми спробуємо неформально пояснити це.

Почнемо зі статистик середнього положення.

Приклад 4.3.1. Нехай дані складаються з вимірювань температури плавлення деякої речовини, отриманих у чотирьох експериментах[11]: $-110°$C, $-111°$C, $-117°$C, $-118°$C. Ми хочемо охарактеризувати середнє положення навколо якого коливаються вимірювання. Чи можна використати для цього, скажімо, середнє гармонічне? Воно дорівнює $-113.8903°$C:

```
> x<-c(-110,-111,-117,-118)
> 4/sum(1/x) # середнє гармонійне у шкалі Цельсія:
```

[11] Цифри умовні, але це міг би бути етиловий спирт (етанол). Його температура плавлення за звичайних умов $-114.15°$C.

```
[1] -113.8903
```

△

На перший погляд, результат має природний вигляд. Але згадаєм, що температури можна вимірювати не тільки у градусах Цельсія, а і в інших шкалах, наприклад — Кельвіна (K). Для перерахунку температури з шкали °C у шкалу K треба додати 273.15:

$$t_{\mathrm{K}} = t_{\mathrm{C}} + 273.15.$$

Як повинно змінитись середнє положення у вибірці, якщо ми перейдемо у шкалу K? Природно було б, щоб воно також збільшилось на 273.15, як і кожен елемент вибірки. Але ні:

▽

```
> y<-x+273.15
> y # значення темпереатур у шкалі Кельвіна:

[1] 163.15 162.15 156.15 155.15

> 4/sum(1/y) # середнє гармонічне у шкалі Кельвіна:

[1] 159.0715

> 4/sum(1/y)-4/sum(1/x) # зсув середнього гармонічного:

[1] 272.9618
```

△

— зсув вийшов на 272.9618.

А якщо на роль середнього положення у вибірці використовувати вибіркове середнє, то воно буде зсуватись точно так само, як зсунувся початок відліку на шкалі вимірювання:

▽

```
> sum(y)/4-sum(x)/4

[1] 273.15
```

△

Отже, гармонічне середнє недоцільно використовувати для усереднення даних, які можуть вимірюватись у різних шкалах, що відрізняються положенням початку вимірювання. ◀

Сформулюємо це у загальному випадку. Нехай, для довільної вибірки $\mathbf{X} = (X_1, \ldots, X_n)$, $T(\mathbf{X})$ — деяка статистика (тобто функція від $\mathbf{X}$). Для $c \in \mathbb{R}$ позначимо $\mathbf{X}+c = (X_1+c, \ldots, X_n+c)$.

Статистику T називають **еквіваріантною відносно додавання**[12], якщо для всіх $c \in \mathbb{R}$ і всіх можливих вибірок $\mathbf{X}$:

$$T(\mathbf{X} + c) = T(\mathbf{X}) + c.$$

Інакше кажучи, еквіваріантна статистика зсувається так само, як зсуваються всі елементи вибірки.

Легко перевірити, що вибіркове середнє, зрізане середнє, вибіркова медіана і середина діапазону є статистиками еквіваріантними відносно зсуву. Їх можна використовувати для визначення середнього положення спостережень, які вимірюються "відносно довільно вибраного початку відліку". (Як температура, координати у просторі, астрономічний час тощо)[13].

Геометричне та гармонічне середнє не є еквіваріантними відносно зсуву. Чи випливає з цього, що їх не можна використовувати як статистики середнього положення? Це залежить від того, що має описувати статистика. У прикладі 4.1.2 ми бачили, що геометричне середнє природне при усередненні складних відсотків. Але відсотки — це характеристика, яка має абсолютний початок відліку — 0%. Він відповідає відсутності змін. Якщо початок відліку перенести в іншу точку, відсоток перестане бути відсотком. Отже, у цьому випадку відсутність еквіваріантності не є недоліком.

Крім положення початку шкали відліку можуть відрізнятись також масштабною одиницею. Наприклад, вагу можна вимірювати

[12] Або еквіваріантною відносно зсуву шкали вимірювання.

[13] Шкалу вимірювання, у якої положення 0 можна обирати довільно, називають *інтервальною шкалою*.

у кілограмах, а можна — у тонах. При переході від кілограмів до тон всі спостережувані значення у вибірці зменшаться у 1000 разів. Природно, щоб і значення середнього положення також зменшилось на цей коефіцієнт.

Статистика $T(\mathbf{X})$ називається **еквіваріантною відносно множення**[14], якщо для всіх $a > 0$ і всіх можливих $\mathbf{X}$:

$$T(a\mathbf{X}) = aT(\mathbf{X}).$$

Розглянуті у підрозд. 4.1 статистики еквіваріантні відносно множення. Отже, кожну з них можна використовувати для усереднення даних, які можна вимірювати у різних одиницях вимірювання.

Таким чином, для статистик середнього положення природно вимагати еквіваріантності відносно додавання і множення.

Перейдемо тепер до статистик розкиду. Якщо всі елементи вибірки збільшити вдвічі, то і розкид їх збільшиться вдвічі. Тобто від статистик розкиду природно вимагати еквіваріантності відносно множення. Саме з цих міркувань замість вибіркової дисперсії використовують корінь з неї — вибіркове середньоквадратичне відхилення. Дійсно,

$$S^2(a\mathbf{X}) = a^2 S^2(\mathbf{X})$$

— сталий множник виноситься з під знака дисперсії з квадратом, тобто дисперсія не є еквіваріантною відносно множення. А для середньоквадратичного відхилення $S(a\mathbf{X}) = aS(\mathbf{X})$ при всіх додатних a.

Всі інші характеристики розкиду, розглянуті у підрозд. 4.2, є еквіваріантними відносно множення.

Додавання одного й того числа до всіх спостережень у вибірці на розкид, вочевидь, не впливає. Отже, статистика розкиду T має бути **інваріантною відносно додавання**:

$$T(\mathbf{X} + c) = T(\mathbf{X})$$

[14] Еквіваріантною відносно зміни масштабу.

для всіх $c \in \mathbb{R}$ і всіх можливих вибірок $\mathbf{X}$. Всі статистики розкиду, які ми розглядали, задовольнять цю умову.

4.4 Статистики форми розподілу

Грубо кажучи, все, що не є середнім положенням чи розкидом, можна назвати формою розподілу вибірки. Ми розглянемо тут кілька популярних статистик, що використовуються для опису форми розподілу.

Коефіцієнт варіації для вибірки $\mathbf{X}$ — це середньоквадратичне відхилення, поділене на вибіркове середнє:

$$\mathrm{CV}(\mathbf{X}) = \frac{S(\mathbf{X})}{\bar{\mathbf{X}}}.$$

(Інколи CV вказують у відсотках, тобто домножають на 100). Цю статистику рекомендується використовувати лише для спостережень, які можуть набувати тільки додатних значень. Вона показує, наскільки великим є розкид вибірки порівняно з середнім значенням. Інколи кажуть, що це відносна характеристика розкиду.

Застосування коефіцієнта варіації базується на припущенні про приблизну пропорційність досліджуваних явищ, подібну до тої, яку ми обговорювали, розглядаючи вибіркові середні.

Нехай, наприклад, ми досліджуємо розкид цін на каву певного бренду у різних магазинах однієї країни. При дослідженні США ціни будуть у доларах, при дослідженні Японії — у єнах. Оскільки долар коштує приблизно 100 єн[15], можна сподіватись, що і середні японські ціни в єнах будуть десь у 100 разів вищі. Але природно очікувати, що і розкид японських цін буде більшим, ніж американських: різниця цін у 5 єн, це зовсім не те, що різниця у 5 дол. Якщо відмінності цін цілком визначаються валютним курсом, то японське середньоквадратичне відхилення має бути у 100 разів більшим, ніж американське. При переході від доларів до

[15] 103.1370 за курсом Google 07.10.2016 о 18:03.

єн множник 100 у чисельнику і знаменнику коефіцієнта варіації скоротиться, тобто CV японських і американстких цін мають бути приблизно однаковими.

Звичайно, для порівняння можна було б перерахувати всі ціни в одній валюті і потім порівнювати, наприклад, середньоквадратичні відхилення. Але при такому підході виникає питання про вибір курсу валют (офіційний, біржовий, поточний, середній за останній рік. . .). Використання CV для порівняння знімає такі питання "стандартизації шкали", оскільки CV — безрозмірна величина.

Таким чином, близькість коефіцієнтів розсіювання цін для різних країн — очікуване явище, а от при виявленні відмінностей це могло б свідчити про якісь специфічні відмінності економік відповідних країн. З аналогічних міркувань коефіцієнти варіації використовують у кліматології, наприклад, при аналізі коливань річкових стоків.

Зауважимо, що $\mathrm{CV}(\mathbf{X})$ не змінюється при зміні масштабу вимірювання (є інваріантною відносно множення), але змінюється при зміні початку відліку шкали. Тому, наприклад, CV температур, виміряних у градусах Цельсія, буде відрізнятись від CV тих самих температур у градусах Кельвіна. Відповідно, застосовувати CV для опису таких даних слід обережно.

Коефіцієнти варіації часто використовують для порівняння точності різних методів вимірювання медико-біологічних показників стану організму, психометричних шкал, що характеризують особистість тощо. Наприклад, однією з загальних характеристик роботи імунної системи організму є швидкість осідання еритроцитів (ШОЕ). Існують різні методики його вимірювання: методи Вестергрена (стандартна) та Вінтробе базуються на вимірюванні глибини, на яку осідають еритроцити у капілярі протягом одної години, автоматичні системи, які використовують ценрифугування (дають результат за 5 хв) та ін. При аналізі роботи таких методів одну порцію крові розділяють на кілька зразків, які аналізують окремо. Після цього визначають коефіцієнт варіації значень ШОС, отриманих за цими зразками. Чим менший CV для даного методу, тим

метод точніший. Такий підхід дозволяє порівнювати точність безпосередніх вимірювань без зведення їх до єдиної стандартизованої шкали.

Асиметрія (skewness) вибірки $\mathbf{X} = (X_1, \ldots, X_n)$ визначається як

$$\gamma_1(\mathbf{X}) = \frac{1}{S(\mathbf{X})^3} \frac{1}{n} \sum_{j=1}^{n} (X_j - \bar{X})^3.$$

Ця величина показує, наскільки симетрично розташовані вибіркові дані навколо свого середнього значення. Якщо симетрія ідеальна (для кожного j знайдеться таке i, що $X_j - \bar{X} = \bar{X} - X_i$) то асиметрія дорівнює 0.

Нормування на $S(\mathbf{X})^3$ введено для того, щоб асиметрія була інваріантною відносно масштабування. Крім того, завдяки відніманню $\bar{X}$ вона є також інваріантною відносно вибору початку відліку. В цьому розумінні асиметрія є типовим представником характеристик форми розподілу: їх зазвичай визначають так, щоб вони не змінювались при лінійних змінах шкали вимірювання. При такому підході характеристики форми доповнюють інформацію про розподіл вибіркових даних, яку дають характеристики середнього положення та розкиду.

4.5 Групування та навантаження

Групування. Серед даних у вибірці можуть зустрічатись однакові значення. Якщо різних значень, яких набуває змінна, порівняно небагато, і більшість з них зустрічається у вибірці кілька разів, то зручно не виписувати всю вибірку, а перелічити ці різні значення і вказати їхні частоти (кількість повторень). Запис вибірки у такому вигляді називають *групуванням*, а саму вибірку — *згрупованою* (grouped).

Нехай $x_1 < \cdots < x_K$ — всі різні значення, яких може набувати досліджувана змінна (варіанти), K — кількість цих значень.

Абсолютна частота n_i варіанти x_i у вибірці $X = (X_1, \ldots, X_n)$ — це кількість номерів[16] $j = 1, \ldots, n$, для яких $X_j = x_i$.

Груповані дані можна записувати у вигляді таблиці, яку називають рядом розподілу вибірки:

Варіанти	x_1	x_2	$\ldots$	x_K
Частоти	n_1	n_2	$\ldots$	n_K

"Груповані вибірки" природно виникають, наприклад, тоді, коли досліджувана величина є цілочисловою за своїм змістом. Скажімо, це може бути кількість бракованих виробів, виявлених на контролі протягом одного дня виробництва. У інших випадках групування виникає внаслідок обмеженої точності вимірювання досліджуваних величин: якщо довжину комах вимірювати лінійкою, на якій є лише міліметрові поділки, то результат вимірювання у міліметрах буде цілим числом, хоча справжні довжини можуть набувати будь-яких додатних значень.

Нарешті, інколи виникає потреба[17] провести "примусове групування" (*grouping*, або *binning*[18]), коли дані спеціально огрублюють. У цьому випадку весь інтервал $[a, b]$ можливих значень змінної розбивають на K підінтервалів $A_k = [t_{k-1}, t_k)$, де $a = t_0 < t_1 < \cdots < t_K = b$ — деякі точки. (Напр., при рівномірному розбитті беруть $t_k = a + kh$, де $h = (b - a)/K$ — ширина інтервалу розбиття). Якщо спостережуване значення X_j потрапляє в інтервал A_k, його заміняють на значення середини цього інтервалу $x_k = (t_k + t_{k-1})/2$. Утворену огрублену вибірку групують. Зрозуміло, що в такій вибірці n_k — це кількість тих спостережень, які потрапили в інтервал A_k.

Статистики групованих вибірок.

Вибіркове середнє за групованою вибіркою можна обчислити

[16] Кількість об'єктів у вибірці

[17] Наприклад, при побудові гістограм, див. підрозд. 7.1, або при застосуванні тестів типу χ^2, див. підрозд. 9.6.

[18] інтервали розбиття називають *bins* — кошики, відповідно, примусово групована вибірку — *binned sample*.

так:

$$\bar{X}_w = \frac{1}{n}\sum_{j=1}^{n} X_j = \frac{1}{n}\sum_{k=1}^{K} n_k x_k = \sum_{k=1}^{K} w_k x_k,$$

де $w_k = n_k/n$ — вагові коефіцієнти[19].

Аналогічно, середнє геометричне обчислюється за формулою

$$\mathrm{GM}_w(X) = \left(\prod_{k=1}^{K} (x_k)^{n_k}\right)^{1/n} = \prod_{k=1}^{K} (x_k)^{w_k}.$$

Дещо складніше визначити медіану групованої вибірки. Для цього треба знайти таке значення k, для якого $\sum_{i:x_i<x_k} w_i < 1/2$ і $\sum_{i:x_i>x_k} w_i \le 1/2$. Тоді x_k буде вибірковою медіаною групованої вибірки. (При такому означенні вибіркова медіана, обчислена за примусово групованою вибіркою, може трохи відрізнятись від медіани, обчисленої без групування).

Груповане середнє абсолютне відхилення можна обчислювати як

$$\mathrm{MAD}_w(X) = \sum_{k=1}^{K} w_k |x_k - \bar{X}_w|.$$

Групована вибіркова дисперсія має вигляд

$$S_w^2(X) = \sum_{k=1}^{K} w_k (x_k - \bar{X}_w)^2.$$

Якщо вибірка є природно групованою (наприклад, коли змінна набуває лише цілочислових значень) то $S_w^2(X)$ це в те ж саме, що звичайна вибіркова дисперсія. Але, якщо дані були огрублені примусовим групуванням, то групована дисперсія є огрубленням справжньої вибіркової. Якщо розбиття при групуванні було рівномірним, можна ввести спеціальну поправку, яка дозволяє більш

[19] Навантаження, ваги, weights.

точно наблизити справжню дисперсію:

$$S^2_{corr}(X) = S^2_w(X) + \frac{h^2}{12}, \tag{4.7}$$

де h — ширина інтервалу розбиття. Величина $h^2/12$ називається *поправкою Шеппарда* (*Sheppard's correction*).

Навантажені статистики. Навантажені середні вигляду

$$\bar{X}_w = \sum_k w_k X_k$$

природно використовувати не тільки для групованих даних. Такі суми виникають при аналізі інших статистичних даних.

Приклад 4.5.1. *Нормою прибутку підприємства* називають прибуток, отриманий ним протягом року, ділений на обсяг капіталу, інвестованого у це підприємство. Нехай результатом спостережень n підприємств є розмір прибутку p_j та розмір інвестованого капіталу c_j для j-го обстеженого підприємства ($j = 1, \ldots, n$). Як визначити середню норму прибутку цих підприємств?

Можливі два варіанти. По-перше, можна обчислити норми прибутку по кожному підприємству окремо:

$$r_j = \frac{p_j}{c_j}$$

й усереднити їх, отримавши

$$\bar{r} = \frac{1}{n} \sum_{j=1}^{n} r_j.$$

По-друге, можна знайти сумарний прибуток всіх обстежених підприємств одразу і розділити його на сумарний капітал цих підприємств:

$$\bar{r}_w = \frac{\sum_{j=1}^n p_j}{\sum_{j=1}^n c_j} = \sum_{j=1}^{n} w_j r_j,$$

де

$$w_j = \frac{c_j}{\sum_{i=1}^{n} c_i}.$$

Тобто у цьому випадку ми отримали навантажене середнє з ваговими коефіцієнтами, пропорційними капіталам підприємств. Це і зрозуміло, можна сподіватись, що чим більший капітал підприємства, тим більшим повинен бути його внесок у економіку, отже, при підсумовуванні його варто враховувати з більшою вагою.

Який варіант середнього є більш "правильним" для цих даних? Відповідь залежить від задачі, яка стоїть перед дослідником. Якщо дослідження проводиться, наприклад, для міністерства фінансів, яке хоче оцінити можливий майбутній прибуток підприємств країни залежно від вкладених інвестицій, то слід орієнтуватись на навантажене середнє. Якщо дослідження виконується для фіскальної служби, яка має на меті виявити підприємства з аномальними значеннями норми прибутку, то, можливо, більш правильним орієнтиром нормального підприємства буде просте вибіркове середнє. А можливо, для визначення середнього положення норми прибутку у цьому випадку краще скористатись вибірковою медіаною. ◀

Існує багато інших задач, у яких природним буде застосування навантажених середніх. Для повного розуміння того, чому у цих задачах навантаження набуває певної форми, треба описати відповідні дані певними ймовірнісними моделями, які обговорюються пізніше. Тому тут ми лише побіжно згадаємо найбільш поширені варіанти навантажень.

Приклад 4.5.2. Нехай проводяться вимірювання однієї і тієї ж фізичної величини різними приладами. Позначимо X_j — результат вимірювання j-м приладом, $j = 1, \ldots, n$. Точність вимірювань різна у різних приладів. Дисперсія похибки[20] j-го приладу дорівнює σ_j^2. Можна довести, що у цьому випадку найбільш точною

[20]Маємо на увазі дисперсію, вказану у паспорті приладу, яка характеризує точність вимірювань цим приладом, визначену при його сертифікації.

оцінкою справжнього значення вимірюваної величини[21] є

$$\bar{X}_\sigma = \frac{1}{\sum_{j=1}^{n} 1/\sigma_j^2} \sum_{j=1}^{n} \frac{X_j}{\sigma_j^2} = \sum_{j=1}^{n} w_j X_j,$$

де

$$w_j = \frac{1/\sigma_j^2}{\sum_{i=1}^{n} 1/\sigma_i^2}.$$

Інтуїтивний зміст цієї формули зрозумілий: чим більша дисперсія похибки, тим менша точність відповідного вимірювання, тому спостереження, що мають більші дисперсії, включаються у сумарну оцінку з меншими коефіцієнтами. ◀

Приклад 4.5.3. Нехай досліджувані об'єкти мають різні шанси потрапити до вибірки, причому ці шанси пов'язані з характеристикою, що досліджується. Такі вибірки називають зміщеними.

Наприклад, об'єктами можуть бути риби, виловлені у ставку, а характеристикою — довжина рибини. Чим більшою є рибина, тим більше у неї шансів потрапити до рибальської сітки. Якщо метою дослідження є оцінювання середньої довжини риб у ставку, то середнє довжин виловлених риб буде завищеною оцінкою цієї характеристики. Тобто оцінка за зміщеною вибіркою є зміщеною.

Для виправлення вибіркового зміщення використовують навантажені середні з ваговими коефіцієнтами, обернено пропорційними ймовірності того, що дане спостереження потрапить до вибірки. Такі вагові коефіцієнти називають *коефіцієнтами Горвіца — Томпсона*. ◀

Окрім навантажених середніх можуть використовуватись також інші навантажені статистики, такі як навантажена медіана або навантажена дисперсія. Формули для цих статистик використовуються такі ж, як наведено вище для групованих даних, але вагові коефіцієнти мають інший зміст.

[21] Це оцінка методу найбільшої вірогідності у випадку нормального розподілу даних, див. приклад 8.3.3. Її дисперсія найменша в класі всіх лінійних незміщених оцінок.

Інколи змістовні вагові коефіцієнти у формулах для навантажених статистик не є нормованими, тобто їх сума не дорівнює 1. У такому випадку нормують саму статистику. Наприклад, якщо $\sum_{j=1}^{n} w_j \neq 1$, то навантажене вибіркове середнє слід обчислювати за формулою

$$\bar{X} = \frac{1}{\sum_{j=1}^{n} w_j} \sum_{j=1}^{n} w_j X_j,$$

а навантажене геометричне середнє — за формулою

$$\mathrm{GM}(X) = \left(\prod_{j=1}^{n} (X_j)^{w_j} \right)^{1/\sum_{j=1}^{n} w_j}.$$

4.6 Обчислення описових статистик у R

Обчислення більшості основних описових статистик у R реалізовано у вигляді функцій однотипної структури. Для менш поширених статистик часто можна написати просту команду, яка їх обчислює. Зведення за цими функціями наведено у табл. 4.2.

Таблиця 4.2. Функції для обчислення описових статистик

Статистика	Позначення	Функція
Вибіркове середнє	$\bar{X}$	`mean(x)`
Геометричне середнє	$\mathrm{GM}(X)$	`prod(x)^(1/length(x))`
Гармонійне середнє	$\mathrm{HM}(X)$	`1/mean(1/x)`
Зрізане середнє	$\mathrm{TM}_{2a}(X)$	`mean(x,trim=a)`
Медіана	$\mathrm{med}(X)$	`median(x)`
Виправлена вибіркова дисперсія	$S_0^2(X)$	`var(x)`
Середньоквадратичне відхилення	$S_0(X)$	`sd(x)`
Середнє абсолютне відхилення	$\mathrm{MAD}(X)$	`mean(abs(x-mean(x)))`
Інтерквартильний розмах	$\mathrm{IQ}(X)$	`IQR(x)`

У всіх цих функцій першим параметром `x` є вибірка, за якою обчислюється відповідна статистика. Цей параметр може бути числовим вектором або матрицею. В обох випадках результатом виконання функції є одне число — значення відповідної статистики розраховане за всіма елементами `x`. Винятком із цього правила є функція `var`. Якщо її аргументом `x` є матриця, вона обчислює матрицю вибіркових коваріацій[22] для стовпчиків `x`.

Наприклад:

```
> x=cbind(1:3,4:6)
> x

     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

[22]Означення вибіркових коваріацій і коваріаційної матриці див. підрозд. 5.2.

```
> mean(x)

[1] 3.5

> sd(x)

[1] 1.870829

> var(x)

     [,1] [,2]
[1,]    1    1
[2,]    1    1
```

Інколи необхідно обчислити середні значення по кожному стовпчику матриці окремо. Це можна зробити, використовуючи функцію `colMeans()`. Результатом розрахунку є іменований вектор значень статистик для всіх змінних. Функцію `colMeans()` можна також застосовувати до фреймів даних:

```
>  x<-c(1:5)
> y<-rep(3,5)
> z<-x<y
> dat<-data.frame(x,y,z)
> colMeans(dat)

  x   y   z
3.0 3.0 0.4
```

(звичайна функція `mean()` видає повідомлення про помилку при застосуванні до фреймів). Якщо на змінних фрейма треба розрахувати якісь інші статистики крім середнього, можна скористатись функцією `apply()` — див. приклад 4.6.1.

У всіх розглянутих функцій є також логічний параметр-опція `na.rm`. Якщо вказати `na.rm=T`, то перед розрахунком статистики

з вибірки будуть вилучатись всі пропущені значення (статистика розраховується тільки за не пропущеними). Стандартно `na.rm=F`, у цьому випадку за наявності пропущених значень у вибірці значенням функції також буде `NA`.

У R є функція `mad()`. Але вона обчислює не MAD(X) — середнє абсолютне відхилення, як можна було б сподіватись, а медіанне абсолютне відхилення — MedAD(X) у наших позначеннях. У цій функції є параметр `constant` — константа, на яку домножається розраховане медіанне абсолютне відхилення. Стандартне значення `constant=1.4826`. При використанні такого множника `mad(x)`, буде консистентною оцінкою для середньоквадратичного відхилення вибірки з нормальним розподілом. Якщо потрібне справжнє значення MedAD(X), слід задати `constant=1`.

У стандартній поставці R немає окремих функцій для обчислення навантажених статистик. Але більшість з них неважко записати, безпосередньо використовуючи формули для їхнього обчислення:

```
> x<-1:5 # вибірка
> w<-c(2,2,2,1,1) # вагові коефіцієнти
> #
> # навантажене вибіркове середнє:
> sum(x*w)/sum(w)

[1] 2.625

> # навантажене гармонічне середнє:
> (prod(x^w))^(1/sum(w))

[1] 2.27597
```

Складніше запрограмувати навантажену медіану. Якщо вагові коефіцієнти набувають лише цілих значень, це можна зробити відтворивши початкову вибірку за навантаженою і застосувавши звичайну функцію `median()`:

```
> x<-1:5 # вибірка
> w<-c(2,2,2,1,1) # вагові коефіцієнти
> #
> median(rep(x,w)) # навантажена медіана
[1] 2.5
```

Тут функція `rep(x,w)` кожен елемент x_j вибірки `x` розмножила w_j разів. Після цього `median()` обчислила медіану цієї розмноженої вибірки. Можна сказати, що ми трактували нашу вибірку як груповану і за цією групованою вибіркою відновили початкову (із повторами).

Зрозуміло, що такий спосіб розрахунку навантаженої медіани є дуже неефективним, особливо коли вагові коефіцієнти великі. У пакеті `laeken` є функція , котра обчислює навантажену медіану при довільних вагових коефіцієнтах.

Часто описові статистики використовуються, коли треба порівняти багато вибірок однотипних даних. Якщо ці вибірки зібрані у матрицю, то виникає потреба розраховувати статистики окремо для кожного стовпчика (або рядка) матриці. Це можна зробити, використовуючи функцію `apply()` так, як описано у п. 2.2.5.

Приклад 4.6.1. Наприклад, змінна `fm` містить значення концентрацій формальдегіду (мг на м^3) в атмосферному повітрі, виміряні на Бесарабській площі міста Києва у різні години доби (о 1-й, 7-й, 13-й і 19-й годинах) за період з 15 по 21 жовтня 2015 р. (дані з сайта ЦГО України `http://www.cgo.kiev.ua/`). Рядок матриці відповідає одній добі спостережень, стовпчик — певній годині доби. Нас може цікавити, наскільки змінюються концентрації протягом доби і наскільки вони змінюються при вимірюванні у певний час у різні дні спостережень. Вибравши на роль характеристики розкиду середньоквадратичне відхилення, обчислимо його по кожному рядку і кожному стовпчику. Для цього спочатку запишемо числові дані у змінну `fm`, а потім застосуємо функцію `sd()` за допомогою функції `apply()`:

```
> # Концентрації формальдегіду по днях
> d15=c(0.005,0.008,0.010,0.005)
> d16=c(0.004,0.005,0.015,0.008)
> d17=c(0.004,0.010,0.012,0.009)
> d18=c(NA,NA,NA,NA)
> d19=c(0.008,0.011,0.014,0.015)
> d20=c(0.009,0.011,0.014,0.007)
> d21=c(0.007,0.009,NA,NA)
> # Створюємо матрицю концентрацій:
> fm=rbind(d15,d16,d17,d18,d19,d20,d21)
> colnames(fm)<-(c("t01","t07","t13","t19"))
> apply(fm,1,sd,na.rm=T)

        d15         d16         d17         d18
0.002449490 0.004966555 0.003403430          NA
        d19         d20         d21
0.003162278 0.002986079 0.001414214

> apply(fm,2,sd,na.rm=T)

        t01         t07         t13         t19
0.002136976 0.002280351 0.002000000 0.003768289
```

Бачимо, що у різні дні та у різних місцях розкид даних може відрізнятись вдвічі. ◀

Як параметр `fun` у функції `apply()` можна використовувати не тільки ім'я функції, а і її опис. Наприклад, якщо за даними `fm` треба підрахувати гармонічні середні за кожну добу спостережень, це можна зробити так:

```
apply(fm,1,function(x)(prod(x)^(1/length(x))))
```

Інколи буває, що всі дані для аналізу зібрані в одному фреймі, причому досліджувана характеристика є однією зі змінних цього

фрейма. Розбиття на окремі підвибірки треба зробити за іншими змінними-факторами, що характеризують приналежність досліджуваних об'єктів до різних груп. У таких випадках зручно використовувати функцію . Вона призначена для застосування певної функції-статистики окремо до кожної підвибірки, заданої комбінацією певних факторів. Значенням функції є таблиця значень статистики для всіх можливих комбінацій факторів.

Приклад 4.6.2. У фреймі даних `ToothGrowth` містяться дані про дослідження впливу різних дієт на швидкість росту зубів у свиней. Всього у фреймі 60 спостережень, кожне відповідає одній свині. Змінна `len` вказує довжину зубів, `sup` — харчову добавку, яку використовували для внесення у раціон свині вітаміну C (VC — хімічна аскорбінова кислота, OJ — помаранчовий сік), `dose` — щоденна доза вітаміну, яку отримувала свиня із цією добавкою (лише три варіанти доз: 0.5, 1 або 2 мг). Нас цікавить, як відрізняються середні значення та середньоквадратичні відхилення `len` при різних комбінаціях факторів `sup` і `dose`.

```
> # Таблиця вибіркових середніх:
> tapply(ToothGrowth$len,
+        list(ToothGrowth$supp,ToothGrowth$dose),mean)

     0.5     1     2
OJ 13.23 22.70 26.06
VC  7.98 16.77 26.14

> # Таблиця середньоквадратичних відхилень:
> tapply(ToothGrowth$len,
+        list(ToothGrowth$supp,ToothGrowth$dose),sd)

        0.5        1        2
OJ 4.459709 3.910953 2.655058
VC 2.746634 2.515309 4.797731
```

◀

Розділ 5

Опис залежностей

У попередньому розділі ми розглянули дані, в яких кожному спостереженню відповідає одна числова величина — змінна. У більшості прикладних статистичних дослідженнях кожен спостережуваний об'єкт характеризується кількома різними змінними, причому досліднику треба описати залежність між цими змінними. Такі задачі вивчає статистика багатовимірних даних. У цьому розділі ми розглянемо багатовимірну описову статистику, тобто будемо намагатись досліджувати залежності, не виходячи з якоїсь теоретичної моделі, а намагаючись виявити внутрішню структуру даних. Дослідження залежностей між числовими змінними на основі регресійних моделей розглядається далі у розд. 10. Метод перевірки залежності між двома змінними не числової природи описано у п. 9.6.4.

5.1 Діаграми розсіювання

Нехай є фрейм (набір) даних, в якому містяться значення різних числових змінних (характеристик) для n об'єктів[1]. Перше,

[1] У фреймі можуть бути також і дані не числової природи, але зараз нас цікавить, у першу чергу, робота з числовими змінними.

з чого варто почати дослідження — це подивитись на дані. Якщо об'єктів багато, проглядання числової таблиці мало що допомагає.

Значно краще дають уявлення про особливості даних рисунки, на яких кожному об'єкту відповідає точка на площині з координатами, визначеними певними змінними цього об'єкта. Такі рисунки у статистиці називають **діаграмами розсіювання** (англ. *scatterplot*).

Проста діаграма розсіювання утворюється, коли з двох змінних, що описують об'єкт, перша відкладається по горизонталі, а друга — по вертикалі.

Приклад 5.1.1. Фрейм даних `faithful`, що входить у стандартну поставку `R`, містить дані про виверження одного гейзера з Йелоустонського парку у США[2]. У даних для 272 послідовних вивержень гейзера записано тривалість виверження (змінна `eruptions`) і тривалість інтервалу до наступного виверження (змінна `waiting`). Нас може цікавити, чи є залежність між цими змінними.

Побудуємо діаграму розсіювання, відкладаючи `eruptions` по горизонталі, а `waiting` — по вертикалі[3] (рис. 5.1):

```
> # Діаграма розсіювання:
> plot(faithful$waiting~faithful$eruptions,
+      xlab="erruptions",        ylab="waiting")
> #         Підгонка прямою:
>         abline(lm(faithful$waiting~faithful$eruptions),
>          col="red")
```

[2]Цей гейзер має назву Old Faithful geyser. Він є одним з найбільших та популярніших у парку.

[3]Відображення точок на площині функцією `plot()` описано у підрозд. 3.2. Опис функції `lm()` (підгонка рівняння прямої, шо описує залежність між змінними "в середньому", за методом найменших квадратів) відкладемо на потім (див. підрозд. 10.2.).

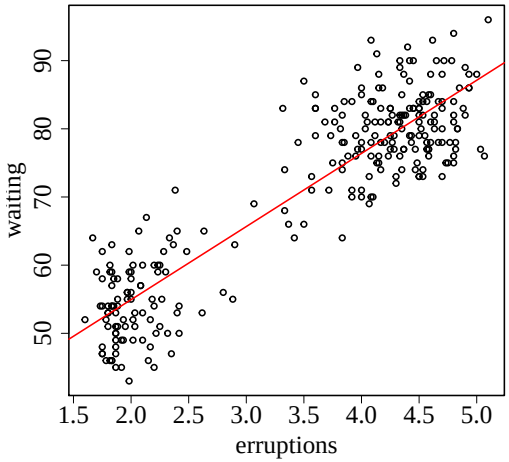

Рис. 5.1. Діаграма розсіювання Old Faithful geyser

На рис. 5.1 видно, що строгої залежності між досліджуваними змінними немає, але більшим значенням `waiting` відповідають в середньому більші `erruptions`. Це приклад статистичної залежності між змінними. Точки на діаграмі розташовуються навколо червоної прямої, тобто наша залежність “приблизно лінійна”, не видно особливостей, які вимагали б застосування нелінійних моделей для опису залежності.

На рис. 5.1 також помітно, що точки досить чітко розділяються на дві групи — одна ліворуч внизу, друга — праворуч вгорі. Між ними лежать кілька точок, які важко віднести до певної групи. Але більшість класифікується однозначно. Такі окремі групи спостережень називають **кластерами**. Цікаво, що проведена нами пряма досить добре відображає залежність між змінними в обох кластерах одразу, хоча б на перший погляд. Наскільки можна довіряти цьому першому погляду, ми обговоримо пізніше, у прикладі 10.4.3.

◀

Якщо треба проаналізувати більше ніж дві змінні, використо-

вуються так звані **матричні діаграми розсіювання** (*matrix plot*).

На матричній діаграмі зображаються попарні діаграми розсіювання всіх пар числових змінних. Діаграми розташовуються на рисунку у вигляді матриці, кожен рядок якої відповідає одній змінній відкладеній по вертикалі, а кожен стовпчик — змінній по горизонталі. Назва відповідної змінної вказується у діагональному елементі матриці.

Найпростіше відобразити матричну діаграму просто викликавши функцію `plot()` з першим параметром — фреймом даних, які треба відобразити. При цьому кожна пара відображається на двох діаграмах, симетричних відносно діагоналі матриці. Якщо вам досить однієї діаграми на кожну пару, можна залишити лише частину матриці, яка розташована над діагоналлю. У наступному прикладі показано, як це робиться за допомогою функції `pairs()`

Приклад 5.1.2. Розглянемо знову дані про розміри квітів-півників (Iris) з прикладу 2.2.1. У фреймі `iris` містяться дані про 150 квіток. Для кожної квітки перші чотири змінні є числовими (позначають різні розміри), а п'ята (`Species`) вказує один з трьох видів півників, якому належить дана квітка. Зобразимо відповідну матричну діаграму (рис. 5.2):

```
> # вибираємо кольори для рисування різних видів:
> color<-c("red","green","blue")
> # рисуємо діаграму перших чотирьох змінних iris:
> pairs(iris[,1:4],col=color[as.numeric(iris[,5])],
+       lower.panel = NULL,
+       pch=as.numeric(iris[,5]))
> # Наступна команда дозволяє рисувати легенду
> # за межами основного рисунка:
> par(xpd=TRUE)
> # рисуємо легенду:
> legend(0.02,0.4,title="Species:",
+ legend=c('setosa', 'versicolor', 'virginica'),
+ pch=c(1,2,3), # символи для різних видів
```

```
+ col=color)
```

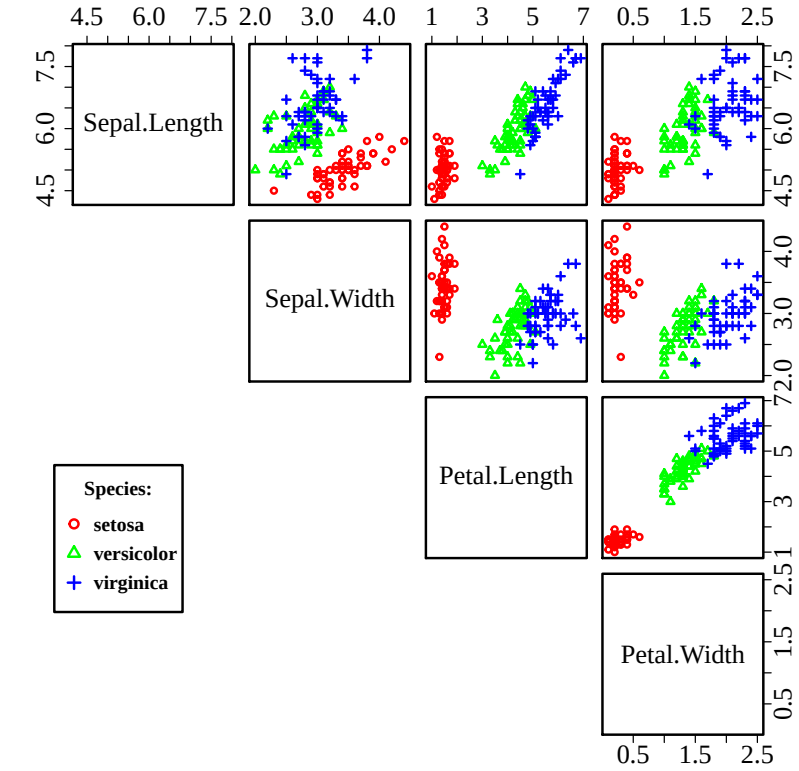

Рис. 5.2. Матрична діаграма для квітів-ірисів

На діаграмах розсіювання різними кольорами та різними символами відмічені точки, що відповідають квітам різних видів. Ліворуч внизу виведено легенду, що пояснює, яким видам відповідають ці символи і кольори. Тепер одразу можна бачити, що червоні кола, що відповідають квітам виду setosa, скупчились у окремий кластер-хмарку, яка помітно відрізняється за характеристиками від інших

видів. Хмарки точок для видів versicolor (зелені трикутники) і virginica (сині хрестики) частково перекриваються, але також явно відрізняються за розташуванням. Отже, залежность між числовими характеристиками різних видів ірисів краще досліджувати окремо, не об'єднуючи їх в одну вибірку.

Наприклад, на діаграмі у третьому прямокутнику у верхньому рядку по горизонталі відкладено довжину пелюсток (змінна `Petal.Length`), а по вертикалі — довжину чашолистків (змінна `Sepal.Length`). Хмари точок для versicolor і virginica помітно витягнуті вздовж прямих, що проходять під помітним кутом до горизонталі. Це свідчить про залежність між цими характеристиками у даних видів півників. Ніяких особливостей, що вимагали б нелінійної моделі для опису такої залежності непомітно. (Виразні нелінійності на діаграмах розсіювання можна побачити на рис. 10.14 і 10.22). Не помітні і викиди, тобто окремі точки, які відхилялися б від основної маси спостережень. (Приклади діаграм розсіювання з викидами — на рис. 10.6 і 10.8).

За хмарою точок для setosa жодної виразної залежності між `Petal.Length` і `Sepal.Length` не помітно. Можливо, нам заважають її побачити точки для інших видів, через які setosa скупчились у лівому нижньому кутку рисунка? Нарисуємо їх на окремій діаграмі розсіювання (рис. 5.3):

```
> setosa<-iris[iris$Species=="setosa",]
> plot(setosa$Sepal.Length~setosa$Petal.Length,
+ xlab="petal length",ylab="sepal length")
> abline(lm(setosa$Sepal.Length~setosa$Petal.Length),
+        col="red")
```

Червона пряма на рис. 5.3 відповідає залежності середньої довжини пелюстки від значення довжини чашолистків. Точки показують значення цих характеристик для конкретних квіток. Бачимо, що розкид довжини пелюстки навколо середнього значення значно переважає зміну цього середнього під впливом зміни довжини

чашолистка. (Рухаючись вздовж червоної лінії зліва направо ми підніземося значно менше, ніж розкидані точки навколо). Отже, якщо залежність між змінними і є, вона мало помітна на фоні їх розкиду. Строго горизонтальна лінія відповідала б повній відсутності залежності. Спостережуване незначне відхилення від горизонтальності може бути наслідком випадкових коливань. Для перевірки цього слід застосувати відповідний статистичний тест, про що див. далі у прикладі 5.5.1.

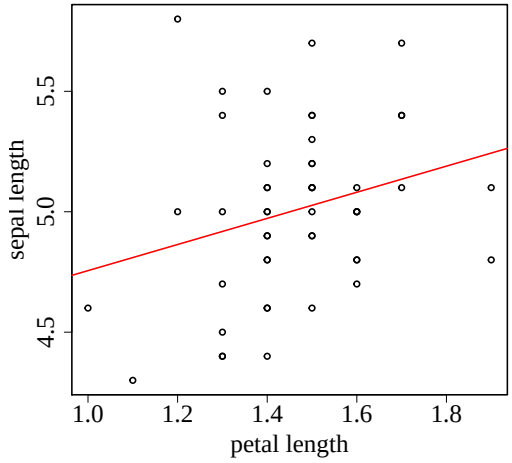

Рис. 5.3. Діаграма розсіювання для ірисів виду setosa

◀

Матричні діаграми можна будувати не тільки з використанням `plot()` і `pairs()`. Спеціальні функції для цього є у різних бібліотеках `R`. Вони дозволяють виводити більше важливої інформації на таких діаграмах. Застосування функції `scatterplotMatrix()` з бібліотеки `car` розглядається у прикладі 10.2.1.

Матричні діаграми розумно використовувати, якщо треба дослідити залежність невеликої кількості змінних — трьох-п'яти, в

усякому випадку, не більше десяти. Коли досліджуються дані з десятками або сотнями змінних, доцільно спочатку спробувати вибрати серед них найбільш залежні, а вже потім аналізувати залежність між вибраними. Для цього потрібна якась проста числова характеристика залежності двох змінних. Такими характеристиками є коефіцієнти кореляції, що розглядаються у наступних підрозділах.

5.2 Коефіцієнт кореляції Пірсона

Найбільш популярною мірою залежності між двома змінними є (парний) **коефіцієнт кореляції Пірсона** (*Pearson correlation*)[4]. У цьому підрозділі ми спочатку дамо його формальне означення, коротко опишемо основні властивості, потім покажемо, як він обчислюється в R, а вже потім спробуємо пояснити, яку саме залежність він описує.

Означення коваріації і кореляції. Нехай для кожного спостережуваного об'єкта вимірюються дві числові характеристики-змінні — X і Y. Позначимо X_j, Y_j — значення цих характеристик для j-го об'єкта у вибірці ($j = 1, \ldots, n$).

Вибірковою коваріацією (*sample covariance*) змінних X і Y називають величину

$$\operatorname{cov}(X, Y) = \frac{1}{n} \sum_{j=1}^{n} (X_j - \bar{X})(Y_j - \bar{Y}),$$

де $\bar{X}$, $\bar{Y}$ — середні значення відповідних змінних, n — обсяг вибірки.

[4]Існує багато інших коефіцієнтів кореляції, але, якщо цей термін вживається без додаткових означень або просто кажуть "кореляція дорівнює", то, у більшості випадків, йдеться саме про парний коефіцієнт кореляції Пірсона.

Виправленою вибірковою коваріацією називають[5]

$$\operatorname{cov}_0(X,Y) = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})(Y_j - \bar{Y}),$$

Коефіцієнтом кореляції Пірсона називають

$$r(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{S^2(X)S^2(Y)}} = \frac{\operatorname{cov}_0(X,Y)}{\sqrt{S_0^2(X)S_0^2(Y)}},$$

де $S^2(X)$, $S^2(Y)$ — вибіркові дисперсії X і Y, $S_0^2(X)$, $S_0^2(Y)$ — виправлені вибіркові дисперсії (див. (4.6)).

Властивості коефіцієнтів кореляції Пірсона.

1. Коефіцієнт кореляції за абсолютним значенням не перевищує одиницю:

$$|r(X,Y)| \le 1.$$

2. Коефіцієнт кореляції дорівнює ± 1 тоді і тільки тоді, коли між X і Y у вибірці існує строга лінійна залежність, тобто є такі числа b_0, b_1, що

$$Y_j = b_0 + b_1 X_j.$$

(У цьому випадку всі точки на діаграмі розсіювання будуть лежати на одній прямій). Знак $r(X,Y)$ відповідає знаку b_1.

3. Коефіцієнт кореляції не змінюється при лінійній зростаючій зміні шкали вимірювання X і Y: якщо, наприклад, розглянути $X'_j = a_1 X_j + a_0$, $a_1 > 0$, то $r(X',Y) = r(X,Y)$. (Можна сказати, що $r(X,Y)$ є еквіваріантним відносно додавання і множення по обох своїх аргументах — пор. підрозд. 4.3). (Якщо $a_1 < 0$, кореляція змінює знак: $r(X',Y) = -r(X,Y)$).

4. Якщо змінні X і Y незалежні, то, при великих обсягах вибірки, $r(X,Y)$ буде близьким до 0. Незалежність тут треба розуміти у статистичному значенні: вважається, що Y і X — незалежні, якщо знання X ніяк не допомагає прогнозувати значення Y.

[5]Виправлена вибіркова коваріація є незміщеною оцінкою для теоретичної коваріації за кратною вибіркою.

Зауважимо, що коефіцієнт кореляції Пірсона не є робастною статистикою: одне забруднення-викид може сильно змінити його. Тому для аналізу залежності за забрудненими даними краще користуватись іншими статистиками.

Обчислення кореляцій. У R коефіцієнт кореляції Пірсона між змінними X і Y можна обчислювати використовуючи функцію `cor()`. Вектори значень X і Y при цьому вказують як перший і другий аргументи функції.

Приклад 5.2.1. У прикладі 3.4.4 ми розглянули коливання інтересу до джинсових і карго-шортів у різних штатах США. З порівняння на географічних картах виявилось, що перевага інтересу до джинсових шортів може бути пов'язана з низьким рівнем урбанізації штату. Подивимось, як корельовані ці показники.

Дані візьмемо з файла `shortU.txt`, де у змінній `urban` містяться рівні урбанізації штатів, у змінній `jean` — кількість гуглзапитів на джинсові шорти, у `cargo` — кількість запитів на карго-шорти у даному штаті.

Для порівняння з `urban` створимо змінну

`x<-jean/(jean+cargo)`

— частка інтересу до джинсових шортів у загальному інтересі до шортів. Обчислимо коефіцієнт кореляції:

```
> tb<-read.table("c:/rem/term/shortU.txt",header=T)
> x<-tb$jean/(tb$jean+tb$cargo)
> cor(x,tb$urban)

[1] -0.3234017
```

Отримали кореляцію $r = -0.3234017$. Це невелике значення, але помітно відмінне від 0. Воно від'ємне, тобто зростання урбанізації штату в середньому приводить до зменшення інтересу до джинсових шортів.

Чи можна впевнено сказати, що ми встановили наявність залежності між інтересом до джинсових шортів і урбанізацією? Ні. Таке

значення коефіцієнта кореляції могло б бути наслідком випадкових коливань у даних, що не пов'язані зі справжньою залежністю. Наскільки це ймовірно? Ми повернемось до цього питання у прикладах 10.1.2 і 10.2.3. ◀

Часто кожен об'єкт у вибірці описується не двома, а значно більшою кількістю змінних. У цьому випадку для аналізу залежностей знаходять попарні кореляції для всіх можливих пар змінних. Отримані значення записують у вигляді таблиці, яку називають **вибірковою кореляційною матрицею**.

Позначимо змінні, що описують об'єкт $X^1, \dots X^d$. Тоді кореляційна матриця цього набору змінних — це матриця складена з попарних коефіцієнтів кореляції Пірсона[6]:

$$\mathbf{R} = \begin{pmatrix} r(X^1, X^1) & r(X^1, X^2) & \dots & r(X^1, X^d) \\ r(X^2, X^1) & r(X^2, X^2) & \dots & r(X^2, X^d) \\ \vdots & \vdots & \ddots & \vdots \\ r(X^d, X^1) & r(X^d, X^2) & \dots & r(X^d, X^d) \end{pmatrix}$$

Для обчислення кореляційної матриці також можна використати функцію `cor()`, причому досить передати їй один параметр — фрейм даних, для змінних якого розраховуються кореляції.

Приклад 5.2.2. Обчислимо кореляційну матрицю за даними про розміри квітів-півників з прикладу 5.1.2. У даному прикладі обмежимось обчисленням для квітів виду Setosa.

```
> cor(iris[iris$Species=="setosa",1:4])

             Sepal.Length Sepal.Width Petal.Length
Sepal.Length    1.0000000   0.7425467    0.2671758
Sepal.Width     0.7425467   1.0000000    0.1777000
Petal.Length    0.2671758   0.1777000    1.0000000
Petal.Width     0.2780984   0.2327520    0.3316300
```

[6] Аналогічну матрицю, складену з попарних коваріацій, називають *вибірковою коваріаційною матрицею*.

```
             Petal.Width
Sepal.Length   0.2780984
Sepal.Width    0.2327520
Petal.Length   0.3316300
Petal.Width    1.0000000
```

△

З таблиці бачимо, що всі кореляції додатні — при зростанні одного з вимірів квітки, в середньому, зростають і інші. (Цього і слід було сподіватись, півники — досить пропорційні квіти). Найбільш корельованими виявились ширина і довжина чашолистка — кореляція 0.7425467. (Такий рівень кореляції свідчить про чітко виражену залежність). Для ширини і довжини пелюстки кореляція вдвічі менша — 0.3316300. (Подивившись на квітку півника, легко зрозуміти — чому). Кореляції між характеристиками чашолистка і характеристиками пелюсток іще менші, хоча однозначно назвати їх незначущими не можна. ◄

Інколи буває необхідно обчислити попарні кореляції всіх змінних одного набору з усіма змінними іншого. Це також можна зробити, використовуючи функцію `cor()`. Наприклад,

```
cor(iris[iris$Species=="setosa",1:2],
    iris[iris$Species=="setosa",3:4])
```

підраховує кореляції між характеристиками чашолистків і характеристиками пелюсток Iris Setosa (перевірте).

Трактовка кореляцій. Розберемось тепер, як саме кореляція Пірсона характеризує залежність між змінними. Ми розглянемо три трактовки $r(X, Y)$.

1. Якщо треба характеризувати залежність між двома змінними X_j, Y_j, що описують об'єкти у вибірці, то природно обрати міру залежності так, щоб вона не змінювалась при зміні одиниць вимірювання цих змінних. Тобто міра залежності має бути інваріантною відносно множення. Аналогічно, природно вимагати і інваріантності відносно додавання. Тому зручно перейти від почa-

ткових змінних до нормованих:

$$\tilde{X}_j = \frac{X_j - \bar{X}}{S(X)},\ \tilde{Y}_j = \frac{Y_j - \bar{Y}}{S(Y)}.$$

(Ми відняли від кожного значення середнє за всіма спостереженнями і розділили на стандартне відхилення. Можна сказати, що ми вибрали шкалу вимірювання наших змінних так, щоб у ній вони мали нульове середнє й одиничну дисперсію).

Міру залежності можна тепер визначати спеціально для нормованих змінних. Для цього спробуємо спрогнозувати $\tilde{Y}_j$, використовуючи як прогноз bX_j, де b — число, яке нам треба підібрати так, щоб прогноз вийшов найточнішим. Точність будемо вимірювати середнім квадратів відхилень прогнозу від справжніх значень, тобто шукатимем таке b, при якому

$$\frac{1}{n}\sum_{j=1}^{n}(\tilde{Y}_j - b\tilde{X}_j)^2 = 1 - 2br(\tilde{X}, \tilde{Y}) + b^2$$

буде найменшою. (Нагадаємо, що n — кількість спостережень у вибірці).

Можна показати, що мінімум досягається якраз при
$b = r(\tilde{X}, \tilde{Y}) = r(X, Y)$.

Отже коефіцієнт кореляції Пірсона між двома змінними — це найкращий коефіцієнт пропорційності для прогнозування однієї нормованої змінної за допомогою другої. Якщо, наприклад, $r(X, Y) = 0.5$, то при зростанні $\tilde{X}$ в 6 разів, $\tilde{Y}$ зросте приблизно втричі.

2. Нехай ми використовуємо безпосередньо X_j для прогнозування Y_j за лінійною формулою

$$Y_j \approx b_1 X_j + b_0.$$

Підгонку коефіцієнтів прогнозу зробимо за методом найменших квадратів (див. підрозд. 10.1). Отримаємо прогноз

$$\hat{Y}_j = \hat{b}_1 X_j + \hat{b}_0,$$

де $\hat{b}_0$, $\hat{b}_1$ — підігнані значення коефіцієнтів.

Точність отриманого прогнозу характеризується відношенням дисперсії його помилок до дисперсії прогнозованої змінної:

$$\frac{S^2(Y-\hat{Y})}{S^2(Y)} = \frac{\sum_{j=1}^{n}(Y_j-\hat{Y}_j)^2}{\sum_{j=1}^{n}(Y_j-\bar{Y})^2} = 1-(r(X,Y))^2.$$

Таким чином, квадрат коефіцієнта кореляції характеризує точність лінійного прогнозу однієї змінної за другою. Чим ближчий $(r(X,Y))^2$ до 1, тим менший розкид помилки прогнозу порівняно з розкидом прогнозованої змінної.

3. Розглянемо центровані вектори даних $\mathbf{X}' = (X_1', \ldots, X_n')$, $\mathbf{Y} = (Y_1', \ldots, Y_n')$, де

$$X_j' = X_j - \bar{X},\ Y_j' = Y_j - \bar{Y}.$$

(Нормувати дисперсіями не будемо). Тоді

$$r(X,Y) = \frac{\langle \mathbf{X}', \mathbf{Y}' \rangle}{\|\mathbf{X}'\| \cdot \|\mathbf{Y}'\|} = \cos(\alpha),$$

де $\langle \mathbf{X}', \mathbf{Y}' \rangle$ — скалярний добуток векторів $\mathbf{X}'$, $\mathbf{Y}'$ у n-вимірному просторі $\mathbb{R}^n$, $\|\cdot\|$ позначає довжину (евклідову норму) вектора в $\mathbb{R}^n$, α — кут між векторами $\mathbf{X}'$ і $\mathbf{Y}'$ в $\mathbb{R}^n$.

Отже, коефіцієнт кореляції визначає кут між центрованими векторами змінних у просторі спостережень. $r(X,Y)=0$ відповідає ортогональності (перпендикулярності) векторів $\mathbf{X}'$ і $\mathbf{Y}'$. Значення $r(X,Y)=\pm 1$ свідчать про те, що вектори $\mathbf{X}'$ і $\mathbf{Y}'$ колінеарні.

Як бачимо, всі розглянуті трактування коефіцієнта кореляції Пірсона пов'язані з розглядом залежності "схожої на лінійну". І дійсно, якщо між змінними є сильна нелінійна залежність, кореляція Пірсона може її не помітити.

Приклад 5.2.3. Нехай X_j — арифметична прогресія від -1 до 1 з n елементів, а $Y_j = (X_j)^2$. Вочевидь, між X і Y є строга функціональна залежність. Але $r(X,Y) = 0$ для будь-якого n. Наприклад:

```
> n<-1000
> X<-seq(from=-1,to=1,length.out =n)
> Y<-X^2
> cor(X,Y)

[1] -3.261627e-17
```

З цього не треба робити висновок, що кореляція Пірсона зовсім не бачить нелінійних залежностей. Але для аналізу таких залежностей її треба застосовувати обережно.

Ми повернемось до розгляду цього питання у п. 9.7.3. ◄

5.3 Візуалізація кореляцій

Якщо треба дослідити залежність між багатьма різними змінними, то аналіз таблиці попарних кореляцій (кореляційної матриці) стає нетривіальною проблемою. Для цього зручно користуватись спеціальними методами відображення кореляцій на відповідних рисунках. Ми розглянемо два призначені для цього засоби R — карту кореляцій і кореляційну мережу.

Карта кореляцій — це зображення кореляцій на площині клітинками різного кольору. Для цього зручно використовувати функцію `corrplot()` з бібліотеки `corrplot`. Як це робиться ми покажемо у наступних прикладах.

Приклад 5.3.1. У стандартну поставку R входить фрейм даних `mtcars`, що містить дані про характеристики 32 моделей автомобілів[7], такі як кількість циліндрів (`cyl`) та потужність двигуна (`hp`), кількість карбюраторів (`carb`) тощо. Для того, щоб проаналізувати залежності між цими змінними, обчислимо матрицю попарних кореляцій Пірсона і відобразимо їх карту, використовуючи `corrplot()`. Результат відображення — на рис. 5.4.

[7] Дані досить старі, тому робити з них висновки про сучасні моделі не варто.

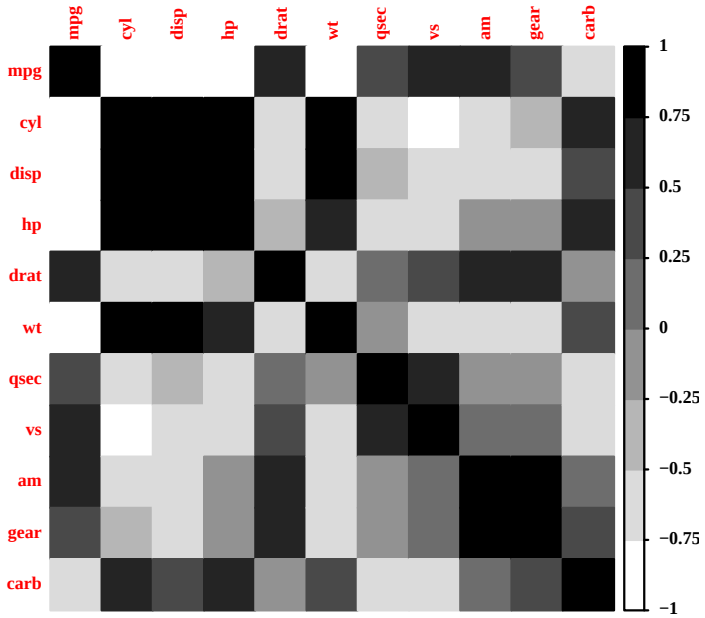

Рис. 5.4. Карта кореляцій фрейма mtcar

Рисунок 5.4 відображений таким скриптом:

```
>  library(corrplot)
>  M <- cor(mtcars)
>  corrplot(M, method="color"
+           ,cl.cex=1,cl.align.text="l",#*
+           col=gray((7:0)/7)            #*
+           )
```

На цьому рисунку кожній парі змінних відповідає квадратик. Колір, яким він зафарбований, вибрано залежно від кореляції. Шка-

ла відповіності кольорів та кореляцій відображена праворуч від основної діаграми. На ній чорний колір відповідає кореляціям, близьким до 1, білий — близьким до -1. Такий варіант визначений опцією `col=gray((7:0)/7)`. При роботі на комп'ютері краще застосовувати різнокольорову шкалу. Для цього вилучіть з наведеного вище скрипта рядки, відмічені `#*`.

Ви отримаєте рисунок, аналогічний рис. 5.4, але тепер додатній кореляції відповідає синій, від'ємній — брунатний. Чим сильніша кореляція, тим насиченіший колір. На рисунку видно, наприклад, що змінні `disp`, `cyl` і `hp` утворюють групу з сильною додатною кореляцією між собою. А змінні `qsec`, `vs`, `am` і `gear` — негативно корельовані з цією групою. Змінна `mpg` сильно негативно корельована з `disp`– `hp` і помірно додатно корельована з групою `qsec`–`gear` і т.д. Для спеціаліста з дизайну автомобілів це може бути корисною інформацією.

При цьому рисунок є симетричним відносно головної діагоналі, а на самій діагоналі завжди стоять клітинки, що відповідають 1: кореляція змінної з собою — абсолютна. Тому для аналізу досить виводити лише трикутник над (або під) діагоналлю. Це дозволяє об'єднати графічну інформацію з числовою, як це зроблено у такому прикладі за допомогою функції `corrplot.mixed()` (рис. 5.5):

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

Тут кружечки над діагоналлю відтворюють кореляції не тільки насиченістю кольору, а й розміром. А під діагоналлю записані числові значення кореляцій. Використовуючи опції `upper` і `lower` можна замовити функції `corrplot.mixed()` виводити над та під діагоналлю числа, кружечки, квадратики, еліпси. На мій погляд, така картинка інформативніша, ніж попередня.

Однак прихильники рис. 5.4 скажуть, що на ньому краще помітні великі сині квадрати навколо діагоналі, що відповідають групам (кластерам) сильно корельованих змінних. Так, але це дає змогу виділити групу лише, якщо змінні, які до неї входять, розташовані

у фреймі поруч. Наприклад, `mpg` природно об'єднати в одну групу з `disp`, `cyl` і `hp`, але у фреймі стовпчик `mpg` розташований далеко від інших змінних групи, тому помітити це за рисунком не так просто.

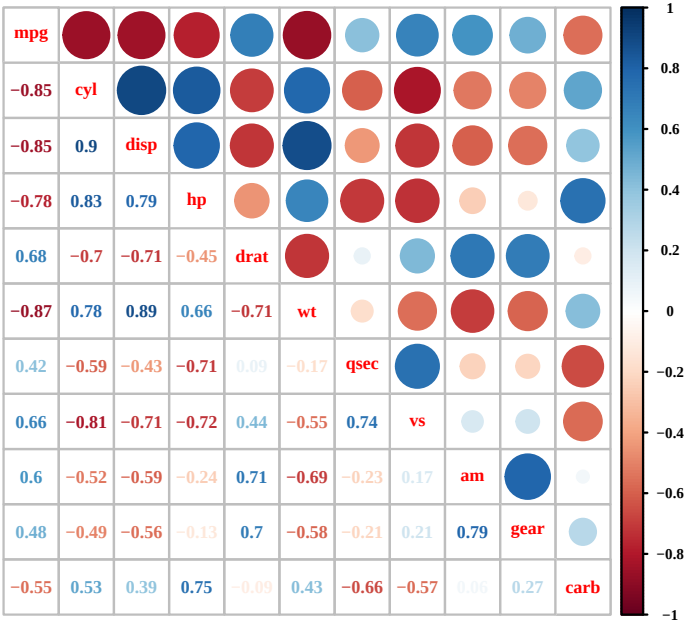

Рис. 5.5. Мішаний графік кореляцій для фрейма mtcar

Щоб спростити цю задачу, у функціях `corrplot.mixed()` і `corrplot()` можна змінити порядок, в якому змінні відображаються на діаграмі. Це робиться опцією `order`. Значення цієї опції `"original"` (як у фреймі) та `"alphabet"` (у алфавітному порядку) тривіальні. Але можна задати також значення `"AOE"`, `"FPC"`, `"hclust"`, що відповідають різним евристичним процедурам по-

шуку найбільш відповідного порядку змінних, щоб сильно корельовані розташовувались поруч. Перші два значення відповідають алгоритмам, що використовують спектральний розклад коваріаційної матриці, третій — `"hclust"` безпосередньо шукає кластери змінних методом ієрархічної кластеризації. Користувач може перепробувати їх всі й обрати той, який дасть найбільш переконливий результат.

При використанні `order="hclust"` у `corrplot()` можна вказати опцію

`addrect=`*кількість кластерів*

— тоді на рисунку буде виділена відповідна кількість кластерів, знайдених методом ієрархічної кластеризації.

У нашому прикладі з даними `mtcars` це можна зробити так:

```
> corrplot.mixed(M,order="hclust")
> corrplot(M,order="hclust",addrect=3)
```

(результат на рис. 5.6). Якщо задати `addrect=2`, то два нижні кластери на нижньому рисунку об'єднаються в один великий квадрат. Бачимо, що змінні досить впевнено розбились на два (можливо, три) кластери, всередині яких вони майже всі позитивно корельовані між собою. А кореляції між кластерами переважно негативні. ◀

Приклад 5.3.2. Повернемось до даних про квіти-півники з фрейма `iris`, що розглядались у прикладі 5.1.2. Щоб порівняти відмінності кореляцій між розмірами квітки у квітів різних видів використайте такий скрипт:

```
>  Spec<-c("setosa","versicolor","virginica")
>  for(sp in Spec){
+   corrplot(cor(iris[iris$Species==sp,1:4]),
+   type="upper",cl.pos="n",diag=F); title(sub=sp)
+                 }
```

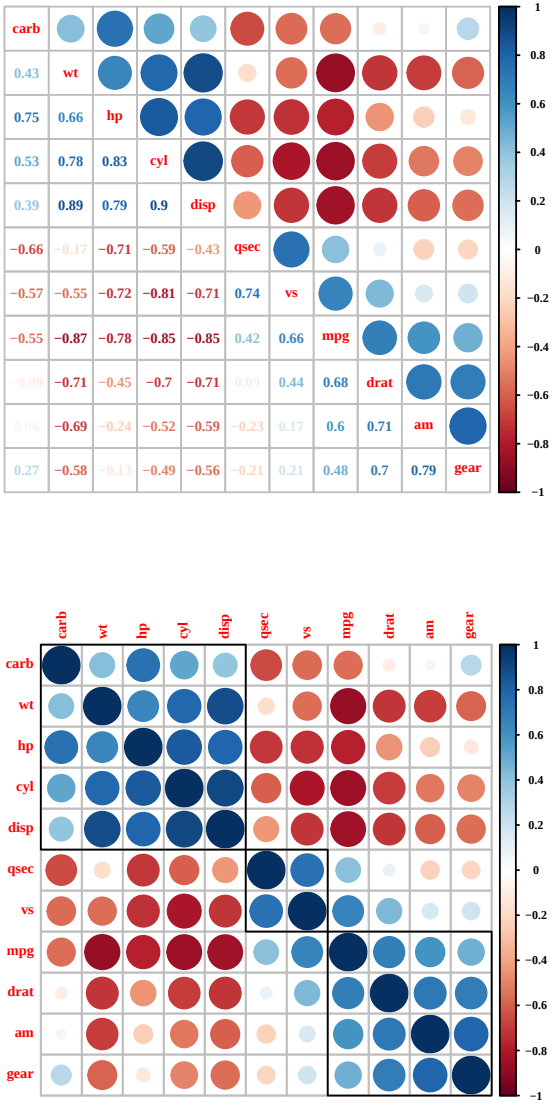

Рис. 5.6. Мішаний графік кореляцій для фрейма mtcar з перестановкою змінних та виділенням кластерів

На отриманому рисунку можна буде побачити, що кореляції для видів versicolor і virginica досить схожі та помітно відрізняються від кореляцій для setosa.

◀

Кореляційні мережі (*correlation networks*). Інший спосіб відобразити кореляції — це граф, в якому кожній змінній відповідає одна вершина, а величина та знак кореляції передаються кольором і товщиною ребер. Такі графи називають *кореляційними мережами*.

Для зображення кореляційної мережі можна скористатись функцією `qgraph()` з бібліотеки `qgraph`. Цій функції треба передати кореляційну матрицю, що буде відображатись. Якщо не вказувати інших параметрів, то на графі вершини будуть розташовані по колу, в тому ж порядку, в якому змінні йдуть у фреймі даних. Можна також задати свої власні значення для координат вершин.

Приклад 5.3.3. Кореляції для даних `mtcars`, що розглядались у прикладі 5.3.1, можна відобразити за допомогою наступного скрипта:

```
> library("qgraph")
>          # рахуємо кореляції для даних про автомобілі:
> M <- cor(mtcars)
> qgraph(M,negDashed=TRUE) # виводимо верхній рисунок
> #
> # список номерів груп:
> grp<-as.factor(c(3,1,1,1,3,1,2,2,3,3,1))
> # виводимо нижній рисунок:
> qgraph(M,negDashed=TRUE,layout="groups",
+        groups=grp,minimum=0.5)
```

Результат — на рис. 5.7). Як бачимо, функція `qgraph()` скоротила назви змінних до трьох літер. Це необхідно для того, щоб вони вміщувались у кружечках-вершинах графа. Якщо вам не подобаються автоматичні скорочення, можна задати власні імена вершин у опції `nodeNames`.

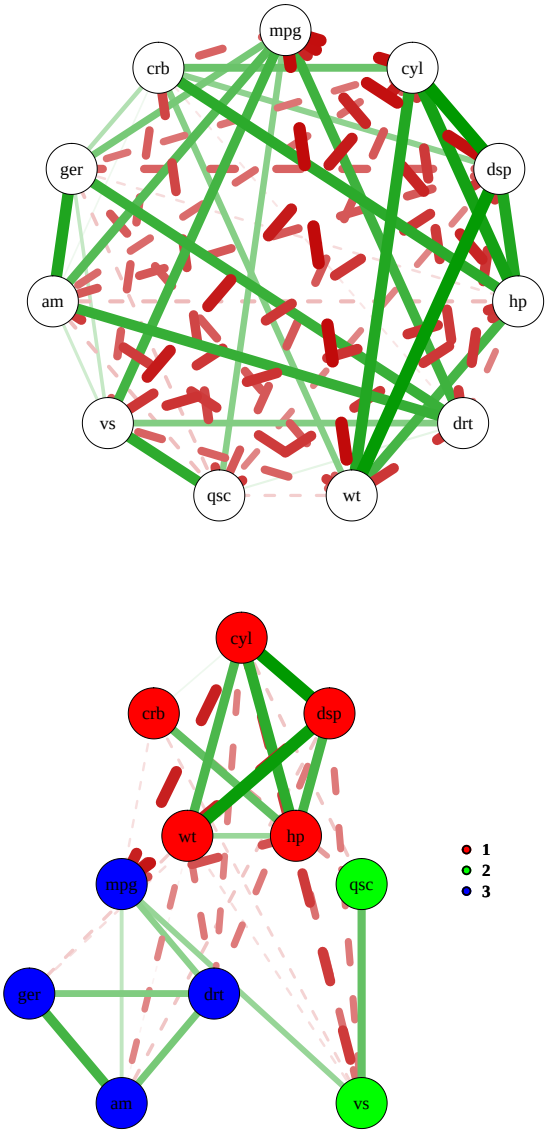

Рис. 5.7. Кореляційна мережа для фрейма mtcar

На рис. 5.7 зверху — граф, який `qgraph()` створює автоматично, якщо їй передати лише матрицю кореляцій. Розібратись у ньому досить важко. (На кольоровому рисунку додатним кореляціям відповідають зелені ребра на графі, від'ємним — брунатні. Для кращого сприйняття чорно-білого варіанту ми задали опцію `negDashed=TRUE`, яка вказує, що ребра між від'ємно корельованими змінними відображаються переривчастими лініями.)

Для зручності сприйняття доцільно згрупувати вершини-змінні приблизно так, як ми зробили на рис 5.6. Там ми виділили три групи, всередині яких майже всі кореляції додатні, а кореляції між групами — від'ємні. Для відображення такого групованого графа функцією `qgraph()` можна задати опції:

`layout="groups"` — яка вказує, що граф буде розбито на групи, причому елементи кожної групи знову розташовуються по колу;

`groups` — склад груп: це може бути вектор факторів, які відповідають різним групам, або список кількох цілочислових векторів, де у кожному векторі вказуються номери змінних, що належать відповідній групі.

На рис. 5.7 (знизу) ми задали три групи, що відповідають групам з рис. 5.6. Крім того, для зручності сприйняття, на цьому графі відображаються лише "сильні" кореляції, тобто такі, які за абсолютною величиною перевищують 0.5. Це зроблено за допомогою опції `minimum`.

У опції `layout` можна також задати безпосередньо координати вершин на площині у вигляді матриці з двома стовпчиками (координати по горизонталі і вертикалі), кожен рядок якої відповідає одній вершині. Іще одне можливе значення цієї опції `layout="spring"` відповідає автоматичному вибору положень вершин у такий спосіб, щоб структура графа найкраще сприймалась візуально (використовується алгоритм Фрухтермана — Рейнголда [28]). Читач може самостійно перевірити, який результат дасть виконання команди

```
qgraph(M,layout="spring",minimum=0.5)
```

◀

Рисунки з кореляційною мережею значно компактніші, ніж карти кореляцій, які ми розглянули вище. Однак розібратись в них може бути важче, хоча деякі риси системи залежностей він дозволяє відобразити більш опукло.

Приклад 5.3.4. Скрипт для порівняння кореляцій довжин трьох видів квітів-півників за допомогою кореляційних мереж можна записати так:

```
>     library("qgraph")
> Spec<-c("setosa","versicolor","virginica")
> for(sp in Spec){
+ qgraph(cor(iris[iris$Species==sp,1:4]))
+ text(-0.75,1,sp,cex=1.5)
+ }
```

Результат — на рис. 5.8. ◄

5.4 Ранги та рангові кореляції

Як ми бачили у підрозд. 5.2, коефіцієнт кореляції Пірсона призначений для виявлення залежностей, подібних до лінійних. Якщо залежність, котру намагаються виявити, не є лінійною, використання цього коефіцієнта може бути недоречним. Крім того, r Пірсона не є робастною характеристикою: один викид може різко змінити висновки, зроблені на основі цього коефіцієнта.

Тому досить часто для аналізу залежностей використовують інші коефіцієнти кореляції, що базуються на рангах спостережень. Такі коефіцієнти називають *ранговими*. Далі ми спочатку введемо поняття рангу, а потім розглянемо два відповідних коефіцієнта: ρ Спірмена і τ Кендалла.

Ранги. Нехай для об'єктів у вибірці спостерігається деяка числова характеристика (змінна) X. Позначимо X_j — значення цієї характеристики у j-того об'єкта.

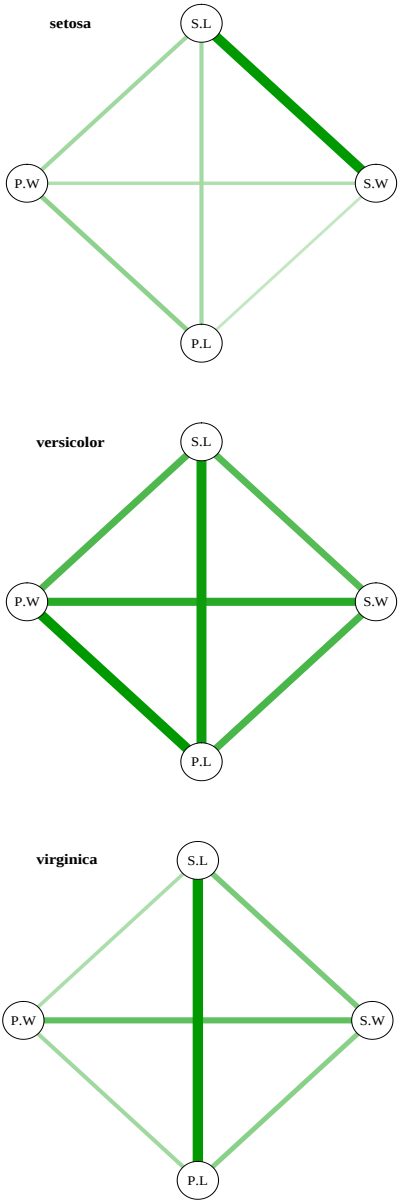

Рис. 5.8. Порівняння кореляцій розмірів для півників

Спочатку введемо означення рангів для випадку, коли всі спостережувані значення X_j — різні.

Рангом R_j^X j-го об'єкта щодо змінної X називають номер цього об'єкта у вибірці, впорядкованій за зростанням X.

Наприклад, нехай спостерігаються такі значення:

j	1	2	3	4	5
X	3.5	-1.2	4.8	1.2	0

Після перестановки в порядку зростання отримаємо такий варіаційний ряд:

-1.2; 0; 1.2; 3.5; 4.8.

У ньому елемент, який у початковій невпорядкованій вибірці мав номер 1, опинився на 3-му місці. Отже, $R_1^X = 3$. Аналогічно, $R_2^X = 1$, $R_3^X = 5$.

Якщо у вибірці присутні кілька елементів з однаковими значеннями змінної X, то при перестановці у порядку зростання їх можна поставити на різні місця. У цьому випадку кажуть, що їх ранги пов'язані (англ *tied ranks*). Зазвичай, пов'язані ранги замінюють на їхнє середнє значення.

Наприклад, нехай спостерігається така вибірка.

j	1	2	3	4	5
X	0	2	0	2	0

У цій вибірці перший, третій і п'ятий елементи — пов'язані. Їх можна розташувати на перших трьох місцях у варіаційному ряді в довільному порядку. Тобто їх ранги мали б бути 1, 2, 3. Середнє цих рангів — 2. Тому всім об'єктам з номерами 1, 3, 5 присвоюють ранг 2. Аналогічно, другому та четвертому об'єктам присвоюють ранг $4.5 = (4 + 5)/2$. Остаточно маємо наступну таблицю рангів.

j	1	2	3	4	5
R_j^X	2	4.5	2	4.5	2

Якщо для кожного об'єкта у вибірці спостерігаються декілька змінних, він може мати різні ранги за різними змінними.

ρ **Спірмена.** Нехай для кожного з n об'єктів у вибірці спостерігаються змінні X і Y. Ранговий коефіцієнт кореляції Спірмена (англ. *Spearman's rank correlation*) визначається як коефіцієнт ко-

реляції Пірсона між рангами спостережень:

$$\rho(X.Y) = r(R^X, R^Y) = \frac{\sum_{j=1}^n (R_j^X - \bar{R}^X)(R_j^Y - \bar{R}^Y)}{\sqrt{\sum_{j=1}^n (R_j^X - \bar{R}^X)^2 \sum_{j=1}^n (R_j^Y - \bar{R}^Y)}}, \tag{5.1}$$

де $\bar{R}^X$, $\bar{R}^Y$ — середні значення рангів за X і Y по всій вибірці.

Якщо пов'язані ранги відсутні, то R_j^X, при $j = 1, \ldots, n$, пробігають всі цілі значення від 1 до n по одному разу. Тому у цьому випадку $\bar{R}^X = \bar{R}^Y = (n+1)/2$ і

$$\sum_{j=1}^n (R_j^X - \bar{R}^X)^2 = \sum_{j=1}^n (R_j^Y - \bar{R}^Y)^2 = \frac{(n-1)n(n+1)}{12}.$$

Використовуючи цей факт легко отримати, що за відсутності пов'язаних рангів,

$$\rho(X.Y) = 1 - \frac{6 \sum_{j=1}^n (R_j^X - R_j^Y)}{n(n^2-1)}. \tag{5.2}$$

Ця формула більш популярна, ніж (5.1), але вона не дає правильних результатів, якщо є пов'язані ранги. У цьому випадку для обчислення ρ слід користуватись (5.1).

У `R` для обчислення коефіцієнта Спірмена можна скористатись функцією `cor` (див. підрозд. 5.2), вказавши в ній опцію

`method = "spearman"`.

Приклад 5.4.1. У прикладі 5.2.1 ми використали коефіцієнт кореляції Пірсона для вимірювання рівня залежності між інтересом до джинсових шортів і рівнем урбанізації у різних штатах США. Скористаємось тепер для цього коефіцієнтом ρ Спірмена.

▽

```
> tb<-read.table("c:/rem/term/shortU.txt",
>                header=T)
>  x<-tb$jean/(tb$jean+tb$cargo)
>  cor(x,tb$urban,method = "spearman")
```

```
[1] -0.4387847
```

△

◀

τ **Кенделла** (англ. *Kendall's tau coefficient*). Як показує формула (5.2), у коефіцієнті ρ Спірмена підсумовуються квадрати різниць між рангами об'єкта за першою та другою змінними. Чим більші ці різниці, тим більше ρ. Кореляція Кендалла побудована на іншій ідеї: вона тим більша, чим більше пар досліджуваних об'єктів розташовано в однаковому порядку по першій і другій змінній.

Точніше, нехай пов'язані ранги відсутні. Переберемо всі пари індексів (i, j), де $1 \le i < j \le n$. Будемо казати, що пара (i, j) узгоджена за змінними X та Y, якщо

$$(X_i - X_j)(Y_i - Y_j) > 0,$$

тобто за порядком зростання X елементи цієї пари розташовані так само, як і за порядком зростання Y. Якщо порядок за X протилежний порядку за Y, тобто

$$(X_i - X_j)(Y_i - Y_j) < 0,$$

будемо казати, що пара (i, j) не узгоджена.

Позначимо кількість усіх узгоджених пар n_+, а всіх неузгоджених — n_-. Тоді $n_+ + n_- = n(n-1)/2 = n_0$ — кількість всіх можливих пар індексів.

Кореляція Кендалла між X і Y (за відсутності зв'язаних рангів) визначається як

$$\tau(X, Y) = \frac{n_+ - n_-}{n_0},$$

тобто це нормована різниця між кількістю узгоджених і неузгоджених пар. Нормування (ділення на n_0) обрано так, щоб значення $\tau(X, Y)$ за абсолютною величиною не перевищувало 1.

Існує багато узагальнень коефіцієнта Кендалла на випадок пов'язаних рангів. Ми обмежимось одним з них (найбільш популярним), який позначають τ_b (читається "тау-бе Кендалла").

Отже, нехай у вибірці наявні пов'язані ранги. Позначимо $t_1,\ldots,$ t_k — кількості елементів у групах з пов'язаними рангами по змінній X. (Тобто у вибірці наявні точно t_1 елементів у яких X набуває одне і те ж фіксоване значення, t_2 елементів, у яких X має інше фіксоване значення і т.д.) $u_1,\ldots,u_m$ — аналогічно, по змінній Y.

$$n_1 = \sum_{i=1}^{k} t_i(t_i-1)/2,$$

$$n_2 = \sum_{i=1}^{k} u_i(u_i-1)/2,$$

n_+, n_- — кількості узгоджених та неузгоджених пар (при цьому, якщо ранги по хоча б одній змінній у парі пов'язані, така пара не враховується ні серед узгоджених, ні серед неузгоджених), $n_0 = n(n-1)/2$. Тоді

$$\tau_b(X,Y) = \frac{n_+ - n_-}{\sqrt{(n_0-n_1)(n_0-n_2)}}.$$

(Нормування знову обрано так, щоб максимальне і мінімальне значення коефіцієнта кореляції дорівнювали ± 1).

Для обчислення τ у `R` можна скористатись функцією `cor()`, вказавши опцію `method="kendall"`

Приклад 5.4.2. Продовжуючи приклад прикладі 5.4.1 обчислимо коефіцієнт кореляції Кендала для залежності між інтересом до джинсових шортів і рівнем урбанізації у різних штатах США.

```
> cor(x,tb$urban,method = "kendall")

[1] -0.2869042
```

◀

Таким чином, для цього прикладу ми отримали кореляцію Пірсона $r = -0.3234017$, кореляцію Спірмена $\rho = -0.4387847$ і кореляцію Кендалла $\tau = -0.2869042$.

Звичайно, виникає питання, яку з цих кореляцій доцільно обрати для характеризації залежності між рівнем урбанізації штату та зацікавленням джинсовими шортами серед його жителів?

Взагалі, як правильно обирати коефіцієнт кореляції? Для того, щоб зрозуміти це, порівняємо їх властивості. Для кореляції Пірсона основні властивості перелічені на с. 173.

Властивості рангових коефіцієнтів кореляції. Наступні властивості однакові для кореляцій Спірмена і Кендалла.

1. Коефіцієнт кореляції за абсолютним значенням не перевищує одиницю:

$$-1 \leq \rho(X,Y), \tau(X,Y) \leq 1.$$

2. Ранговий коефіцієнт кореляції дорівнює ± 1 тоді і тільки тоді, коли залежність між X і Y у вибірці є монотонною. Якщо коефіцієнт дорівнює 1, ця залежність — зростаюча, тобто з $X_i < X_j$ випливає $Y_i \leq Y_j$. Якщо коефіцієнт дорівнює -1, залежність — спадна: з $X_i < X_j$ випливає $Y_i \geq Y_j$.

3. Рангові коефіцієнти кореляції не змінюються при монотонній зміні шкали вимірювання X і Y: якщо, наприклад, розглянути $X'_j = f(X_j)$, де f — строго зростаюча функція, то $\rho(X',Y) = \rho(X,Y)$ і $\tau(X',Y) = \tau(X,Y)$. (Якщо f — монотонно спадна функція, кореляція змінить знак).

Наприклад, рангові коефіцієнти не змінюються при переході до логарифмічної шкали вимірювання.

4. Якщо змінні X і Y незалежні, то при великих обсягах вибірки коефіцієнт кореляції буде близьким до 0. Незалежність тут треба розуміти у статистичному значенні: вважається, що Y і X — незалежні, якщо знання X ніяк не допомагає прогнозувати значення Y[8].

5. Коефіцієнти Спірмена і Кендалла є робастними: наявність забруднень-викидів не може дуже сильно вплинути на значення коефіцієнта. При цьому τ Кендалла є більш стійким щодо забруднень,

[8] Більш формальне визначення поняття незалежності і того, як воно пов'язане з коефіцієнтами кореляції, ми розглянемо далі.

ніж ρ Спірмена.

Проілюструємо це наступним прикладом.

Приклад 5.4.3. Нехай незабруднена вибірка складається зі значень (X_j, Y_j) таких, що $Y_j = 1 - X_j$, $X_j = 0.1(j-1)$, $j = 1, \ldots, 11$. Зрозуміло, що всі розглянуті нами кореляції (r Пірсона, ρ Спірмена, τ Кендалла) для такої вибірки будуть дорівнювати -1.

Внесемо у вибірку забруднення, а саме, змінимо значення для 5-го елемента на $X'_j = c$, $Y'_j = c$, де c — деяке число. Всі інші елементи вибірки залишимо без змін. Якщо взяти c достатньо великим, можна зробити $r(X', Y')$ як завгодно близьким до 1. Отже висновок про кореляцію між X і Y може змінитись на протилежний завдяки зміні одного елемента вибірки. Коефіцієнти Спірмена і Кендела, звичайно, також зміняться, але не так радикально: τ не можна зробити більшим ніж -0.6363636, а ρ — більшим ніж -0.5, яке б велике c ми ні обрали. Тобто негативна кореляція між основною масою спостережень буде помічена обома цими коефіцієнтами, причому коефіцієнт Кендалла виявляє дещо більшу стійкість щодо забруднень.

У наступному скрипті-ілюстрації $c = 3$:

```
> x<-seq(0,1,0.1)
> y<-1-x
> plot(x,y)
> cor(x,y)

[1] -1

> x[5]<-3
> y[5]<-3
> cor(x,y)

[1] 0.6783182

> plot(x,y)
> cor(x,y,method = "kendall")
```

```
[1] -0.6363636

> cor(x,y,method = "spearman")

[1] -0.5
```

◄

Діаграми розсіювання даних відображені цим скриптом див. рис. 5.9.

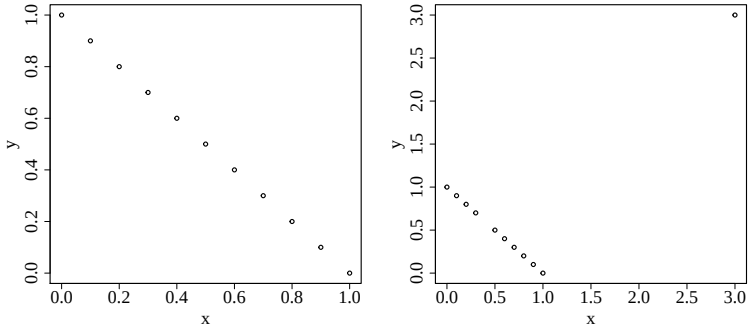

Рис. 5.9. Кореляція без викиду (ліворуч) і з (праворуч)

Рекомендації з використання коефіцієнтів кореляції.

1. Якщо у вас є теоретична модель даних, за якою змінні пов'язані лінійною залежністю, розмитою невеликими випадковими відхиленнями (напр., внаслідок похибок вимірювання), то для пошуку таких залежностей можна рекомендувати кореляцію Пірсона.

2. Кореляцію Пірсона доцільно також використовувати для характеризації сили залежності, якщо на діаграмі розсіювання не помітно інших залежностей, крім лінійних.

3. Якщо у даних можливі забруднення-викиди, краще використовувати кореляцію Кендалла, оскільки вона є найбільш робастною серед популярних коефіцієнтів кореляції. Кореляція Спір-

мена також стійка щодо забруднень і її можна використовувати у такому випадку, якщо для цього є якісь додаткові причини.

4. Рангові кореляції природно використовувати для аналізу даних, якщо шкала вимірювань деяких змінних визначена лише "з точністю до порядку" (порядкова шкала). Наприклад, рівень розумового розвитку людей (IQ) прийнято визначати у балах, що коливаються в межах від 60 до 140. Можна стверджувати, що людина з IQ 120 є більш інтелектуально розвиненою, ніж та, що має IQ 80. Але навряд чи має якийсь сенс твердження, що перша людина у 1.5 раза розумніша, ніж друга. Тобто сам IQ можна було б вимірювати і в інших одиницях, важливий тільки порядок на цій шкалі.

5. При виборі коефіцієнта кореляції важливу роль відіграє традиція у даній предметній області. Наприклад, у психологічних дослідженнях з початку XX ст. прийнято використовувати кореляцію Спірмена. Зрозуміло, що порівнювати нові наукові чи практичні результати зі старими психологам зручніше, якщо для аналізу нових використовується та ж техніка, яка була застосована у попередніх роботах. Тому перехід на нову техніку кореляційного аналізу не варто проводити без достатньо поважних причин. Якщо у такому випадку статистик вважає потрібним використати кореляційну техніку, відмінну від звичної для його замовників, краще зробити одразу два варіанти: старий і новий. І пояснити, чим новий варіант кращий старого.

5.5 Сила і значущість кореляції

Коефіцієнти кореляції використовують для того, щоб побачити, наскільки сильною є залежність між двома змінними. У деяких підручниках можна навіть знайти таку або подібну таблицю для визначення сили кореляції за значенням коефіцієнта:

$|r(X, Y)| = 1$ — змінні пов'язані лінійною функціональною залежністю;

$0.95 \leq |r(X,Y)| < 1$ — зв'язок дуже сильний, майже функціональний;

$0.75 \leq |r(X,Y)| < 0.95$ — зв'язок тісний (сильний);

$0.5 \leq |r(X,Y)| < 0.75$ — зв'язок середній (помірний);

$0.2 \leq |r(X,Y)| < 0.5$ — зв'язок слабкий;

$|r(X,Y)| < 0.2$ — зв'язку майже немає.

(див. [19])[9]

Якогось серйозного наукового змісту такі таблиці не мають, вони призначені для тих користувачів, яким зручніше оперувати словесними назвами, ніж числами. Але з їхнім використанням пов'язана певна небезпека. Інколи від замовника статистичного аналізу при обговоренні результатів можна почути: "Тут кореляція 0.485, а мені сказали, що такі кореляції слабкі і нікого не цікавлять. Чи не можна якось порахувати так, щоб вийшло більше 0.5?" Така постановка питання є, вочевидь, цілком неприйнятною.

З іншого боку, сама лише сильна корельованість двох змінних не обов'язково свідчить про наявність зв'язку між ними. Наприклад, якщо вибірка складається лише з двох елементів ($n = 2$), то всі вибіркові кореляції між змінними дорівнюватимуть ± 1.[10] Але, звичайно, звідси не можна робити висновки про залежності між змінними. Зрозуміло, за двома спостереженнями взагалі не можна робити будь-які обґрунтовані статистичні висновки. Однак при великій кількості спостережень навіть порівняно невеликі значення коефіцієнта кореляції можуть переконливо свідчити про наявність залежності між змінними.

---

[9] У російськомовному інтернеті такі таблиці можна зустріти під назвою "шкала Чеддока". При спробі знайти оригінал англійською мовою, гугл на запит "Chaddock correlation" видає чимало англомовних статей авторів з російськими прізвищами, надрукованих у сміттєвих псевдо-міжнародних журналах. Гіпотетичним автором "шкали" міг би бути Charles Gilbert Chaddock, якого можна знайти у вікіпедії. Він був американським невропатологом і психіатром XIX ст., зокрема, написав роботу "A suggestion for the statistical classification of insanity", але кореляції там не згадуються. Втім, у деяких російських статтях ця ж або подібна табличка фігурує під назвами "шкала Чертока" або "шкала Е.П.Голубкова".

[10] Якщо значення досліджуваних змінних різні на різних елементах вибірки.

Таким чином, при виявленні залежностей за допомогою коефіцієнтів кореляції, слід враховувати не тільки величину коефіцієнта, а і кількість спостережень, за якими він розрахований. Для цього застосовується стандартна схема тестів перевірки статистичних гіпотез, описана у підрозд. 9.1. Тут ми лише спрощено пояснимо її застосування в аналізі кореляцій. Більш докладно див. у п. 9.7.2.

Для простоти розглянемо тест для перевірки значущості кореляції з використанням коефіцієнта кореляції Пірсона $r(X, Y)$ (для інших коефіцієнтів процедура аналогічна).

Значущість кореляції визначається для певного рівня значущості[11] α. Рівень значущості — це ймовірність[12], з якою наша процедура перевірки буде помилково виявляти залежність змінних, які насправді є незалежними. Величину α обирає замовник статистичного дослідження, виходячи з того, наскільки небезпечними для нього є такі помилки. Наприклад, якщо покласти $\alpha = 0.05$ (це класичний вибір для біології, соціології, медицини та психології), то в середньому 1 раз на 20 випадків, коли залежності між змінними насправді немає, наша процедура помилково її виявить.

За значенням α визначається поріг тесту C_α. Цей поріг є найменшим числом C таким, щоб за відсутності залежності між X і Y ймовірність того, що $|r(X, Y)| > C$ не перевищувала α.

Тестова процедура має вигляд:

> Якщо $|r(X, Y)| > C_\alpha$ — вважаємо, що значуща залежність між X і Y виявлена.
> Якщо $|r(X, Y)| \leq C_\alpha$ — вважаємо, що залежність не виявлена (статистичні дані не підтверджують гіпотезу про залежність).

Чим меншим вибрати α, тим більшим буде C_α, тобто тим більшою має бути кореляція, щоб тестова процедура визнала її зна-

[11] *significance level.*

[12] Тобто середнє значення, навколо якого коливаються відносні частоти помилок при застосуванні тестової процедури.

чущою. Тому, вибираючи менше α, ми зменшуємо ймовірність помилково виявити залежності, яких немає, але перестаємо бачити не дуже сильні залежності, які насправді є.

Порогове значення C_α залежить також від n — кількості елементів у вибірці. Із зростанням n поріг C_α зменшується. Тобто чим більший обсяг вибірки, тим меншою може бути значуща кореляція.

На комп'ютерах зазвичай використовується інша процедура для реалізації таких тестів. А саме, розглянемо поріг $C_\alpha = C(\alpha)$ як функцію від можливих значень α на інтервалі $[0, 1]$. Позначимо $p(C)$ — функцію, обернену до $C(\alpha)$ на цьому інтервалі: $p(C_\alpha) = \alpha$. Оскільки $C(\alpha)$ — спадна функція, то і $p(C)$ також буде спадною. Тому нерівність

$$|r(X, Y)| > C_\alpha$$

еквівалентна нерівності

$$p(|r(X, Y)|) < \alpha.$$

Величина $p = p(|r(X, Y)|)$ називається **досягнутим рівнем значущості**[13] для перевірки незалежності змінних X і Y за допомогою кореляцій Пірсона.

З використанням досягнутого рівня значущості тестову процедуру можна оформити так:

> Якщо $p < \alpha$ — вважаємо, що значуща залежність між X і Y виявлена.
> Якщо $p \geq \alpha$ — вважаємо, що залежність не виявлена (статистичні дані не підтверджують гіпотезу про залежність).

Зрозуміло, що ця процедура цілком еквівалентна попередній: на однакових даних обидві процедури будуть давати однакові результати. Тому кажуть, що ці процедури реалізують один і той самий статистичний тест.

[13] *significance,* або *p-value*.

Зручність використання досягнутого рівня значущості p полягає в тому, що його можна порівнювати з будь-яким рівнем значущості α, заданим користувачем. Зокрема, якщо при публікації дослідник вказує, яке p він отримав за своїми даними, то читач може сам порівняти це значення з тим α, яке він вважає прийнятним, не роблячи додаткових обчислень.

При використанні коефіцієнтів Спірмена і Кендалла тестові процедури будуються за тією ж схемою, але, звичайно, порогові значення C_α розраховуються за іншими формулами.

У R розраховувати досягнуті рівні значущості і робити перевірку залежності на основі кореляцій можна, використовуючи функцію `cor.test()`. Перші два параметри цієї функції — `x`, `y` — це змінні, за якими розраховується кореляція (мають бути векторами однакової довжини). Опція `method` дозволяє обрати тип коефіцієнта кореляції: `"pearson"` (стандартно), `"kendall"` або `"spearman"`.

Досягнутий рівень значущості для перевірки залежності функція вказує у результаті виконання в атрибуті `$p.value`.

Приклад 5.5.1. У прикладі 5.1.2 ми розглянули залежність між довжиною пелюсток і чашолистків у квітів-півників виду setosa. На рис. 5.3 видно, що виразної залежності між цими змінними немає. Але пряма, яку ми підігнали за методом найменших квадратів для опису залежностей "в середньому" начеб-то показує, що із зростанням довжини пелюсток зростає і середня довжина чашолистків.

Чи можна вважати цю залежність значущою?

Обчислимо кореляції між цими змінними:

```
> x<-iris[iris$Species=="setosa","Petal.Length"]
> y<-iris[iris$Species=="setosa","Sepal.Length"]
> cor(x,y)

[1] 0.2671758

> cor(x,y,method = "kendall")

[1] 0.2173273
```

```
> cor(x,y,method = "spearman")

[1] 0.2788849
```

Отримали кореляції $r(X,Y) = 0.2671758$, $\tau(X,Y) = 0.2173273$, $\rho(X,Y) = 0.2788849$. Всі значення коефіцієнтів відповідають варіанту "зв'язок слабкий" з таблиці, що наведена на початку цього підрозділу.

Застосуємо тепер статистичні тести, які ми тільки що розглянули. Виберемо стандартний рівень значущості $\alpha = 0.05$. Розрахуємо досягнуті рівні значущості:

```
> cor.test(x,y)$p.value

[1] 0.06069778

> cor.test(x,y,method = "kendall")$p.value

[1] 0.0447895

> cor.test(x,y,method = "spearman")$p.value

[1] 0.04985095
```

Бачимо, що досягнутий рівень значущості, розрахований на основі коефіцієнта Пірсона r, дорівнює $p_r = 0.06069778 > 0.05$, тобто кореляцію слід визнати незначущою: така кореляція не дозволяє стверджувати, що між змінними є якась залежність. На основі коефіцієнта Кендалла τ отримуємо $p_\tau = 0.0447895 < \alpha$. Тобто якщо робити висновки на основі τ, слід визнати кореляцію значущою[14]. На основі ρ Спірмена ми знову мусимо прийняти незначущість кореляції: $p_\rho = 0.04985095$.

[14] Хоча $r(X,Y) > \tau(X,Y)$!

Такий результат дещо спантеличує. Але слід розуміти, жоден, навіть найбільш адекватний статистичний тест не гарантує від помилок: ми можемо лише сподіватись, що помилок не буде занадто багато.

У даному випадку можна сказати, що спостережень явно не вистачає для того, щоб зробити остаточний висновок: чи є така кореляція свідченням справжньої залежності між змінними, чи вона виникла внаслідок випадкового збігу обставин саме у цій вибірці і не буде виявлятись на інших спостереженнях. ◀

Коли структуру залежностей змінних відображають у вигляді кореляційної мережі, ребра графа, що відповідають незначущим кореляціям доцільно не відображати. У функції `qgraph`, яку ми розглядали в підрозд. 5.3, є для цього спеціальна опція `minimum`. Якщо кореляція між змінними менша, ніж значення, вказане у `minimum` (стандартно — 0), то ці змінні на графі не з'єднуються ребром. Можна також вказати `minimum="sig"`, тоді не будуть відмічатись незначущі кореляції. У цьому випадку, для того, щоб функція працювала правильно, їй треба задати іще опції `alpha` — рівень значущості і `sampleSize` — обсяг вибірки, за якою були розраховані кореляції.

Приклад 5.5.2. От як можна змінити скрипт з прикладу 5.3.4, щоб на кореляційній мережі виводились лише кореляції (Пірсона) значущі з рівнем $\alpha = 0.05$:

```
> library("qgraph")
> Spec<-c("setosa","versicolor","virginica")
> for(sp in Spec){
+   qgraph(cor(iris[iris$Species==sp,1:4]),
+                minimum="sig",alpha=0.05,
+                sampleSize=
+            nrow(iris[iris$Species==sp,1:4]))
+         text(-0.75,1,sp,cex=1.5)
+           }
```

Результат — на рис. 5.10.

Відрізаючи ребра, що відповідають не значущим кореляціям, ми привертаємо більше уваги до значущих. Зокрема, це полегшує порівняння кореляційної структури для різних наборів даних. Щоб переконатись у цьому, порівняйте рис. 5.8 з рис. 5.10. ◄

Описані у цьому підрозділі тести є дуже популярними, але ними не вичерпуються техніки порівняння залежностей між змінними. Про інші підходи до аналізу залежностей див. у підрозд. 9.7 і 9.7.3. На основі описаних у цих підрозділах підходів також можна будувати графи, що описують наявність і силу виявленої залежності між парами досліджуваних змінних.

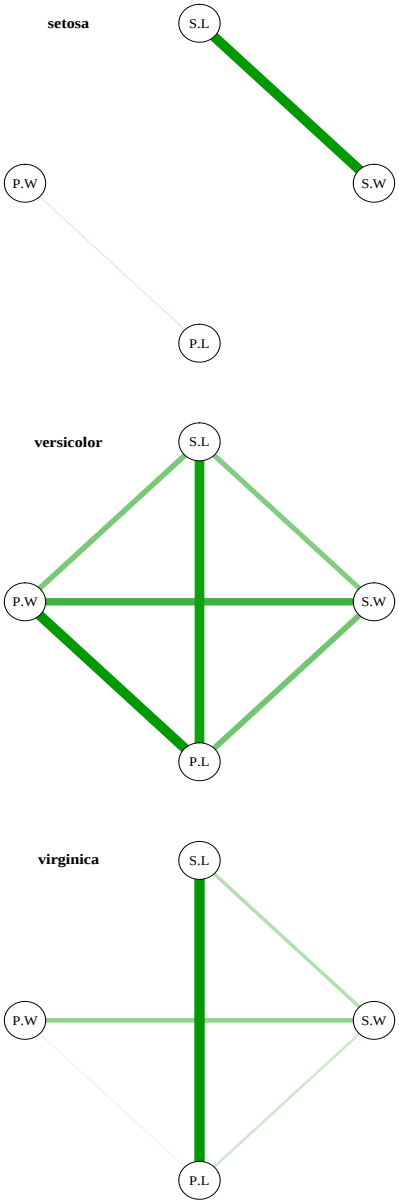

Рис. 5.10. Порівняння кореляцій розмірів для півників

Розділ 6

Основні ймовірнісні розподіли

У цьому розділі обговоримо теоретичні моделі, якими у математичній статистиці описують розподіли даних і те, як обчислення за цими теоретичними моделями можна проводити в R. Крім того, у підрозд. 6.5 ми побачимо, як можна створювати штучні дані, які задовольняють такі теоретичні моделі. Штучно згенеровані дані часто використовуються для перевірки якості статистичних алгоритмів. У наступних розділах книги описані моделі розподілів застосовуються для аналізу статистичних даних.

Я намагався писати цей розділ так, щоб його можна було читати з мінімальними уявленнями про теорію ймовірностей. За потреби читач може звернутись до Додатка Б де нагадуються основні поняття що стосуються ймовірностей, випадкових величин, розподілів та математичних сподівань. Більш докладно про ці та інші поняття теорії ймовірностей можна прочитати у книгах [2, 9, 18]

6.1 Загальні поняття та схема використання основних розподілів у R

У математичній статистиці дані прийнято розглядати як випадкові об'єкти — випадкові величини або вектори, процеси, поля, множини... Статистичні характеристики даних природно описувати в термінах імовірнісних розподілів цих об'єктів.

Розподіл будь-якої випадкової величини (в.в.) ξ можна задати, вказуючи функцію розподілу (ф.р.), тобто $F_\xi(x) = \mathsf{P}\{\xi \le x\}$. Якщо існує така функція $f_\xi(x)$, що $F_\xi(x) = \int_{-\infty}^{x} f_\xi(t)dt$ при всіх $x \in \mathbb{R}$, то кажуть, що розподіл є абсолютно неперервним, а f_ξ, називають щільністю розподілу. Щільність також однозначно задає розподіл.

Якщо розподіл є дискретним, тобто існує зліченний набір $T = \{t_1, t_2, \dots, \} \in \mathbb{R}$, такий, що $\mathsf{P}\{\xi \in T\} = 1$, то функцію $f_\xi(x) = \mathsf{P}\{\xi = x\}$ можна трактувати як щільність розподілу ξ відносно лічильної міри. Цю функцію інколи також називають розподілом (або рядом розподілу) дискретної випадкової величини (probability mass function).

Квантилем $Q^\xi(\alpha)$ розподілу випадкової величини ξ, рівня α називають найменше серед чисел[1] x, для яких $F_\xi(x) \ge \alpha$. Якщо існує функція $F_\xi^{-1}(x)$, обернена до функції розподілу, то $Q^\xi(\alpha) = F_\xi^{-1}(\alpha)$.

Для опису розподілу даних та функцій від них (статистик) часто використовуються параметричні моделі, у яких функція розподілу вважається відомою з точністю до деяких параметрів. У наступних підрозділах цього розділу найбільш вживані моделі розподілів розглянуті детальніше. Зараз ми обмежимось загальною схемою організації ймовірнісних обчислень за допомогою R.

У R для ряду найбільш поширених параметричних моделей реалізовані функції, що обчислюють функцію розподілу, щільність, квантилі для заданого розподілу та генерують псевдовипадкову

[1] Точніше, $Q^\xi(\alpha) = \inf\{x \in \mathbb{R} \,:\, F_\xi(x) \ge \alpha\}$.

величину із заданим розподілом. Ці функції організовані за єдиною схемою (табл. 6.1).

Таблиця 6.1. Імена функцій для основних імовірнісних розподілів

Розподіл	Мнемоніка	Параметри
бета	beta	shape1, shape2
біноміальний	binom	size, prob
гамма	gamma	shape, rate
геометричний	geom	prob
гіпергеометричний	hyper	m, n, k
експоненційний	exp	rate
Коші	cauchy	location, scale
логістичний	logis	location scale
логнормальний	lnorm	meanlog, sdlog
негативний біноміальний	nbinom	size, prob
нормальний	norm	mean, sd
Пуассона	pois	lambda
рівномірний	unif	min, max
Вейбула	weibull	shape, scale
Вілкоксона	wilcox	m,n
χ^2	chisq	df
F-Фішера	f	df1, df2
T-Стьюдента	t	df

Ім'я функції утворюється з *мнемоніки* — скороченого імені розподілу та префікса, який вказує, що обчислює дана функція. Префікси можуть бути такими:

p — обчислення функції розподілу (probability). Наприклад, `pnorm(1.96)` — функція стандартного нормального розподілу у точці 1.96 (п. 6.2.1);

d — обчислення щільності (density) розподілу (для абсолютно неперервних випадкових величин) або ймовірності поптрапляння у точку (для дискретних). Наприклад,

`dbinom(1,size=1,prob=0.5)` — ймовірність того, що випадкова величина з біноміальним розподілом дорівнює 1, якщо ймовірність успіху 0.5, а кількість випробувань — 1 (п. 6.3.1);

q — обчислення квантиля (quantile) заданого рівня: значенням функції `qnorm(c(0.025,0.975))` буде вектор квантилів рівня 0.025 і 0.975 для стандартного нормального розподілу, тобто (-1.959964, 1.959964);

r — генерація псевдовипадкових чисел (random number) із заданим розподілом: `rnorm(100)` генерує 100 псевдовипадкових значень, що моделюють вибірку з незалежних стандартних нормальних випадкових величин.

У функцій з префіксами `p`, `d` і `q` першим параметром є вектор значень аргументів, для яких треба обчислити відповідну функцію (функція розподілу, щільність, квантиль). У функцій з префіксом r (псевдовипадкових генераторів) перший аргумент — розмір вибірки, тобто кількість генерованих величин.

Наступні параметри є параметрами розподілу. Вони різні для різних розподілів (див. третій стовпчик табл. 6.1), але однакові для всіх функцій, пов'язаних з даним розподілом. Наприклад, для нормального розподілу, параметри `mean` та `sd` вказують математичне сподівання та стандартне відхилення (корінь квадратний з дисперсії).

У всіх функцій з префіксом `p` і `q` є логічний параметр-опція `lower.tail`. Його стандартне значення — FALSE. Якщо задати `lower.tail=T`, то p-функція буде замість функції розподілу обчислювати функцію виживання $\mathsf{P}\{\xi > x\} = 1 - F_\xi(x)$, а q-функція — верхній квантиль, тобто $Q^\xi(1-\alpha)$.

Якщо у функції `p` задати опцію `log.p=T`, то вони будуть обчислювати логарифм функції розподілу:

```
> pnorm(-1.96,log.p=T)

[1] -3.688964

> log(pnorm(-1.96))
```

```
[1] -3.688964
```

△

(Насправді `pnorm(-1.96,log.p=T)` не обчислює спочатку ф.р., а потім її логарифм, а одразу шукає цей логарифм за спеціальним алгоритмом наближеного обчислення. Тому такий варіант виклику функції працює швидше і дає точніший результат ніж `log(pnorm(-1.96))`, хоча для більшості застосувань різниця майже непомітна).

6.2 Неперервні розподіли на прямій

У цьому підрозділі розглядаються найбільш вживані розподіли, шо описують поведінку випадкових величин, які можуть набувати довільних числових значеннь на всій прямій, або на деякому інтервалі (неперервних випадкових величин).

6.2.1 Одновимірний гауссів (нормальний) розподіл

Гауссів (нормальний) розподіл[2] є, мабуть, найбільш вживаною моделлю розподілу неперервної випадкової величини. Кажуть, що в.в. ξ має нормальний розподіл з параметрами μ, σ^2 (позначається $\xi \sim N(\mu, \sigma^2)$), якщо щільність розподілу ξ має вигляд

$$f_\xi(x) = f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Гауссів розподіл з параметрами $\mu = 0$, $\sigma^2 = 1$ називають стандартним. Його щільність позначають

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

[2] Gaussuian, normal.

Функція розподілу ξ не записується аналітично

$$F_\xi(x) = F(x, \mu, \sigma) = \int_{-\infty}^{x} f(t, \mu, \sigma)dt,$$

але її можна виразити через **стандартну функцію нормального розподілу**[3]

$$\Phi(x) = F(x, 0, 1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) dt.$$

Для будь-яких $\mu \in \mathbb{R}$ і $\sigma > 0$

$$F(x, \mu, \sigma^2) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

Якщо $\xi \sim N(\mu, \sigma^2)$, параметр μ є математичним сподіванням та медіаною розподілу ξ, σ^2 — дисперсія ξ. Графіки щільностей нормального розподілу при різних значеннях параметрів зображено на рис. 6.1. Графіки на рис. 6.1 побудовані таким скриптом:

```
> plot(c(-4,4),c(0,0.8),type="n",xlab="x",ylab="")
> curve(dnorm(x,mean=0,sd=1),-4,4,lwd=2,add=T)
> curve(dnorm(x,mean=0,sd=0.5),-4,4,col="red",
+       lwd=2,lty="dashed",add=T)
> curve(dnorm(x,mean=0,sd=1.5),-4,4,col="brown",
+       lwd=2,lty="longdash",add=T)
> curve(dnorm(x,mean=1,sd=1),-4,4,col="blue",
+       lwd=2,lty="dotdash",add=T)
> legend(-4,0.75,
+        legend=c("N(0,1)","N(0,0.25)",
```

[3] Ця функція виражається через спеціальну функцію erf (гауссова функція помилок, інтеграл помилок): $\Phi(x) = (1 + \mathrm{erf}(x/\sqrt{2}))/2$, де

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

```
+                 "N(0,1.75)","N(1,1)"),
+       lty=c("solid","dashed","longdash","dotdash"),
+       lwd=c(2,2,2,2),
+       col=c(1,"red","brown","blue"))
```

△

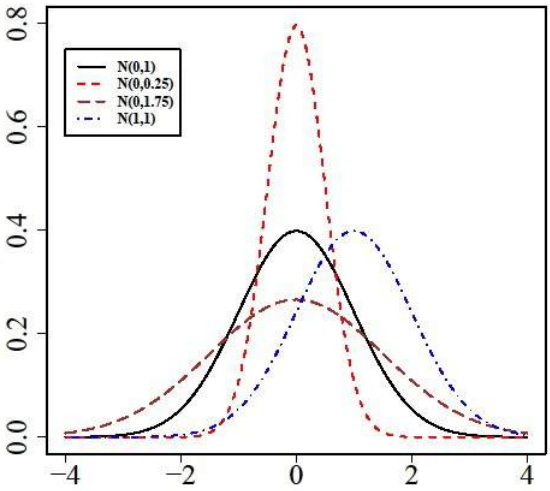

Рис. 6.1. Щільність нормального розподілу

На графіках видно, що параметр μ визначає положення точки максимуму (піка) гауссової щільності, а σ^2 — гостроту піка (чим менше σ^2, тим пік гостріший).

У R мнемоніка нормального розподілу — `norm`, параметр μ позначається як `mean` — середнє, σ — як `sd` (скорочення від standard deviation — середньоквадратичне відхилення).

Важливою особливістю гауссового розподілу є те, що сума кількох незалежних гауссових випадкових величин також є гауссовою.

Точніше, якщо $\xi_i \sim N(\mu_i, \sigma_i)$, $i = 1, \ldots, n$ — незалежні, то

$$\sum_{i=1}^{n} \xi_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

Своєю популярності у статистиці гауссів розподіл завдячує центральній граничній теоремі. Грубо кажучи, ця теорема стверджує, що сума великої кількості незалежних випадкових величин, дисперсії яких не дуже сильно відрізняються, має приблизно гауссів розподіл. Таким чином, цей розподіл природно застосовувати для опису випадкової поведінки таких величин, які складаються під дією багатьох різних факторів, не дуже залежних між собою, якщо їхні впливи на дану величину підсумовуються.

Наприклад, при стрільбі з лука у мішень відхилення точки, в яку влучить стріла, від центра мішені утворюється внаслідок неточного прицілювання, коливань тіла стрілка при стрільбі, поривів вітру тощо. Таких факторів багато, вони мало пов'язані один з одним, а їхні впливи накладаються, створюючи результат — координату по горизонталі (або вертикалі), точки потрапляння стріли у мішень. Тому таку координату природно спробувати описати гауссовим розподілом.

Зовсім не обов'язково, щоб результати реальних стрільб дійсно описувались цим розподілом. Чому може виникати відхилення від гауссового розподілу? Негауссовість часто виникає, якщо одна причина серед тих, що формують спостереження, є значно впливовішою, ніж інші. Скажімо, стрілок може цілитись, заплющуючи праве або ліве око. Якщо він лівим оком бачить значно краще, ніж правим, то різні варіанти заплющування очей будуть приводити до різних розподілів координат точки потрапляння. Якщо спостерігати ці результати разом, не знаючи, яке око заплющував стрілок при кожному конкретному пострілі, то ми отримаємо суміш двох наборів спостережень з різними розподілами: для лівого і правого ока окремо (про суміші див. п. 6.4.3). Навіть, коли розподіл кожного з цих наборів є гауссовим, розподіл їх суміші не буде гауссовим.

Аналіз спостережень, що описуються гауссовим розподілом — один з найбільш розвинених розділів математичної статистики. У цій книзі оцінюванню за гауссовими спостереженнями присвячені приклади 8.4.2, 8.4.3, 8.5.3, а перевірці гіпотез — підрозд. 9.5.

Багато розподілів, пов'язаних з гауссовим (півнормальний, логнормальний, хі-квадрат, Т-Стьюдента, F-Фішера), розглянемо далі. Крім того, особливого розгляду заслуговує багатовимірний гауссів розподіл.

6.2.2 Півнормальний розподіл

Півнормальним (half-normal) називають розподіл, який має випадкова величина $\eta = |\xi|$ де $\xi \sim N(0, \sigma^2)$. Позначення $\eta \sim \mathrm{HalfN}(\sigma^2)$. Цей розподіл природно застосовувати для моделювання поведінки величин, що визначаються як відхилення чогось, що має нормальний розподіл, від середнього положення, без урахування напрямку відхилення.

Щільність півнормального розподілу

$$f(x) = f(x, \sigma) = \begin{cases} \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) & \text{при } x \geq 0, \\ 0 & \text{при } x < 0. \end{cases}$$

Функція розподілу

$$F(x) = F(x; \sigma^2) = \begin{cases} 2\Phi(x/\sigma) - 1 & \text{при } x > 0 \\ 0 & \text{при } x < 0. \end{cases}$$

Інколи для параметризації півнормальних розподілів замість параметра σ^2 використовують

$$\vartheta = \frac{\sqrt{\pi}}{\sigma\sqrt{2}}.$$

Математичне сподівання та дисперсія півнормального розподілу

$$\mathsf{E}\,\eta = \frac{\sigma\sqrt{2}}{\sqrt{\pi}} = \frac{1}{\vartheta},\ \mathsf{D}\,\eta = \left(1 - \frac{2}{\pi}\right)\sigma^2.$$

Медіана

$$\mathrm{med}(\eta) = \sigma\Phi^{-1}(3/4).$$

6.2.3 Логнормальний розподіл

Логнормальним (log-normal) називають розподіл, який має випадкова величина $\eta = e^{\xi}$, де $\xi \sim N(\mu, \sigma^2)$. Інакше кажучи, логарифм логнормальної величини має нормальний розподіл. Позначення $\eta \sim \mathrm{LN}(\mu, \sigma^2)$.

Щільність розподілу логнормальної величини з параметрами (μ, σ^2):

$$f_\eta(x) = f(x; \mu, \sigma) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{\log x - \mu)^2}{2\sigma^2}\right) & \text{при } x > 0, \\ 0 & \text{при } x \le 0. \end{cases}$$

Графіки цієї щільності при різних значеннях параметрів зображені на рис. 6.2.

Функція розподілу:

$$F_\eta(x) = \begin{cases} \Phi\left(\frac{\log x - \mu}{\sigma}\right) & \text{при } x > 0, \\ 0 & \text{при } x \le 0. \end{cases}$$

Математичне сподівання, дисперсія і медіана:

$$\mathsf{E}\,\eta = \exp(\mu + \sigma^2/2),\ \mathsf{D}\,\eta = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2},\ \mathrm{med}(\eta) = e^{\mu}.$$

У `R` логнормальний розподіл має ім'я (мнемоніку) `lnorm`, його параметри μ — `meanlog`, σ — `sdlog`.

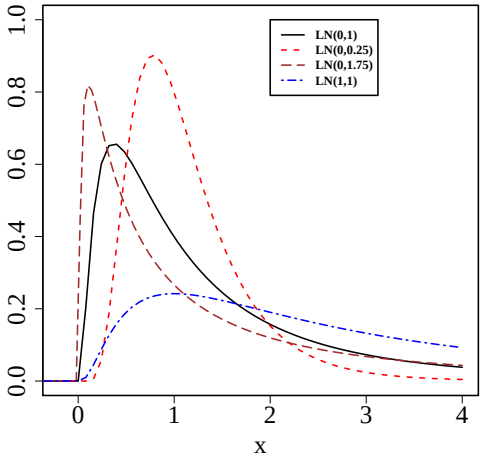

Рис. 6.2. Щільність логнормального розподілу

Поширеність логнормального розподілу серед моделей реальних даних пояснюється центральною граничною теоремою аналогічно тому, як це було для гауссового розподілу. Дійсно, якщо деяка спостережувана величина η формується як добуток великої кількості незалежних випадкових величин, то її логарифм буде сумою логарифмів співмножників. Отже, внаслідок центральної граничної теореми можна сподіватись, що розподіл $\log \eta$ буде близьким до гауссового, а розподіл самого η — логнормальним.

Таким чином, якщо впливи різних причин, що формують досліджувану величину додаються, можна очікувати нормального розподілу, а якщо перемножаються — логнормального. Зрозуміло, що у більшості реальних досліджень конкретний спосіб взаємодії різних причин не описується такими простими формулами, як сума чи добуток. Це лише дуже спрощений спосіб міркувань, що можуть навести на ідею вибору відповідного розподілу. Зокрема, можна сподіватись логнормальності там, де розмір, досягнутий раніше, підсилює можливості подальшого зростання.

Наприклад, якщо досліджувана змінна — розмір капіталу різних підприємств деякої галузі, то можна припускати, що цей розмір складається під дією багатьох причин і змінюється з часом. Але, якщо підприємство має капітал в 1 тис. грн., навіть у сприятливих умовах, йому важче протягом року збільшити цей капітал на 10 тис., ніж підприємству, що має капітал 1 млн. грн. Також і втрати під впливом несприятливих обставин у більшого підприємства будуть більші. Досягнутий підприємством розмір підсилює вплив випадкових причин. У цьому прикладі природно характеризувати прирости не "на стільки-то гривень", а "у стільки-то разів", тобто внески випадкових впливів перемножаються а не додаються. Отже, скоріше слід сподіватись логнормального розподілу для розміру капіталів.

Логнормальний розподіл належить до класу розподілів, що породжують викиди. Тобто для вибірки з логнормальним розподілом типовою є наявність викидів навіть тоді, коли забруднень сторонніми спостереженнями немає.

Далі у прикладі 10.2.2 ми побачимо, що у регресійній формулі, яка пов'язує вагу і довжину оселедців, похибки мають логнормальний розподіл.

6.2.4 Експоненційний і гамма розподіли та розподіл Лапласа

Експоненційним (показниковим, exponential) називають розподіл випадкової величини ξ зі щільністю

$$f_\xi(x) = f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{при } x \geq 0, \\ 0 & \text{при } x < 0, \end{cases}$$

де $\lambda > 0$ — параметр, що називається інтенсивністю (`rate` в `R`). Позначення[4] $\mathrm{Exp}(\lambda)$

[4]Інколи використовують параметризацію експоненційних розподілів параметром $\vartheta = 1/\lambda$, який дорівнює математичному сподіванню експоненційної ви-

Функція розподілу:

$$F_\xi(x) = F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{при } x \geq 0, \\ 0 & \text{при } x < 0. \end{cases}$$

Таким чином, експоненційно розподілені випадкові величини можуть набувати лише додатних значеннь.

Експоненційний розподіл часто використовують для опису часу очікування яких-небуть подій. Це пов'язано з властивістю відсутності післядії, яку серед неперервних розподілів має лише експоненційний. А саме, нехай час очікування ξ деякої події. Припустимо, що пройшов час t, але подія не відбулась. Яким буде розподіл величини $(\xi - t)$ — часу очікування, який залишився? Якщо ξ має експоненційний розподіл, то розподіл залишкового часу чекання той самий, що й у ξ:

$$\mathsf{P}\{\xi - t < x \mid \xi \geq t\} = \mathsf{P}\{\xi < x\}.$$

Наприклад, нехай ξ — час від вкручування до перегоряння електричної лампочки. Якщо перегоряння відбувається внаслідок дії таких випадкових зовнішніх причин, як стрибки напруги у мережі, що не пов'язані з тривалістю роботи лампочки, то природно сподіватись саме відсутності післядії: скільки б часу не довелось лампочці пропрацювати, а шансів перегоріти протягом найближчої години буде стільки ж, як коли вона була зовсім нова. У цьому випадку розподіл ξ має бути експоненційним[5].

Зрозуміло, що у реальних лампочок крім випадкових зовнішніх причин є ще старіння внаслідок інтенсивної роботи. Якщо

падкової величини. Аналогічна параметризація можлива для гамма-розподілу та розподілу Лапласа.

[5] Це правильно якщо час роботи розглядати як неперервну випадкову величину. Якщо ξ — дискретна величина (скажімо, кількість днів, які лампочка пропрацювала до перегоряння), то єдиний можливий розподіл без післядії для неї — геометричний.

ефект старіння помітний порівняно з ефектами випадкових причин, експоненційний розподіл не буде хорошою моделлю для опису відповідного часу чекання.

Математичне сподівання, дисперсія та медіана експоненційної випадкової величини ξ:

$$\mathsf{E}\,\xi = \frac{1}{\lambda},\ \mathsf{D}\,\xi = \frac{1}{\lambda^2},\ \operatorname{med}(\xi) = \frac{\log 2}{\lambda}.$$

Методи статистичного аналізу на основі експоненційного розподілу розглянуто далі у прикладах 8.4.1, 8.5.1, 9.2.1.

Сума n незалежних експоненційних випадкових величин ξ_i з однаковою інтенсивністю λ має **гамма-розподіл**

$$\xi_1 + \xi_2 + \cdots + \xi_n \sim \Gamma(n, \lambda).$$

Щільність гамма-розподілу $\Gamma(n, \lambda)$:

$$f(x; n, \lambda) = \begin{cases} \frac{\lambda^n x^{n-1}}{\Gamma(n)} e^{-\lambda x} & \text{при } x \geq 0, \\ 0 & \text{при } x < 0 \end{cases}$$

(тут $\Gamma(n)$ — гамма-функція від n, параметри n і λ можуть бути довільними додатними числами).

Графіки щільності експоненційного та гамма-розподілів при різних значеннях параметрів зображені на рис. 6.3.

Цей розподіл для цілих n також називають розподілом Ерланга.

У `R` експоненційний розподіл має мнемоніку `exp`, а інтенсивність λ відповідає параметру `rate`. Мнемоніка гамма-розподілу — `gamma`, параметри n =`shape`, λ =`rate`.

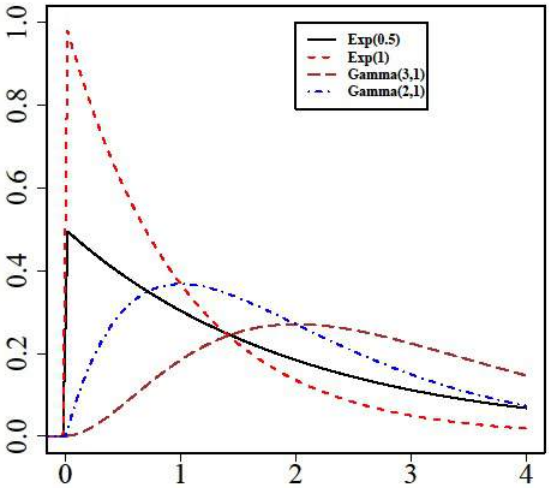

Рис. 6.3. Щільність експоненційного та гамма розподілів

Різниця двох незалежних експоненційних випадкових величин $\xi_1 - \xi_2$ з інтенсивністю λ має розподіл Лапласа (позначення: Laplace$(0, \lambda)$). Щільність розподілу Лапласа Laplace(μ, λ):

$$f_\xi(x) = f(x; \mu, \lambda) = \frac{\lambda}{2} e^{-\lambda|x-\mu|}.$$

Щодо утворення викидів експоненційний розподіл вважається проміжним. Розподіли, у яких функція розподілу спадає на нескінченності повільніше, ніж експоненційна, породжують викиди (це розподіли з важкими хвостами, як Т-розподіл Стьюдента або розподіл Парето). Якщо функція розподілу спадає швидше, ніж експоненційна, такий розподіл викидів не породжує (розподіли з легкими хвостами — такі як нормальний розподіл). Вибірки з експоненційного розподілу на гістограмі часто утворюють окремі стовпчики, що не дуже сильно відхиляються від основної маси

спостережень. Їх інколи називають “помірними викидами”.

У `R` функції для роботи з розподілом Лапласа є у бібліотеці `rmutil`. Вони мають мнемоніку `laplace` і працюють аналогічно до інших ймовірнісних функцій `R`. Параметр μ у цих функціях позначається `m`, параметр λ — `s`.

6.2.5 Розподіли екстремальних типів: Вейбулла, Фреше і Гумбеля

Як було зазначено, нормальні випадкові величини природно виникають при підсумовуванні, логнормальні — при множенні впливів. Але крім підсумовування та множення можливе також формування спостережуваного результату як максимуму або мінімуму деякого набору незалежних випадкових величин. Наприклад, міцність ланцюга на розрив дорівнює міцності найслабшої його ланки. Аналогічно центральній граничній теоремі існує теорема про можливі граничні розподіли максимумів та мінімумів великої кількості незалежних випадкових величин. Існують три типи таких розподілів: Вейбулла, Фреше і Гумбеля.

Розподіл Вейбулла визначається функцією розподілу:

$$F_\xi(x) = \begin{cases} 1 - e^{-(x/\lambda)^k} & \text{при } x \geq 0, \\ 0 & \text{при } x < 0. \end{cases}$$

Тут $k > 0$, $\lambda > 0$ — параметри розподілу. Параметр λ називають параметром масштабу (`scale`), а k — параметром форми (`shape`). При використанні у матеріалознавстві k називають вейбулловим параметром матеріалу. Виявляється, що міцність виробу або час роботи приладу до відмови часто мають розподіл Вейбулла, причому параметр k залежить переважно від матеріалу, з якого зроблений виріб, а не від його конструктивних особливостей.

Щільність розподілу Вейбулла:

$$f_\xi(x) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & \text{при } x \geq 0, \\ 0 & \text{при } x < 0. \end{cases}$$

Математичне сподівання і дисперсія випадкової величини ξ з розподілом Вейбулла:

$$\mathsf{E}\,\xi = \lambda\Gamma\left(1+\frac{1}{k}\right),\ \ \mathsf{D}\,\xi = \lambda^2\left[\Gamma\left(1-\frac{2}{k}\right) - \left(\Gamma\left(1+\frac{1}{k}\right)\right)^2\right].$$

У `R` розподіл Вейбулла має мнемоніку `weibull`.

Застосування розподілу Вейбулла для опису тривалості роботи вальниць розглянуто у прикладах 8.6.2 і 9.3.2.

Розподіли Гумбеля і Фреше менше застосовуються як моделі розподілів реальних даних, ніж розподіл Вейбулла. Наведемо їх функції розподілу. Для розподілу Гумбеля (лог-вейбулів розподіл):

$$F_\xi(x) = \exp\left(-e^{-(x-\mu)/\beta}\right),$$

де μ, β — параметри розподілу.

Для розподілу Фреше:

$$F_\xi(x) = \begin{cases} \exp(-(x/s)^{-\alpha}) & \text{при } x \geq 0, \\ 0 & \text{при } x < 0. \end{cases}$$

де α і s — параметри розподілу.

6.2.6 Рівномірний розподіл

Випадкова величина ξ має рівномірний (uniform) розподіл на інтервалі $[a, b]$, якщо її щільність

$$f_\xi(x) = f(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{при } a \leq x \leq b, \\ 0 & \text{при } x \notin [a, b]. \end{cases}$$

Функція розподілу

$$F_\xi(x) = \begin{cases} \frac{x-a}{b-a} & \text{при } a \leq x \leq b, \\ 0 & \text{при } x < a, \\ 1 & \text{при } x > b. \end{cases}$$

Позначення $\xi \sim \mathrm{Unif}[a,b]$.

Математичне сподівання, дисперсія, медіана:

$$\mathsf{E}\,\xi = \mathrm{med}(\xi) = \frac{1}{2}(a+b),\ \mathsf{D}\,\xi = \frac{1}{12}(b-a)^2.$$

У `R` мнемоніка рівномірного розподілу — `unif`, параметри $a =$`min`, $b =$`max`.

Рівномірний розподіл використовують для опису "похибок округлення". Наприклад, якщо довжину яких-небудь досить довгих предметів вимірювати лінійкою, що має лише сантиметрові поділки, і брати округлення до найближчої поділки, то розподіл похибки (у сантиметрах) буде рівномірним на $[-1/2, 1/2]$. Якщо округлювати, відкидаючи дробову частину, рівномірним на $[0,1]$. У страховій справі при контрактах з виплатою страхової суми наприкінці року інтервал між страховою подією та виплатою за нею вважають рівномірно розподіленим протягом року.

Статистика рівномірного розподілу має свої особливості, часом несподівані для тих, хто звик до аналізу даних з неперервною щільністю розподілу, див. приклад 8.4.5.

Нехай ξ_1, ξ_2 — незалежні, рівномірно розподілені випадкові величини на $[0,1]$. Тоді $\eta = \xi_1 - \xi_2$ має симетричний трикутний розподіл на $[-1,1]$ (розподіл Сімпсона) з щільністю

$$f_\eta(x) = (1-|x|)\mathbb{1}\{|x|<1\},$$

$\xi_1 + \xi_2$ має симетричний трикутний розподіл на $[0,2]$, а $|\xi_1 - \xi_2|$ має трикутний розподіл з щільністю

$$f_{|\eta|}(x) = 2(1-x)\mathbb{1}\{0<x<1\}.$$

У загальному випадку трикутний розподіл (triangular distribution) на інтервалі $[a,b]$ з модою (вершиною) у $c \in [a,b]$ задається своєю щільністю

$$f_\xi(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & \text{при } a \le x \le c, \\ \frac{2(b-x)}{(b-a)(b-c)} & \text{при } c < x \le b, \\ 0 & \text{при } x \notin [a,c]. \end{cases}$$

Позначення $\mathrm{Triangle}(a, b, c)$.

Математичне сподівання, дисперсія та медіана трикутного розподілу:

$$\mathsf{E}\,\xi = \frac{a+b+c}{3},\ \mathsf{D}\,\xi = \frac{a^2+b^2+c^2-ab-ac-bc}{18},$$

$$\mathrm{med}(\xi) = \begin{cases} a + \sqrt{\frac{(b-a)(c-a)}{2}} & \text{при } c \geq (a+b)/2, \\ b - \sqrt{\frac{(b-a)(c-a)}{2}} & \text{при } c \leq (a+b)/2. \end{cases}$$

Графіки щільності рівномірного та трикутного розподілів при різних значеннях параметрів зображені на рис. 6.4.

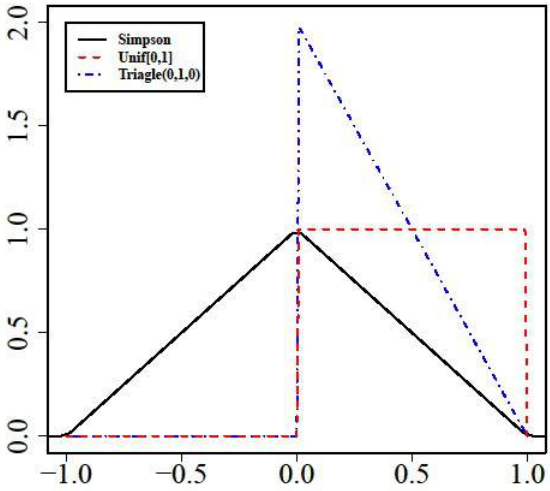

Рис. 6.4. Щільності рівномірного і трикутного розподілу

Трикутний розподіл у загальній формі інколи використовують для підгонки розподілу даних, зосереджених на скінченному інтервалі, особливо, якщо даних небагато і відомо, що щільність

має бути унімодальною (тобто мати єдиний максимум). Вибір саме трикутного розподілу для таких задач не має теоретичного обґрунтування і пов'язаний переважно з простотою запису щільності та функції розподілу.

6.2.7 Розподіли, пов'язані з гауссовим: χ^2, T і F

Є три типи розподілів, які не часто застосовуються для опису реальних даних, але дуже часто виникають при використанні статистичних алгоритмів: χ^2-розподіл, T-розподіл Стьюдента та F-розподіл Фішера. Наведемо їхні означення і коротко опишемо їхні властивості.

Розподіл χ^2. Кажуть, що випадкова величина ξ має розподіл χ^2 з ν ступенями вільності, якщо щільність її розподілу має вигляд:

$$f_\xi(x) = \begin{cases} \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)} & \text{якщо } x > 0, \\ 0 & \text{якщо } x < 0. \end{cases}$$

Розподіл χ^2 — це частковий випадок Γ-розподілу.

Позначення: $\xi \sim \chi^2_\nu$.

Якщо $\eta_1, \ldots, \eta_\nu$ — незалежні стандартні гауссові випадкові величини, то

$$\xi = \sum_{i=1}^{\nu} (\eta_i)^2$$

має χ^2-розподіл з ν ступенями вільності.

$$\mathsf{E}\,\xi = \nu,\ \mathsf{D}\,\xi = 2\nu.$$

Мнемоніка розподілу χ^2 у `R` — `chisq`, параметр `df=`ν — кількість ступенів вільності ν.

T-розподіл Стьюдента. Кажуть, що випадкова величина ξ має T-розподіл Стьюдента з ν ступенями вільності, якщо щільність її розподілу має вигляд

$$f_\xi(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Ця щільність є симетричною навколо 0. Розподіл має важкі хвости — вибірки з спостережень з T-розподілом часто мають викиди навіть тоді, коли вони не забруднені.

Позначення $\xi \sim \mathrm{T}_\nu$.

Якщо $\eta_0, \eta_1, \ldots, \eta_\nu$ — незалежні стандартні гауссові випадкові величини, то випадкова величина

$$\xi = \frac{\eta_0}{\sqrt{\frac{1}{\nu}\sum_{i=1}^{\nu}(\eta_i)^2}}$$

має T-розподіл Стьюдента з ν ступенями вільності.

$$\mathsf{E}\,\xi = 0, \;\; \mathrm{med}(\xi) = 0.$$

Якщо $\nu > 2$, то $\mathsf{D}\,\xi = \nu/(\nu-2)$. При $\nu \le 2$ скінченної дисперсії у ξ немає.

Мнемоніка T-розподілу Стьюдента у `R`: `t`. Параметр `df=`ν — кількість ступенів вільності.

F-розподіл Фішера. Цей розподіл має два параметри ν_1 — кількість ступенів вільності чисельника і ν_2 — кількість ступенів вільності знаменника. (Позначається $\xi \sim F(\nu_1, \nu_2)$). Така назва параметрів пов'язана з тим, що $F(\nu_1, \nu_2)$-розподіл має випадкова величина

$$\xi = \frac{\frac{1}{\nu_1}\sum_{i=1}^{\nu_1}(\eta_i)^2}{\frac{1}{\nu_2}\sum_{i=1}^{\nu_2}(\zeta_i)^2},$$

де η_i, ζ_j — незалежні в сукупності стандартні гауссові випадкові величини.

Щільність $F(\nu_1, \nu_2)$-розподілу:

$$f_\xi(x) = \frac{1}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/\nu_2} x^{\nu_1/2-1} \left(1 + \nu_1 x/\nu_2\right)^{-(\nu_1+\nu_2)/2}$$

при $x > 0$. (Тут $B(a,b)$ — бета-функція).

Математичне сподівання і дисперсія $F(\nu_1, \nu_2)$-розподілу:

$$\mathsf{E}\,\xi = \frac{\nu_2}{\nu_2 - 2} \text{ (при } \nu_2 > 2),$$

$$\mathsf{D}\,\xi = \frac{2(\nu_2)^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)} \text{ (при } \nu_2 > 4).$$

Мнемоніка F-розподілу у `R`: `f`. Параметри: `df1`=ν_1 — кількість ступенів вільності чисельника, `df2`=ν_2 — кількість ступенів вільності знаменника.

6.3 Дискретні розподіли на прямій

Дискретними називають випадкові величини, які можуть приймати лише значень з деякої фіксованої скінченної чи хоча б зліченної множини. Найбільш популярні дискретні розподіли відповідають випадковим величинам зі значеннями у множині цілих невід'ємних чисел.

Нагадаємо, що у дискретних розподілів немає щільності у звичайному розумінні (тобто відносно міри Лебега), але можна розглядати їхні щільності відносно лічильної міри. Таку щільність для дискретної випадкової величини ξ ми позначатимемо $f_\xi(x) = \mathsf{P}\{\xi = x\}$. Цю функцію називають розподілом ймовірностей для ξ.

6.3.1 Біноміальний розподіл

Біноміальний (binomial) розподіл традиційно описують як розподіл кількості успіхів у серії з n незалежних випробувань, якщо ймовірність успіху у кожному випробуванні дорівнює p. Позначення $\xi \sim \mathrm{Binom}(n,p)$. Зрозуміло, що n — натуральне число, $0 < p < 1$.

Розподіл ймовірностей ξ:

$$f_\xi(k) = \mathsf{P}\{\xi = k\} = C_n^k p^k (1-p)^{n-k},\ k = 0,\dots,n,$$

де

$$C_n^k = \frac{n!}{k!(n-k)!}$$

— кількість комбінацій з n по k (біноміальний коефіцієнт).

У `R` мнемоніка біноміального розподілу — `binom`, параметри — n =`size`, p =`prob`.

Математичне сподівання та дисперсія:

$$\mathsf{E}\,\xi = np,\ \mathsf{D}\,\xi = np(1-p).$$

Якщо $\xi_j \sim \mathrm{Binom}(n_i, p)$ — незалежні випадкові величини, то $\sum_{j=1}^{m} \xi_j \sim \mathrm{Binom}(\sum_{j=1}^{m} n_j, p)$.

Зрозуміло, що поняття "успіх" в означенні біноміального розподілу можна трактувати досить широко. Скажімо, нехай у деякому технологічному процесі ймовірність виробити бракований виріб дорівнює p і брак у окремих виробах виникає незалежно від інших виробів. Тоді кількість бракованих у серії n виробів буде мати $\mathrm{Binom}(n, p)$ розподіл, хоча отримання бракованої продукції важко назвати успіхом.

Якщо кількість випробувань n достатньо велика, а ймовірність успіху p не є близькою до 0 або 1, то біноміальний розподіл є близьким до нормального:

$$\mathsf{P}\{\xi < x\} \approx \Phi\left(\frac{x - nx}{\sqrt{np(1-p)}}\right)$$

(це наслідок центральної граничної теореми).

Якщо n — велике, а p — мале, причому величина $\lambda = np$ має помірне значення, біноміальний розподіл добре наближається розподілом Пуассона:

$$\mathsf{P}\{\xi = k\} \approx \frac{\lambda^k}{k!} e^{-\lambda},\ k = 0, 1, 2, \ldots$$

(це гранична теорема Пуассона).

Ми застосуємо біноміальний розподіл для аналізу цікавої історичної проблеми у прикладі 9.2.3.

6.3.2 Розподіл Пуассона

Розподіл Пуассона (Poisson distribution) задається ймовірностями

$$f_\xi(k) = \frac{\lambda^k}{k!} e^{-\lambda},\ k = 0, 1, 2, \ldots$$

де $\lambda > 0$ — параметр розподілу. Позначення $\xi \sim \mathrm{Poiss}(\lambda)$. У `R` мнемоніка розподілу Пуассона — `pois`, параметр λ =`lambda`.

Математичне сподівання та дисперсія:

$$\mathsf{E}\,\xi = \lambda,\ \mathsf{D}\,\xi = \lambda.$$

Якщо $\xi_j \sim \mathrm{Poiss}(\lambda_j)$ — незалежні випадкові величини, то

$$\sum_{j=1}^{m} \xi_j \sim \mathrm{Poiss}\left(\sum_{j=1}^{k} \lambda_j\right).$$

Розподіл Пуассона називають “розподілом рідкісних подій”. Це пов’язано з граничною теоремою Пуассона, яку неформально можна сформулювати таким чином.

Якщо кожна з великої кількості (n) незалежних подій може відбутись з малою ймовірністю p, то кількість подій що відбулись має приблизно пуассонів розподіл з параметром $\lambda = np$.

Наприклад, спонтанний розпад ядер радіоактивних елементів відбувається випадково, причому ядра розпадаються зі сталою інтенсивністю незалежно одне від одного. Ймовірність того, що конкретне ядро розпадеться, мала, але кількість ядер у зразку речовини — велика. Тому кількість радіоактивних розпадів у зразку, зареєстрована рахівником Гейгера протягом певного часу, має Пуассонів розподіл. Параметр λ у цьому прикладі можна трактувати як середнє значення, навколо якого будуть коливатись кількості розпадів у довгій серії однотипних вимірювань. Якщо розпадів багато, за центральною граничною теоремою розподіл їхньої кількості має бути приблизно нормальним.

І дійсно, розподіл Пуассона наближається до нормального при великих λ. Тому використовувати його для опису даних доцільно лише при невеликих λ, коли кількості спостережуваних подій що відбулись невеликі (не перевищують 2-3 десятки). Звідси і назва — розподіл рідкісних подій.

У прикладі 9.6.3 ми побачимо, що цей розподіл добре описує результати обстрілу Лондона ракетами Фау-2 у II світовій війні.

6.3.3 Геометричний розподіл

Геометричний розподіл (geometric distribution) — це розподіл кількості невдач у серії незалежних випробувань, які треба провести до першого успіху, якщо ймовірність успіху у кожному випробуванні дорівнює p. Позначення $\xi \sim \mathrm{Geom}(p)$.

Ймовірності геометричного розподілу:

$$f_\xi(k) = \mathsf{P}\{\xi = k\} = p(1-p)^k \; k = 0, 1, 2, \ldots$$

Інколи використовують інше означення геометричного розподілу, в якому розраховується кількість всіх випробувань, потрібних для отримання успіху, включаючи те останнє, яке виявилось успішним, тобто розглядається величина $\eta = \xi + 1$ з розподілом $f_\eta(x) = \mathsf{P}\{\xi = k\} = p(1-p)^{k-1}$, $k = 1, 2, \ldots$. Яке означення використовується, можна зрозуміти з контексту.

Математичне сподівання та дисперсія геометричного розподілу:

$$\mathsf{E}\,\xi = \frac{1-p}{p}, \; \mathsf{D}\,\xi = \frac{1-p}{p^2}.$$

У `R` мнемоніка геометричного розподілу `geom`, параметр (ймовірність успіху) p =`prob`.

Геометричний розподіл відіграє для дискретних випадкових величин роль, аналогічну ролі експоненційного розподілу для неперервних випадкових величин. Він є єдиним дискретним розподілом

з властивістю відсутності післядії. Тому його природно використовувати для опису часу очікування подій, які відбуваються внаслідок зовнішніх причин, не пов'язаних зі змінами досліджуваного об'єкта.

Скажімо, нехай у прикладі з тривалістю роботи електричної лампочки від вкручування до перегоряння з п. 6.2.4 час вимірюється у цілих днях, які лампочка пропрацювала. Тоді відповідна випадкова величина матиме геометричний розподіл, якщо лампочка перегоряє під впливом зовнішніх випадкових обставин, незалежно від часу її попередньої роботи. У такій трактовці "успіх" — це перегоряння лампочки у даний день, p — ймовірність цієї події, котра вважається сталою.

6.4 Комбінації та перетворення розподілів

Звичайно, можливі розподіли випадкових даних не обмежуються переліченими у попередніх підрозділах. Одна з можливостей більш гнучкого моделювання розподілу реальних даних — комбінування кількох стандартних розподілів або перетворення випадкових величин із заданим розподілом. Інколи таку техніку використовують без змістовного обґрунтування: перетворення/комбінацію обирають так, щоб вона відповідала реально спостережуваному, наприклад, на гістограмі розподілу даних. Якщо інших можливостей немає, така техніка допустима.

Проте ми зараз зосередимось на іншому підході, коли модель розподілу обирається на основі певних теоретичних міркувань про природу випадковості досліджуваних даних.

6.4.1 Зрізані розподіли

Нехай випадкова величина ξ має розподіл F. Зрізаним[6] (обмеженим) на інтервал $[a, b)$ розподілом називають умовний розподіл ξ

[6] Truncated distribution.

за умови, що $\xi \in [a, b)$. Функція розподілу для зрізаного розподілу

$$F_{[a,b)}(x) = \mathsf{P}\{\xi < x \mid \xi \in [a.b)\} = \frac{F_\xi(x)}{F_\xi(b) - F_\xi(a)},$$

де $F_\xi(x) = \mathsf{P}\{\xi < x\}$ — функція розподілу ξ.

Якщо ξ має щільність розподілу f_ξ, то щільність зрізаного розподілу

$$f_{[a,b)}(x) = \frac{f_\xi(x)}{F_\xi(b) - F_\xi(a)}.$$

Приклад 6.4.1. Зрізаний експоненційний розподіл.

Нехай дані з ремонтної майстерні являють собою тривалість роботи деяких приладів від моменту продажу до повернення для ремонту по гарантії виробника. Гарантія дається на 3 роки з моменту продажу. Визначимо, яким може бути розподіл такої тривалості.

Зазвичай гарантія видається на такий термін, за який ефекти старіння приладу не створюють додаткових загроз його функціонуванню. Тому основними причинами відмови мають бути випадкові зовнішні обставини. У цьому випадку, як ми бачили у п. 6.2.4, розподіл часу від початку експлуатації до відмови природно описувати експоненційним розподілом. Але наші дані містять тривалості роботи лише тих приладів, у яких відмова сталась до закінчення дії гарантії. Як працювали прилади, що пережили гарантійний термін — нам невідомо.

Отже, розподіл спостережуваних даних — це умовний розподіл часу до відмови, за умови, що він птрапляє у інтервал $[0, 3)$ (якщо час вимірюється у роках). Отримуємо зрізаний експоненційний розподіл з функцією розподілу

$$F_\xi(x) = F(x; \lambda, c) = \begin{cases} 0 & \text{при } x < 0, \\ \frac{1-e^{-\lambda x}}{1-e^{-\lambda c}} & \text{при } 0 < x \le c, \\ 1 & \text{при } x > c, \end{cases}$$

де c — "поріг зрізання" (у нашому прикладі $c = 3$). Щільність

цього розподілу

$$f_\xi(x) = f(x;\lambda,c) = \begin{cases} 0 & \text{при } x < 0, \\ \frac{\lambda e^{-\lambda x}}{1-e^{-\lambda c}} & \text{при } 0 < x \le c, \\ 0 & \text{при } x > c. \end{cases}$$

Щільність зрізаного експоненційного розподілу порівняно з експоненційною щільністю див. на рис. 6.5.

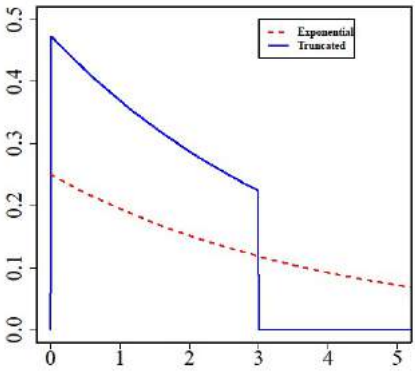

Рис. 6.5. Щільність експоненційного (Exponential) та зрізаного експоненційного (Truncated) розподілів

Для генерації псевдовипадкових чисел зі зрізаним експоненційним розподілом можна використати означення цього розподілу як умовного. Функцію `rexptr()`, що реалізує цю ідею, можна задати таким скриптом:

```
> rexptr<-function(lambda=1,trun=1){
+  repeat{ x<-rexp(1,lambda); if(x<trun) break}
+ }
```

— у циклі `repeat` генерація експоненційно розподілених псевдовипадкових чисел з інтенсивністю `lambda` продовжується доти, доки чергове число не виявиться меншим, ніж поріг зрізання `trun`. Це число стає значенням, яке повертає функція. ◀

Ми використаємо зрізаний експоненційний розподіл у ряді прикладів, зокрема, у прикладі 8.1.4.

Зрозуміло, що зрізання можливе не тільки для експоненційного, а і для будь-якого іншого розподілу. Наприклад, півнормальний розподіл можна трактувати як зрізаний нормальний на інтервалі $[0, \infty)$.

Приклад 6.4.2. Пуассонів розподіл зі зрізаним нулем. Розглянемо знову майстерню з ремонту деяких приладів. Нехай для кожного ремонтованого приладу у майстерні записують, скільки різних дефектів у ньому було виявлено. Яким може бути розподіл кількості дефектів в одному приладі?

Якщо вважати, що дефекти виникають незалежно один від одного, причому ймовірність появи кожного дефекту мала, то ми маємо справу з кількістю рідкісних подій, тобто можна сподіватись розподілу Пуассона. Але зрозуміло, що прилад з нульовою кількістю дефектів не потрапить до ремонтної майстерні. Тому природно розглянути як кандидата на роль розподілу таких даних пуасонів, але зі зрізаним нулем[7]:

$$\mathsf{P}\{\text{кількість дефектів} = k\} = \mathsf{P}\{\xi = k \,|\, \xi \neq 0\} = \frac{\lambda^k}{k!(e^\lambda - 1)},$$

де $\xi \sim \mathrm{Poiss}(\lambda)$.

Математичне сподівання і дисперсія випадкової величини η з цим розподілом:

$$\mathsf{E}\,\eta = \frac{\lambda}{1 - e^{-\lambda}}, \mathsf{D}\,\eta = \frac{\lambda}{1 - e^{-\lambda}} \left(1 - \frac{\lambda}{e^\lambda - 1}\right).$$

Статистика таких розподілів обговорюється у прикладах 8.1.5 та 9.4.1. ◀

[7] Zero-truncated Poisson distribution.

6.4.2 Суми незалежних випадкових величин

Інколи буває природно трактувати спостережувану випадкову величину як суму/різницю двох або більше незалежних випадкових величин. Наприклад, якщо деяка велична ξ вимірюється з похибкою вимірювання ϵ, причому розподіл похибки не залежить від значення ξ, то спостережувана величина $X = \xi + \varepsilon$ буде сумою двох незалежних випадкових величин.

Нехай ξ та η — незалежні випадкові величини, розподіл ξ є F, а розподіл η — G. Тоді розподіл H їх суми X є згорткою[8] F і G (позначається $H = F \star G$). Для функцій розподілу згортка визначається як

$$H(x) = \int_{-\infty}^{\infty} F(x-t)G(dt) = \int_{-\infty}^{\infty} G(x-t)F(dt).$$

Для щільностей розподілу:

$$h(x) = \int_{-\infty}^{\infty} f(x-t)g(t)dt = \int_{-\infty}^{\infty} g(x-t)f(t)dt.$$

Якщо ξ та η — дискретні випадкові величини, які набувають лише цілих значеннь, то

$$\mathsf{P}\{X = k\} = \sum_{j=-\infty}^{\infty} \mathsf{P}\{\xi = k-j\}\,\mathsf{P}\{\eta = j\} =$$

$$= \sum_{j=-\infty}^{\infty} \mathsf{P}\{\eta = k-j\}\,\mathsf{P}\{\xi = j\}.$$

Розглянемо кілька прикладів, коли природно використовувати згортки для опису розподілу спостережень. Ці приклади матимуть дещо штучний вигляд, оскільки підбирались з демонстраційною метою, однак їхні більш складні аналоги досить часто зустрічаються у медико-біологічній та економічній статистиці.

Приклад 6.4.3. Досліджується вплив малих доз іонізуючого опромінення на утворення певної мутації у культурі клітин. Зразки

[8] Convolution.

культури опромінюються з низькою інтенсивністю протягом тривалого часу. Мутація виявляється у момент поділу мутантної клітини. Спостережувана величина — час від початку опромінення до виявлення мутації.

Мутації під дією опромінення виникають тоді, коли опромінення потрапляє у "мішень" — молекулу ДНК. Це може відбутись випадково у будь-який момент часу. При цьому розподіл залишкового часу до моменту мутації не залежить від того, скільки на неї чекали. Отже, для часу від початку експерименту до виникнення мутації (ξ) природним буде експоненційний розподіл. Але після виникнення повинен іще пройти час до поділу клітини, коли мутація буде виявлена. Нехай кожна клітина ділиться точно через два тижні після попереднього поділу. Тоді час від виникнення мутації до моменту поділу (η) природно вважати рівномірно розподіленим на двотижневому інтервалі. Часові інтервали ξ і η є незалежними[9] і їх треба додати, щоб отримати спостережувану величину — час від початку експерименту до виявлення мутації $X = \xi + \eta$.

Маємо $\xi \sim \mathrm{Exp}(\lambda)$, $\eta \sim \mathrm{Unif}[0, c]$. Щільність розподілу X

$$f^X(x) = \int f^\xi(x-t) f^\eta(t) dt = \int_0^c \frac{1}{c} \lambda e^{-\lambda(x-t)} \mathbb{1}\{x - t > 0\} dt =$$

$$= \frac{\lambda}{c} e^{-\lambda x} \int_0^{\min(c,x)} e^{\lambda t} dt =$$

$$= \frac{1}{c} e^{-\lambda x} (e^{\lambda \min(c,x)} - 1).$$

Графік f^X для $\lambda = 1/2$, $c = 2$ зображено на рис. 6.6.

[9] Якщо мутація ніяк не впливає на початок поділу.

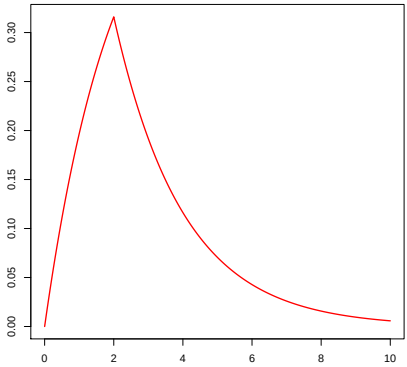

Рис. 6.6. Щільність суми експоненційної та рівномірної випадкових величин

Цей графік можна відобразити таким скриптом:

```
> l<-0.5
> c0<-2
> x<-seq(0,10,0.01)
> y<-sapply(x,function(x)
+              {exp(-l*x)*(exp(l*min(c0,x))-1)/c0}
+           )
> plot(x,y,lwd=2,col="red",
+ type="l",xlab="x",ylab="density")
```

◀

Приклад 6.4.4. Два приятелі Андрій і Василь стоять на пероні метро, чекаючи на поїзди. Вони збираються їхати у протилежних напрямках. В обох напрямках поїзди йдуть з інтервалом 7 хв, причому інтервал між приходом поїзда "туди" і наступним поїздом "назад"[10] рівномірно розподілений на [0,7].

[10]Цей інтервал, вочевидь, є одним і тим самим для всіх поїздів, що дотримую-

Спостерігається величина X — час, який пройде між від'їздом Андрія і від'їздом Василя, причому, якщо Андрій виїхав першим, ця величина додатна, і від'ємна, якщо першим поїхав Василь. Яким буде розподіл X?

Якщо приятелі прийшли не спеціально до відправлення якогось поїзда, а у випадковий момент часу, не пов'язаний з розкладом руху, то час чекання до відправлення кожного з них є рівномірним на [0,7]. За умовою задачі ці часи незалежні між собою і X — їхня різниця. Отже, X має симетричний трикутний розподіл на інтервалі $[-7, 7]$. ◄

Приклад 6.4.5. Турист, що подорожує автостопом, чекає на шосе попутну машину, щоб проїхати до міста, на вокзал. Яким буде розподіл часу X від початку чекання до приїзду на вокзал?

Цей час складається з двох доданків: ξ — час чекання на зупинку попутної машини і η — час, витрачений на дорогу. Для ξ природно припускати експоненційний розподіл (попутні машини їздять по шосе незалежно від того, чекає їх турист чи ні), для η — нормальний розподіл (затримки у дорозі можуть виникати з різних, більш-менш незалежних між собою причин і підсумовуються у загальній тривалості поїздки). Величини ξ і η природно вважати незалежними.

Отже, $\xi \sim \mathrm{Exp}(\lambda)$, $\eta \sim N(\mu, \sigma^2)$, $X = \xi + \eta$.

Щільність розподілу X

$$f^X(x) = \int_0^\infty f^\xi(t) f^\eta(t-x)dt =$$

$$= \frac{\lambda}{\sqrt{2\pi}\sigma} \int_0^\infty e^{-\lambda t} e^{-\frac{(x-t-\mu)^2}{2\sigma^2}} dt =$$

$$= \lambda \exp\left(\frac{\lambda}{2}(2\mu + \lambda\sigma^2 - 2x)\right)\left(1 - \Phi\left(\frac{\mu + \lambda\sigma^2 - x}{\sigma}\right)\right).$$

Графік f^X для $\lambda = 3/4$, $\mu = 1$, $\sigma = 0.5$ зображено на рис. 6.7.

ться цього розкладу. Припустимо, що він випадково встановлюється вранці, коли запускають перші поїзди.

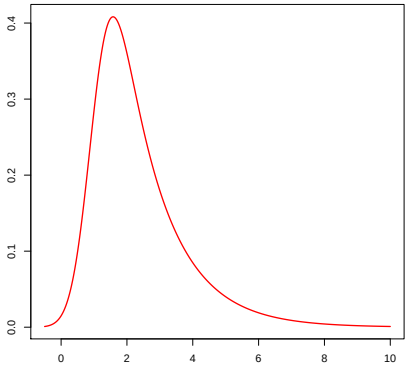

Рис. 6.7. Щільність суми експоненційно та нормально розподіленої випадкових величин

Цей графік можна відобразити таким скриптом:

```
> l<-0.75
> m<-1
> s<-0.5
> x<-seq(-0.5,10,0.01)
> y<-sapply(x,function(x){
+   l*exp(0.5*l*(2*m+l*s^2-2*x))*
+     (1-pnorm((m+l*s^2-x)/s))
+ })
> plot(x,y,lwd=2,col="red",
+      type="l",xlab="x",ylab="density")
```

◀

Приклад 6.4.6. Нехай випадкова величина ξ, що має рівномірний розподіл на $[0, c]$ вимірюється з похибкою $\eta \sim N(\mu, \sigma^2)$. Тоді

щільність розподілу спостережуваної величини $X = \xi + \eta$ буде

$$f^X(x) = \frac{1}{c}\left(\Phi\left(\frac{c+m-x}{\sigma}\right) - \Phi\left(\frac{m-x}{\sigma}\right)\right).$$

Графік f^X для $c = 5$, $\mu = 1$, $\sigma = 0.5$ зображено на рис. 6.8.

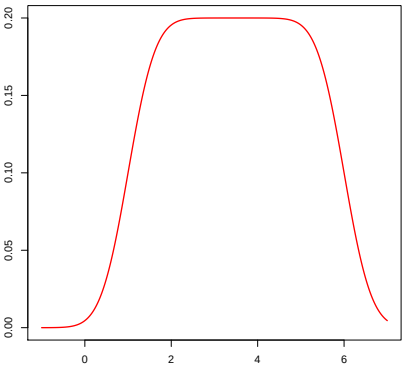

Рис. 6.8. Щільність суми рівномірно та нормально розподіленої випадкових величин

◄

6.4.3 Суміші кількох розподілів

Нехай кожен досліджуваний об'єкт O належить одній з M різних популяцій $\mathcal{P}_1, \ldots \mathcal{P}_M$. Ми спостерігаємо певну характеристику об'єкта $\xi = \xi(O)$. Припустимо, що розподіл $\xi(O)$ залежить від того, до якої популяції належить досліджуваний об'єкт. Позначимо F_m функцію розподілу $\xi(O)$ для об'єктів, що належать m-й популяції:

$$F_m(x) = \mathsf{P}\{\xi(O) < x \,|O \in \mathcal{P}_m\}.$$

З якої саме популяції отримано об'єкт — невідомо. Але ми припускаємо, що популяція $\mathcal{P}_m$ обирається випадково з ймовірністю p_m:

$$p_m = \mathsf{P}\{O \in \mathcal{P}_m\}.$$

Тоді розподіл спостережуваного ξ буде сумішшю (mixture) розподілів F_m:

$$F^{\xi}(x) = \mathsf{P}\{\xi < x\} = \sum_{m=1}^{M} p_m F_m(x).$$

Популяції $\mathcal{P}_m$ та розподіли F_m називають компонентами суміші, p_m — змішуючими ймовірностями (mixing probabilities). Зрозуміло, що повинні виконуватись умови

$$\sum_{m=1}^{M} p_m = 1,\ p_m \geq 0,\ m = 1, \ldots, M.$$

Якщо F_m мають щільності f_m, то ξ має щільність

$$f^{\xi}(x) = \sum_{m=1}^{M} p_m f_m(x).$$

Ці рівності задають ймовірнісну модель скінченної суміші. Наприклад, на рис. 6.9 зображено графіки щільності двокомпонентної суміші з гауссовими компонентами.

Блакитна (переривчаста) крива на рисунку відповідає щільності

$$f(x) = \frac{1}{2}\varphi(x-1) + \frac{1}{2}\varphi(x+1),$$

(φ — щільність $N(0,1)$) тобто суміші двох нормальних розподілів з одиничною дисперсією та математичними сподіваннями -1 і 1. Змішуючі ймовірності $p_1 = p_2 = 1/2$.

Червона (суцільна) крива відповідає суміші двох розподілів: $F_1 \sim N(-2, 1/4)$ і $F_2 \sim N(2, 3)$ зі змішуючими ймовірностями $p_1 = 0.4$, $p_2 = 0.6$.

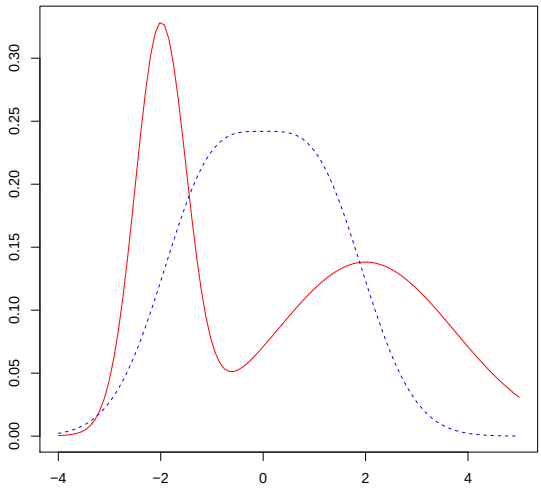

Рис. 6.9. Щільності двокомпонентних сумішей з нормального розподілу

На цій кривій помітні два “піки” — локальні максимуми, розділені мінімумом. Щільності, що мають більше одного локального максимуму називають *багатомодальними*[11] Багатомодальність часто (хоча і не завжди) пов’язана з наявністю у даних суміші кількох компонент.

Статистичні задачі, у яких спостережувані дані природно описувати такими моделями суміші, розглянуті далі у прикладах 8.1.6 і 9.2.2.

6.5 Генерація псевдовипадкових чисел

У статистиці часто виникає потреба отримати послідовність чисел, які описуються певною ймовірнісною моделлю. У прості-

[11] А щільності з єдиним максимумом — *унімодальними*.

шому випадку це може бути послідовність незалежних, однаково розподілених випадкових величин. З точки зору класичної теорії ймовірностей, випадковість є властивістю не конкретної числової послідовності, а способу, у який ця послідовність була отримана.

Наприклад, нехай проводиться дослід з вимірюванням кількості радіоактивних розпадів у певному зразку речовини за 1 хв. Протягом 8 хв досліду зафіксовані значення 3 — за першу хвилину, 1 — за другу і далі 4, 1, 5, 9, 2, 6. У сучасній фізиці вважається, що спонтанний розпад атомних ядер відбувається випадково, незалежно у різних ядрах зі сталою інтенсивністю. Тому отримана у досліді послідовність є випадковою. Ця сама послідовність, знайдена комп'ютерною програмою при обчисленні знаків числа π, випадковою не є.

З цієї точки зору, всі послідовності чисел, які можна згенерувати на звичайному комп'ютері без використання яких-небудь зовнішніх джерел випадковості, не є випадковими[12]. Але можна розглядати алгоритми, що генерують послідовності, які імітують випадковість, тобто мають основні властивості, притаманні послідовностям випадкових величин. Такі алгоритми і програми, що їх реалізують, називають *генераторами* (датчиками) псевдовипадкових чисел (pseudorandom numbers generators). Префікс псевдо- інколи пропускають і кажуть про генерацію випадкових чисел. Це не є помилкою, якщо пам'ятати про імітаційний характер такої випадковості.

Останнім часом набула розвитку техніка генерування квазівипадкових чисел (quasirandom numbers) — послідовностей, що поєднують деякі властивості випадкових з такими особливостями, яких справжні випадкові послідовності мати не можуть. Зокрема, такі числа використовуються при наближеному інтегруванні багатовимірних функцій за методом Монте-Карло (див. підрозд. 7.7 у книзі [43]). У даній книзі ця тематика не розглядається.

Зазвичай генерація псевдовипадкових чисел починається зі

[12] Про альтернативні підходи, які дозволяють визначити, чи є випадковою певна нескінченна числова послідовність, див. [4].

створення рівномірних чисел, тобто послідовності, яка імітує поведінку послідовності незалежних, однаково розподілених випадкових величин з рівномірним розподілом на $[0, 1]$. Потім, використовуючи ті чи інші перетворення цієї послідовності, отримують псевдовипадкові послідовності із заданим розподілом, наприклад, нормальні або такі, що утворюють ланцюг Маркова із заданими ймовірностями переходу.

Генерація рівномірних псевдовипадкових послідовностей має вже більш ніж 70-річну історію, тут відібрані найкращі генератори, які і реалізовані у базовому `R`. Намагатись самостійно покращити їх без глибокого знання відповідної теорії та власного досвіду у цій області не варто. Але я включив у цю книгу елементарні відомості про таку генерацію, щоб читач мав змогу, по-перше, зрозуміти, як відбувається генерація у стандартних програмах, а по-друге, при бажанні створити свій власний генератор, якщо раптом виникне недовіра до стандартного. Ці відомості вміщені у п. 6.5.1.

У `R` реалізовані також функції, що дозволяють отримати послідовності, які імітують кратні вибірки з такими основними ймовірнісними розподілами, як нормальний, експоненційний, пуассонів тощо. Але цих функцій може бути недостатньо, якщо вам треба згенерувати псевдовипадкову послідовність з яким-небудь менш поширеним розподілом, наприклад, з розподілом Парето. Тому розуміння загальних підходів до такої генерації є важливим елементом роботи статистика. З елементарними відомостями про це можна ознайомитись у п. 6.5.2. Коротко про те, як ці техніки генерації реалізовані у `R`, можна прочитати у п.6.5.3.

6.5.1 Генератори рівномірних псевдовипадкових чисел

Отже, рівномірні псевдовипадкові числа — це числові послідовності, які відтворюють основні властивості послідовностей незалежних однаково розподілених випадкових величин з рівномірним розподілом на $[0, 1]$. Для створення таких послідовностей зазвичай використовують рекурсивну техніку. При цьому задаються

деякі значення початкових елементів послідовності x_1, x_2,... x_k і функція $f(t_1, \ldots, t_k)$, що породжує наступний елемент послідовності. Після цього послідовність визначається як

$$x_{k+1} = f(x_1, \ldots, x_k),$$

$$x_{k+2} = f(x_2, \ldots, x_{k+1}),$$

...

$$x_n = f(x_{n-k}, \ldots, x_{n-1}),$$

...

У найпростішому випадку $k = 1$, тобто кожен наступний елемент послідовності визначається за попереднім:

$$x_n = f(x_{n-1}).$$

При виборі функції f у першу чергу керуються міркуваннями простоти ралізації та швидкості виконання. Найбільш поширена сукупність генераторів — лінійні конгруентні генератори, у яких функція f будується з використанням лінійної залежності зі сталими коефіцієнтами. Розрізняють два типи генераторів: з цілочисловою та дійснозначною арифметикою.

У генераторі з дійснозначною арифметикою x_1 вибирають з інтервалу $(0, 1)$ і послідовність породжується за правилом

$$x_n = \{ax_{n-1} + c\},\ n = 2, 3, \ldots,$$

де $\{x\}$ — дробова частина числа x. Тут a і c — фіксовані дійсні числа.

У генераторі з цілочисловою арифметикою спочатку будується допоміжна послідовність натуральних чисел

I_1, I_2,... I_n,...

Початкове число I_1 вибирають з інтервалу 1,...,$m - 1$, послідовність формується за правилом

$$I_n = (aI_{n-1} + c) \mod m,\ n = 2, 3, \ldots,$$

де a, c та m — фіксовані натуральні числа. Якщо $c = 0$, генератор називають мультиплікативним. Послідовність дійсних чисел з інтервалу $[0, 1]$ отримують з I_n діленням на m:

$$x_n = \frac{I_n}{m}.$$

Число a називають множником, c — приростом, а m — модулем генератора.

У сучасних генераторах зазвичай використовують цілочислові схеми, оскільки правила округлення у дійснозначній арифметиці відрізняються на різних комп'ютерах. Тому один і той самий дійснозначний генератор може на одному комп'ютері давати хорошу послідовність, а на іншому — погану. Цілочислова арифметика на всіх комп'ютерах реалізована однаково (якщо організувати обчислення без переповнень). З цієї точки зору цілочислові генератори є більш надійними.

Нехай I_j, $j = 1, 2, \ldots$ — послідовності, згенеровані лінійним конгруентним генератором з цілочисловою арифметикою. Зрозуміло, що якщо при деяких n і k, $I_n = I_{n-k}$, то для всіх $i = 1, 2, \ldots$ буде виконано $I_{n+i} = I_{n-k+i}$, тобто послідовність буде циклічно повторюватись. Найменше k, при якому це буде виконуватись, називають періодом (або довжиною періоду) генератора. Очевидно, що циклічно повторювана послідовність не може вважатись випадковою, тому генератори не доцільно використовувати для породження послідовностей з довжиною, більшою, ніж період. Отже, хороший генератор мусить мати великий період.

Оскільки між 0 і $m - 1$ є рівно m чисел, період цілочислового лінійного конгруентного генератора не може бути більшим, ніж m. Відомі умови на параметри генератора, при яких він має найбільший період (тобто m, див. [10]):

Теорема 6.5.1. *Для того, щоб цілочисловий лінійний конгруентний генератор мав період m, необхідно і достатньо, щоб виконувались умови:*

1. c і m взаємно прості.

2. Всі прості дільники m є дільниками $a-1$.

3. Якщо m кратне 4, то $a-1$ також кратне 4.

Але не всякий генератор з максимальним періодом буде вдалим. Дійсно, послідовність 1, 2,... m має період m, але на випадкову вона не схожа. Тому для оцінки якості генератора необхідно проводити спеціальні "тести на випадковість". Такі тести будують за звичайною схемою статистичних тестів для перевірки того, що певна послідовність даних відповідає обраній імовірнісній моделі. Ми зупинимось зараз лише на двох елементарних графічних способах перевірки якості генератора псевдовипадкових чисел.

Нехай $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ — послідовність незалежних, рівномірно на $[0,1]$ розподілених псевдовипадкових чисел. Емпіричною функцією розподілу даних $\mathbf{X}$ називають

$$\hat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{\xi_j < x\}.$$

Зрозуміло, що $\hat{F}_n(x)$ — це відносна частота $(-\infty, x)$ у вибірці. За законом великих чисел, при великих n, $\hat{F}_n(x) \approx F(x)$, де $F(x)$ — функція розподілу для рівномірного розподілу на $[0,1]$, тобто

$$F^{U[0.1]}(x) = \mathsf{P}\{\xi_1 < x\} = \begin{cases} 0 & \text{при } x < 0, \\ x & \text{при } 0 \le x \le 1, \\ 1 & \text{при } x > 1. \end{cases}.$$

Для графічної перевірки якості генератора можна відобразити на одному графіку емпіричну функцію розподілу згенерованої псевдовипадкової послідовності та $F^{U[0.1]}(x)$. Якщо вони будуть близькими одна до одної, генератор пройшов це випробування. Якщо помітне систематичне відхилення емпіричної функції від теоретичної, генератор не адекватний.

Наступний приклад демонструє, як працює лінійний конгруентний генератор з цілочисловою арифметикою і параметрами $a = 65539$, $c = 0$, $m = 2^{31}$ з початковим значенням $I_1 = 2^{15} + 2$. Кількість спостережень $n = 200$.

Цей генератор був досить популярним у 60-70-ті роки XX ст. під назвою RANDU, зокрема, використовувався як стандарт на комп'ютерах фірми IBM[13].

```
>   n<-200   # кількість чисел
>   a<-65539 # RANDU параметри
>   c0<-0     #
>   m<-2^31  #
>   I<-numeric(n) # цілочислова послідовність
>   I[1]<-2^15+2
>   for(i in 2:n){
+   I[i]<-(a*I[i-1]+c0)%% m
+   }
>   x<-I/m # псевдовипадкові числа
>   plot(1:n,x,cex=0.3) # рисуємо діаграму чисел
>   sx<-sort(x)
>   # рисуємо емпіричну функцію розподілу:
>   plot(sx,(1:n)/n,type="s",xlim=c(0,1),ylim=c(0,1))
>   # графік теоретичної функції розподілу:
>   abline(a=0,b=1,col="red")
```

Результати роботи відображені на рис. 6.10. Зверху — діаграма, у якій координати точок по горизонталі відповідають номеру псевдовипадкового числа, а по вертикалі — його значенню. Знизу — емпірична функція розподілу, побудована за псевдовипадковою послідовністю.

[13] Допитливий читач може помітити, що тут ми, оперуючи з цілими числами, використовуємо дійснозначну арифметику, оскільки тип `I` не `integer`, а `numeric`. Це зроблено тому, що звичайний тип `integer` у `R` передбачає занадто короткі числа для даного прикладу. У `R` є пакети, що дозволяють вводити цілі довільної довжини, але у цій книзі ми не будемо їх використовувати. Для розглядуваних ілюстративних прикладів генерації псевдовипадкових чисел дійснозначна арифметика дає достатньо адекватні результати.

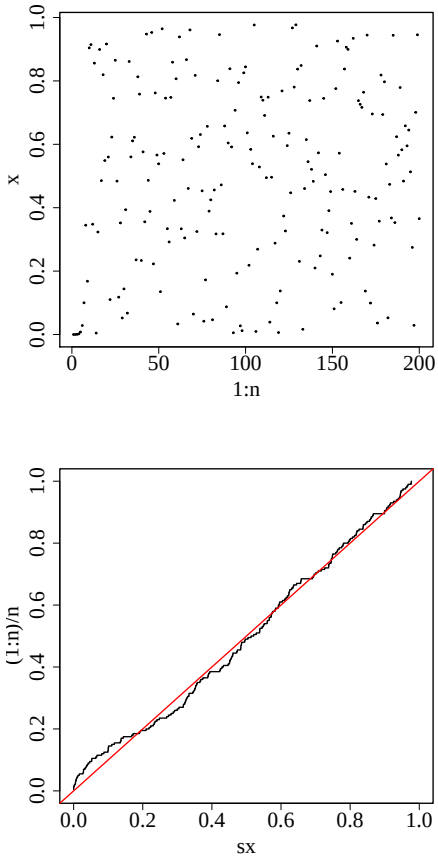

Рис. 6.10. Генератор RANDU: розкид та емпірична функція розподілу

Рисунок зверху демонструє “випадкову” поведінку послідовності: не помітно яких-небудь закономірностей, що свідчили б про невипадковість. Рисунок знизу показує рівномірність розподілу — емпірична функція розподілу коливається навколо теоретичної. При збільшенні довжини послідовності відхилення емпіричної функції від теоретичної стають все менш помітними. Можна

вважати, що цей тест генератор RANDU пройшов.

Іще один важливий вид тестів — графічна перевірка залежності двох або трьох сусідніх елементів послідовності на графіку пар/трійок. Для того, щоб побачити залежності, на площині рисують точки з координатами

$(x_j, x_{j+1}), j = 1, \ldots, n-1$

або у тривимірному просторі — з координатами

$(x_j, x_{j+1}, x_{j+2}), j = 1, n-2.$

На відповідних діаграмах намагаються знайти закономірності, що відрізняють поведінку послідовонсті від справжньої випадкової. Для лінійних конгруентних генераторів такою закономірністю є розташування точок вздовж невеликої кількості прямих ліній на площині або площин — у тривимірному просторі. Зрозуміло, що така особливість генератора свідчить про невипадковість.

Продовжуючи попередній приклад, ці тести можна реалізувати так:

```
x1<-x[1:(n-2)]
x2<-x[2:(n-1)]
x3<-x[3:n]
library(rgl)
plot3d(x1,x2,x3)     # 3D-графіка
plot(x1,x3,cex=0.3) # точки на площині
```

Результати тестів — на рис. 6.11. На двовимірній діаграмі розсіювання не видно закономірностей, що характеризували б послідовність як не випадкову: точки розкидані хаотично і заповнюють квадрат з приблизно однаковою щільністю. Отже, цей тест пройдений. На тривимірній картинці також спочатку закономірності не були помітні, але після повороту вдалось отримати те, що зображено на рис. 6.11 знизу: точки розташовані на кількох (приблизно 15) площинах всередині кубу. Зрозуміло, що така поведінка не відповідає уявленням про незалежні випадкові величини з рівномірним розподілом, отже, цей тест генератор RANDU не пройшов. Саме

тому його зараз не використовують для генерації псевдовипадкових чисел.

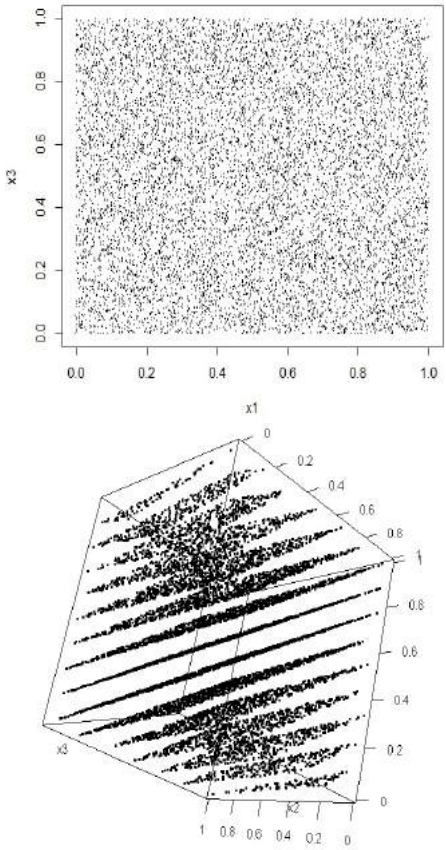

Рис. 6.11. Діаграми розсіювання пар та трійок для RANDU

Насправді, всі лінійні конгруентні генератори дають послідовності, що породжують тривимірні структури, подібні до виявлених нами у генератора RANDU. Але у хороших генераторів кількість

площин, на яких розташовуються точки, велика, і ці площини розміщені поруч одна від одної, тому такі генератори проходять цей тест.

У книзі [43] як "мінімальний стандарт" рекомендовано використовувати генератор Парка та Мілера з $a = 7^5$, $c = 0$, $m = 2^{31} - 1$. Цей генератор проходить описані нами тести, а також більшість тестів, які прийнято застосовувати до таких генераторів. Його період $2^{31} - 2 \approx 2.1 \times 10^9$. Це велике число, але для деяких застосувань воно може бути недостатнім.

Існують складніші техніки генерації псевдовипадкових послідовностей, що мають значно більші періоди. Наприклад, у підрозд. 7.1 книги [43] розглядається техніка комбінування двох лінійних конгруентних генераторів з різними періодами, яка дозволяє отримати послідовність з періодом, не меншим ніж найменше спільне кратне комбінованих генераторів.

Ще один спосіб генерації псевдовипадкових чисел, що набув популярності останнім часом, — генератори Фібоначчі із запізненням (lagged Fibonacci generator), у яких для породження чергового елемента послідовності використовується не один попередній елемент, а два, взяті з фіксованим запізненням. Наприклад, адитивний генератор Фібоначчі має вигляд

$$I_n = (I_{n-k} + I_{n-l}) \mod m,$$

де $k < l$ фіксовані числа (лаги). Для створення послідовності цим генератором треба задати не один, а l початкових елементів, після чого можна використовувати генеруючу формулу. Модуль m традиційно вибирають ступенем двійки: $m = 2^b$. При правильному виборі лагів цей генератор дозволяє отримати період $2^{b-1}(2^l - 1)$. Прикладами "хороших" лагів є $k = 7$, $l = 10$ або $k = 5$, $l = 17$.

Подальші відомості про генератори рівномірних послідовностей можна знайти у книзі Д. Кнута [10]. У книзі [43] описані програмні реалізації таких генераторів мовою C та приклади застосування.

6.5.2 Генерація псевдовипадкових чисел із заданим розподілом

Якщо деяким генератором створена псевдовипадкова послідовність з рівномірним розподілом, то отримати з неї послідовність, що імітує незалежні випадкові величини з іншим розподілом можна, використовуючи різні перетворення. При цьому початкова послідовність лише імітує випадковість, але це ігнорується. Тобто у цьому підрозділі ми будемо трактувати початкову послідовність $\eta_1, \ldots \eta_n, \ldots$ як послідовність незалежних однаково розподілених випадкових величин з певним розподілом G. Цей розподіл назвемо початковим. (Поки що ми вміємо генерувати лише послідовності з рівномірним розподілом, але далі нам інколи буде зручно використовувати як початковий який-небудь інший розподіл).

Наша мета — побудувати послідовність $\xi_1, \ldots, \xi_n, \ldots$ незалежних випадкових величин із заданим розподілом F. Цей розподіл називають цільовим. Методи генерації таких послідовностей розрізняються залежно від того, в якій формі заданий цільовий розподіл.

Квантильне перетворення. Нехай задана функція розподілу для цільового розподілу $F(x) = \mathsf{P}\{\xi_1 < x\}$, причому $F(x)$ є неперервною і строго зростаючою там, де вона не дорівнює 0 або 1. Розглянемо випадкову величину $\xi = F^{-1}(\eta)$, де випадкова величина η рівномірно розподілена на [0,1], F^{-1} — функція, обернена до F. (Нам досить, щоб рівність $F(F^{-1}(x)) = x$ виконувалась для всіх x між 0 та 1, тобто F^{-1} повинна бути оберненою до F на тому інтервалі, де F — неперервна і строго зростаюча).

Тоді для функції розподілу ξ виконуються рівності

$$F_\xi(x) = \mathsf{P}\{\xi < x\} = \mathsf{P}\{F^{-1}(\eta) < x\} = \mathsf{P}\{\eta < F(x)\} = F(x),$$

тобто саме ξ має цільовий розподіл.

Отже, отримати випадкову послідовність з функцією розподілу F можна, застосувавши перетворення $x \to F^{-1}(x)$ до кожного елемента рівномірної початкової послідовності η_j окремо:

$\xi_j = F^{-1}(\eta_j)$. Оскільки випадкові величини початкової послідовності були незалежними між собою, незалежними будуть і отримані ξ_j.

Це перетворення називають квантильним, тому що $F^{-1}(\alpha)$ — квантиль рівня α для розподілу F.

Приклад 6.5.1. Нехай треба згенерувати послідовність незалежних, однаково розподілених випадкових величин з експоненційним розподілом. Функція розподілу — $F_\lambda(x) = 1 - e^{-\lambda x}$ при $x \geq 0$. Функція $F^{-1}(y) = -\log(1-y)/\lambda$ є оберненою для $F(x)$ при $x \geq 0$. Якщо η рівномірно розподілена на $[0, 1]$, то і $(1 - \eta)$ також. Тому з рівномірної початкової послідовності $\eta_1, \ldots, \eta_n$ цільову експоненційну послідовність можна отримати перетворенням

$$\xi_j = -\frac{\log \eta_j}{\lambda}.$$

Згенеруємо у R вибірку з n експоненційних псевдовипадкових величин і відобразимо графіки її емпіричної та теоретичної функцій розподілу:

```
> n<-100    # кількість спостережень
> lambda=0.5 # інтенсивність exp розподілу
> # Використовуємо генератор Парка-Міллера:
> a<-7^5
> c0<-0
> m<-2^31-1
> y<-numeric(n)
> y[1]<-1000
> for(i in 2:n){
+ y[i]<-(a*y[i-1]+c0)%% m
+ }
> y<-y/m            # рівномірна послідовність
> x<--log(y)/lambda # квантильне перетворення
> #
> # рисуємо емпіричну функцію розподілу:
```

```
> sx<-sort(x)
> plot(sx,(1:n)/n,type="s")
> # графік теоретичної функції розподілу:
> lines(sx,pexp(sx,rate=lambda),col="red")
```

△

Тут ми скористались генератором Парка і Міллера для отримання рівномірної послідовності `y`, а потім застосували квантильне перетворення, щоб отримати цільову послідовність `x`. Графік її емпіричної функції розподілу порівняно з відповідною теоретичною функцією відображений на рис. 6.12.

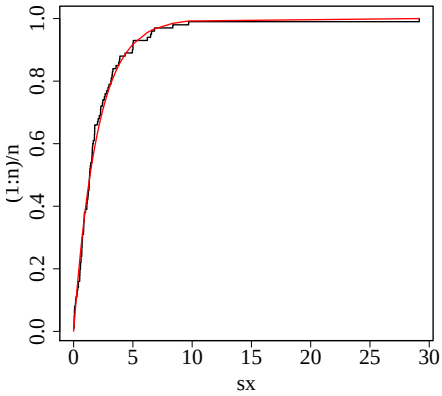

Рис. 6.12. Емпірична функція розподілу для експоненційного генератора випадкових чисел

◀

Метод проріджування. Квантильне перетворення дозволяє отримати незалежні випадкові величини з будь-яким розподілом. Але для цього потрібна функція, що знаходить квантилі цільового розподілу. Часто такі функції важко записати у явному вигляді, а чисельний підрахунок квантилів становить самостійну задачу. Метод проріджування дозволяє генерувати послідовності із заданим

розподілом, використовуючи для цього не квантилі, а щільності розподілу. Пояснимо ідею цього методу.

Нехай випадкова величина η має щільність розподілу g, а нам потрібна випадкова величина з щільністю f. Припустимо, що для всіх x $f(x) \le Cg(x)$ для деякого фіксованого числа $0 < C < \infty$. Введемо ще одну випадкову величину u, що має рівномірний розподіл на $[0, 1]$ і незалежна від η.

Знайдемо умовну ймовірність

$$\mathsf{P}\left\{\eta < x \mid u < \frac{f(\eta)}{Cg(\eta)}\right\} = \frac{\mathsf{P}\{\eta < x,\ u < \frac{f(\eta)}{Cg(\eta)}\}}{\mathsf{P}\{u < \frac{f(\eta)}{Cg(\eta)}\}}.$$

Для чисельника маємо

$$\mathsf{P}\{\eta < x,\ u < \frac{f(\eta)}{Cg(\eta)}\} = \int_{-\infty}^{x} \int_{0}^{f(y)/(Cg(y))} dtg(y)dy =$$

$$= \frac{1}{C} \int_{-\infty}^{x} f(y)dy.$$

Аналогічно для знаменника

$$\mathsf{P}\{u < \frac{f(\eta)}{Cg(\eta)}\} = \frac{1}{C}.$$

Отже, функція розподілу для розподілу η за умови $u < \frac{f(\eta)}{Cg(\eta)}$, дорівнює

$$\mathsf{P}\left\{\eta < x \mid u < \frac{f(\eta)}{Cg(\eta)}\right\} = \int_{-\infty}^{x} f(y)dy,$$

тобто це функція розподілу цільового розподілу зі щільністю f.

Ідея методу проріджування полягає в тому, щоб згенерувати послідовність пар (η_1, u_1), (η_2, u_2), ..., де η_j мають щільність g, u_j — рівномірні на [0,1] і всі випадкові величини незалежні в сукупності, а потім відібрати з елементів цієї послідовності ті, які

задовольняють умову $u_j < f(\eta_j)/(Cg(\eta_j))$. Послідовність, створена відібраними η_j, буде мати цільовий розподіл.

Приклад 6.5.2. Розглянемо задачу генерації послідовності з півнормальним розподілом з параметром $\sigma = 1$. Нагадаємо, що це розподіл випадкової величини $|\zeta|$, де ζ — стандартна гауссова випадкова величина. Його функція розподілу

$$F(x) = \mathsf{P}\{|\zeta| < x\} = \begin{cases} 2\Phi(x) - 1 & \text{при } x \geq 0, \\ 0 & \text{при } x < 0. \end{cases}$$

Щільність розподілу —

$$f(x) = \begin{cases} \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) & \text{при } x \geq 0, \\ 0 & \text{при } x < 0. \end{cases}$$

Щільність цільового розподілу записується у явному вигляді, а квантилі ні. Тому природно скористатись для генерації методом проріджування. Оскільки $f(x) > 0$ для всіх додатних x, то рівномірний розподіл не можна використати як початковий. Але можна взяти як початкові експоненційно розподілені випадкові величини з інтенсивністю $\lambda = 1$. Щільність цього розподілу на додатній півосі $g(x) = \exp(-x)$.

Бачимо, що $f(x) \leq Cg(x)$ для $C = \sqrt{2e/\pi}$ і

$$\frac{f(x)}{Cg(x)} = \exp\left(-\frac{(x-1)^2}{2}\right).$$

Для генерації експоненційно розподіленої послідовності використаємо квантильне перетворення, як у прикладі 1. Оформимо знаходження чергового елемента псевдовипадкової послідовності у вигляді окремої функції. У скрипті, що наведений нижче, `rand()` — функція, яка генерує одне чергове рівномірне [0,1] число. (При цьому відповідне значення цілочислової послідовності `I` записується у глобальну змінну за допомогою глобального привласнення

`I<<-` всередині тіла функції (див. п. 2.7.1). Функція, що генерує півнормальне число, називається `rhnorm`.

```
> n<-300    # кількість спостережень
> a<-7^5     #   параметри генератора
> m<-2^31-1  #   Парка і Мілера
> I<-500     # початкове значення для генератора
> #
> # генератор рівномірної послідовності:
> rand<-function(){I<<-(a*I)%%m; I/m}
> #
> # генератор півнормальної послідовності:
> rhnorm<-function()
+ {
+ repeat{
+ u<-rand()
+ x<--log(rand()) # x - експоненційне
+ if(u<exp(-0.5*(x-1)^2))return(x)
+ }
+ }
> # генеруємо півнормальну послідовність:
> x<-replicate(n,rhnorm())
> # рисуємо графік емпіричної функції розподілу
> sx<-sort(x)
> plot(sx,(1:n)/n,type="s",lwd=2,xlab="",ylab="")
> # графік теоретичної функції розподілу:
> lines(sx,2*pnorm(sx)-1,col="red",lwd=2,lty=2)
> # рисуємо гістограму відносних частот:
> hist(x, density=20, breaks=15, prob=TRUE,
+ xlab="",ylab="",main="")
> # рисуємо графік півнормальної щільності:
> curve(2*dnorm(x),
+ col="darkblue", lwd=2, add=TRUE, yaxt="n")
```

Результати графічної перевірки якості генерації зображені на рис. 6.13,. Тут на рисунку ліворуч емпірична функція розподілу порівнюється з теоретичною, а рисунку праворуч — гістограма відносних частот[14] згенерованої послідовності з щільністю півнормального розподілу (синя крива).

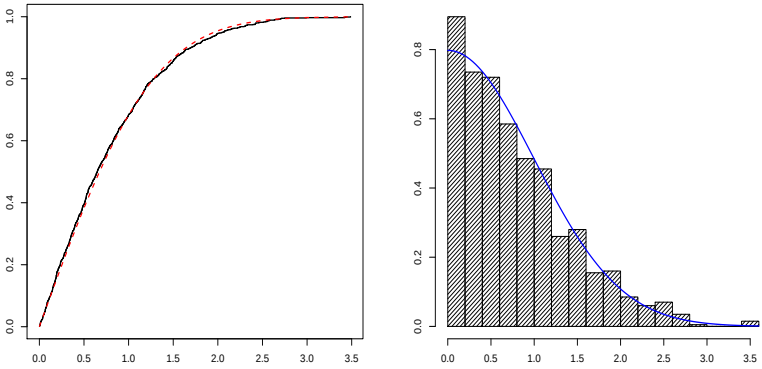

Рис. 6.13. Емпірична функція розподілу для півнормального генератора випадкових чисел

Як бачимо, щільність та функція розподілу півнормального розподілу добре відтворюються нашим генератором. ◀

6.5.3 Випадкові числа в R

У базовому R реалізовані генератори псевдовипадкових послідовностей з основними ймовірнісними розподілами, вказаними у табл. 6.1. Назви всіх цих функцій починаються з літери `r`, після чого йде назва розподілу: `rnorm()` генерує нормальні послідовності, `rexp()` — експоненційні і т.п.

[14] Про гістограму як оцінку щільності див. підрозд. 7.1.

Першим параметром всіх цих функцій є кількість елементів послідовності. Після цього параметра можна вказувати параметри розподілу. Наприклад:

`rnorm(10)` — генерує вектор з 10 псевдовипадкових стандартних нормальних чисел;

`rnorm(5, mean=1,sd=0.5)` — вектор з 5 нормальних чисел з математичним сподіванням 1 та дисперсією 0.25;

`rexp(1,rate=0.5)` одне число з експоненційним розподілом з інтенсивністю $\lambda = 0.5$.

Окрім функцій, що імітують послідовності незалежних, однаково розподілених випадкових величин, у R є функції, що дозволяють імітувати вибір навмання елементів з деякої скінченної сукупності. Найпростіша з цих функцій — `sample()`. Її параметри:

`x` — вектор, що містить сукупність, з якої будуть обиратись елементи;

`size` — кількість елементів, що буде відібрано;

`replace` — якщо задати цей параметр як `FALSE`, відібрані елементи не будуть повертатись до сукупності (тобто відібрати їх можна лише один раз — відбір без повернення). Якщо задати `replace=TRUE`, то всі `n` елементів будуть відібрані з усієї сукупності `x` (відбір з поверненням).

Генерація псевдовипадкових чисел у стандартних функціях базового R організована за схемою, подібною до прикладу 2 з п. 6.5.2. Використовується одна цілочислова послідовність, на основі якої будуються значення всіх псевдовипадкових чисел, які генеруються під час сеансу роботи з R. Чергове значення цілочислової послідовності зберігається у глобальній змінній і змінюється при виконанні кожної функції-генератора.

Початкове значення цілочислової послідовності називається зерниною — seed. Ця зернина стандартно вибирається на початку сеансу роботи з R за останніми цифрами часу, який на цей момент показує годинник комп'ютера. Таким чином, кожного разу, коли ви запускаєте R, генерується нова послідовність псевдовипадкових чисел.

Вибір зернини за годинником зручний, якщо ви перевіряєте статистичні особливості ваших алгоритмів: кожна нова перевірка відбувається на нових даних. Але на етапі відлагоджування програми, коли вам треба пересвідчитись, що її робота відповідає теоретичному алгоритму і виловити невідповідності, така генерація створює незручності. Помилка програми, яка була помітною на одній послідовності, може загубитись при повторному тестуванні. Щоб усунути цей ефект бажано при відладці кожного разу запускати програму на тій самій псевдовипадковій послідовності. Це можна зробити, зафіксувавши зернину.

Вибір зернини робить функція `set.seed()`. Як параметр цієї функції можна вказати будь-яке ціле додатне число. За цим числом буде обрана зернина. Далі у цій книзі при використанні генераторів псевдовипадкових чисел зернина фіксується. Це зроблено для того, щоб опис результатів у книзі відповідав тому, що видає скрипт. При самостійній роботі зі скриптами з цієї книги фіксувати зернину не треба, якщо ви хочете подивитись на випадковий розкид результатів.

Розділ 7

Методи графічного аналізу одновимірних даних

У розд. 6 ми познайомились з основними теоретичними моделями розподілів даних. Тепер розглянемо графічні засоби перевірки того, що розподіл спостережуваних даних відповідає теоретичній моделі. У підрозд. 7.4 також побачимо, як порівнювати розподіли двох різних наборів даних, не маючи попередніх теоретичних моделей для них.

Окрім графічних методів перевірки розподілу існують і інші, наприклад, за допомогою спеціальних статистичних тестів. Про один з таких тестів див. п. 9.6.3.

7.1 Гістограми

Гістограма є найбільш популярним способом графічного відображення розподілу числових даних. Розрізняють гістограми абсолютних та відносних частот.

Перш ніж описувати побудову гістограм, введемо важливі по-

значення. Нехай спостерігаються значення змінної X у n об'єктів. Зберемо їх в один набір $(X_1, \dots, X_n)$. Задамо деякий інтервал $[a, b]$, на якому розміщені всі спостережувані значення. Розіб'ємо цей інтервал на K підінтервалів $A_1, \dots A_K$ однакової ширини $h = (b - a)/K$. Інтервали A_i, $i = 2, \dots, K$ визначаються як $A_i = (t_{i-1}, t_i]$, де $t_i = a + ih$, $A_1 = [t_1, t_2]$.

Позначимо $n_i = \sum_{j=1}^{n} \mathbb{1}\{X_j \in A_i\}$ — кількість спостережуваних значень, що потрапили на інтервал A_i. Величину n_i називають *абсолютною частотою* (absolute frequency, count) інтервалу A_i у вибірці X. Величину $\nu_i = n_i/n$ називають *відносною частотою* (relative frequency). [1]

Гістограма абсолютних частот будується так. На горизонтальній осі відкладаються інтервали A_i і над кожним інтервалом будується стовпчик висоти n_i (див. ліву частину рис. 7.1).

У гістограмі відносних частот висота стовпчика визначається як $f_i = \nu_i/h = n_i/(nh)$. Таким чином, на рисунку гістограма відносних частот відрізняється від гістограми абсолютних лише масштабом по вертикалі (рис. 7.1). Нормуючий множник $1/(nh)$ для гістограми відносних частот обраний так, щоб її можна було використовувати як оцінку для щільності розподілу вибірки.

Дійсно, нехай $X = (X_1, \dots, X_n)$ — вибірка з незалежних однаково розподілених випадкових величин (кратна вибірка), що мають щільність розподілу f. За законом великих чисел, при великому обсязі вибірки n:

$$\nu_i = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{X_j \in [t_{i-1}, t_i)\} \approx \mathsf{P}\{X_1 \in [t_{i-1}, t_i)\} = \int_{t_{i-1}}^{t_i} f(t)dt.$$

[1] Зауважимо, що при нашому виборі відкритих зліва інтервалів A_i, спостереження, яке опинилось на межі двох інтервалів, потрапляє до інтервалу, що лежить ліворуч. (Так реалізовано обчислення частот для гістограм в `R`). Інколи навпаки, задають інтервали розбиття, відкриті справа. Іще один можливий варіант, коли спостереження, що лежить на межі двох інтервалів, враховується у частотах обох, але з вагою 1/2. При великій кількості спостережень без повторень ці відмінності не важливі, але у деяких випадках можуть бути важливими для розуміння поведінки гістограми.

Якщо $x \in [t_{i-1}, t_i)$, f — гладенька функція і h маленьке, то $\int_{t_{i-1}}^{t_i} f(t)dt \approx f(x)h$. Отже, $f_i \approx f(x)$, тобто функція

$$\hat{f}(x) = \begin{cases} f_1 & \text{якщо } x \in A_1 \\ f_2 & \text{якщо } x \in A_2 \\ \dots & \\ f_K & \text{якщо } x \in A_K \\ 0 & \text{якщо } x \notin [a, b] \end{cases}$$

є хорошим наближенням для $f(x)$. Гістограму відносних частот можна розглядати як графік цієї функції, а саму $\hat{f}(x)$ називають гістограмною оцінкою щільності розподілу.

Таким чином, якщо гістограму зображають, щоб побачити щільність розподілу даних, доцільно використовувати саме гістограму відносних частот. Однак певні переваги має і гістограма абсолютних частот: за висотою її стовпчиків одразу можна побачити, скільки спостережень потрапило в той чи інший інтервал розбиття.

У `R` для зображення гістограм використовується стандартна функція `hist(x, ...)`. Перелічимо деякі параметри/опції цієї функції:

`x` — набір даних (вибірка), за яким будується гістограма;

`breaks` — параметр, що контролює вибір точок розбиття. Якщо він не заданий, то стандартно кількість точок розбитття обирається за формулою Стургеса: $K = \lfloor \log_2 n + 1 \rfloor$, де n — кількість елементів `x`. Якщо `breaks` — це одне число, то функція бере його як кількість інтервалів розбиття. При цьому кінцеві точки всього інтервалу, на якому будується гістограма, визначаються як `min(x)`, `max(x)`. Якщо `breaks` — числовий вектор, його розглядають як набір точок розбиття $t_0 < t_1 < \dots < t_K$.

`probability` — логічна опція, стандартно — `FALSE`. Якщо вона дорівнює `TRUE`, будується гістограма відносних частот, інакше — абсолютних.

`right` — логічна опція, якщо вона TRUE, то інтервали розбиття вважаються замкненими справа, відкритими зліва.

`density, angle, col, border` — параметри, що контролюють штриховку та колір прямокутників гістограми так само, як у функції `rect()`.

`main, xlab, ylab` — параметри, що задають основну назву та назви осей гістограми.

`plot` — якщо цей параметр зробити `FALSE`, гістограма відображатись не буде. Але функція `hist` розрахує всі параметри, необхідні для зображення гістограми (інтервали розбиття та висоти стовпчиків) і видасть їх як результат своєї роботи. Тобто значенням функції є об'єкт, що містить ці параметри. Його можна зберегти для подальшого використання. (Скажімо, для відображення пізніше на іншому рисунку).

Приклад 7.1.1. У файлі `tips.csv` містяться дані про чайові, які отримував один офіціант ресторану у США протягом двох з половиною місяців роботи у 1990 р. Розмір чайових, отриманих за кожне обслуговування, записаний у змінній `tip`, змінна `sex` вказує стать особи, що оплачувала рахунок ("F" — жінка, "M" — чоловік). Щоб отримати гістограми розміру чайових, виконаємо такі команди:

```
> z<-read.csv("c:/rem/rstat/data/tips.csv")
> hist(z$tip,main="",xlab="",ylab="")
> hist(z$tip,probability=T,
+ main="",xlab="",ylab="")
```

Спочатку ми прочитали дані за допомогою функції `read.csv` (на моєму комп'ютері файл `tips.csv` розміщений у каталозі `c:/rem/rstat/data/`). Потім ми вивели гістограму абсолютних частот і гістограму відносних частот. Результат виконання зображений на рис. 7.1.

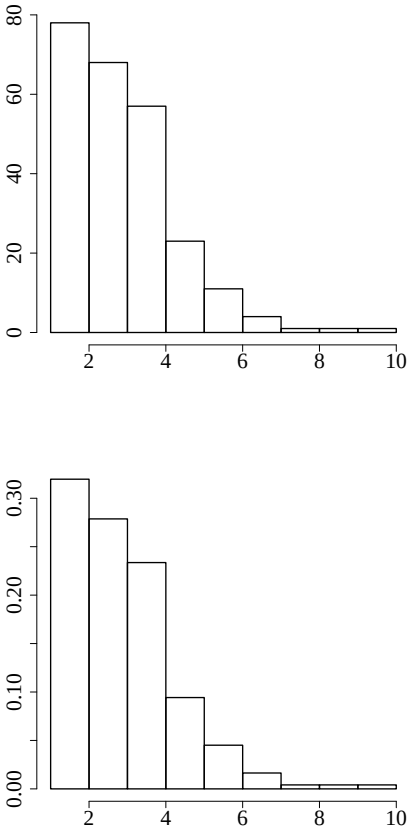

Рис. 7.1. Гістограми абсолютних (зверху) та відносних (знизу) частот

З рис. 7.1 можна зробити висновок, що щільність розподілу розміру чайових є монотонно спадною. Зсунемо початкову точку гістограми [2] на 1/2 (рис. 7.2, зверху). Тепер рисунок має такий вигляд, наче щільність спочатку зростає, а потім починає спадати.

Зменшимо ширину інтервалу розбиття — покладемо $h = 0.125$

[2]Origin, тобто лівий кінець інтервалу, на якому побудована гістограма.

— отримуємо картинку на рис. 7.2 знизу.

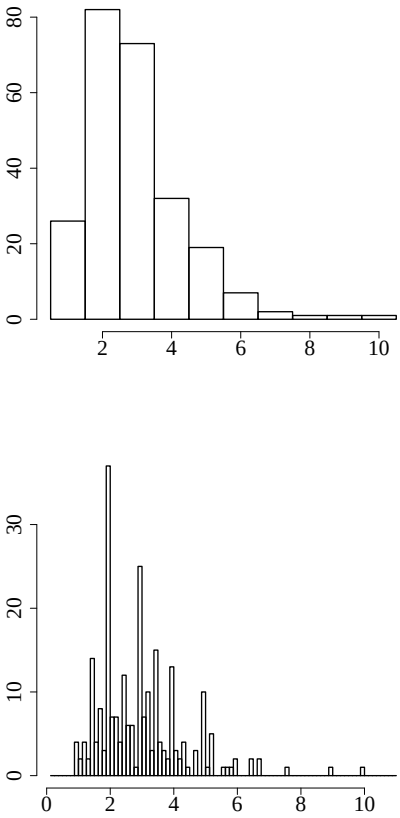

Рис. 7.2. Гістограми для tip: початкова точка та ширина інтервалу

Команди, якими відображені рисунки, мають такий вигляд:

```
> z<-read.csv("c:/rem/rstat/data/tips.csv")
> hist(z$tip,main="",xlab="",ylab="",
```

```
+ breaks=(1:11)-0.5)
> hist(z$tip,main="",ylab="",
+ xlab="",breaks=(1:88)*0.125)
```

△

Якщо уважно придивитись до останнього рисунка, то можна побачити, що піки на гістограмі відповідають цілим розмірам чайових (2, 3, 4, 5 доларів) а також цілим значенням плюс півдолара. Крім того, праворуч від основної маси спостережень розташовані окремі невисокі стовпчики, що відповідають аномально великим чайовим. Ці спостереження легко пояснити з соціально-психологічних міркувань: людина може залишити "на чай" дрібні монети здачі, або дати гроші з свого гаманця. У другому випадку, зазвичай, залишають круглу суму. Більшість людей дотримуються загальноприйнятого розміру чайових, але дехто часом виявляє аномальну щедрість. Таким чином, у даному випадку не можна казати про якусь спільну щільність розподілу даних, що описує всі спостереження. Радше дані треба описувати моделлю суміші двох компонент: абсолютно неперервної (здачу залиште собі) і дискретної (два-три-чотири долари на чай). Проте, гістограма абсолютних частот дає можливість візуально проаналізувати такі дані і зробити певні висновки про їхній розподіл. ◄

Істотні відмінності висот сусідніх стовпчиків гістограми не обов'язково свідчить про наявність дискретної компоненти. При зменшенні ширини інтервалів розбиття h розкид висот стовпчиків зростає і тоді, коли дані являють собою кратну вибірку з розподілу, що має гладеньку щільність. Це легко зрозуміти: відносна частота інтервалу у вибірці наближається до ймовірності потрапляння у цей інтервал лише при великій кількості спостережень. Але, якщо інтервал малий, то мала і ймовірність потрапити на нього, отже, на нього потрапить мало спостережень і його частота буде помітно коливатись навколо ймовірності. Це добре видно у наступному прикладі (рис. 7.3):

▽

```
> set.seed(3)
```

```
> z<-rnorm(200)
> hist(z,10,main="",xlab="",ylab="")
> hist(z,50,main="",xlab="",ylab="")
```

△

На рис. 7.3 зверху — гістограма, побудована з 10-ма широкими інтервалами, знизу — з 50-ма вузенькими.

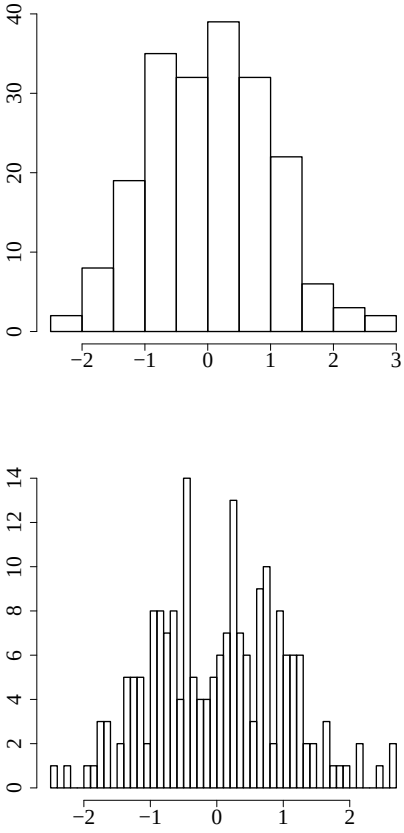

Рис. 7.3. Гістограми нормального розподілу

Як і у попередньому прикладі, звуження інтервалів привело до появи піків та стовпчиків, розміщених окремо від основної маси спостережень. Але у розміщенні піків не помітно якої-небудь закономірності, а стовпчики, що стоять окремо, розташовані досить близько від сусідніх. Висоти всіх стовпчиків невеликі, тобто спостережень недостатньо для надійної оцінки щільності на кожному інтервалі. Тому ці ефекти природно трактувати, як випадкові. У даному прикладі ми знаємо, що вони дійсно є випадковими, оскільки спостереження `z` були створені генератором псевдовипадкових чисел зі стандартним нормальним розподілом. Але у загальному випадку відрізнити випадкові ефекти від значущих особливостей на гістограмі може бути складно.

7.2 Графічна перевірка узгодженості розподілу. P-P- та Q-Q-діаграми

Одне з найбільш поширених застосувань гістограми — візуальне визначення типу розподілу та перевірка узгодженості даних з цим розподілом. Як ми з'ясували у попередньому підрозділі, гістограма відносних частот є оцінкою щільності розподілу за кратною вибіркою[3]. Зобразивши таку гістограму разом з теоретичною щільністю на одному рисунку, можна побачити, наскільки теоретична модель відповідає реальним даним[4].

Приклад 7.2.1. у наборі даних `airquality` містяться дані щоденних вимірювань метеорологічної станції у Нью-Йорку з травня по вересень 1973 р. Зокрема, у змінній `airquality$Wind` записана сила вітру у відповідний день. Ми хочемо перевірити, чи є розподіл

[3] У цьому підрозділі як теоретичну модель для опису даних ми розглядаємо лише кратні вибірки.

[4] Якщо теоретичний розподіл має невідомі параметри, їх треба оцінити, перш ніж відображати графік щільності. Як це робиться у простіших випадках, розказано у п. 8.6. Щоб правильно побудувати оцінку у складніших випадках, треба ознайомитись із загальною теорією оцінювання, якій присвячений весь розд. 8.

цієї характеристики нормальним. Наведемо два варіанти програми відображення відповідної гістограми та щільності розподілу:

```
> # 1. гістограма відносних частот.
> #
> g = airquality$Wind
> m<-mean(g)
> std<-sqrt(var(g))
> hist(g, density=20, breaks=10, prob=TRUE,
+ xlab="", ylim=c(0, 0.15),
+ main="",,ylab="")
> curve(dnorm(x, mean=m, sd=std),
+ col="darkblue", lwd=2, add=TRUE, yaxt="n")
> #
> # 2. гістограма абсолютних частот
> #
> hi<-hist(g, density=20, breaks=10,
+ xlab="", ylim=c(0, 45),
+ main="",,ylab="")
> curve(dnorm(x, mean=m, sd=std)
+ *length(g)*(hi$breaks[2]-hi$breaks[1]),
+ col="darkblue", lwd=2, add=TRUE, yaxt="n")
```

У першому варіанті будується гістограма відносних частот (параметр `prob=TRUE`) і нормальна щільність, параметри якої оцінюються відповідно середнім та коренем з вибіркової дисперсії змінної `g`. Результат зображено на рис. 7.4 зверху. Як бачимо, посередині гістограми є провал там, де мав бути пік щільності. Чи можна вважати його випадковим, чи це дійсно відхилення від нормальності розподілу даних сили вітру? За гістограмою відносних частот вирішити це неможливо. На гістограмі абсолютних частот можна побачити, скільки спостережень припало на цей провал, але масштаб цієї гістограми не відповідає масштабу графіку щільності.

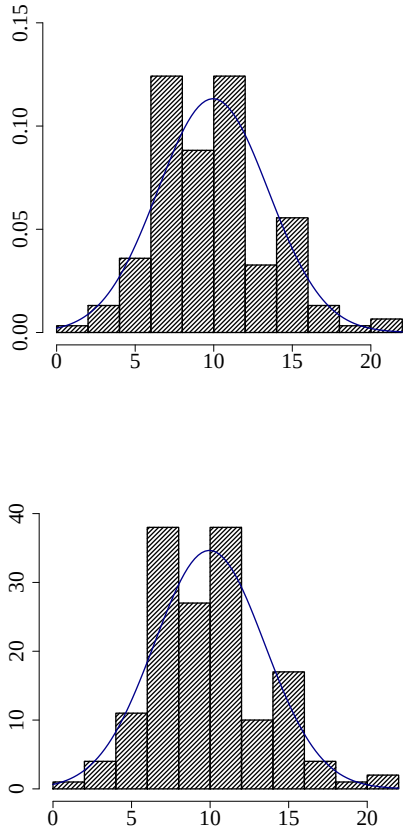

Рис. 7.4. Гістограми з графіком щільності

На рис. 7.4 знизу відображений графік щільності, помноженої на нормуючий множник nh, де n — обсяг вибірки, h — ширина підінтервалу розбиття. Це дозволяє порівнювати гістограму абсолютних частот з обраним (нормальним) розподілом. Щоб правильно визначити h, ми у скрипті зберегли значення результату функції `hist()` у змінній `hi`. Цей результат є об'єктом класу `histogram`

і має атрибут `hi$breaks`, у якому містяться значення точок розбиття для побудованої гістограми. Різниця між сусідніми точками дорівнює h.

З гістограми абсолютних частот на рис. 7.4) видно, що кількість спостережень, які припадають на інтервал між двома піками, становить близько 25, а кожному піку відповідає близько 40 спостережень. Це великі обсяги даних і помітна відмінність між піками та провалом. Навряд чи вона викликана випадковим відхиленням. Радше, така гістограма свідчить про те, що розподіл даних не є нормальним[5]. ◄

Перевірка розподілу даних на основі гістограм зручна тим, що за формою гістограми можна вгадати розподіл: гістограму, що відповідає нормальному розподілу не сплутаєш із гістограмою експоненційно розподілених даних. Але гістограми мають і недоліки: неправильно обравши ширину інтервалів розбиття або початок діапазону гістограми, можна отримати невдалий результат. Тому поруч з гістограмами використовуються інші техніки графічної перевірки того, наскільки розподіл даних узгоджується з певною теоретичною моделлю: P-P (ймовірність проти ймовірності) та Q-Q (квантиль проти квантиля) діаграми. Ці діаграми побудовані на порівнянні емпіричної функції розподілу або емпіричних квантилів з відповідними характеристиками теоретичної моделі. Вони не потребують задання додаткових параметрів налаштування, подібних до ширини інтервалу розбиття для гістограми. Але їхнім недоліком є те, що теоретичний розподіл треба визначити наперед: за формою діаграми його вгадувати не можна.

Почнемо з розгляду P-P-діаграм.

Нехай $X = (X_1, \ldots, X_n)$ — набір даних. Дослідник трактує X як кратну вибірку і хоче перевірити гіпотезу H_0 про те, що X_j мають функцію розподілу F. Якщо ця гіпотеза правильна, то для будь-якого $x \in \mathbb{R}$ емпірична функція розподілу вибірки $\hat{F}_n(x)$ є

[5]Крім такої перевірки "на око" доцільно також провести перевірку узгодженості нормального розподілу, використовуючи відповідні статистичні тести. У прикладі 9.6.4 показано, як це зробити за допомогою тесту χ^2.

близькою до F:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{X_j \le x\} \approx F(x)$$

при великих обсягах вибірки.

Підставимо у $\hat{F}_n(x)$ та $F(x)$ вибіркові значення $X_j, j = 1, \ldots, n$ і зобразимо нарисуємо точки з координатами $(F(X_j), \hat{F}_n(X_j))$. Це і є P-P-діаграма. Якщо гіпотеза H_0 правильна, то ордината та абсциса кожної точки повинні бути близькими одна до одної, отже, точки мають вишикуватись поблизу від бісектриси першого координатного кута, як це зображено на рис. 7.5 ліворуч. Якщо це не так, гіпотезу H_0 слід відхилити. Рис. 7.5 праворуч ілюструє випадок, коли для підгонки розподілу даних була обрана функція розподілу з не правильною (завищеною) дисперсією.

Припустимо, що всі значення X_j у вибірці є різними і впорядкуємо їх у порядку зростання, отримавши варіаційний ряд: $X_{[1]} < X_{[2]} < \cdots < X_{[n]}$. Тоді $\hat{F}_n(X_{[j]}) = j/n$, отже, P-P-діаграма складається з точок $(F(X_{[j]}), j/n)$, $j = 1, \ldots, n$.

У R P-P-діаграму, наприклад, для стандартного нормального розподілу, можна зобразити таким чином:

```
> # Генеруємо дані для прикладу
> set.seed(3)
> n<-100
> x<-rnorm(n)
> y<-rnorm(n,sd=3)
> # P-P для x з стандартним нормальним розподілом
> plot(pnorm(sort(x)),(1:length(x))/length(x),asp=1,
+ ylab="", xlab="")
> # Виводимо бісектрису координатного кута
> abline(0,1,col=2)
> # P-P для y з стандартним нормальним розподілом
> plot(pnorm(sort(y)),(1:length(y))/length(y),asp=1,
```

```
+ ylab="", xlab="")
> abline(0,1,col=2)
```

△

(Тут у `plot()` опції `xlab`, `ylab` задають текст написів при осях координат, опція `asp=1` забезпечує однаковий масштаб по вертикалі та горизонталі). Результат на рис. 7.5.

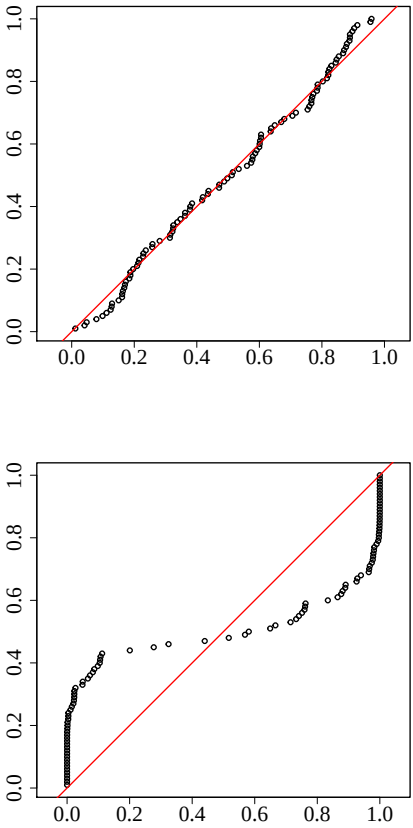

Рис. 7.5. P-P-діаграми

Побудова Q-Q-діаграми аналогічна, але по горизонталі та вертикалі відкладаються відповідно теоретичні та емпіричні квантилі. Точніше, роль емпіричних квантилів відіграють порядкові статистики $X_{[j]}$, яким відповідають теоретичні квантилі $Q^F(p_j)$, де $p_j = j/n - 1/(2n)$. (Значення p_j відповідає середині стрибка емпіричної функції розподілу $\hat{F}_n(x)$ у точці $x = X_{[j]}$). Таким чином, на Q-Q-діаграмі відображаються точки з координатами $(Q^F(p_j), X_{[j]})$, $j = 1, \ldots, n$. Якщо розподіл даних описується функцією розподілу F, ці точки мають бути поблизу від бісектриси першого координатного кута.

Q-Q-діаграма має важливу перевагу над P-P-діаграмою. Її зручно використовувати, коли теоретична функція розподілу відома з точністю до невідомих параметрів зсуву та масштабу. Тобто відомо, що $F(x) = F_0((x - a)/s)$, де a (зсув) і s (масштаб) — невідомі параметри. (Напр., для нормального розподілу F_0 може бути функцією розподілу стандартного нормального розподілу, a — математичним сподіванням, s — середньоквадратичним відхиленням). У цьому випадку $Q^F(\alpha) = sQ^{F_0}(\alpha) + a$, отже, якщо на Q-Q-діаграмі відобразити точки з координатами $(Q^{F_0}(p_j), X_{[j]})$, вони розташуються поблизу від прямої з рівнянням $y = sx + a$. Це дозволяє перевіряти гіпотезу про розподіл даних, не оцінюючи параметри зсуву та масщтабу. Більше того, ці параметри можна оцінити візуально за Q-Q-діаграмою.

Для нормального розподілу Q-Q-діаграму у `R` можна побудувати, використовуючи функції `qqnorm()` та `qqline()`:

```
> x<-rnorm(200,mean=1,sd=0.5)
> qqnorm(x,ylab="", xlab="",main="")
> qqline(x)
```

(У `x` створена вибірка з нормального розподілу з середнім 1 та дисперсією 0.25, потім функція `qqnorm()` будує Q-Q-діаграму, у якій по осі абсцис відкладені квантилі стандартного нормального розподілу, функція `qqline()` оцінює математичне сподівання a та

стандартне відхилення s за даними і проводить на діаграмі пряму $y = sx + a$. Результат виконання цих команд зображено на рис. 7.6. Зверніть увагу, що побудована пряма не є бісектрисою першого координатного кута, але точки розташовані біля неї. Так і має бути, оскільки розподіл даних є нормальним, але не стандартним нормальним.

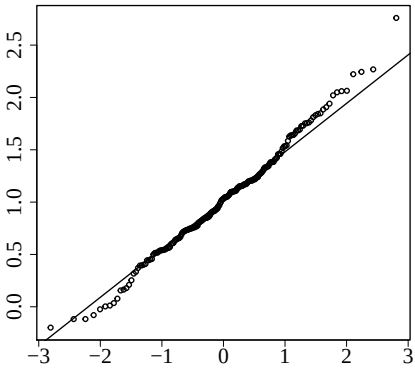

Рис. 7.6. Q-Q-діаграма

Якщо теоретичний розподіл не є нормальним, значення квантилів слід обчислювати, використовуючи відповідну функцію для даного розподілу. Наприклад, перевірку того, що розподіл даних є логістичним можна виконати так:

```
> set.seed(3)
> x<-rnorm(200,mean=1,sd=0.5)
> plot(qlogis(ppoints(x)),sort(x),
+ ylab="", xlab="",main="")
> abline(lm(sort(x)~qlogis(ppoints(x)))$coefficients)
```

Результат див. рис. 7.7.

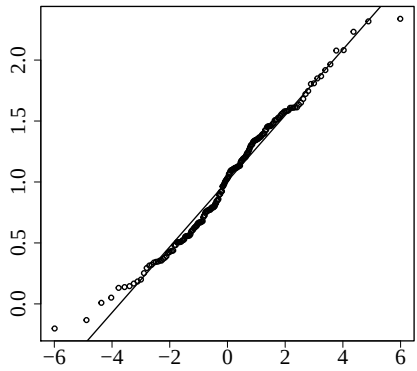

Рис. 7.7. Q-Q-діаграма

У цьому прикладі дані генеруються з нормальним розподілом, а перевірка проводится для теоретичного логістичного розподілу. Функція `ppoints(x)` обчислює значення рівнів квантилів p_j, отже, значенням виразу `qlogis(ppoints(x))` є вектор теоретичних квантилів, що відкладаються по горизонталі. Функція `abline()` рисує пряму лінію, коефіцієнти якої отримуються підгонкою за методом найменших квадратів (функція `lm`).

Зазначимо, що за цією Q-Q-діаграмою помітити відмінність розподілу даних (нормального) від логістичного майже неможливо.

7.3 Q-Q-діаграма з прогнозними інтервалами

Розглядаючи Q-Q-діаграми, можна побачити, що навіть коли розподіл даних відповідає теоретичному, точки на діаграмі відхиляються від бісектриси першого координатного кута, хоча і не дуже сильно. Причому у різних частинах діаграми такі випадкові відхилення можуть бути різними. Зазвичай, відхилення крайніх точок

помітніші, ніж відхилення точок всередині діаграми. Тому бажано крім бісектриси зобразити також інтервали, у які з великою ймовірністю можуть потрапляти точки на діаграмі, якщо теоретичний розподіл правильно описує дані.

Стандартні функції R не надають такої можливості. Розглянемо спосіб побудови таких прогнозних інтервалів за допомогою імітаційного моделювання.

Нехай треба побудувати інтервал, у який потраплятиме точка, що відповідає j-й порядковій статистиці із заданою ймовірністю $1-\alpha$. Ідея полягає в тому, щоб згенерувати багато (K) вибірок з розподілом, який відповідає теоретичному. Всі згенеровані вибірки повинні мати однаковий обсяг n, який дорівнює обсягу тієї реальної вибірки, що досліджується. По кожній такій вибірці візьмемо j-ту порядкову статистику. Отримаємо K значень $X^{(k)} = (X^k_{[j]}, k = 1, \ldots, K)$, де $X^k_{[j]}$ — j-та статистика для k-ї вибірки. За цими значеннями знайдемо емпіричні квантилі $X^-_j = Q^{X^{(k)}}(\alpha/2)$, $X^+_j = Q^{X^{(k)}}(1-\alpha/2)$. В інтервалі (X^-_j, X^+_j) буде приблизно $(1-\alpha)K$ елементів $X^{(k)}$. За законом великих чисел, при великих K, ймовірність для j-ї порядкової статистики потрапити у цей інтервал приблизно дорівнює $1-\alpha$.

Зрозуміло, що для побудови діаграми разом з інтервалами такі розрахунки треба повторити для всіх $j = 1, \ldots, n$. Модельовані вибірки можуть бути ті ж самі для різних j.

Реалізуємо цю ідею у вигляді функції `QQplot`, яка перевіряє узгодженість розподілу даних зі стандартним нормальним розподілом:

```
> QQplot<-function(x,K=1000,alpha=0.05){
+ n<-length(x)
+ normQ<-qnorm((1:n-0.5)/n)
+ sx<-sort(x)
+ W<-matrix(rnorm(K*n),nrow=n,ncol=K)
+ W<-apply(W,2,sort)
+ tops<-apply(W,1,quantile,probs=1-alpha/2)
```

```
+ bots<-apply(W,1,quantile,probs=alpha/2)
+ plot(c(normQ,normQ,normQ),c(tops,bots,sx),type="n",
+      xlab="theoretical quantiles",
+      ylab="empirical quantiles")
+ points(normQ,sx,col=2)
+ segments(normQ,bots,normQ,tops,col=4)
+ abline(0,1,col=1)
+ }
> x<-rnorm(100)
> QQplot(x)
```

△

Результат роботи програми див. на рис. 7.8.

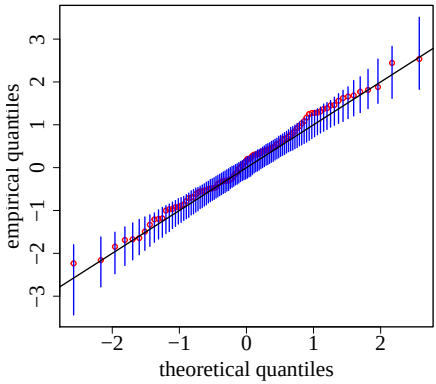

Рис. 7.8. Q-Q-діаграма з прогнозними інтервалами

Розберемо роботу функції. Її параметри:

`x` — вибірка, для якої будується Q-Q діаграма;

`K` — кількість псевдовипадкових вибірок, що будуть згенеровані для отримання прогнозних інтервалів (стандартне значення `K=1000`);

`alpha` — ймовірність, з якою точка на діаграмі може вийти з прогнозного інтервалу (стандартно `alpha=0.05`).

У тілі функції спочатку обчислюються абсциси точок на діаграмі — у векторі `normQ`. Створюється варіаційний ряд даних — `sx`. Потім генерується матриця `W`, стовпчиками якої є K псевдовипадкових вибірок зі стандартного нормального розподілу. Команда `W<-apply(W,2,sort)` впорядковує стовпчики `W` у порядку зростання. Тепер вони містять варіаційні ряди модельованих вибірок. Кожен (j-й) рядок матриці `W` складається тепер з порядкових статистик модельованих вибірок з індексом j. Ми шукаємо X_j^- і X_j^+ як відповідні квантилі для j-го рядка та вміщуємо їх у вектори `bots` і `tops` для всіх $j = 1, \ldots, n$. Далі йде виведення рисунка. Спочатку виводиться тільки рамка з підписами, підігнана так, щоб у ній розмістились всі елементи рисунка. Після цього `points()` виводить точки діаграми, `segments()` — інтервали, `abline` — бісектрису координатного кута.

7.4 Порівняння розподілів кількох наборів даних

У статистиці часто виникає задача порівняння розподілів різних наборів однотипних даних. Скажімо, за даними податкової інспекції можна поставити питання: чи відрізняється розподіл доходів населення у минулому та позаминулому роках? Для порівняння розподілів двох наборів даних можна використовувати рисунки, на яких зображено дві гістограми одразу, або Q-Q-діаграми, де по горизонталі відкладено квантилі одного набору, а по вертикалі іншого.

Наприклад, розглянемо дані про чайові з набору `tips.csv`, який ми вже використовували у підрозд. 7.1. Ми хочемо перевірити, чи відрізняються розподіли чайових залежно від того, хто їх сплачує — чоловік чи жінка? Гістограми та Q-Q-діаграми для такої перевірки можна вивести таким чином:

```
> # читаємо дані з файла:
> z<-read.csv("c:/rem/rstat/data/tips.csv")
> #
> # Будуємо дві гістограми на одному рисунку
> #
> hist(z$tip[z$sex=="M"],breaks=10,probability=T,
+ angle=0,density=12,xlim=c(0,10),ylim=c(0,0.45),
+ ylab="", xlab="",main="")
> hist(z$tip[z$sex=="F"],probability=T,
+ breaks=10,angle=90,density=12, xlim=c(0,10),add=T)
> #
> # Q-Q діаграма
> #
> qqplot(z$tip[z$sex=="F"],z$tip[z$sex=="M"],
+        xlab="females",ylab="males")
>  abline(0,1)
```

У цьому скрипті першою виводиться гістограма розподілу чайових для клієнтів-чоловіків (`z$sex=="M"`). Її стовпчики заштриховані вертикально. Потім на тому самому рисунку виводиться гістограма для жінок з горизонтальою штриховкою. Ми обрали для порівняння гістограми відносних частот, тому що вибірки мають помітно різний обсгяг (чоловіки розплачувались частіше, ніж жінки). Якби порівнювались абсолютні частоти, "жіноча" гістограма була б майже непомітна на фоні "чоловічої" (перевірте).

І гістограма, і Q-Q-діаграма свідчать, що принципової різниці у розподілі чайових не помітно для основної маси спостережень. Але для чоловіків помітно кілька випадків з аномально великими чайовими, для жінок таких випадків немає.

Коли стовпчики кількох гістограм перекриваються, це створює незручність для їхнього візуального аналізу. Більш зручним може бути застосування діаграм, на яких стовпчики розташовані поруч (рис. 7.10).

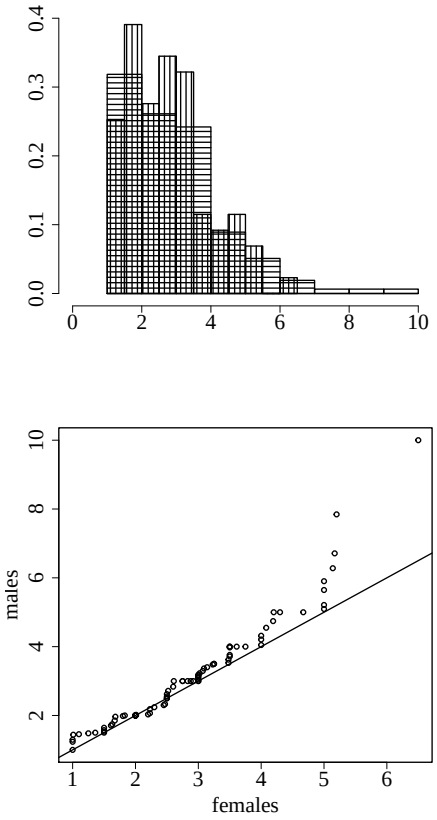

Рис. 7.9. Порівняння двох розподілів

Як ми бачили у підрозд. 3.1, такі діаграми можна зображати, використовуючи функцію `barplot`:

```
> z<-read.csv("c:/rem/rstat/data/tips.csv")
> ctip<-cut(z$tip,breaks=1:10,labels=(1:9)+0.5)
> counts<-table(z$sex,ctip)
```

```
> counts["F",]=counts["F",]/sum(counts["F",])
> counts["M",]=counts["M",]/sum(counts["M",])
> barplot(counts,beside=T,col=c("darkblue","yellow"))
> legend(x=16,y=0.31,c("Female","Male"),
+        fill=c("darkblue","yellow"))
>   #
```

Результат — на рис. 7.10).

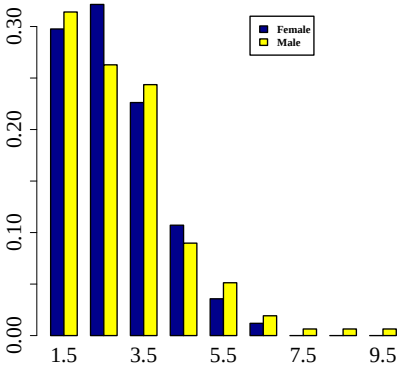

Рис. 7.10. Гістограма через barplot

Функція `cut()` у цьому скрипті використана для групування даних: отримуючи на вході числовий вектор `z$tip`, вона видає вектор, елементами якого є фактори, що показують, в який інтервал розбиття потрапило відповідне значення `z$tip`. Функція `table(z$sex,ctip)` складає таблицю (матрицю) частот появ пар значень факторів `(z$sex,ctip)`:

```
> table(z$sex,ctip)

   ctip
    1.5 2.5 3.5 4.5 5.5 6.5 7.5 8.5 9.5
```

```
F  25  27  19   9   3   1   0   0   0
M  49  41  38  14   8   3   1   1   1
```

△

— жінок (F), що дали чайові в інтервалі від 2 до 3 (позначений 2.5) було 25 і т.д.

Далі функція `barplot()` рисує стовпчикову діаграму, як описано у підрозд. 3.1, а функція `legend()` виводить пояснення-легенду.

7.5 Скриньки з вусами

Гістограми дають найбільш повне уявлення про розподіл одновимірних даних. Однак, коли треба порівняти розподіли багатьох (більше трьох) наборів даних, зображення гістограм усіх цих наборів на одному рисунку стає занадто складним для візуального сприйняття. Тому для забезпечення можливості графічного аналізу даних статистик змушений пожертвувати частиною інформації, відображаючи для кожного набору не гістограму, а лише найбільш характерні властивості розподілу.

Діаграми, які відтворюють мінімальний набір характеристик розподілу англійською мовою називається box-whisker plot, або просто boxplot. Українською це можна перекласти як *“скринька з вусами”*.

Для набору одновимірних даних скринька з вусами будуються за схемою, зображеною на рис. 7.11. На цьому рисунку значення даних відображаються по вертикальній осі. Можливе також горизонтальне розташування скриньки з вусами.

Прямокутник (скриньку) рисують від нижнього квартиля Q_1 (тобто квантиля рівня $1/4$) до верхнього квартиля Q_3 (квантиля рівня $3/4$), обчислених за даними. Лінія, що розрізає прямокутник, відповідає медіані med. Вусики, що стирчать зі скриньки, відмічають діапазон розташування даних, які не є викидами. Тобто верхній вусик відповідає найбільшому не викиду, нижній — найменшому. (На рис. 7.11 нижній вусик позначено min, а верх-

ній — max). Кожен кружечок поза діапазоном відповідає одному індивідуальному значенню-викиду.

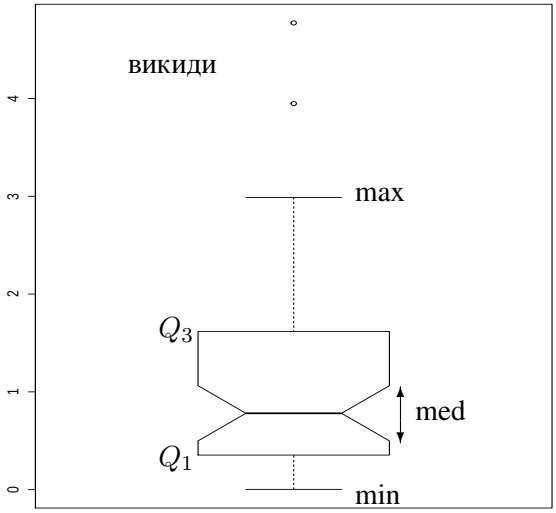

Рис. 7.11. Скринька з вусами

Для визначення того, які спостереження слід віднести до викидів, є різні підходи, що мають евристичний характер. При найбільш поширеному, викидами вважають ті значення, що перевищують $Q_3 + 1.5\text{IQ}$ або є меншими ніж $Q_1 - 1.5\text{IQ}$, де IQ=$Q_3 - Q_1$ — інтерквартильний розмах. Іноді окремо виділяють "далекі" викиди, або екстремальні значення, тобто ті значення даних, які виходять за межі інтервалу $[Q_1 - 3\text{IQ}, Q_3 + 3\text{IQ}]$. Якщо цей підхід використовується, то екстремальні значення позначають на діаграмі хрестиками, а помірні викиди (тобто такі, які не є екстремальними) — кружечками. Множники 1.5 та 3 у цих формулах не мають якогось науково-математичного або потаємно-містичного змісту,

а використовуються лише за домовленістю.

Інколи у стінках скриньки роблять трикутні зарубки (англ. notches), зовнішні краї яких відповідають довірчому інтервалу для медіани розподілу даних з рівнем значущості 0.95[6]. (На рис. 7.11 такий довірчий інтервал позначений стрілками).

Зазвичай ширина прямокутника-скриньки та вусиків обирається так, щоб рисунок було зручно сприймати на око, інформації про дані вона не містить. Але інколи ширину скриньки вибирають пропорційно кореню квадратному з кількості елементів у наборі даних, за яким вона побудована — чим ширша скринька, тим більше у ній даних.

Рисунок з однією скринькою для єдиного набору даних несе небагато інформації. Але розмістивши декілька таких скриньок паралельно для різних наборів, можна одразу помітити характерні відмінності розподілів даних у різних наборах.

Для рисування кількох скриньок з вусами у R можна використовувати функцію `boxplot()`. Першим (основним) параметром цієї функції є список наборів (векторів) даних, для яких будуються скриньки з вусами. Наприклад:

```
> set.seed(20)
> a<-rexp(200)
> b<-rnorm(100,2,1)
> c<-rchisq(40,5)
> x<-list(a,b,c)
> boxplot(x,notch=T,varwidth=T,
+         names=c("exp","norm","chisq"))
```

Тут ми згенерувалитри вибірки з різними розподілами: експоненційним, нормальним та хі-квадрат, склали їх в один список і відобразили за допомогою `boxplot` (рис. 7.12).

[6]Про довірчі інтервали див. підрозд. 8.5.

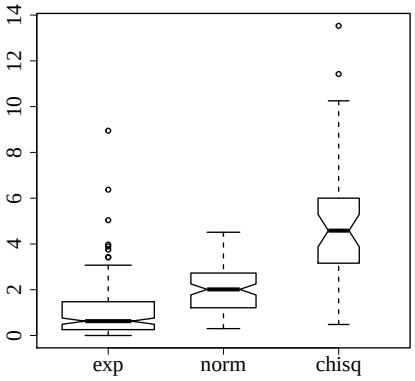

Рис. 7.12. Порівняння трьох розподілів

На рис. 7.12 можна помітити симетрію нормальної вибірки, асиметрію експоненційної. Хі-квадрат розподіл є асиметричним, але на рисунку ця асиметрія виражена не сильно. Викиди не відмічені у нормальній вибірці, два викиди — у хі-квадрат. Сім "викидів" зафіксовано у експоненційній вибірці, але за їхнім розташуванням можна твердити, що більшість з них не далеко відійшли від основної маси спостережень, тобто трактування їх як викидів є питанням смаку.

Ми скористались опцією `notch=T` для того, щоб відобразити довірчі інтервали для медіан у вигляді зарубок на скриньках. За цими інтервалами можна зробити попередній висновок, що медіани теоретичних розподілів вибірок є різними[7].

Опція `varwidth=T` вказує, що ширину скриньок слід обирати пропорційно до кореня з обсягу вибірки — тому скринька для `exp` вийшла помітно ширшою, ніж інші.

Опція `names` задає імена, що будуть підписані під скринька-

[7] Довірчі інтервали для них не мають спільних точок, докладніше див. підрозд. 9.4.

ми. Аналогічно можна використовувати опцію `col`, щоб задавати кольори скриньок.

У комп'ютерній статистиці часто виникають задачі аналізу даних, що записані у єдиному фреймі, причому одна змінна містить певну числову характеристику (відгук) об'єктів, що досліджуються, а інша (фактор) — клас, до якого належить даний об'єкт. При цьому питання полягає в тому, щоб проаналізувати залежність розподілу відгука від фактора. У таких випадках для опису задачі у `boxplot()` перший параметр можна задати формулою вигляду *відгук* $\sim$ *фактор*.

При такому запиті функція `boxplot()` розділить весь набір даних на окремі підвибірки. Кожна підвибірка складатиметься з об'єктів, що відповідають певному фіксованому значенню фактора. Скриньки з вусами будуються окремо для кожної підвибірки і відображаються на одному рисунку[8].

Приклад 7.5.1. у фреймі даних `InsectSprays` містяться дані про випробування якості різних видів інсектицидів. Один рядок даних відповідає одному випробуванню. У кожному випробуванні обчислювалась кількість комах, що загинули під дією інсектициду — змінна `count`. У змінній `spray` вказується тип інсектициду (літера A-F). Нас цікавить, як розподіл `count` пов'язаний зі `spray`. Відповідні скриньки задає програма

```
> boxplot(count ~ spray, data = InsectSprays,
+          col = "lightgray",horizontal=T)
```

Тут `data` задає фрейм даних, з якого вибирають змінні, а значення опції `horizontal=T` вказує, що скриньки повинні розміщуватись горизонтально.

На рис. 7.13 бачимо, що інсектициди C, D, E виявились значно менш ефективними ніж інші, інскетицид F у деяких експериментах

[8]Можна вказати декілька факторів, напр.: *відгук* $\sim$ *фактор1+фактор2*. Тоді набір даних буде розбитий на підвибірки, що відповідають різним комбінаціям можливих значень *фактор1* і *фактор2*.

виявив себе найкращим, але найкраща медіана — у B і т.д.

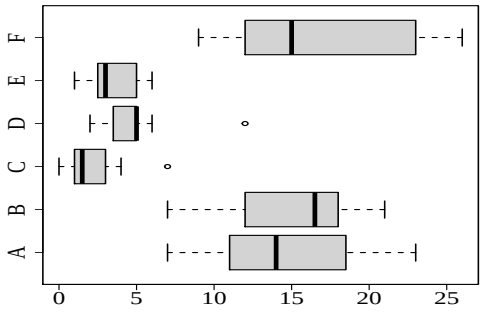

Рис. 7.13. Порівняння ефективності інсектицидів

Наскільки статистично обґрунтованим є це відчуття переваги одних інсектицидів над іншими, що виникає при розгляді такого рисунка? Це питання обговорюється у прикладі 9.7.1.

Зауважимо, що аналогічну діаграму можна отримати, якщо записати `plot(count~spray,data=InsectSprays)`. ◄

Розділ 8

Оцінювання параметрів

У цьому розділі описані основні підходи до задачі підгонки розподілу спостережуваних даних певною теоретичною моделлю. Вважається, що у цій теоретичній моделі є декілька невідомих параметрів, які треба оцінити за даними. У перших трьох підрозділах описано три різних підходи до такого оцінювання: метод моментів, оцінювання на основі порівняння емпіричних та теоретичних квантилів і метод найбільшої вірогідності. Далі у підрозд. 8.4 ми розглянемо питання про те, як охарактеризувати та порівняти точність різних оцінок. У підрозд. 8.5 показано, як будувати довірчі інтервали та довірчі еліпсоїди для невідомих параметрів.

Приклади, що розглядаються у підрозд. 8.1 — 8.5, використовують порівняно складні моделі розподілів (вимірювання з похибками, зрізані розподіли, суміші). Тому реалізація відповідних оцінок вимагає написання власних програм на R.

У більшості випадків статистичної обробки даних теоретичну модель обирають з невеликого набору стандартних розподілів (нормальний, експоненційний, пуассонів і т.д.). Для підгонки таких моделей у R є готові засоби. Читачі, яких цікавить саме така підгонка, якщо теорія статистичного оцінювання їм у загальних рисах відома (або не цікава), можуть звернутись одразу до підрозд. 8.6, де описано, як оцінювати параметри простих стандартних розподілів

та будувати довірчі інтервали за допомогою функції `fitdistr()`.

Для розуміння теоретичних міркувань, що наведені у цьому та наступних розділах книги, вимагається значно глибше знайомство з теорією ймовірностей та векторною алгеброю, ніж це було досі. Мінімальну необхідну інформацію про вектори і матриці вміщено у Дод. А. Відомостям з теорії ймовірностей присвячений Дод. Б. Зокрема, у підрозд. Б.3 дано означення багатовимірного нормального (гауссового) розподілу, а у підрозд. Б.4 описано основні види ймовірнісної збіжності (за ймовірністю, слабко, майже напевно) та наведені приклади граничних теорем.

Втім, я намагався організувати виклад так, аби книгу можна було читати і не відчуваючи відмінності між, скажімо, збіжностями за ймовірністю і майже напевно. Щоб отримати загальне уявлення, про що йдеться, читачеві досить мати загальне уявлення про збіжність як про наближення чогось до чогось "коли n стає все більше і більше". Зрозуміло, що таким загальним поняттям не варто обмежуватись. Бажаючим зануритись у теорію глибше рекомендую книги з теорії ймовірностей і математичної статистики [3, 2, 9, 18].

8.1 Оцінки методу моментів

Нехай спостережувані дані являють собою **кратну вибірку**, тобто набір $\mathbf{X} = (\xi_i, \ldots, \xi_n)$, де $\xi_j \in \mathbb{R}^p$ — незалежні випадкові вектори з розподілом

$$\mathbf{P}_\vartheta(A) = \mathbf{P}^\xi_\vartheta(A) = \mathsf{P}\{\xi_j \in A\},$$

де $\vartheta \in \Theta \in \mathbb{R}^d$ — d-вимірний невідомий параметр, Θ — множина можливих значень невідомого параметра. (Вектор ϑ можна трактувати, як набір d числових невідомих параметрів).

Для того, щоб оцінити ϑ, задамо деяку вимірну функцію $\mathbf{h} : \mathbb{R}^p \to \mathbb{R}^d$ так, щоб для всіх $\mathbf{t} \in \Theta$ було скінченним математичне сподівання

$$\mathbf{H}(\mathbf{t}) = \mathsf{E}_\mathbf{t}\, \mathbf{h}(\xi_1) = \int_{\mathbb{R}^p} \mathbf{h}(\mathbf{x}) \mathbf{P}_\mathbf{t}(d\mathbf{x}).$$

Внаслідок закону великих чисел, за великих обсягів вибірки n:

$$\hat{\mathbf{h}}_n = \frac{1}{n}\sum_{j=1}^{n} \mathbf{h}(\xi_j) \approx \mathbf{H}(\vartheta).$$

Прирівняємо

$$\hat{\mathbf{h}}_n = \mathbf{H}(\mathbf{t}) \tag{8.1}$$

і виберемо як оцінку ϑ таку статистику[1] $\hat{\vartheta} = \hat{\vartheta}(\mathbf{X})$, щоб при підстановці її у (8.1) замість $\mathbf{t}$ це рівняння перетворювалось на рівність майже напевно. Таку оцінку $\hat{\vartheta}_n$ називають оцінкою методу моментів (моментною оцінкою) для ϑ з моментною функцією $\mathbf{h}$. Функцію $\mathbf{H}(\vartheta)$ називають (узагальненим) теоретичним моментом (або вектором моментів) розподілу $\mathbf{P}_\vartheta$, а $\hat{\mathbf{h}}_n$ — емпіричним моментом вибірки $\mathbf{X}$. У випадку одновимірних спостережень ($p = 1$) при $h(x) = x^k$, $H(\vartheta) = \mathsf{E}_\vartheta\, \xi^k$ називають k-м теоретичним моментом, а $\hat{h}_n = \frac{1}{n}\sum_{j=1}^{n} \xi_j^k$ — k-м емпіричним моментом.

Якщо рівняння (відносно $\mathbf{t}$):

$$\mathbf{H}(\mathbf{t}) = \mathbf{x} \tag{8.2}$$

має єдиний корінь для всіх x, що належать множині можливих значень функції $\mathbf{h}$, то $\hat{\vartheta} = \mathbf{H}^{-1}(\hat{\mathbf{h}})$, де $\mathbf{H}^{-1}$ — функція, обернена до функції $\mathbf{H}$. (При цьому необхідно, щоб $\mathbf{H}^{-1}$ була вимірною функцією).

Якщо рівняння (8.1) має декілька коренів, то оцінка методу моментів визначена неоднозначно: будь-який з коренів можна використовувати як оцінку.

Приклад 8.1.1. Нехай $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ кратна вибірка з експоненційного розподілу з невідомою інтенсивністю λ, тобто щільність розподілу ξ

$$f_\lambda(x) = \lambda e^{-\lambda x} \mathbb{1}\{x > 0\}.$$

[1] Тобто вимірну функцію від даних $\mathbf{X}$.

Задача полягає в оцінці $\lambda \in (0, \infty)$. Виберемо за моментну функцію $h^{(1)}(x) = x$. Тоді

$$H(\lambda) = \mathsf{E}_\lambda\, h^{(1)}(\xi_1) = \int_0^\infty x f_\lambda(x) dx = \frac{1}{\lambda}.$$

Отже, оцінка методу моментів з цією моментною функцією має вигляд

$$\hat{\lambda}_n^{(1)} = \frac{1}{\hat{h}_n^{(1)}} = \frac{1}{\bar{\xi}} = \frac{n}{\sum_{j=1}^n \xi_j}.$$

Якщо обрати моментну функцію $h^{(2)}(x) = x^2$, отримуємо іншу оцінку:

$$\mathsf{E}_\lambda(\xi_1)^2 = \frac{2}{\lambda^2},$$

тому оцінка методу моментів, що відповідає $h^{(2)}$, має вигляд

$$\hat{\lambda}_n^{(2)} = \sqrt{\frac{2}{\hat{h}_n^{(2)}}}.$$

◄

Приклад 8.1.2. Нехай $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ — кратна вибірка з нормального розподілу з невідомим математичним сподіванням μ та невідомою дисперсією σ^2. Позначимо невідомий векторний параметр $\vartheta = (\mu, \sigma^2)^T \in \Theta = \mathbb{R} \times (0, +\infty)$. Позначимо також $\mathbf{t} = (m, s^2)^T$. Виберемо за моментну функцію $\mathbf{h}(x) = (x, x^2)^T$. Тоді $\mathbf{H}(\mathbf{t}) = (m, s^2 + m^2)^T$. Отже, оцінка методу моментів визначається як розв'язок системи рівнянь

$$\begin{cases} \bar{\xi} = & m, \\ \overline{\xi^2} = & s^2 + m^2, \end{cases}$$

де $\bar{\xi} = \frac{1}{n}\sum_{j=1}^n \xi_j$, $\overline{\xi^2} = \frac{1}{n}\sum_{j=1}^n (\xi_j)^2$ — перший і другий вибіркові моменти.

Таким чином, $\hat{\vartheta}_n = (\bar{\xi}, \overline{\xi^2} - (\bar{\xi})^2)^T$, тобто оцінками для μ та σ^2 є вибіркове середнє та (не виправлена) вибіркова дисперсія.

Нагадаємо, що виправлена вибіркова дисперсія дорівнює

$$S_0^2(\mathbf{X}) = \frac{n}{n-1}(\overline{\xi^2} - (\bar{\xi})^2).$$

Як оцінка для теоретичної дисперсії вона дещо точніша, ніж не виправлена, тому на практиці доцільно використовувати саме її. ◀

Легко перевірити, що всі оцінки у прикладах 8.1.1 — 8.1.2 є сильно констстентними. Наступна теорема дає достатні умови консистентності моментних оцінок.

Теорема 8.1.1. *Нехай* $\mathbf{X}$ *— кратна вибірка,* $\mathbf{H}(\mathbf{t})$ *існує для всіх* $\mathbf{t} \in \Theta$, $\mathbf{H}^{-1}$ *існує і є неперервною на множині всіх можливих значень моментної функції. Тоді*

$$\hat{\vartheta}_n = \mathbf{H}^{-1}(\hat{\mathbf{h}}_n) \to \vartheta \text{ м.н. при } n \to \infty.$$

Доведення безпосередньо випливає з посиленого закону великих чисел.

Метод моментів інколи можна узагальнити на випадок неоднаково розподілених спостережень.

Приклад 8.1.3. Нехай випадкові величини, що самі мають нормальний розподіл, вимірюються різними приладами, які мають певні похибки вимірювання. Таким чином, результати вимірювання можна зобразити у вигляді

$$\xi_j = \eta_j + \varepsilon_j,$$

де η_j — справжнє значення величини, виміряної у j-му досліді, ε_j — похибка вимірювання. Тут η_j, ε_j $j = 1, \ldots, n$ вважаються незалежними в сукупності, $\eta_j \sim N(\mu, \sigma^2)$, $\varepsilon_j \sim N(0, s_j^2)$, де s_j^2 — відома дисперсія похибки при j-му вимірюванні, μ і σ^2 — невідомі параметри, які треба оцінити за даними $\mathbf{X} = (\xi_1, \ldots, \xi_n)$.

Будемо вважати, що дисперсії похибок обмежені зверху: $\sigma_j^2 < S < \infty$.

Хоча ξ_j не є однаково розподіленими випадковими величинами, але $\bar{\xi} = \bar{\eta} + \bar{\varepsilon} \to \mu$ при $n \to \infty$ м.н., оскільки $\bar{\eta} \to \mu$ за

підсиленим законом великих чисел, а $\bar{\varepsilon} \sim N(0, \sum_{j=1}^{n} \sigma_j^2/n)$ збігається до 0 м.н. (Це легко довести, використовуючи лему Бореля — Кантеллі).

Отже, $\bar{\xi}$ є незміщеною та консистентною оцінкою μ. ◀

Задача. Побудуйте консистентну оцінку σ^2 у цьому прикладі.

Розглянемо тепер приклад застосування R для обчислення оцінок методу моментів у випадку, коли розв'язати моментне рівняння (8.1) аналітично не вдається.

Приклад 8.1.4. Нехай дані $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ являють собою кратну вибірку зі зрізаного експоненційного розподілу з функцією розподілу

$$F_\xi(x) = F(x; \lambda, C) = \begin{cases} 0 & \text{при } x < 0, \\ \frac{1-\exp(-\lambda x)}{1-\exp(-\lambda C)} & \text{при } 0 \le x < C, \\ 1 & \text{при } x \ge C. \end{cases} \qquad (8.3)$$

Вважаємо поріг зрізання C відомим, а $\lambda > 0$ — невідомим параметром розподілу, який треба оцінити.

Як моментну функцію виберемо $h(x) = x$. Тоді

$$H(\lambda) = \mathsf{E}_\lambda\, \xi_1 = \frac{C}{1 - \exp(C\lambda)} + \frac{1}{\lambda}.$$

Позначимо розв'язок рівняння $\bar{\xi} = H(l)$ (відносно l) через $\hat{\lambda}_n^{MM}$ — це і буде оцінка методу моментів для λ. Оскільки розв'язати моментне рівняння аналітично не можна, то для знаходження оцінки застосуємо техніку наближеного обчислення кореня цього рівняння. Наприклад, для цього можна використати функцію `nleqslv` з бібліотеки `nleqslv`. Найпростіший варіант виклику цієї функції —

```
nleqslv(x, fn),
```

де `x` — початкове наближене значення для кореня,
`fn` — функція, корінь якої треба знайти. (Тобто ми шукаємо розв'язок рівняння `fn(x)=0`).

Значенням функції `nleqslv` є об'єкт, що має багато атрибутів, зокрема у атрибуті `$x` міститься отримане наближене значення кореня, у атрибуті `$fvec` — значення функції у точці `x` (якщо корінь знайдено правильно, це значення має бути близьким до 0).

Оформимо обчислення оцінки за даними `X` у вигляді функції:

```
> library(nleqslv)
> # функція eqv задає рівняння H(l)=Mx
> # trun - поріг зрізання експоненційного розподілу
> eqv<-function(l,Mx,trun){
+  trun/(1-exp(trun*l))+1/l-Mx
+ }
> # функція EstMM рахує оцінку lambda за даними X
> # методом моментів
> EstMM<-function(x,trun){
+  Mx<-mean(x)
+  nleqslv(1/Mx,eqv,Mx=Mx,trun=trun)$x
+ }
```

Тут ми спочатку створили функцію `eqv`, коренем якої буде наша оцінка, а потім — функцію `EstMM`, яка обчислює оцінку. Аргументами цієї функції є `x` — вибірка, за якою будується оцінка і `trun` — параметр зрізання (відомий).

Функція `EstMM` спочатку знаходить вибіркове середнє і записує його як змінну `Mx`, а потім викликає функцію `nleqslv` для розв'язування моментного рівняння. При цьому як початкове наближення для кореня рівняння вибрано `1/Mx`, тобто моментну оцінку для інтенсивності не зрізаного експоненційного розподілу.

Перевіримо, чи правильно працює наша функція на модельованих даних, які мають зрізаний експоненційний розподіл. Для цього треба згенерувати дані з таким розподілом, а потім викликати функцію `EstMM`. Спочатку ми створили функцію `rexptr`, яка генерує одне псевдовипадкове число, використовуючи генератор експоненційного розподілу `rexp()` та зрізаючи його результат доти, доки він не стане меншим, ніж поріг зрізання. Потім генеруємо вибірку, використовуючи `replicate()` і розраховуємо оцінку для інтенсивності за цією вибіркою:

```
> set.seed(2)
> # Генерація псевдовипадкових даних
> U<-2     # поріг зрізання
> l<-0.5   #  інтенсивність
> n<-10000 # обсяг вибірки
> # функція rexptr генерує одне псевдовипадкове
> # число зі зрізаним експоненційним розподілом
> # з інтенсивністю lambda та порогом зрізання trun
> rexptr<-function(lambda=1,trun=1){
+  repeat{
+    x<-rexp(1,lambda)
+    if(x<trun) break
+  }
+  x
+ }
> # Генеруємо вектор зрізаних експоненційних в.в.
> X<-replicate(n,rexptr(l,U))
> # Рахуємо оцінку
> EstMM(X,U)

[1] 0.5033029
```

Справжня інтненсивність `l=0.5`, оцінка — 0.5033029.

На перший погляд результат задовільний. Але, звичайно, якість алгоритму оцінювання не можна визначити на основі оцінки, обчисленої за однією вибіркою. Більш детальний аналіз буде проведено у прикладі 8.4.4.

Зазначимо, що розв'язок моментного рівняння може бути від'ємним для деяких вибірок. Оскільки за змістом λ додатне, немає рації використовувати від'ємне значення як його оцінку. У такому випадку можна лише стверджувати, що λ настільки мале, що його неможливо оцінити точно. При $\lambda \to 0$ функція розподілу зрізаного нормального розподілу перетворюється на рівномірну на інтервалі $[0, C]$. Якщо для реальних даних, які розглядаються, модель рівномірного розподілу допустима, то як оцінку λ, можна вибрати величину $\hat{\lambda}_n^{MMtr} = \max(\hat{\lambda}_n^{MM}, 0)$, вважаючи, що нульовому значенню оцінки відповідає рівномірний розподіл. ◄

Приклад 8.1.5. У лабораторії, де перевіряють жорсткі диски комп'ютерів, що надійшли на гарантійний ремонт, відмічають кількість дефектів, виявлених на кожному диску. У прикладі 6.4.2 запропоновано використовувати пуассонів розподіл зі зрізаним нулем для опису розподілу таких даних. Цей розподіл має один параметр — інтенсивність (λ). Дані лабораторії містять статистику кількості дефектів для однотипних дисків, вироблених трьома різними фірмами-виробниками (ці фірми позначимо A, B і C) , наведену у табл. 8.1.

Таблиця 8.1. Дані про дефекти жорстких дисків

К-сть дефектів	A	B	C
1	20	25	33
2	13	20	16
3	11	15	4
4	6	7	1
5	2	0	0
6	0	1	0

Тут у першому стовпчику вказана кількість дефектів на диску, а у наступних — скільки дисків з такою кількістю дефектів було виявлено серед дисків виробництва відповідної фірми. За цими даними треба оцінити інтенсивність появи дефектів у моделі пуасонівського розподілу для кожного виробника окремо.

Скористаємось методом моментів з моментною функцією $h(x) = x$. Ми маємо справу з груповаными вибірками — у наявних даних вказані не кількості дефектів для конкретних дисків, а частоти появи дисків з певною кількістю дефектів. Тому, як це описано у підрозд. 4.5:

$$\hat{h}_n = \hat{\mu} = \frac{1}{n}\sum_{k=1}^{K} x_k n_k,$$

де $x_k = k$ — кількість дефектів на диску, n_k — кількість дисків з такою кількістю дефектів, K — найбільша кількість дефектів, яке зустрічається у даних. (Зрозуміло, що розрахунки треба проводити окремо для кожного виробника). За п. 6.4.1 відповідний теоретичний момент дорівнює $H(\lambda) = \lambda/(1-\exp(-\lambda))$. Моментна оцінка буде розв'язком рівняння

$$\hat{\mu} = \frac{\lambda}{1-e^{-\lambda}}.$$

Це рівняння не розв'язується аналітично, тому скористаємось функцією `nleqslv`, взявши як початкове наближення λ значення $\hat{\mu}$ (яке було б моментною оцінкою λ у простій пуассоновій моделі без зрізання нуля). Для наших даних скрипт що реалізує оцінки може мати такий вигляд:

▽

```
> library(nleqslv)
> # Вводимо дані:
> A<-c(20,13,11,6,2,0)
> B<-c(25,20,15,7,0,1)
> C<-c(33,16,4,1,0,0)
```

```
> x<-1:6
> # моментна оцінка
> # перший момент як функція від інтенсивності l
> # мінус m - емпіричний момент
> moment<-function(l,m){
+ l/(1-exp(-l))-m
+ }
> # Оцінка інтенсивності
> # x - вектор значень спостережуваної змінної
> # w - частоти значень у вибірці
> EstP<-function(x,w){
+ m<-weighted.mean(x,w)
+ nleqslv(m,moment,m=m)$x
+ }
> # Оцінюємо за даними:
> EstP(x,A)

[1] 1.821516

> EstP(x,B)

[1] 1.749454

> EstP(x,C)

[1] 0.8742175
```

△

Таким чином, наша оцінка інтенсивності утворення дефектів на дисках фірми A — 1.821516, фірми B — 1.749454, фірми C — 0.8742175. За цими результатами можна сказати, що якість дисків фірм A та B приблизно однакова, а на дисках фірми C дефекти утворюються з вдвічі нижчою інтенсивністю.

Але, звичайно, це тільки оцінки. Навіть, якщо справжні інтенсивності утворення дефектів однакові для всіх трьох фірм, значення оцінок, знайдених за даними, майже завжди будуть різними.

Більше того, якщо провести оцінювання за даними інших лабораторій або за інший період часу, можна отримати інші оцінки тих самих інтенсивностей. Тому важливо вміти визначати, наскільки спостережувані відмінності оцінок відповідають справжнім відмінностям оцінюваних параметрів, а не є результатом випадкових коливань. Ми повернемось до цього питання пізніше у підрозділах, присвячених довірчим інтервалам та перевірці гіпотез[2]. ◄

Приклад 8.1.6. Розглянемо дані вимірювання певної числової характеристики ξ тварин (це може бути, напр., довжина тіла у мишей). Для досліду відібрано набір мишей, кожна з яких має один з двох можливих типів генотипу (назвемо їх Г1 і Г2). Ймовірність того, що у даної миші Г1 дорівнює 1/2 (така ж, вочевидь, і ймовірність Г2). Дослідник не може з'ясувати, який саме генотип має кожна миша, але вважає, що розподіл ξ залежить від генотипу. Точніше, розподіл $\xi \sim N(\mu_i, \sigma^2)$, якщо миша має i-й генотип. Значення μ_1, μ_2, σ^2 — невідомі, їх треба оцінити за вибіркою $\mathbf{X} = (\xi_1, \ldots, \xi_n)$, де ξ_j — значення ξ для j-ї спостережуваної миші. (Тобто у цьому прикладі різним генотипам відповідають різні середні значення характеристики ξ, але на її розкид генотип не впливає).

Таким чином, ми маємо вибірку з незалежних спостережень ξ_j, розподіл яких є сумішшю двох нормальних[3]:

$$\mathsf{P}\{\xi < x\} = F(x) = \frac{1}{2}\Phi\left(\frac{x-\mu_1}{\sigma}\right) + \frac{1}{2}\Phi\left(\frac{x-\mu_2}{\sigma}\right),$$

де $\Phi(x)$ — функція розподілу стандартного нормального розподілу, $\vartheta = (\mu_1, \mu_2, \sigma^2)^T$ — невідомий параметр.

Зрозуміло, що, помінявши місцями μ_1 та μ_2, ми не змінимо розподіл наших даних. Отже, за даними неможливо визначити, якому з двох генотипів відповідає середнє значення μ_1, а якому — μ_2, навіть, якщо нам вдасться оцінити ці параметри. Тому надалі

[2] Про довірчі інтервали — у прикладі 8.5.2, про перевірку гіпотез — у прикладі 9.3.3.

[3] Див. п. 6.4.3 про суміші кількох розподілів.

будемо вважати, що $\mu_1 \le \mu_2$, тобто першим генотипом буде той, якому відповідає менше середнє для характеристики ξ.

Для оцінки ϑ скористаємось методом моментів. Розподіл даних є симетричним навколо математичного сподівання $\mu = \mathsf{E}\,\xi_j = \frac{1}{2}(\mu_1 + \mu_2)$. Тому для оцінювання зручно використовувати трохи іншу параметризацію, ввівши параметр $\Delta = \mu_2 - \mu$. Тоді для опису розподілу ξ можна використати набір параметрів $\tilde{\vartheta} = (\mu, \Delta, \sigma^2)^T$. При цьому ξ_j можна трактувати як суму незалежних випадкових ведичин:

$$\xi_j = \mu + \eta + \zeta,$$

де $\eta \sim N(0, \sigma^2)$,

$$\zeta = \begin{cases} \Delta & \text{з ймовірністю } 1/2, \\ -\Delta & \text{з ймовірністю } 1/2. \end{cases}$$

Це зображення дозволяє дуже просто знаходити моменти ξ_j:

$$\mathsf{E}\,\xi_j = \mu,$$

$$\mathsf{D}\,\xi_j = \mathsf{E}(\xi_j - \mu)^2 = \sigma^2 + \Delta^2,$$

$$\mathsf{E}(\xi_j - \mu)^3 = 0,$$

$$\mathsf{E}(\xi_j - \mu)^4 = \mathsf{E}(\eta + \zeta)^4 = 3\sigma^4 + 6\sigma^2\Delta^2 + \Delta^4.$$

(У останній рівності ми скористались тим, що $\mathsf{E}\,\eta^4 = 3\sigma^4$).

Рівняння для третього центрованого моменту виявилось непридатним для оцінювання, оскільки цей момент не залежить від невідомих параметрів. Використаємо рівняння для першого, другого і четвертого моментів, замінивши у них теоретичні моменти емпіричними:

$$\bar{\xi} = \frac{1}{n}\sum_{j=1}^{n} \xi_j \text{ замість } \mathsf{E}\,\xi_j,$$

$$M_2 = \frac{1}{n}\sum_{j=1}^{n}(\xi_j - \bar{\xi})^2 \text{ замість } \mathsf{E}(\xi_j - \mu)^2,$$

$$M_4 = \frac{1}{n}\sum_{j=1}^{n}(\xi_j - \bar{\xi})^4 \text{ замість } \mathsf{E}(\xi_j - \mu)^4.$$

Отримуємо такі рівняння для знаходження моментних оцінок $(\hat{\mu}, \hat{\Delta}, \hat{\sigma}^2)^T$:

$$\bar{\xi} = \hat{\mu},$$

$$M_2 = \hat{\sigma}^2 + \hat{\Delta}^2,$$

$$M_4 = 3\hat{\sigma}^4 + 6\hat{\sigma}^2\hat{\Delta}^2 + \hat{\Delta}^4.$$

Якщо $3(M_2)^2 < M_4$, ці рівняння не мають розв'язку. Інакше — мають єдиний розв'язок:

$$\hat{\mu}^{MM} = \bar{\xi},$$

$$\hat{\Delta}^{MM} = \sqrt[4]{\frac{3(M_2)^2 - M_4}{2}},$$

$$\hat{\sigma}^{2,MM} = M_2 - (\hat{\Delta}^{MM})^2.$$

Це і є наші моментні оцінки. (Оскільки $(M_2)^2 < M_4$, то оцінка $\hat{\sigma}^{2.MM}$ не може бути від'ємною).

Для визначеності, при виконанні $3(M_2)^2 < M_4$ будемо вважати $\hat{\Delta}^{MM} = 0$, $\hat{\sigma}^{2,MM} = M_2$. Якщо дані ξ_j дійсно мають розподіл F, то для їхніх теоретичних моментів повинно виконуватись $3(\mathsf{D}\,\xi_j)^2 > \mathsf{E}(\xi_j - \mu)^4$, отже, при зростанні n ймовірність того, що $3(M_2)^2 < M_4$, буде прямувати до 0. Але при невеликих обсягах вибірки можливе виконання $3(M_2)^2 < M_4$ за рахунок відхилення емпіричних моментів від теоретичних. Тому наше розширення визначення моментної оцінки на цей випадок не є недоречним.

Зауважимо, що при роботі з реальними даними треба завжди враховувати можливість того, що теоретична модель зовсім не придатна для їхнього опису. Не виконання умови $3(M_2)^2 > M_4$ при значному обсязі даних може бути вказівкою саме на такий випадок.

Тепер оцінки для μ_1 та μ_2 можна визначити як:
$\hat{\mu}_1^{MM} = \hat{\mu}^{MM} - \hat{\Delta}^{MM}$,
$\hat{\mu}_2^{MM} = \hat{\mu}^{MM} + \hat{\Delta}^{MM}$.

У наступному скрипті спочатку генеруються дані з розподілом F, а потім за цими даними обчислюються моментні оцінки параметрів.

```
> set.seed(2)
> # Задаємо параметри:
> n<-300 # кількість спостережень
> m1<-1  # математичне сподівання для Г1
> m2<-5  # математичне сподівання для Г2
> s<-1   # стандартне відхилення для обох генотипів
> # Генерація даних:
> m<-c(m1,m2)
> # (y ind - номери генотипів для кожної миші)
> ind<-1+as.numeric(runif(n)<1/2)
> mx<-m[ind]
> xMixt<-rnorm(n,mx,s) # згенерована вибірка
> # Оцінка методу моментів за вибіркою x
> # результат - вектор оцінок для (m1,m2,s^2)
> EstMixMom<-function(x){
+ M1<-mean(x)
+ x0<-x-M1
+ M2<-mean(x0^2)
+ M4<-mean(x0^4)
+ delta<-(max(c(3*M2^2-M4)/2,0))^0.25
+ s2<-M2-delta^2
+ EstM1<-M1-delta
+ EstM2<-M1+delta
+ c(EstM1,EstM2,sqrt(s2))
+ }
> # значення оцінки на моделованій вибірці xMixt:
> EstMixMom(xMixt)
```

```
[1] 1.2638559 5.1202688 0.9613953
```

△

Триста спостережень згенеровані зі значеннями параметрів $\mu_1 = 1$, $\mu_2 = 5$, $\sigma^2 = 1$. Отримані оцінки $\hat{\mu}_1^{MM} = 1.2638559$, $\hat{\mu}_2^{MM} = 5.1202688$, $\hat{\sigma}^{2,MM} = 0.9613953$. Результат задовільний. ◀

8.2 Оцінки методу квантилів

Як ми бачили у розд. 4.1, вибіркове середнє і дисперсія є не робастними характеристиками вибірки. Те ж стосується будь-яких функціональних моментів з необмеженою моментною функцією. Тому коли припускається, що дані можуть бути забруднені викидами, доцільно замість моментів використовувати для оцінювання більш робастні статистики. Такими статистиками є вибіркові квантилі, якщо їхні рівні не є близькими до 0 або 1. Найбільш робастною статистикою є вибіркова медіана, тобто квантиль рівня 1/2.

Нехай $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ — кратна вибірка з неперервною функцією розподілу F_ϑ спостереження ξ_j, $\vartheta \in \Theta \subseteq \mathbb{R}$ — невідомий параметр. Позначимо $Q^{\mathbf{X}}(\alpha)$ — вибіркову квантиль рівня α, $Q^{F_\vartheta}(\alpha)$ — теоретичну квантиль розподілу F_ϑ. Тоді для всіх α, таких, що $F_\vartheta(\cdot)$ є строго зростаючою у деякому околі $Q^{F_\vartheta}(\alpha)$, виконується збіжність

$$Q^{\mathbf{X}}(\alpha) \to Q^{F_\vartheta}(\alpha), \text{ м.н. при } n \to \infty.$$

Нехай при деякому α функція $q(t) = Q^{F_t}(\alpha)$ має неперервну обернену $q^{-1}(u)$ на множині можливих значень $Q^{\mathbf{X}}(\alpha)$ (для всіх можливих значень $\mathbf{X}$). Покладемо $\hat{\vartheta}^Q = q^{-1}(Q^{\mathbf{X}}(\alpha))$. Тоді, якщо при справжньому значенні невідомого параметра ϑ функція $F_\vartheta(\cdot)$ є строго зростаючою в околі $Q^{F_\vartheta}(\alpha)$, то $\hat{\vartheta}^Q$ — строго консистентна оцінка ϑ.

Приклад 8.2.1. Нехай $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ має експоненційний розподіл з невідомою інтенсивністю λ. Тоді

$$F_\lambda(x) = (1 - \exp(-\lambda x))\mathbb{1}\{x > 0\},$$

отже $Q^{F_\lambda}(1/2) = (\log 2)/\lambda$. Як оцінку для λ можна обрати

$$\hat{\lambda}^{med} = \frac{\log 2}{\mathrm{med}(\mathbf{X})}$$

Ця оцінка є сильно консистентною і робастною. Її називають медіанною оцінкою інтенсивності експоненційного розподілу. ◄

Приклад 8.2.2. Нехай $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ — кратна вибірка з розподілу $F \sim N(\mu, \sigma^2)$, параметри μ та σ^2 — невідомі, їх треба оцінити. Оскільки щільність нормального розподілу симетрична навколо μ, то μ є медіаною цього розподілу, отже, як оцінку для нього можна взяти вибіркову медіану $\hat{\mu}_n^{med} = \mathrm{med}(\mathbf{X})$.

Для оцінки σ скористаємось тим, що

$$Q^{N(\mu,\sigma^2)}(\alpha) = \mu + \sigma Q^{N(0,1)}(\alpha).$$

Тому, для будь-якого α:

$$\sigma = \frac{Q^F(1-\alpha) - Q^F(\alpha)}{2\lambda_\alpha},$$

де $\lambda_\alpha = Q^{N(0,1)}(1-\alpha)$. Традиційно для побудови оцінки вибирають $\alpha = 1/4$ і отримують

$$\hat{\sigma}_n^{IQ} = \frac{Q^{\mathbf{X}}(3/4) - Q^{\mathbf{X}}(1/4)}{2\lambda_{1/4}} \approx \frac{\mathrm{IQ}(\mathbf{X})}{1.34898},$$

де $\mathrm{IQ}(\mathbf{X})$ — інтерквартильний розмах вибірки $\mathbf{X}$. Ця оцінка називається інтерквартильною оцінкою середньоквадратичного відхилення.

Оцінки $\hat{\mu}_n^{med}$ та $\hat{\sigma}_n^{IQ}$ є сильно консистентними. ◄

Приклад 8.2.3. Розглянемо дані спостережень нормальних випадкових величин з нормальною похибкою, описані у прикладі 3 розд. 8.1: $\mathbf{X} = (\xi_1, \ldots, \xi_n)$, $\xi_j \sim N(\mu, \sigma^2 + \sigma_j^2)$, спостереження незалежні.

Хоча спостереження не є однаково розподіленими, але у всіх ξ_j медіани однакові і дорівнюють μ. Використовуючи це, при додатковій умові $\sigma_j^2 < S < \infty$ можна показати, що $\mathrm{med}(\mathbf{X})$ буде консистентною оцінкою μ. ◀

Фактично ми, визначили квантильну оцінку як розв'язок рівняння

$$Q^{F_t}(\alpha) = Q^{\mathbf{X}}(\alpha) \tag{8.4}$$

відносно t. Зазвичай функцію $Q^{F_t}(\alpha)$ неможливо записати у явному вигляді і розв'язування цього рівняння становить самостійну проблему. У таких випадках можна переписати (8.4) в еквівалентному вигляді

$$F_t(Q^{\mathbf{X}}(\alpha)) = \alpha \tag{8.5}$$

і шукати оцінку, як розв'язок цього рівняння відносно t.

Приклад 8.2.4. Розглянемо знову кратну вибірку зі зрізаного експоненційного розподілу $\mathbf{X} = (\xi_1, \ldots, \xi_n)$, описану у прикладі 8.1.4. Для медіани рівняння (8.5) перетворюється на $F(\mathrm{med}(\mathbf{X}), \lambda, C) = 1/2$, де $F(x, \lambda, C)$ задано (8.3). Отже, медіанна оцінка для λ є коренем рівняння (відносно l):

$$\frac{1 - \exp(-l\,\mathrm{med}(\mathbf{X}))}{1 - \exp(-lC)} = 1/2.$$

У R оформити обчислення таких оцінок можна так само, як це було зроблено для моментних оцінок:

```
> # функція eqvmed задає медіанне рівняння
> #   F(medi,l)=1/2
> # medi - медіана вибірки, l - оцінка інтенсивності
> eqvmed<-function(l,medi,trun){
+     (1-exp(-l*medi))/(1-exp(-l*trun))-1/2
+  }
>  # функція EstMmed рахує оцінку lambda за даними X
>  # методом медіан
>  EstMed<-function(x,trun){
```

```
+     Mx<-median(x)
+     nleqslv(log(2)/Mx,eqvmed,medi=Mx,trun=trun)$x
+ }
```

△

На даних, згенерованих у підрозд. 8.1, функція `EstMed()` дає значення оцінки 0.5057097 (при справжньому $\lambda = 1/2$. Це трохи менш точно, ніж результат моментного оцінювання, але також досить добре.

Ця оцінка може набувати від'ємних значень, як і оцінка методу моментів у цій задачі, розглянута у прикладі 8.1.4. Такі значення можна замінити 0.

Зауважимо, що у цьому прикладі забруднення даних дуже великими викидами неможливе: спостереження, що потрапили за межі інтервалу $[0, C]$ не можуть належати зрізаному експоненційному розподілу. Такі спостереження, якщо вони потраплять до вибірки, слід трактувати не як забруднення, а як грубі помилки, і вилучати з розгляду. (Або відмовитись від моделі зрізаного експоненційного розподілу для таких даних). Тому застосування медіанної оцінки у цій задачі навряд чи можна обґрунтувати, посилаючись на вимогу робастності. ◄

Приклад 8.2.5. У задачі оцінювання інтенсивності пуассоновного розподілу зі зрізаним нулем (яка описана у прикладі 8.1.5) застосування медіанної техніки не приведе до консистентної оцінки. Дійсно, спостереження у цій задачі набувають лише цілих значень. Тому їхня медіана може бути лише цілим або півцілим числом. Зрозуміло, що корені оціночного рівняння, побудованого на такій медіані, не можуть наближати довільні додатні значення, яких може набувати інтенсивність. ◄

Це стосується й інших задач з дискретними даними: у більшості випадків квантильні оцінки для них застосовувати не можна.

Приклад 8.2.6. У задачі аналізу суміші двох нормальних розподілів з прикладу 8.1.6 квантильні оцінки побудувати можна. Зокрема, для оцінювання $\bar{\mu}$ можна скористатись вибірковою медіаною $\mathrm{med}(\mathbf{X})$. Щоб оцінити Δ і σ можна використати інтер-

квартильний розмах та різницю квантилів $Q^{\mathbf{X}}(1-\alpha) - Q^{\mathbf{X}}(\alpha)$ при якому-небудь $\alpha \neq 1/4$. Отримані квантильні рівняння не розв'язуються в явному вигляді, тому розв'язки доведеться шукати якимнебудь чисельним алгоритмом. Читачі можуть розглянути це як корисну вправу. ◄

8.3 Оцінки методу найбільшої вірогідності

На відміну від методів моментів і квантилів, метод найбільшої вірогідності у загальному випадку не потребує однорідних незалежних спостережень. Але при використанні цього методу необхідно, щоб розподіл даних описувався оцінюваними параметрами однозначно. Отже, нехай дані $\mathbf{X}$ розглядаються як випадковий елемент деякого простору можливих значень даних $\mathcal{X}$, що має розподіл $P_\vartheta^{\mathbf{X}}(A) = \mathsf{P}\{\mathbf{X} \in A\}$, $\vartheta \in \Theta \subseteq \mathbb{R}^d$ — невідомий параметр цього розподілу.

Припустимо, що існує міра μ на просторі $\mathcal{X}$ і сукупність функцій $f_\vartheta^{\mathbf{X}}(\mathbf{x})$, $f_\vartheta : \mathcal{X} \to \mathbb{R}$, $\vartheta \in \Theta$, така, що

$$P_\vartheta^{\mathbf{X}}(A) = \int_A f_\vartheta^{\mathbf{X}}(\mathbf{x})\mu(d\mathbf{x})$$

для всіх вимірних підмножин $A \in \mathcal{X}$ та всіх $\vartheta \in \Theta$.

Функція $f_\vartheta^{\mathbf{X}}$ називається щільністю розподілу $\mathbf{X}$ відносно міри μ. Якщо $\mathcal{X} \subseteq \mathbb{R}^n$, а міра μ є мірою Лебега, функцію $f_\vartheta^{\mathbf{X}}$ називають сумісною щільністю елементів вектора $\mathbf{X}$ (спостережень).

Функцією вірогідності називають випадкову функцію від невідомого параметра, яка отримується при підстановці даних замість аргументу у щільність розподілу:

$$L(\mathbf{t}) = f_{\mathbf{t}}(\mathbf{X}),\ \mathbf{t} \in \Theta.$$

Логарифмічна функція вірогідності — це логарифм $L(\mathbf{t})$, тобто $l(\mathbf{t}) = \log L(\mathbf{t})$.

Оцінкою методу найбільшої вірогідності для параметра ϑ називають таку статистику $\hat{\vartheta}_n^{ML}$, на якій досягається максимум функції вірогідності:

$$L(\hat{\vartheta}_n^{ML}) = \max_{\mathbf{t}\in\Theta} L(\mathbf{t}).$$

Зрозуміло, що оцінка найбільшої вірогідності є також точкою максимуму логарифмічної функції вірогідності.

У випадку, коли дані $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ являють собою набір незалежних спостережень ξ_j, функція вірогідності є добутком щільностей окремих спостережень:

$$L(\mathbf{t}) = \prod_{j=1}^{n} f_{\mathbf{t}}^j(\xi_j),$$

де $f_\vartheta^j(\mathbf{x})$ — щільність розподілу спостереження ξ_j в припущенні, що справжнє значення невідомого параметра дорівнює ϑ.

Для кратної вибірки $f_\vartheta^j(\mathbf{x}) = f_\vartheta(\mathbf{x})$ не залежить від j.

Приклад 8.3.1. Знову розглянемо кратну вибірку

$$\mathbf{X} = (\xi_1, \ldots, \xi_n)$$

з експоненційним розподілом. Щільність розподілу одного спостереження $f_\lambda(x) = \lambda e^{-\lambda x}\mathbb{1}\{x > 0\}$. Параметр λ невідомий, його треба оцінити. Запишемо логарифмічну функцію вірогідності:

$$l(\lambda) = \log\left(\prod_{j=1}^{n} f_\lambda(\xi_j)\right) = n\log(\lambda) - \lambda\sum_{j=1}^{n}\xi_j.$$

Бачимо, що максимум цієї функції за λ досягається при

$$\hat{\lambda}_n^{MLE} = \frac{1}{\overline{\xi}}.$$

Таким чином, у цій задачі оцінка методу найбільшої вірогідності дорівнює моментній оцінці з моментною функцією $h(x) = x$. ◄

Приклад 8.3.2. Розглянемо кратну вибірку $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ з нормального розподілу з невідомими математичним сподіванням

μ та дисперсією σ^2. Щільність одного спостереження має вигляд

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Логарифмічна функція вірогідності має вигляд

$$l(\mu, \sigma^2) = -n(\log(2\pi)/2 + \log\sigma) - \frac{\sum_{j=1}^{n}(\xi_j - \mu)^2}{2\sigma^2}.$$

Тривимірний графік для функції вірогідності, побудованої за вибіркою з 1000 спостережень з $N(1, 2.25)$-розподілом, див. на рис. 8.1.

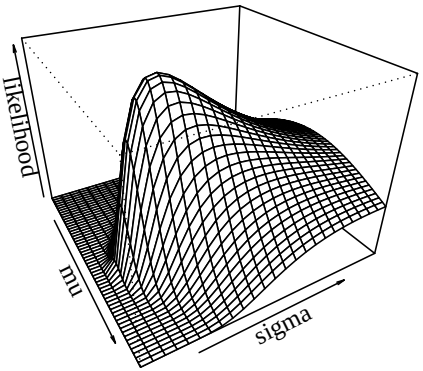

Рис. 8.1. Функція вірогідності для параметрів нормального розподілу

Взявши похідні від цієї функції за обма аргументами і прирівнявши їх до 0, знаходимо точку максимуму, яка і буде оцінкою методу найбільшої вірогідності

$$\hat{\mu}_n^{MLE} = \bar{\xi},\ \hat{\sigma}_n^{2\,MLE} = S^2(\mathbf{X}).$$

Отже, і у цьому випадку оцінки методу найбільшої вірогідності дорівнюють отриманим у прикладі 8.1.2 оцінкам методу моментів.

Лінії рівня для функції вірогідності зображені на рис. 8.2. Хрестиком відмічено положення точки максимуму, що відповідає оцінці $(\hat{\mu}_n^{MLE}, \hat{\sigma}_n^{MLE})$.

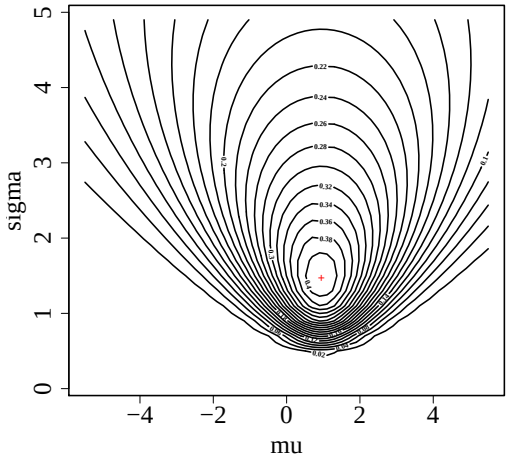

Рис. 8.2. Лінії рівня функції вірогідності для параметрів нормального розподілу

◄

Приклад 8.3.3. Як побудувати оцінку методу найбільшої вірогідності у задачі оцінювання математичного сподівання та дисперсії гауссового розподілу за спостереженнями з неоднорідними похибками з прикладу 8.1.3? У цьому випадку щільність розподілу одного спостереження ξ_j:

$$f^j_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma_j^2)}} \exp\left(-\frac{(x-\mu)^2}{2(\sigma^2+\sigma_j^2)^2}\right).$$

Логарифмічна функція вірогідності має вигляд

$$l(\mu, \sigma) = \prod_{j=1}^{n} \log f^{j}_{\mu,\sigma}(\xi_j).$$

Перетворюючи цей вираз отримуємо, що точки максимуму функції $l(\mu, \sigma)$ є точками мінімуму функції

$$r(\mu, \sigma) = \sum_{j=1}^{n} \log(\sigma^2 + \sigma_j^2) + \sum_{j=1}^{n} \left(\frac{(\xi_j - \mu)^2}{(\sigma^2 + \sigma_j^2)} \right).$$

При фіксованому σ мінімум цієї функції за μ досягається при[4]

$$\mu = \mu(\sigma) = \frac{\sum_{j=1}^{n} \frac{\xi_j}{\sigma^2+\sigma_j^2}}{\sum_{j=1}^{n} \frac{1}{\sigma^2+\sigma_j^2}}.$$

Таким чином, щоб отримати оцінку найбільшої вірогідності параметрів μ та σ можна спочатку знайти оцінку $\hat{\sigma}_n^{MLE}$ як точку мінімуму функції $r(\mu(s), s)$ за s, а потім отримати оцінку для μ як $\hat{\mu}_n^{MLE} = \mu(\hat{\sigma}_n^{MLE})$.

Реалізуємо цю ідею в `R`. Мінімізувати функцію $r(\mu(s), s)$ аналітично не можна, тому скористаємось наближеним методом Ньютона. Цей метод реалізує функція `nlm()`[5]. Виклик цієї функції: `nlm(f,p,...)`, де `f` — числова функція векторного аргументу, яку треба мінімізувати, `p` — вектор початкових значень для точки мінімуму. Функція `f` повинна мати першим параметром вектор, за яким відбувається мінімізація, він має бути тієї ж довжини, що і `p`. Замість `...` у виклику `nlm()` можна вказати значення інших параметрів функції `f`, якщо вони потрібні. Значення точки мінімуму функція `nlm()` повертає у атрибуті `$estimate`.

[4]Зауважимо, що коли $\sigma = 0$, тобто в усіх дослідах вимірюється одна і та сама фізична величина μ, ми отримуємо формулу для навантаженого середнього, еквівалентну запропонованій у прикладі 4.5.2.

[5]Можна також використати функцію `optim()`, яка дозволяє більше варіантів вибору методу чисельної оптимізації.

Оцінку (μ, σ) можна організувати так:

```
> ll<-function(s,x,sigm)
+ {
+  ss<-s^2+sigm^2
+  mu<-sum(x/ss)/sum(1/ss)
+  sum(log(ss))+sum((x-mu)^2/ss)
+ }
> EstMLEGauss<-function(x,sigm)
+ {
+  sEst<-nlm(ll,sd(x),x=x,sigm=sigm)$estimate
+  ss<-sEst^2+sigm^2
+  muEst<-sum(x/ss)/sum(1/ss)
+  c(muEst,sEst)
+ }
```

Тут функція `ll(s,x,sigm)` забезпечує обчислення $r(\mu(s), s)$. Параметр `x` — це вибірка, за якою проводиться оцінювання, `sigm` — вектор значень стандартних відхилень помилок $(\sigma_1, \ldots, \sigma_n)$ (він повинен мати таку ж довжину, як і `x`).

Функція `EstMLEGauss` знаходить точку мінімуму функції `ll()`, використовуючи як початкове наближення стандартне відхилення вибірки (це, вочевидь, завищена оцінка, оскільки у неї входять дисперсії похибок).

Перевіримо роботу цієї оцінки на модельованих даних. Стандартні відхилення похибок σ_j для моделювання виберемо так, щоб вони рівномірно збільшувались від 1 на початку до 3 наприкінці спостережень. Оцінюване стандартне відхилення виберемо рівним $\sigma = 0.5$, математичне сподівання $\mu = 1$:

```
> set.seed(2)
> n<-1000     # обсяг вибірки
> mu<-1       # математичне сподівання
> sigma0<-0.5 # стандартне відхилення
```

```
> # стандартні відхилення похибок:
> sigm<-seq(1,3,length.out = n)
> # генерація даних:
> x<-rnorm(n,mu,sigma0)+sigm*rnorm(n)
> res=EstMLEGauss(x,sigm) # підрахунок оцінки
> res # значення оцінок для мат.спод. та ст. відх.:

[1]  1.0718504 -0.4625247

> # графік функції r(s):
> s<-seq(-0.8,0.8,length.out=100)
> y<-sapply(s,ll,x=x,sigm=sigm)
> plot(s,y,type="l")
> abline(v=res[2],col="red")
```

△

Графік функції $r(s)$ для цього прикладу зображений на рис. 8.3. На ньому червоною вертикальною прямою відмічено знайдене нами положення точки мінімуму — оцінки $\hat{\sigma}_n^{MLE} = -0.4625247$.

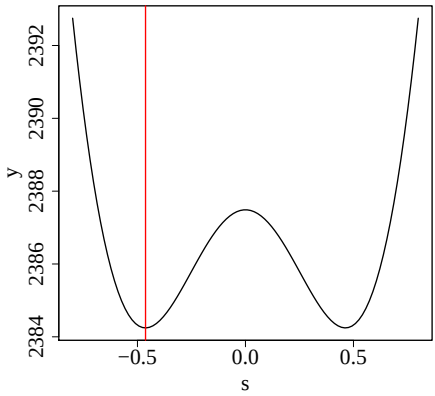

Рис. 8.3. Перетворена функція вірогідності для даних з похибками

Але ж вона від'ємна? Так, насправді ми всюди при оцінці використовували не s, а s^2, тому, якщо s, точка мінімуму $r(s)$, то і $-s$ є точкою мінімуму. Тому алгоритм наближеного пошуку може знайти або додатну, або від'ємну точку. Якщо потрібне саме додатне значення, не забудьте взяти модуль від оцінки.

Оцінка μ за методом найбільшої вірогідності у цьому прикладі дорівнює $\hat{\mu}_n^{MLE} = 1.0718504$. ◀

Приклад 8.3.4. Тепер розглянемо оцінку методу найбільшої вірогідності для інтенсивності λ зрізаного експоненційного розподілу з прикладу 8.1.4.

Логарифм функції вірогідності у цій задачі має вигляд

$$l(\lambda) = n(\log(\lambda) - \log(1 - e^{-C\lambda}) - \lambda\bar{\xi}).$$

Обчислення оцінки можна організувати аналогічно тому, як це зроблено у прикладі 8.3.3:

```
> # функція ll рахує - log(вірогідність) з точністю
> # до константи. Mx - вибіркове середнє,
> # trun - поріг зрізання експоненційного розподілу
> ll<-function(l,Mx,trun){
+  -log(l/(1-exp(-l*trun)))+l*Mx
+ }
> # функція EstMLtr рахує оцінку lambda за даними x
> # методом найбільшої вірогідності
> EstMLEtr<-function(x,trun) {
+  Mx<-mean(x)
+  nlm(ll,1/Mx,Mx=Mx,trun=trun)$estimate
+ }
```

Обчисливши цю оцінку на тих самих даних, на яких перевірялась робота моментних оцінок $\hat{\lambda}_n^{MM}$, можна пересвідчитись, що значення оцінок збігаються з точністю до округлення. І дійсно, продиференціювавши функцію вірогідності та прирівнявши її до

0 для знаходження екстремуму, отримуємо в точності моментне рівняння для $\hat{\lambda}_n^{MM}$. Таким чином, ми фактично отримали дві алгоритмічні реалізації однієї і тієї ж оцінки: в першому випадку за допомогою чисельного розв'язування нелінійного рівняння, у другому — з використанням чисельної нелінійної оптимізації. Яка з цих реалізацій виявиться кращою (швидкодійнішою, стабільнішою, точнішою) залежить від того, як запрограмовані відповідні методи розв'язування рівнянь та мінімізації. ◄

Приклад 8.3.5. Для задачі оцінки інтенсивності пуассонового розподілу зі зрізаним нулем (приклад 8.1.5) також можна застосувати метод найбільшої вірогідності. Оскільки дані є дискретними, то звичайної ймовірнісної щільності (відносно міри Лебега) вони не мають. Але у даному випадку для побудови функції вірогідності можна скористатись щільністю розподілу спостережень відносно лічильної міри. Ця "дискретна щільність" дорівнює ймовірності того, що випадкова величина набуває заданого значення:

$$f_\lambda(k) = \mathsf{P}\{\xi_j = k\}$$

$$= \frac{\lambda^k}{k!(e^\lambda - 1)},$$

де ξ_j — випадкова величина з пуассоновим розподілом зі зрізаним нулем та інтенсивністю λ.

Таким чином, логарифмічна функція вірогідності матиме вигляд

$$l(\lambda) = \sum_{k=1}^{K} n_k \log\left(\frac{\lambda^k}{k!(e^\lambda - 1)}\right)$$

$$= n(\hat{\mu}\log(\lambda) - \log(e^\lambda - 1)) - \sum_{k=1}^{K} n_k \log(k!).$$

(тут використані ті ж позначення, що і у прикладі 8.1.5). Диференціюючи $l(\lambda)$ за λ і прирівнюючи похідну до 0, отримуємо рівняння

для оцінки методу найбільшої вірогідності

$$n\left(\frac{\hat{\mu}}{\lambda}-\frac{e^{\lambda}}{e^{\lambda}-1}\right)=0.$$

Це рівняння еквівалентне моментному рівнянню, отриманому у прикладі 8.1.5, отже, оцінка методу моментів і в цьому прикладі виявилась оцінкою найбільшої вірогідності. ◀

Приклад 8.3.6. Застосуємо метод найбільшої вірогідності для оцінки параметрів у моделі суміші двох гауссових розподілів з прикладу 8.1.6. Дані є вибіркою $\mathbf{X}=(\xi_1,\ldots,\xi_n)$ з незалежних, однаково розподілених випадкових величин зі щільністю розподілу

$$f(x)=\frac{1}{2\sigma}\left(\varphi\left(\frac{x-\mu_1}{\sigma}\right)+\varphi\left(\frac{x-\mu_2}{\sigma}\right)\right).$$

Логарифмічну функцію вірогідності можна записати у вигляді

$$l(\mu_1,\mu_2,\sigma)=n(\log(\sigma^2)/2-\log(2))+$$

$$+\sum_{j=1}^{n}\left(\varphi\left(\frac{\xi_j-\mu_1}{\sigma}\right)+\varphi\left(\frac{\xi_j-\mu_2}{\sigma}\right)\right).$$

Точку максимуму цієї функції знайти аналітично неможливо, тому скористаємось функцією `nlm()` для чисельної максимізації. Скрипт, що реалізує оцінку методу найбільшої вірогідності, має вигляд:

```
> # ll - логарифмічна функція вірогідності
> # (з точністю до несутєвих констант)
> # theta - невідомий параметр у форматі
> #         (mu1,mu2,sigma)
> # x - вибірка
> ll<-function(theta,x){
+ M1<-theta[1]
+ M2<-theta[2]
```

```
+ s<-theta[3]
+ log(s^2)/2-mean(log(exp(-(x-M1)^2/(2*s^2))+
+                        exp(-(x-M2)^2/(2*s^2))))
+ }
> #
> # EstMixML Оцінка методу найбільшої вірогідності
> # аргумент x - вибірка,
> # результат - вектор, значення оцінки у форматі
> #         (mu1,mu2,sigma)
> EstMixML<-function(x)
+            nlm(ll,EstMixMom(x),x=x)$estimate
> # обчислення оцінки на вибірці xMixt, згенерованій
> # у прикладі моментної оцінки:
>  EstMixML(xMixt)

[1] 1.2451458 5.0869098 0.9762926
```

△

Як початкові значення для пошуку точки максимуму використовуються оцінки методу моментів, обчислені функцією `EstMixMom()` з прикладу 8.1.6. Оцінки розраховані на даних, згенерованих у тому ж прикладі. Як ми пам'ятаємо, при генерації використовувались значення $\mu_1 = 1$, $\mu_2 = 5$, $\sigma = 1$. Отримані оцінки методу найбільшої вірогідності $\hat{\mu}_1 = 1.245$, $\hat{\mu}_2 = 5.0869$, $\hat{\sigma} = 0.976$. Це трохи точніше, ніж те, що було отримано методом моментів у прикладі 8.1.6. Звичайно, було б передчасно робити висновки про точність оцінок методу моментів та методу найбільшої вірогідності лише за результатом одного цього експерименту. ◀

8.4 Асимптотична нормальність і матриця розсіювання оцінок

У попередніх підрозділах описано три способи побудови оцінок невідомих параметрів. Їхнє застосування, як ми бачили, приводить до різних оцінок. Наприклад, для оцінювання інтенсивності λ

експоненційного розподілу ми отримали три різних оцінки: оцінку на основі першого моменту $\hat{\lambda}_n^{(1)}$ (вона також є оцінкою найбільшої вірогідності), оцінку на основі другого моменту $\hat{\lambda}_n^{(2)}$ та медіанну оцінку $\hat{\lambda}_n^{med}$.

Яка з цих оцінок краща? Поки що ми можемо лише стверджувати, що медіанна оцінка є робастною, а моментні — ні. Інакше кажучи, якщо дані забруднені спостереженнями, що мають не такий розподіл, як основна маса, на моментні оцінки покладатись не варто, а медіанна може давати більш відповідний результат.

А яка з цих оцінок точніша, якщо наша модель повністю відповідає даним? Для того, щоб відповісти на це запитання можна провести комп'ютерний експеримент: згенерувати дані із заданим розподілом, підрахувати різні оцінки і порівняти їх із справжнім значенням параметра. Зрозуміло, що за одним набором випадкових даних результат буде один, за іншим — інший. Тому в експерименті треба згенерувати багато різних наборів даних з одним і тим самим розподілом, по кожному набору обчислити всі оцінки, які порівнюються. Після цього можна порівнювати розподіли отриманих оцінок: які з них мають більший розкид навколо середнього, і на скільки середнє оцінок відхиляється від оцінюваного параметра.

Проведення таких експериментів на сьогодні є обов'язковим елементом розробки нових алгоритмів статистичного оцінювання. Але, звичайно, у такий спосіб неможливо перевірити роботу оцінок для всіх можливих значень оцінюваних параметрів.

Виявляється, що задача теоретичного порівняння оцінок значно спрощується, якщо розглядати їх поведінку при нескінченному зростанні обсягу даних. У багатьох випадках оцінки виявляються асимптотично нормальними, тобто розподіл їх відхилення від справжнього значення стає близьким до нормального розподілу з нульовим середнім. Оскільки такий розподіл в одновимірному випадку характеризується одним числом — дисперсією, то і порівнювати різні асимптотично нормальні оцінки можна лише за цією дисперсією — коефіцієнтом розсіювання.

Опишемо цей підхід більш детально, розглядаючи одразу випадок d-вимірного невідомого параметра

$$\vartheta = (\vartheta_1, \ldots, \vartheta_d)^T \in \Theta \subseteq \mathbb{R}^d$$

та відповідної консистентної оцінки

$$\hat{\vartheta}_n = (\vartheta_{1n}, \ldots, \vartheta_{dn})^T.$$

З консистентності оцінки випливає збіжність $\hat{\vartheta}_n - \vartheta \to 0$ (за ймовірністю), коли $n \to \infty$. Для характеризації точності оцінки важливо знати, як швидко ця різниця прямує до 0. Швидкість збіжності досліджують, домножаючи $\hat{\vartheta}_n - \vartheta$ на нормуючу послідовність a_n, що прямує до нескінченності. Цю послідовність підбирають так, щоб $a_n(\hat{\vartheta}_n - \vartheta)$ прямувало не до 0 і не до нескінченності, а до деякого проміжного значення.

Можна довести, що за досить широких умов при правильному виборі нормування, розподіл такої нормованої різниці прямує до нормального з нульовим математичним сподіванням — $N(0, \mathbf{V}_{\hat{\vartheta}}(\vartheta))$. Тут $\mathbf{V}_{\hat{\vartheta}}(\vartheta)$ — коваріаційна матриця граничного нормального розподілу, що залежить від справжнього значення невідомого параметра ϑ. Цю матрицю називають **матрицею розсіювання** оцінки $\hat{\vartheta}_n$. У багатьох задачах оцінювання для кратних вибірок правильним нормуванням є $a_n = \sqrt{n}$.

У одновимірному випадку $d = 1$, коли оцінюваний параметр — це одне число, матриця розсіювання також складається з одного елемента — дисперсії $v_{\hat{\vartheta}}(\vartheta)$ граничного нормального розподілу нормованої оцінки. Це число називають **коефіцієнтом розсіювання**.

Отже, в одновимірному випадку зі збіжності $\sqrt{n}(\vartheta_n - \vartheta)$ до $N(0, v_{\hat{\vartheta}}(\vartheta))$ випливає, що для будь-якого $\lambda > 0$:

$$\mathsf{P}\left\{ \frac{|\sqrt{n}(\hat{\vartheta}_n - \vartheta)}{\sqrt{v_{\hat{\vartheta}}(\vartheta)}} \le \lambda \right\} \to \mathsf{P}\{|\zeta| \le \lambda\} = 1 - 2\Phi(-\lambda), \qquad (8.6)$$

де $\zeta \sim N(0,1)$, Φ — функція розподілу $N(0,1)$. Поклавши $\lambda_\alpha = Q^\Phi(1-\alpha)$, отримуємо

$$\mathsf{P}\left\{|\hat{\vartheta}_n - \vartheta| \le \frac{\sqrt{v_{\hat{\vartheta}}(\vartheta)}\lambda_{\alpha/2}}{\sqrt{n}}\right\} = 1-\alpha. \tag{8.7}$$

Таким чином, при великих обсягах вибірки ширина інтервалу, у який відхилення оцінки від оцінюваного значення потрапляє із заданою ймовірністю $1-\alpha$, прямо пропорційна $\sqrt{v_{\hat{\vartheta}}(\vartheta)}$ (для всіх $\alpha > 0$). Тому точність асимптотично нормальних оцінок прийнято характеризувати за допомогою коефіцієнта розсіювання: чим він менший, тим оцінка точніша.

У багатовимірному випадку також, чим "менша" матриця $\mathbf{V}_{\hat{\vartheta}}(\vartheta)$, тим оцінка $\hat{\vartheta}_n$ точніша. Порівняння матриць тут виконується Льовнером: $\mathbf{A} < \mathbf{B}$ рівносильно тому, що $\mathbf{B} - \mathbf{A}$ є невід'ємно визначеною матрицею.

З'ясуємо тепер, як обчислювати матриці розсіювання. Розглянемо випадок, коли дані являють собою кратну вибірку $\mathbf{X} = (\xi_1, \ldots, \xi_n)$, невідомий параметр $\vartheta = (\vartheta_1, \ldots, \vartheta_d)$ є d-вимірним і його оцінка $\hat{\vartheta}_n = (\vartheta_{1n}, \ldots, \vartheta_{dn})$ також.

Матриця розсіювання моментної оцінки. Нехай використовується моментна оцінка з моментною функцією

$$\mathbf{h}(\xi) = (h_1(\xi), \ldots, h_d(\xi))^T,$$

і вектором теоретичних моментів —

$$\mathbf{H}(\mathbf{t}) = (H_1(\mathbf{t}), \ldots, H_d(\mathbf{t}))^T = \mathsf{E}_\mathbf{t}\,\mathbf{h}(\xi_1), \mathbf{t} = (t_1, \ldots, t_d)^T \in \Theta.$$

Позначимо $\mathbf{H}'(\mathbf{t})$ матрицю перших похідних від $\mathbf{H}(\mathbf{t})$ (матрицю Якобі):

$$\mathbf{H}'(\mathbf{t}) = \frac{\partial}{\partial \mathbf{t}^T}\mathbf{H}(\mathbf{t}) = \begin{pmatrix} \frac{\partial H_1(\mathbf{t})}{\partial t_1} & \cdots & \frac{\partial H_1(\mathbf{t})}{\partial t_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial H_d(\mathbf{t})}{\partial t_1} & \cdots & \frac{\partial H_d(\mathbf{t})}{\partial t_d} \end{pmatrix}$$

Теорема 8.4.1. *Нехай виконані такі умови.*

1. Елементи коваріаційної матриці $\mathbf{D}_\vartheta = \operatorname{cov}(\mathbf{h}(\xi_1))$ *є скінченними.*

2. Існує обернена функція $\mathbf{H}^{-1}$.

3. Функція $\mathbf{H}'(\mathbf{t})$ *є неперервною за* $\mathbf{t}$ *у деякому околі* ϑ.

Тоді консистентна моментна оцінка $\hat{\vartheta}_n$*, яка задовольняє рівняння* $\mathbf{H}(\hat{\vartheta}_n) = \hat{\mathbf{h}}_n$ *є асимптотично нормальною з матрицею розсіювання*

$$\mathbf{V}_{\hat{\vartheta}}(\vartheta) = (\mathbf{H}'(\vartheta))^{-T}\mathbf{D}_\vartheta(\mathbf{H}'(\vartheta))^{-1}. \tag{8.8}$$

У одновимірному випадку формула (8.8) перетворюється на

$$v_{\hat{\vartheta}} = \frac{\mathsf{D}_\vartheta\, h(\xi_1)}{(H'(\vartheta))^2}. \tag{8.9}$$

З'ясуємо, звідки взялась формула (8.8). Замінимо моментне рівняння його наближенням, використовуючи розклад $\mathbf{H}$ за формулою Тейлора в околі точки ϑ:

$$\mathbf{H}(\vartheta) + \mathbf{H}'(\tau)(\hat{\vartheta}_n - \vartheta) = \hat{\mathbf{h}}_n,$$

де τ — проміжна точка між ϑ і $\hat{\vartheta}_n$. Враховуючи, що $\mathbf{H}(\vartheta) = \mathsf{E}\,\hat{\mathbf{h}}_n$, отримуємо

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta) = (\mathbf{H}(\tau))^{-1}\sqrt{n}(\hat{h}_n - \mathsf{E}\,\hat{h}_n). \tag{8.10}$$

За центральною граничною теоремою, розподіл $\sqrt{n}(\hat{h}_n - \mathsf{E}\,\hat{h}_n)$ збігається до розподілу випадкового вектора $\zeta \sim N(0, \mathbf{D}_\vartheta)$. Враховуючи неперервність $\mathbf{H}'(\mathbf{t})$, отримуємо звідси формулу (8.8).

Матриця розсіювання оцінки найбільшої вірогідності. Нехай розподіл спостережень має щільність $f_\vartheta(\mathbf{x})$ відносно деякої міри μ. Позначимо

$$\mathbf{I}(\vartheta) = \mathsf{E}_\vartheta\, \frac{\partial}{\partial\vartheta} \ln f_\vartheta(\xi_1) \left(\frac{\partial}{\partial\vartheta} \ln f_\vartheta(\xi_1)\right)^T =$$

$$= \left(\int \frac{\frac{\partial}{\partial \vartheta_i} f_\vartheta(\mathbf{x}) \frac{\partial}{\partial \vartheta_k} f_\vartheta(\mathbf{x})}{f_\vartheta(x)} \mu(dx) \right)_{i,k=1}^{d}$$

— інформаційна матриця Фішера для параметра ϑ за одним спостереженням ξ_1.

Міркування, подібні розглянутим для моментних оцінок, приводять до наступної формули для матриці розсіювання оцінок методу найбільшої вірогідності $\hat{\vartheta}_n$:

$$\mathbf{V}_{\hat{\vartheta}}(\vartheta) = (\mathbf{I}(\vartheta))^{-1} \tag{8.11}$$

— матриця розсіювання є матрицею, оберненою до інформаційної.

У одновимірному випадку для коефіцієнта розсіювання отримуємо

$$v_{\hat{\vartheta}}(\vartheta) = \frac{1}{I(\vartheta)},$$

де

$$I(\vartheta) = \int \frac{\left(\frac{\partial}{\partial \vartheta} f_\vartheta(x)\right)^2}{f_\vartheta(x)} \mu(dx)$$

— інформація Фішера про параметр ϑ, що міститься в одному спостереженні.

Матриця розсіювання для квантильних оцінок. Нехай знову, дані являють собою кратну вибірку $\mathbf{X}$ випадкових величин ξ_j з функцією розподілу F_ϑ та щільністю $f_\vartheta(x)$, $\vartheta \in \Theta \in \mathbb{R}^d$. Зафіксуємо набір рівнів $\alpha = (\alpha_1, \ldots, \alpha_d)$, $0 < \alpha_i < 1$. Позначимо $\mathbf{q}^\alpha(\vartheta) = (Q^{F_\vartheta}(\alpha_1), \ldots, Q^{F_\vartheta}(\alpha_1))$ — вектор теоретичних квантилей, $\hat{\mathbf{q}}_n^\alpha = (Q^{\mathbf{X}}(\alpha_1), \ldots, Q^{\mathbf{X}}(\alpha_1))$ — набір емпіричних квантилей. Нехай для всіх α_i виконано $f_\vartheta(Q^{F_\vartheta}(\alpha_i)) > 0$. Тоді з наслідка 1 підрозд. 7 розд. 1 [3] випливає, що $\sqrt{n}(\hat{\mathbf{q}}_n^\alpha - \mathbf{q}^\alpha(\vartheta))$ збігається за розподілом до $N(0, \mathbf{C})$, де $\mathbf{C} = (c_{i,k})_{i,k=1}^d$:

$$c_{ik} = \frac{\min(\alpha_i, \alpha_k) - \alpha_i \alpha_k}{f_\vartheta(Q^{F_\vartheta}(\alpha_i)) f_\vartheta(Q^{F_\vartheta}(\alpha_k))}. \tag{8.12}$$

Нехай квантильна оцінка ϑ_n^α для ϑ є розв'язком рівняння

$$\mathbf{q}^\alpha(\mathbf{t}) = \hat{\mathbf{q}}_n^\alpha$$

відносно $\mathbf{t}$.

Тоді міркування, аналогічні до тих, які ми використали для моментних оцінок, приводять до наступного виразу для матриці розсіювання оцінки $\hat{\vartheta}_n^\alpha$:

$$\mathbf{V}_{\hat{\vartheta}^\alpha}(\vartheta) = \mathbf{Q}^{-T}\mathbf{C}\mathbf{Q}^{-1}, \tag{8.13}$$

де $\mathbf{Q} = \frac{\partial}{\partial \vartheta^T}\mathbf{q}^\alpha(\vartheta)$.

Зокрема, для медіанної оцінки ϑ_n^{med}, що є розв'язком рівняння

$$q^{1/2}(t) = \mathrm{med}(X),$$

коефіцієнт розсіювання дорівнює

$$v_{\vartheta^{med}} = \frac{1}{4(f_\vartheta(\mathrm{med}(\xi_1))(q^{1/2}(\vartheta))')^2}. \tag{8.14}$$

Подивимось, як записати коефіцієнт розсіювання квантильної оцінки, якщо вона визначається як розв'язок рівняння (8.5). Формулу, яку ми отримаємо, можна вивести з (8.13), але ми зробимо це безпосередньо.

Отже, нехай для оцінки $\hat{\vartheta}_n$ виконується рівняння

$$F_{\hat{\vartheta}_n}(\hat{q}_n) = \alpha,$$

де $\hat{q}_n = Q^{\mathbf{X}}(\alpha)$. В умовах, що вказані вище, $\sqrt{n}(\hat{q}_n - q_\alpha)$ збігається за розподілом до $N(0, c)$, де

$$q_\alpha = Q^{F_\vartheta}(\alpha),\ c = \alpha(1-\alpha)/(f_\vartheta(q_\alpha))^2.$$

Розкладаючи ліву частину цієї рівності в околі точки (ϑ, q_α), отримуємо

$$F_\vartheta(q_\alpha) + \frac{\partial}{\partial t}F_t(q)(\hat{\vartheta}_n - \vartheta) + \frac{\partial}{\partial q}F_t(q)(\hat{q}_n - q_\alpha) = \alpha,$$

де t — проміжна точка між $\hat{\vartheta}_n$ і ϑ, q — проміжна точка між $\hat{q}_n$ і q_α. Звідси отримуємо

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta) \sim \frac{\frac{\partial}{\partial q}F_\vartheta(q_\alpha)}{\frac{\partial}{\partial \vartheta}F_\vartheta(q_\alpha)}\sqrt{n}(\hat{q}_n - q_\alpha).$$

Використовуючи асимптотичну нормальність $\hat{q}_n$, отримуємо коефіцієнт розсіювання $\hat{\vartheta}$:

$$v_{\hat{\vartheta}_n} = \left(\frac{\frac{\partial}{\partial q}F_\vartheta(q_\alpha)}{\frac{\partial}{\partial \vartheta}F_\vartheta(q_\alpha)}\right)^2 \frac{\alpha(1-\alpha)}{(f_\vartheta(q_\alpha))^2}. \tag{8.15}$$

Приклад 8.4.1. Повернемося до розгляду задачі оцінки інтенсивності λ експоненційного розподілу за кратною вибіркою $\mathbf{X} = (\xi_1, \ldots, \xi_n)$. У попередніх розділах були введені три оцінки:

$$\hat{\lambda}_n^{(1)} = 1/\bar{\xi},\ \hat{\lambda}_n^{(2)} = \sqrt{\frac{2n}{\sum_{j=1}^n \xi_j^2}},\ \hat{\lambda}_n^{\mathrm{med}} = \frac{\log 2}{\mathrm{med}(X)}.$$

Перші дві оцінки отримані методом моментів з моментними функціями $h_1(x) = x$ та $h_2(x) = x^2$. Враховуючи, що

$$\mathsf{D}_\lambda\, h_1(\xi_1) = \frac{1}{\lambda^2},\ \mathsf{D}_\lambda\, h_2(\xi_1) = \frac{23}{\lambda^4},$$

за (8.9) отримуємо коефіцієнти розсіювання цих оцінок:

$$v_{\hat{\lambda}^{(1)}} = \lambda^2, v_{\hat{\lambda}^{(2)}} = \frac{23}{16}\lambda^2.$$

Третя оцінка — медіанна. Теоретична медіана експоненційного розподілу $\mathrm{med}(\xi_1) = \log 2/\lambda$, а щільність розподілу у медіані — $f_\lambda(\mathrm{med}(\xi_1)) = \lambda/2$. Тому коефіцієнт розсіювання цієї оцінки

$$v_{\hat{\lambda}^{med}} = \frac{\lambda^2}{(\log 2)^2}.$$

Оскільки $23/16 \approx 1.4375 < 2.08137 \approx 1/(\log 2)^2$, ці результати показують, що найбільш точною при великих обсягах вибірок є оцінка $\hat{\lambda}_n^{(1)}$, наступною — $\hat{\lambda}_n^{(2)}$, а найменш точною з трьох розглянутих є медіанна оцінка.

Відношення коефіцієнтів варіації двох різних оцінок одного параметра називають їх **відносною асиптотичною ефективністю** (asymptotic relative efficiency, ARE). Наприклад,

$$v_{\hat{\lambda}^{med}}/v_{\hat{\lambda}^{(1)}} = 1/(\log 2)^2 \approx 2.08137$$

— ARE оцінки найбільшої вірогідності порівняно з медіанною оцінкою. ARE має простий статистичний зміст, який легко зрозуміти враховуючи (8.7). Якщо ми раніше користувались оцінкою $\hat{\lambda}_n^{(1)}$, а тепер замість неї хочемо використати $\hat{\lambda}_n^{(med)}$, то для забезпечення такої ж точності, як і раніше, нам доведеться збільшити обсяг вибірки у два (точніше у 2.08137) рази. Це варто робити, якщо вигоди від робастності медіанної оцінки перевищують додаткові витрати на збільшення обсягу спостережень. Інакше слід використовувати оцінку найбільшої вірогідності.

Те, що найкращою виявиться $\hat{\lambda}_n^{(1)}$, можна було сказати вже тоді, коли виявилось, що це оцінка найбільшої вірогідності. Справа в тому, що при виконанні досить широких умов[6] ОНВ є асимптотично нормальними оцінками з коефіцієнтом розсіювання, найменшим серед всіх "правильних" (т. зв. регулярних) оцінок. ◀

Приклад 8.4.2. Нехай тепер оцінюються математичне сподівання μ і дисперсія σ^2 за кратною вибіркою гауссових спостережень $\mathbf{X}$. Ми отримали по дві оцінки для кожного параметра: метод моментів дав той самий результат, що і метод найбільшої вірогідності (приклади 8.1.2 і 8.3.2):

$$\hat{\mu}^{MLE} = \bar{\xi},\ \hat{\sigma}_n^{2\,MLE} = S^2(\mathbf{X}),$$

[6] Умов регулярності, [3], розді. 2, підрозд. 16.

а метод квантилів (приклад 8.2.2) —

$$\hat{\mu}_n^{med} = \operatorname{med} \mathbf{X},\ \hat{\sigma}^{2\ IQ} = \left(\frac{Q^{\mathbf{X}}(3/4) - Q^{\mathbf{X}}(1/4)}{2\lambda_{\alpha/4}} \right)^2 .$$

(Тут, як і раніше, $\lambda_\alpha = Q^{N(0,1)}(1-\alpha)$).

Для обчислення матриці розсіювання оцінок найбільшої вірогідності знайдемо інформаційну матрицю для $\vartheta = (\mu, \sigma^2)^T$. Бачимо, що[7]

$$\frac{\partial}{\partial \mu} f_\vartheta(\xi_1) = \frac{\xi_1 - \mu}{\sigma^2}, \quad \frac{\partial}{\partial \sigma^2} f_\vartheta(\xi_1) = -\frac{1}{2\sigma^2} - \frac{(\xi_1 - \mu)^2}{2\sigma^4}.$$

Отже, інформаційна матриця для одного спостереження має вигляд

$$\mathbf{I}(\vartheta) = \mathsf{E} \begin{pmatrix} \frac{(\xi_1-\mu)^2}{\sigma^4} & \frac{\xi_1-\mu}{2\sigma^4} + \frac{(\xi_1-\mu)^3}{2\sigma^6} \\ \frac{\xi_1-\mu}{2\sigma^4} + \frac{(\xi_1-\mu)^3}{2\sigma^6} & \frac{((\xi_1-\mu)^2-\sigma^2)^2}{4\sigma^8} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

Таким чином, матриця розсіювання оцінок найбільшої вірогідності

$$\mathbf{V}_{\hat{\vartheta}^{MLE}}(\vartheta) = \mathbf{I}^{-1}(\vartheta) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}. \tag{8.16}$$

Ми отримали, що коефіцієнт розсіювання $\hat{\mu}_n^{MLE}$ дорівнює σ^2, а коефіцієнт розсіювання $\hat{\sigma}_n^{2\ MLE}$ дорівнює $2\sigma^4$. Ці оцінки є асимптотично некорельованими.

Знайдемо коефіцієнти розсіювання квантильних оцінок. Для $\hat{\mu}_n^{med}$ це можна зробити безпосередньо за формулою (8.14):

$$v_{\hat{\mu}^{med}} = \frac{1}{4(f_\vartheta(\mu))^2} = \frac{\pi\sigma^2}{2}.$$

[7] Тут $f_\vartheta(x)$ — щільність нормального розподілу з параметрами μ, σ^2, причому диференціюючи за σ^2 слід розуміти це як єдиний символ, а не як квадрат σ.

Для $\hat{\sigma}_n^{2\ IQ}$ міркування дещо складніші. Почнемо з визначення граничної коваріаційної матриці для вектора

$$\mathbf{z}_n = (z_n^1, z_n^2)^T = \sqrt{n}(\hat{\mathbf{q}}_n - \mathbf{q}),$$

де

$$\hat{\mathbf{q}}_n = (Q^{\mathbf{X}}(1/4), Q^{\mathbf{X}}(3/4))^T,$$
$$\mathbf{q} = (Q^{N(\mu,\sigma^2)}(1/4), Q^{N(\mu,\sigma^2)}(3/4))^T.$$

За (8.12) коваріаційна матриця розподілу двовимірного нормального вектора $\mathbf{z}$, до якого збігається розподіл $\mathbf{z}_n$, дорівнює

$$\mathbf{C} = (c_{ik})_{i,k=1}^2 \frac{1}{(f_\vartheta(\mu + \sigma\lambda_{1/4}))^2} \begin{pmatrix} \frac{3}{16} & \frac{1}{16} \\ \frac{1}{16} & \frac{3}{16} \end{pmatrix}.$$

Звідси отримуємо, що послідовність $\tilde{z}_n = (z_n^1 - z_n^2)/(2\lambda_{1/4})$ також є асимптотично нормальною з асимптотичною дисперсією

$$\tilde{c} = \frac{1}{(2\lambda_{1/4})^2}(c_{11} - 2c_{12} + c_{22}) = \frac{\pi e^{-\lambda_{1/4}^2}\sigma^2}{8\lambda_{1/4}^2}.$$

Оскільки при великих n:

$$\sqrt{n}(\hat{\sigma}_n^{2\ IQ} - \sigma^2) \sim 2\sigma\tilde{z},$$

то

$$v_{\hat{\sigma}^2\ IQ} = 4\sigma^2\tilde{c} = \frac{\pi e^{-\lambda_{1/4}^2}\sigma^4}{2\lambda_{1/4}^2} \approx 5.44184\sigma^4.$$

Таким чином, відносна асимптотична ефективність оцінки найбільшої вірогідності для μ порівняно з медіанною

$$v_{\hat{\mu}^{med}}/v_{\hat{\mu}^{MLE}} = \pi/2 \approx 1.5708.$$

Для ОНВ дисперсії порівняно з квартильною відносна асимптотична ефективність

$$v_{\hat{\sigma}^2\ IQ}/v_{\hat{\sigma}^2\ MLE} = \frac{\pi e^{-\lambda_{1/4}^2}}{4\lambda_{1/4}^2} \approx 2.72092.$$

Тобто при використанні квартильної оцінки потрібно у 2.72 раза більше спостережень, ніж при використанні звичайної вибірокової дисперсії для досягнення однакової точності оцінювання.

Зрозуміло, що вся ця асимптотична теорія працює лише при достатньо великих обсягах вибірки. Наскільки великих? Що буде для невеликих обсягів? Щоб відповісти на такі запитання, проводять спеціальні імітаційні експерименти (simulation study). Подивимось, як це зробити у нашому прикладі.

Ми згенеруємо `B=1000` різних вибірок з одним і тим самим нормальним розподілом з параметрами `mu=1` (математичне сподівання) і `sigma=1` (стандартне відхилення). По кожній вибірці будуть обчислені чотири оцінки, які вміщуються у масиви оцінок:

$\hat{\mu}_n^{MLE}$ у `EstMuMom`, $\hat{\mu}_n^{med}$ у `EstMuMed`, $\hat{\sigma}_n^{2\ MLE}$ у `EstSMom`, $\hat{\sigma}_n^{2\ IQ}$ у `EstSMed`.

По кожному з цих масивів ми обчислюємо вибіркове середнє, що має наближати математичне сподівання відповідної оцінки і віднімаємо від нього справжнє значення оцінюваного параметра. Отримуємо приблизне значення зміщення оцінки. Це значення домножається на $\sqrt{n}$. Асимптотичний розподіл нормованої оцінки має нульове математичне сподівання, тому можна сподіватись, що при досить великих n таке нормоване зміщення буде близьким до 0.

Далі ми обчислюємо вибіркові дисперсії по масивах оцінок і домножаємо на n. Ця величина має приблизно дорівнювати коефіцієнту розсіювання оцінки. Якщо це не так, можна запідозрити, що наші теоретичні розрахунки неадекватні або обсяг вибірки недостатньо великий для застосування асимптотичної теорії.

Наведемо скрипт, що реалізує цю ідею для обсягу вибірки `n=200`:

```
> set.seed(3)
> B<-1000  # кількість вибірок
> mu<-1    # математичне сподівання
> n<-200   # обсяг вибірки
> sigma<-1 # середнє квадратичне відхилення
```

```
> EstMuMom<-numeric(B)
> EstMuMed<-numeric(B)
> EstSMom<-numeric(B)
> EstSMed<-numeric(B)
> # Генеруємо вибірки і обчислюємо оцінки:
> for(i in 1:B)
+ {
+ x<-rnorm(n,mu,sigma)
+ EstMuMom[i]<-mean(x)
+ EstMuMed[i]<-median(x)
+ EstSMom[i]<-var(x)
+ EstSMed[i]<-(IQR(x)/1.34898)^2
+ }
> # Моментна оцінка мат. сподівання:
> (mean(EstMuMom)-mu)*sqrt(n) # зміщення

[1] -0.001903953

> n*var(EstMuMom)   # розсіювання

[1] 0.9389821

> sigma^2 # теоретичний коеф. розсіюввання

[1] 1

> # Медіанна оцінка для мат. сподівання:
> (mean(EstMuMed)-mu)*sqrt(n) # зміщення

[1] -0.005031941

> n*var(EstMuMed)              # розсіювання

[1] 1.459559

> 3.1415*sigma^2/2     # теоретичний коеф. розсіювання
```

```
[1] 1.57075

> # Моментна оцінка дисперсії:
> (mean(EstSMom)-sigma^2)*sqrt(n) # зміщення

[1] 0.05725523

> n*var(EstSMom)   # розсіювання

[1] 2.010686

> 2*sigma^4  # теоретичний коеф. розсіювання

[1] 2

> # Квартильна оцінка дисперсії
> (mean(EstSMed)-sigma^2)*sqrt(n) # зміщення

[1] 0.01135744

> n*var(EstSMed)  # розсіювання

[1] 5.561874

> 5.44184*sigma^4 # теоретичний коеф. розсіювання

[1] 5.44184
```

△

Як бачимо, при $n = 200$ результати імітаційного моделювання непогано (хоча і не ідеально) узгоджуються з асимптотичними формулами. Наприклад, теоретичний коефіцієнт розсіювання медіанної оцінки 1.57075, а його аналог, отриманий моделюванням — 1.459559. Нормоване зміщення дорівнює -0.005031941. Це дуже мало, порівняно з дисперсією, тому зміщенням як джерелом похибки можна знехтувати і характеризувати цю оцінку лише дисперсією. (Насправді медіанна оцінка у даному випадку незміщена,

тобто відхилення зміщення від 0 — це результат неточності нашого імітаційного експерименту).

Слід пам'ятати, що зі збіжності розподілів не випливає збіжність моментів. Тому навіть у асимптотично нормальних оцінок дисперсія нормованої оцінки при зростанні обсягу вибірки не обов'язково прямує до коефіцієнта розсіювання. Зокрема, так може бути, коли у маленькому відсотку випадків оцінка дозволяє грубі відхилення від справжнього значення параметра. Такі відхилення можуть відіграти роль викидів, що спотворюють дисперсію оцінки при скінченному обсязі вибірки.

У такому випадку доречно використати для наближення коефіцієнта розсіювання яку-небудь робастну оцінку дисперсії. Такою оцінкою може бути квантильна оцінка, якою ми щойно скористались у нашому прикладі. Крім того, доцільно перевірити, чи дійсно розподіл вибірки з оцінок добре узгоджується з нормальним. Розглянемо скрипт, що реалізує цю ідею:

```
> # Quantile estimate for dispersion:
> (sum(IQR(EstSMed))/1.34898)^2*n

[1] 5.04423

>  # histogram
>  hist(EstSMed,xlab=" ",ylab=" ",main=" ")
```

Як бачимо, значення цього робастного наближення — 5.04423, помітно відрізняється як від значення нормованої вибіркової дисперсії оцінок — 5.561874, так і від теоретичного коефіцієнта розсіювання — 5.44184 (вони, доречі, досить добре узгоджуються). Це може пояснюватись тим, що, внаслідок порівняно малого обсягу вибірки n, розподіл оцінок недостатньо добре наближається нормальним. І дійсно, гістограма оцінок на рис. 8.4 вказує на помітне відхилення від нормальності, зокрема, на асиметрію розподілу.

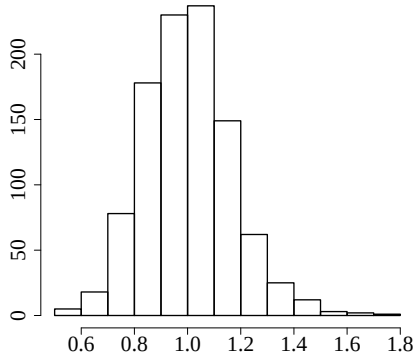

Рис. 8.4. Гістограма вибірки з оцінок

Можна зробити висновок, що у цій задачі оцінювання при таких невеликих n цілком покладатись на асимптотичні формули не варто. ◄

Приклад 8.4.3. Перейдемо до розгляду задачі оцінки параметрів нормального розподілу за спостереженнями з неоднорідними похибками з прикладу 8.1.3. У прикладі 8.3.3 ми розібрались як обчислювати оцінки найбільшої вірогідності для середнього та дисперсії у цій моделі. Чи можна скористатись нашою асимптотичною теорією, щоб охарактеризувати точність таких оцінок? Наприклад, нас може цікавити, наскільки вона погіршилась порівняно з випадком прикладу 8.4.2, коли дані спостерігались без похибок.

Безпосередньо формулу (8.11) для матриці розсіювання у даному випадку застосувати не можна, оскільки у цьому прикладі спостереження не є однаково розподіленими. Однак відомо, що асимптотична нормальність виконана і для ОНВ у багатьох моделях з неоднаковим розподілом спостережень (див. [3], підрозд. 66). Зокрема, за умови обмеженості дисперсій похибок, вона буде виконуватись у нашій моделі. В таких випадках матрицю розсіювання можна обчислювати використовуючи “середню інформацію

на одне спостереження", тобто

$$\bar{\mathbf{I}}_n(\vartheta) = \frac{1}{n}\sum_{j=1}^{n}\mathbf{I}_j(\vartheta),$$

де $\mathbf{I}_j(\vartheta)$ — інформаційна матриця для j-го спостереження. Матриця розсіювання ОНВ дорівнює

$$\mathbf{V}_{\hat{\vartheta}^{MLE}}(\vartheta) = \lim_{n\to\infty}(\bar{\mathbf{I}}_n(\vartheta))^{-1}.$$

Для практичних наближень замість границі при $n \to \infty$ беруть значення $(\bar{\mathbf{I}}_n(\vartheta))^{-1}$ при тому обсязі вибірки n, для якого робляться розрахунки. (Зрозуміло, що n має бути достатньо великим, інакше наша асимптотична теорія працювати не буде).

Аналогічно тому, як це було зроблено у прикладі 8.4.2, отримуємо, що інформаційна матриця для j-го спостереження (воно у нашій моделі має розподіл $N(\mu, \sigma^2 + \sigma_j^2)$), має вигляд

$$\mathbf{I}_j(\vartheta) = \begin{pmatrix} \frac{1}{\sigma^2+\sigma_j^2} & 0 \\ 0 & \frac{1}{2(\sigma^2+\sigma_j^2)^2} \end{pmatrix}.$$

(Тут, як і раніше, $\vartheta = (\mu, \sigma^2)$). Отже, коефіцієнти розсіювання оцінок дорівнюють

$$v_{\hat{\mu}^{MLE}}(\sigma^2) = \frac{n}{\sum_{j=1}^{n}(\sigma^2+\sigma_j^2)^{-1}},$$

$$v_{\hat{\sigma}^2\ MLE}(\sigma^2) = \frac{2n}{\sum_{j=1}^{n}(\sigma^2+\sigma_j^2)^{-2}},$$

◄

Приклад 8.4.4. Проведемо порівняння ефективності оцінок інтенсивності λ зрізаного експоненційного розподілу, отриманих у прикладах попередніх підрозділів. У прикладі 8.1.4 була побудована оцінка методу моментів (позначимо її $\hat{\lambda}^{MM}$), яка виявилась

також оцінкою найбільшої вірогідності у прикладі 8.3.4. У прикладі 8.2.4 введена медіанна оцінка (позначимо її $\hat{\lambda}^{med}$). Обидві оцінки обчислюються як розв'язки відповідних рівнянь, у явному вигляді їх виразити не можна. Тим більш цікаво, що їхні коефіцієнти розсіювання цілком можна знайти аналітично.

Почнемо з $\hat{\lambda}^{MM}$. Оскільки це оцінка найбільшої вірогідності, її коефіцієнт розсіювання можна знайти як $1/I(\lambda)$, де

$$I(\lambda) = \int_0^C \frac{(f'_\lambda(x))^2}{f_\lambda(x)} dx$$

— інформація за Фішером на одне спостереження,

$$f_\lambda(x) = \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda C}}$$

— щільність одного спостереження при $x \in [0, C]$,

$$f'_\lambda(x) = \frac{\partial}{\partial\lambda} f_\lambda(x) = \frac{e^{\lambda(C-x)}(\lambda(x-C) + e^{\lambda C}(\lambda x - 1) - 1)}{(e^{\lambda C} - 1)^2}.$$

Формулу для інформації можна отримати звичайним інтегруванням:

$$I(\lambda) = \frac{1}{\lambda^2} + \frac{C^2}{2 - (e^{\lambda C} + e^{-\lambda C})}.$$

Про всяк випадок перевіримо правильність цієї формули і покажемо, як можна наближено обчислювати інтеграли в `R` у тих випадках, коли для них немає явних виразів.

У `R` для наближеного обчислення інтегралів використовується функція

```
integrate(f,lower,upper),
```

де

`f` — дійсна функція дійсного аргументу, від якої підраховується інтеграл[8],

[8] Ця функція має бути векторизованою, тобто `R` має розуміти, як її застосовувати до масивів поелементно. Якщо `R` цього не розуміє — застосуйте векторизацію, як описано у п. 2.7.2.

`lower`, `upper` — нижня та верхня межі інтегрування.

Покажемо, як можна використати її для перевірки результату інтегрування у нашому випадку:

```
> l<-0.5 # інтенсивність
> U<-1   # поріг зрізання
> # щільність розподілу
> f<-function(x,l,U){l*exp(-l*x)/(1-exp(-l*U))}
> # похідна щільності розподілу за l
> fp<-function(x,l,U){exp(l*(U-x))*(l*(x-U)
+    +exp(l*U)*(-l*x+1)-1)/(exp(l*U)-1)^2}
> # підінтегральна функція для інформації
> g<-function(x){(fp(x,l,U))^2/f(x,l,U)}
> # аналітичний вираз для інформації
> inf<-function(l,U){1/l^2+
+      U^2/(2-(exp(U*l)+exp(-U*l)))}
> #
> inf(l,U) # інформація за формулою

[1] 0.08230191

> integrate(g,0,U) # наближений інтеграл для інформації

0.08230191 with absolute error < 9.1e-16
```

Як бачимо, значення інформації за нашою формулою та значення, отримане наближеним інтегруванням, однакові. Отже, при $\lambda = 0.5$, $C = 1$ коефіцієнт розсіювання моментної оцінки $v_{\hat{\lambda}_{MM}}(\lambda) = 1/I(\lambda) = 1/0.08230191 = 12.15039$.

Тепер обчислимо коефіцієнт розсіювання медіанної оцінки. Для цього скористаємось формулою (8.15). У нашому випадку $\alpha = 1/2$ медіана

$$q_\alpha = -\frac{1}{\lambda}(\log(1 + e^{-C\lambda}) - \ln 2),$$

$$\frac{\frac{\partial}{\partial x}F_\lambda(x)}{\frac{\partial}{\partial \lambda}F_\lambda(x)} = \frac{\lambda(e^{\lambda C}-1)}{C(1-e^{\lambda x})+x(e^{\lambda C}-1).}$$

Підставляючи це у формулу (8.15), отримуємо коефіцієнт розсіювання $\hat{\lambda}^{med}$:

$$v_{\hat{\lambda}^{med}}(\lambda) = \frac{\lambda^2(e^{\lambda C}-1)^2}{(\lambda C+(e^{\lambda C}+1)\log((e^{-\lambda C}+1)/2))^2}.$$

Підставляючи значення $\lambda = 0.5$, $C = 1$, отримуємо коефіцієнт розсіювання $v_{\hat{\lambda}^{med}}(\lambda) = 16.3344$.

Отже, при цих значеннях параметрів відносна асимптотична ефективність оцінки найбільшої вірогідності порівняно з медіанною становть $16.3344/12.15039 = 1.34435$. При використанні медіанних оцінок треба використовувати вибірки на 34% більші порівняно з вибірками для оцінок найбільшої вірогідності, якщо ми хочемо забезпечити однакову точність оцінювання.

На відміну від прикладів 8.4.1 і 8.4.2, відносна асимптотична ефективність (ARE) медіанної оцінки та ОНВ залежить тепер від невідомого параметра λ. Ми можемо подивитись на графіку, як вона змінюється при різних λ (рис. 8.5).

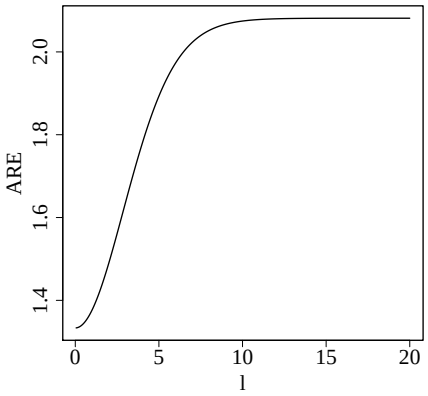

Рис. 8.5. Асимптотична відносна ефективність медіанної оцінки

Цей рисунок відображений таким скриптом:

```
> vmm<-function(l){1/inf(l,U)}
> # коефіцієнт розсіювання для медіанної оцінки
> vmed<-function(l){
+ (l*(exp(l*U)-1)/
+    (l*U+(1+exp(l*U))*log((1+exp(-l*U))/2)))^2
+ }
> l<-(1:400)/20
> ARE<-sapply(l,vmed)/sapply(l,vmm)
> plot(l,ARE,type="l")

> vmed(0.01)/vmm(0.01)

[1] 1.333337

> vmed(20)/vmm(20)

[1] 2.081368
```

Як бачимо, ARE зростає із зростанням λ від 1.333 до 2.081368. Тобто у найгіршому випадку (при дуже великих λ) медіанна оцінка вимагає вдвічі більше спостережень, ніж ОНВ для забезпечення еквівалентної точності оцінювання. Цікаво, що такий самий ефект ми отримали і для експоненційного розподілу без зрізання, тільки там така ARE була для всіх можливих значень λ. ◀

Досі ми розглядали лише оцінки, які є асимптотично нормальними. Але у деяких задачах оцінювання доцільно використовувати оцінки, що не мають цієї властивості. Зокрема, так часто буває при оцінюванні параметрів, що пов'язані з розривами щільності розподілу спостережень. Тут ми обмежимось лише одним прикладом такого оцінювання.

Приклад 8.4.5. Нехай спостереження являють собою кратну вибірку $\mathbf{X} = (\xi_1, \dots, \xi_n)$ з рівномірним розподілом на інтервалі $[a, b]$. Треба оцінити $\mu = \mathsf{E}\,\xi_1 = (b - a)/2$. Для цього можна

запропонувати три оцінки:

$$\hat{\mu}_n^{mean} = \frac{1}{n}\sum_{j=1}^{n} \xi_j$$

— вибіркове середнє,

$$\hat{\mu}_n^{med} = \mathrm{med}(\mathbf{X})$$

— вибіркова медіана,

$$\hat{\mu}_n^{MR} = \frac{1}{2}(\min(\mathbf{X}) + \max(\mathbf{X}))$$

— середина діапазону. (Це, до речі, класичні статистики середнього положення, які ми розглядали у підрозд. 4.1). Зрозуміло, що медіана буде робастною, а середина діапазону — найбільш чутливою до викидів оцінкою. А яка з цих оцінок є більш точною?

Середнє і медіана є асимптотично нормальними оцінками. Їхні коефіцієнти розсіювання дорівнюють

$$v^{mean} = \frac{b-a}{12},\ v^{med} = \frac{b-a}{4},$$

тобто вибіркове середнє у цій задачі оцінювання втричі ефективніше, ніж вибіркова медіана.

Але середина діапазону не є асимптотично нормальною, а нормування $\sqrt{n}$ не є відповідним для цієї оцінки:

$$\sqrt{n}(\hat{\mu}^{MR} - \mu) \to 0 \text{ за ймовірністю},$$

тобто $\hat{\mu}^{MR}$ збігається до μ швидше, ніж з порядком $1/\sqrt{n}$. Виявляється ([3]), що відповідним нормуванням у цьому випадку буде множення на n: послідовність $n(\hat{\mu}^{MR} - \mu)$ збігається за розподілом до деякого невиродженого розподілу.

На рис. 8.6 зображено графіки залежності дисперсій цих трьох оцінок від обсягу вибірки n (у логарифмічному масштабі по обох

осях). Графіки отримані в імітаційному експерименті, подібному до описаних вище. Дані генерувались з рівномірним розподілом на $[0, 1]$. Експеримент підтверджує, що середина діапазону має найменшу дисперсію при всіх обсягах вибірки.

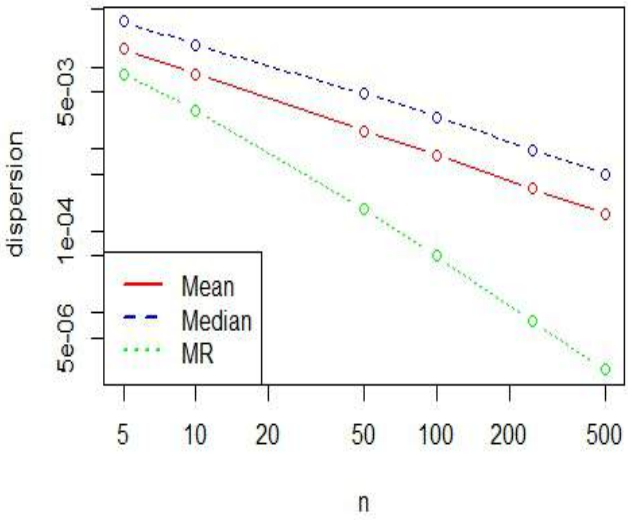

Рис. 8.6. Коефіцієнти розсіювання оцінок для математичного сподівання рівномірного розподілу в залежності від обсягу вибірки n. Логарифмічний масштаб по обох осях

Таким чином, у цьому прикладі середина діапазону виявляється значно точнішою оцінкою математичного сподівання, ніж вибіркове середнє або медіана. Це може бути важливим аргументом на користь його використання у відповідних прикладних задачах оцінювання, якщо робастність не є важливою. ◄

8.5 Довірчі інтервали та еліпсоїди

Зрозуміло, що оцінка $\hat{\vartheta}_n$, побудована за випадковими даними $\mathbf{X}$ майже ніколи не дорівнює справжньому значенню невідомого параметра ϑ. Як далеко може відхилитись оцінка від справжнього ϑ? Щоб охарактеризувати область можливих значень одновимірного параметра використовують техніку довірчих інтервалів. А саме, замість однієї оцінки $\hat{\vartheta}_n$ використовують пару статистик[9] ϑ_n^-, ϑ_n^+, таких, що $\vartheta_n^- < \vartheta_n^+$ і

$$\mathsf{P}\{\vartheta \in [\vartheta_n^-, \vartheta_n^+]\} = 1 - \alpha, \tag{8.17}$$

де α — задане статистиком мале число, яке називають рівнем значущості.

Таким чином, довірчий інтервал — це інтервал, побудований за спостережуваними даними, який покриває невідомий параметр із заданою ймовірністю $1 - \alpha$.

Якщо (8.17) виконується точно для заданого обсягу даних n і всіх $\vartheta \in \Theta \subseteq \mathbb{R}$, то $[\vartheta_n^-, \vartheta_n^+]$ називають точним (або строгим) довірчим інтервалом. Якщо рівність у (8.17) досягається лише асимптотично, тобто

$$\lim_{n\to\infty} \mathsf{P}\{\vartheta \in [\vartheta_n^-, \vartheta_n^+]\} = 1 - \alpha, \tag{8.18}$$

то довірчий інтервал називають асимптотичним. Нарешті, якщо виконується нерівність

$$\mathsf{P}\{\vartheta \in [\vartheta_n^-, \vartheta_n^+]\} \geq 1 - \alpha,$$

довірчий інтервал називають нестрогим.

8.5.1 Побудова асимптотичних довірчих інтервалів

Теорія асимптотичної нормальності оцінок дозволяє будувати асимптотичні довірчі інтервали з використанням коефіцієнтів роз-

[9] Тобто вимірних функцій від даних $\mathbf{X}$.

сіювання оцінок. Дійсно, нехай для невідомого параметра ϑ існує асимптотично нормальна оцінка $\hat{\vartheta}_n$ з коефіцієнтом розсіювання $v(\vartheta) = v_{\hat{\vartheta}_n}(\vartheta)$. Припустимо, що $v(\vartheta)$ є неперервною функцією $\vartheta \in \Theta$. Тоді $v(\vartheta)/v(\hat{\vartheta}_n) \to 1$ при $n \to \infty$ (за ймовірністю) і з (8.7) випливає, що[10]

$$\lim_{n\to\infty} \mathsf{P}\left\{ \frac{\sqrt{n}|\hat{\vartheta}_n - \vartheta|}{\sqrt{v(\hat{\vartheta}_n)}} \le \lambda_{\alpha/2} \right\} = 1 - \alpha. \tag{8.19}$$

Покладемо

$$\vartheta_n^{\pm} = \hat{\vartheta}_n \pm \lambda_{\alpha/2}\sqrt{\frac{v(\hat{\vartheta}_n)}{n}}. \tag{8.20}$$

Оскільки (8.19) еквівалентно

$$\lim_{n\to\infty} \mathsf{P}\{\vartheta \in [\vartheta_n^-, \vartheta_n^+]\} = 1 - \alpha,$$

то $[\vartheta_n^-, \vartheta_n^+]$ є асимптотичним довірчим інтервалом з рівнем значущості $1 - \alpha$.

Таку техніку можна використовувати і у тому випадку, коли крім параметра ϑ у розподілу даних є й інші невідомі параметри. Для побудови довірчого інтервалу потрібна лише асимптотично нормальна оцінка ϑ та консистентна оцінка його коефіцієнта розсіювання. Якщо вони визначені, то довірчий інтервал можна будувати за формулою (8.20).

Зрозуміло, що чим менше коефіцієнт розсіювання оцінки, тим вужчим буде довірчий інтервал, побудований за цією технікою. Тому при побудові довірчих інтервалів природно обирати оцінки з найменшим коефіцієнтом розсіювання, якщо немає інших важливих вимог (як робастність або еківаріантність).

Приклад 8.5.1. Знову розглянемо задачу оцінки інтенсивності експоненційного розподілу λ за кратною вибіркою $\mathbf{X}$. У прикладі 8.1.1, оцінкою з найменшим коефіцієнтом розсіювання для λ виявилась $1/\bar{\xi}$, причому її коефіцієнт розсіювання $v(\lambda) = \lambda^2$.

[10] Тут, як і раніше, $\lambda_\alpha = Q^{N(0.1)}(1 - \alpha)$.

Нехай треба побудувати довірчий інтервал для λ з рівнем значущості $\alpha = 0.05$. Зазначимо, що[11]

$$\lambda_{\alpha/2} = Q^{N(0,1)}(0.975) \approx 1.96.$$

Отже, за (8.20) отримуємо межі довірчого інтервалу:

$$\lambda^- = \frac{1}{\bar{\xi}} - \frac{1.96}{\bar{\xi}\sqrt{n}},\ \lambda^+ = \frac{1}{\bar{\xi}} + \frac{1.96}{\bar{\xi}\sqrt{n}}.$$

Рівень значущості довірчого інтервалу 0.05 означає, що в середньому на 100 задач оцінювання у 95 випадках такий довірчий інтервал покриє справжнє значення інтенсивності. Довірчий інтервал асимптотичний, отже, це має виконуватись "при достатньо великих обсягах вибірки". Спробуємо в імітаційному експерименті подивитись, наскільки точним виявиться це передбачення для вибірок помірного обсягу:

```
> set.seed(2)
> l<-0.5      # інтенсивність експ. розподілу
> B<-10000    # кількість вибірок
> n<-100      # обсяг вибірки
> lambda<-qnorm(0.975)
> set.seed(2)
> l<-0.5      # інтенсивність експ. розподілу
> B<-10000    # кількість вибірок
```

[11] Існує певна незручність в тому, що літера λ позначає і невідомий параметр, і квантиль нормального розподілу. У цих позначеннях формули для довірчого інтервалу будуть незрозумілими. Можна було б спеціально для цього прикладу ввести якесь особливе позначення для інтенсивності або для квантилі. Але обидва позначення є стандартними і відступ від них також заплутав би справу. Я вийшов з положення, зафіксувавши $\alpha = 0.05$. Це дуже популярний рівень значущості і відповідне йому $\lambda_{\alpha/2} = 1.96$ більшість статистиків знає напам'ять. У багатьох прикладних книгах число 1.96 з'являється без пояснень, як магічна константа. Його зв'язок з рівнем значущості залишається для користувачів таємницею. Довірчий інтервал, який ми отримаємо, може бути прикладом таких формул.

```
> n<-100      # обсяг вибірки
> lambda<-qnorm(0.975)
> #res - масив результатів випробувань
> res<-replicate(B,
+ {
+ x<-rexp(n,l)
+ estl<-1/mean(x)
+ ifelse(abs(l-estl)<lambda*estl/sqrt(n),1,0)
+ })
> err=1-mean(res) # частота помилок
> err

[1] 0.0474
```

△

Тут вибрано $\lambda = 0.5$, моделюється $B = 10000$ вибірок і по кожній вибірці перевіряється, чи потрапить 0.5 у довірчий інтервал (точніше, чи є різниця $\hat{\lambda} - \lambda$ меншою, ніж половина ширини інтервалу). Якщо для i-ї модельованої вибірки ця умова виконана, на i-му місці у масиві результатів `res` записується 1, інакше - 0. Потім частота потраплянь визначається як середнє `res`.

Як бачимо, результат експерименту показує помірне узгодження з теорією — частота 0.0474 при теоретичній ймовірності 0.05. При обсязі вибірки $n = 1000$ цей самий скрипт дасть 0.0505 — чудова узгодженість.

Цікаво, що у даному прикладі можна побудувати точний довірчий інтервал, використовуючи той факт, що випадкова величина $S = \sum_{j=1}^{n} \xi_j$ має розподіл $\Gamma(n, \lambda)$ (див. п. 6.2.4). Отже, $\lambda S \sim \Gamma(n, 1)$. Тому

$$\mathsf{P}\{Q^{\Gamma(n,1)}(\alpha/2) \le \lambda S \le Q^{\Gamma(n,1)}(1-\alpha/2)\} = 1 - \alpha.$$

Звідси

$$\mathsf{P}\{\lambda^- \le \lambda \le \lambda^+\} = 1 - \alpha,$$

де

$$\lambda^- = Q^{\Gamma(n,1)}(\alpha/2)/S,\ \lambda^+ = Q^{\Gamma(n,1)}(1-\alpha/2)/S.$$

Перевірка цього довірчого інтервалу на даних, згенерованих як у попередньому скрипті, дає для $n = 100$ частоту помилок 0.0494. Тобто при цьому обсязі вибірки точний довірчий інтервал виявляється справді помітно точнішим, ніж розглянутий вище асимптотичний. При $n = 1000$ маємо частоту помилок 0.0501, тобто при такому обсязі вибірки відмінність від асимптотичного інтервалу вже несуттєва. ◀

Звичайно, на практиці, там, де є можливість знайти точні довірчі інтервали, краще використовувати саме їх.

Приклад 8.5.2. Розглянемо задачу аналізу якості жорстких комп'ютерних дисків з прикладів 8.1.5 — 8.3.5. Для інтенсивностей утворення дефектів λ кожного виробника можна за даними з прикладу 8.1.5 побудувати асимптотичні довірчі інтервали, використовуючи стандартну схему. Дійсно, за (8.9) отримуємо, що коефіцієнт розсіювання моментної оцінки, визначеної у прикладі 8.1.5, дорівнює

$$v_{\hat\lambda}(\lambda) = \frac{e^{-\lambda}(1-e^{\lambda})^2}{e^{\lambda}-\lambda-1}.$$

Отже, реалізувати підрахунок значень лівої та правої меж довірчого інтервалу можна таким чином:

```
> alpha=0.05 # рівень значущості довірчого інтервалу
> # Коефіцієнт розсіювання
> # як функція від інтенсивності l:
> VarTh<-function(l)(
+   exp(-l)*(1-exp(l))^2)*l/(exp(l)-1-l
+                    )
> # CIPoisZ рахує межі довірчого інтервалу
> # для інтенсивності
> # x - значення варіант у вибірці
> # w - частота варіант
> # результат: набір (li, est, ui), де
> # li - ліва межа, ui - права межа інтервалу,
> # est - точкова оцінка.
```

```
> CIPoisZ<-function(x,w,alpha=0.05){
+ lambda<-qnorm(1-alpha/2)
+ EstL<-EstP(x,w)
+ dif<-lambda*sqrt(VarTh(EstL)/sum(w))
+ c(EstL-dif,EstL,EstL+dif)
+ }
```

△

(Функція `EstP()` описана у прикладі 8.1.5). Значення меж довірчих інтервалів та точкові оцінки можна тепер розрахувати таким чином:

▽

```
>  # вводимо дані:
>  A<-c(20,13,11,6,2,0)
>  B<-c(25,20,15,7,0,1)
>  C<-c(33,16,4,1,0,0)
>  x<-1:6
> ci<-rbind(CIPoisZ(x,A),CIPoisZ(x,B),CIPoisZ(x,C))
> colnames(ci)<-c("li","est","ui")
> rownames(ci)<-c("A","B","C")
> ci

         li       est       ui
A 1.4044474 1.8215158 2.238584
B 1.3899723 1.7494543 2.108936
C 0.5630007 0.8742175 1.185434
```

△

Отже, наприклад, для λ дисків виробника B довірчий інтервал має вигляд $[1.3899723, 2.108936]$.

Для того, щоб відобразити всі три інтервали на одному рисунку, можна скористатись функцією `plotCI()` з бібліотеки `plotrix`. Основні параметри цієї функції:

`x`, `y` — координати точок, що будуть відповідати точковим оцінкам, навколо яких будується довірчий інтервал (основна точка);

`uiw` (`liv`) — відстані від основних точок до нижньої (верхньої) границі довірчого інтервалу. (Якщо вказано лише `uiw`, то `liv=uiw`, тобто основна точка розташована посередині інтервалу);

Альтернативний варіант задання границь:

`ui` (`li`) — абсолютні координати нижніх (верхніх) границь інтервалу. (Ці параметри використовуються лише тоді, коли `uiw` і `liv` не задані);

`add` — логічний параметр, що показує, чи треба додавати інтервали на попередньому рисунку (`add=T`), чи рисувати на тому, який вже є (`add=F`).

Застосуємо `plotCI()` для відображення довірчих інтервалів для інтенсивностей у нашому прикладі:

```
> library(plotrix)
> plotCI(1:3,y=ci[,2],ui=ci[,3],li=ci[,1],
+ xlab=" ",ylab="intensities",xlim=c(1,3.2),xaxt="n")
> axis(1,at=1:3,labels=c("A","B","C"))
```

(Опція `xaxt="n"` відміняє рисування стандартної горизонтальної осі, а функція `axis` рисує нестандартну вісь, на якій відмічені назви виробників замість координат по горизонталі).

Результат роботи скрипта зображено на рис. 8.7 Як ми бачимо, довірчий інтервал для інтенсивності, отриманий за даними про диски виробника C, не перетинається з довірчими інтервалами для інших виробників (тобто на рисунку не можна провести горизонтальну лінію, яка проходила б через цей інтервал та який-небудь з інтервалів для A чи B). Отже, немає такого значення λ, яке могло б бути спільним для дисків виробництва фірми C та дисків A (або B). Для фірм A і B існує багато кандидатів на роль спільного значення λ — це може бути будь-яке число, що належить перетину відповідних інтервалів. Тому можна зробити висновок, що інтенсивності для A і B — однакові і відрізняються від інтенсивності C.

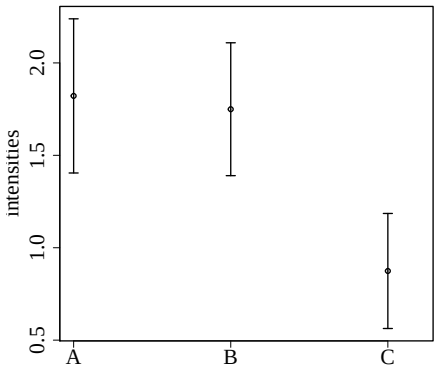

Рис. 8.7. Довірчі інтервали для інтенсивностей утворення дефектів на жорстких дисках

◄

Насправді запропонована техніка є занадто грубою, хоча її досить часто використовують. Ми повернемось до цього питання у розділі про перевірку гіпотез (приклад 9.3.3).

8.5.2 Побудова довірчих еліпсоїдів

Якщо модель розподілу має кілька (d) невідомих параметрів, то можна будувати довірчі інтервали для кожного параметра окремо, використовуючи техніку, описану вище. Але можливий і інший підхід — набір невідомих параметрів $\vartheta = (\vartheta_1, \ldots, \vartheta_d)^T \in \Theta \subseteq \mathbb{R}^d$ можна розглядати як точку у d-вимірному просторі і поставити задачу побудови підмножини цього простору, яка покриває ϑ із заданою ймовірністю. Точніше, довірчою множиною для ϑ із заданим рівнем значущості α називають таку випадкову (побудовану за спостережуваними данними $\mathbf{X}_n$) множину $\hat{\Theta} = \hat{\Theta}(\mathbf{X}_n)$, для якої

$$\mathsf{P}\{\vartheta \in \hat{\Theta}\} = 1 - \alpha.$$

Відповідно, асимптотичною довірчою множиною називають множину $\hat{\Theta}(\mathbf{X}_n)$, для якої

$$\lim_{n\to\infty} \mathsf{P}\{\vartheta \in \hat{\Theta}(\mathbf{X}_n)\} = 1 - \alpha.$$

Взагалі кажучи, для різних задач статистичного аналізу можуть бути корисними довірчі множини різного вигляду. Ми обмежимось розглядом техніки побудови асимптотичних довірчих еліпсоїдів на основі асимптотично нормальних оцінок.

Отже, нехай для невідомого (d-вимірного) параметра ϑ за даними $\mathbf{X}_n$ можна побудувати асимптотично нормальну оцінку $\hat{\vartheta}_n$, яка має матрицю розсіювання $\mathbf{V}$. Припустимо, що ця матриця невироджена і для неї є консистентна оцінка $\hat{\mathbf{V}}_n$.

Для довільного $\mathbf{t} \in \mathbb{R}^d$ розглянемо випадкову величину

$$S_n(\mathbf{t}) = n(\mathbf{t} - \hat{\vartheta}_n)^T \hat{\mathbf{V}}_n^{-1} (\mathbf{t} - \hat{\vartheta}_n).$$

Можна довести, що розподіл $S_n(\vartheta)$ прямує до розподілу χ^2_d при $n \to \infty$[12].

Для рівня значущості α покладемо $h^d_\alpha = Q^{\chi^2_d}(1-\alpha)$,

$$\hat{\Theta}_n = \{\mathbf{t} \in \mathbb{R}^d \ : \ S_n(\mathbf{t}) \le h^d_\alpha\}.$$

Тоді

$$\lim_{n\to\infty} \mathsf{P}\{\vartheta \in \hat{\Theta}_n\} = \lim_{n\to\infty} \mathsf{P}\{S_n(\vartheta) \le h^d_\alpha\} = 1 - \alpha,$$

тобто $\hat{\Theta}_n$ є асимптотичною довірчою множиною для ϑ рівня α.

При достатньо великих n, рівняння

$$S_n(\mathbf{t}) = h^d_\alpha \tag{8.21}$$

[12] Дійсно, з консистентності $\hat{\mathbf{V}}_n$ та асимптотичної нормальності $\hat{\vartheta}_n$ випливає збіжність за розподілом $\zeta_n = \sqrt{n}\hat{\mathbf{V}}_n^{-1/2}(\hat{\vartheta}_n - \vartheta) \to N(0, \mathbb{E})$ при $n \to \infty$, де $\mathbb{E}$ — одинична $d \times d$-матриця. Оскільки $S_n(\vartheta) = \|\zeta_n\|^2$, отримуємо потрібну збіжність.

відносно $\mathbf{t}$ задає у $\mathbb{R}^d$ поверхню другого порядку, яка є еліпсоїдом з центром у $\hat{\vartheta}_n$ (оскільки матриця $\mathbf{V}$ додатно-визначена, а $\hat{\mathbf{V}}_n \to \mathbf{V}$). $\hat{\Theta}_n$ — тіло, обмежене цією поверхнею. Тому цю множину називають (асиптотичним) довірчим еліпсоїдом для ϑ.

Приклад 8.5.3. Розглянемо кратну вибірку $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ з гауссовим розподілом: $\xi_j \sim N(\mu, \sigma^2)$. Нехай обидва параметри μ та σ є невідомими. У прикладах 8.1.2 і 8.3.2 показано, що метод моментів та метод найбільшої вірогідності тут дають однакові оцінки:

$$\hat{\mu}_n = \bar{\xi} = \frac{1}{n}\sum_{j=1}^{n} \xi_j, \ \hat{\sigma}_n = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\xi_j - \bar{\xi})^2},$$

тобто вибіркове середнє та вибіркове середньоквадратичне відхилення.

Як видно з прикладу 8.4.2, вектор оцінок $\hat{\vartheta}_n^{MLE} = (\hat{\mu}_n, \hat{\sigma}_n^2)^T$ є асимптотично нормальним як оцінка для $\vartheta = (\mu, \sigma^2)^T$ з матрицею розсіювання, що задається (8.16).

Тепер побудуємо довірчі інтервали для μ, σ та довірчий еліпсоїд для $\theta = (\mu, \sigma)^T$ на основі оцінок $\hat{\theta}_n = (\hat{\mu}_n, \hat{\sigma}_n^2)^T$. (Зверніть увагу, що еліпсоїд будується саме для σ, а не σ^2). Використовуючи (8.16), бачимо, що $\hat{\theta}_n$ також є асимптотично нормальною з матрицею розсіювання

$$\mathbf{V} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2/2 \end{pmatrix}.$$

Як оцінку цієї матриці за даними природно використати

$$\hat{\mathbf{V}}_n = \begin{pmatrix} \hat{\sigma}_n^2 & 0 \\ 0 & \hat{\sigma}_n^2/2 \end{pmatrix}.$$

Таким чином, можемо написати асимптотичні довірчі інтервали окремо для параметрів

$$\mathsf{P}\{\mu \in [\mu_n^-(\alpha), \mu_n^+(\alpha)]\} \to 1 - \alpha, \text{ при } n \to \infty,$$

де

$$\mu_n^{\pm}(\alpha) = \hat{\mu}_n \pm \frac{\lambda_{\alpha/2}\hat{\sigma}_n}{\sqrt{n}}.$$

Зазвичай на практиці використовують більш точний, неасимптотичний інтервал для μ, враховуючи те, що

$$\frac{\hat{\mu}_n - \mu}{\sqrt{n}S_0(\mathbf{X})}$$

має T-розподіл Стьюдента з $n-1$ ступенем вільності. Тому, поклавши

$$t_{\alpha/2}^{n-1} = Q^{T_{n-1}}(1-\alpha),\ \hat{\mu}_n^{\pm}(\alpha) = \hat{\mu}_n \pm \frac{t_{\alpha/2}^{n-1}S_0(\mathbf{X})}{\sqrt{n}},$$

отримуємо

$$\mathsf{P}\{\mu \in [\hat{\mu}_n^-, \hat{\mu}_n^+]\} = 1-\alpha.$$

Отже, межі точного довірчого інтервалу відрізняються від асимптотичних тим, що у них замість нормальних квантилів використовуються квантилі T-розподілу, а вибіркова дисперсія береться виправлена. При $n > 100$ відмінність між цими інтервалами не є суттєвою, наприклад $t_{0.025}^{99} \approx 1.98, \lambda_{0.025} \approx 1.96$. Якщо довірчі інтервали використовуються для графічного зображення результатів статистичного аналізу, така відмінність у третьому знаку майже не буде помітною.

Асимптотичний довірчий інтервал для σ матиме вигляд

$$\mathsf{P}\{\sigma \in [\sigma_n^-, \sigma_n^+]\} \to 1-\alpha, \text{ при } n \to \infty,$$

де

$$\sigma_n^{\pm} = \hat{\sigma}_n \pm \frac{\lambda_{\alpha/2}\hat{\sigma}_n}{\sqrt{2n}}.$$

Для σ також можна написати точний довірчий інтервал, використовуючи той факт, шо

$$\zeta = \frac{n\sigma_n^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Дійсно, поклавши $h^- = Q^{\chi^2_{n-1}}(\alpha/2)$, $h^+ = Q^{\chi^2_{n-1}}(1-\alpha/2)$, отримуємо

$$\mathsf{P}\{h^- \le \zeta \le h^+\} = 1 - \alpha.$$

Підставляючи сюди явний вираз для ζ і розв'язуючи отримані нерівності відносно σ, отримуємо

$$\mathsf{P}\{\sigma \in [\hat\sigma^-, \hat\sigma^+]\} = 1 - \alpha,$$

де

$$\hat\sigma^{\pm} = \sqrt{\frac{n\hat\sigma^2}{h^{\mp}}}.$$

Подивимось, як ці (та й інші) довірчі інтервали можна відображати, використовуючи R.

У наступному прикладі ми спочатку створюємо три вибірки x, y, z з нормальним розподілом, а потім будуємо довірчі інтервали для їхнього математичного сподівання та стандартного відхилення:

```
> library(plotrix)
> set.seed(3)
> n<-100       # обсяг вибірки
> alpha<-0.05  # рівень значущості
> # квантилі для визначення меж інтервалів
> lambda<-qnorm(1-alpha/2)
> tq<-qt(1-alpha/2,n-1)
> hm<-qchisq(alpha/2,n-1)
> hp<-qchisq(1-alpha/2,n-1)
> # генерація трьох вибірок
> x<-rnorm(n,0.72,1)
> y<-rnorm(n,0.72,1)
> z<-rnorm(n,1.1,0.9)
> D<-cbind(x,y,z)
> # оцінки математичних сподівань
> EstMu<-apply(D,2,mean)
```

```
> # оцінки дисперсій
> EstS<-apply(D,2,sd)
> # межі асимптотичного інтервалу для
> # математичних сподівань:
> MuPlusN<-EstMu+EstS*lambda*sqrt(n-1)/n
> MuMinusN<-EstMu-EstS*lambda*sqrt(n-1)/n
> # межі точного інтервалу для
> # математичних сподівань:
> MuPlusT<-EstMu+EstS*tq/sqrt(n)
> MuMinusT<-EstMu-EstS*lambda/sqrt(n)
> # рисуємо асимптотичні інтервали для
> # математичних сподівань:
> plotCI(1:3,y=EstMu,ui=MuPlusN,li=MuMinusN,
+     col="red",xlab=" ",ylab="means",slty="dashed",
+     xlim=c(1,3.2),xaxt="n")
> # рисуємо точні інтервали для
> # математичних сподівань:
> plotCI((1:3)+0.2,y=EstMu,ui=MuPlusT,
+     li=MuMinusT,col="blue",add=T)
> # рисуємо горизонтальну вісь координат з
> # назвами вибірок, за якими побудовані інтервали
> axis(1,at=(1:3)+0.1,labels=c("X","Y","Z"))
> # аналогічно будуємо інтервали
> # для стандартних відхилень:
> SPlusN<-EstS+EstS*lambda*sqrt(n-1/2)/n
> SMinusN<-EstS-EstS*lambda*sqrt(n-1/2)/n
> SPlusC<-sqrt((n-1)/hm)*EstS
> SMinusC<-sqrt((n-1)/hp)*EstS
> plotCI(1:3,y=EstS,ui=SPlusN,
+     li=SMinusN,col="red",slty="dashed",
+     xlab=" ",ylab="standard deviations",
+     xlim=c(1,3.2),xaxt="n")
> plotCI((1:3)+0.2,y=EstS,ui=SPlusC,
+     li=SMinusC,col="blue",add=T)
```

```
> axis(1,at=(1:3)+0.1,labels=c("X","Y","Z"))
```

△

Отримані довірчі інтервали можна побачити на рис. 8.8 (зверху — для математичних сподівань, знизу — для стандартних відхилень). Червоним кольором (штриховою лінією) відображені асимптотичні довірчі інтервали, синім (суцільною лінією) — точні.

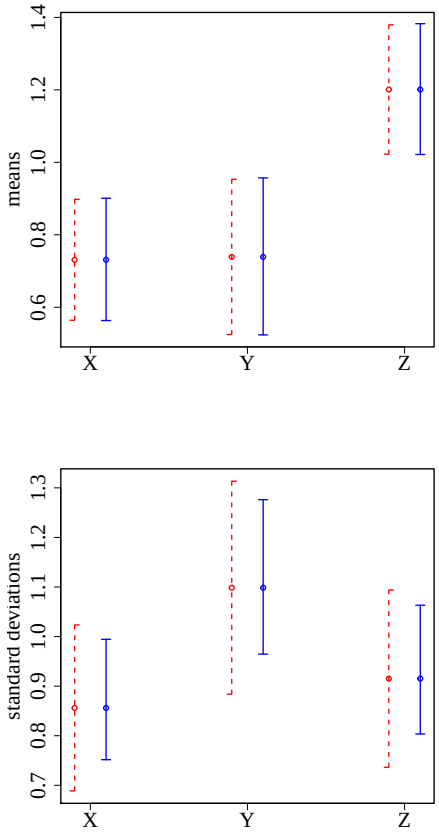

Рис. 8.8. Довірчі інтервали для середніх та стандартних відхилень

Як бачимо, при обсязі вибірки $n = 100$ різниця між точним і асимптотичним довірчими інтервалами для математичних сподівань майже не помітна. Для стандартних відхилень точні інтервали трохи вужчі, ніж асимптотичні.

На рис. 8.8 зверху бачимо, що довірчі інтервали для математичних сподівань (як асимптотичні, так і точні) вибірок X та Y перекриваються: є спільне значення, яке належить обом інтервалам. (На рисунку інтервали рознесені по горизонталі, тому це означає, що є горизонтальна лінія, яка перетинає обидва інтервали). Інтервал для математичного сподівання Z лежить значно вище і не має спільних точок з двома попередніми. Це можна трактувати, як підтвердження припущення про те, що математичні сподівання розподілів вибірок X і Y — однакові та відрізняються від математичного сподівання Z. (Саме так і були згенеровані ці вибірки).

На рис. 8.8 знизу всі довірчі інтервали для стандартних відхилень мають непорожній перетин — точка $\sigma = 0.98$ належить усім інтервалам. Отже, можна припустити, що стандартні відхилення для всіх трьох розподілів однакові. Це припущення хибне, у X та Y $\sigma = 1$, у Z — $\sigma = 0.9$. Можна сказати, що ця відмінність виявилась занадто малою, щоб бути поміченою за допомогою наших довірчих інтервалів.

Така техніка використання довірчих інтервалів для порівняння параметрів різних вибірок і перевірки припущень про них є дуже поширеним. С точки зору строгої теорії, у ній є певний недолік, яка обговорюється і виправляється п. 9.3.3. Але як спосіб грубої попередньої оцінки ця техніка допустима. Якщо треба одночасно проаналізувати відмінності обох параметрів — μ та σ для різних вибірок, природно побудувати для них довірчі еліпси (еліпсоїди у двовимірному просторі). У даному випадку рівняння (8.21) відносно $\mathbf{t} = (\mu, \sigma)$ набуває вигляду

$$\frac{n}{h_\alpha^2 \hat{\sigma}_n^2}(\mu - \hat{\mu}_n)^2 + \frac{2n}{h_\alpha^2 \hat{\sigma}_n^2}(\sigma - \hat{\sigma}_n)^2 = 1,$$

де $h_\alpha^2 = Q^{\chi_2^2}(1-\alpha)$. Це рівняння еліпса з осями, паралельними осям координат, причому горизонтальна піввісь (відповідає параметру μ):

$$a = \hat{\sigma}_n \sqrt{\frac{h_\alpha^2}{n}},$$

а вертикальна (за параметром σ) —

$$b = \hat{\sigma}_n \sqrt{\frac{h_\alpha^2}{2n}}.$$

Центр еліпса розміщений у точці $(\hat{\mu}_n, \hat{\sigma}_n)$.

Для відображення цього еліпса скористаємось функцією `draw.ellipse()` з бібліотеки `plotrix`. Основні параметри цієї функції:

`x,y` — координати центра еліпса;

`a,b` — довжини півосей;

`angle` — кут повороту еліпса відносно системи координат.

У цій функції можна також використовувати звичайні опції функцій для відображення геометричних об'єктів, як, наприклад, `col` (колір), `density` (щільність штриховки). Ці опції функція `draw.ellipse()` передає без змін функції `polygon()`[13], яка використовується при зображення еліпса.

У поданому нижче скрипті зображення одного довірчого еліпса оформлено у вигляді окремої функції `ConfEllipse()`[14]. Використовуються вибірки `x, y, z`, згенеровані у попередньому скрипті.

```
> ConfEllipse<-function(x,alpha=0.05,label=NULL,...)
+ {
+ n<-length(x)
```

[13] Зображення багатокутника за координатами його вершин.

[14] Ця функція призначена для побудови довірчого еліпса для параметрів нормального розподілу. Більш загальна функція `PlotEllipse()` для зображення довірчих еліпсів за оцінками та їхніми (оціненими) коваріаційними матрицями описана у підрозд. 8.6.

```
+ EstMuMom<-mean(x)
+ EstSMom<-sd(x)
+ q<-qchisq(1-alpha,2)
+ a<-EstSMom*sqrt(q/n)
+ b<-EstSMom*sqrt(q/(2*n))
+ draw.ellipse(EstMuMom,EstSMom,a,b,0,...)
+ text(x=EstMuMom,y=EstSMom,labels=label,cex=2)
+ }
> plot(c(0.5,1.5),c(0.5,1.5),
+    type="n",xlab="mean",ylab="sd")
> ConfEllipse(x,label="X",col="red",density=15)
> ConfEllipse(y,label="Y",col="green",density=20)
> ConfEllipse(z,label="Z",col="blue",density=10)
```

△

Отримані еліпси можна побачити на рис. 8.9. Оскільки довірчі еліпси для вибірок x та y перетинаються, то можна зробити висновок, що параметри нормального розподілу цих вибірок однакові. Еліпс для z не перетинається з іншими, отже, розподіл цієї вибірки відрізняється від розподілу двох інших. ◄

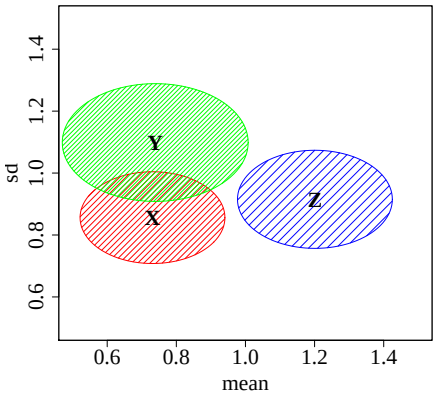

Рис. 8.9. Довірчі еліпсоїди для середніх та стандартних відхилень

8.6 Оцінювання для стандартних розподілів

Приклади з підрозд. 8.1 — 8.5 показують, як можна отримувати оцінки невідомих параметрів у випадку досить складних моделей розподілу даних. Якщо підганяється порівняно проста стандартна модель (напр., експоненційного або пуассонового розподілу), то писати власну програму оцінювання не має сенсу. Можна скористатись стандартними засобами системи R.

Для підгонки стандартних моделей можна скористатись функцією `fitdistr()` з бібліотеки `MASS`. Ця функція має такі параметри:

`x` — вектор з даними (вибірка);

`densfun` — тут можна вказати вид розподілу, який підганяється: `beta`, `cauchy`, `chi-squared`, `exponential`, `f`, `gamma`, `geometric`, `log-normal` (синонім: `lognormal`), `logistic`, `normal`, `negative binomial`, `Poisson`, `t` або `weibull`. У цьому параметрі можна також явно задати щільність теоретичного розподілу як функцію від спостереження та невідомих параметрів;

`start` — початкові значення для пошуку оцінки найбільшої вірогідності чисельними методами.

Якщо оцінку методу найбільшої вірогідності для даної моделі розподілу можна обчислити в явному вигляді (як для експоненційного або пуассонового розподілу), то `fitdistr()` використовує явні формули. Якщо це неможливо, проводиться чисельна максимізація функції вірогідності за допомогою функції `optim()`. Для деяких розподілів (гамма, логістичного та ін.) початкові значення для пошуку точки максимуму можна задати з евристичних міркувань і `fitdistr()` це робить, якщо параметр `start` не вказаний. Для інших розподілів вказувати `start` необхідно. Початкові значення слід задавати у вигляді іменованого списку.

Результатом роботи функції є об'єкт з атрибутами:

`estimate` — вектор оцінок невідомих параметрів;

`sd` — оцінки для середньоквадратичних відхилень цих оцінок;

`vcov` — оцінка для коваріаційної матриці вектора оцінок;

`loglik` — логарифм відношення вірогідності у точці максимуму.

`sd` та `vcov` розраховуються на основі оцінок для матриці розсіювання (див. підрозд. 8.4). Їх можна використовувати для побудови асимптотичних довірчих інтервалів та еліпосїдів. Значення `loglik` може бути корисним для перевірки гіпотез про невідомі параметри (див. підрозд. 9.3, приклад 9.3.2).

Функції `fitdistr()` можна задавати і інші параметри. Вони будуть передаватись функції `optim()`, яка шукає максимум функції вірогідності, а через неї можуть потрапити й у саму функцію віроігдності. Наприклад, якщо серед параметрів розподілу є відомі, їхні значення можна передати у вигляді додаткових опцій у `fitdistr()`. Тоді ця функція шукатиме оцінки лише для невідомих параметрів.

Приклад 8.6.1. Продемонструємо роботу функції `fitdistr()` при обчисленні оцінок для параметрів гамма-розподілу. У скрипті спочатку генерується вибірка `x` обсягу 500 з параметрами `shape=5`, `rate=0.1`. Потім за цією вибіркою проводиться оцінювання:

```
> library(MASS)
> n<-500
> s<-5
> r<-0.1
> set.seed(123)
> x <- rgamma(n, shape = s, rate = r)
> fitted<-fitdistr(x, "gamma")
> fitted

      shape          rate
  5.158923002   0.106554667
 (0.316202041) (0.006859197)
```

Отримано досить точну оцінку `shape`≈ 5.1589 і `rate`≈ 0.1065. В останньому рядку в дужках вказані відповідні оцінки середньоквадратичних відхилень. Їх можна використати для побудови довірчих

інтервалів з урахуванням асимптотичної нормальності оцінок параметрів. А саме, якщо $\hat{\vartheta}$ — оцінка для параметра ϑ, а $\hat{\sigma}$ — оцінка для дисперсії $\hat{\vartheta}$, то межі довірчого інтервалу з рівнем значущості α можна визначити як

$$\vartheta^- = \hat{\vartheta} - \lambda_{\alpha/2}\hat{\sigma},\ \vartheta^+ = \hat{\vartheta} + \lambda_{\alpha/2}\hat{\sigma}.$$

Таким чином, продовжуючи цей скрипт, ліву і праву межі 95%-довірчого інтервалу для `shape` можна знайти як:

```
> alpha<-0.05 # рівень значущості
> lambda<-qnorm(1-alpha/2)
> Est<-fitted$estimate["shape"] # значення оцінки
> # оцінка середньоквадратичного відхилення:
> sdEst<-fitted$sd["shape"]
> # межі довірчого інтервалу:
> Est-lambda*sdEst # ліва:

   shape
4.539178

> Est+lambda*sdEst # права:

   shape
5.778668
```

Як бачимо, у цьому прикладі справжнє значення параметра потрапляє до побудованого довірчого інтервалу.

Припустимо тепер, що справжнє значення параметра інтенсивності гамма-розподілу `rate=0.1` відоме, а треба лише оцінити `shape`. У цьому випадку доведеться задати початкове значення для пошуку оцінки у параметрі `start` функції `fitdistr()`. Значення параметра `rate=r` передається у функцію, яка обчислює щільність гамма-розподілу. Оцінку можна обчислити за допомогою такого скрипта:

```
> options(warn = -1)
> fitdistr(x,"gamma",start=list(shape=1),rate=r)

     shape
  4.87187500
 (0.09370892)
```

(Якщо цей скрипт запустити без команди `options(warn=-1)`, він видасть попередження про те, що метод оптимізації, використаний для знаходження точки максимуму у `fitdistr()`, не є надійним. Проте ми отримуємо адекватну оцінку, отже, шукати кращого методу не будемо).

Як бачимо, отримана оцінка з середньоквадратичним відхиленням значно меншим, ніж у випадку, коли параметр `rate` був невідомим. Тобто знання цього параметра дозволяє оцінити `shape` точніше.

Наведемо скрипт, який будує довірчий еліпсоїд для обох параметрів одразу (він використовує результати роботи першого скрипта з цього прикладу):

```
>   # confidence ellipse plot
> PlotEllipse<-function(parest,S,
+                       alpha=0.05,numpoint=50){
+ Sinv<-ginv(S)
+ # draw the confidence region
+ # get points for a circle with radius r
+ r=sqrt(qchisq(1-alpha,2)*2)
+ theta=seq(0,2*pi,length.out=numpoint)
+ z=cbind(r*cos(theta),r*sin(theta))
+ # transform points of circle into points of ellipse
+ # using svd of inverse covariance matrix
+ Sinv_svd=svd(Sinv) # inverse of covariance matrix
+ # transform from circle to ellispse:
```

```
+ xt=t(Sinv_svd$v)%*%diag(1/sqrt(Sinv_svd$d))%*%t(z)
+ x=t(xt)
+ # translate the ellipse so that center is
+ # the estimated parameter value:
+ x=x+matrix(rep(as.numeric(parest),numpoint),
+         nrow=numpoint,byrow=T)
+ plot(x[,1],x[,2],type="l",xlab="shape",
+   ylab="rate",lwd=1)
+ points(parest[1],parest[2],pch=20,col="blue",cex=2)
+ }
> PlotEllipse(fitted$estimate,fitted$vcov)
```

Отриманий еліпс зображено на рис. 8.10.

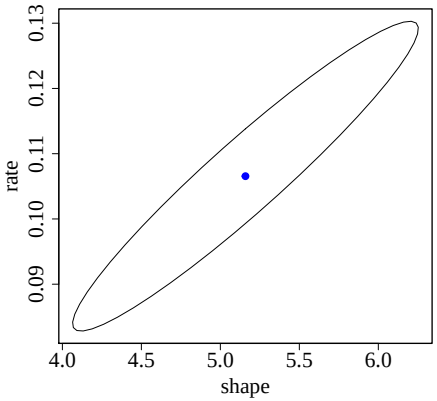

Рис. 8.10. Довірчий еліпсоїд для параметрів гамма-розподілу

У скрипті вводиться функція `PlotEllipse`, призначена для рисування довірчого еліпса для двох невідомих параметрів на основі рівняння (8.21). Її параметри:

`parest` — двовимірний вектор оцінок пари невідомих параметрів;

`S` — оцінка для коваріаційної матриці `parest`;
`alpha` — рівень значущості довірчого еліпса;
`numpoint` — кількість точок для рисування еліпса по точках[15].

◄

Приклад 8.6.2. (Вальниці і розподіл Вейбула) При аналізі надійності для опису розподілу тривалості роботи пристроїв часто використовують розподіл Вейбула. Зокрема, ще з 40-х років XX ст. цей розподіл застосовують для аналізу надійності кулькових вальниць (підшипників). Розглянемо приклад даних з випробувань вальниць (приклад 3.3.1 з [37], с. 98). Ці дані містять результати 21 випробування вальниць, для кожної вальниці вказано кількість мільйонів обертів, після яких вона зламалась. Дані містяться у наборі `bearings` з бібліотеки `reliaR`. Оцінимо параметри `shape` і `scale` розподілу Вейбула для підгонки моделі цих даних і подивимось, наскільки модель відповідає даним.

```
> library(reliaR)
> library(MASS)
> data("bearings") # підключаємо дані про вальниці
> # знаходимо оцінки параметрів розподілу Вейбула:
> fitted<-fitdistr(bearings, "weibull")
> fitted$estimate

   shape    scale
 2.10221 81.85425

>   # рисуємо гістограму даних
>   hist(bearings,breaks=5,probability=TRUE,
+  xlab=" ",ylab=" ",main=" ")
>   # на гістограмі  щільність розподілу Вейбула
>   #                        з оціненими параметрами
```

[15] Алгоритм, реалізований у цій функції, взято з `www.r-bloggers.com/learning-r-parameter-fitting-for-models-involving-differential-equations/`

```
>   curve(dweibull(x,shape=fitted$estimate["shape"],
+   scale=fitted$estimate["scale"]),col="blue",add=TRUE)
>   # рисуємо Q-Q діаграму:
>   plot(qweibull(ppoints(bearings),
+              shape=fitted$estimate["shape"],
+   scale=fitted$estimate["scale"]),sort(bearings),
+   xlab="",ylab="")
>   abline(a=0,b=1,col="red")
```

△

Таким чином, підгонка за методом найбільшої вірогідності дає оцінку `shape`$\approx$ 2 і `scale`$\approx$ 82. Гістограма (рис. 8.11) та QQ-діаграма (рис. 8.12) не виявляють суттєвих відхилень розподілу даних від теоретичної вейбулової моделі. (Зауважимо, що при такому малому обсязі даних, гістограми зазвичай бувають мало інформативними).

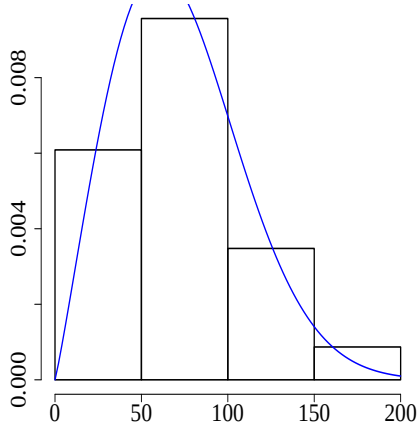

Рис. 8.11. Підгонка розподілу часу роботи вальниць. Гістограма

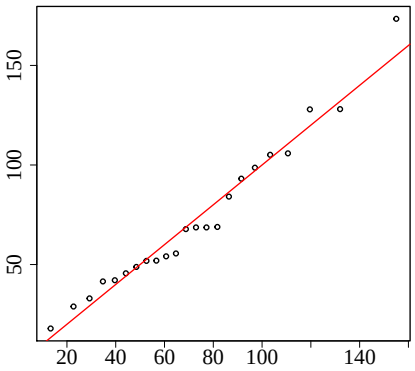

Рис. 8.12. Підгонка розподілу часу роботи вальниць. Q-Q-діаграма

◄

Розділ 9

Перевірка гіпотез

9.1 Загальні відомості

Поруч із задачами оцінювання параметрів у статистиці велику роль відіграють задачі перевірки гіпотез. Тут ми не намагаємось оцінити невідомий параметр якомога точніше. Задача полягає в перевірці того, наскільки наші дані підтверджують або суперечать певним припущенням про досліджуване явище.

Статистичні гіпотези — це припущення про розподіл спостережуваних статистичних даних. На практиці вони пов'язані з деякими змістовними гіпотезами про природу досліджуваного явища, об'єкта, процесу. Правильний підхід до аналізу даних полягає в тому, щоб почати з висунення змістовної гіпотези. Далі слід сформулювати ймовірнісну модель та статистичну гіпотезу, яка відповідає змістовній. Після цього — вибрати правильний алгоритм перевірки обраної статистичної гіпотези (статистичний тест), застосувати його та інтерпретувати отримані результати.

Досить поширена протилежна практика: до статистичних даних намагаються застосувати різноманітні тести, сподіваючись, що їх результати наштовхнуть дослідника на змістовні гіпотези. На мою думку, такий підхід є невдалим. Якщо у статистика немає розумної ймовірнісної моделі даних, доцільно скористатись де-

скриптивними методами і спробувати відшукати її. Тільки після цього є сенс висувати статистичні гіпотези та застосовувати тести для їх перевірки.

У цьому підрозділі розглядається формальна постановка задачі перевірки статистичних гіпотез у загальному вигляді. Як ці гіпотези можуть бути пов'язані із змістовними гіпотезами у різних прикладних областях, буде видно з прикладів, що розглядаються далі.

Нехай весь набір статистичних даних $\mathbf{X}$ є випадковим елементом простору даних $\mathcal{X}$. Розподіл даних, тобто набір ймовірностей $\mathbf{P}_\vartheta^{\mathbf{X}}(A) = \mathsf{P}\{\mathbf{X} \in A\}$ відомий з точністю до невідомого параметра $\vartheta \in \Theta$, де Θ — деяка множина (простір) можливих значень параметра. Наприклад, $\mathbf{X}$ може бути кратною вибіркою обсягу n, а ϑ — числовим параметром. Тоді $\mathcal{X} = \mathbb{R}^n$, $\vartheta \in \Theta \subseteq \mathbb{R}$. Будемо вважати, що ϑ однозначно задає розподіл даних $\mathbf{X}$. Тоді гіпотези про цей розподіл є гіпотезами про можливі значення ϑ. Ми обмежимось розглядом двоальтернативних гіпотез. Будемо вважати, що простір параметрів Θ розбитий на дві множини, що не перетинаються: $\Theta = \Theta_0 \cup \Theta_1$, $\Theta_0 \cap \Theta_1 = \varnothing$. Гіпотезою H_i назвемо припущення про те, що $\vartheta \in \Theta_i$, $i = 0, 1$.

При класичному підході до задачі перевірки гіпотез, гіпотези H_0 і H_1 не є рівноправними. Гіпотеза H_0 вважається основною, тобто її відхиляють лише тоді, коли дані переконливо показують, що вона є хибною. Гіпотеза H_1 вважається альтернативною, її приймають лише тоді, коли дані переконливо свідчать на її користь.

Для перевірки статистичних гіпотез за даними використовують алгоритми, які називають статистичними тестами[1]. З формальної точки зору тест можна розглядати як функцію, що будь-якому можливому значенню з простору даних зіставляє номер гіпотези, яку тест приймає при цьому значенні $\mathbf{X}$. Таким чином, тест — це вимірна функція $\pi : \mathcal{X} \to \{0, 1\}$. Якщо $\pi(\mathbf{X}) = 0$, тест приймає

[1] У англомовній літературі прийнята назва *test*, у російськомовній — *критерий*. В українській літературі вживають як назву тест, так і назву критерій. Інколи "критерієм" називають те, що ми далі називаємо статистикою тесту.

основну гіпотезу, якщо $\pi(\mathbf{X}) = 1$ — альтернативу.

На практиці тести часто мають вигляд

$$\pi(\mathbf{X}) = \mathbb{1}\{S(X) > C\}, \text{ або } \pi(\mathbf{X}) = \mathbb{1}\{S(X) < C\},$$

де $S(\mathbf{X})$ — статистика, тобто вимірна функція від даних, а C — деяке фіксоване число. У такому випадку $S(X)$ називають статистикою тесту, а C — порогом.

Характеризація якості тестів. Якість тесту характеризується ймовірністю того, що тест прийме неправильну гіпотезу (помилиться). При використанні тесту можливі помилки двох типів:

1. **Помилка першого типу:** правильна основна гіпотеза, а тест її відхиляє, тобто $\vartheta \in \Theta_0$, але $\pi(\mathbf{X}) = 1$. Ймовірність такої помилки

$$\alpha_\pi(\vartheta) = \alpha(\vartheta) = \mathsf{P}_\vartheta\{\pi(\mathbf{X}) = 1\} = \mathsf{E}_\vartheta\, \pi(\mathbf{X}),\ \vartheta \in \Theta_0.$$

2. **Помилка другого типу:** основна гіпотеза хибна, а тест її приймає, тобто $\vartheta \in \Theta_1$, але $\pi(\mathbf{X}) = 0$. Ймовірність такої помилки

$$\beta_\pi(\vartheta) = \beta(\vartheta) = \mathsf{P}_\vartheta\{\pi(\mathbf{X}) = 0\} = 1 - \mathsf{E}_\vartheta\, \pi(\mathbf{X}),\ \vartheta \in \Theta_1.$$

Ймовірність того, що тест правильно прийме альтернативу, коли вона є правильною, називають **потужністю** (power) тесту і позначають

$$\varphi_\pi(\vartheta) = 1 - \beta_\pi(\vartheta) = \mathsf{E}_\vartheta\, \pi(\mathbf{X}),\ \vartheta \in \Theta_1.$$

Найбільше можливе значення ймовірності помилки першого типу для тесту π називають **рівнем значущості тесту** (test signficance level):

$$\alpha_\pi = \sup_{\vartheta \in \Theta_0} \alpha_\pi(\vartheta).$$

Серед усіх можливих тестів бажано вибрати такий, який матиме найменшу ймовірність помилки. Проте зазвичай при зменшенні α_π збільшується β_π, і навпаки. Тому у статистиці прийнятий наступний підхід до вибору тестів.

Фіксують деяке мале додатне число $\alpha = \alpha_0$, яке називають **стандартним рівнем значущості** і розглядають лише тести π, для яких

$$\alpha_\pi \leq \alpha_0 \tag{9.1}$$

(тобто ймовірність помилково відхилити основну гіпотезу не перевищує стандартного рівня значущості). Серед таких тестів вибирають тест з найбільшою потужністю $\varphi_\pi(\vartheta)$.

Якщо вдається знайти такий тест π^* що $\alpha_{\pi^*} \leq \alpha_0$ і для всіх інших тестів π, які задовольняють умову (9.1) виконано

$$\varphi_{\pi^*}(\vartheta) \geq \varphi_\pi(\vartheta) \text{ для всіх } \vartheta \in \Theta_1,$$

то тест π^* називають **рівномірно найбільш потужним тестом** (р.н.п.) рівня α для перевірки гіпотези H_0 проти альтернативи H_1.

Р.н.п. тести існують не для всіх гіпотез. Часто їх шукають не в класі всіх можливих тестів, а лише в якомусь класі "правильних" тестів. Якщо р.н.п. тест у даній задачі перевірки гіпотез знайти не вдається, використовують тести, потужність яких є достатньою для практичних потреб.

Вибір основної гіпотези та рівня значущості. Оскільки H_0 і H_1 не є рівноправними, то важливо правильно визначитись, яке з двох альтернативних припущень вважати основним при перевірці гіпотез. Нехай, наприклад, дослідник проводить експеримент з метою виявити певний ефект. Скажімо, ефектом може бути вплив ліків на протікання хвороби. В результаті експерименту отримані дані $\mathbf{X}$. Якщо правильним є припущення про відсутність ефекту, $\mathbf{X}$ має розподіл R, якщо ефект є — розподіл S. Як основну можна взяти гіпотезу $F^{\mathbf{X}} = R$ або гіпотезу[2] $F^{\mathbf{X}} = S$. Який вибір кращий?

Якщо дослідник хоче своїми даними переконати колег у тому, що ефект виявлений, він повинен за основну взяти гіпотезу про те, що ефекту немає — $H_0 : F^{\mathbf{X}} = R$. Тоді, якщо відповідний

[2]У цьому випадку невідомим параметром можна вважати сам розподіл, який набуває значень з двоелементної множини $\Theta = \{R, S\}$.

статистичний тест прийме $H_1 : F^{\mathbf{X}} = S$, це буде підтвердженням наявності ефекту на основі даних експерименту, а не внаслідок того, що гіпотеза $F^{\mathbf{X}} = S$ апріорі була вибрана основною. І навпаки, якщо досліди проводять, наприклад, для перевірки відсутності негативних побічних ефектів ліків, основною повинна бути гіпотеза про те, що такі ефекти є.

Наступним важливим кроком в організації перевірки гіпотез є вибір стандартного рівня значущості α. Цей рівень також визначається не з математичних, а з прикладних міркувань. Наприклад, нехай у випадку первірки лікувального ефекту досліджуваної речовини основна гіпотеза — відсутність ефекту. Виберемо $\alpha = 0.05$. Тоді перевіряючи різні ліки, що насправді не дають ефекту, будемо хибно його виявляти в середньому один раз на двадцять випадків. І, відповідно, рекомендувати ці ліки для подальшого використання. Якщо це нас не влаштовує, можна зменшити α, встановивши його, наприклад, 0.01. Але тоді ми ризикуємо частіше не помічати тих ефектів, які ліки справді дають.

З цих міркувань у різних предметних областях встановлюються різні стандартні рівні значущості. Найменший з рекомендованих рівнів 0.05 прийнятий у соціології, економіці, медицині. Більш строгі рівні 0.01 або 0.001 прийняті в експериментальній фізиці та інженерних науках.

Досягнутий рівень значущості тесту[3]. (attained significance, p-value) Оскільки вибір рівня значущості може змінюватись залежно від обставин, статистичні тести прийнято одразу розробляти так, щоб вони могли працювати з будь-яким обраним стандартним рівнем значущості. При цьому і результат роботи тесту на конкретних даних часто зручно подавати так, щоб по ньому можна було

[3] Як стандартний рівень значущості 0.05, так і використання досягнутих рівнів значущості для перевірки статистичних гіпотез є предметом дискусій, що не вщухають вже майже століття. З сучасним критичним поглядом на це можна ознайомитись у [42]. Американська статистична асоціація нещодавно опублікувала спеціальну заяву з цього приводу [53]. Застереження щодо можливих неправильних трактувань досягнутого рівня значущості див. [30].

одразу сказати, при якому рівні значущості приймається основна гіпотеза, а при якому — альтернатива. Для такого запису використовується спеціальна статистика $p(\mathbf{X})$.

Справа в тому, що один і той самий тест зазвичай можна записати у багатьох еквівалентних формах. Нехай, наприклад, тест має вигляд

$$\pi(\mathbf{X}) = \mathbb{1}\{S(\mathbf{X}) > C_\alpha\}, \tag{9.2}$$

де поріг C_α і статистика тесту $S(\mathbf{X})$ підібрані так, що рівень значущості тесту $\alpha_\pi = \alpha$. Візьмемо довільну строго зростаючу функцію h. Тоді (9.2) еквівалентно

$$\pi(\mathbf{X}) = \mathbb{1}\{h(S(\mathbf{X})) > h(C_\alpha)\}.$$

Таким чином, пара (статистика,поріг) $(h(S(\mathbf{X})), h(C_\alpha))$ є еквівалентною парі $(S(\mathbf{X}), C_\alpha)$. Вони породжують один і той самий тест.

Серед всіх еквівалентних статистик тесту є одна — $p(\mathbf{X})$, при використанні якої тест набуває вигляду

$$\pi(\mathbf{X}) = \mathbb{1}\{p(\mathbf{X}) < \alpha\}.$$

Ця статистика $p(\mathbf{X})$ і називається досягнутим рівнем значущості тесту.

Таким чином, якщо на конкретних даних тест дає значення досягнутого рівня значущості p, то основна гіпотеза приймається при $p \geq \alpha$ і відхиляється при $p < \alpha$.

Розглянемо важливий частковий випадок, коли статистика $S(\mathbf{X})$ тесту $\pi(\mathbf{X})$, заданого (9.2), підібрана так, щоб її розподіл був тим самим при всіх значеннях параметра, що відповідають основній гіпотезі[4]. Позначимо функцію розподілу для цього розподілу $G(s)$:

$$G(s) = \mathsf{P}_\vartheta\{S(\mathbf{X}) < s\} \text{ для всіх } \vartheta \in \Theta_0.$$

[4]Такі тести називають *незалежними від розподілу*. Зрозуміло, що на альтернативі розподіл статистики має відрізнятись від розподілу на основній гіпотезі, інакше від такого тесту користі не буде.

Обмежимось випадком, коли G — неперервна строго зростаюча функція. Тоді тільки поріг $C_\alpha = Q^G(1-\alpha)$ забезпечує рівень значущості α для тесту π.

Таким чином, тест з рівнем значущості α має вигляд

$$\pi(\mathbf{X}) = \mathbb{1}\{S(\mathbf{X}) > G^{-1}(1-\alpha)\} = \mathbb{1}\{1 - G(S(\mathbf{X})) < \alpha\}.$$

Отже, досягнутий рівень значущості цього тесту можна обчислити за формулою

$$p(\mathbf{X}) = 1 - G(S(\mathbf{X})). \tag{9.3}$$

9.2 Тест відношення вірогідності для перевірки простих гіпотез

Статистична гіпотеза називається простою, якщо вона однозначно визначає розподіл даних. Розглянемо випадок, коли і основна гіпотеза, і альтернатива є простими. У цьому випадку параметричну модель розподілу даних можна не описувати, а задавати гіпотези безпосередньо, вказуючи відповідний розподіл.

Як і раніше, позначимо набір статистичних даних через $\mathbf{X}$. Нехай гіпотезі H_i $(i = 0, 1)$ відповідає розподіл даних $F_i^{\mathbf{X}}$. Припустимо, що для деякої міри μ на просторі даних $\mathcal{X}$ існують щільності розподілів $F_i^{\mathbf{X}}$ відносно μ, які ми позначимо $f_i^{\mathbf{X}}$:

$$F_i^{\mathbf{X}}(A) = \int_A f_i^{\mathbf{X}}(x)\mu(dx).$$

Відношенням вірогідності (likelihood ratio) для перевірки простої гіпотези H_0 проти простої альтернативи H_1 називають статистику

$$\mathrm{LR}(\mathbf{X}) = \frac{f_1^{\mathbf{X}}(\mathbf{X})}{f_0^{\mathbf{X}}(\mathbf{X})}.$$

Тестом відношення вірогідності з порогом C називають тест

$$\pi(\mathbf{X}) = \pi_C(\mathbf{X}) = \begin{cases} 1 & \text{якщо } \mathrm{LR}(\mathbf{X}) > C, \\ 0 & \text{якщо } \mathrm{LR}(\mathbf{X}) \le C. \end{cases}$$

У випадку простих основної та альтернативної гіпотез ймовірності помилок першого і другого типів для будь-якого тесту π виражаються кожна одним числом (а не функцією від невідомого параметра ϑ, як у загальному випадку) і позначаються відповідно α_π та β_π.

Теорема 9.2.1. *Якщо при виконанні H_0 випадкова величина $LR(\mathbf{X})$ має неперервну функцію розподілу G, то тест відношення вірогідності $\pi^*_\alpha(\mathbf{X}) = \pi_C(\mathbf{X})$ з $C = C_\alpha = Q^G(1-\alpha)$ буде найбільш потужним[5] тестом для перевірки H_0 проти H_1 з рівнем значущості α.*

Досягнутий рівень значущості цього тесту

$$p(\mathbf{X}) = 1 - G(LR(\mathbf{X})).$$

(Доведення див. теорему 1 п. 42 у [3] або п. 3.17 у [9]).

Зрозуміло, що такий найбільш потужний тест можна будувати, використовуючи не безпосередньо статистику відношення вірогідності $\mathrm{LR}(\mathbf{X})$, а будь-яку монотонну функцію від неї, — $h(\mathrm{LR}(\mathbf{X}))$. Для визначення порогу c_α у такому тесті $\pi(h(\mathrm{LR}(\mathbf{X})))$ можна скористатись умовою

$$\alpha = \alpha_\pi = \mathsf{E}_0\, \pi(h(\mathrm{LR}(\mathbf{X}))).$$

(математичне сподівання береться в припущенні, що правильною є H_0).

У випадку, коли $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ є кратною вибіркою, щільність всього набору даних $f_i^{\mathbf{X}}$ записується як добуток щільностей окремих спостережень ξ_i. Тому

$$\mathrm{LR}(\mathbf{X}) = \prod_{j=1}^{n} \frac{f_1(\xi_j)}{f_0(\xi_j)},$$

де f_i — щільність розподілу ξ_j при виконанні гіпотези H_i.

[5] У термінології попереднього підрозділу — рівномірно найбільш потужним, але зараз потужність це не функція, а одне число, тому казати "рівномірно" немає сенсу.

На практиці працювати з сумами буває зручніше, ніж з добутками, тому для кратних вибірок використовують також логарифмічне відношення вірогідності

$$\mathrm{lr}(\mathbf{X}) = \log \mathrm{LR}(\mathbf{X}) = \sum_{j=1}^{n} \log\left(\frac{f_1(\xi_j)}{f_0(\xi_j)}\right).$$

У багатьох випадках практичного застосування статистичних тестів описати аналітично розподіл відношення вірогідності даних при виконанні основної гіпотези (або альтернативи) не вдається. Але підібрати правильний поріг тесту та визначити ймовірність помилки другого типу можна, використовуючи імітаційне моделювання.

Для визначення порога тесту згенеруємо спочатку велику кількість B незалежних між собою вибірок обсягу n з розподілом, що відповідає основній гіпотезі. Позначимо ці вибірки $\mathbf{X}^{(1;0)},\ldots,\mathbf{X}^{(B;0)}$. Обчислимо значення статистики lr на кожній з цих вибірок: $\mathrm{lr}_j^0 = \mathrm{lr}(\mathbf{X}^{(j;0)})$. Набір $\mathbf{L}^{(0)} = (\mathrm{lr}_1^0,\ldots,\mathrm{lr}_B^0)$ є кратною вибіркою з розподілом, який відповідає розподілу G статистики $\mathrm{lr}(\mathbf{X})$ при виконанні основної гіпотези. Тому вибірковий квантиль $\hat{c}_\alpha = Q^{\mathbf{L}^0}(1-\alpha)$ є хорошою оцінкою для порога тесту $c_\alpha = Q^G(1-\alpha)$, що відповідає рівню значущості α. Поріг $\hat{c}_\alpha$ можна використовувати при реалізації тесту для перевірки гіпотез на конкретних даних $\mathbf{X}$. При цьому тест π працює таким чином:

> Прийняти гіпотезу H_0, якщо $\mathrm{lr}(\mathbf{X}) \le \hat{c}_\alpha$ і відхилилити, якщо $\mathrm{lr}(\mathbf{X}) > \hat{c}_\alpha$.

Досягнутий рівень значущості тесту можна оцінити як

$$\hat{p}(\mathbf{X}) = 1 - \hat{F}_B^{\mathbf{L}^0}(\mathrm{lr}(\mathbf{X})) = \frac{1}{B}\sum_{j=1}^{B} \mathbb{1}\{\mathrm{lr}_j^0 > \mathrm{lr}(\mathbf{X})\}.$$

Для того, щоб оцінити ймовірність помилки другого типу цього тесту, треба згенерувати вибірки з розподілом, що відповідає

альтернативі: $\mathbf{X}^{(1;1)},\ldots,\mathbf{X}^{(B;1)}$ і обчислити статистику lr на них: $\mathrm{lr}_j^1 = \mathrm{lr}(\mathbf{X}^{(j;1)})$, $\mathbf{L}^{(1)} = (\mathrm{lr}_1^1,\ldots,\mathrm{lr}_B^1)$.

Оцінкою для β_π буде частота помилок тесту π на вибірках $\mathbf{X}^{(j;1)}$:

$$\hat{\beta}_\pi = \frac{1}{B}\sum_{j=1}^{B} \mathbb{1}\{\mathrm{lr}_j^1 < \hat{c}_\alpha\}.$$

Приклад 9.2.1. Розглянемо наступну умовну задачу. Нехай деяким підприємством була закуплена партія n електричних ламп. Всі лампочки були використані і було зареєстровано час роботи кожної лампи до перегоряння — ξ_j, $j = 1,\ldots,n$. Відомо, що фірмові лампи мають експоненційний розподіл часу роботи до перегоряння з інтенсивністю λ_0, а час роботи дешевого аналогу фірмових виробів — також експоненційний з інтенсивністю $\lambda_1 > \lambda_0$. Лампи були закуплені як фірмові, але за спостереженнями виникла підозра, що це дешевий аналог. Треба вирішити, чи слід виставляти рекламацію постачальнику товару.

Гіпотеза — A: "лампи є фірмовими" не вимагає додаткових дій. Гіпотеза — B: "лампи є дешевим аналогом" приводить до необхідності подавати рекламацію. Отже, для прийняття гіпотези B потрібне обґрунтування на основі спостережуваних даних. Тому як основну слід обрати гіпотезу A, і дотримуватись її доти, поки дані не змусять нас прийняти B.

Таким чином, $\mathbf{X} = (\xi_1,\ldots,\xi_n)$ — кратна вибірка з експоненційного розподілу з інтенсивністю λ. Треба за цією вибіркою перевірити гіпотезу H_0: $\lambda = \lambda_0$ проти альтернативи H_1: $\lambda = \lambda_1$, причому $\lambda_0 < \lambda_1$.

Логарифмічне відношення вірогідності для цих гіпотез дорівнює

$$\mathrm{lr}(\mathbf{X}) = \sum_{j=1}^{n} \log\left(\frac{\lambda_1 e^{-\lambda_1\xi_j}}{\lambda_0 e^{-\lambda_0\xi_j}}\right) = n(\log\lambda_1 - \log\lambda_0) + (\lambda_0 - \lambda_1)\sum_{j=1}^{n}\xi_j.$$

Оскільки $\sum_{j=1}^{n}\xi_j$ є монотонно спадною функцією від $\mathrm{lr}(\mathbf{X})$, тест відношення вірогідності можна записати в еквівалентній формі

$\pi_c(\mathbf{X}) = \mathbb{1}\{\sum_{j=1}^n \xi_j < c\}$. Для заданого рівня значущості α відповідний поріг $c = c_\alpha$ можна вибрати з умови

$$\alpha = \mathsf{E}_0\, \pi_c(\mathbf{X}) = \mathsf{P}_{\lambda_0} \left\{ \sum_{j=1}^n \xi_j < c \right\}.$$

Якщо ξ_j — експоненційно розподілені з інтенсивністю λ незалежні випадкові величини, то $\sum_{j=1}^n \xi_j$ має Γ-розподіл з інтенсивністю λ і параметром форми n, тобто $\Gamma(n, \lambda)$. Отже, $c_\alpha = Q^{\Gamma(n,\lambda_0)}(\alpha)$.

Ймовірність помилки другого типу для цього тесту

$$\beta_\pi = 1 - \mathsf{E}_{\lambda_1}\, \pi(\mathbf{X}) = \mathsf{P}\{\sum_{j=1}^n \xi_j > c_\alpha\} = F^{\Gamma(n,\lambda_1)}(Q^{\Gamma(n,\lambda_0)}(\alpha)).$$

◄

Приклад 9.2.2. Цей приклад також є умовним. Нехай досліджується генотип деякої рослини. Дослідника цікавить певний ген, який впливає на конкретну ознаку рослини ξ (напр., співвідношення довжини і ширини листка). Відомо, що є два алеля (варіанти) цього гена — домінантний A та рецесивний — a. Дослідника цікавить конкретна рослина Z, генотип якої може бути Aa або aa. Для з'ясування того, яким насправді є генотип Z, цю рослину схрещують з рослиною, яка точно має генотип aa. Якщо генотип Z є aa, то всі нащадки також будуть мати генотип aa. Якщо генотип Z — Aa, то кожен з нащадків з ймовірністю $p = 1/2$ отримає генотип Aa, і з ймовірністю $1 - p$ — генотип aa.

Відомо, що ознака ξ у рослин з генотипом aa має щільність розподілу f_{aa}, а у рослин з генотипом Aa — f_{Aa}. Отримано n нащадків, значення ξ у j-го нащадка ξ_j. Треба за вибіркою $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ визначити, яким був генотип рослини Z. Основною є гіпотеза про генотип aa.

Зрозуміло, що щільність, яка відповідає основній гіпотезі — це $f_0(x) = f_{aa}(x)$. Щільність, що відповідає альтернативі, є сумішшю двох щільностей з ймовірностями p та $1 - p$:

$$f_1(x) = p f_{Aa}(x) + (1 - p) f_{aa}(x).$$

Логарифмічне відношення вірогідності має вигляд

$$\mathrm{lr}(\mathbf{X}) = \sum_{j=1}^{n} \log\left(\frac{p f_{Aa}(\xi_j) + (1-p) f_{aa}(\xi_j)}{f_{aa}(\xi_j)}\right).$$

Аналітично розподіл цієї величини виразити не можна. Тому скористаємось імітаційним моделюванням, як описано вище. У прикладі розглядається випадок, коли f_{aa} та f_{Aa} — гауссові щільності з математичним сподіванням і середньоквадратичним відхиленням `m0=2`, `s0=0.4` для f_{aa} і `m1=3`, `s1=0.75` для f_{Aa}. Кількість спостережень (кількість нащадків досліджуваної рослини) `n=12`:

```
> set.seed(5)
> m0<-2    # середнє для aa
> s0<-0.4 # сер. кв. відх. для aa
> m1<-3    # середнє для Aa
> s1<-0.75 # сер. кв. відх. для Aa
> p<-0.5   # ймовірність aa при альтернативі
> n<-12    # обсяг вибірки
> # логарифм щільності, що відповідає основній гіпотезі
> f0<-function(x){log(dnorm(x,m0,s0))}
> # логарифм щільності, що відповідає альтернативі
> f1<-function(x){log(p*dnorm(x,m0,s0)+
+                     (1-p)*dnorm(x,m1,s1))}
> # логарифмічне відношення вірогідності (x - вибірка)
> lr<-function(x)sum(sapply(x,f1)-sapply(x,f0))
> # генератор однієї вибірки при H0
> gen0<-function(n)rnorm(n,m0,s0)
> # генератор однієї вибірки при H1
> gen1<-function(n){
+ z<-rbinom(n,size=1,prob=p)
+ rnorm(n,z*m0+(1-z)*m1,z*s0+(1-z)*s1)
+ }
> alpha<-0.05 # стандартний рівень значущості
```

```
> B<-10000    # кількість модельованих вибірок
> lr0<-replicate(B,lr(gen0(n))) # масив значень lr при H0
>  # поріг тесту, що відповідає рівню alpha
> Ca<-quantile(lr0,1-alpha)
> Ca

      95%
-1.326246

> lr1<-replicate(B,lr(gen1(n))) # масив значень lr при H1
> # оцінка ймовірності помилок другого роду
> mean(lr1<Ca)

[1] 0.0264

> # відображення гістограм для lr
> mi<-min(c(lr0,lr1))
> mx<- 60 #max(c(lr0,lr1))
> hist(lr0,breaks=15,probability=T,
+ angle=0,density=12,xlim=c(mi,mx),ylim=c(0,0.33),
+ col="red",xlab="lr",main=" ")
> hist(lr1,probability=T,
+ breaks=15,angle=90,density=12, xlim=c(mi,mx),
+ col="blue",add=T)
> abline(v=Ca)
```

△

Як бачимо, для рівня значущості $\alpha = 0.05$ ми отримали поріг тесту $c_\alpha = -1.326246$. При цьому ймовірність помилки другого роду оцінюється як $\beta_\pi = 0.0264$. Ця оцінка вийшла навіть меншою, ніж встановлена нами ймовірність помилки першого роду.

На рис. 9.1 зображені гістограми значень (логарифмічного) відношення вірогідності для основної гіпотези (червоним кольором, горизонтальна штриховка) та альтернативи (синім кольором, вертикальна штриховка). Вертикальна лінія відмічає положення порогу c_α. Площа тієї частини гістограми для H_0, що лежить праворуч

від c_α, відповідає ймовірності помилки першого роду $\alpha_\pi = 0.05$. Площа частини гістограми для H_1, яка лежить ліворуч від c_α, відповідає ймовірності помилки другого типу β_π. Зміщуючи положення порогу праворуч, можна зменшити α_π, але β_π при цьому збільшиться. І навпаки, зменшуючи поріг, ми збільшуємо ймовірність помилки першого типу та збільшуємо другого. Таким чином, отримані гістограми дозволяють побачити, наскільки хорошим може бути тест для розпізнавання цих двох гіпотез. Чим ближче ці гістограми одна до одної, тим менше шансів побудувати хороший тест.

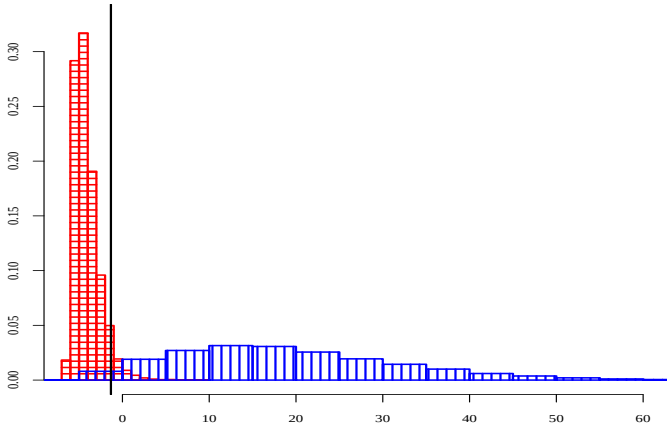

Рис. 9.1. Гістограми для відношень вірогідності з прикладу 9.2.2. Горизонтальна штриховка (червоним) - при основній гіпотезі, вертикальна штриховка (синім) - при альтернативі. Вертикальна лінія відповідає порогу тесту

Нехай в ході експерименту було отримано такі значення ξ:

2.6	1.7	1.9	3.1	3.9	1.7
1.6	1.9	1.6	2.9	2.6	1.4

Перевіримо, на користь якої гіпотези свідчать ці дані:

```
> # Вибірка для перевірки гіпотези:
> x<-c(2.6,1.7,1.9,3.1,3.9,1.7,
+ 1.6,1.9,1.6,2.9,2.6,1.4)
> # статистика відношення вірогідності:
> lr(x)

[1] 9.685149

> # досягнутий рівень значущості:
> mean(lr0>lr(x))

[1] 0
```

Тут ми спочатку ввели вибірку і позначили її `x`. Потім обчислили логарифмічне відношення вірогідності — воно виявилось рівним 9.685149. Це більше, ніж обчислене нами c_α, отже, при рівні значущості 0.05 слід прийняти альтернативу: генотип досліджуваної рослини Aa. Далі ми обчислили досягнутий рівень значущості, і отримали значення 0. (Насправді, звичайно, 0 — це лише наша оцінка, справжнє $p(\mathbf{X})$ додатне, але настільки мале, що наша техніка оцінювання не дозволяє помітити його відмінність від 0). Таким чином, при всіх розумних рівнях значущості ці дані свідчать на користь альтернативи. ◄

Приклад 9.2.3. Покажемо, як статистичний аналіз дозволяє аналізувати автентичність літературного твору[6]. Як відомо, "Сло-

[6] У цьому прикладі розглядаються реальні дані та цікава історична проблема. Але ми обмежимось лише розглядом дуже малої частинки цієї проблеми, тому, звичайно, наше дослідження матиме переважно навчальний характер. Докладний статистичний аналіз всіх даних, що стосуються "Слова", потребував би, мабуть, монографії обсягу більшого, ніж вся ця книга.

во о полку Ігоревім" було опубліковано у 1800 році А.І. Мусіним-Пушкіним як передрук древньоруського рукопису. В українській та російській науці прийнято вважати "Слово" автентичним твором невідомого древньоруського автора XII — XIII ст. (Гіпотеза А). Однак значної популярності (особливо у західному літературознавстві) набули гіпотези про те, що "Слово" є імітацією/підробкою, виконаною у XVIII ст. (Гіпотеза І). Деякі дослідники називають також можливих кандидатів на роль автора-імітатора. Зокрема, американський славіст Едвард Кінан обстоює авторство чеського філолога Йосифа Добровського (1753 — 1859) [35] (Гіпотеза ІД). Французький славіст Андре Мазон [40] та російський історик А.А. Зімін [8] вважали автором архімандрита Іоїля Биковського, від якого отримав рукопис Мусін-Пушкін [40, 8] (Гіпотеза ІБ).

Російський філолог А.А. Зализняк провів лінгвістичний аналіз мови "Слова", порівнюючи її з мовою різних східнослов'янських літературних джерел та мовою берестяних грамот, які, починаючи з середини XX ст. знаходять археологи при розкопках давньоруських міст (у Новгороді, Пскові, Старій Русі, Торжку, Звенигороді Галицькому, Києві та ін.). Результати його досліджень викладені у книзі [7]: мова "Слова" має специфічні особливості давньоруської мови XII — XIII ст., які не були відомі літераторам та лінгвістам XVIII ст.[7] Ці особливості не могли бути відтворені авторами XVIII ст. випадково або несвідомо. Тому А.А. Зализняк приходить до висновку про правильність гіпотези А.

У цьому прикладі розглянуто лише один аргумент з книги А.А. Зализняка, що стосується положення у реченні частки *ся*. Як відомо, у сучасних літературних українській та російській мовах,

[7]Точніше, за формулюванням А.А. Зализняка: "... текст СПИ был создан в конце XII — начале XIII века и переписан где-то на северо-западе в XV или XVI веке" [7]. Зрозуміло, що "Слово" не могло з'явитись раніше походу Ігоря 1185 року, отже, за спогадами учасників подій його могли б написати або наприкінці XII, або у XIII ст. Далі я буду вказувати XIII ст. як крайню межу, після якої "Слово" вже не можна розглядати як автентичний твір.

ся з дієсловами вживається як "постфікс", тобто частина слова, що стоїть наприкінці (постпозиція *ся*, напр.: *стосується*, *вживається*). У сучасних діалектах української мови та у деяких інших слов'янських мовах *ся* вживають також як самостійну частку (клітику), яка зазвичай розміщується біля початку речення, або його логічної частини:

Хто по кладці мудро ступає, той ся в болоті не купає[8].

(препозиція *ся*).

За спостереженнями А.А.Зализняка, у давньоруській мові вживання *ся* розрізнялось залежно від роду тексту та часу його написання. Стародавні церковнослов'янські тексти використовують *ся* у постпозиції, але у живій розмовній мові переважала препозиція *ся*. Це видно, зокрема, з мови берестяних грамот. Мова грамот значно ближча до тодішньої живої, розмовної, ніж мова літописів та церковних пам'яток письменства. У XII-XIII ст., коли, за гіпотезою А, було створено "Слово", під впливом живої мови, вживання *ся* у препозиції зустрічалось у літературних текстах, але не часто. Зокрема, частіше зустрічається препозиція там, де автор тексту намагається передати живу мову персонажів:

Аз уже бородат, а ти ся еси родил

(Я вже був бородатим, коли ти народився) — каже у літописі князь В'ячеслав Мономахович Юрієві Долгорукому.

Досить часто препозиція зустрічається і у "Слові", наприклад:

А чи диво ся братіє стару помолодити

(а чи дивно, браття, коли старий омолоджується).

У різних граматичних конструкціях частота вживання препозиції була різною, причому вона залежала також від типу тексту: світського чи церковного. Пізніше, під впливом літературної мови, постпозиція *ся* стала загальноприйнятою і у живій мові, а препозиція збереглась лише у діалектах.

Таким чином, на кінець XVIII ст., коли "Слово" мало бути написане за гіпотезою І, стандартний правопис церковнослов'ян-

[8] Зі збірки М. Номис Українські приказки, прислівъя и таке инше (1864).

ської мови не передбачав можливості препозиції *ся*, але людина, яка хотіла б імітувати стародавній рукопис, могла бачити зразки такої препозиції в існуючих текстах. Однак зрозуміти, що причиною появ препозиції є вплив живої мови, філолог XVIII ст. не міг би, оскільки для цього потрібно було б знайомство з берестяними грамотами, які почали знаходити та аналізувати лише у XX ст. Більше того, граматичне правило, за яким у слов'янських (та інших індоєвропейських мовах) частки-клітики розташовуються на початку фрази, було виявлено Я. Вакернагелем лише наприкінці XIX ст. Таким чином, навіть найдосвіченіший філолог кінця XVIII ст. не зміг би правильно відтворити частоту вживання препозиції *ся* у фразах різного типу.

А.А. Залізняк розглядає сім різних типів (розрядів) вживання *ся* залежно від побудови фрази. Для аналізу "Слова" важливими є розряди 2-4, по яких існує достатній обсяг інформації про частотність вживання препозиції/постпозиції *ся* у фразах такого типу у XII-XIII ст. як у світській, так і у церковній мові. У табл. 9.1 наведена частина даних з таблиці на с. 67 книги [7].

Таблиця 9.1. Частоти препозиції *ся* у фразах з різних літературних джерел

	Розряд 2	Розряд 3	Розряд 4
Пряма мова	81%	81%	57%
Авторський текст	49%	12%	3%
"Слово"	1/1	3/3	6/10

Перші два рядки містять частоту фраз з препозицією *ся* (серед всіх фраз відповідного розряду) у Київському літописі за 1118–1200 роки (за Іпатіївським списком). Як відмічає А.А. Залізняк, ці частоти помітно відрізняються в авторському тексті літопису (коли літописець веде власну розповідь про події) та у тих місцях, де літописець передає пряму мову світських персонажів (напр., промови князів). Відповідно, у першому рядку таблиці містяться відносні частоти препозиції *ся* у прямій мові світських

персонажів, а у другому, авторському тексті та у промовах церковних діячів, процитованих у літописі.

Як бачимо, частота препозицій *ся* у відтвореній літописцем мові світських персонажів помітно більша, ніж у авторській мові. Це можна пояснити бажанням літописця передати особливості живої мови, які виявляються також у берестяних грамотах[9]. Якщо "Слово" є твором світської літератури XII-XIII ст., можна очікувати, що частоти препозиції *ся* у ньому будуть близькими до таких частот у книжній передачі живої мови, тобто у прямій мові Київського літопису. Частоти "Слова" вміщені у третьому рядку табл. 9.1 у аі дробів, де чисельник — кількість препозицій *ся*, знаменник — кількість всіх фраз з *ся* з даного розряду, які зустрічаються у "Слові". Вочевидь, частоти "Слова" цілком подібні частотам прямої мови літопису і сильно відрізняються від частот авторської.

Дані табл. 9.1 є сильним аргументом на користь гіпотези А про автентичність "Слова": філологи XVIII ст. не знали про особливості живої мови XII — XIII ст. і не могли б так точно відтворити частотні характеристики її передачі у письмових текстах того часу. Але як статистичний аргумент, ці міркування мають важливий недолік: малу кількість спостережень. Маємо всього 14 випадків вживання *ся*, причому на перший розряд припадає лише один випадок. Опоненти А.А. Зализняка (зокрема М.Мозер, [41]) зауважують, що при такому обсязі даних спостережуваний розподіл препозицій міг скластись подібно до живої мови XII-XIII ст. цілком випадково.

Наша мета у даному прикладі — перевірити статистичну значущість спостережень А.А. Зализняка з препозиції *ся* у "Слові". Для цього розглянемо дану задачу як перевірку статистичної гіпотези і застосуємо тест відношення вірогідності. Нехай правильною є гіпотеза А — аутентичність "Слова". Будемо вважати, що у цьому випадку для кожної фрази з відповідного розряду автор міг обрати форму з препозицією *ся* з імовірністю, яка відповідає частоті у

[9] У книзі [7] наведено значно більше статистичної інформації по різних джерелах, включаючи грамоти, але у цьому прикладі ми обмежимось лише аналізом даних табл. 9.1.

прямій мові Київського рукопису (тобто першому рядку табл. 9.1). Позначимо набір цих ймовірностей[10]:

$$\mathbf{p}_A = (p_{1,A}, p_{2,A}, p_{3,A}) = (0.81, 0.81, 0.57).$$

Припустимо, що вибір пре- або постпозиції для кожної фрази відбувався незалежно[11]. Тоді розподіл кількості фраз з препозицією у i-му розряді буде біноміальним з ймовірністю успіху p_i^0 та кількістю випробувань n_i, де

$$\mathbf{n} = (n_1, n_2, n_3) = (1, 3, 10)$$

— загальні кількості фраз відповідних розрядів у "Слові". Отже, якщо позначити ζ_i можливу кількість фраз з препозицією *ся* у "Слові" при виконанні гіпотези А, то

$$\mathsf{P}\{\zeta_i = k\} = C_{n_i}^k (p_{i,A})^k (1 - p_{i,A})^{n_i - k}.$$

Позначимо $\mathbf{k} = (k_1, k_2, k_3) = (1, 3, 6)$ — спостережувані кількості препозицій у відповідних розрядах у "Слові". Тоді, враховуючи незалежність фраз, функція вірогідності для параметра $\mathbf{p}_A$ матиме вигляд

$$L(\mathbf{p}_A) = \prod_{i=1}^{3} C_{n_i}^{k_i} (p_{i,A})^{k_i} (1 - p_{i,A})^{n_i - k}.$$

Припустимо тепер, що правильною є гіпотеза ІД (про те, що автором є чеський філолог Й. Добровський). Оскільки Добровський, перебуваючи у Санкт-Петербурзі, ознайомився з давньоруськими літописами, він міг помітити різний розподіл препозицій *ся* і (хоча на теоретичному рівні ці закономірності не були виявлені до XX

[10] Оскільки досліджуються лише розряди 2–4, нумерація розрядів у нас зсунута на 1: другому розряду відповідає ймовірність $p_{1,A}$ і т.д.

[11] Це досить сильне і, мабуть, не цілком правильне припущення, але якщо від нього відмовитись, треба буде висловити якесь припущення про ймовірнісний зв'язок між формою різних фраз. Така модель була б ще менш обґрунтованою, ніж припущення про незалежність. Тому краще вже дотримуватись його.

ст.) інтуїтивно відтворити їх у своєму творі. Але, звичайно, він не міг би виділити пряму мову світських персонажів як особливий, більш "живий" різновид, тому, що для цього у нього не було інших порівняльних матеріалів (як-от — берестяні грамоти). Тому його відтворення мало відповідати усередненому книжному стилю літопису. Оскільки пряма мова займає менше 10% тексту літопису, то цей усереднений стиль мав би приблизно такі ж частотні характеристики, як авторський текст літопису, тобто відповідав би ймовірностям

$$\mathbf{p}_I = (p_{1,I}, p_{2,I}, p_{3,I}) = (0.49, 0.12, 0.03).$$

Для того, щоб обґрунтувати статистичну значущість даних Залізняка як аргументу на користь гіпотези А проти гіпотези ІД, слід як основну гіпотезу обрати ІД. Тоді, якщо дані будуть суперечити цій гіпотезі, ми зможемо стверджувати, що вони свідчать на користь А.

Таким чином, нашою основною гіпотезою H_0 буде ІД, а альтернативою, H_1 — А. Логарифмічне відношення вірогідності має вигляд

$$\mathrm{lr} = \log\left(\frac{L(\mathbf{p}_A)}{L(\mathbf{p}_I)}\right) = S + K,$$

де

$$S = \sum_{i=1}^{3} k_i \log\left(\frac{p_{i,A}(1-p_{i,I})}{p_{i,I}(1-p_{i,A})}\right),$$

K — константа (незалежна від спостережуваних частот k_i), яка не впливає на результати тесту. Надалі як статистку нашого тесту будемо використовувати S.

Як і раніше, при значеннях S, менших або рівних порогу c_α, слід прийняти гіпотезу H_0, якщо ж S більше c_α, то H_0 слід відхилити. Поріг c_α обирається за заданим рівнем значущості α як найменше c, при якому $\mathsf{P}_{H_0}\{S > c\} \le \alpha$. Такі ймовірності можна підрахувати в явному вигляді, але ми для реалізації тесту скористаємось тією ж технікою імітаційного моделювання, яка була застосована у попередньому прикладі. А саме, згенеруємо $B = 10000$

наборів даних з розподілу, що відповідає основній гіпотезі, обчислимо на кожному наборі статистику S_j, $j = 1, \ldots, B$ і знайдемо вибірковий квантиль рівня $1 - \alpha$ отриманої вибірки з B значень S. Це і буде поріг c_α. Досягнутий рівень значущості тесту визначається як відносна частота події $S > S_j$, де S — значення нашої статистики на справжніх даних. Ця ідея реалізована у такому скрипті:

```
> set.seed(3)
> alpha<-0.05 # стандартний рівень значущості
> B<-10000    # кількість модельованих наборів даних
> m<-3        # кількість частот у наборі
> # ймовірності для авторської мови у літописі:
> pI<-c(0.49,0.12,0.03)
> # ймовірності для прямої мови світських осіб:
> pA<-c(0.81,0.81,0.57)
> # частоти у "Слові":
> k<-c(1,3,6)
> # кількості фраз з відповідних розрядів у "Слові":
> n<-c(1,3,10)
> # статистика тесту для даних "Слова":
> S<-sum(log(pA*(1-pI)/(pI*(1-pA)))*k)
> # Масив значень статистики на модельованих даних
> S0<-replicate(B,
+ {
+ k0<-rbinom(m,n,pI)
+ sum(log(pA*(1-pI)/(pI*(1-pA)))*k0)
+ })
> # поріг тесту:
> quantile(S0,1-alpha)

     95%
8.690406
```

```
> # значення статистики:
> S

[1] 34.36504

> # досягнутий рівень значущості:
> mean(S0>S)

[1] 0
```

△

Таким чином, статистика нашого тесту на реальних даних дорівнює приблизно 34.36504, поріг тесту для рівня значущості 0.05 дорівнює 8.690406. Основну гіпотезу слід відхилити, дані суперечать припущенню про авторство Добровського. Більше того, оскільки наша оцінка досягнутого рівня значущості — 0, то гіпотезу про авторство Добровського слід відхилити при будь-якому розумному рівні значущості.

Аналогічно можна розглянути гіпотезу ІБ про авторство Іоїля Биковського. Архімандрит Іоїль не був філологом, але читав багато стародавніх рукописів, тому міг помітити, що у більш древніх текстах частіше зустрічається препозиція *ся*. Якщо він хотів своєму тексту надати древнього колориту, то, мабуть, ставив би випадково *ся* у пре- та постпозиції з приблизно однаковими ймовірностями. Виходячи з цих міркувань, щоб оцінити вірогідність авторства Биковського, візьмемо значення

$$\mathbf{p}_I = (p_{1,I}, p_{2,I}, p_{3,I}) = (0.5, 0.5, 0.5)$$

і виконаємо ті ж дії, що у попередньому випадку. Отримуємо $S = 7.491148$, $c_{0.05} = 6.927445$, досягнутий рівень значущості — 0.0098. Отже, дані суперечать також і гіпотезі про те, що "Слово" є імітацією, написаною людиною XVIII ст., яка, не будучи професійним філологом, була добре обізнана зі старовинними рукописами і намагалась навмання відтворити їх стиль (як це міг би зробити Іоїль Биковський).

Таким чином, незважаючи на малий обсяг наявних даних, вони є статистично значущим аргументом на користь гіпотези про автентичність "Слова о полку Ігоревім"[12]. ◀

9.3 Тест відношення вірогідності для складних гіпотез

Випадки, коли треба перевіряти просту гіпотезу проти простої альтернативи, зустрічаються на практиці менш часто, ніж випадки перевірки складних гіпотез. У таких складних випадках тест відношення вірогідності не завжди є оптимальним, але часто — це хороший вибір. У цьому підрозділі ми спочатку розглянемо загальну схему тесту відношення вірогідності для складних гіпотез, а потім зупинимось на випадку так званих вкладених гіпотез, коли цю схему можна реалізувати порівняно просто.

9.3.1 Загальна схема тестів відношення вірогідності

Розглянемо задачу перевірки гіпотез за даними $\mathbf{X}$ з розподілом $F^{\mathbf{X}}(A) = F_\vartheta\{A\}$, де $\vartheta \in \Theta \subseteq \mathbb{R}^d$ — невідомий параметр. Основна гіпотеза H_0 полягає в тому, що $\vartheta \in \Theta_0$, альтернатива H_1: $\vartheta \in \Theta_1$, де $\Theta_i \subseteq \Theta$, Θ_1 і Θ_0 не мають спільних точок. Оскільки ми не розглядаємо інших гіпотез крім H_0 і H_1, природно вважати, що $\Theta_0 \cup \Theta_1 = \Theta$. Інколи буває зручно використовувати параметризацію, для якої це припущення не виконується, але у даному підрозділі ми таких випадків не розглядаємо.

[12] Зрозуміло, що цей аргумент не вирішує остаточно поставленого історичного питання. З цього приводу нагадаю історію, яку оповідає С.П. Новіков у своїх спогадах [14] про А.М. Колмогорова. У молодості Андрій Миколайович хотів стати істориком і навіть написав статтю про економіку древнего Новгорода, де висловив цікаву історичну гіпотезу. Коли цю статтю показали рецензенту, той сказав, що на підкріплення гіпотези наведено дуже мало доказів, треба зібрати інші. Колмогоров розчарувався в історичній науці й остаточно присвятив себе математиці, де досить навести одне доведення, щоб теорема вважалась істиною.

Тест відношення вірогідності для перевірки H_0 проти H_1 можна побудувати, коли існує така міра μ, що для всіх можливих значень $\vartheta \in \Theta_0 \cup \Theta_1$ $F_\vartheta(A)$ має щільність $f_\vartheta(x)$ відносно μ:

$$F_\vartheta(A) = \int_A f_\vartheta^{\mathbf{X}}(x)\mu(dx).$$

Відношенням вірогідності для перевірки H_0 проти H_1 за даними $\mathbf{X}$ називають

$$\mathrm{LR}^*(\mathbf{X}) = \frac{\sup_{\vartheta_1\in\Theta_1} f_{\vartheta_1}^{\mathbf{X}}(\mathbf{X})}{\sup_{\vartheta_0\in\Theta_0} f_{\vartheta_0}^{\mathbf{X}}(\mathbf{X})}.$$

Позначимо $\hat\vartheta^{(i)} = \operatorname{argmax}_{\vartheta\in\Theta_i} f_\vartheta^{\mathbf{X}}(\mathbf{X})$ — оцінка методу найбільшої вірогідності для параметра ϑ в припущенні, що виконана гіпотеза H_i. Якщо ці оцінки існують, то

$$\mathrm{LR}^*(\mathbf{X}) = \frac{f_{\hat\vartheta^{(1)}}^{\mathbf{X}}(\mathbf{X})}{f_{\hat\vartheta^{(0)}}^{\mathbf{X}}(\mathbf{X})}. \tag{9.4}$$

Таким чином, відношення вірогідності — це відношення найбільшого можливого значення функції вірогідності на альтернативі до її найбільшого значення при основній гіпотезі.

Логарифмічне відношення вірогідності — це

$$\mathrm{lr}^*(\mathbf{X}) = \log \mathrm{LR}^*(\mathbf{X}) = \log f_{\hat\vartheta^{(1)}}^{\mathbf{X}}(\mathbf{X}) - \log f_{\hat\vartheta^{(0)}}^{\mathbf{X}}(\mathbf{X}).$$

Якщо дані $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ є кратною вибіркою зі щільністю розподілу одного спостереження $f_\vartheta(x)$, то логарифмічне відношення вірогідності записується як

$$\mathrm{lr}^*(\mathbf{X}) = \sum_{j=1}^{n} \log\left(\frac{f_{\hat\vartheta^{(1)}}(\xi_j)}{f_{\hat\vartheta^{(0)}}(\xi_j)}\right).$$

Тест відношення вірогідності для H_0 проти H_1 має вигляд

$$\pi(\mathbf{X}) = \pi_C(\mathbf{X}) = \begin{cases} 1 & \text{якщо } \mathrm{LR}^*(\mathbf{X}) > C, \\ 0 & \text{якщо } \mathrm{LR}^*(\mathbf{X}) \le C, \end{cases}$$

де $C = C_\alpha$ — поріг тесту, який вибирають так, щоб рівень значущості π дорівнював заданому рівню α. Інакше кажучи, C_α — це такий поріг C, при якому

$$\sup_{\vartheta\in\Theta_0} \mathsf{P}_\vartheta\{\mathrm{LR}^*(\mathbf{X}) > C\} = \alpha.$$

Зрозуміло, що тест відношення вірогідності можна реалізувати, використовуючи логарифмічне відношення вірогідності або будь-яку іншу статистику, яка отримується з $\mathrm{LR}^*(\mathbf{X})$ монотонним перетворенням.

Зауважимо, що $\mathrm{LR}^*(\mathbf{X}) < 1$ (або, що те саме, $\mathrm{lr}^*(\mathbf{X}) < 0$) відповідає тому, що максимум функції вірогідності досягається на значенні параметра, який відповідає основній гіпотезі ($\sup_{\vartheta\in\Theta_0} f_\vartheta^{\mathbf{X}}(\mathbf{X}) > \sup_{\vartheta\in\Theta_1} f_\vartheta^{\mathbf{X}}(\mathbf{X})$), тобто основна гіпотеза є більш вірогідною, ніж альтернатива. У такому випадку відхиляти основну гіпотезу немає сенсу[13]. Тому поріг C_α доцільно завжди обирати більшим 1 (відповідно, поріг для $\mathrm{lr}^*(\mathbf{X})$ — більшим 0). Якщо ви отримали $C_\alpha < 1$, це, радше свідчить про невдалий вибір рівня значущості — ви дозволили собі занадто велику ймовірність похибки першого роду α.

Якщо дотримуватись обмеження $C > 1$, відношення вірогідності можна визначити як

$$\mathrm{LR}(\mathbf{X}) = \frac{\sup_{\vartheta_1\in\Theta} f_{\vartheta_1}^{\mathbf{X}}(\mathbf{X})}{\sup_{\vartheta_0\in\Theta_0} f_{\vartheta_0}^{\mathbf{X}}(\mathbf{X})} = \frac{f_{\hat\vartheta^{MLE}}^{\mathbf{X}}(\mathbf{X})}{f_{\hat\vartheta^{(0)}}^{\mathbf{X}}(\mathbf{X})}, \tag{9.5}$$

де $\hat\vartheta^{MLE}$ — оцінка методу найбільшої вірогідності по всьому простору можливих значень невідомого параметра $\vartheta = \Theta = \Theta_1 \cup \Theta_0$. У сучасних підручниках часто (9.5) приймають як основне означення відношення вірогідності, хоча (9.4) більш інтуітивно зрозуміле і більше відповідає традиції. Якщо $\mathrm{LR}^*(\mathbf{X}) > 1$, то

[13]Ми домовились відхиляти H_0 лише тоді, коли дані переконливо їй суперечать, а у цьому випадку дані її підтверджують.

$\mathrm{LR}(\mathbf{X}) = \mathrm{LR}^*(\mathbf{X})$. Якщо $\mathrm{LR}^*(\mathbf{X}) \le 1$, то $\mathrm{LR}(\mathbf{X}) = 1$.

Приклад 9.3.1. Розглянемо знову задачу перевірки якості електричних лампочок з прикладу 9.2.1. Там ми припускали, що тривалість роботи лампи описується експоненційним розподілом, причому якісним (брендовим) лампам відповідає інтенсивність $\lambda = \lambda_0$, а неякісним — $\lambda = \lambda_1 > \lambda_0$, і λ_1 та λ_0 вважались відомими. Такі випадки можливі, але мабуть більш поширеним є інший випадок, коли якісними вважаються лампи, що перегоряють з інтенсивністю $\lambda \le \lambda_0$, де λ_0 — деякий поріг, зафіксований у технічних умовах для даного типу ламп. Якщо $\lambda > \lambda_0$, лампи вважаються неякісними.

Нехай спостереження тривалості роботи ламп $\mathbf{X} = (\xi_1, \dots, \xi_n)$ являють собою кратну вибірку з експоненційного розподілу з невідомою інтенсивністю λ. Побудуємо тест відношення вірогідності для перевірки гіпотези $H_0 : \lambda \le \lambda_0$ (якісні лампи) проти $H_1 : \lambda > \lambda_0$ (неякісні) за даними $\mathbf{X}$.

Логарифмічна функція вірогідності для даних $\mathbf{X}$ має вигляд

$$\log f_\lambda^{\mathbf{X}}(\mathbf{X}) = n(\log\lambda - \lambda\bar{\xi}),$$

де $\bar{\xi} = \frac{1}{n}\sum_{j=1}^n \xi_j$.

Позначимо

$$\hat{\lambda}^{(0)} = \operatorname*{argmax}_{\lambda\le\lambda_0} \log f_\lambda^{\mathbf{X}}(\mathbf{X}),$$

$$\hat{\lambda}^{(1)} = \operatorname*{argmax}_{\lambda\ge\lambda_0} \log f_\lambda^{\mathbf{X}}(\mathbf{X}).$$

Оскільки функція $\log f_\lambda^{\mathbf{X}}(\mathbf{X})$ при $\lambda \in (0, +\infty)$ має єдину точку максимуму у $\lambda = 1/\bar{\xi}$, то

$$\hat{\lambda}^{(0)} = \begin{cases} 1/\bar{\xi}, & \text{якщо } 1/\bar{\xi} \le \lambda_0, \\ \lambda_0 & \text{якщо } 1/\bar{\xi} > \lambda_0, \end{cases}, \quad \hat{\lambda}^{(1)} = \begin{cases} \lambda_0 & \text{якщо } 1/\bar{\xi} \le \lambda_0, \\ 1/\bar{\xi}, & \text{якщо } 1/\bar{\xi} > \lambda_0, \end{cases}.$$

Тому

$$\mathrm{lr}^*(\mathbf{X}) = \begin{cases} n(\log\lambda_0 - \lambda_0\bar{\xi} - \log(1/\bar{\xi}) + 1) \text{ якщо } 1/\xi \le \lambda_0, \\ n(\log(1/\bar{\xi}) - 1 - \log\lambda_0 + \lambda_0\bar{\xi}) \text{ якщо } 1/\bar{\xi} > \lambda_0. \end{cases}$$

Якщо $1/\bar{\xi} < \lambda_0, \mathrm{lr}^*(\mathbf{X}) < 0$ — у цьому випадку ми домовились приймати H_0. Якщо $1/\bar{\xi} > \lambda_0$, $\mathrm{lr}^*(\mathbf{X})$ є монотонно спадною функцією $\bar{\xi}$ (це легко перевірити за знаком похідної). Отже, тест відношення вірогідності можна записати в еквівалентній формі

$$\pi(\mathbf{X}) = \begin{cases} 1 & \text{якщо } \bar{\xi} < c_\alpha, \\ 0 & \text{якщо } \bar{\xi} \geq c_\alpha, \end{cases}$$

де c_α слід обрати, виходячи з умови

$$\sup_{\lambda \leq \lambda_0} \mathsf{P}_\lambda\{\bar{\xi} < c_\alpha\} = \alpha. \tag{9.6}$$

Оскільки $n\lambda\bar{\xi}$ має гамма-розподіл з інтенсивністю 1 і показником форми n, то $\mathsf{P}_\lambda\{\bar{\xi} < c\} = \mathsf{P}_\lambda\{n\lambda\bar{\xi} < n\lambda c\} = F^{\Gamma(n,1)}(n\lambda c)$ є монотонно зростаючою функцією λ і умова (9.6) перетворюється на $F^{\Gamma(n,1)}(n\lambda_0 c_\alpha) = \alpha$, звідки

$$c_\alpha = \frac{Q^{\Gamma(n,1)}(\alpha)}{n\lambda_0}.$$

Фактично, побудований нами тест відношення вірогідності для складних H_0 і H_1 тотожний тесту, отриманому у підрозд. 9.2 для простих гіпотез. За теоремою 9.2.1 він буде найбільш потужним тестом рівня α для перевірки гіпотези $\lambda = \lambda_0$ проти гіпотези $\lambda = \lambda_1$ для всіх $\lambda_1 > \lambda_0$. Тому він є рівномірно найбільш потужним для $H_0 : \lambda \in (0, \lambda_0]$ проти $H_1 : \lambda \in (\lambda_0, +\infty)$. ◀

Цей приклад показує, що тест відношення вірогідності у випадку складних гіпотез може бути простим у реалізації і мати певні властивості оптимальності. Багато стандартних статистичних тестів отримано саме як окремий випадок застосування техніки найбільшої вірогідності. Але у загальному випадку знаходження порогу та перевірка оптимальності для таких тестів становлять досить складну проблему. У наступному підрозділі ми розглянемо важливий окремий випадок, коли для вибору порогу тесту відношення вірогідності при великій кількості спостережень можна використовувати порівняно просту асимптотичну теорію.

9.3.2 Асимптотика тесту відношення вірогідності для вкладених гіпотез

Досить часто основна гіпотеза, яку треба перевірити за даними, є частковим випадком загальної моделі даних, виокремленим додатковими умовами на невідомі параметри. Тобто дані $\mathbf{X}$ взагалі описуються певним розподілом F_ϑ, $\vartheta \in \Theta \subseteq \mathbb{R}^d$, а основна гіпотеза H_0 полягає в тому, що для ϑ виконуються додаткові умови (обмеження) вигляду $\mathbf{g}(\vartheta) = 0$, де $\mathbf{g} : \Theta \to \mathbb{R}^r$ — деяка гладенька функція:

$$H_0 : \ \vartheta \in \Theta \text{ і виконано } \mathbf{g}(\vartheta) = 0. \tag{9.7}$$

Такі гіпотези називають обмеженими.

Наприклад, нехай $\mathbf{X}$ — кратна вибірка з нормального розподілу $N(\mu, \sigma^2)$, де $\vartheta = (\vartheta_1, \vartheta_2)^T = (\mu, \sigma^2)^T \in \Theta = (\mathbb{R}, (0, +\infty))$ — невідомий параметр. Необхідно перевірити гіпотезу про те, що коефіцієнт варіації вибірки дорівнює 0.7. Це означає, що $H_0 : \sigma/\mu = 0.7$. У термінах параметра ϑ це можна записати як

$$H_0 : \ \sqrt{\vartheta_2} - 0.7\vartheta_1 = 0,$$

тобто у цьому випадку $\mathbf{g}(\vartheta) = \sqrt{\vartheta_2} - 0.7\vartheta_1$. Тут $r = 1$. В загальному випадку, r це кількість скалярних рівнянь, які задають гіпотезу H_0 (кількість обмежень).

Зазвичай набір всіх значень, які ϑ може набувати при обмеженій гіпотезі вигляду (9.7), можна описати, використовуючи набір нових параметрів $\theta = (\theta_1, \ldots, \theta_q)^T$, як

$$\Theta_0 = \{\mathbf{h}(\theta), \theta \in \Theta' \subseteq \mathbb{R}^q\}. \tag{9.8}$$

Так, в нашому прикладі з коефіцієнтом варіації можна $\mu \in \mathbb{R}$ взяти як новий параметр θ, тоді, при виконанні H_0, $\vartheta_2 = \sigma^2 = 0.49\mu^2$. Отже, у цьому випадку функція $\mathbf{h}(\theta) = (\theta, 0.49\theta^2)^T$, $\Theta' = \mathbb{R}$.

Кажуть, що гіпотеза (модель) вигляду $H_0 : \vartheta \in \Theta_0$, де Θ_0 визначено (9.8), є *вкладеною* (nested) у загальну модель $H_U : \ \vartheta \in \Theta$.

Зрозуміло, що для еквівалентності гіпотези $H_0 : \vartheta \in \Theta_0$ і гіпотези H_0, визначеної (9.7), повинно виконуватись $\mathbf{g}(\mathbf{h}(\theta)) = 0$ для всіх $\theta \in \Theta'$. Крім того, для всіх $\vartheta \in \Theta$, які не можна представити у вигляді $\vartheta = \mathbf{h}(\theta)$, повинно виконуватись $\mathbf{g}(\vartheta) \neq 0$. Якщо у системі r рівнянь $\mathbf{g}(\vartheta) = 0$ немає таких, що виражаються через інші, Θ має вимірність d а Θ' має вимірність q, то маємо $q = d - r$. (Кількість параметрів, необхідних для опису моделі, зменшується на кількість обмежень, що накладаються на цю модель).

Розглянемо задачу перевірки гіпотези $H_0 : \vartheta \in \Theta_0$ проти альтернативи $H_1 : \vartheta \in \Theta \setminus \Theta_0$, де Θ_0 визначено (9.8). (Тобто альтернатива полягає в тому, що виконана загальна модель H_U, але не виконується хоча б одне з обмежень, які задають основну гіпотезу H_0). Відношення вірогідності (логарифмічне) для перевірки цих гіпотез можна обчислювати як і у попередньому підрозділі:

$$\mathrm{lr}(\mathbf{X}) = \log \mathrm{LR}(\mathbf{X}) = \log f^{\mathbf{X}}_{\hat{\vartheta}}(\mathbf{X}) - \log f^{\mathbf{X}}_{\mathbf{h}(\hat{\theta})}(\mathbf{X}),$$

де $\hat{\vartheta}$ — оцінка найбільшої вірогідності для ϑ у загальній моделі H_U, $\hat{\theta}$ — оцінка найбільшої вірогідності для θ у вкладеній моделі $H_0 : \vartheta = \mathbf{h}(\theta)$. ($f_\vartheta(x)$ як і раніше, позначає щільність розподілу даних відносно деякої фіксованої міри μ).

Нехай дані $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ являють собою кратну вибірку обсягу n. Виявляється, що при виконанні досить загальних умов, якщо виконана H_0, то

$$\mathsf{P}\{2\mathrm{lr}(\mathbf{X}) < t\} \to F^{\chi^2_r}(t), \text{ при } n \to \infty. \tag{9.9}$$

(Тут $F^{\chi^2_r}$ — функція розподілу χ^2-розподілу з кількістю ступенів вільності r, що дорівнює кількості обмежень, які визначають основну гіпотезу).

Таким чином, тест відношення вірогідності

$$\pi(\mathbf{X}) = \begin{cases} 1 & \text{якщо } 2\mathrm{lr}(\mathbf{X}) > c_\alpha, \\ 0 & \text{якщо } 2\mathrm{lr}(\mathbf{X}) \leq c_\alpha \end{cases}$$

з $c_\alpha = Q^{\chi^2_r}(1-\alpha)$ буде мати асимптотичний рівень значущості α.

Відповідно, досягнутий рівень значущості цього тесту можна визначити як

$$p(\mathbf{X}) = 1 - F^{\chi^2_r}(2\mathrm{lr}(\mathbf{X})).$$

Строге формулювання умов, за яких виконується (9.9), можна знайти у теоремі 6.5 книги [46]. Опишемо основні з них неформально:

1. Функція $\mathbf{h}$ має бути неперервно диференційовною.

2. Справжнє значення параметра ϑ, що відповідає розподілу даних, повинно бути внутрішньою точкою Θ. Тобто для деякого ε всі вектори в $\mathbb{R}^d$, що лежать на відстані від ϑ, меншій ніж ε, також мають належати Θ. Аналогічно, θ, для якого $\vartheta = \mathbf{h}(\theta)$, повинно бути внутрішньою точкою Θ'.

3. Інформаційна матриця для ϑ за спостереженням ξ_j у загальній моделі має існувати і бути невиродженою. Аналогічно, має бути невиродженою інформаційна матриця для θ у обмеженій моделі.

4. Мають виконуватись додаткові умови регулярності з теореми 4.36 у [46]. (Ці умови виконуються у більшості звичайних ймовірнісних моделей з щільностями, двічі неперервно диференційовними за невідомими параметрами).

Приклад 9.3.2. Розглянемо знову дані про тривалість роботи вальниць з прикладу 8.6.2. Як ми бачили, розподіл часу роботи вальниць добре описується розподілом Вейбулла. У підручниках параметр `shape` цього розподілу (вейбулів параметр) при розрахунку надійності сталевих кулькових вальниць рекомендують приймати рівним 1.3. У прикладі 8.6.2 ми оцінили `shape` за даними й отримали значення 2.10221. Це помітно відрізняється від 1.3. Але ця оцінка побудована лише за 23 спостереженнями. Чи є її відхилення від "стандартного" значення істотним, чи це наслідок випадкового розкиду даних і ним можна знехтувати? Для перевірки застосуємо тест найбільшої вірогідності.

Основна гіпотеза полягає в тому, що значення вейбулового параметра є стандартним: H_0 : `shape=1.3`, альтернатива — H_1 : `shape`$\neq 1.3$. При цьому параметр `rate` невідомий і може набувати

будь-яких додатних значеннь. Отже, загальна модель H_U (розподіл Вейбулла з повністю невідомими параметрами) визначається двома параметрами ($d = 2$), а у гіпотезі H_0 накладається одне обмеження ($r = 1$). Подвоєну логарифмічну статистику відношення вірогідності обчислюємо, використовуючи функцію `fitdistr()`, описану у п. 8.6. Як ми пам'ятаємо, її результатом є об'єкт, що в атрибуті `loglik` містить максимальне досягнуте значення логарифмічної функції вірогідності. Отже,

```
> library(reliaR)
> library(MASS)
> data("bearings")
> options(warn=-1)
> # Підгонка у загальній моделі:
> fitted<-fitdistr(bearings, "weibull")
> # Підгонка у обмеженій моделі:
> fitted1<-fitdistr(bearings, "weibull",shape=1.3)
> # Подвоєне логарифмічне відношення вірогідності:
> lr2<-2*(fitted$loglik-fitted1$loglik)
> # Досягнутий рівень значущості:
> p_level=1-pchisq(lr2,1)
> cat("Статистика=",lr2,";
+     Досягнутий рівень зн.=",p_level)

Статистика= 7.367821 ;
    Досягнутий рівень зн.= 0.006640136
```

Таким чином, досягнутий рівень значущості тесту відношення вірогідності на наших даних — 0.006640136, менше 1%, тому відхилення оцінки вейбуллового параметра від стандартного значення слід вважати значущим. ◀

9.3.3 Багатовибіркові задачі

Хоча у попередньому підрозділі асимптотичне співвідношення (9.9) введене для даних, що утворюють одну кратну вибірку, його можна також застосовувати і для перевірки гіпотез у випадку, коли дані складаються з кількох кратних вибірок, причому розподіли у різних вибірках можуть бути різними. Задачі перевірки гіпотез, пов'язані з такими даними, називають *багатовибірковими*.

Найбільш поширений приклад багатовибіркових задач — перевірка однорідності. У таких задачах вважається, що для всіх вибірок виконується одна загальна модель розподілу, але кожній вибірці можуть відповідати свої особливі значення невідомих параметрів цієї моделі. Задача полягає у перевірці гіпотези H_0 про те, що значення параметрів однакові для всіх вибірок (тобто що вибірки однорідні). Розглянемо приклад застосування тесту відношення вірогідності для перевірки такої гіпотези.

Приклад 9.3.3. Повернемось до аналізу задачі, описаної у прикладі 8.1.5. Розглядаються дані гарантійної майстерні про кількість дефектів на жорстких дисках, виготовлених трьома фірмами A, B і C. Розіб'ємо дані на три вибірки, по кожній фірмі-виробнику окремо. Розподіл даних у кожній вибірці — пуассонів зі зрізаним нулем. Інтенсивності цього розподілу $(\lambda_A, \lambda_B, \lambda_C)$ для різних вибірок можуть відрізнятись. Перевіряється гіпотеза про те, що насправді всі інтенсивності однакові: $H_0 : \lambda_A = \lambda_B = \lambda_C$. Застосуємо для перевірки H_0 тест відношення вірогідності.

Для цього треба обчислити найбільшу логарифмічну вірогідність в умовах загальної моделі (різні інтенсивності). Вона дорівнюватиме сумі максимальних значень логарифмічних вірогідностей для кожної вибірки окремо. Далі, щоб знайти максимум логарифмічної вірогідності при H_0, треба об'єднати всі три вибірки в одну і підганяти пуассонову модель зі зрізаним нулем по об'єднаній вибірці. Різниця між найбільшим значенням при H_0 і найбільшим значенням для загальної моделі буде потрібною нам статистикою логарифмічного відношення вірогідності.

Згідно з прикладом 8.3.5 оцінки найбільшої вірогідності для цієї моделі ті ж, що й отримані у прикладі 8.1.5 оцінки методу моментів. Отже, визначимо:

```
> library(nleqslv)
> # Вводимо дані для аналізу:
> x<-1:6 # кількість дефектів на одному диску
> # частоти дисків з такою кількістю дефектів
> # серед продукції фірм A, B і C:
> A<-c(20,13,11,6,2,0)
> B<-c(25,20,15,7,0,1)
> C<-c(33,16,4,1,0,0)
> # моментна оцінка
> # перший момент як функція від інтенсивності l
> # мінус m - емпіричний момент
> moment<-function(l,m){
+ l/(1-exp(-l))-m
+ }
> # Оцінка інтенсивності
> # x - вектор значень спостережуваної змінної
> # w - частоти значень у вибірці
> EstP<-function(x,w){
+ m<-weighted.mean(x,w)
+ nleqslv(m,moment,m=m)$x
+ }
> # логарифмічна функція вірогідності
> # lambda - аргумент функції (інтенсивність)
> # x - вектор значень спостережуваної змінної
> # w - частоти значень у вибірці
> llP<-function(lambda,x,w){
+ m<-weighted.mean(x,w)
+ n<-sum(w)
+ K<-length(x)
+ n*(m*log(lambda)-log(exp(lambda) -1))
```

```
+   -sum(w*log(factorial(1:K)))
+ }
```

Обчислимо подвоєне логарифмічне відношення вірогідності:

```
> 2*(llP(EstP(x,A),x,A)+llP(EstP(x,B),x,B)
+    +llP(EstP(x,C),x,C)-llP(EstP(x,A+B+C),x,A+B+C))

[1] 0
```

Цю статистику треба порівняти з пороговим значенням, яке є квантилем розподілу χ^2 з $r = 3 - 1 = 2$ ступенями вільності, оскільки у загальній моделі підганялись 3 параметри, а у обмеженій H_0 — один. (Це також зрозуміло з того, що H_0 задається двома рівняннями-умовами: $\lambda_A = \lambda_B$, $\lambda_B = \lambda_C$). Для $\alpha = 0.05$ маємо:

```
> qchisq(1-0.05,2)

[1] 5.991465
```

Оскільки статистика більша, ніж поріг, то слід прийняти альтернативу: вибірки не є однорідними, значення параметрів суттєво відрізняються. ◄

Аналізувати багатовибіркові дані за цією схемою доцільно, коли розміри всіх вибірок приблизно однакового порядку. Якщо, наприклад, одна з вибірок у 1000 разів більша, ніж інші, то значення параметрів, отримані оцінкою по цій вибірці, краще вважати встановленими точно (оскільки похибка їх оцінки буде несуттєвою порівняно з похибками оцінок за іншими вибірками).

9.4 Тести за довірчими інтервалами

У підрозд. 8.5 ми розглядали застосування довірчих інтервалів для перевірки гіпотез про невідомі параметри розподілу даних.

Розберемо це питання з позицій загальної теорії перевірки статистичних гіпотез, якій присвячено даний розділ.

Нехай $\vartheta \in \mathbb{R}$ — невідомий числовий параметр розподілу даних $\mathbf{X}$. За даними побудовано довірчий інтервал $[\vartheta^-(\alpha), \vartheta^+(\alpha)]$ для для цього параметра, який покриває справжнє значення ϑ із заданою ймовірністю $1 - \alpha$:

$$\mathsf{P}\{\vartheta \in [\vartheta^-(\alpha), \vartheta^+(\alpha)]\} = 1 - \alpha.$$

(Зрозуміло, що $\vartheta^+(\alpha)$ і $\vartheta^-(\alpha)$ є також функціями від $\mathbf{X}$, хоча ми цього явно не записуємо.)

Треба перевірити гіпотезу H_0 про те, що $\vartheta \in \Theta_0$, де $\Theta_0 \subset \mathbb{R}$ — деяка фіксована множина. Для цього можна скористатись довірчим інтервалом, застосувавши такий тест:

$$\pi(\mathbf{X}) = \begin{cases} 0 & \text{якщо } [\vartheta^-(\alpha), \vartheta^+(\alpha)] \cap \Theta_0 \neq \varnothing, \\ 1 & \text{якщо } [\vartheta^-(\alpha), \vartheta^+(\alpha)] \cap \Theta_0 = \varnothing, \end{cases}$$

тобто

> Тест приймає основну гіпотезу, якщо довірчий інтервал перетинає множину Θ_0 і відхиляє її, якщо перетину немає.

Інтуїтивний зміст тесту очевидний: будь-яка спільна точка довірчого інтервалу і Θ_0 — кандидат на роль справжнього значення ϑ, для якого виконується основна гіпотеза. Якщо таких кандидатів немає, H_0 природно відхилити.

Яким буде рівень значущості такого тесту? Щоб зрозуміти це, оцінимо ймовірність похибки першого типу: тест відхиляє H_0, коли вона правильна. Якщо H_0 виконана, то $\vartheta \in \Theta_0$ і

$$\mathsf{P}\{[\vartheta^-(\alpha), \vartheta^+(\alpha)] \cap \Theta_0 = \varnothing\} \leq \mathsf{P}\{\vartheta \notin [\vartheta^-(\alpha), \vartheta^+(\alpha)]\} = \alpha.$$

Отже, поклавши $\alpha = \alpha_0$, де α_0 — заданий стандартний рівень значущості, отримаємо тест з рівнем, який не перевищує стандартного. Тобто наш тест π задовольняє мінімальну вимогу до тестів

(9.1), висунуту у підрозд. 9.1. Не очевидно, чи буде цей тест найкращим можливим, але за відсутності кращих, користуватись ним можна.

У прикладі 8.5.3 описана більш складна техніка перевірки гіпотез, яка використовує декілька довірчих інтервалів. А саме, ми розглядаємо m невідомих параметрів ϑ_i, $i = 1, \dots, m$ і для кожного маємо довірчий інтервал $[\vartheta_i^-(\alpha), \vartheta_i^+(\alpha)]$, тобто

$$\mathsf{P}\{\vartheta_i \in [\vartheta_i^-(\alpha), \vartheta_i^+(\alpha)]\} = 1 - \alpha \text{ для всіх } i \in 1 \dots, m. \quad (9.10)$$

Треба перевірити гіпотезу $H_0 : \vartheta_1 = \vartheta_2 = \cdots = \vartheta_m$ (однорідність). Для цього застосовується такий тест:

$$\pi^m(\alpha) = \begin{cases} 0 & \text{якщо } \bigcap_{i=1}^m [\vartheta_i^-(\alpha), \vartheta_i^+(\alpha)] \neq \varnothing, \\ 1 & \text{якщо } \bigcap_{i=1}^m [\vartheta_i^-(\alpha), \vartheta_i^+(\alpha)] = \varnothing, \end{cases}$$

— ми приймаємо гіпотезу про рівність всіх параметрів, якщо є точка, яка належить всім довірчим інтервалам, і відхиляємо, якщо такої точки немає. Знову інтуїтивний зміст зрозумілий — спільна точка всіх інтервалів є природним кандидатом на роль єдиного спільного значення для всіх параметрів.

На жаль, для цього тесту рівень значущості може бути більшим, ніж α. Дійсно, при виконанні H_0 досить, щоб один з невідомих параметрів не був покритий відповідним довірчим інтервалом і тест може прийняти альтернативу. А (9.10) гарантує лише те, що для кожного інтервалу окремо ймовірність покриття відповідного ϑ_i буде дорівнювати α. Такі набори довірчих інтервалів називають неодночасними.

Щоб перевіряти гіпотези за тестом $\pi^m(\alpha)$ із заданим рівнем значущості, потрібен набір **одночасних** довірчих інтервалів, тобто таких інтервалів $[\tilde{\vartheta}_i^-(\alpha), \tilde{\vartheta}_i^+(\alpha)]$, для яких

$$\mathsf{P}\left[\bigcap_{i=1}^m \{\vartheta_i \in [\tilde{\vartheta}_i^-(\alpha), \tilde{\vartheta}_i^+(\alpha)]\}\right] = 1 - \alpha$$

— ймовірність того, що всі довірчі інтервали одночасно покривають відповідні параметри дорівнює $1-\alpha$ (або, для нестрогого набору довірчих інтервалів, не перевищує $1-\alpha$).

Існують декілька способів побудови таких одночасних наборів. Тут ми розглянемо два.

Перший спосіб застосовується, якщо можна побудувати набір неодночасних довірчих інтервалів, які задовольняють (9.10) і є незалежними між собою. (Напр., якщо кожен з цих інтервалів будується по окремій вибірці і ці вибірки незалежні, як у прикладі 8.5.2). Тоді для того, щоб отримати набір одночасних довірчих інтервалів з рівнем значущості α_0, можна взяти неодночасні, але з меншим рівнем

$$\alpha = 1 - \sqrt[m]{1-\alpha_0}, \tag{9.11}$$

де m — кількість інтервалів у наборі.

Справді, внаслідок незалежності:

$$\mathsf{P}\left[\bigcap_{i=1}^{m}\{\vartheta_i \in [\vartheta_i^-(\alpha), \vartheta_i^+(\alpha)]\}\right] = \prod_{i=1}^{m}\mathsf{P}\left\{\vartheta_i \in [\vartheta_i^-(\alpha), \vartheta_i^+(\alpha)]\right\} =$$

$$= \prod_{i=1}^{n}(1-\alpha) = 1-\alpha_0.$$

Таким чином, набір $[\vartheta_i^-(\alpha), \vartheta_i^+(\alpha)]$ є набором одночасних довірчих інтервалів рівня α_0.

Другий спосіб можна використовувати і тоді, коли неодночасні довірчі інтервали є залежними. У цьому випадку вибираємо

$$\alpha = \frac{\alpha_0}{m.} \tag{9.12}$$

(Такий підхід називають **методом Бонферроні**). При такому виборі α:

$$\mathsf{P}\left[\bigcap_{i=1}^{m}\{\vartheta_i \in [\vartheta_i^-(\alpha), \vartheta_i^+(\alpha)]\}\right] = 1-\mathsf{P}\left[\bigcup_{i=1}^{m}\{\vartheta_i \notin [\vartheta_i^-(\alpha), \vartheta_i^+(\alpha)]\}\right] \geq$$

$$\geq 1 - \sum_{i=1}^{m} \mathsf{P}\{\vartheta_i \notin [\vartheta_i^-(\alpha), \vartheta_i^+(\alpha)]\} = 1 - \alpha_0.$$

Метод Бонферроні дає більш широкі довірчі інтервали, ніж метод, що використовує незалежність. Тому при його застосуванні важче помітити відхилення від основної гіпотези, коли воно насправді є. Таким чином, метод Бонферроні доцільно використовувати лише тоді, коли початкові неодночасні довірчі інтервали є залежними. Хоча, при стандартному рівні значущості $\alpha_0 = 0.05$ і невеликій кількості інтервалів, відмінність між цими двома типами інтервалів на око майже непомітна.

Приклад 9.4.1. Зобразимо набір одночасних довірчих інтервалів для інтенсивностей зрізаного пуасонового розподілу з прикладу 8.1.5. На рис. 9.2 вони зображені червоним кольором (суцільною лінією) разом з неодночасними довірчими інтервалами (чорним кольором, переривчастою лінією), отриманими у прикладі 8.5.2.

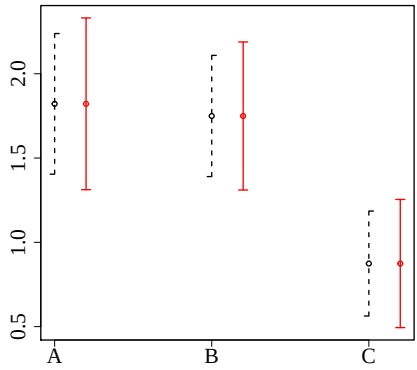

Рис. 9.2. Довірчі інтервали для інтенсивностей (одночасні — суцільною лінією, не одночасні — переривчастою)

Програма на R для відображення цього рисунка майже та сама, що у прикладі 8.5.2, лише α для одночасних інтервалів обчислює-

ться за (9.11) з $\alpha_0 = 0.05$. Як бачимо, довірчі одночасні інтервали трохи ширші, ніж неодночасні, але у даному випадку це не вплинуло на висновок про неоднорідність вибірок: значення λ для вибірки C суттєво відрізняється від λ для A і B, які можуть бути однаковими[14].

Цей приклад ілюструє перевагу тесту на основі довірчих інтервалів над тестом відношення вірогідності, який ми застосували у прикладі 9.3.3 . Довірчі інтервали дозволяють не тільки виявити неоднорідність вибірок, а і зрозуміти, яка саме вибірка відрізняється від інших. Можна сказати, що за допомогою набору одночасних довірчих інтервалів ми перевіряємо не пару гіпотез — H_0 : однорідність проти H_1 : неоднорідність, а цілий набір з 5 гіпотез:

$H_0 : \lambda_A = \lambda_B = \lambda_C$;

$H_A : \lambda_A \neq \lambda_B = \lambda_C$;

$H_B : \lambda_A = \lambda_C \neq \lambda_B$;

$H_C : \lambda_A = \lambda_B \neq \lambda_C$;

і H_{ABC} : всі λ — різні. ◀

Робити перевірку багатьох гіпотез за допомогою картинки з довірчими інтервалами значно зручніше, ніж перевіряти такі гіпотези попарно, використовуючи тести, подібні до тесту відношення вірогідності. Але при цьому слід мати на увазі, що і рисунки із довірчими інтервалами не завжди вдається інтерпретувати просто і однозначно. Наприклад, якщо рисунок з довірчими інтервалами має такий вигляд, як на рис. 9.3, то можна зробити висновок, що вибіркам A і C відповідають різні значення параметра, а параметр для вибірки B може бути рівним або параметру для A, або параметру для C. (Обом одразу не може, бо вони різні!) Якому? Відповісти на це за довірчими інтервалами неможливо.

[14] Зрозуміло, що цей висновок є ghfdbkmybv лише за умови, що модель пуасонового розподілу зі зрізаним нулем адекватно описує дані. Для перевірки цього можна скористатись тестом χ^2, див. приклад 9.6.2.

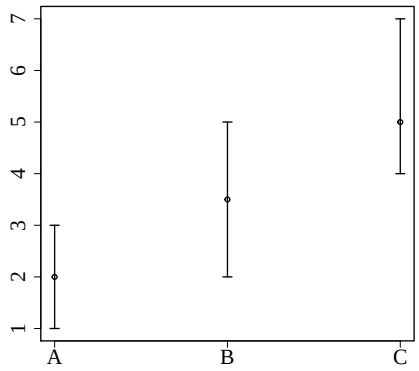

Рис. 9.3. Довірчі інтервали з невизначеним висновком

З іншого боку, тест відношення вірогідності має певні переваги над висновками за довірчими інтервалами. Він не дозволяє сказати, в чому саме полягає неоднорідність, але саму неоднорідність вловлює акуратніше, ніж це можна зробити за довірчими інтервалами. Тест помічає такі малі відхилення, яких довірчі інтервали не помітять. Тому при аналізі реальних даних доцільно поєднувати проведення числових тестів на зразок тесту відношення вірогідності із графічними засобами статистичного аналізу, як довірчі інтервали.

За тією ж логікою можна використовувати й інші довірчі множини, наприклад, довірчі еліпсоїди. Зокрема, у прикладі 8.5.3 при застосуванні довірчих еліпсоїдів для перевірки гіпотези про одночасну однорідність і математичних сподівань, і дисперсій доцільно зменшити номінальний рівень значущості еліпсоїдів за формулою (9.11).

Довірчі еліпсоїди можна також використовувати для перевірки більш складних гіпотез.

Приклад 9.4.2. Нехай у медичній лабораторії впроваджується нова методика вимірювання швидкості осідання еритроцитів

(ШОЕ) у пробах крові. Стандартна методика, що використовувалась раніше, забезпечує коефіцієнт варіації вимірювань 10% (тобто CV=0.1). Треба перевірити, чи гарантує нова методика такий самий або кращий (менший) CV.

Для цього одну пробу крові розділили на 6 окремих зразків і по кожному визначили ШОЕ за новою методикою. Отримали такі результати (у міліметрах на годину[15]):

N	1	2	3	4	5	6
ШОЕ	7.3	7.8	8.2	7.2	8.1	8.

За цими даними легко обчислити $\mathrm{CV}(\mathbf{X}) = 0.05442358$. Це майже вдвічі краще, ніж за стандартною методикою. Але наша вибірка складається лише з 6 спостережень. Чи можемо ми покладатись на висновок з такої малої кількості даних?

Припустимо, що коливання вимірювань мають нормальний розподіл з невідомими нам математичним сподіванням μ і середньоквадратичним відхиленням σ. Нам треба перевірити гіпотезу $H_0 : \mathrm{CV} = \sigma/\mu > 0.1$ (Як основну слід взяти саме цю гіпотезу, а не протилежну, якщо ми хочемо своїми даними переконати замовників в тому, що $\mathrm{CV} \leq 0.1$). У координатах (μ, σ) область, що відповідає H_0, буде півплощиною, яка лежить вище прямої $\sigma > 0.1\mu$.

Зобразимо цю півплощину і довірчий еліпсоїд для (μ, σ) із заданим рівнем значущості $\alpha_0 = 0.05$. Переконливо відхилити H_0 можна лише тоді, коли цей еліпсоїд не має спільних точок з півплощиною. (Ймовірність такої події при виконанні H_0 не більше α_0,

[15] У стандартній методиці Вестергрена ШОЕ визначається як ширина прошарку чистої плазми над еритроцитами, що осіли за годину і вимірюється на око, з точністю до міліметра. Але сучасні автоматичні методи можуть давати числа з формально більшою точністю. Наскільки така підвищена точність може бути корисною для реальної діагностики — питання непросте. Зокрема, якщо коливання результатів вимірювань для одного зразка такі, як у нас, знаки після коми навряд чи є значущими.

оскільки вона можлива лише тоді, коли справжнє значення параметрів опиняється за межами еліпсоїда). Якщо еліпсоїд і півплощина мають хоча б одну спільну точку, H_0 слід прийняти.

Для зображення еліпсоїда використаємо введену нами у прикладі 8.5.3 функцію `ConfEllipse()`. Півплощину зобразимо функцією `polygon()`:

```
> library(plotrix)
> x<-c(7.3, 7.8, 8.2, 7.2, 8.1, 8.)
> plot(c(5,10),c(0.1,1),type="n",xlab="mean",ylab="sd")
> points(mu,sigma)
> ConfEllipse(x,col="blue",density=15)
> polygon(c(4,11,11,4),c(0.4,1.1,1.1,1.1),
+         col="red",density=5)
> abline(0,0.1,lwd=2)
```

Результат відображено на рис. 9.4. Довірчий еліпсоїд (синій) та множина значень параметрів, що відповідає H_0 (заштрихована червоним), не перетинаються. Отже, H_0 слід відхилити, тобто коефіцієнт варіації нової методики менший 0.1.

Тут корисно зробити зауваження. По-перше, довірчий еліпсоїд, який ми побудували, є асимптотичним, йому можна довіряти лише при великих обсягах вибірок. 6 спостережень навряд чи можна назвати "великим обсягом".

По-друге, нормальний розподіл коливань значень вимірювань взятий нами "зі стелі". Дійсно, при правильній побудові методики вимірювання, похибки вимірювань складаються внаслідок дій багатьох неконтрольованих факторів і серед цих факторів не повинно бути "домінуючих", таких, які переважають всі інші[16]. Усереднення впливів таких факторів за центральною граничною теоремою

[16] Якщо домінуючі фактори є, правильна методика повинна враховувати і контролювати їхню дію. Скажімо, якщо на осідання еритроцитів великий вплив має температура навколишнього середовища, то треба робити відповідну поправку на температуру при визначенні результату.

має давати приблизно нормальний розподіл. Але ми досліджуємо результати вимірювання за нестандартною методикою. Хто може гарантувати нормальність розподілу її похибок?

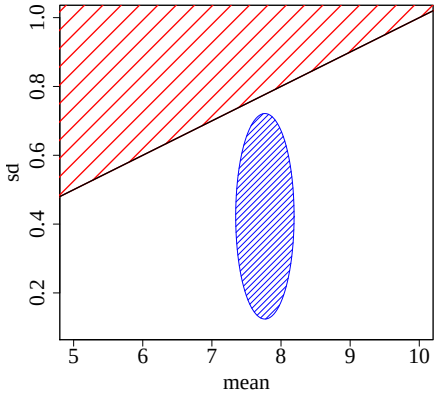

Рис. 9.4. Довірчий еліпсоїд для даних ШОЕ

Отже, наш висновок про виконання умови CV<0.1 можна обґрунтовано критикувати. Проте рис. 9.4 додає впевненості в тому, що нова методика визначення ШОЕ є перспективною. ◄

9.5 Тести для нормального розподілу

Цей підрозділ присвячено тестам, які застосовуються для перевірки гіпотез про математичні сподівання або дисперсії даних, що за припущенням мають нормальний (гауссів) розподіл. Наприклад, для математичного сподівання розподілу однієї кратної вибірки можна розглядати задачу перевірки гіпотези про те, що справжнє значення математичного сподівання μ дорівнює даному фіксованому числу μ_0, або що воно не перебільшує μ_0. Зазвичай у таких задачах дисперсія σ^2 розподілу невідома, але бувають і задачі, в яких точне значення σ^2 відоме.

Якщо спостереження складаються з двох вибірок, то може виникнути питання про те, чи є однаковими їхні математичні сподівання (гіпотеза однорідності середніх[17]). Дисперсії при цьому можуть бути відомими для обох вибірок або невідомими. Якщо треба робити висновки про математичне сподівання, а дисперсія невідома, то її називають заважаючим параметром у цій задачі.

Аналогічно можна сформулювати відповідні гіпотези для дисперсій. Наприклад, для двовибіркових даних часто виникає задача перевірки гіпотези про те, що дисперсії обох вибірок однакові. Для перевірки гіпотез про дисперсії математичне сподівання буде заважаючим параметром, якщо воно невідоме.

Для перевірки гіпотез такого роду ще у першій половині XX ст. розроблений набір стандартних тестів, які прийнято вважати оптимальними. Більшість з них є варіантами тесту відношення вірогідності для відповідної задачі. Зазвичай вони є рівномірно найбільш потужними серед тестів, які природно застосовувати у даній задачі перевірки гіпотез. У тих випадках, коли відповідного рівномірно найбільш потужного тесту не існує, стандартний тест дає результати розумно близькі до найкращих можливих у тих випадках, яких природно сподіватись у практичних задачах. Тому немає сенсу намагатись розробляти свої власні тести у даній області, якщо тільки ви не є спеціалістом саме з цього типу тестів і не хочете довести перевагу свого підходу іншим спеціалістам в галузі статистичної теорії. Спеціалісти з прикладної статистики та практики у предметних областях, де застосовується статистика, все одно не повірять у доречність застосування нестандартного тесту до нормально розподілених даних.

Всі тести, описані у даному підрозділі, крім тесту Велча — Сатерсвайта, є неасимптотичними, тобто їх можна застосовувати для всіх обсягів вибірок, більших 1. Можливість робити висновки

[17]Математичне сподівання теоретичного розподілу часто трактують, як середнє значення досліджуваної величини у генеральній сукупності, з якої відібрано вибірку. Тому гіпотезу про рівність математичних сподівань також називають гіпотезою про однорідність середніх.

всього лиш по чотирьох (а то і по двох) спостереженнях — велика перевага. Але слід розуміти, що вона ґрунтується на апріорному припущенні про нормальність розподілу цих спостережень. Якщо воно неправильне, висновки можуть бути цілком хибними.

Для вибірок великих обсягів припущення про нормальність не є критичним, якщо перевіряється гіпотеза про математичні сподівання. Висновки відповідних стандартних тестів будуть достатньо надійними завдяки центральній граничній теоремі[18]. Для стандартних тестів щодо дисперсій це не правильно — на не нормальних даних вони можуть давати хибні результати навіть при великих обсягах вибірок.

Далі ми спочатку подивимось, як у `R` перевіряти гіпотези, використовуючи стандартні функції, а лише потім, на основі яких формул працюють ці функції. Таким чином, читач, якого не цікавлять деталі, може обмежитись знайомством з технікою використання стандартних функцій `R` у підрозд. 9.5.1 та 9.5.2, не входячи у розгляд математики, на яку вони спираються.

9.5.1 T-тест. Перевірка гіпотез про середні. Дисперсія — заважаючий параметр

Для того, щоб перевіряти гіпотези про значення математичного сподівання нормального розподілу, у випадку, коли дисперсія невідома, можна використовувати функцію `t.test`. Вона реалізує різні версії T-тесту Стьюдента.

Функція `t.test` має такі параметри:

`x`, `y` — набори даних. Якщо перевіряється гіпотеза про одну вибірку, досить вказати тільки `x`, а `y` не задають. При порівнянні двох вибірок одну з них вміщують в `x`, а другу — в `y`. Можливий також варіант запису двовибіркової задачі у вигляді формули. Тоді перший параметр функції записують як `data~factor`, де `data` — числовий вектор, що містить першу і другу вибірки, `factor` —

[18] Зрозуміло, що для цього розподіл даних повинен мати скінченну дисперсію, інакше центральна гранична теорема виконуватись не буде.

вектор факторів тієї ж довжини, що `data`, елементи якого можуть набувати лише двох різних значень, які вказують, до якої вибірки належить відповідний елемент `data`.

`alternative` — вказує альтернативу, яка перевіряється. Стандартно це `"two.sided"` — двостороння альтернатива. Тобто основною гіпотезою вважається $H_0 : \mu = $`mu`, де `mu` — число, задане у відповідному параметрі функції, а $H_1 : \mu \neq$`mu`. У цьому випадку тест помічає відхилення від основної гіпотези в обидва боки — у більший і менший.

Якщо обрати `alternative="greater"`, то тест для однієї вибірки `x` з невідомим математичним сподіванням μ буде перевіряти гіпотезу $H_0 : \mu \leq$`mu`, проти альтернативи $H_1 : \mu >$`mu`, тобто помічатиме лише відхилення у більший бік, а всі значення $\mu <$`mu` розглядатиме як відповідні основній гіпотезі. При `alternative="less"` навпаки, $H_1 : \mu <$`mu`, проти $H_0 : \mu \geq$`mu`. (Досить вказати лише першу літеру назви альтернативи — `t`, `l` або `g`).

Якщо задано дві вибірки[19] `x` та `y` з невідомими математичними сподіваннями μ_x і μ_y, то перевіряються гіпотези про різницю $\mu_y - \mu_x$, аналогічні тим, що розглянуті вище для одновибіркового тесту. Наприклад, при `alternative="greater"`, альтернатива має вигляд $H_1 : \mu_y > \mu_x +$`mu`, а основна гіпотеза — $H_0 : \mu_y \leq \mu_x +$`mu`.

`mu` — як ми тільки що розібрали, це параметр, де вказується значення, або одностороння межа для μ в одновибіркових тестах, або для $\mu_y - \mu_x$ — у двовибіркових.

`paired` — цей параметр використовується тільки у двовибіркових тестах. Для вибірок `x` та `y` можливі два випадки: вони можуть містити дані про два набори різних об'єктів, не пов'язаних між собою (незалежні або непов'язані вибірки), або у `x` можуть бути дані про деяку характеристику досліджуваних об'єктів, у `y` — про іншу характеристику цих самих об'єктів (залежні, або "пов'язані вибірки"). Тести для перевірки рівності математичних сподівань

[19] Або формула, за якою дані розбиваються на дві такі вибірки.

різні для зв'язаних і незв'язаних вибірок. Тести для зв'язаних вибірок називаються "парними тестами". Якщо вам потрібен парний тест, обирайте `paired=TRUE`. Якщо вибірки непов'язані — `paired=FALSE`. (Стандартно `FALSE`).

`var.equal` — організація двовибіркового тесту суттєво різна залежно від того, чи вважаємо ми однаковими (невідомі) дисперсії в обох вибірках. Якщо є підстави вважати дисперсії однаковими, вибираємо `var.equal=TRUE`, тоді тесту буде легше помітити відмінність математичних сподівань. Якщо впевненості немає, краще обрати `var.equal=FALSE` — такий вибір не буде помилкою у будь-якому випадку (це стандартний варіант[20]).

`conf.level` — крім перевірки гіпотези тест також вказує довірчий інтервал для математичного сподівання ($\mu_y - \mu_x$ для двовибіркових тестів). Цей параметр задає його надійність ($1 - \alpha_0$). Якщо обрана двостороння гіпотеза, будується звичайний довірчий інтервал, якщо одностороння — вказується тільки одна одностороння межа, яка дозволяє перевірити відповідну гіпотезу (як другий кінець інтервалу, вказують $\pm$`Inf`).

Результатом виконання функції є об'єкт, що має такі атрибути:

`statistic` і `parameter` — значення статистики тесту (це T-статистика, див. п. 9.5.5) і кількість ступенів вільності для неї;

`p.value` — значення досягнутого рівня значущості тесту (порівняти його із стандартним і прийняти ту чи іншу гіпотезу ви маєте самі);

`conf.int` — межі довірчого інтервалу.

Результат функції `t.test()` має ще кілька атрибутів, що відіграють допоміжну роль, їх можна подивитись у help.

Приклад 9.5.1. В автоматі, що розливає каву по склянках, є дозатор, який регулює об'єм налитої кави. Ми встановили його

[20] Але слід мати на увазі, що при `var.equal=TRUE` використовується не асимптотичний Т-тест Стьюдента, який можна застосовувати при будь-яких обсягах вибірки більше 2. При `var.equal=FALSE` використовується тест Велча — Сатерсвайта, який є наближеним і дає хороші результати лише при достатньо великих обсягах вибірки.

на поділку 100 мл і хочемо перевірити, чи правильно він працює. Для цього п'ять разів замовили автомату каву і кожного разу виміряли об'єм, який він налив. Отримали значення: 95, 90, 105, 90, 85. Бачимо, що недолив зустрічається значно частіше, ніж перевищення замовленого об'єму. Чи свідчать ці дані про систематичне відхилення середнього об'єму кави, яку наливає автомат, від номінального?

Для перевірки використаємо T-тест. Як основну гіпотезу візьмемо припущення, що автомат працює правильно[21] — математичне сподівання об'єму налитої у склянку кави $H_0 : \mu = 100$. Перевіряємо наявність відхилень в обидва боки (двостороння гіпотеза):

```
> t.test(c(95, 90, 105, 90, 85),mu=100)

        One Sample t-test

data:  c(95, 90, 105, 90, 85)
t = -2.0642, df = 4, p-value = 0.1079
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
  83.58462 102.41538
sample estimates:
mean of x
       93
```

R вивів таблицю результатів, де вказано, що застосовувався одновибірковий T-тест, написано, по яких даних проводились підрахунки. Значення T-статистики виявилось рівним -2.0642 (чим ближче статистика до 0, тим більше це свідчить на користь основної гіпотези). Досягнутий рівень значущості — 0.1079, тобто основну гіпотезу про те, що автомат в середньому дає 100 мл кави,

[21] Саме так треба робити, якщо ми хочемо, щоб у хибності роботи автомата нас переконували дані спостережень, а не апріорні міркування.

слід прийняти за результатами Т-тесту при будь-якому розумному стандартному рівні значущості.

Довірчий інтервал — $[83.58462, 102.41538]$ (з рівнем значущості 0.05). Оскільки 100 потрапляє до цього інтервалу, то у нас немає причин відхилити гіпотезу $\mu = 100$. Тест на основі довірчого інтервалу рекомендує прийняти H_0. (У даному випадку тест на основі довірчого інтервалу і Т-тест еквівалентні. При однаковому рівні значущості вони на будь-яких даних будуть видавати одні і ті самі результати).

Далі у таблиці вказано, що середнє за даними дорівнює 93. Це, звичайно, менше, ніж 100, але таке відхилення виявилось недостатнім для того, щоб зробити висновок про систематичну помилку автомата. Маємо розглядати це як результат випадкових коливань навколо правильного середнього положення 100 мл.

Нехай тепер нас цікавить тільки можливість недоливу, а перелив ми готові полишити на совість автомата. Якщо за даними бажано побачити можливий недолив, то як основну треба взяти гіпотезу про перелив: $H_0 : \mu > 100$.

```
> t.test(c(95, 90, 105, 90, 85),mu=100,
+        alternative="less")


        One Sample t-test

data:  c(95, 90, 105, 90, 85)
t = -2.0642, df = 4, p-value = 0.05397
alternative hypothesis: true mean is less than 100
95 percent confidence interval:
     -Inf 100.2294
sample estimates:
mean of x
       93
```

Як бачимо з таблиці, для одностороннього тесту `p-value` = 0.05397. При стандартному рівні значущості $\alpha = 0.05$ нам знову слід прийняти гіпотезу про відсутність недоливу[22]. Але досягнутий рівень значущості тепер менший, ніж у попередньому випадку. Недолив легше помітити, якщо не відволікатись на пошук можливого, але не цікавого для нас переливу.

Дехто може сказати, що 0.05397 — це майже 0.05. Іще б трохи, і ми виявили б недолив! Зрештою, при дуже великому бажанні, можна дозволити собі підняти допустиму ймовірність помилки першого роду до 0.1. Тоді наче можна буде прийняти альтернативу. Але якщо ви з таким результатом спробуєте потягнути продавців кави до суду за недолив, то їхні адвокати зроблять з вас посміховище. Краще спробувати отримати більше даних на користь вашого висновку, наприклад, провести з автоматом іще кілька дослідів.

Тут є делікатна проблема: якщо ви будете наливати склянки по одній і перевіряти кожного разу гіпотезу про недолив за всіма даними отриманими до даного моменту, то рано чи пізно тест повідомить вас про недолив навіть тоді, коли насправді його немає (при $\mu = 100$). Правильний підхід: визначити потрібну кількість експериментів заздалегіть. За наявності деякої попередньої інформації це можливо. ◀

Приклад 9.5.2. У прикладі 3.4.4 ми розглядали дані про інтерес до шортів двох різних фасонів. Відобразивши ці дані на карті США ми помітили, що переважання інтересу до джинсових шортів над карго-шортами характерне для штатів, що розташовані у басейні Міссісіпі — Міссурі. Звичайно, картинка на карті для когось може бути переконливою, а комусь здаватись безглуздою. Наскільки статистично значущим є виявлений нами ефект? Щоб перевірити

[22]Серед результатів роботи функції `t.test()` знаходимо також "односторонній довірчий інтервал" $(-\infty, 100.2294]$ для μ з рівнем значущості $\alpha = 0.05$. Оскільки множина значень μ, що відповідають H_0, тобто $[100, +\infty)$ має спільні точки з цим інтервалом, мусимо прийняти H_0 за правилом з підрозд. 9.4. Цей тест на основі одностороннього довірчого інтервалу завжди дає той самий результат, що і відповідний односторонній Т-тест Стьюдента.

це, спробуємо застосувати T-тест для перевірки рівності математичних сподівань двох вибірок.

Нагадаємо, що у фреймі даних, який розглядається, змінна `cargo` містить кількість запитів про карго-шорти у даному штаті, а змінна `jean` — кількість запитів про джинсові шорти[23]. Щоб охарактеризувати переважання інтересу до джинсових шортів, введемо змінну `x`, що визначається як

`x=jean/(jean+cargo)`.

Чим більше у даному штаті цікавляться джинсовими шортами, тим більше `x`. У змінній `miss` міститься значення 1, якщо штат лежить у басейні Міссісіпі — Міссурі (MM-штат), і 0 — якщо ні (не-MM штат)[24]. Чи можна вважати, що змінна `x` має нормальний розподіл? Розглянемо Q-Q-діаграму для неї окремо по MM-штатах та не-MM штатах (див. рис. 9.5, 9.6).

```
> tb<-read.table("c:/rem/term/shortU.txt",header=T)
> tb$x<-tb$jean/(tb$jean+tb$cargo)
> qqnorm(tb$x[tb$miss==0],main=" ",
+        xlab=" ",ylab=" ")
> qqline(tb$x[tb$miss==0])
> qqnorm(tb$x[tb$miss==1],main=" ",
+        xlab=" ",ylab=" ")
> qqline(tb$x[tb$miss==1])
```

На діаграмі для не-MM штатів (рис. 9.5) точки добре вкладаються на пряму, розподіл можна вважати нормальним. На діаграмі MM штатів (рис. 9.6) помітний викид, що відповідає штату Південна Дакота, для якого значення `x`$=1$ (жителі цього штату не робили

[23] Фрейм даних завантажується з файла `shortU.txt` і у програмах далі фігурує під назвою `tb`.

[24] Я створив цю змінну сам за картою США. Є деяка непевність щодо того, як відмічати штати, які частково лежать у басейні цих рік, а частково — ні. Можливо, моя класифікація не зовсім методично коректна, але можемо собі таке дозволити, оскільки тут цей приклад має навчальний характер.

жодного запиту про карго-шорти). За винятком цього викиду, точки на QQ-діаграмі добре вкладаються на пряму — розподіл також можна вважати нормальним.

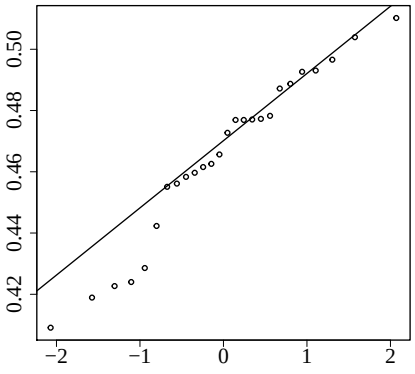

Рис. 9.5. Q-Q-діаграма для не-ММ штатів

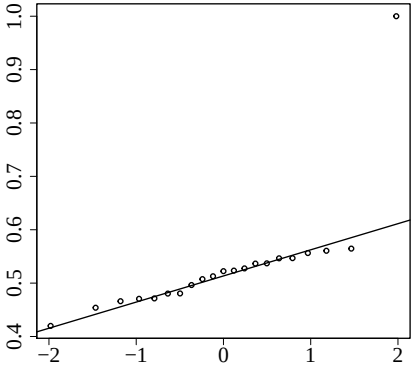

Рис. 9.6. Q-Q-діаграма для ММ штатів

Пояснити особливість Південної Дакоти не берусь, тому далі будемо проводити аналіз даних як з викидом, так і без нього. Подивимось, чи вплине викид на статистичні висновки.

Застосуємо двовибірковий Т-тест для незв'язаних вибірок для перевірки того, чи є значуща відмінність між середніми значеннями (математичними сподіваннями) змінної `x` у ММ та не-ММ штатах[25]:

```
> t.test(tb$x~tb$miss)$p.value   # з викидом

[1] 0.01502045

> tb1<-tb[!(tb$x==1),]           # видалили викид
> t.test(tb1$x~tb1$miss)$p.value # без викиду

[1] 0.0002063404
```

Як бачимо, у варіанті з викидом досягнутий рівень значущості — 0.01502, без викиду — 0.0002063. При $\alpha_0 = 0.05$ треба приймати альтернативу (математичні сподівання `x` по різних групах штатів значущо відрізняються). При $\alpha_0 = 0.01$ прийняти альтернативу треба, якщо не враховувати викид. Можна сказати, що цей викид (якщо трактувати його як забруднення) маскує відмінність середніх. Цікаво, що при врахуванні викиду різниця між вибірковими середніми зростає. Здавалося б, більшу різницю легше помітити? Але насправді тест порівнює різницю між середніми з розкидом спостережень. Альтернатива приймається тоді, коли різницю середніх не можна пояснити випадковим розкидом. А викид збільшує також і розкид вибірки,у даному випадку значніше, ніж різницю середніх.

Оскільки ми приймаємо стандартне $\alpha_0 = 0.05$, то нашим остаточним висновком буде прийняття альтернативи незалежно від

[25] Стандартно функція `t.test()` використовує двосторонній тест.

трактування викиду у Південній Дакоті. ◄

Приклад 9.5.3. Виробники акумуляторів для слухових апаратів випустили акумулятори нової моделі і стверджують, що вони працюють до розряду в середньому на 2 год. довше, ніж старі у тих самих апаратах. Для перевірки цього твердження взяли п'ять слухових апаратів і спочатку ввімкнули їх зі старими акумуляторами та дали працювати до розряду, а потім — з новими. Результати — тривалість роботи до розряду у годинах записані у таблиці:

Номер апарата	1	2	3	4	5
Старі акумулятори	12	14	11.5	13	10
Нові акумулятори	14.5	15	13	14	11.5

Середній час роботи нових акумуляторів 13.6 год. Це лише на 1.5 год. більше ніж середній час роботи старих (12.1 год.). Чи можна, спираючись на ці дані стверджувати, що виробник неправильно інформує про переваги нових акумуляторів? Для перевірки значущості результату застосуємо Т-тест. Оскільки тривалість роботи акумулятора може залежати від слухового апарата, у якому він працює, вибірки для нових і старих акумуляторів є зв'язаними. Отже, встановлюємо опцію `paired=TRUE`. Основною має бути гіпотеза, що виробник дає достовірну інформацію — середня різниця між часом роботи нового і старого акумуляторів більше 2. Проведемо парний Т-тест, вибравши односторонню альтернативу "різниця середніх менше `mu=2`" (`l`):

```
> y<-c(12,14,11.5,13,10)   # старі акумулятори
> x<-c(14.5,15,13,14,11.5) # нові акумулятори
> t.test(x,y,mu=2,alternative="l",paired=TRUE)


        Paired t-test

data:  x and y
t = -1.8257, df = 4, p-value = 0.07096
alternative hypothesis: true difference in means is
```

```
less than 2 95 percent confidence interval:
    -Inf 2.08383
sample estimates:
mean of the differences
                   1.5
```

△

Тест дає досягнутий рівень значущості 0.07096, цього не досить, щоб сваритись із виробником. Приймаємо основну гіпотезу. ◄

9.5.2 F-тест. Перевірка гіпотез про дисперсії. Заважаючий параметр — математичне сподівання

Розглянемо тепер задачу перевірки гіпотез про дисперсії двох нормально розподілених вибірок при невідомому математичному сподіванні. Для перевірки таких гіпотез використовують F-тест (тест Фішера). У `R` цей тест реалізовано в функції `var.test()`. Логіка її використання схожа на логіку `t.test()` (див. п. 9.5.1) для двовибіркових гіпотез, але є і відмінності. Параметри цієї функції:

`x`, `y` — вибірки, за якими перевіряється гіпотеза. Як і у `t.test()`, можна задати `x`, `y` або вказати формулу типу

`data~factor`.

Вибірки `x`, `y` вважаються нормальними з невідомими дисперсіями σ_x^2, σ_y^2 відповідно. Математичні сподівання також вважаються невідомими.

`ratio` відношення дисперсій σ_x^2/σ_y^2, що відповідає нульовій гіпотезі.

`alternative` — вказує, для якої альтернативи робиться перевірка. Якщо вказано `alternative="two.sided"` — перевіряється гіпотеза $H_0 : \sigma_x^2/\sigma_y^2 =$`ratio` проти альтернативи

$H_1 : \sigma_x^2/\sigma_y^2 \neq$`ratio`.

Варіанти `alternative="greater"` або `"less"` відповідають одностороннім альтернативам $H_1 : \sigma_x^2/\sigma_y^2 >$`ratio` (відповідно — $<$).

`conf.level` — задає рівень надійності $(1-\alpha)$ довірчого інтервалу для відношення σ_x^2/σ_y^2.

В результаті виконання функція створює об'єкт, що має такі атрибути:

`statistic` — значення статистики тесту (F-статистика Фішера);

`parameter` — пара чисел, що відповідають ступеням вільності чисельника і знаменника у розподілі Фішера для статистики тесту;

`p.value` — досягнутий рівень значущості тесту;

`conf.int` — довірчий інтервал для σ_x^2/σ_y^2;

`estimate` — оцінка для σ_x^2/σ_y^2.

Приклад 9.5.4. Продовжимо розгляд даних про інтерес до шортів з прикладів 9.5.2 і 3.4.4. Перевіримо, чи відрізняються дисперсії змінної `x`[26] для ММ і не-ММ штатів:

```
> tb<-read.table("c:/rem/term/shortU.txt",header=T)
> tb$x<-tb$jean/(tb$jean+tb$cargo)
> var.test(tb$x~tb$miss)$p.value   # з викидом

[1] 9.85907e-10

> tb1<-tb[!(tb$x==1),]             # видалили викид
> var.test(tb1$x~tb1$miss)$p.value # без викиду

[1] 0.0752459
```

З урахуванням викиду — Північної Дакоти виявляється, що відмінність між дисперсіями значуща на будь-якому розумному рівні значущості — досягнутий рівень значущості $p = 9.86 \times 10^{-10}$. Без викиду маємо $p = 0.075$. Це і не дивно: з викидом вибіркова дисперсія `x` по ММ штатах дорівнює 0.013, а без викиду — 0.0016, тобто на порядок менше (і ближче до дисперсії по не-ММ штатах — 0.00075). Я б сказав, що дані не підтверджують гіпотезу про

[26] Рівень інтересу до джинсових шортів.

відмінність дисперсій характеристики `x` у ММ і не-ММ штатах. Втім, навіть без викиду вибіркові дисперсії відрізняються приблизно вдвічі. Тому впевеності у рівності відповідних теоретичних дисперсій немає.

Якщо прийняти гіпотезу про рівність дисперсій, то можна повторити тест для рівності математичних сподівань `x`, враховуючи це припущення, тобто з опцією `var.equal=TRUE`:

```
> t.test(tb1$x~tb1$miss,
+        var.equal=TRUE)$p.value # дані без викиду

[1] 6.868349e-05
```

Отримали досягнутий рівень значущості $p = 6.87 \times 10^{-5}$, тобто відмінність середніх у штатах різних груп є значущою. Нагадаю, що той самий результат ми отримали і без припущення про рівність дисперсій.

Власне, результат цієї останньої перевірки можна було передбачити без проведення тесту: якщо ми побачили відмінність середніх, не використовуючи припущення про рівність дисперсій, то тим більше воно буде помітне з таким припущенням. У протилежному випадку, коли Т-тест з `var.equal=FALSE` не бачить невеликих відмінностей, тест з `var.equal=TRUE` може їх трактувати як значущі. ◄

9.5.3 Z-тест для гіпотез про середнє без заважаючих параметрів

Перейдемо тепер до теоретичного опису тестів. Почнемо з тестів для математичного сподівання.

Нехай дані $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ являють собою кратну вибірку з розподілу $N(\mu, \sigma^2)$, причому σ^2 відоме, а μ — ні. Потрібно перевірити гіпотезу $H_0 : \mu \leq \mu_0$ проти альтернативи $H_1 : \mu > \mu_0$, де

μ_0 — деяке фіксоване число. Стандартний рівень значущості — α_0.

Для перевірки використаємо статистику

$$Z = \frac{\sqrt{n}(\bar{\xi} - \mu_0)}{\sigma},$$

де $\bar{\xi}$ — вибіркове середнє $\mathbf{X}$, n — кількість спостережень. (Z називають Z-статистикою).

Як і раніше, позначимо $\lambda_{\alpha_0} = Q^{N(0,1)}(1 - \alpha_0)$ — це буде поріг нашого тесту. Тест побудуємо таким чином:

Якщо $Z > \lambda_{\alpha_0}$ — приймаємо H_1, інакше — приймаємо H_0.

Цей тест називають одностороннім Z-тестом.

Якщо справжнє $\mu = \mu_0$, то $Z \sim N(0,1)$ і $\mathsf{P}\{Z > \lambda_{\alpha_0}\} = \alpha_0$. Якщо $\mu < \mu_0$, то

$$\mathsf{P}\{Z > \lambda_{\alpha_0}\} = \mathsf{P}\{\sqrt{n}(\bar{\xi} - \mu_0)/\sigma > (\mu_0 - \mu)/\sigma + \lambda_{\alpha_0}\} \le$$

$$\le \mathsf{P}\{(\bar{\xi} - \mu_0)/\sigma > \lambda_{\alpha_0}\} \le \alpha_0.$$

Отже, для всіх $\mu \le \mu_0$ ймовірність помилки першого роду нашого тесту не перевищує α_0, тобто рівень значущості Z-тесту дорівнює α_0.

Можна показати, що односторонній Z-тест є рівномірно найбільш потужним у класі всіх тестів із заданим рівнем значущості (див. [46], приклад 6.6, або [3], п. 5 розд. 3). Він також буде тестом відношення вірогідності. Досягнутий рівень значущості цього тесту $p = 1 - \Phi(Z)$, де $\Phi(x)$ — функція розподілу $N(0,1)$.

Приклад 9.5.5. Вагу досліджуваного зразка виміряли тричі на вагах, що мають стандартне відхилення похибки вимірювання $\sigma = 0.5$ мг. Отримані результати: 11 мг, 10.3 мг, 10.6 мг. Чи можна стверджувати, що вага зразка більша, ніж 10 мг? Перевірку гіпотези зробити з рівнем значущості $\alpha_0 = 0.01$.

Оскільки всі спостереження більше 10, то на перший погляд дані підтверджують гіпотезу. Але відхилення виміряних значень від 10 невелике порівняно з стандартним відхиленням похибки. Тому необхідно зробити акуратну перевірку з використанням відповідного тесту. Ми будемо трактувати справжнє значення ваги μ як математичне сподівання вимірювань, а відхилення від нього, як випадкову похибку з нормальним розподілом, що має нульове математичне сподівання та дисперсію σ^2. (Вимірювальні прилади спеціально виробляють і градуюють так, щоб їхні похибки були такими).

Основною буде гіпотеза, протилежна тій, яку ми хочемо підтвердити даними: $H_0 : \mu \leq 10$. Якщо тест прийме протилежну гіпотезу $H_1 : \mu > 10$, це буде її переконливим обґрунтуванням. Обчислимо досягнутий рівень значущості Z-тесту для цієї пари гіпотез:

```
> x<-c(11,10.3,10.6)
> 1-pnorm(sqrt(length(x))*(mean(x)-10)/0.5)

[1] 0.01412018
```

Отримали $p = 0.0141$. Це більше, ніж 0.01, тест приймає основну гіпотезу. Отже, проведених вимірювань недостатньо, щоб підтвердити, що $\mu > 10$ мг. ◄

Зазначимо, що якщо при використанні Z-тесту завищити дисперсію (тобто взяти для нормування не справжню дисперсію, а її оцінку зверху), то рівень його значущості зменшиться. Тобто його можна буде використовувати, хоча при цьому зросте ймовірність помилок другого роду: тест перестане помічати альтернативні значення, які лежать близько до порогу μ_0.

Якщо треба перевірити гіпотезу $H_0 : \mu = \mu_0$ проти двосторонньої альтернативи $H_1 : \mu \neq \mu_0$, застосовують двосторонній Z-тест, який працює таким чином:

Прийняти H_0, якщо $|Z| < \lambda_{\alpha_0/2}$ і відхилити у протилежному випадку.

Цей тест є рівномірно найбільш потужним у класі всіх незміщених тестів рівня α_0. Досягнутий рівень значущості для нього $p = (1 - \Phi(|Z|))/2$.

Двосторонній Z-тест еквівалентний тесту відношення вірогідності для відповідної гіпотези і тесту на основі довірчого інтервалу для μ. У даному випадку довірчий інтервал рівня α_0 для μ має вигляд

$$A_n(\alpha_0) = [\bar{\xi} - \lambda_{\alpha_0/2}\sigma/\sqrt{n}, \bar{\xi} + \lambda_{\alpha_0/2}\sigma/\sqrt{n}]. \tag{9.13}$$

Зрозуміло, що умова $|Z| < \lambda_{\alpha_0/2}$ еквівалентна $\mu \in A_n(\alpha_0)$.

Досить часто виникає потреба розглядати гіпотезу $\mu = \mu_0$ не як основну, а як альтернативну. Скажімо, якщо за даними кількох зважувань ми хочемо перевірити, чи є дана гиря стандартною гирею вагою 1 кг, нам треба як основну взяти гіпотезу про те, що вона нестандартна: $H_0 : \mu \neq 1$ кг, $H_1 : \mu = 1$ кг.

У такій постановці розумний тест для перевірки цих гіпотез запропонувати не можна. Множина значень μ, яка відповідає альтернативі, настільки вузька, що найкраща оцінка для μ, тобто $\bar{\xi}$, потрапляє у неї з ймовірністю 0. А коли $\bar{\xi}$ потрапляє у множину, яка відповідає основній гіпотезі H_0, у нас немає причин відхилити H_0.

Більш розумним буде встановити певну область допустимих значень навколо стандарту, які відповідатимуть альтернативі. Наприклад, нехай стандартна гиря повинна мати вагу 1 кг$\pm$1 г. Тоді $H_0 : \mu \notin [0.999, 1.001]$, $H_1 : \mu \in [0.999, 1.001]$.

У загальному вигляді можна записати $H_0 : \mu \notin [a, b]$, де $a < b$ — фіксовані числа. Для перевірки такої гіпотези проти альтернативи $H_1 : \mu \in [a, b]$, можна застосувати тест на основі довірчого інтервалу:

H_1 приймається, якщо $A_n(\alpha_0)$ повністю лежить у $[a, b]$. Якщо є точки $A_n(\alpha_0)$, які не належать $[a, b]$, приймається H_0.

Цей тест є рівномірно найбільш потужним. Він еквівалентний тесту відношення вірогідності для цієї пари гіпотез.

9.5.4 Знову про тести для дисперсії

Гіпотези про дисперсії природно перевіряти, використовуючи суми квадратів відхилення спостережень від математичного сподівання (якщо воно відоме) або від середнього.

Нехай, як і раніше, $\mathbf{X} = (\xi_1, \dots, \xi_n)$ — кратна вибірка з розподілу $N(\mu, \sigma^2)$. Тоді

$$\frac{1}{\sigma^2}\sum_{j=1}^{n}(\xi_j - \mu)^2 \sim \chi^2_n,\ \frac{1}{\sigma^2}\sum_{j=1}^{n}(\xi_j - \bar{\xi})^2 = \frac{(n-1)S_0^2(\mathbf{X})}{\sigma^2} \sim \chi^2_{n-1}.$$

Завдяки цьому можна будувати односторонні тести для перевірки гіпотез про дисперсії, аналогічні розглянутим вище для математичного сподівання. При цьому квантилі χ^2-розподілу відіграють ту ж роль, яку для середніх відігравали квантилі нормального розподілу. Відмінність в тому, що розподіл χ^2 — не симетричний, тому його верхні квантилі не можна виразити через нижні.

Приклад 9.5.6. У прикладі 9.5.5 ми розглядали ваги, що визначають вагу з середньоквадратичним відхиленням похибки $\sigma = 0.5$ мг. Звідки взялось це значення? Мабуть хтось перевірив роботу цих вагів на еталонних зразках і визначив дисперсію похибок. На практиці важливо, щоб ця дисперсія не перевищувала деякої межі. Отже, нехай для еталонного зразка з точно відомою вагою μ було проведено n зважувань $(\xi_1, \dots, \xi_n)$. Відповідною перевіркою підтверджена нормальність розподілу та відсутність систематичної похибки (тобто перевірено, що математичне сподівання дорівнює μ). Треба перевірити, що дисперсія не перевищує σ_0^2.

Таким чином, маємо задачу перевірки гіпотези $H_0 : \sigma^2 > \sigma_0^2$ проти альтернативи $H_1 : \sigma^2 \le \sigma_0^2$. Статистика тесту

$$C = \frac{1}{\sigma_0^2} \sum_{j=1}^{n} (\xi_j - \mu)^2,$$

поріг — $C_{\alpha_0} = Q^{\chi^2_n}(\alpha_0)$:

Тест приймає H_0, якщо $C \ge C_{\alpha_0}$ і відхиляє, якщо $C < C_{\alpha_0}$.

Досягнутий рівень значущості тесту визначається на основі статистики C як $p = F^{\chi^2_n}(C)$. ◀

Розглянемо тепер двосторонній тест для перевірки гіпотези $H_0 : \sigma^2 = \sigma_0^2$ проти альтернативи $H_1 : \sigma^2 \ne \sigma_0^2$. Ту саму статистику C, що визначена у прикладі 9.5.6, можна використати і для цього тесту, але порівнювати її треба з двома порогами — $C^-_{\alpha_0} = Q^{\chi^2_n}(\alpha_0/2)$ і $C^+_{\alpha_0} = Q^{\chi^2_n}(1 - \alpha_0/2)$.

Тест:

Якщо $C^-_{\alpha_0} \le C \le C^+_{\alpha_0}$ приймаємо основну гіпотезу, інакше — відхиляємо.

Досягнутий рівень значущості цього тесту: $p = 1 - |1 - 2F^{\chi^2_n}(C)|$.

Приклад 9.5.7. Псевдовипадкові числа, створені генератором `rnorm(n,mean=1,sd=2)`, повинні мати нормальний розподіл з математичним сподіванням 1 і дисперсією 4. Перевіримо, чи є значущим відхилення дисперсії згенерованих чисел від 4? Основна гіпотеза — відхилення немає, $H_0 : \sigma^2 = 4$:

▽

```
> set.seed(2)
> n<-1000000
> x<-rnorm(n,mean=1,sd=2)
```

```
> s0<-4
> 1-abs(1-2*pchisq(sum((x-1)^2)/s0,n))

[1] 0.7328603
```

△

Досягнутий рівень значущості $p = 0.7328603$ — відхилення не виявлено за мільйоном спостережень. Якщо у цьому скрипті підставити `s0<-3.98`, отримаємо $p = 0.0013$, а при `s0<-4.02` — $p = 0.0001$, тобто при рівні значущості $\alpha_0 = 0.01$ слід прийняти альтернативу. Наш тест помічає відхилення, що складають піввідсотка справжньої дисперсії. Але для цього знадобилось 1 000 000 спостережень. ◀

Для того, щоб порівнювати дисперсії двох вибірок, застосовують F-тести. Ці тести будуються на нормованому відношенні C-статистик, що відповідають окремим вибіркам. Таке відношення називають F-відношенням (F-статистикою).

Нехай дані складаються з двох вибірок $\mathbf{X}^1 = (\xi_1^1, \ldots, \xi_{n_1}^1)$ і $\mathbf{X}^2 = (\xi_1^2, \ldots, \xi_{n_1}^2)$, причому спостереження з $\mathbf{X}^1$ мають розподіл $N(\mu_1, \sigma_1^2)$, а з $\mathbf{X}^2$ — $N(\mu_2, \sigma_2^2)$. Значення μ_1, μ_2 — відомі, треба перевірити гіпотезу про рівність дисперсій[27] $H_0 : \sigma_1^2 = \sigma_2^2$. У цьому випадку F-статистика має вигляд

$$F = \frac{C(\mathbf{X}^1)}{C(\mathbf{X}^2)} = \frac{\frac{1}{n_1}\sum_{j=1}^{n_1}(\xi_j^1 - \mu^1)^2}{\frac{1}{n_2}\sum_{j=1}^{n_2}(\xi_j^2 - \mu^2)^2}.$$

При виконанні H_0 ця статистика має F-розподіл Фішера з n_1 ступенями вільності чисельника і n_2 ступенями вільності знаменника. Відповідно, для перевірки (двосторонньої) H_0 статистика порівнюється з двома порогами: $F_{\alpha_0}^- = Q^{F(n_1,n_2)}(\alpha_0/2)$, $F_{\alpha_0}^+ = Q^{F(n_1,n_2)}(1 - \alpha_0/2)$.

Тест має вигляд:

[27] Можна також розглядати односторонні гіпотези, або тести, у яких перевіряються гіпотези про відношення двох дисперсій.

Якщо $F_{\alpha_0}^- \le F \le F_{\alpha_0}^+$ — прийняти H_0, інакше — відхилити.

Досягнутий рівень значущості цього тесту

$$p = 1 - |1 - 2G(F)|, \tag{9.14}$$

де G — функція розподілу $F(n_1, n_2)$-розподілу[28].

Приклад 9.5.8. Порівнюють дві методики зважування[29]. Для цього один і той самий зразок вагою 1 г зважують тричі обома методиками і отримують значення 1.008571, 1.000319, 1.006722 — за першою методикою, 0.9986845, 1.0020705, 0.9997553 — за другою. Основною є гіпотеза про те, що обидві методики дають однакову дисперсію похибок. Але вибіркова дисперсія другої вибірки приблизно у шість разів більша, ніж першої. Чи можна твердити, що дані суперечать основній гіпотезі? Застосуємо F-тест:

```
>     y<-c(1.008571, 1.000319, 1.006722)
>     x<-c(0.9986845, 1.0020705, 0.9997553)
> F<-sum(length(y)*(x-1)^2)/(length(x)*sum((y-1)^2))
> 1-abs(1-2*pf(F,length(x),length(y)))

[1] 0.03593744
```

Досягнутий рівень значущості $p = 0.03593744$. Якщо використовувати 5% стандартний рівень значущості, основну гіпотезу слід відхилити, методики дають різну точність вимірювання.

Спробуємо перевірити ту ж гіпотезу, застосовуючи стандартний F-тест для порівняння дисперсій у двох вибірках:

```
> var.test(x,y,ratio=1)
```

[28] У наших загальних позначеннях слід було б написати $p = 1 - |1 - 2F^{F(n_1,n_2)}(F)|$, але тут літера F виходить занадто перенавантаженою.

[29] Це можуть бути не зважування, а вимірювання будь-якої фізичної величини.

```
        F test to compare two variances
data:  x and y
F = 0.15973, num df = 2, denom df = 2, p-value =
0.2755
alternative hypothesis: true ratio of variances
is not equal to 1
95 percent confidence interval:
 0.004095665 6.229506005
sample estimates:
ratio of variances
         0.1597309
```

△

— досягнутий рівень значущості цього тесту $p = 0.2755$ — слід прийняти основну гіпотезу!

Чому наш тест прийняв альтернативу, а стандартний — ні? При побудові нашого тесту ми виходили з того, що математичне сподівання вимірювань відоме (1 г). У стандартному тесті математичні сподівання оцінюються вибірковими середніми. Це зменшує точність оцінок для дисперсій, тому стандартний тест менш схильний помічати їх відмінності, частіше буде змушений приймати основну гіпотезу, не зважаючи на різницю вибіркових дисперсій. Таким чином, враховуючи додаткову інформацію ми можемо зробити наш тест більш чутливим. ◄

А як працює стандартний тест двовибірковий F-тест для перевірки рівності дисперсій[30], у якому μ_1 і μ_2 вважаються невідомими? Статистика цього тесту має вигляд

$$F = \frac{\frac{1}{n_1}\sum_{j=1}^{n_1}(\xi_j^1 - \bar{\xi}^1)^2}{\frac{1}{n_2}\sum_{j=1}^{n_2}(\xi_j^2 - \bar{\xi}^2)^2}.$$

При виконанні основної гіпотези $H_0 : \sigma_1^2 = \sigma_2^2$ ця статистика має F-розподіл Фішера з $n_1 - 1$ ступенем вільності чисельника і $n_2 - 1$ ступенем вільності знаменника. (Тобто заміна при

[30]Тобто тест, реалізований у функції `var.test()`, див. п. 9.5.2.

обчисленні статистики справжнього математичного сподівання його оцінкою зменшує кількість ступенів вільності на 1). Відповідно змінюються пороги тесту: $F^-_{\alpha_0} = Q^{F(n_1-1,n_2-1)}(\alpha_0/2)$, $F^+_{\alpha_0} = Q^{F(n_1-1,n_2-1)}(1-\alpha_0/2)$.

Сам тест тепер з новими порогами має той самий вигляд, як і у попередньому випадку. Досягнутий рівень значущості також розраховується за (9.14), де G тепер — це функція розподілу для $F(n_1-1, n_2-2)$.

9.5.5 Знову про тести для математичних сподівань

Нам залишилось розібратись теоретично, як працюють T-тести для перевірки гіпотез про математичні сподівання при невідомих дисперсіях[31].

Для одновибіркових задач, коли гіпотези про математичне сподівання μ перевіряють за нормальною кратною вибіркою $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ з невідомою дисперсією, T-тести використовують статистику

$$T = T(\mathbf{X}) = \frac{(\bar{\xi} - \mu_0)\sqrt{n}}{S_0(\mathbf{X})},$$

де μ_0 — порогове значення математичного сподівання для гіпотези, що перевіряється, $S_0^2(\mathbf{X}) = \frac{1}{n-1}\sum_{j=1}^n (\xi_j - \bar{\xi})^2$ — виправлена вибіркова дисперсія даних, $S_0(\mathbf{X}) = \sqrt{S_0^2(\mathbf{X})}$. Тобто при побудові T-статистики справжня дисперсія даних замінюється її оцінкою по вибірці.

Якщо справжнє математичне сподівання даних $\mu = \mu_0$, то T має T-розподіл Стьюдента з $n-1$ ступенем вільності. Тому тести для перевірки гіпотез про μ можна будувати на основі T-статистики аналогічно описаним у п. 9.5.3, але використовуючи як пороги тесту квантилі T-розподілу замість квантилів нормального розподілу.

[31] Як на практиці користуватись цими тестами в R розповідається у п. 9.5.1.

Наприклад, тест для перевірки гіпотези $H_0 : \mu = \mu_0$ має такий вигляд:

Якщо $|T| < Q^{T_{n-1}}(1 - \alpha_0/2)$, приймаємо H_0, інакше відхиляємо.

При аналізі двовибіркових задач із залежними вибірками для порівняння математичних сподівань фактично використовується Т-статистика, побудована за різницями спостережень з першої і другої вибірок. Точніше, нехай спостерігаються дані $\mathbf{X}^1 = (\xi_1^1, \ldots, \xi_n^1)$, $\mathbf{X}^2 = (\xi_1^2, \ldots, \xi_n^2)$, причому спостереження з однаковими номерами у першій і другій вибірках, тобто вектори (ξ_j^1, ξ_j^2) є незалежними, гауссовими векторами з (можливо) корельованими координатами. Треба перевірити гіпотезу[32] H_0 про те, що $\mu_1 = \mu_2$, де $\mu_i = \mathsf{E}\,\xi_j^i$. Позначимо $\xi_j = \xi_j^1 - \xi_j^2$. Оскільки ξ_j є нормально розподіленими з математичним сподіванням $\mu = \mathsf{E}\,\xi_j = \mu_1 - \mu_2$, то перевірка H_0 зводиться до перевірки гіпотези $\mu = 0$. Для цього використовується статистика $T(\mathbf{X}) = T(\mathbf{X}^1 - \mathbf{X}^2)$:

якщо $|T(\mathbf{X})| \le Q^{T_{n-1}}(1 - \alpha_0/2)$ — H_0 приймається, інакше відхиляється.

Залишились ще двовибіркові тести для незв'язаних вибірок, коли $\mathbf{X}^1 = (\xi_1^1, \ldots, \xi_{n_1}^1)$, $\mathbf{X}^2 = (\xi_1^2, \ldots, \xi_{n_2}^2)$ незалежні між собою з можливо різними математичними сподіваннями μ_1 і μ_2. Дисперсії σ_1^2 і σ_2^2 вважаються невідомими. Розрізняють два випадки.

Перший випадок: вважається, що $\sigma_1^2 = \sigma_2^2$. Тоді можна для перевірки $H_0 : \mu_1 = \mu_2$ використати Т-статистику

$$T(\mathbf{X}^1, \mathbf{X}^2) = \frac{\bar{\xi}^1 - \bar{\xi}^2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S^2(\mathbf{X}^1, \mathbf{X}^2)}},$$

[32] Або інші аналогічні гіпотези.

де

$$S^2(\mathbf{X}^1, \mathbf{X}^2) = \frac{(n_1 - 1)S_0^2(\mathbf{X}^1) + (n_2 - 1)S_0^2(\mathbf{X}^2)}{(n_1 + n_2 - 2)}$$

— об'єднана оцінка для спільної дисперсії обох вибірок, обчислена з урахуванням різних математичних сподівань.

При виконанні гіпотези H_0 статистика $T(\mathbf{X}^1, \mathbf{X}^2)$ має T-розподіл Стьюдента з $n_1 + n_2 - 2$ ступенями вільності. Відповідно поріг тесту для перевірки H_0 обирають як $Q^{T_{n_1+n_2-2}}(1 - \alpha_0/2)$.

Гіпотеза H_0 приймається, якщо $|T(\mathbf{X}^1, \mathbf{X}^2)| \leq Q^{T_{n_1+n_2-2}}(1 - \alpha_0/2)$ і відхиляється при виконанні протилежної нерівності.

Другий випадок: дисперсії вибірок не вважаються однаковими (можливо $\sigma_1^2 \neq \sigma_2^2$). У цьому випадку для перевірки гіпотези $\mu_1 = \mu_2$ використовується t-тест Велча — Сатерсвайта (Welch — Satterthwaite test). У цьому тесті використовується статистика

$$T^{WS} = \frac{\bar{\xi}^1 - \bar{\xi}^2}{\sqrt{\frac{S_0^2(\mathbf{X}^1)}{n_1} + \frac{S_0^2(\mathbf{X}^2)}{n_2}}}.$$

Якби тут замість $S_0^2(\mathbf{X}^i)$ стояли справжні σ_i^2, то така статистика мала б стандартний нормальний розподіл при H_0. Але справжні дисперсії нам невідомі. Заміна їх оцінками змінює розподіл статистики. Цей розподіл не описується явно, але ще у 1946 р. Ф.Е. Сатерсвайт запропонував його наближення T-розподілом Стьюдента з кількістю ступенів вільності

$$\nu = \frac{\left(\frac{S_0^2(\mathbf{X}^1)}{n_1} + \frac{S_0^2(\mathbf{X}^2)}{n_2}\right)^2}{\frac{(S_0^2(\mathbf{X}^1)/n_1)^2}{n_1 - 1} + \frac{(S_0^2(\mathbf{X}^2)/n_1)^2}{n_2 - 1}}.$$

Отже, тест працює так:

Приймати основну гіпотезу $H_0 : \mu_1 = \mu_2$ слід, якщо $|T^{WS}| \le Q^{T_\nu}(1 - \alpha_0/2)$. Якщо нерівність виконується у протилежний бік, то слід прийняти альтернативу.

Як було зазначено, тест Велча — Сатерсвайта є асимптотичним — він добре працює лише при достатньо великих обсягах вибірки. Всі інші тести, описані у цьому підрозділі можна застосовувати для вибірок будь-якого обсягу, більшого 1.

9.6 Тести χ^2

Тести χ^2 можна розглядати як наближену версію тестів відношення вірогідності для спостережень, що набувають значень лише з деякого скінченного набору (тобто описуються категорійними змінними, яким у R відповідає тип фактор, див. п. 2.2.3). З теоретичної точки зору якихось переваг над тестами відношення вірогідності вони не мають. Але для значної кількості застосувань саме використання певної версії тестів χ^2 стало практичним стандартом. Тому важливо вміти їх застосовувати та інтерпретувати їхні результати. Почнемо з простішого варіанта тесту χ^2, потім введемо загальну схему і розглянемо два найбільш поширені застосування — перевірка узгодженості розподілу і перевірка залежності.

9.6.1 Тест χ^2 для простих основних гіпотез

Розглянемо таку задачу: дані являють собою кратну вибірку $\mathbf{X} = (\xi_1, \dots, \xi_n)$, причому спостереження ξ_j набувають значень з фіксованого набору $(x_1, \dots, x_r)$. Позначимо $q_k = \mathsf{P}\{\xi_j = x_k\}$, $\mathbf{q} = (q_1, \dots, q_r)$ — вектор справжнього розподілу спостережень. Цей вектор вважається невідомим. Треба перевірити гіпотезу про те, що цей розподіл дорівнює деякому фіксованому розподілу $\mathbf{p} = (p_1, \dots, p_r)$. При цьому ми вважаємо, що всі $p_i > 0$.

Отже, розподіл наших даних задається параметром $\mathbf{q}$, який

набуває значеннь з набору всіх можливих імовірнісних розподілів

$$Q = \{\mathbf{q} \,:\, q_1 + \cdots + q_r = 1,\ q_i \geq 0, i = 1, \ldots, r\}.$$

Основна гіпотеза, яку ми перевіряємо —

$$H_0 : \mathbf{q} = \mathbf{p},$$

альтернатива — $\mathbf{q} \neq \mathbf{p}$, причому $\mathbf{q}$ може набувати будь-яких інших значеннь з Q.

Ідея тесту полягає в тому, щоб порівняти (абсолютні) частоти появ значень x_i у вибірці з тими теоретичними значеннями, які вони в середньому повинні мати, якщо H_0 правильна. Нагадаємо[33], що абсолютна частота x_i — це

$$n_i = \#\{j \,:\, \xi_j = x_i\} = \sum_{j=1}^{n} \mathbb{1}\{\xi_j = x_i\}$$

— кількість тих спостережень, у яких спостерігалось x_i (при використанні тесту χ^2 ці частоти називають **емпіричними** або спостережуваними (англ. *observed frequency*) та інколи позначають O_i).

При виконанні H_0 математичне сподівання емпіричної частоти дорівнює $\bar{n}_i = np_i$. Ці величини називають **теоретичними частотами** (англ. *expected frequency* — очікувані частоти, позначення — E_i).

Якщо H_0 виконана, то $n_i \approx \bar{n}_i$ при великих n. У тесті χ^2 використовується статистика, що є навантаженою сумою квадратів відхилень n_i від $\bar{n}_i$:

$$\chi^2_{emp} = \sum_{i=1}^{r} \frac{(n_i - \bar{n}_i)^2}{\bar{n}_i}.$$

Цю статистику називають *статистикою хі-квадрат*, або хі-квадрат емпіричним (*observed chi-squared*). Якщо виконана основна

[33] Див. підрозд. 4.5

гіпотеза, природно сподіватись, що χ^2 буде малим[34].

При виконанні H_0 статистика χ^2_{emp} для великих n має приблизно χ^2-розподіл з $r-1$ ступенем вільності. Це дозволяє за заданим стандартним рівнем значущості α визначити поріг відповідного тесту

$$\chi^2_{th} = Q^{\chi^2_{r-1}}(1-\alpha)$$

який називають **хі-квадрат теоретичне**. Тест має такий вигляд:

> Якщо $\chi^2_{emp} \le \chi^2_{th}$, приймаємо H_0 (розподіл даних задається **p**),
> якщо $\chi^2_{emp} > \chi^2_{th}$ — відхиляємо H_0 (виявлено значуще відхилення розподілу даних від теоретичного).

Досягнутий рівень значущості тесту обчислюється як

$$p = 1 - F^{\chi^2_{r-1}}(\chi^2_{emp}). \tag{9.15}$$

У R реалізувати тест χ^2 для простих гіпотез можна, використовуючи функцію `chisq.test()`. Для цього їй у параметрі `x` треба передати набір емпіричних частот $(n_1, \dots, n_r)$, а у параметрі `p` — набір теоретичних ймовірностей $(p_1, \dots, p_r)$, що відповідають основній гіпотезі. Якщо цей параметр не заданий, стандартно використовується $p_i = 1/r$ (дискретний рівномірний розподіл).

Як результат функція видає досягнутий рівень значущості для перевірки основної гіпотези[35].

Приклад 9.6.1. У спорті відомий так званий "ефект відносного віку": спортсмени, що народились у певні "щасливі" місяці, частіше досягають високих успіхів у спорті, ніж народжені у ін-

[34] Ділення кожного доданка на $\bar{n}_i$ на евристичному рівні можна пояснити тим, що без нього відмінності емпіричних і теоретичних частот для великих $\bar{n}_i$ заважали б помічати і враховувати такі відмінності для малих $\bar{n}_i$.

[35] Точніше, значенням функції є об'єкт з результатами перевірки, досягнутий рівень значущості — у його атрибуті `p.value`.

ші, “не щасливі” місяці[36]. Ми перевіримо, чи виявляється схожий ефект у вивченні математики: чи можна стверджувати, що люди, які народились у певну пору року, мають кращі можливості вивчати математику, ніж ті, яким не повезло народитись невдалої пори?

Для перевірки скористаємось результатами опитування студентів-математиків однієї академічної групи четвертого курсу механіко-математичного факультету Київського національного університету, що наведені у табл. 9.2.

Таблиця 9.2. Дані про пору року народження студентів

Пора року	зима	весна	літо	осінь
Кількість народжених	7	3	2	8

Тут вказано, скільки студентів групи народилось у відповідну пору року. Навчання на факультеті вимагає значного напруження математичних здібностей та умінь, тому ті, хто до цього не готовий, або не вступають сюди, або відсіюються на молодших курсах. Якщо, скажімо, нарождення восени сприяє розвитку математичних здібностей більше, ніж народження влітку, то можна сподіватись, що серед нашої групи буде більше відповідних студентів.

Придивляючись до таблиці, бачимо, що саме так і є. Чи є цей факт значущим доказом на користь гіпотези про ефект відносного віку для математичних успіхів? Не обов’язково. Навіть, якщо всі пори року однаково сприяють народженню математиків, випадковий розкид приведе до того, що у різних групах будуть різні пропорції студентів за різними сезонами народження[37]. Тому розумно формалізувати задачу і зробити перевірку, використовуючи

[36] Англ. relative age effect. Це не є результатом астрологічного впливу зірок, а пов’язано з організацією навчання у спортивних школах (див., напр., [29]).

[37] Теоретично треба ще враховувати можливість різного розподілу загальної народжуваності за порами року. Але в Україні народжуваність мало відрізняється у різні сезони.

відповідний статистичний тест.

Ми будемо розглядати нашу академічну групу як вибірку, кожного студента — як окремий елемент цієї вибірки, а пору року, коли він народився — як випадкову характеристику студента[38]. Тоді таблиця 9.2 — це набір емпіричних частот, за яким можна перевіряти гіпотези про розподіл випадкових спостережень.

Якщо ми хочемо виявити ефект відносного віку за даними, використовуючи статистичний тест, то як основну слід взяти гіпотезу про те, що ефекту немає. Цьому відповідає рівномірний розподіл: ймовірності народитись у будь-яку пору для студента-математика однакові і дорівнюють $1/4$. Це і буде гіпотеза H_0. Отже, застосуємо тест χ^2 для її перевірки:

```
> chisq.test(x=c(7,3,2,8))


        Chi-squared test for given probabilities

data:  c(7, 3, 2, 8)
X-squared = 5.2, df = 3, p-value = 0.1577
```

Функція повідомляє, що значення $\chi^2_{emp} = 5.2$, для перевірки використовується χ^2-розподіл з `df`$= 3$ ступенями вільності і досягнутий рівень значущості тесту `p-value` $= 0.1577$. Висновок: при будь-якому розумному стандартному рівні значущості слід прийняти основну гіпотезу. Спостережувані у цих даних відхилення частот від рівномірних недостатні для того, щоб зробити висновок про теоретичну нерівномірність розподілу.

Як ми вже казали, тест χ^2 є асимптотичним: він дає адекватні результати лише для вибірок великого обсягу. На практиці не рекомендують використовувати стандартний тест χ^2, якщо хоча б одна

[38] Для мене, звичайно, дата мого народження не є випадковою. Але, якщо розглядати навмання вибрану людину з деякої популяції, дата її народження буде випадковою величиною.

з емпіричних частот $n_i < 5$. Тому у нашому прикладі застовність цього тесту може викликати сумніви.

Для таких випадків недоцільно використовувати формулу (9.15), яка застосовує асимптотичне наближення, але можна скористатись імітаційним моделюванням для розрахунку досягнутого рівня значущості, що відповідає статистиці χ^2_{emp} для вибірки заданого малого обсягу. Схожу техніку ми застосовували у прикладах 9.2.2 та 9.2.3 для статистики відношення вірогідності. Функція `chisq.test()` реалізує такий алгоритм імітаційного обчислення досягнутого рівня значущості тесту при застосуванні опції `simulate.p.value=TRUE`:

```
> set.seed(3)
> chisq.test(x=c(7,3,2,8),simulate.p.value=TRUE)


        Chi-squared test for given probabilities with
        simulated p-value (based on 2000 replicates)

data:  c(7, 3, 2, 8)
X-squared = 5.2, df = NA, p-value = 0.2064
```

— отримали приблизне значення `p-value`$= 0.2064$, тобто і за результатами цього тесту треба прийняти основну гіпотезу. При використанні інших псевдовипадкових послідовностей[39] значення отриманого наближення для `p-value` буде змінюватись, але не дуже сильно, рішення про прийняття H_0 залишатиметься в силі.

Повертаючись до табл. 9.2 можна помітити, що спостережувані частоти для зими та осені відрізняються від теоретичних рівномірних в один бік, а для весни та літа — в протилежний. Виникає бажання об'єднати пори року у два "суперсезони" — осінь-зима і весна-літо та перевірити відмінності частот від рівномірних по цих

[39] Тобто якщо встановити іншу зернину `set.seed()` або не встановлювати жодної.

двох сезонах. Якщо це зробити, отримаємо `p-value`$= 0.02535$ — гіпотезу про рівномірний розподіл слід відхилити (при $\alpha = 0.05$).

Це типовий приклад методичної помилки при застосуванні статистичного тесту: ми спочатку підігнали дизайн тесту під дані так, щоб вони найвиразніше свідчили на користь однієї з гіпотез, а тоді вже застосували тест. Зрозуміло, що він показав те, чого ми хотіли, а не те, що є насправді. Правильний підхід полягає в тому, щоб вибирати потрібну версію тесту не дивлячись на дані, а виходячи з поставленої теоретичної задачі. В ідеалі дослідник має спочатку сформулювати гіпотезу, потім розробити дизайн експерименту та обрати тест, яким буде проведена перевірка. Потім провести дослідження, отримати дані і застосовувати до них лише ту тестову процедуру, яка була визначена наперед.

На жаль, у практичних дослідженнях витримати таку послідовність важко: гіпотези змінюються в ході досліджень, висуваються нові теорії на основі даних, і їх також треба перевіряти. Такі перевірки не матимуть сили остаточного доказу, але можуть допомогти у пошуці істини. А остаточною має бути контрольна перевірка на нових даних, отриманих після того, як гіпотеза набула завершеної форми. Тому при описі результатів перевірки необхідно завжди вказувати, наскільки висування гіпотези було пов'язане з тими даними, за якими вона перевіряється.

Завершуючи розгляд цього прикладу, додам, що він, звичайно, умовний, але дуже зручний для лекційної демонстрації. Я проводив такі опитування на заняттях близько 25 разів і лише в одному випадку застосування тесту χ^2 з $\alpha = 0.05$ привело до прийняття альтернативи. Саме так і має бути, якщо залежності успіху у математиці від пори народження немає. ◀

9.6.2 Тест χ^2 для складної основної гіпотези

Досить часто основна гіпотеза визначає розподіл даних не однозначно, а з точністю до деякого невідомого параметра. Позначимо такий параметр $\tau \in T \subseteq \mathbb{R}^d$. Використовуючи позначення

попереднього підрозділу, можна записати

$$H_0: \text{ Існує таке } \tau \in T, \text{ що } \mathbf{q} = \mathbf{p}(\tau),$$

де $\mathbf{p}(\tau)$ — розподіл даних, який відповідає значенню параметра τ.

Відповідно, альтернатива має вигляд

$$H_1: \text{ Для всіх } \tau \in T, \mathbf{q} \neq \mathbf{p}(\tau).$$

У цьому випадку неможливо обчислити теоретичні частоти за формулою $\bar{n}_i = np_i(\tau)$, оскільки τ невідоме. Але можна замінити це невідоме значення деякою оцінкою $\hat{\tau}_n$ за спостереженнями $\mathbf{X}$. Зрозуміло, що, будуючи оцінку, слід виходити з припущення про правильність H_0[40]. Рекомендується використовувати оцінку методу найбільшої вірогідності, тобто

$$\hat{\tau}_n^{MLE} = \underset{\tau \in T}{\operatorname{argmax}} \sum_{i=1}^{r} n_i \ln(p_i(\tau)), \tag{9.16}$$

або оцінку методу мінімуму χ^2:

$$\hat{\tau}_n^{MCH} = \underset{\tau \in T}{\operatorname{argmin}} \sum_{i=1}^{r} \frac{(n_i - np_i(\tau))^2}{np_i(\tau)}.$$

Ці оцінки є асимптотично еквівалентними в тому розумінні, що $\sqrt{n}(\hat{\tau}_n^{MLE} - \hat{\tau}_n^{MCH}) \to 0$ за ймовірністю при $n \to \infty$. Тому, з точки зору асимптотичної теорії, їхнє використання дає однаковий ефект.

Таким чином, у випадку складної основної гіпотези статистика тесту χ^2 визначається як

$$\chi^2_{emp} = \sum_{i=1}^{r} \frac{(n_i - np_i(\hat{\tau}_n))^2}{np_i(\hat{\tau}_n)},$$

[40] Якщо H_0 неправильна, то τ не має змісту.

де як $\hat{\tau}_n$ можна використовувати оцінку найбільшої вірогідності, оцінку мінімуму χ^2, або будь-яку іншу оцінку, асимптотично еквівалентну цим двом.

Якщо H_0 правильна, то при виконанні досить широких умов розподіл χ^2_{emp} прямує до χ^2-розподілу з

$$\mathrm{df} = r - d - 1 \tag{9.17}$$

ступенем вільності. Тут r — кількість різних значень, які з ненульвою ймовірністю можуть приймати спостереження, d — кількість незалежних[41] невідомих числових параметрів, що описують розподіл даних при H_0. Тобто r — найменша вимірність векторного параметра τ, який потрібен для однозначного задання розподілу при H_0.

Виходячи з цього поріг тесту визначається як

$$\chi^2_{th} = Q^{\chi^2_{r-d-1}}(1-\alpha),$$

де α — стандартний рівень значущості.

У таких позначеннях тестова процедура має той самий вигляд, як і у попередньому випадку:

Якщо $\chi^2_{emp} \le \chi^2_{th}$, приймаємо H_0 (розподіл даних задається **p**),
якщо $\chi^2_{emp} > \chi^2_{th}$ — відхиляємо H_0 (виявлено значуще відхилення розподілу даних від теоретичного).

Досягнутий рівень значущості тесту обчислюється як

$$p = 1 - F^{\chi^2_{r-d-1}}(\chi^2_{emp}).$$

Приклад 9.6.2. У прикладі 8.1.5 ми розглядали дані про кількості дефектів жорстких комп'ютерних дисків, які були повернені

[41] Коли кажуть про **незалежні** параметри, то мають на увазі, що жоден з них не можна виразити через інші.

по гарантії. Там припускалося, що ці дані описуються розподілом Пуассона зі зрізаним нулем. Перевіримо цю гіпотезу за даними зі стовпчика A табл. 8.1, використовуючи тест χ^2.

Теоретичні ймовірності для цього розподілу були визначені у прикладі 6.4.2:

$$p_i(\lambda) = \frac{\lambda^i}{i!(e^\lambda - 1)}. \tag{9.18}$$

Емпіричні частоти з табл. 8.1 відрізняються від 0 лише для $i = 1, \ldots, 5$. Теоретичні ймовірності не нульові для всіх натуральних i, але тест χ^2 не можна застосовувати до спостережень, які теоретично можуть набувати нескінченної кількості значень. У таких випадках рекомендують об'єднувати в одну "комірку" найбільше значення, яке зустрічається у вибірці та всі теоретично можливі значення більші, ніж це. У нашому випадку об'єднуються всі $i = 5, 7, \ldots$. Теоретична ймовірність потрапити у цю комірку

$$p'_5(\lambda) = 1 - \sum_{i=1}^{4} p_i(\lambda). \tag{9.19}$$

Використаємо цей варіант для застосування тесту χ^2 у наступному скрипті[42]:

```
> x<-1:5 # значення, що зустрічаються у вибірці
> A<-c(20,13,11,6,2) # емпіричні частоти
> lambda<-EstP(x,A) # оцінка параметра
> n<-sum(A) # кількість спостережень у вибірці
> # теоретичні ймовірності:
> p<-lambda^x/(factorial(x)*(exp(lambda)-1))
> p[length(p)]<-1-sum(p[-length(p)]) # корекція p[5]
> np<-n*p # теоретичні частоти
```

[42] Функція `EstP()`, що обчислює оцінку параметра λ, визначена у прикладі 8.1.5. У прикладі 8.3.5 показано, що це оцінка методу найбільшої вірогідності. Щоправда, ця оцінка обчислюється за моделлю (9.19) і не враховує корекцію (9.18), але практики зазвичай нехтують такими дрібними неточностями.

```
> chi2<-sum((A-np)^2/np) # хі-квадрат емпіричне
> # досягнутий рівень значущості:
> 1-pchisq(chi2,df=length(x)-2)

[1] 0.6782084
```

△

Ми отримали в результаті досягнутий рівень значущості 0.6782084, тобто треба прийняти основну гіпотезу про те, що розподіл даних відповідає пуасоновій моделі зі зрізаним нулем.

Як і у прикладі 9.6.1, можна помітити, що частота $n_5 = 2$ менша 5. За стандартними правилами у таких випадках використання тесту χ^2 не рекомендується. На жаль, реалізувати техніку імітаційного моделювання аналогічно прикладу 9.6.1 у нашій теперішній задачі не вдасться: оскільки ми не знаємо λ, то неможливо згенерувати дані з розподілом, що відповідає основній гіпотезі.

У таких випадках стандартна рекомендація — об'єднувати комірки справа наліво з малою кількістю спостережень доти, поки відповідна частота не стане більшою або рівною 5. (Теоретичну частоту треба буде обчислювати як суму ймовірностей для об'єднаних комірок і при визначенні ступенів свободи r брати рівним кількості утворених об'єднаних комірок).

Ця рекомендація нагадує об'єднання зими з осінню у прикладі 9.6.1, яке ми визнали методично некоректним. Але є і відмінність: тепер ми не підганяємо дизайн за даними під гіпотезу, що нам подобається. Якщо рішення про об'єднання комірок за чітко визначеним алгоритмом приймається до того, як дані будуть отримані, і проводиться незалежно від того, чи сприятиме воно певній гіпотезі, чи ні — ми маємо право це робити. Це не є підгонкою статистики під бажаний результат. Звичайно, при такому підході ми відступаємо від теоретично обґрунтованої схеми тесту χ^2 і це може привести до певних неточностей. Практики стверджують, що при роботі з реальними даними ці неточності несуттєві. ◄

Для того, щоб перевірити узгодженість з одним із стандартних розподілів (пуассоніським, біноміальним, негативним біноміальним), можна скористатись функцією `goodfit()` з бібліотеки `vcd`.

Ця функція автоматично оцінює параметри розподілу, обчислює теоретичні частоти і проводить тест χ^2 для перевірки відповідності даних теоретичному розподілу.

Параметри цієї функції:

`x` — дані для аналізу — це може бути вектор, що містить вибірку з спостережуваними значеннями, або групована вибірка, записана у матриці або фреймі з двома стовпчиками: у першому мають бути частоти, а у другому — відповідні значення спостережуваної змінної;

`type` — тип розподілу, який перевіряється:

`"poisson"`, `"binomial"` або `"nbinomial"`;

`method` — метод перевірки гіпотези та оцінювання невідомих параметрів: `"ML"` відповідає тесту відношення вірогідності для тестування і методу найбільшої вірогідності для оцінки; `"MinChisq"` — тесту χ^2 для перевірки гіпотези і методу мінімуму χ^2 для оцінювання;

`par` — список значень параметрів розподілу (ті параметри, які не вказані у списку, оцінюються за даними).

Приклад 9.6.3. (Дані для цього прикладу взяті з [26], підрозд. 2.13, де вони наводяться з посиланням на роботу 1946 р. [23]).

Під час Другої світової війни німецька армія бомбардувала Британію крилатими ракетами Фау-1. Зокрема, у південну частину Лондона потрапило 535 ракет. Автора [23] цікавить, чи вдавалось німецьким військовим націлити ракети на якусь конкретну ціль, чи вони в межах південного Лондона падали цілком випадково?[43]

[43] Сьогодні таке питання звучить цілком академічно, але під час війни, коли, власне, проводилось це дослідження, воно мало цілком практичну мету. Якщо подивитись карту Лондона з позначеними на ній місцями падіння крилатих ракет, то можна помітити значні райони, куди не влучила ні одна ракета. З цього приводу висловлювались різні гіпотези, наприклад, про те, що у таких районах розміщені німецькі шпигунські об'єкти [34]. Тому важливо було переконатись, чи є хоч якась закономірність у цих падіннях, чи така картинка виникла випадково. (На жаль, я не знайшов у інтернеті карту для Фау-1, але аналогічний ефект "районів, вільних від бомбардування" можна помітити і на карті для Фау-2: `https://londonist.com/2013/06/v2`).

Для перевірки вся карта південного Лондона була розбита на 576 квадратиків площею 1/16 кв. км кожен. Всі місця попадання Фау-1 нанесли на карту і підрахували, скільки ракет потрапило у кожен квадратик. Результати вміщені у таблиці:

Кількість попадань	0	1	2	3	4	> 5
Кількість квадратиків	229	211	93	35	7	1

Можливі дві гіпотези:

(A) — велика кількість попадань у деякі квадратики пов'язана з тим, що ракети були спрямовані на конкретні цілі, розташовані у цих квадратиках (прицільне бомбардування);

(B) — різні кількості попадань у різні квадратики склались випадково, для всіх квадратиків ймовірність бути враженим ракетою була однакова (неприцільне бомбардування).

Приймемо гіпотезу (B) як основну — H_0. Їй відповідає припущення про те, що кількості попадань мають розподіл Пуассона (див. п. 6.3.2). Для перевірки цього скористаємось функцією `goodfit()`. У наступному скрипті ми спочатку створюємо фрейм з групованою вибіркою, а потім робимо перевірку гіпотези:

```
> library(vcd)
> hits<-0:5
> freq<-c(229, 211, 93, 35, 7, 1)
> x<-data.frame(freq,hits)
> gf<-goodfit(x,type= "poisson",method= "MinChisq")
> summary(gf)

  Goodness-of-fit test for poisson distribution
              X^2 df P(> X^2)
Pearson 1.168532  4 0.883252

> gf$par

$lambda
[1] 0.9312896
```

Результати перевірки гіпотези H_0 функцією `goodfit()` записані в об'єкт `gf`. Функція `summary(gf)` вивела їх на екран: досягнутий рівень значущості тесту дорівнює $p = 0.883252$ — слід прийняти основну гіпотезу про те, що бомбардування не було прицільним.

Оцінки значень невідомих параметрів функція `goodfit()` вказує в атрибуті `$par` свого результату, він виводиться останньою командою скрипта. Отже, наші дані добре описуються розподілом Пуасона з параметром $\lambda = 0.9312896$. ◄

9.6.3 Тест χ^2 з групуванням для перевірки узгодженості

Хоча χ^2-тести призначені для аналізу дискретних даних, їх можна застосовувати і для перевірки розподілу даних, що набувають довільних дійсних значень. Для цього початкові спостереження огрублюють, застосовуючи техніку групування, описану у підрозд. 4.5, і до групованих даних застосовують звичайний тест χ^2. Опишемо це детально.

Нехай дані $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ являють собою кратну вибірку, ξ_j — спостереження, що можуть набувати довільних числових значень. Функція розподілу ξ_j, тобто $G(x) = \mathsf{P}\{\xi_j < x\}$ вважається невідомою. Треба перевірити припущення, що ця функція належить деякому набору розподілів F_τ, $\tau \in T \subseteq \mathbb{R}^d$. Тут τ — невідомий параметр, або набір параметрів. Функції розподілу F_τ є неперервними.

Наприклад, F_τ може бути набором гауссових розподілів: $F_\tau(x) = \Phi((x - \mu)/\sigma)$, $\tau = (\mu, \sigma^2) \in \mathbb{R} \times (0, +\infty)$. У такому випадку відповідний тест буде тестом для перевірки нормальності (гауссовості).

Отже, ми перевіряємо основну гіпотезу H_0 : існує таке $\tau \in T$, що $G(x) = F_\tau(x)$ для всіх x.

Виберемо інтервал $[a, b]$, якому належать всі спостережувані значення ξ_j. Розіб'ємо цей інтервал на K підінтервалів $A_1, \ldots A_K$ однакової ширини $h = (b-a)/K$. Інтервали A_i, $i = 2, \ldots, K$ визна-

чаються як $A_i = (t_{i-1}, t_i]$, де $t_i = a + ih$, $A_1 = [t_1, t_2]$. Ці інтервали називають комірками (англ. *bin*). Підрахуємо $n_i = \#\{j : \xi_j \in A_i\}$ $i = 1, \ldots, K$ — емпіричні частоти комірок, тобто кількості попадань вибіркових значень у відповідні інтервали.

Теоретичні ймовірності визначаються як

$$p_i = F_{\hat{\tau}}(t_i) - F_{\hat{\tau}}(t_{i-1})$$

для $i = 2, \ldots, K-1$. Крайні інтервали при розрахунку ймовірностей розширюють так, щоб сума всіх ймовірностей дорівнювала 1:

$$p_1 = F_{\hat{\tau}}(t_1),\ p_K = 1 - F_{\hat{\tau}}(t_{K-1}).$$

Тут $\hat{\tau}$ — оцінка для невідомого параметра τ в припущенні, що виконується основна гіпотеза. Як $\hat{\tau}$ рекомендують обирати оцінку методу найбільшої вірогідності за групованими даними, тобто τ_n^{MLE}, визначене (9.16). Альтернативою може бути оцінка методу найбільшої вірогідності за початковими даними:

$$\hat{\tau}'_n = \operatorname*{argmax}_{\tau \in T} \sum_{j=1}^{n} \log f_\tau(\xi_j).$$

Використання τ_n^{MLE} теоретично більш коректне, оскільки забезпечує правильну асимптотичну поведінку відповідної статистики χ^2. Але на практиці, якщо $\hat{\tau}'_n$ підрахувати простіше, використовують саме її.

За теоретичними ймовірностями знаходять відповідні теоретичні частоти

$$\bar{n}_i = np_i.$$

Тепер статистика тесту і порогове значення розраховуються так само, як у випадку дискретних спостережень:

$$\chi^2_{emp} = \sum_{i=1}^{r} \frac{(n_i - np_i(\hat{\tau}_n))^2}{np_i(\hat{\tau}_n)},\ \chi^2_{th} = Q^{\chi^2_{K-d-1}}(1-\alpha).$$

Тест приймає H_0, якщо $\chi^2_{emp} \leq \chi^2_{th}$ і відхиляє, якщо $\chi^2_{emp} > \chi^2_{th}$.

Досягнутий рівень значущості тесту:

$$p = 1 - F^{\chi^2_{K-d-1}}(\chi^2_{emp}).$$

Приклад 9.6.4. У прикладі 7.2.1 ми розглядали дані про силу вітру з набору даних `airquality` і за гістограмою помітили можливі відхилення розподілу сили вітру від нормального. Перевіримо, наскільки значущими є ці відхилення, використовуючи тест χ^2.

У наступному скрипті для групування даних застосована функція `hist()`, яку зазвичай використовують для відображення гістограми (див. підрозд. 7.1). При встановленні опції `plot=FALSE` гістограма не виводиться, але функція робить всі розрахунки і її результатом є об'єкт, що в атрибуті `$breaks` містить точки розбиття на комірки (t_i), а в атрибуті `$counts` — емпіричні частоти комірок.

Математичне сподівання `m` та середньоквадратичне відхилення `sd` у скрипті оцінюються за групованими даними, причому при оцінці `sd` використовується поправка Шеппарда (4.7). Такий підхід є стандартним при перевірці нормальності:

```
> g = airquality$Wind # дані для аналізу
> # виконуємо групування:
> r<-hist(g,  breaks=10,plot = FALSE)
> nn<-r$counts   # емпіричні частоти
> tt<-r$breaks   # межі комірок
> h<-tt[2]-tt[1] # ширина комірки
> x<-tt[-length(tt)]+h/2 # середини комірок
> # оцінка математичного сподівання:
> m<-sum(x*nn)/sum(nn)
> # оцінка середньоквадратичного відхилення
```

```
> # з поправкою Шеппарда:
> s<-sqrt(sum((x-m)^2*nn)/sum(nn)+h^2/12)
> pp<-pnorm(tt,mean=m,sd=s) # теоретичні ймовірності
> # розширюємо крайні комірки:
> pp[c(1,length(tt))]<-c(0,1)
> # теоретичні частоти:
> nth<-length(g)*(pp[-1]-pp[-length(pp)])
> # статистика тесту хі-квадрат:
> chi2emp<-sum((nn-nth)^2/nth)
> # досягнутий рівень значущості:
> 1-pchisq(chi2emp,df=length(tt)-4)

[1] 0.0007148321
```

△

Ми отримали досягнутий рівень значущості $p = 0.0007148$, отже, основну гіпотезу треба відхилити: розподіл даних значущо відрізняється від нормального. ◀

Аналогічно можна проводити перевірку гіпотез про узгодженість розподілу даних з іншими теоретичними розподілами.

9.6.4 Перевірка незалежності двох змінних тестом χ^2

Ще одна дуже важлива сфера застосування тестів χ^2 — перевірка наявності чи відсутності залежності між спостережуваними змінними.

Нехай для кожного об'єкта, що досліджується, спостерігаються дві змінні A і B, кожна з яких може набувати лише значень з фіксованого набору — $a_1, \ldots, a_I$ для змінної A і $b_1, \ldots, b_K$ для змінної B. Треба за кратною вибіркою зі спостережень n об'єктів $\mathbf{X} = \{(A_j, B_j),\ j = 1, \ldots, n\}$ перевірити гіпотезу H_0: A і B є незалежними випадковими величинами.

Позначимо $p_{ik} = \mathsf{P}\{A = a_i,\ B = b_k\}$, $i = 1, \ldots, I$, $k = 1, \ldots, K$ — невідомий спільний розподіл пари (A, B). Гі-

потеза H_0 еквівалентна твердженню

$$H_0 : p_{ik} = p_{i\cdot}p_{\cdot k}, \text{ для всіх } i = 1, \dots, I,\ k = 1, \dots, K,$$

де $p_{i\cdot} = \mathsf{P}\{A = a_i\} = \sum_{k=1}^{K} p_{ik}$, $p_{\cdot k} = \mathsf{P}\{B = b_k\} = \sum_{i=1}^{I} p_{ik}$.

Таким чином, основній гіпотезі відповідає розподіл, що визначається параметром $\tau = (p_{i\cdot}, i = 1, \dots, I; p_{\cdot k}, k = 1, \dots, K)$. Використовуючи загальну схему тестів χ^2, розраховуємо емпіричні частоти

$$n_{ik} = \#\{j :\ A_j = a_i, B_j = b_k\}, i = 1, \dots, I,\ k = 1, \dots, K,$$

— кількість тих спостережень, для яких пара змінних (A_j, B_j) дорівнює значенню (a_i, b_k). За цими даними оцінюємо невідомий параметр τ, використовуючи метод найбільшої вірогідності. Отримуємо оцінки

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n},\ \hat{p}_{\cdot k} = \frac{n_{\cdot k}}{n}, i = 1, \dots, I,\ k = 1, \dots, K.$$

Розраховуємо теоретичні частоти: $\bar{n}_{ik} = n\hat{p}_{i\cdot}\hat{p}_{\cdot k}$. Складаємо статистику тесту:

$$\chi^2_{emp} = \sum_{i=1}^{I}\sum_{k=1}^{K} \frac{(n_{ik} - \bar{n}_{ik})^2}{\bar{n}_{ik}}.$$

При визначенні порогу тесту (та досягнутого рівня значущості) слід враховувати, що компоненти невідомого параметра τ не є незалежними: $\sum_{i=1}^{I} p_{i\cdot} = 1$ і $\sum_{k=1}^{K} p_{\cdot k} = 1$. Тому маємо лише $d = I + K - 2$ незалежних параметрів. Кількість різних комірок (значень, яких можуть набувати спостережувані пари змінних) $r = IK$. Отже, за формулою (9.17) кількість ступенів вільності для χ^2 визначається як

$$\mathrm{df} = r - d - 1 = (I - 1)(K - 1).$$

Поріг тесту

$$\chi^2_{th} = Q^{\chi^2_{\mathrm{df}}}(1 - \alpha).$$

Тест приймає основну гіпотезу (про незалежність), якщо $\chi^2_{emp} \le \chi^2_{th}$ і відхиляє її, якщо $\chi^2_{emp} > \chi^2_{th}$.

Досягнутий рівень значущості

$$p = 1 - F^{\chi^2_{\text{df}}}(\chi^2_{emp}).$$

У R реалізувати χ^2-тест для перевірки незалежності можна, використовуючи функцію `chisq.test()`. При цьому задавати її параметри треба не так, як ми це робили у підрозд. 9.6.1. Розглянемо це на прикладі з реальними даними.

Приклад 9.6.5. У фреймі даних `survey` з бібліотеки `MASS` містяться дані опитування 237 студентів університету Аделаїди. Зокрема, у змінній `Smoke` містяться відповіді цих студентів на запитання про те, як багато вони палять сигарет (`Heavy` — багато, `Regul` — регулярно, `Occas` — зрідка і `Never` — ніколи), а у змінній `Exer` — відповіді про те, як часто вони роблять фізичні вправи (`Freq` — часто, `Some` — інколи, `None` — ніколи). Нас цікавить: чи є залежність між звичкою до паління і регулярністю виконання фізичних вправ?

Щоб з'ясувати це побудуємо табличку, де вказуються частоти всіх можливих пар відповідей на ці запитання. Зробимо це за допомогою функції `table()`:

```
> library(MASS)
> tbl = table(survey$Smoke, survey$Exer)
> tbl

        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7
```

Тепер у змінній `tbl` міститься таблиця (матриця) емпіричних частот пар відповідей. Наприклад, варіант відповідей `(Never,Some)` обрали 84 студенти серед всіх 237, що взяли участь у опитуванні. Такі таблиці прийнято називати **таблицями спряженості** змінних/ознак (англ. *contingency table*). Якщо змінні незалежні, то розподіли студентів за змінною `Exer` при фіксованій змінній `Smoke` мають бути приблизно однаковими для всіх значень `Smoke`. Тобто всі рядки таблиці повинні бути приблизно пропорційними (відрізнятись лише сталими множниками). Те ж має виконуватись і для стовпчиків.

Для аналізу таблиць спряженості зручно використовувати графічне зображення, яке називають мозаїчною діаграмою (*mosaic plot*). На таких діаграмах частоти зображаються у вигляді прямокутників таким чином. Спочатку рисують вертикальні стовпчики, ширина яких пропорційна частотам $n_{i\cdot}$. (Нагадаємо, що $n_{i\cdot}$ це кількість об'єктів у вибірці, що мають i-те значення першої характеристики — у нашому прикладі це `Smoke`). Потім кожен (i-й, $i = 1, \ldots, I$) стовпчик розбивається на прямокутники, висоти яких пропорційні n_{ik}, $k = 1, \ldots, K$. Отримуємо "мозаїку", що відображає таблицю спряженості. У `R` це можна зробити, використовуючи функцію `\mosaicplot()` з пакета `MASS`[44]:

```
>    # Мозаїчна діаграма для даних з прикладу:
> # Скорочуємо назви факторів до однієї літери:
> levels(survey$Exer)<-c("F","N","S")
> levels(survey$Smoke)<-c("H","N","O","R")
> # Відображаємо мозаїчну діаграму:
> mosaicplot(~Smoke+Exer,data=survey,
+       col=c("red","yellow","blue"),
+       main="",cex=1.5)
>    # Генерація незалежних даних
```

[44] Тут ми застосували варіант цієї функції, в якому таблиця задається формулою вигляду `~x+y`, де `x` `y` — змінні, залежність яких досліджується. Параметр `data` вказує фрейм даних, у якому містяться ці змінні.

```
>    # і відображення на діаграмі:
>    set.seed(3)
>    independent<-data.frame(x<-rbinom(500,3,0.6),
+                            y<-rbinom(500,2,0.4))
>    mosaicplot(~x+y,data=independent,
+          col=c("red","yellow","blue"),
+          main="",cex=1.5)
```

△

(Результат на рис. 9.7, 9.8). На рис. 9.7 мозаїка для даних з нашого прикладу. (Щоб не робити рисунок занадто великим, назви рівнів факторів замінені їхніми першими літерами — `Heavy` замінено на `H` і т.д.).

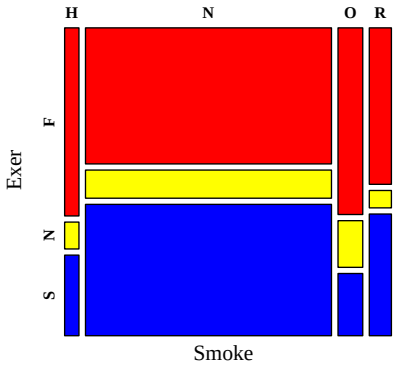

Рис. 9.7. Мозаїчна діаграма для даних про куріння і фізичні вправи

Для порівняння, на рис. 9.8 — мозаїка, що утворюється за змінними, які були згенеровані незалежно одна від одної. Незалежним змінним відповідають мозаїки, в яких горизонтальні сторони відповідних прямокутників утворюють майже суцільні прямі лінії.

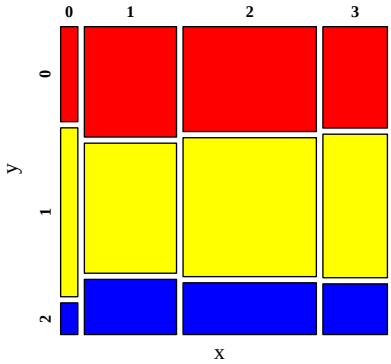

Рис. 9.8. Мозаїчна діаграма для модельованих незалежних змінних

Проглядаючи таблицю і мозаїчну діаграму, бачимо, що строгої пропорційності немає, але і надзвичайно сильних відхилень не помітно.

Застосуємо тест χ^2:

```
> chisq.test(tbl)


        Pearson's Chi-squared test

data:  tbl
X-squared = 5.4885, df = 6, p-value = 0.4828
```

Функція обчислила статистику $\chi^2_{emp} = 5.4885$, визначила кількість ступенів вільності `df`=6 і знайшла досягнутий рівень значущості `p-value` =0.4828. Отже, треба прийняти основну гіпотезу про незалежність між звичками до паління та до виконання фізичних вправ.

Можна було передати функції як параметр не таблицю спряженості, а безпосередньо змінні, що нас цікавлять:

```
chisq.test(survey$Smoke, survey$Exer)
```

Функція сама обчислить емпіричні частоти і видасть ті ж значення, що й у попередньому випадку.

Зазначимо, що серед емпіричних частот зустрічаються менші за 5, тому покладатись на результат стандартного тесту χ^2 не можна. Доцільно використати техніку імітаційного наближеного розрахунку p-value:

```
> set.seed(3)
> chisq.test(tbl,simulate.p.value =T)


        Pearson's Chi-squared test with simulated
        p-value (based on 2000 replicates)

data:  tbl
X-squared = 5.4885, df = NA, p-value = 0.4653
```

Отримали p-value=0.4653, тобто знову слід прийняти гіпотезу про відсутність залежності. ◄

9.7 Перевірка залежності двох змінних

У п. 9.6.4 ми розглянули варіант тесту χ^2, який дозволяє перевірити, чи є залежними дві змінні, кожна з яких набуває лише фіксованої кількісті значень (категорійні змінні). Дуже часто у прикладних дослідженнях виникає необхідність перевірити залежність між змінними числової природи або між числовою і категорійною змінною. Звичайно, у таких випадках можна групувати числові змінні для отримання категорійних, як ми це робили у

п. 9.6.3. Але при такому підході дані огрублюються і частину корисної інформації можна втратити. Тому доцільно мати спеціальні тести для перевірки залежності між некатегорійними змінними.

Далі у п. 9.7.1 розглянемо тести для перевірки того, чи залежить математичне сподівання або дисперсія деякої числової змінної (відгуку) від значень категорійної змінної (фактора). (Це називають перевіркою впливу фактора на середні значення або дисперсію відгуку). Розділ статистики, присвячений таким перевіркам, називають дисперсійним аналізом.

У п. 9.7.2 розглядається задача кореляційного аналізу, коли треба перевірити, чи є залежність між змінними числової природи або між змінними, що задають порядок на спостережуваному наборі даних (рангами). З кореляціями як дескриптивними характеристиками залежності ми вже ознайомились у розд. 5, зокрема, у підрозд. 5.5 був неформально описаний тест на основі кореляції Пірсона. Тепер ми повернемось до питання перевірки залежності на більш глибокій ймовірнісній основі.

Нарешті, у п. 9.7.3 ми порівняємо роботу різних тестів на модельованих даних і спробуємо пояснити, в яких випадках який тест краще застосовувати.

9.7.1 Однофакторний дисперсійний аналіз

Дисперсійний аналіз (скорочено ДА, англ. *Analysis of variance, ANOVA*[45])— це сукупність статистичних методів, призначених для того, щоб аналізувати залежність числової змінної, що характеризує досліджуваний об'єкт, (відгуку) від факторів категорійної природи, які впливають на цей об'єкт. Наприклад, вага (відгук) однорічної свині, яку відгодовують, може залежати від її породи (перший фактор) і дієти, яка використовується при відгодівлі (другий фактор). У цьому підрозділі ми обмежимось виявленням залежностей від одного фактора, відповідні методи називають одно-

[45]Скорочення ANOVA часто використовується без пояснення в українсько- та російськомовній літературі, тому його варто пам'ятати.

факторним ДА (*one-way ANOVA*). Ми будемо розглядати модель однофакторного ДА з фіксованими ефектами.

Отже, нехай спостерігається n об'єктів, на які впливає фактор, що може набувати значень з набору M різних значень. У ДА такі різні можливі значення фактора називають рівнями[46]. Нас цікавить вплив цього фактора на числову змінну Y, значення якої відоме для всіх спостережуваних об'єктів.

Перенумеруємо всі рівні фактора індексом $i = 1, \ldots, M$. Розіб'ємо весь набір спостережуваних об'єктів (вибірку) на групи (підвибірки), що відповідають кожному можливому рівню фактора. Нехай n_i — кількість об'єктів в i-й групі. Тоді $n = \sum_{i=1}^{M} n_i$. Перенумеруємо об'єкти в кожній групі індексом $j = 1, \ldots, n_i$. Таким чином, кожен об'єкт характеризується парою індексів i, j. Позначимо Y_{ij} — значення змінної Y у j-го об'єкта з підвибірки, що відповідає i-тому рівню фактора. (Напр., якщо досліджується вплив породи на вагу свиней, Y_{ij} — вага j-ї свині серед свиней i-ї породи).

У класичній моделі однофакторного ДА з фіксованими ефектами вважається, що спостереження $Y_{ij}, i = 1, \ldots, M, j = 1, \ldots, n_i$ є незалежними гауссовими випадковими величинами, причому дисперсія всіх Y_{ij} однакова (її позначають σ^2), а математичні сподівання μ_i $i = 1, \ldots, M$ однакові всередині кожної групи, але для різних груп можуть бути різними.

Таким чином,

$$Y_{ij} \sim N(\mu_i, \sigma^2). \tag{9.20}$$

Якщо досліджуваний фактор не впливає на розподіл відгуку, то всі математичні сподівання мають бути однаковими:

$$\text{Гіпотеза } H^{\mu}: \quad \mu_1 = \mu_2 = \cdots = \mu_M. \tag{9.21}$$

Припущення (9.21) називають **гіпотезою про однорідність (рівність) середніх**. Така назва пов'язана з тим, що μ_i трактують як

[46] Звідси походить і відповідна термінологія `R`.

середнє значення змінної Y по "всій популяції" об'єктів, які відповідають i-му рівню фактора[47]. Скажімо, у нашому прикладі, μ_i — середнє значення, навколо якого коливається вага однорічних свиней i-ї породи після відгодівлі.

Обчислимо вибіркові середні по кожній підвибірці:

$$\bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \ .$$

Зрозуміло, що $\bar{Y}_{i\cdot}$ будуть різними навіть при виконанні гіпотези H^{μ} внаслідок випадкових коливань спостережень. Отже, для того, щоб можна було обґрунтовано відхилити H^{μ}, треба, щоб $\bar{Y}_{i\cdot}$ досить сильно відрізнялись одне від одного при деяких i. F-Тест Фішера, який ми будемо використовувати, можна розглядати саме як спосіб визначити, коли відмінності між $\bar{Y}_{i\cdot}$ слід вважати достатньо сильними, щоб від них не можна було відмахнутись, пославшись на випадкові коливання. З іншого боку, цей тест вкладається у загальну схему статистичних тестів: він є тестом відношення вірогідності для перевірки основної гіпотези $H_0 = H^{\mu}$ проти альтернативи

$$H_1 : \text{ існує така пара індексів } i \neq k, \text{ що } \mu_i \neq \mu_k.$$

Тест Фішера для однофакторного ДА використовує статистику, яка називається F-емпіричне:

$$F_{emp} = \frac{\frac{1}{M-1} \sum_{i=1}^{M} n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{\frac{1}{n-M} \sum_{i=1}^{M} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2}. \tag{9.22}$$

Тут $\bar{Y}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{M} \sum_{j=1}^{n_i} Y_{ij}$ — загальне середнє відгуку за всіма спостереженнями (середня вага всіх свиней всіх порід у нашій вибірці).

[47] Тобто всіх можливих таких об'єктів, а не тільки тих, що потрапили до вибірки.

Коли H_0 виконано, то при достатньо великих n_i всі $\bar{Y}_{i\cdot} \approx m_i$ будуть приблизно однаковими і близькими до $\bar{Y}_{\cdot\cdot}$. Тому сума у чисельнику (9.22) буде близькою до 0. Ця сума називається **міжгруповою сумою квадратів**. Вона показує, наскільки відхиляються одне від одного середні, розраховані за групами (підвибірками), що відповідають різним рівням фактора.

Сума у знаменнику (9.22) називається **внутрішньогруповою сумою квадратів**. Вона показує, як сильно можуть коливатись відхилення значень відгуку від середнього у кожній групі окремо.

Таким чином, обчислюючи F_{emp}, ми порівнюємо міжгруповий розкид середніх з внутрішньогруповими коливаннями спостережень. Великі значення F_{emp} мають свідчити на користь альтернативи, малі — на користь основної гіпотези. З іншої точки зору, статистика F_{emp} еквівалентна статистиці відношення вірогідності для задачі перевірки H_0 проти H_1 (див. [11]).

При виконанні H_0 статистика F_{emp} має розподіл Фішера з $M-1$ ступенем вільності чисельника і $n-M$ ступенями вільності знаменника. Тому, для того, щоб тест мав заданий рівень значущості, її треба порівнювати з порогом

$$F_{theor} = Q^{F(M-1,n-M)}(1-\alpha),$$

який називається F-теоретичне.

Остаточно тестову процедуру можна сформулювати так:

Якщо $F_{emp} \le F_{theor}$ — прийняти H_0 (залежність середніх від фактора не виявлена),
якщо $F_{emp} > F_{theor}$ — прийняти H_1 — залежність виявлена.

Допитливий читач міг помітити, що у випадку, коли фактор має точно два рівні, гіпотеза про однорідність середніх перетворюється на гіпотезу $H_0 : \mu_1 = \mu_2$, яку ми у підрозд. 9.5.5 перевіряли Т-тестом Стьюдента. Можна показати, що у цьому випадку Т-тест і

F-тест еквівалентні.

Приклад 9.7.1. У прикладі 7.5.1 ми розглядали дані випробувань різних інсектицидів з фрейма `InsectSprays`. Нагадаємо, що у `InsectSprays` містяться змінні `spray` — тип інсектициду (A-F) та `count` — кількість комах, які були вбиті даним інсектицидом у відповідному досліді. Чим більша кількість вбитих комах, тим кращою можна вважати дію інсектициду. На рис. 7.13 ми бачили, що деякі типи інсектицидів виявились помітно ефективнішими, ніж інші. Перевіримо, наскільки статистично значущим є це спостереження, застосовуючи техніку однофакторного ДА[48].

Для застосування F-тесту Фішера скористаємось функцією `aov()`. Модель для аналізу задається формулою у форматі *відгук*~*фактор*, а фрейм, в якому розміщені дані, вказується в опції `data`. Результат роботи функції можна зберегти у якій-небудь змінній, а можна одразу вивести функцією `summary()`:

```
> summary(aov(count~spray,data=InsectSprays))

            Df  Sum Sq Mean Sq F value Pr(>F)
spray        5    2669   533.8    34.7 <2e-16 ***
```

[48] Строго кажучи, для цього треба бути впевненим, що дані відповідають класичній моделі (9.20). Однорідність дисперсій ми перевіримо пізніше. Припущення щодо нормальності розподілу вбитих комах є природним. Якщо вважати, що кожна комаха має певну ймовірність загибелі у досліді і гине незалежно від інших, нормальність буде наслідком центральної граничної теореми за умови, що загинуло досить багато комах. Але, якщо кількості загиблих в одному досліді невеликі, доцільно використовувати пуассонівську граничну теорему, тобто вважати що дані мають розподіл Пуассона. Поглянувши на дані (а це перше, що має зробити статистик!) бачимо, що для інсектицидів C, D і E варто застосовувати саме пуассонове наближення. Однак ми продовжимо використання класичного ДА у цьому прикладі. Такий підхід доволі поширений, оскільки ДА є популярною стандартною технікою, до якої звикло багато спеціалістів у прикладних областях. У більшості випадків він не приводить до занадто хибних висновків (див. зауваження наприкінці цього підрозділу). Читач може самостійно провести дослідження цих даних на основі пуассонової моделі, використовуючи метод найбільшої вірогідності, як описано у п. 9.3.3.

```
Residuals   66   1015    15.4
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

△

У таблиці результатів стовпчик, позначений `Df`, містить ступені вільності чисельника $M - 1 = 5$ і знаменника $n - M = 66$. У стовпчику `Sum Sq` записані міжгрупова (2669) і внутрішньогрупова (1015) суми квадратів. Чисельник (533.8) і знаменник (15.4) F_{emp} вміщені у наступному стовпчику, а сама статистика $F_{emp} = 34.7$ записана далі.

Для спеціаліста у прикладній області ці значення можуть бути цікаві, але основний результат тесту вміщений у останньому стовпчику праворуч — це досягнутий рівень значущості тесту `<2e-16`, тобто майже 0. Таким чином, для будь-якого розумного рівня значущості за F-тестом Фішера слід прийняти альтернативу — математичні сподівання кількості вбитих комах є різними для інсектицидів різного типу.

Звичайно, тепер доцільно подивитись, як саме відрізняються ці характеристики ефективності інсектицидів. Для цього ми у підрозд. 9.4 застосовували системи одночасних довірчих інтервалів. Для однофакторного ДА такі інтервали можна будувати, використовуючи функцію `groupwiseMean()` з пакета `rcompanion`, як це зроблено у наступному скрипті:

▽

```
> library(plotrix)
> library(rcompanion)
> # Знаходимо середні значення
>      # і межі довірчих інтервалів:
> CI<-groupwiseMean(count~spray,data=InsectSprays,
+                   conf=(0.95)^(1/6))
> # Відображаємо інтервали:
> plotCI(1:6,y=CI$Mean,ui=CI$Trad.upper,
+ li=CI$Trad.lower,
```

```
+ xlab=" ",ylab="counts",xaxt="n")
> # виводимо горизонтальну вісь
> # з позначеннями рівнів фактора:
> axis(1,at=1:6,labels=levels(CI$spray))
```

△

(Зверніть увагу, що ми побудували одночасні інтервали з рівнем значущості $\alpha = 0.05$ скориставшись формулою (9.11) з $m = 6$ для визначення параметра `conf`$=1-\alpha$).

На рис. 9.9 зображені отримані довірчі інтервали.

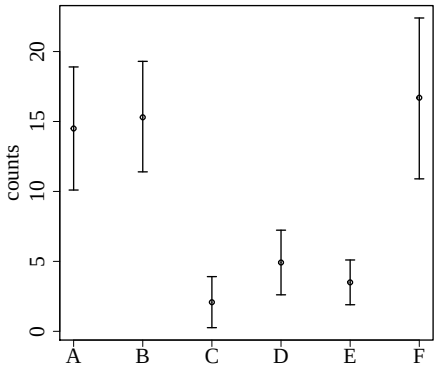

Рис. 9.9. Довірчі інтервали для математичних сподівань

Помітно, що інсектициди розбились на дві групи: (1) A, B, F і (2) C, D, E. Інтервали для інсектицидів з різних груп не перетинаються, отже, їхні ефективності (математичні сподівання кількості загиблих комах) різні. Всередині кожної групи інтервали мають спільну точку, отже, у нас немає підстав для того, щоб розрізняти ефективності відповідних інсектицидів. У першої групи ефективність значущо, більша ніж у другої. ◄

F-тест базується на припущенні, що дисперсія відгуку однакова для всіх рівнів фактора. Часто це припущення не можна апріорі

ні прийняти, ні відхилити. У таких випадках природно спробувати перевірити гіпотезу про однорідність дисперсій.

А саме, припустимо, що спостереження описуються моделлю

$$Y_{ij} \sim N(\mu_i, \sigma_i^2),$$

в якій дисперсії σ_i^2, що відповідають різним рівням фактора, є невідомими і можуть набувати довільних значень.

Основна гіпотеза, яку ми перевіряємо — **гіпотеза про однорідність дисперсій**:

$$H_0 = H^\sigma : \ \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_M^2.$$

Альтернатива H_1 полягає в тому, що існують хоча б дві не рівні одна одній дисперсії.

Для перевірки такої гіпотези можна застосовувати різні тести, з яких найбільш популярним є тест Левена. В основі цього тесту лежить використання статистики F_{emp}, в якій замість початкових даних Y_{ij} використовуються абсолютні відхилення Y_{ij} від медіан відповідних підвибірок. Більш докладно цей тест описаний у підрозд. 2.5 книги [11]. Тут ми обмежимось лише прикладом його використання в R.

Приклад 9.7.2. Продовжимо розгляд даних з прикладу 9.7.1. Перевіримо, чи є однаковими дисперсії кількостей загиблих комах для різних типів інсектицидів. Для цього використаємо функцію `leveneTest()` з пакета `car`. Застосування цієї функції подібне до застосування `aov()`, але перевіряє вона однорідність дисперсій:

```
> library(car)
> leveneTest(count~spray,data=InsectSprays)

Levene's Test for Homogeneity of Variance
(center = median)
      Df F value   Pr(>F)
group  5  3.8214 0.004223 **
```

```
      66
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Значення статистики тесту дорівнює 3.8214, досягнутий рівень значущості p =0.004223. При традиційному для нас $\alpha = 0.05$ слід прийняти гіпотезу про неоднорідність дисперсій.

Таким чином, у нас є серйозні підстави для сумнівів у результаті F-тесту. Можливо, його висновок про неоднорідність середніх не є обґрунтованим. Але використана нами техніка перевірки однорідності середніх за допомогою довірчих інтервалів не залежить від однорідності дисперсій (кожен довірчий інтервал будується на основі вибіркової дисперсії по своїй підвибірці і не пов'язаний з іншими підвибірками). Тому ми можемо зробити остаточний висновок про неоднорідність середніх. ◄

Після цього прикладу у читача може залишитись враження, що F-тест взагалі не потрібен: адже тест довірчих інтервалів працює у значно більш широкій області. Але насправді довірчі інтервали є значно менш точним інструментом, ніж F-тест. Вони дають можливість помітити лише досить грубі відмінності. F-тест дозволяє помітити менш сильно виражену неоднорідність, але, на жаль, не може підказати, які саме математичні сподівання відрізняються.

Зауваження. Зупинимось на ще одному питанні: чи можна використовувати F-тест у випадку, коли розподіл спостережень не є нормальним, як ми це робили у прикладі 9.7.1? Відповідь умовно-позитивна: якщо обсяги всіх підвибірок, які відповідають різним рівням фактора, достатньо великі, то розподіл статистики F_{emp} при H_0 буде близьким до розподілу Фішера[49]. Тому F-тест і в цьому випадку буде забезпечувати правильний рівень значущості.

Однак слід мати на увазі, що для нормально розподілених даних F-тест є найбільш потужним у правильному класі тестів (тобто

[49]Точніше, необхідно іще, щоб дисперсії відгуку існували і були однаковими для всіх підвибірок.

найкращим з точки зору мінімізації ймовірності помилки). Для даних, які мають інші розподіли, можуть існувати тести однорідності, більш потужні, ніж F-тест. Наприклад, хорошим претендентом на роль такого більш потужного тесту може бути тест відношення вірогідності.

9.7.2 Тести кореляцій

Для того, щоб перевірити наявність чи відсутність залежності між двома числовими змінними, можна скористатись тестами на основі коефіцієнтів кореляції. Ми вже познайомились із такими тестами у підрозд. 5.5, тому тут залишається тільки розглянути більш формальну їхню теорію.

Отже, нехай у кожного об'єкта, що спостерігаються, є дві числові змінні X_j і Y_j. Ми хочемо перевірити, чи є ці змінні незалежними.

Якщо вважати, що (X_j, Y_j) $j = 1, \ldots, n$ незалежні, однаково розподілені вектори з сумісним нормальним розподілом, то найкращий тест для перевірки незалежності будується на основі коефіцієнта кореляції Пірсона.

Для сумісно нормальних[50] X_1 і Y_1 гіпотеза про їхню незалежність як випадкових величин, еквівалентна припущенню про **некорельованість**:

$$H_0 : \operatorname{cor}(X_1, Y_1) = 0,$$

де

$$\operatorname{cor}(X_1, Y_1) = \frac{\mathsf{E}(X_1 - \mathsf{E}\, X_1)(Y_1 - \mathsf{E}\, Y_1)}{\sqrt{\mathsf{D}\, X_1 \, \mathsf{D}\, Y_1}}$$

— теоретичний коефіцієнт кореляції[51].

[50] Тобто коли (X_1, Y_1) — гауссів нормальний вектор (див. Додаток Б.3).

[51] Оскільки (X_j, Y_j) вважаються однаково розподіленими векторами, $\operatorname{cor}(X_j, Y_j)$ не залежить від j. Для визначеності позначаємо цю кореляцію $\operatorname{cor}(X_1, Y_1)$.

Вибірковий коефіцієнт кореляції Пірсона

$$r(X,Y) = \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})(Y_j - \bar{Y})}{\sqrt{S^2(X)S^2(Y)}}$$

можна розглядати як оцінку для "справжньої" теоретичної кореляції $\mathrm{cor}(X_1, Y_1)$.

Тест для перевірки H_0 проти альтернативи $\mathrm{cor}(X_1, Y_1) \neq 0$ (залежність між змінними є) полягає у порівнянні кореляції Пірсона з пороговим значенням C_α: гіпотеза про незалежність приймається, якщо $|r(X,Y)| \leq C_\alpha$.

За нашою загальною схемою побудови статистичних тестів поріг C_α треба обрати так, щоб при виконанні H_0 справжнювалась рівність

$$\mathsf{P}\{|r(X,Y)| \geq C_\alpha\} = \alpha,$$

де α — заданий стандартний рівень значущості тесту.

Можна показати, що у випадку гауссових спостережень цю умову задовольняє

$$C_\alpha = \sqrt{\frac{f_\alpha/(n-2)}{1 + f_\alpha/(n-2)}}, \tag{9.23}$$

де $f_\alpha = Q^{F(1,n-2)}(1-\alpha)$ (див. [11], підрозд. 2.4).

Таким чином, тест набуває такого вигляду:

> Якщо $|r(X,Y)| > C_\alpha$ — приймаємо гіпотезу про залежність X і Y.
> Якщо $|r(X,Y)| \leq C_\alpha$ — вважаємо, що залежність не виявлена.

Це — той самий тест, що був описаний у підрозд. 5.5.

Досягнутий рівень значущості цього тесту

$$p(r(X,Y)) = 1 - G\left(\frac{r^2(X,Y)(n-2)}{1 - r^2(X,Y)}\right),$$

де G — функція розподілу F-розподілу Фішера $F(1, n-2)$.

Якщо розподіл даних не є гауссовим, то з некорельованості не випливає незалежність: теоретичний коефіцієнт кореляції може бути нульовим у залежних величин. Тому тест на основі кореляції Пірсона не помічає багато різних реально існуючих залежностей. Але його можна використовувати як тест для перевірки саме некорельованості, тобто відсутності залежностей, близьких до лінійних (див. обговорення у підрозд. 5.2). При цьому вибір порогу C_α за (9.23) забезпечує приблизний рівень значущості α для великих обсягів вибірки n.

Для негауссових спостережень зручно застосовувати тести на основі рангових коефіцієнтів кореляції. Такі тести називають ранговими тестами незалежності. Найбільш популярні з них використовують ρ Спірмена або τ Кендалла, які ми розглядали у підрозд. 5.4. Влаштовані вони аналогічно розглянутому вище тесту з кореляцією Пірсона: беруть абсолютну величину того чи іншого рангового коефіцієнта кореляції і порівнюють її з порогом. Поріг обирають так, щоб рівень значущості тесту дорівнював заданому α.

Наприклад, при великих обсягах вибірки n ($n > 30$) поріг для тесту на основі ρ Спірмена можна визначити за формулою (9.23), підставляючи в неї $\rho(X, Y)$ замість $r(X, Y)$. Аналогічно визначається досягнутий рівень значущості тесту.

Зручність цих тестів пов'язана з тим, що, при виконанні гіпотези про незалежність X і Y, їхні розподіли не залежать від функцій розподілу X та Y[52]. Тому для того, щоб знайти поріг рангового тесту, не треба знати, яким є розподіл кожної змінної. Тести, що мають цю властивість називають **тестами, незалежними від розподілу**.

Але що насправді перевіряють такі тести? Можна показати, що при зростанні $n \to \infty$:

$$\rho(X, Y) \to \operatorname{cor}(F_X(X_1), F_Y(Y_1)) =$$

[52] Якщо ці функції неперервні.

$$= 3(2\,\mathsf{P}[(X_2 - X_1)(Y_3 - Y_1) > 0] - 1) = \rho_\infty(X, Y),$$

і

$$\tau(X, Y) \to 2\,\mathsf{P}[(X_2 - X_1)(Y_2 - Y_1) > 0] - 1 = \tau_\infty(X, Y),$$

де F_X і F_Y — функції розподілу X і Y, (X_j, Y_j) — незалежні копії вектора (X, Y) (див. статтю "Spearman rank correlation coefficient" у [27]).

Таким чином, тест на основі кореляції Спірмена можна розглядати як тест для перевірки основної гіпотези

$$H_0^\rho : \ \rho_\infty(X, Y) = 0,$$

проти альтернативи $\rho_\infty(X, Y) \neq 0$. Легко перевірити, що для незалежних випадкових величин $\rho_\infty(X, Y) = 0$. Але це співвідношення може виконуватись і для залежних X і Y.

Тому тест незалежності на основі коефіцієнта Спірмена не буде помічати деякі залежності. Аналогічні міркування правильні і для τ Кендалла. Далі у п. 9.7.3 ми розглянемо кілька демонстраційних прикладів того, які саме залежності можна чи не можна побачити за допомогою рангових кореляцій.

Особливо корисні рангові тести незалежності у випадку, коли досліджувана змінна насправді вказує лише порядок елементів у вибірці, а її точні числові значення не мають глибокого змісту (виміряна у "порядковій шкалі", див. підрозд. 5.4).

У `R` тести незалежності на основі r, ρ і τ реалізуються за допомогою функції `cor.test()`. Її застосування до реальних даних розглянуто у прикладі 5.5.1. Приклади роботи цієї функції на модельованих даних див. у п. 9.7.3.

9.7.3 Порівняння різних підходів до пошуку залежності

Отже, для перевірки залежності двох числових змінних можна використовувати:

— тести на основі кореляцій Пірсона, Спірмена або Кендалла (п. 9.7.2),

— F-тест для перевірки однорідності середніх з групуванням однієї змінної (п. 9.7.1),

— тест Левена для перевірки однорідності дисперсій з групуванням однієї змінної (п. 9.7.1),

— тест χ^2 для перевірки незалежності з групуванням обох змінних (9.6.4).

Який з цих тестів кращий? Відповідь залежить від ряду обставин: від обсягу вибірки, від того, яку саме залежність ми сподіваємось побачити, від можливої наявності викидів-забруднень. Далі ми розглянемо кілька прикладів аналізу спеціально згенерованих даних, що показують особливості застосування тестів незалежності у різних випадках. Аналізовані змінні у прикладах мають назви `x` і `y` (у скриптах — з номерами, що відповідають номеру прикладу).

В усіх цих прикладах використовуються стандартні тести на основі кореляцій, що реалізовані у функції `cor.test()`.

Для тестів дисперсійного аналізу (F-тесту і тесту Левена) ми проводимо групування за змінною `x` на три групи спостережень: з малими (`low`), проміжними (`mid`) і великими (`high`) значеннями `x`. Для цього вся область спостережуваних значень `x` розбивається на три підінтервали однакової довжини. До кожної групи належать спостереження, у яких `x` потрапляє на відповідний інтервал. Далі F-тест перевіряє гіпотезу про рівність середніх `y` у спостережень з цих трьох груп, а тест Левена — рівність відповідних дисперсій.

Групування виконується за допомогою функції `cut()` (див. п. 2.2.3).

Для χ^2-тесту аналогічне групування проводиться також за змінною `y` і використовуються дані груповані по обох змінних. Таким чином, тест перевіряє, чи залежить розподіл спостережень по групах за змінною `y` від того, в яку групу за `x` потрапляють ці спостереження.

У наступному скрипті знаходження досягнутих рівнів значущості всіх цих тестів зібране в одній функції `CompareTests()`,

котра видає їх у вигляді іменованого вектора:

```
> CompareTests<-function(x,y){
+ xc<-cut(x,breaks=seq(min(x),max(x),length.out=4),
+ labels=c("low","mid","high"),include.lowest=T)
+ yc<-cut(y,breaks=seq(min(y),max(y),length.out=4),
+ labels=c("low","mid","high"),include.lowest=T)
+ res<-c(
+ cor.test(x,y)$p.value,
+ cor.test(x,y,method="spearman")$p.value,
+ cor.test(x,y,method="kendall")$p.value,
+ summary(aov(y~xc))[[1]][1,5],
+ leveneTest(x~yc)[[3]][1],
+ chisq.test(table(xc,yc),
+            simulate.p.value =T)$p.value
+ )
+ names(res)<-c("Pearson","rho","tau",
+               "ANOVA","Leven","chi2")
+ res
+ }
```

Параметрами цієї функції є аналізовані змінні x і y.

Далі ми будемо застосовувати цю функцію у всіх прикладах до модельованих даних. При цьому стандартним рівнем значущості будемо вважати $\alpha = 0.05$. Спочатку розглянемо випадок малої кількості спостережень.

Приклад 9.7.3. У цьому прикладі згенеруємо лише 6 спостережень, причому x матиме рівномірний розподіл на $[0, 1]$, а y генерується як сума x і маленької гауссової похибки:

```
> set.seed(3)
> n<-6
> x1<-runif(n)
> y1<-x1+0.25*rnorm(n)
```

```
> plot(x1,y1)
> grid(3,3,col="red",lty=1,lwd=1)
```

— діаграму розсіювання для цих даних див. на рис. 9.10. Червоні горизонтальні і вертикальні прямі на цьому рисунку показують, як групуються дані за змінними `x` і `y`. Ці прямі відображені функцією `grid()`.

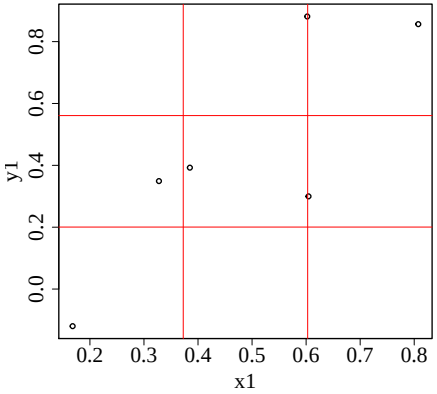

Рис. 9.10. Дані з приблизно лінійною залежністю

На око можна помітити, що із зростанням `x` в середньому зростає і `y`, хоча це правило не є абсолютним — залежність помітна, але не переконлива.

Застосуємо наш набір тестів до цих даних:

```
> print(CompareTests(x1,y1),digits=4)

Pearson     rho     tau   ANOVA   Leven    chi2
0.03685 0.29722 0.27222 0.31032 0.67158 0.56922
```

Як бачимо, лише тест на основі кореляції Пірсона помітив залежність: $p = 0.03685 < 0.05$. за всіма іншими тестами значущої залежності не виявлено.

Це типовий випадок, коли тест Пірсона має абсолютну перевагу над іншими: дуже мала кількість спостережень, залежність близька до лінійної. На таких даних лише на цей тест і можна покладати надії, якщо треба виявити залежність. Але якщо у цьому прикладі збільшити кількість спостережень `n` хоча б до 10, залежність між `x` і `y` стане помітною для всіх тестів, крім тесту Левена (перевірте!). ◀

Приклад 9.7.4. Розглянемо тепер випадок, коли між змінними є сильно виражена нелінійна, але монотонна залежність

$$y = x^6$$

— див. рис. 9.11.

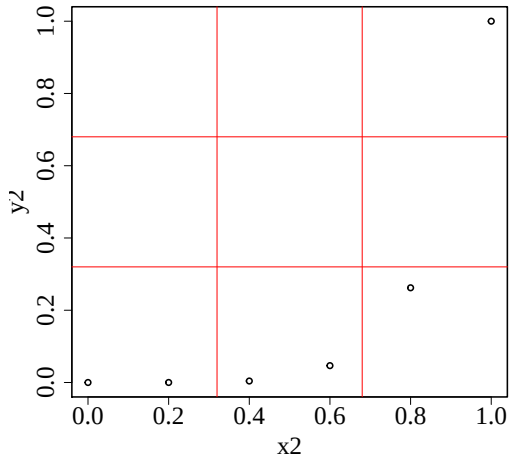

Рис. 9.11. Дані з нелінійною залежністю

Змінна `x` формується як арифметична прогресія від 0 до 1,

всього n= 6 спостережень:

```
> x2<-seq(0,1,length.out = 6)
> y2<-x2^6
> plot(x2,y2)
> grid(3,3,col="red",lty=1,lwd=1)
```

На діаграмі розсіювання рис. 9.11 монотонна залежність цілком помітна, але даних так мало, що статистична значущість цієї залежності сумнівна. Застосуємо нашу батарею тестів:

```
> print(CompareTests(x2,y2),digits=4)

 Pearson       rho       tau     ANOVA     Leven      chi2
0.063214 0.002778 0.002778 0.205774 0.260575       NaN
```

Лише рангові тести (Спірмена і Кендалла з $p = 0.002778$) змогли побачити цю залежність. Близьким до успіху був тест на основі кореляції Пірсона, у нього $p = 0.063214$ — поруч з 0.05 але все ж більше. Тест χ^2 видав невизначений результат: оскільки у вибірці не знайшлось жодного значення y в інтервалі проміжних значень, статистику тесту обчислити неможливо. Тести дисперсійного аналізу не виявили залежності.

Якщо збільшити кількість спостережень до 10, залежність стане помітною тесту з кореляцією Пірсона. При n= 20 всі наші тести її помітять (перевірте).

Отже рангові тести можуть підтвердити значущість монотонної або дуже близької до монотонної залежності по малих вибірках, там, де іншим тестам для цього не вистачає спостережень. ◄

Таким чином, при дуже малих обсягах спостережень не доцільно використовувати які-небудь тести незалежності, крім тестів на основі кореляцій.

Перейдемо тепер до випадків, коли деякі тести не помічають залежності по вибірках як завгодно великого обсягу. У наступному

прикладі розглядається випадок, подібний до обговореного нами у прикладі 5.2.3.

Приклад 9.7.5. Нехай x має рівномірний розподіл на $[-1,1]$, а $y = x^2$. Це цілком очевидна і чудово помітна на рисунку (рис. 9.12) залежність, яку не вміють помічати коефіцієнти кореляції.

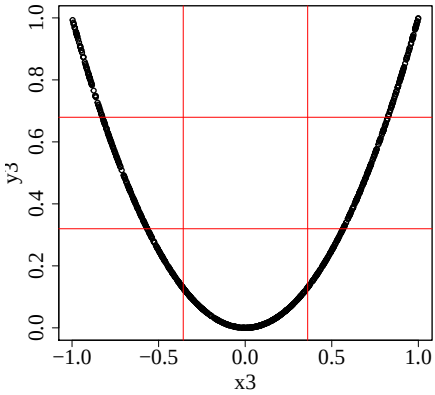

Рис. 9.12. Дані з симетричною залежністю

Подивимось, як працюватимуть відповідні тести на вибірці обсягу $n = 1000$ спостережень:

```
> set.seed(3)
> x3<-runif(1000,-1,1)
> y3<-x3^2
> plot(x3,y3)
> grid(3,3,col="red",lty=1,lwd=1)
> print(CompareTests(x3,y3),digits=4)

   Pearson        rho        tau      ANOVA      Leven
 1.891e-01  2.677e-01  2.177e-01 2.101e-141  1.447e-73
```

```
      chi2
4.998e-04
```

△

Як бачимо, незважаючи на великий обсяг даних, жодний тест кореляцій не зміг побачити залежність. Всі тести, що використовують групування, успішно виявили залежність.

Цей ефект пов'язаний саме з симетрією нашої картинки, де монотонно спадна залежність при $x < 0$ збалансована монотонним зростанням при $x > 0$. Якщо взяти `x` рівномірно розподіленим на $[-1, 0.9]$, то всі тести кореляцій будуть помічати залежність. (Причому тест на основі коефіцієнта Пірсона навіть краще, ніж рангові тести, тобто при менших обсягах вибірки).

Таким чином, тести на основі кореляцій не здатні бачити залежності у випадку, коли позитивна корельованість однієї частини даних компенсується негативною для іншої їхньої частини. Тести з групуванням такі залежності можуть помітити. ◀

Приклад 9.7.6. Розглянемо тепер картинку з ідеальною симетрією, в якій дані рівномірно розподілені на колі (див. рис. 9.13).

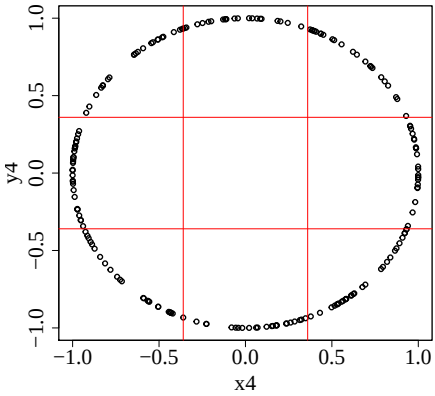

Рис. 9.13. Дані рівномірно розподілені на колі

Нехай не спостережувана змінна t має рівномірний розподіл на $[-\pi, \pi]$, а спостережувані x і y визначаються як

$$x = \sin(t),\ y = \cos(t).$$

Застосуємо до цих даних наші тести:

```
> set.seed(4)
> t<-runif(200,-pi,pi)
> x4<-sin(t)
> y4<-cos(t)
> plot(x4,y4)
> grid(3,3,col="red",lty=1,lwd=1)
> print(CompareTests(x4,y4),digits=4)

  Pearson       rho       tau     ANOVA     Leven
0.5400021 0.6205366 0.9932548 0.8355214 0.0011815
     chi2
0.0009995
```

Як бачимо, тепер і F-тест дисперсійного аналізу не помічає залежності. Це й не дивно. F-тест призначений для виявлення відмінностей середніх у різних групах. А середнє y дорівнює нулю в усіх трьох групах, що відповідають "малим", "проміжним" і "великим" значенням x.

Тести Левена та χ^2 помічають залежність, але за різними ознаками. Тест Левена налаштований бачити відмінності дисперсій у різних групах. І дійсно, у групі, що відповідає проміжним значенням x, розкид y більший, ніж у крайніх групах. Це і дозволяє тесту Левена виявити залежність.

Тест χ^2 налаштований бачити відмінності розподілів у різних рядках (або у різних стовпчиках) таблиці спряженості. Ми групуємо дані по вертикалі та по горизонталі у три групи, отже таблиця спряженості буде розміру 3×3. Виведемо її для групованих даних нашого прикладу:

```
> # Групуємо змінну x:
> xc<-cut(x4,breaks=seq(min(x4),max(x4),length.out=4),
+ labels=c("low","mid","high"),include.lowest=T)
> # Групуємо змінну y:
> yc<-cut(y4,breaks=seq(min(y4),max(y4),length.out=4),
+ labels=c("low","mid","high"),include.lowest=T)
> # Виводимо таблицю спряженості:
> table(xc,yc)

      yc
xc     low mid high
  low   27  29   26
  mid   22   0   19
  high  32  21   24
```

Бачимо, що у першому рядку таблиці в кожну комірку потрапила приблизно однакова кількість спостережень. Аналогічно — у третьому рядку. А от у другому рядку посередині — дірка[53]. Оцю відмінність розподілів у різних рядках і помічає χ^2-тест.

Підсумовуючи, можна сказати, що χ^2-тест здатен помітити залежність будь-якої природи[54], тоді як інші тести націлені на ті чи інші специфічні залежності. Але за універсальність χ^2-тесту треба платити помітно більшим обсягом вибірок, для якого він починає виявляти залежності. ◀

І все ж у цьому прикладі тест Левена помітив залежність. А чи можна навести приклад даних, в якому з усіх наших тестів залежність виявить тільки χ^2? Можна.

Приклад 9.7.7. Згенеруємо `x` як рівномірні на $[0, 1]$ випадкові величини. А розподіл у будемо для кожного спостереження вибирати залежно від `x`. Точніше, цей розподіл буде сумішшю двох

[53] Центр кола, само собою.

[54] Звичайно, для цього може бути не досить групування лише по трьох градаціях по кожній змінній. Щоб побачити справді будь-яку залежність, може знадобитись групування на дуже багато підгруп.

компонент з ймовірністю змішування, що дорівнює x. Розподіли компонент оберемо так, щоб їхні математичні сподівання дорівнювали 0, а дисперсії були однаковими. А саме, перша компонента буде рівномірно розподілена на $[-1, 1]$, а друга — гауссова з нульовим математичним сподіванням і дисперсією $s^2 = 1/3$.

В результаті умовні розподіли y при фіксованому x будуть різними, але їхні математичні сподівання і дисперсії не будуть залежати від x. Діаграма розсіювання таких даних відображена на рис. 9.14. Треба досить добре придивитись до цього рисунка, щоб побачити зміни розподілу y при зміні x. Першим впадає в око зростання кількості точок, що вилітають з сірого прямокутника посередині при зростанні x. Далі можна помітити, що і сам прямокутник потроху розмивається у цьому ж напрямку. Але не очевидно, що це не випадковий ефект.

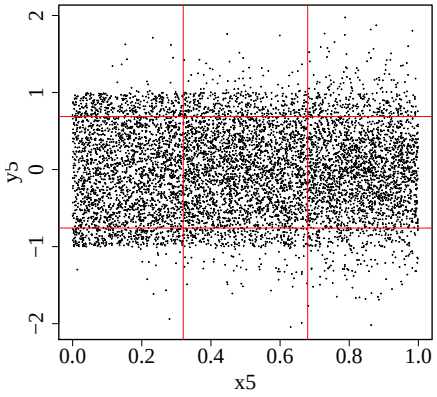

Рис. 9.14. Суміш двох розподілів

Згенеруємо дані з обсягом вибірки $n = 8000$ і проведемо тестування:

```
> set.seed(4)
```

```
> s<-sqrt(1/3)
> n<-8000
> x5<-runif(n,0,1)
> i<-rbinom(n,1,x5)
> y5<-sapply(i,function(i)
+   ifelse(i==0,runif(1,-1,1),rnorm(1,0,s)))
> plot(x5,y5,cex=0.1)
> grid(3,3,col="red",lty=1,lwd=1)
> print(CompareTests(x5,y5),digits=4)

  Pearson       rho       tau     ANOVA     Leven
0.7598570 0.6636689 0.6795313 0.7174002 0.5136237
     chi2
0.0004998
```

△

Як бачимо, тільки тест χ^2 зумів помітити залежність ($p = 0.0004998$). Всі інші вважають, що такі спостереження свідчать про незалежність x і y.

Але для того, щоб тест χ^2 спрацював, нам знадобилось 8000 спостережень. При $n = 7000$ тест залежності не помічає. ◄

Розділ 10

Регресія

Досі ми займались переважно одновимірними статистичними даними, де для кожного досліджуваного об'єкта спостерігалась одна числова характеристика (змінна). Для випадку двох змінних розглядалось лише питання про відсутність або наявність залежності. Тепер ми переходимо до вивчення даних, у яких з кожним об'єктом пов'язано декілька спостережуваних характеристик (змінних). Найбільш поширеною задачею статистики таких даних є дослідження зв'язків між різними змінними, що описують один об'єкт. У математичній статистиці такі дані можна трактувати як випадкові вектори і задавати їхні теоретичні моделі у термінах відповідних функцій розподілу, щільностей тощо. Для побудови оцінок можна використовувати розглянуті у розд. 8 метод моментів, метод найбільшої вірогідності та інші аналогічні методи.

Але такий підхід не завжди є оптимальним. Отримані моделі часом виявляються занадто складнмим і важкими для змістовної інтерпретації. Тому у багатьох статистичних задачах доцільно трактувати зв'язок між досліджуваними змінними як розмитий аналог деякої строгої функціональної залежності. Наприклад, нехай дослідник спостерігає n об'єктів, перенумерованих індексом $j = 1, \ldots, n$, і кожен описується змінними Y_j, X_j^1,..., X_j^m, де m — фіксоване число. Дослідника цікавить, як прогнозувати значен-

ня змінної Y для нових об'єктів (у яких Y не спостерігалось) за значеннями $X^1,\ldots, X^m$. Для цього вводять модель вигляду

$$Y_j = g(X_j^1, \ldots X_j^m) + \varepsilon_j, \tag{10.1}$$

де $g : \mathbb{R}^m \to \mathbb{R}$ — невідома функція, яку називають **функцією регресії**, а ε_j — випадкова **похибка регресії**. Змінну Y, яка прогнозується за іншими, називають **відгуком**, а змінні X^i, $i = 1, \ldots, m$, які використовуються для прогнозування — **регресорами**. Модель (10.1) називають **класичною моделлю регресії**.

Таким чином, ми описуємо зв'язок між відгуком та регресорами як наближену функціональну залежність

$$Y \approx g(X^1, \ldots, X^m)$$

і намагаємось оцінити функцію регресії g, а спостережувані відхилення від цієї залежності трактуємо як не цікаві для нас випадкові похибки[1]. При такому підході дослідник зосереджується на оцінці g, а властивостями похибок цікавиться лише постільки, оскільки вони можуть бути корисними для оцінювання функції регресії, перевірки гіпотез про неї або для прогнозування. Це і називають регресійним підходом до статистичного аналізу багатовимірних даних.

Розрізняють параметричні і непараметричні регресійні моделі. У параметричних моделях вважається, що функція регресії відома з точністю до деякого набору невідомих параметрів $\mathbf{b} = (b_1, \ldots, b_d)$ — **коефіцієнтів регресії**:

$$g(x^1, \ldots, x^m) = g(x^1, \ldots, x^m; \mathbf{b}).$$

У цьому випадку задача оцінювання функції регресії зводиться до оцінки коефіцієнтів регресії.

[1]Можна, наприклад, вважати, що похибка ε_j виникає внаслідок дії неконтрольованих нами причин, які ми не можемо спостерігати, на відміну від регресорів, які ми використовуємо для прогнозу.

У непараметричних моделях вважається, що функція регресії може бути довільною функцією з досить широкого класу. Наприклад, можна розглядати всі двічі неперервно диференційовні функції на заданому інтервалі.

Зрозуміло, що задавши спільний розподіл вектора регресорів і похибок $(X_j^1, \ldots, X_j^m, \varepsilon_j)$ та функцію g у (10.1), можна описати спільний розподіл вектора спостережень

$$\zeta_j = (Y_j, X_j^1, \ldots, X_j^m).$$

Якщо вважати ζ_j незалежними, однаково розподіленими спостереженнями, отримуємо ймовірнісну модель всієї вибірки $\zeta_1, \ldots, \zeta_n$. До такої моделі вже можна буде застосовувати всю ту техніку оцінювання та перевірки гіпотез, яку ми розглядали у попередніх розділах. Моделі такого вигляду називають структурними регресійними моделями.

Інколи буває зручніше трактувати регресори X_j^i як невипадкові фіксовані числа. Наприклад, це можуть бути значення характеристик експерименту, що їх дослідник задає сам. Відповідні регресійні моделі називають функціональними. У них відгук є випадковим, але його випадковість створюється лише похибкою. До таких моделей також можна застосовувати, наприклад, метод найбільшої вірогідності. Але у них відгуки не є однаково розподіленими випадковими величинами. Тому асимптотичні твердження, на яких ґрунтувались алгоритми розд. 8 та 9, для них треба модифікувати. Це приводить і до певної модифікації самих алгоритмів.

Далі у цьому розділі ми розглянемо різні підходи до дослідження статистичних даних на основі регресійних моделей. Спочатку у підрозд. 10.1 можливості цих підходів демонструються на прикладі простої лінійної регресії. Потім застосування різних технік до загальних задач описується більш детально.

10.1 Проста лінійна регресія

Нехай для кожного досліджуваного об'єкта спостерігаються дві змінні — X і Y. Всього спостерігають n об'єктів, значення змінних для j-го об'єкта — X_j і Y_j. Дослідник намагається описати зв'язок між цими змінними, використовуючи наближено лінійну залежність

$$Y_j \approx b_0 + b_1 X_j, \tag{10.2}$$

де b_0 і b_1 — невідомі коефіцієнти, які треба оцінити за даними X_j, Y_j, $j = 1, \ldots, n$. Така модель називається моделлю **простої лінійної регресії**. Її часто використовують для опису даних, прогнозування, перевірки гіпотез.

Класична регресія. Метод найменших квадратів. З точки зору класичного підходу (10.1) наближену формулу (10.2) можна трактувати як модель для прогнозування змінної Y за змінною X на основі лінійної функції регресії, а відхилення від строгої лінійної залежності інтерпретувати як похибки такого прогнозу:

$$Y_j = b_0 + b_1 X_j + \varepsilon_j. \tag{10.3}$$

Тут $g(x, \mathbf{b}) = b_0 + b_1 x$ — лінійна функція регресії, $\mathbf{b} = (b_0, b_1)^T$ — вектор невідомих коефіцієнтів регресії, X_j — регресор, Y_j — відгук, ε_j — похибка регресії.

Щоб оцінити коефіцієнти можна скористатись методом найменших квадратів (МНК)[2]. Для цього складають **функціонал найменших квадратів**, тобто суму квадратів різниць між справжніми значеннями відгуку та значеннями прогнозу, який вийде при використанні коефіцієнтів $\mathbf{b}$:

$$J(\mathbf{b}) = \sum_{j=1}^{n} (Y_j - b_0 - b_1 X_j)^2.$$

[2] Англ. *least squares* (LS).

Оцінкою $\hat{\mathbf{b}} = (\hat{b}_0, \hat{b}_1)^T$ для справжніх коефіцієнтів регресії називають те значення $\mathbf{b}$, на якому досягається мінімум $J(\mathbf{b})$ по всіх можливих $\mathbf{b} \in \mathbb{R}^2$.

Досить просто перевірити, що цей мінімум досягається при

$$\hat{b}_1 = \frac{\widehat{\mathrm{cov}}(X, Y)}{S^2(X)},$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}.$$

Це й є оцінки методу найменших квадратів для коефіцієнтів простої лінійної регресії.

Тут, як і раніше

$\bar{X}$, $\bar{Y}$ — вибіркові середні значень регресора X_j та відгуку Y_j;

$S^2(X)$ — вибіркова дисперсія регресора X;

$\widehat{\mathrm{cov}}(X, Y) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})(Y_j - \bar{Y})$ — вибіркова коваріація відгука з регресором.

Приклад 10.1.1. Людям, що стежать за своєю вагою, корисно мати певний еталон правильної ваги, на який вони могли б орієнтуватись. Часто для цього використовують зріст людини, тобто вважають, що кожному значенню зросту відповідає певна оптимальна вага, яку матиме "людина гарної статури" з таким зростом. Як отримати формулу такої залежності оптимальної ваги від зросту? Можна взяти певний набір людей, про яких відомо, що вони мають гарну статуру, виміряти їх вагу та зріст і далі застосувати метод найменших квадратів для оцінки коефіцієнтів регресії у моделі, де відгуком є вага, а регресором — зріст.

Для цього прикладу я обрав дані про вагу і зріст зірок шоубізнесу (акторів, популярних співаків, фотомоделей), тобто людей, які за професією повинні відповідати критерію "гарної статури". Дані взяті з інтернет-сторінки за адресою

`http://www.celeb-height-weight.psyphil.com/celebrities-height-and-weight-chart-%E2%80%93-celebrity-stats/`.

У таблиці на цій сторінці розміщені дані про зірок (celebrities). Я взяв з цієї таблиці всіх зірок-чоловіків (їх там мало) і частину

зірок-жінок (їх значно більше). Відбір жінок був більш-менш довільним, метою було отримати зручну для пояснень картинку, тому особливо довіряти результатам аналізу не варто, тим більше, що і принципи відбору авторів таблиці невідомі. Вага вказана у фунтах (1 фунт= 0.453592 кг), зріст — у сантиметрах. Надалі я умовно називатиму цих людей "акторами".

Дані вміщені у файлі `actors.csv`, де кожен рядок відповідає одній людині, а стовпчики — змінним, що її характеризують: ім'я та прізвище (`name`), зріст у сантиметрах (`height`), вага у фунтах (`weight`), стать (`gender`: `f` — жіноча, `m` — чоловіча). Відобразимо ці дані на рисунку, де по горизонталі відкладено зріст, по вертикалі — вагу, а у точці з координатами, що відповідають конкретній людині, стоїть літера, яка вказує її стать. (У статистиці такі та аналогічні рисунки називають **діаграмами розсіювання**, ми вже познайомились з ними у підрозд. 5.1) — див. рис. 10.1

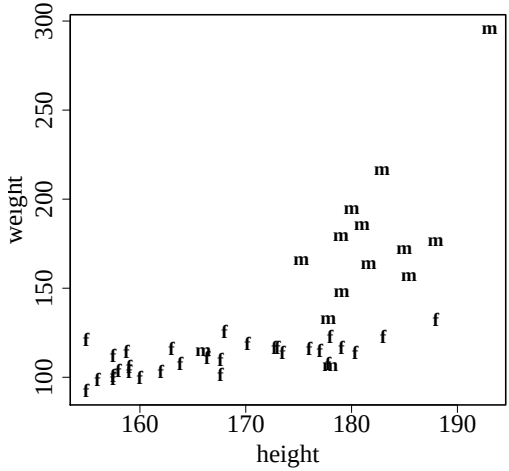

Рис. 10.1. Діаграма розсіювання ваги та зросту акторів (повні дані)

На рис. 10.1 одразу можна побачити, що розподіл даних для

чоловіків та жінок є різним, причому, внаслідок більшого розкиду ваги чоловіків, точки, що відповідають жінкам, збились у нижній частині діаграми. Тому помітити залежність між зростом і вагою жінок на цій діаграмі майже неможливо. Нарисуємо аналогічну діаграму окремо тільки для жінок (рис. 10.2).

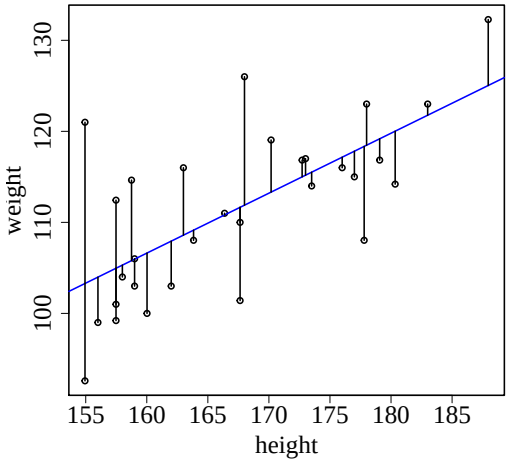

Рис. 10.2. Вага та зросіст акторів. Підгонка за МНК для жінок

На цій діаграмі розсіювання ми вивели також "оцінену лінію регресії", тобто пряму, що має рівняння $y = \hat{b}_0 + \hat{b}_1 x$, де $\hat{b}_0$ і $\hat{b}_1$ — оцінки за методом найменших квадратів для підгонки регресії у моделі

$$\texttt{weight}_j = b_0 + b_1 \times \texttt{height}_j + \varepsilon_j,$$

підігнаній за даними лише для жінок-актрис. Оцінки коефіцієнтів такі: $\hat{b}_0 = 1.222, \hat{b}_1 = 0.6587$. Їх можна спробувати інтерпретувати. Наприклад, $\hat{b}_1$ показує, наскільки зростає ідеальна вага у фунтах при збільшенні зросту на 1 см. Коефіцієнт $\hat{b}_0$ мав би відповідати

ідеальній вазі людини, що має зріст 0 см. Очевидно, таких людей не існує. Якби ми вирішили використовувати одне і те ж лінійне наближення для залежності між зростом і вагою для всіх можливих значень зросту, то природно було б вважати, що нульовому зросту відповідає нульова вага — $b_0 = 0$. Але для всіх наших даних зріст розташований у інтервалі від 155 до 185 см, тому робити за ними статистично обґрунтовані висновки про досліджувану залежність на інших інтервалах значень зросту неможливо.

Формула

$$\widehat{\texttt{weight}}(\texttt{height}) = \hat{b}_0 + \hat{b}_1 \times \texttt{height} \tag{10.4}$$

дає прогноз за методом найменших квадратів для ваги, що відповідає заданому зросту. У цій задачі такий прогноз можна трактувати як ідеальну вагу для актриси, що хоче бути зіркою. Якщо підставити сюди зріст j-ї актриси, отримуємо прогнозну ідеальну вагу для неї $\widehat{\texttt{weight}}_j = \widehat{\texttt{weight}}(\texttt{height}_j)$.

Точки з координатами $(\texttt{height}_j, \widehat{\texttt{weight}}_j)$ лежать на оціненій лінії регресії. На рис. 10.1 праворуч ці точки з'єднані з відповідними точками даних вертикальними відрізками. Довжини цих відрізків характеризують відхилення прогнозів від справжніх спостережуваних значень відгуків. За методом найменших квадратів лінія регресії підбирається так, щоб мінімізувати суму квадратів цих відхилень.

Можна сказати, що МНК мінімізує відхилення по вертикалі точок спостережень від лінії регресії.

Рисунки отримані таким скриптом:

```
> actors<-read.csv2("c:/rem/term/actors.csv",
> header=T)
> # рис. 10.1:
> # рисуємо рамку:
> plot(actors$height,actors$weight,type="n",
+    xlab="height",ylab="weight",sub=" ")
```

```
> # виводимо літери для статі:
> text(actors$height,actors$weight,
+    labels=actors$gender)
> # рис. 10.2:
> actrf<-actors[actors$gender=="f",]
> # рисуємо діаграму розсіювання:
> plot(actrf$height,actrf$weight,sub=" ")
> bls1<-lm(weight~height,data=actrf)$coefficients
>  bls1

(Intercept)      height
  1.2224016   0.6587242

>  abline(bls1,col="blue")
>  segments(actrf$height,
+        actrf$weight,actrf$height,
+        bls1[1]+bls1[2]*actrf$height)
```

△

Тут функція `lm()` реалізує обчислення оцінок коефіцієнтів лінійної регресії за МНК. Вона більш детально описана у підрозд. 10.2. Зараз скажемо лише, що `weight~height` — це формула, яка вказує, що відгуком у нашій моделі є `weight`, а регресором — `height`; опція `data` вказує фрейм, з якого беруться дані для підгонки. Результатом виконання функції `lm()` є об'єкт складної структури, МНК-оцінки коефіцієнтів регресії містяться у ньому в атрибуті `$coefficients`.

Хоча прогнозування ваги за зростом здається найбільш природним способом використання залежності між цими змінними, але не виключено, що у когось виникне обернена задача. Скажімо, Шерлок Холмс може захотіти визначити невідомий йому зріст злочинця за його відомою вагою. Здавалося б, маючи формулу прогнозу (10.4), для отримання оберненої прогнозної формули досить розв'язати відповідне рівняння відносно `height`:

$$\widehat{\texttt{height}}(\texttt{weight}) = -\frac{\hat{b}_0}{\hat{b}_1} + \frac{1}{\hat{b}_1} \times \texttt{weight}. \qquad (10.5)$$

Однак за логікою МНК, такий прогноз буде неправильним. Насправді треба скласти регресійне рівняння

$$\texttt{height}_j = a_0 + a_1 \times \texttt{weight}_j + \delta_j,$$

та оцінювати його коефіцієнти за даними, використовуючи МНК. Отримані оцінки $\hat{a}_0$, $\hat{a}_1$ будуть, у загальному випадку, відрізнятись від коефіцієнтів у формулі (10.5). Але саме їх слід використовувати для прогнозування.

Чому це так? Тому, що при прогнозуванні `height` природно мінімізувати відхилення саме прогнозів `height` від їхніх справжніх значень, а це не еквівалентно мінімізації відхилень прогнозів `weight`, яке ми робили у попередньому випадку. Якщо на діаграмі розсіювання відкладати значення `height` по горизонталі, а `weight` — по вертикалі, то тепер нас будуть цікавити відстані від точок до лінії регресії не по вертикалі, як раніше, а по горизонталі. Результат такої підгонки зображено на рис. 10.3.

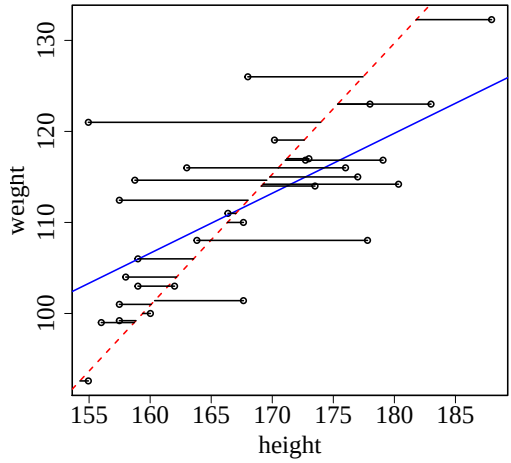

Рис. 10.3. “Обернена” регресія ваги і зросту. Лінія регресії: суцільна — звичайний МНК, штрихова — “обернена”

Тут штрихова червона лінія відповідає регресії з відгуком `height` по регресору `weight` (тобто відгук відкладений по горизонталі, таку регресію можна назвати "оберненою"). Суцільна синя лінія — та сама, що і на попередньому рисунку, тобто звичайна "пряма" МНК регресія. Як бачимо, пряма і обернена регресії помітно відрізняються.

Ортогональна регресія. Інколи буває необхідно встановити за даними лінію регресії так, щоб вона не залежала від вибору системи координат. Наприклад, нехай ви йдете по прямому шосе і час від часу визначаєте своє положення за допомогою GPS, який вказує довготу (X) і широту (Y) місця вашого положення. В результаті ваших спостережень отримується набір точок на топографічній карті (X_j, Y_j), $j = 1, \ldots, n$. Ці точки не будуть лежати всі на одній прямій, тому що GPS вимірює координати з деякими похибками. Вам треба вибрати пряму, яка найкращим чином буде відповідати виміряним даним. Звичайна МНК регресія тут недоречна, тому що, вибираючи різні системи координат на площині карти, ми будемо отримувати різні прямі регресії. А це суперечить логіці задачі.

У таких задачах доцільно застосовувати техніку ортогональної регресії, в якій оцінка отримується мінімізацією суми квадратів відстаней від точок спостережень до відповідних найближчих точок на прямій:

$$J^{TLS}(\mathbf{b}) = \sum_{j=1}^{n} \left[(X_j - \tilde{X}_j(\mathbf{b}))^2 + (Y_j - \tilde{Y}_j(\mathbf{b}))^2 \right],$$

де $(\tilde{X}_j(\mathbf{b}), \tilde{Y}_j(\mathbf{b}))$ — координати точки, що є ортогональною проекцією точки (X_j, Y_j) на пряму, задану рівнянням[3] $y = b_0 + b_1 x$:

$$\tilde{X}_j(\mathbf{b}) = \frac{(X_j + b_1 Y_j - b_0 b_1)}{(1 + b_1^2)}, \; \tilde{Y}_j(\mathbf{b}) = b_0 + b_1 \tilde{X}_j(\mathbf{b}).$$

[3] Строго кажучи, не кожну пряму можна задати таким рівнянням (вертикальну не можна). Але ми зараз не зупиняємось на цьому ускладненні.

Цей функціонал називають повною сумою квадратів. Точка мінімуму цього функціоналу

$$\hat{\mathbf{b}}^{TLS} = \underset{\mathbf{b}\in\mathbb{R}^2}{\operatorname{argmin}} J^{TLS}(\mathbf{b})$$

називається оцінкою мінімуму повних квадратів або ортогональною регрессією[4]

Як приклад ортогональна регресія, застосована до даних про вагу та зріст акторок, зображена на рис. 10.4.

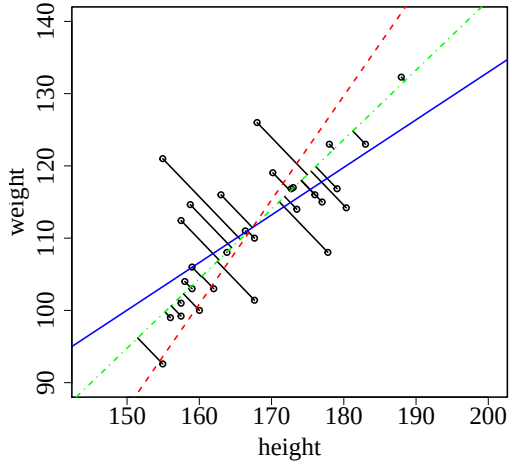

Рис. 10.4. Ортогональна регресія ваги та зросту. Лінія регресії: суцільна — звичайний МНК, штрихова — "обернена", штрихпунктирна — ортогональна.

Лінія, що відповідає ортогональній регресії, зображена зеленим кольором (штрихпунктирною лінією). Вона розташовується між лініями прямої (синя, суцільна лінія) та оберненої (червона

[4]Англ. total least squares (TLS), orthogonal regression, Deming regression.

штрихова лінія) МНК регресії. Всі три лінії перетинаються в одній точці з координатами $(\bar{X}, \bar{Y})$ — у барицентрі[5] набору даних.

Рисунки 10.3 і 10.4 виведено таким скриптом:

```
> #  ``горизонтальна'' регресія
> als2<-lm(height~weight,data=actrf)$coefficients
> # рисуємо діаграму розсіювання:
> plot(actrf$height,actrf$weight,
+ xlab="height",ylab="weight")
> abline(bls1,col="blue") # bls1 з попереднього скрипта
> abline(c(-als2[1]/als2[2],1/als2[2]),
+     col="red",lty="dashed")
> segments(actrf$height,actrf$weight,
+ als2[1]+als2[2]*actrf$weight,actrf$weight)
> #  ортогональна регресія
> library(MethComp)
> tls<-Deming(actrf$height,actrf$weight)
> x<-actrf$height
> y<-actrf$weight
> b0<-tls[1]
> b1<-tls[2]
> # розраховуємо координати проекцій даних
> # на пряму регресії:
> x0<-(x+b1*y-b0*b1)/(1+b1^2)
> y0<-b0+b1*x0
> plot(actrf$height,actrf$weight,asp=1,
+ xlab="height",ylab="weight",
+   xlim=c(155,190),ylim=c(90,140))
> segments(x,y,x0,y0)
> abline(tls[1:2],col="green",lty="dotdash")
> abline(bls1,col="blue")
```

[5]Барицентр — центр мас. Якби у кожній точці діаграми розсіювання даних була розташована маса 1 кг, то центр мас такої системи був би у точці $(\bar{X}, \bar{Y})$.

```
> abline(c(-als2[1]/als2[2],1/als2[2]),
+     col="red",lty="dashed")
```

△

Тут для обчислення коефіцієнтів ортогональної регресії використана функція `Deming()` з бібліотеки `MethComp`. Перший її параметр — це регресор (горизонтальна координата), а другий — відгук (вертикальна координата). При відображенні рисунка функцією `plot()` встановлена опція `asp=1`, яка задає однаковий масштаб як по горизонталі, так і по вертикалі. Інакше відрізки, що з'єднують точки даних з їхніми проекціями на рисунку не були б перпендикулярними до лінії регресії.

Ортогональну регресію можна розглядати як спосіб оцінки коефіцієнтів у регресійній моделі, відмінній від класичної. Дійсно, у прикладі з визначенням "рівняння" шосе, по якому спостерігач рухається, ми маємо справу зі спостереженнями X_j, Y_j справжніх координат точок на шосе (x_j, y_j), виміряними з похибкою, тобто

$$X_j = x_j + \varepsilon_j,\ Y_j = y_j + \delta_j, \tag{10.6}$$

де ε_j, δ_j — відповідні похибки вимірювання. При цьому справжні координати невідомі, але вони точно описуються рівнянням прямої:

$$y_j = b_0 + b_1 x_j. \tag{10.7}$$

Треба за спостережуваними значеннями (X_j, Y_j), $j = 1, \ldots, n$ оцінити коефіцієнти $\mathbf{b} = (b_0, b_1)^T$.

Якщо вважати, що похибки ε_j, δ_j є незалежними між собою і мають однаковий розкид, то природно використати ортогональну регресію для оцінювання.

Моделі, подібні до (10.6) — (10.7), називають **моделями з похибками у змінних**[6]. Їх доцільно використовувати там, де в основі зв'язку між досліджуваними змінними лежить точний фізичний

[6]Errors-in-variables models. Йдеться про те, що похибки наявні як у відгуку, так і у регресорах.

(хімічний, економічний. . .) закон, спотворений похибками вимірювання або подібними ефектами (див. [13, 22]).

Для аналізу даних на зразок ваги-зросту людей такі моделі не є природними. Звичайно, вага і зріст вимірюються неточно, у їх спостережуваних значеннях присутні похибки вимірювання. Але випадкові коливання навколо лінії регресії, які ми спостерігаємо на діаграмах розсіювання рис. 10.1, 10.3, викликані не цими похибками, а зовсім іншими причинами (генетичними, фізіологічними соціальними, психологічними), які приводять до того, що люди з однаковим зростом мають різну вагу[7]. Тому для даних такого роду природніше застосовувати класичну модель регресії, яка дозволяє будувати найбільш точні прогнози для відгуку за спостережуваними значеннями регресорів. ◄

Квантильна регресія. Продовжимо дослідження зв'язку ваги та зросту. З діаграми розсіювання на рис. 10.1 праворуч можна зробити висновок, що для актрис, які потрапили до нашого поля зору, відхилення ваги у 2-3 фунти від передбачуваної за МНК не є великою проблемою, а от відхилення більше 15-ти фунтів майже не зустрічаються. Для людини, що стежить за своєю вагою, встановлення для неї обмежень зверху і знизу може бути важливішим, ніж визначення якогось ідеального фіксованого значення ваги.

Відповідну статистичну задачу можна поставити так: для кожного можливого значення зросту `height` треба встановити таке порогове значення $\widehat{W}_\tau(\texttt{height})$, щоб вага $\tau = 90\%$ актрис-зірок не перевищувала поріг $\widehat{W}_\tau(\texttt{height})$. Це було б розумне обмеження зверху допустимої ваги для людини, що хоче бути актрисою. Аналогічно можна встановити обмеження знизу, поклавши, наприклад, $\tau = 10\%$. Зрозуміло, ми не маємо на увазі τ відсотків актрис, що потрапили до нашої вибірки (серед них може взагалі не бути жодної зі зростом `height`, який нас цікавить). Йдеться про всіх можливих осіб жіночої статі, що мають будову тіла достатньо гарну, як

[7] До речі, для таких характеристик "вага людини" взагалі важко визначити, що саме слід вважати точним значенням: вага до обіду, це зовсім не те, що вага після!

для актриси. У термінах математичної статистики ми оцінюємо квантиль рівня τ теоретичного розподілу ваги особи, що вибрана навмання з популяції всіх (потенційних) актрис, зріст яких дорівнює `height`. Зрозуміло, що оцінку доведеться будувати за тою вибіркою, яка у нас є. Оцінка будується як функція від `height`, тобто на діаграмі розсіювання це буде деяка лінія. За наявною діаграмою навряд чи можна запропонувати якусь складну форму для цієї лінії, отже, вибираємо найпростішу — пряму.

Для оцінювання коефіцієнтів такої прямої застосуємо метод квантильної регресії Р. Кроенкера і Дж. Бассета [36]. Введемо функцію

$$\rho_\tau(y) = \begin{cases} \tau y & \text{якщо } y \geq 0, \\ (1-\tau)|y| & \text{якщо } y < 0. \end{cases} \tag{10.8}$$

Неважко переконатись, що для будь-якої випадкової величини η з неперервним розподілом F квантиль $Q^F(\tau)$ є точкою мінімуму за t функції $\mathsf{E}\,\rho_\tau(\eta - t)$. Якщо замінити математичне сподівання його оцінкою — вибірковим середнім — як точку мінімуму отримаємо вибіркову квантиль. Ідея квантильної регресії полягає в тому, щоб використати для оцінки функціонал, подібний МНК, але із заміною квадратичної функції на функцію ρ_τ. Отже, оцінкою коефіцієнтів $\mathbf{b}$ за методом квантильної регресії буде

$$\hat{\mathbf{b}}_\tau^{qant} = (\hat{b}_{0,\tau}^{quant}, \hat{b}_{1,\tau}^{quant})^T = \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^2} \sum_{j=1}^{n} \rho_\tau(Y_j - b_0 - b_1 X_j). \tag{10.9}$$

Для цих оцінок немає такої явної формули, як для оцінок МНК, але у бібліотеці `quantreg` є функція `rq()`, що підраховує їх чисельними методами. Застосуємо її до даних ваги і зросту актрис (рис. 10.5).

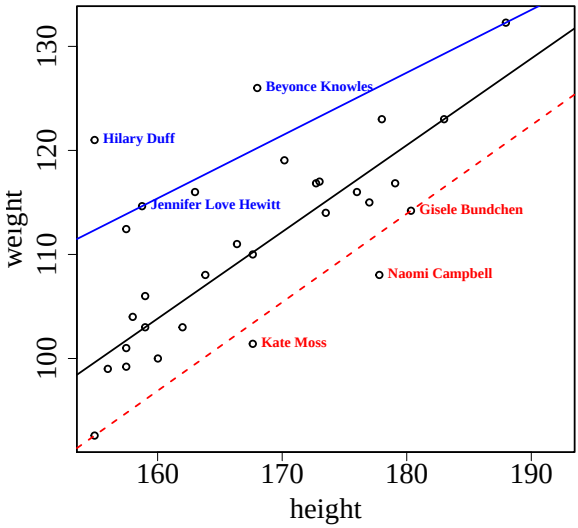

Рис. 10.5. Квантильна регресія

```
> # quantile regression
> library("quantreg")
> plot(actrf$height,actrf$weight,xlim=c(155,192),
+ xlab="height",ylab="weight")
> lmed<-rq(weight~height,tau=0.5,
+         data=actrf)$coefficients
> abline(lmed)
> lq10<-rq(weight~height,tau=0.1,
+         data=actrf)$coefficients
> abline(lq10,col="red",lty="dashed")
> actr10<-actrf[actrf$weight<lq10[1]+
+                 lq10[2]*actrf$height,]
> text(actr10$height,actr10$weight,
+ labels=actr10$name,pos=4,col="red")
```

```
> lq90<-rq(weight~height,tau=0.9,
+        data=actrf)$coefficients
> abline(lq90,col="blue")
> actr90<-actrf[actrf$weight>lq90[1]+
+                lq90[2]*actrf$height,]
> text(actr90$height,actr90$weight,
+ labels=actr90$name,pos=4,col="blue")
```

△

Функція `rq()` організована подібно до функції `lm()`, але має додатковий параметр `tau`, у якому треба задати рівень квантилі τ. На рисунку 10.5 суцільною (блакитною) лінією зображена квантильна регресія рівня $\tau = 0.9$, тобто для справжньої акторки ймовірність опинитись вище цієї лінії дорівнює 0.1. Штрихова (червона) лінія відповідає рівню 0.1, нижче неї опиняться 10% найбільш худорлявих акторок. Посередині чорна лінія з рівнем 0.5, вона відповідає медіані.

Для точок, що вийшли за суцільну та штрихову лінії на рисунку відображені імена та прізвища відповідних акторок. В інтернеті ви можете знайти їхні фотографії та фільми і перевірити, чи дійсно вони виділяються своєю статурою із загальної маси акторок[8].

По середині на рисунку проходить чорна лінія, що відповідає $\tau = 0.5$. Це ліня регресії з коефіцієнтами, що мінімізують функціонал

$$J^{Abs}(\mathbf{b}) = \sum_{j=1}^{n} |Y_j - b_0 - b_1 X_j|.$$

Таку регресію можна було б назвати медіанною, але вона має іншу назву — регресія за **методом найменших модулів, МНМ**[9]. Це одна з найбільш поширених альтернатив класичному методу найменших квадратів. Зокрема, її часто використовують у випадках, коли потрібні робастні оцінки коефіцієнтів регресії.

[8] Їхні фотографії могли б дуже прикрасити цю книгу, але, на жаль, у мене немає права розмістити їх тут.

[9] Least absolute deviations (LAD) regression.

Робастна регресія. Нагадаємо, що робастність — це стійкість оцінок щодо можливих забруднень вибірки спостереженнями, які не мають стосунку до досліджуваного явища. Особливо важливою є стійкість до забруднення викидами, тобто значеннями, які лежать далеко від основної маси спостережень (див. підрозд. 4.1). У випадку регресійних задач слід розрізняти "викиди по вертикалі" — дані про об'єкти з аномально великими відхиленнями значень відгуку і "викиди по горизонталі", коли аномальні значення мають регресори. Метод найменших модулів дає оцінки, стійкі щодо викидів по вертикалі, якщо у даних немає горизонтальних викидів. Тобто якщо забруднення можуть приводити лише до зміни відгуку, а не регресорів, використання МНМ є доцільним. Якщо викиди можуть бути як вертикальними, так і горизонтальними, оцінки МНМ стають не робастними.

В останньому випадку потрібні оцінки, що були б робастими при будь-яких викидах. Прикладом такої оцінки є **оцінка повторних медіан.** Вона будується таким чином.

Нехай спостерігаються регресор X_j та відгук Y_j для j від 1 до n. Розглянемо всі можливі пари точок (X_j, Y_j), (X_i, Y_i), для всіх $i \neq j$. Якщо всі точки різні, то через кожну таку пару проходить тільки одна пряма. Знайдемо кутові коефіцієнти всіх цих прямих:

$$b_1(i,j) = \frac{Y_i - Y_j}{X_i - X_j}.$$

Для кожного i розглянемо набір $b_1(i,j)$, $j = 1, \ldots, n$, $j \neq i$ і обчислимо його вибіркову медіану — $\mu(i)$. Оцінкою $\hat{b}_1^{\mu}$ методу повторних медіан для коефіцієнта регресії b_1 буде вибіркова медіана набору $\mu(i)$, $i = 1, \ldots, n$. Оцінка $\hat{b}_0^{\mu}$ для b_0 визначається як вибіркова медіана набору $Y_j - \hat{b}_1^{\mu} X_j$, $j = 1, \ldots, n$.

Пара $\hat{\mathbf{b}}^{\mu} = (\hat{b}_0^{\mu}, \hat{b}_1^{\mu})^T$ є оцінкою для $\mathbf{b}$, робастною відносно всіх викидів — як вертикальних, так і горизонтальних.

Розглянемо приклад, в якому на модельованих даних продемонстрована поведінка різних оцінок коефіцієнтів регресії за наявності горизонтальних і вертикальних викидів. Дані згенеровані і

оцінки обчислені таким скриптом:

```
>  RepMedian<-function(y,x) # повторна медіана
+ {
+ yy<-rep(y,length(y))-rep(y,each=length(y))
+ xx<-rep(x,length(x))-rep(x,each=length(x))
+ bb<-yy/xx
+ bb<-matrix(bb,nrow=length(x))
+ b1m<-median(apply(bb,1,median,na.rm=T))
+ b0m<-median(y-b1m*x)
+ c(b0m,b1m)
+  }
> # функція рисування:
> FigurePlot<-function(x,y,x1,y1)
+ {
+ plot(x,y)
+ lsc<-lm(y~x)$coefficients
+ lsc1<-lm(y1~x1)$coefficients
+ abline(lsc,col="green",lty="dotdash",lwd=4)
+ abline(lsc1,lty="dotdash",col="green")
+ lad<-rq(y~x,tau=0.5)$coefficients
+ abline(lad,col="blue",lty="solid",lwd=4)
+ lad1<-rq(y1~x1,tau=0.5)$coefficients
+ abline(lad1,col="blue",lty="solid")
+ rm<-RepMedian(y,x)
+ abline(rm,col="red",lty="dashed",lwd=4)
+ rm1<-RepMedian(y,x)
+ abline(rm1,col="red",lty="dashed")
+ }
> set.seed(5)
> # Рисунок вгорі
> x1<-seq(0,1,0.1)
> y1<-2*x1-1+0.5*rnorm(11)
> x<-c(x1,-0.1)
```

```
> y<-c(y1,6)
> FigurePlot(x,y,x1,y1)
> # Рисунок знизу
> x<-c(x1,-3.5)
> y<-c(y1,3.5)
> FigurePlot(x,y,x1,y1)
```

△

Результат — на рис. 10.6 і 10.7.

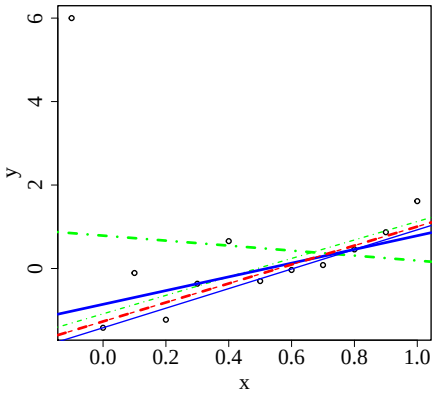

Рис. 10.6. Повторна медіана (червона штрихова) порівняно з МНМ (синя, суцільна) та МНК (зелена, штрихпунктирна). Викид тільки по вертикалі

Тут для двох наборів модельованих даних зображені результати підгонки простої лінійної регресії трьома методами: зелена штрихпунктирна лінія — МНК, синя суцільна — метод найменших модулів, червона штрихова — повторна медіана. Широкими лініями — результати підгонки за повними даними, тонкими — за даними без викиду. На рис. 10.6 — викид вертикальний. Видно, що він помітно вплинув на лінію МНК, дещо змінив МНМ і не вплинув на оцінки повторної медіани. На рис. 10.7 — викид, що

відрізняється від основної маси як по вертикалі, так і по горизонталі. Він радикально змінює оцінки МНК та МНМ, але знову не впливає на лінію, визначену повторною медіаною.

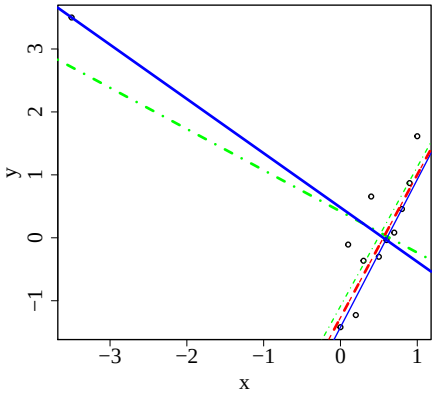

Рис. 10.7. Повторна медіана (червона штрихова) порівняно з МНМ (синя, суцільна) та МНК (зелена, штрихпунктирна). Викид і по вертикалі, і по горизонталі.

У читача може скластись враження, що повторна медіана дає найбільш стабільні оцінки у всіх випадках. Це враження помилкове, що і показує наступний приклад.

Приклад 10.1.2. Розглянемо знову дані `shortU.txt` про інтерес до шортів різних типів у США з прикладу 3.4.4. У прикладі 9.5.2 ми ввели для цих даних змінну `x`, що характеризує переважання інтересу до джинсових шортів порівняно з інтересом до карго-шортів у різних штатах США. У нас була гіпотеза[10] про те, що зміни `x` пов'язані з рівнем урбанізації відповідних штатів (змінна `urban`). Зараз ми розглянемо лише штати, що лежать у басейні Міссісіпі-Міссурі (`miss==1`). Відкладемо для цих штатів по горизонталі `urban`, а по вертикалі — `x` (див. ліву частину на рис. 10.8).

[10]Підтверджена порівнянням карт на рис. 3.20 і 3.22.

Бачимо, що у даних наявний викид — Південна Дакота, де `x`$= 1$. Як вплине цей викид на оцінки коефіцієнтів регресії з відгуком `x` та регресором `urban`?

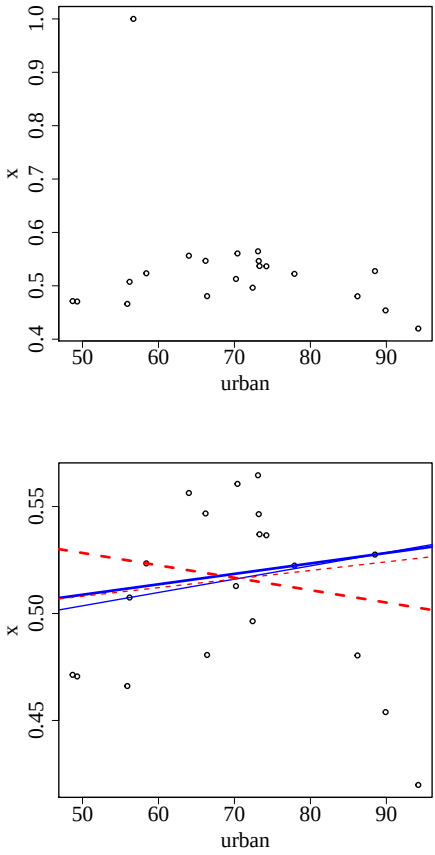

Рис. 10.8. Повторна медіана (червоним пунктиром) порівняно з МНМ (синім, суцільна) для даних про шорти

У верхній частині рис. 10.8 — діаграма розсіювання даних, на

якиій помітний викид. У нижній частині відображені результати підгонки простої лінійної регресії за МНМ (суцільна синя лінія) та за методом повторних медіан (штрихова червона). Як і раніше — тонкі лінії відповідають даним без викиду, широкі — з викидом. (Для зручності порівняння різних оцінок лінії регресії, масштаб обрано так, що сам викид опинився за межами цього рисунка). Ми бачимо, що пряма МНМ майже не змінилась під впливом викиду, а от оцінка повторних медіан змінилась дуже сильно, так, що залежність між `urban` та `x` зі зростаючої перетворилась на спадну.

Чому так сталось? Якщо придивитись до рис. 10.8, то можна побачити, що при малих значеннях `urban` `x` у середньому зростає із зростанням `urban`, а при великих — у середньому спадає. Наша модель намагалась відтворити обидва ці ефекти, однак пряма так поводитись не може. Повторна медіана вибрала один з варіантів за даними без викиду, але викид змусив її перейти до іншого. При використанні МНМ ефекти усереднились, викид великих змін не викликав.

Таким чином, нестабільність оцінок повторної медіани у цьому прикладі викликана тим, що вони застосовані до даних, які не відповідають обраній моделі залежності. Це не можна вважати принциповим недоліком оцінок, але можливість таких ефектів слід мати на увазі при обробці реальних даних.

Ми повернемось до аналізу цих даних у прикладі 10.2.3 ◄

10.2 Лінійний метод найменших квадратів

У цьому підрозділі розглядається стандартна техніка застосування методу найменших квадратів для підгонки лінійних регресійних моделей. Основна увага буде приділена використанню готових простих засобів `R`. Складніші питання відкладемо до наступних підрозділів.

Класична лінійна регресійна модель для опису залежності змін-

ної Y від змінних $X^1,\ldots,X^m$ має вигляд

$$Y_j = b_0 + b_1 X_j^1 + \cdots + b_m X_j^m + \varepsilon_j, \tag{10.10}$$

де індекс $j = 1,\ldots,n$ позначає номер спостереження у вибірці, Y_j — значення відгуку для $j-$го спостереження, $X_j^1,\ldots,X_j^m$ — значення регресорів для j-го спостереження, ε_j — випадкова похибка регресії (не спостерігається), b_i, $i = 0,\ldots,m$ — невідомі коефіцієнти регресії, які треба оцінювати за спостереженнями.

Функціоналом найменших квадратів називають

$$J(\mathbf{b}) = \sum_{j=1}^{n} \left(Y_j - b_0 - \sum_{i=1}^{m} b_i X_j^i \right)^2$$

(тут $\mathbf{b} = (b_0, b_1, \ldots, b_m)^T$ — вектор можливих значень коефіцієнтів регресії).

Оцінка методу найменших квадратів визначається як

$$\hat{\mathbf{b}} = \underset{\mathbf{b}\in\mathbb{R}^{m+1}}{\operatorname{argmin}} J(\mathbf{b}),$$

тобто це такий набір коефіцієнтів регресії, на якому функціонал МНК досягає найменшого значення.

Оцінки МНК існують завжди, але не завжди визначені однозначно[11].

У `R` підгонку лінійних регресійних моделей можна виконувати за допомогою функції `lm()`[12]. Ця функція не тільки обчислює оцінки МНК, а й перевіряє важливі гіпотези про них та оцінює якість прогнозу, який дає регресійна формула.

Найбільш популярна форма виклику функції `lm()` має вигляд:

```
lm(formula,data)
```

[11] Якщо регресія задається (10.10), то для однозначної визначеності оцінок МНК необхідно і достатньо, щоб матриця вибіркових коваріацій набору регресорів була невиродженою.

[12] Ця назва є скороченням для linear model.

тут у параметрі `data` вказують фрейм, з якого беруться дані для побудови регресійної моделі.

Параметр `formula` (**формула**) задає те, що називають **специфікацією моделі**, тобто тут ми пояснюємо комп'ютеру, що треба використати як відгук, а що — як регресори.

Наприклад, формула

```
Y~X1+X2+X3
```

вказує, що змінна `Y` обирається як відгук, а `X1`, `X2`, `X3` — як регресори, тобто модель регресії має вигляд

$$\texttt{Y}_j = b_0 + b_1\texttt{X1}_j + b_2\texttt{X2}_j + b_3\texttt{X3}_j + \varepsilon_j,$$

причому коефіцієнти b_i, $i = 0, \ldots, 3$ невідомі, їх треба оцінити.

У формулах можна використовувати математичні дії:

```
log(Y)~log(X)
```

відповідає регресійній моделі

$$\log(\texttt{Y}_j) = b_0 + b_1 \log(\texttt{X}_j) + \varepsilon_j.$$

Зрозуміло, що коли ви хочете провести регресію, наприклад, з відгуком `Y`, використовуючи як регресор суму змінних `X1` і `X2`, то запис `Y~X1+X2` не дасть бажаного результату. У таких випадках використовують функцію `I()`, яка і сама нічого не робить, і забороняє інтерпретатору формул обробляти свій аргумент (інгибітор інтерпретації):

```
Y~I(X1+X2)
```

— спочатку обчислюється вираз у дужках після `I`, а потім результат використовується як регресор у моделі

$$\log(\texttt{Y}_j) = b_0 + b_1(\texttt{X1}_j + \texttt{X2}_j) + \varepsilon_j.$$

Інколи буває необхідно підігнати регресійну модель, у якій коефіцієнт b_0 відсутній (дорівнює 0)[13]:

$$\texttt{Y}_j = b_1\texttt{X1}_j + b_2\texttt{X2}_j + b_3\texttt{X3}_j + \varepsilon_j.$$

[13] Англійською це називають *regression through origin* — регресія через початок координат.

Щоб пояснити це комп'ютеру, у формулі дописують `-1`:

```
Y~X1+X2-1
```

(Більш повно про позначення у формулах для `lm()` див. підрозд. 10.3).

Результатом роботи функції `lm()` є об'єкт складної структури, що містить всі результати підгонки та перевірки гіпотез про модель. Інформацію про конкретні значення можна отримувати, звертаючись до його атрибутів. Наприклад, якщо виконано привласнення

```
model<-lm(Y~X),
```

то результат підгонки регресії $Y_j = b_0 + b_1 X_j + \varepsilon_j$ буде вміщений у змінну `model`. Зокрема, значення МНК-оцінок коефіцієнтів $(\hat{b}_0, \hat{b}_1)$ можна отримати в атрибуті

```
model$coefficients.
```

Прогнози для значень відгуку у точках спостережень, тобто $\hat{Y}_j = \hat{b}_0 + \sum_{j=1}^{n} \hat{b}_i X_j^i$ містяться у атрибуті

```
model$fitted.values.
```

Різниця між справжнім значенням відгуку та прогнозом називається *залишком* (*residual*) $U_j = Y_j - \hat{Y}_j$. Залишки можна отримати у атрибуті

```
model$residuals.
```

Окрім того, інформацію про результати регресійного аналізу можна подивитись, використовуючи функцію `summary`. Виклик `summary(model)` виводить на екран не тільки значення коефіцієнтів, але і величину коефіцієнта детермінації моделі та результати перевірки залежності відгуку від регресорів.

Коефіцієнтом детермінації лінійної регресійної моделі називають

$$R^2 = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2}{\sum_{j=1}^{n} (Y_j - \bar{Y})^2} = 1 - \frac{\sum_{j=1}^{n} U_j^2}{\sum_{j=1}^{n} (Y_j - \bar{Y})^2}.$$

Ця величина показує, яка частка розкиду (дисперсії) відгуку відтворюється прогнозом на основі даної регресійної моделі. Коефі-

цієнт детермінації завжди невід'ємний і не перевищує одиниці:

$$0 \leq R^2 \leq 1.$$

Чим більший коефіцієнт детермінації, тим точнішим є прогноз на тих даних, за якими підганялась модель регресії. Тому R^2 прийнято використовувати для грубої характеризації якості моделі[14]. Інколи для цього застосовують так званий **виправлений коефіцієнт детермінації**[15]:

$$R^2_{adj} = 1 - \frac{\sum_{j=1}^{n} U_j^2/(n-m-1)}{\sum_{j=1}^{n}(Y_j - \bar{Y})^2/(n-1)}.$$

(При такому виправленні у чисельнику стоїть незміщена оцінка для дисперсії похибок, а у знаменнику — незміщена оцінка для дисперсії відгуку).

Пояснювати загальну техніку підгонки регресійних моделей з використанням графічних засобів, МНК-підгонки та тестів для перевірки гіпотез зручно на конкретних прикладах, до яких ми і перейдемо.

Приклад 10.2.1. Розглянемо дані про автомобілі з фрейма `mtcars`, що входить у стандартну поставку `R`. (Ми вже досліджували ці дані у прикладі 5.3.1). У ньому розглядаються технічні характеристики 32 моделей автомобілей. Зокрема, для кожної моделі вказано такі змінні:

`mpg` — MPG автомобіля, тобто скільки миль він може пройти на одному галоні пального (див. приклад 4.1.4);

`disp` — об'єм циліндрів у кубічних дюймах;

`hp` — потужність двигуна у кінських силах (horspower);

`wt` — вага у тисячах фунтів.

[14] Насправді у більшості випадків прогнозні формули використовують для того, щоб прогнозувати значення відгуку для тих значень регресорів, яких немає у вибірці. Щоб оптимізувати точність такого майбутнього прогнозування, треба використовувати не R^2, а інші характеристики якості.

[15] Англ. *adjusted* R^2.

Серед цих змінних `disp`, `hp` і `wt` є технічними, вони визначаються на стадії розробки автомобіля. А `mpg` — це споживча характеристика, яка залежить від технічних, але не очевидно, чи можна визначити цю залежність з яких-небудь теоретичних міркувань. Тому розумно спробувати підібрати відповідну формулу за емпіричними даними.

Почнемо з того, що відобразимо дані на **матричній діаграмі розсіювання**[16] — рис. рис. 10.9.

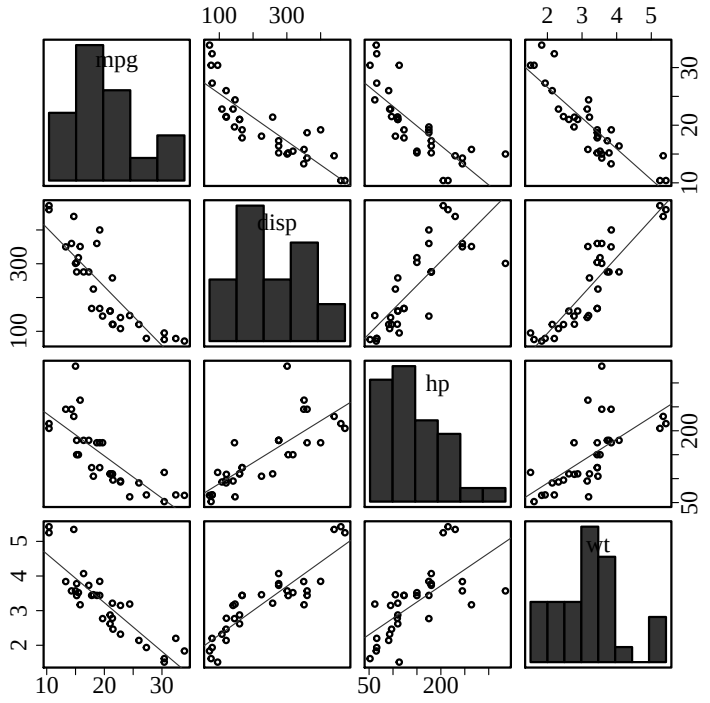

Рис. 10.9. Матрична діаграма для даних про автомобілі

[16] Див. підрозд. 5.1.

Для відображення діаграми використана функція з бібліотеки `car` — `scatterplotMatrix()`:

```
> mc <- mtcars[,c("mpg","disp","hp","wt")]
> library(car)
> scatterplotMatrix(mc,diagonal="histogram",
>                   smoother=F)
```

У квадратах на діагоналі цієї діаграми розміщені гістограми значень змінних. У недіагональних квадратах — діаграми розсіювання відповідних пар змінних. Наприклад, для квадратів у першому рядку по вертикалі відкладається `mpg`. Для квадратів у другому стовпчику — по горизонталі `disp`. Прямі лінії на діаграмах — це лінії регресії на основі МНК для кожної пари змінних. Продивляючись рисунок, переконуємось, що для всіх пар змінних помітна залежність між ними, хоча і не строга функціональна, а розмита випадковими відхиленнями. Немає яких-небудь явних нелінійностей або інших особливостей, скажімо, викидів. Тому природно зупинитись на лінійній регресії для опису залежності та використати МНК для її підгонки.

Виконаємо лінійну регресію, застосовуючи функцію `lm()`:

```
> model <- lm(mpg~disp+hp+wt, data = mtcars)
> summary(model)

Call:
lm(formula = mpg ~ disp + hp + wt, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max
-3.891 -1.640 -0.172  1.061  5.861

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 37.105505   2.110815  17.579  < 2e-16 ***
disp        -0.000937   0.010350  -0.091  0.92851
hp          -0.031157   0.011436  -2.724  0.01097 *
wt          -3.800891   1.066191  -3.565  0.00133 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.639 on 28 degrees of
freedom
Multiple R-squared:  0.8268,
Adjusted R-squared:  0.8083
F-statistic: 44.57 on 3 and 28 DF,  p-value: 8.65e-11
```

△

Ми зберігли результати аналізу у змінній `model` і вивели їх, використовуючи функцію `summary`.

Спочатку виводиться `Call`, тобто опис завдання, яке було поставлено перед функцією `lm()`. Далі — короткий опис залишків прогнозу (`Residuals`): найменше, найбільше значення і квартилі. З них можна скласти перше враження про те, наскільки вдалим вийшов прогноз.

Далі йде таблиця основних результатів підгонки. Кожен рядок таблиці відповідає одному коефіцієнту регресії (перший `Intercept` — $\hat{b}_0$, далі йдуть коефіцієнти при регресорах).

У стовпчику `Estimate` містяться оцінені значення коефіцієнтів, у стовпчику `Std. Error` — середньоквадратичні похибки цих оцінок[17]. Наприклад, оцінка $\hat{b}_3$ для коефіцієнта при `wt` дорівнює -3.800891, а її середньоквадратична похибка $\hat{\sigma}_i$ — 1.066191.

Далі йдуть два стовпчики, пов'язані з перевіркою того, чи є залежність відгуку від відповідного регресора. За формулою (10.10)

[17] Точніше, це оцінки для середньоквадратичних похибок. Їх можна використати для побудови довірчих інтервалів для справжніх коефіцієнтів. Для побудови довірчого еліпсоїда може бути потрібна коваріаційна матриця оцінок. Її можна оцінити, застосовуючи до результатів `lm()` функцію `vcov(model)`.

відгук Y_j залежить від регресора X_j^i тоді і тільки тоді, коли $b_i \neq 0$. Тому фактично перевіряється гіпотеза $H_0 : b_i = 0$, проти альтернативи $H_1 : b_i \neq 0$. Для цього обчислюють Т-статистику Стьюдента (вона виведена у третьому стовпчику таблиці)

$$T_i = \frac{\hat{b}_i}{\hat{\sigma}_i}$$

і порівнюють її з пороговим значенням, що відповідає заданому рівню значущості α. Як це прийнято у комп'ютерній статистиці, результати тестування подані у термінах `p-level` — досягнутого рівня значущості[18] (четвертий стовпчик таблиці результатів). Для того, щоб зробити висновок, `p-level` треба порівняти із α, тобто тестова процедура має такий вигляд:

> Якщо `p-level`$< \alpha$, приймають альтернативу — залежність між Y та X^i є, цей регресор доцільно включити у прогнозну формулу. (Значущий регресор).
> Якщо `p-level`$> \alpha$, приймають основну гіпотезу — залежність між Y та X^i не виявлена, цей регресор, можливо, слід вилучити з формули для прогнозу. (Незначущий регресор).

У таблиці зірочками відмічені рядки, які відповідають значущим регресорам для найбільш популярних рівнів значущості α. Якщо рядок відмічено хоча б однією зірочкою, він є значущим при $\alpha = 0.05$, двома і більше — значущим при $\alpha = 0.01$, трьома — $\alpha = 0.001$.

Таким чином, у нашому прикладі при рівні значущості $\alpha = 0.05$ слід вважати значущими вільний член b_0 (`Intercept`), `hp` і `wt`. Змінну `disp` можна спробувати вилучити з прогнозної формули.

Якщо включити у прогнозну формулу для `mpg` всі регресори, вона матиме вигляд

$$\widehat{\texttt{mpg}} = 37.11 - 0.000937 \times \texttt{disp} - 0.031157 \times \texttt{hp} - 3.80 \times \texttt{wt}$$

[18] Про досягнуті рівні значущості див. підрозд. 9.1.

— mpg автомобіля зменшується при збільшенні об'єму циліндрів, потужності двигуна і ваги.

Наскільки хорошим є цей прогноз? Коефіцієнт детермінації R^2 нашої моделі — `Multiple R-squared` дорівнює 0.8268, тобто прогноз пояснює 82% мінливості mpg. Це не так добре, як хотілося б, але вже досить для того, щоб таку формулу можна було використовувати хоча б для грубої прикидки при проектуванні нового автомобіля[19].

Крім того, у таблиці результатів, виведеній `summary()`, є повідомлення про результати F-тесту Фішера для перевірки того, що відгук залежить від хоча б одного з регресорів[20]. Досягнутий рівень значущості для цього тесту на наших даних дорівнює 8.65×10^{-11}, тобто залежність слід вважати виявленою при будь-якому розумному стандартному рівні значущості. У даному прикладі це не дивно, тому що ми вже виявили значущу залежність від двох регресорів. Можливі випадки, коли залежність від деякого набору взятих разом регресорів виявляється, але жоден регресор не вдається визначити як значущий. Зокрема, це буває коли регресори сильно пов'язані між собою і важко встановити, від якого саме з них залежить відгук[21].

Перевірити якість прогнозування можна, використовуючи діаграму розсіювання, на якій для спостережуваних об'єктів прогноз відкладено по горизонталі, а справжнє значення відгуку — по вертикалі (**діаграма прогноз-відгук**). Для нашого прикладу це можна зробити так:

```
> plot(model$fitted.values,mc$mpg,
+      xlab="mpg forecast",ylab="true mpg")
> abline(c(0,1),col="red")
```

[19] Зазвичай прогнози з $R^2 < 0.8$ не рекомендують для практичного застосування, хоча такі формули часом можуть бути корисні для якісного опису досліджуваного явища і висунення яких-небудь теоретичних гіпотез.

[20] Про тест Фішера для лінійної регресії див. підрозд. 10.4.

[21] Це називається мультиколлінеарністю, див. [17] і підрозд. 3.4. у [11].

(рис. 10.10). Ми бачимо, що точки на діаграмі витягнулись вздовж бісектриси першого координатного кута (червона пряма), отже, прогноз досить добре відтворює справжні значення `mpg`. Придивившись, можна помітити, що при дуже малих і дуже великих значеннях прогнозу точки лежать над червоною прямою, а посередині більше точок під прямою, ніж над нею. Це може свідчити про наявність неврахованої нелінійної залежності відгуку від регресорів.

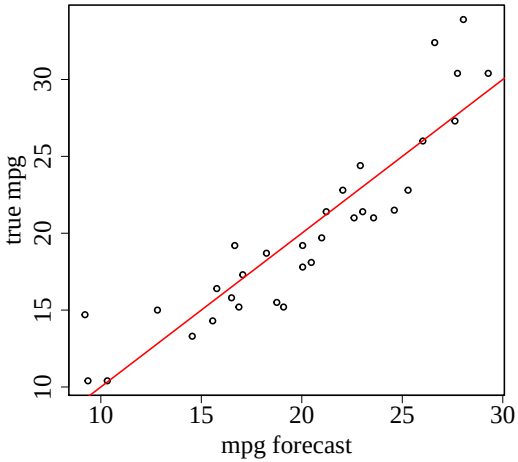

Рис. 10.10. Діаграма прогноз-відгук для mpg

Щоб помітити такі нелінійні ефекти, краще користуватись `діаграмою прогноз-залишки`:

```
> plot(model$fitted.values,model$residuals,
+ xlab="prediction",ylab="residuals")
> abline(0,0,col="red")
```

— див. рис. 10.11. Тут провал посередині відчувається більш виразно. Треба сказати, що при такій малій кількості спостережень в око

часто впадають особливості на діаграмах, що утворились зовсім випадково. У даному випадку однозначно висловитись на користь застосування нелінійної моделі не можна. Відкладемо поки що це питання. На діаграмі не помітно яких-небудь інших особливостей (викидів, розбиття на кластери), точки розкидані більш-менш хаотично. Це свідчить про те, що наша підгонка лінійної регресійної формули виявила всі основні закономірності, які можна було використати для прогнозування. Всі відхилення від прогнозу мають випадковий характер[22].

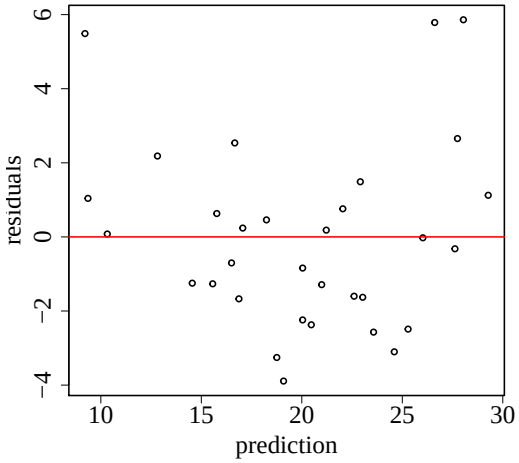

Рис. 10.11. Діаграма прогноз-залишки для mpg

Нарешті, важливим є питання про розподіл похибок регресії. Його прийнято перевіряти за залишками, використовуючи гістограми, Q-Q- та P-P-діаграми[23]. У нашому випадку гістограма бу-

[22] Маю на увазі випадковість в межах нашої моделі. Можливо, якщо врахувати якісь інші параметри автомобіля, наприклад його форму, відхилення, з нашого погляду випадкові, можна було б пояснити і прогноз для `mpg` покращити.

[23] Див. підрозд. 7.2.

де мало інформативною внаслідок малої кількості спостережень. Побудуємо Q-Q-діаграму для перевірки нормальності розподілу похибок:

```
> qqnorm(model$residuals,xlab=" ",ylab=" ",main=" ")
> qqline(model$residuals,col="red")
```

(рис. 10.12). Основна маса спостережень добре вкладається на пряму, тобто гіпотеза про нормальний розподіл похибок підтверджується. Але при великих значеннях квантилів маємо три точки, що помітно відхилились вгору. Це також може бути свідченням невідповідності нашої моделі, хоча б для цих автомобілів. Проте такий ефект міг би скластись випадково. При малому обсязі вибірки обґрунтовано вважати ці спостереження викидами-забрудненнями не можна.

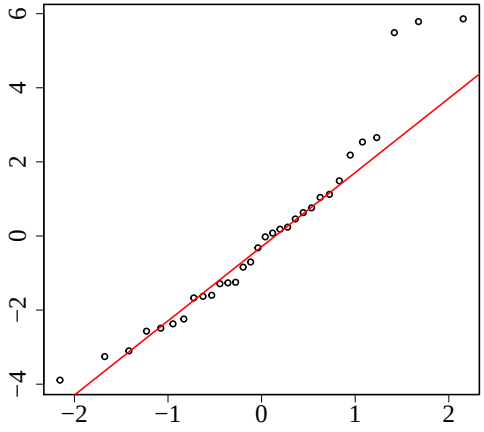

Рис. 10.12. Q-Q-діаграма для залишків прогнозу mpg

Таким чином, у нас є непогана лінійна модель, але є також

і сумнів щодо можливості більш точного прогнозування з використанням нелінійності. На матричній діаграмі розсіювання (рис. 10.9) нелінійності не помітні. Але ця діаграма показує лише попарні залежності між змінними. Можливо, при формуванні `mpg` треба враховувати не тільки вплив кожного регресора окремо, а і їхню взаємодію. Для цього у прогнозну формулу слід включити якусь нелінійну функцію від кількох регресорів одразу. Розкид точок на рис. 10.11 нагадує розмиту параболу. Це дає натяк на використання функцій другого порядку. Квадрати наших змінних використовувати як додаткові регресори не варто, така нелінійність була б помітною на матричній діаграмі. Краще спробувати моделювати взаємодію, використовуючи попарні добутки регресорів.

Вводити у модель багато регресорів небезпечно, особливо, якщо спостережень порівняно мало. При цьому може виникнути ефект перепідгонки, коли прогноз відтворює не загальні закономірності явища, а ті випадкові особливості вибірки, які не будуть повторюватись у майбутньому. Тому, додаючи нові регресори, треба бути обережним і дивитись, чи не можна щось вилучити. Підганяючи лінійну формулу, ми не побачили значущої залежності `mpg` від `disp`. Вилучимо цей регресор, але включимо у модель регресії доданок `hp`×`wt`:

$$\texttt{mpg}_j = b_0 + b_1 \times \texttt{hp}_j + b_2 \times \texttt{wt}_j + b_{12} \times \texttt{hp}{\times}\texttt{wt}_j + \varepsilon_j.$$

Це можна зробити, задавши формулу

```
mpg~hp+wt+I(hp*wt)^2
```

Інший спосіб задати ту ж саму модель:

```
mpg~(hp+wt)^2
```

(див. п. 10.3).

Отже, викликаємо функцію `lm()`:

```
> model2 <- lm(mpg~(hp+wt)^2, data = mc)
> summary(model2)
```

```
Call:
lm(formula = mpg ~ (hp + wt)^2, data = mc)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0632 -1.6491 -0.7362  1.4211  4.5513

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.80842    3.60516  13.816 5.01e-14 ***
hp          -0.12010    0.02470  -4.863 4.04e-05 ***
wt          -8.21662    1.26971  -6.471 5.20e-07 ***
hp:wt        0.02785    0.00742   3.753 0.000811 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.153 on 28 degrees
of freedom
Multiple R-squared:  0.8848,
Adjusted R-squared:  0.8724
F-statistic: 71.66 on 3 and 28 DF,  p-value: 2.981e-13
```

△

Як бачимо, коефіцієнт детермінації цієї моделі помітно кращий, ніж у попередньої — $R^2 = 0.8848$, хоча кількість коефіцієнтів, що оцінювались, залишилась тою самою.

Формула прогнозу має вигляд

$$\widehat{\texttt{mpg}} = 49.80842 - 0.12010 \times \texttt{hp} - 8.21662 \times \texttt{wt} + 0.02785 \times \texttt{hp}{\times}\texttt{wt}.$$

(У формулах R регресор `hp`×`wt` дещо парадоксально позначається `hp:wt`, а `*` зарезервована для іншого).

Всі коефіцієнти у цій формулі значущо відрізняються від 0 при всіх розумних рівнях значущості.

Проведемо графічний аналіз залишків (рис. 10.13).

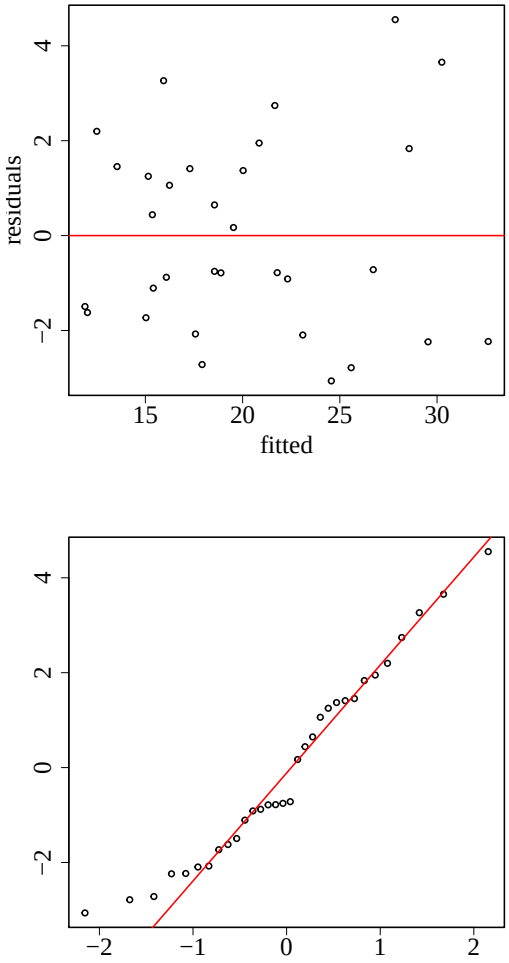

Рис. 10.13. Діаграма розсіювання та Q-Q-діаграма для залишків прогнозу mpg з квадратичним доданком

```
> plot(model2$residuals~model2$fitted.values,
+     xlab="fitted",ylab="residuals")
> abline(0,0,col="red")
> qqnorm(model2$residuals,
+       xlab=" ",ylab=" ",main=" ")
> qqline(model2$residuals,col="red")
```

Тепер вже впевнено можна сказати, що залишки на діаграмі розсіювання розкидані хаотично, а Q-Q-діаграма свідчить про нормальність розподілу похибок регресії. ◄

Приклад 10.2.2. Вага та довжина оселедців. У файлі `herring2.txt` містяться дані про довжину (`LENGTH`) та вагу (`WEIGHT`) оселедців, виловлених у Північному морі. Нас цікавить, чи є залежність між цими змінними і, якщо є, то яка саме.

Почнемо з діаграми розсіювання:

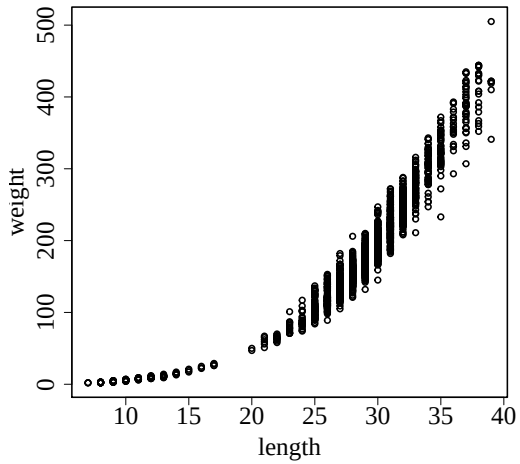

Рис. 10.14. Довжина та вага оселедців

На рис. 10.14 помітна залежність, дещо розмита випадковим розкидом. Лінійність цієї залежності сумнівна.

Дані були прочитані у фрейм `herr` і рис. 10.14 відображений таким скриптом:

```
> herr<-read.table(file=
+                  "c:/rem/term/herring.txt",header=T)
> plot(herr[,"LENGTH"],herr[,"WEIGHT"],
+ xlab = "length",ylab="weight")
```

Проведемо лінійну підгонку за методом найменших квадратів і відобразимо діаграму прогноз-залишки (рис. 10.15).

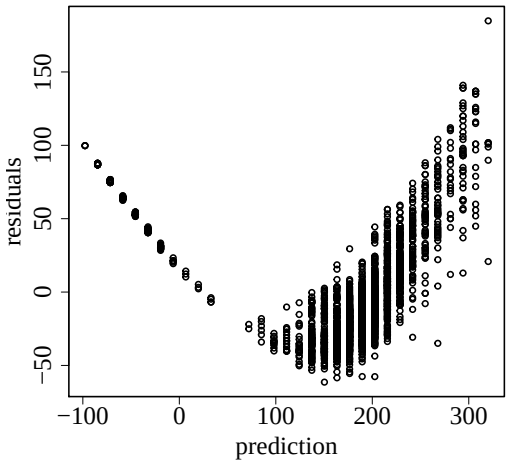

Рис. 10.15. Залишки лінійної моделі для оселедців

Для підгонки моделі був використаний такий скрипт:

```
> resLm<-lm(WEIGHT~LENGTH,data=herr)
```

```
> plot(resLm$fitted.values,resLm$residuals,
+ xlab="prediction",ylab="residuals")
```

△

На цій діаграмі явно простежується нелінійна залежність, що не була врахована нашою моделлю. Окрім того, при зростанні прогнозованих значень WEIGHT зростає розкид залишків. Це наводить на думку про застосування моделі з мультиплікативною похибкою:

$$\text{WEIGHT}_j = C \times \text{HEIGHT}_j^a \times \eta_j,$$

де C і a — невідомі параметри, η — мультиплікативна похибка.[24] У такій моделі із зростанням відгуку зростатиме і його розкид навколо лінії регресії.

Щоб мати змогу застосовувати модель лінійної регресії, перетворимо дані — перейдемо до змінних $\text{LW} = \log(\text{WEIGHT})$, $\text{LL} = \log(\text{LENGTH})$. В результаті модель зводиться до лінійної з адитивною похибкою $\varepsilon = \log(\eta)$:

$$\text{LW}_j = a\text{LL}_j + b + \varepsilon_j,$$

де $b = \log(C)$. Проведемо підгонку цієї моделі. Обчислення можна виконати таким скриптом:

▽

```
> resLog<-lm(log(WEIGHT)~log(LENGTH),data=herr)
> summary(resLog)

Call:
lm(formula = log(WEIGHT) ~ log(LENGTH), data = herr)

Residuals:
     Min       1Q   Median       3Q      Max
```

[24] Параметр a можна трактувати як "вимірність риби": якщо вважати, що вага приблизно пропорційна об'єму, то для риб, що ростуть переважно у довжину, як вугрі, $a \approx 1$, для плоских риб, подібних до камбали $a \approx 2$, для риб, що ростуть у всіх трьох вимірах, $a \approx 3$.

```
-0.33052 -0.05565 -0.00355  0.05627  0.32250

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.984646   0.022678  -263.9   <2e-16 ***
log(LENGTH)  3.309440   0.006833   484.4   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09058 on
2271 degrees of freedom
Multiple R-squared:  0.9904,
Adjusted R-squared:  0.9904
F-statistic: 2.346e+05 on 1 and 2271 DF,
p-value: < 2.2e-16
```

Ми отримали оцінку для коефіцієнта $a = 3.309440$ — це схоже на очікуване значення 3. Отримані значення коефіцієнтів значущо відрізняються від 0.

Діаграма прогноз-залишки для цієї моделі зображена на рис. 10.16. На ній помітні дві хмари даних, які, можливо, пов'язані із двома різними популяціями оселедців (довших та коротших), або коротші оселедці можуть бути молодими рибами (цього року народження), а довші — дорослими. Всередині хмар закономірності не помітні, точки розкидані хаотично. Обидві хмари однаково розкидані по вертикальній осі (що відповідає залишкам) вони не відрізняються. Діаграма виведена таким скриптом:

```
> plot(resLog$fitted.values,resLog$residuals,
+ xlab="prediction",ylab="residuals")
```

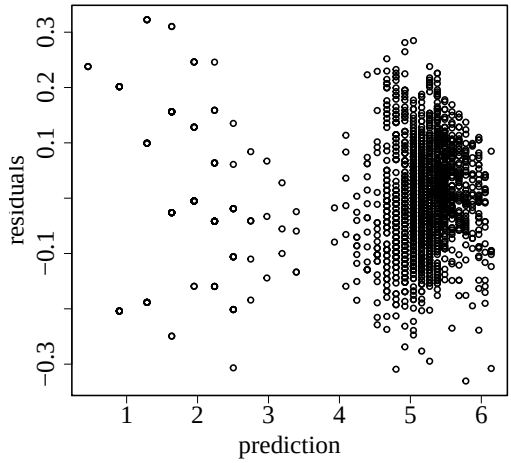

Рис. 10.16. Залишки лінійної моделі для оселедців

Перевіримо нормальність похибок у отриманій лінеаризованій моделі, використовуючи гістограму та QQ-діаграму (рис. 10.17).

```
> # Рисуємо гістограму:
> hi<-hist(resLog$residuals,  breaks=10,
+ xlab=" ", ylim=c(0, 600),
+ ylab=" ",
+ main=" ")
> # Графік нормальної щільності:
> curve(dnorm(x, mean=mean(resLog$residuals),
+ sd=sd(resLog$residuals))
+ *length(resLog$residuals)*(hi$breaks[2]-hi$breaks[1]),
+ col="darkblue", lwd=2, add=TRUE, yaxt="n")
> # Q-Q-діаграма:
> qqnorm(resLog$residuals)
> qqline(resLog$residuals,col="red")
```

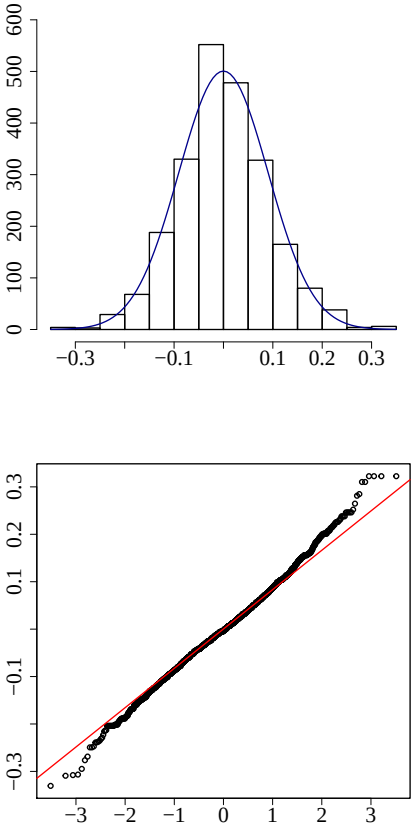

Рис. 10.17. Гістограма та Q-Q-діаграма залишків для оселедців

З гістограми на рис. 10.17 видно, що спостережуваний розподіл залишків для основної маси спостережень добре описується нормальним розподілом. Але на Q-Q-діаграмі помітні невеликі відхилення "хвостів" розподілу від нормальних. Отже, у першому наближенні (ігноруючи негауссову поведінку хвостів розподілу) ми

можемо прийняти таку модель даних

$$\text{WEIGHT}_j = 0.0025171 \times \text{LENGTH}_j^{3.309440}\eta_j,$$

де η — мультиплікативна похибка з логнормальніим розподілом.

Зауваження. Враховуючи те, що ми виділили два кластери — "молодих" та "дорослих" риб, цікаво перевірити, чи відрізнаються коефіцієнти у моделі залежності ваги від довжини для цих категорій риб. ◀

Приклад 10.2.3. Повернемось до аналізу даних прикладу 10.1.2. Розглянемо залежність змінної `x` (інтерес до джинсів) від `urban` (рівень урбанізації) у штатах США, що лежать у басейні Міссісіпі — Міссурі (крім Південної Дакоти). Як ми пам'ятаємо, використання повторної медіани для підгонки простої лінійної регресії дало у цьому прикладі нестабільні результати.

Якщо застосувати просту лінійну регресію з підгонкою за МНК, лінія регресії пройде майже горизонтально (червона лінія на рис. 10.18. Оцінка для коефіцієнта регресії $\hat{b}_1 = 0.0007286$, коефіцієнт детермінації моделі — 0.01128, досягнутий рівень значущості для перевірки впливу `urban` на `x` — 0.656.

Але на діаграмі розсіювання помітний досить виразний максимум посередині — ліворуч від нього залежність схожа на зростаючу, після — на спадну. Для підгонки такої залежності природно спробувати поліном другого порядку:

$$\texttt{x}_j = b_0 + b_1\texttt{urban}_j + b_2(\texttt{urban}_j)^2 + \varepsilon_j.$$

Це приклад так званої **поліноміальної регресії** другого порядку (квадратичної регресії). Для підгонки можна у функції `lm()` використати формулу

```
x~urban+I(urban^2).
```

як це зроблено у такому скрипті:

```
> tb<-read.table("c:/rem/term/shortU.txt",header=T)
> tb$x<-tb$jean/(tb$jean+tb$cargo)
```

```
> tbm<-tb[tb$miss==1,]
> tbm1<-tbm[tbm$x<1,]
> # Підгонка лінійної регресії:
> model<-lm(x~urban,data=tbm1)
> plot(tbm1$urban,tbm1$x,
+ xlab="urban",ylab="x")
> abline(model$coefficients,col="red")
> summary(model)


Call:
lm(formula = x ~ urban, data = tbm1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.081366 -0.033556  0.007042  0.030767  0.056444

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5323215  0.0521274  10.212 6.46e-09 ***
urban       -0.0003302  0.0007286  -0.453    0.656
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04101
on 18 degrees of freedom
Multiple R-squared:  0.01128,
Adjusted R-squared:  -0.04364
F-statistic: 0.2054 on 1 and 18 DF,
 p-value: 0.6558

> # Підгонка квадратичної регерсії:
> model2<-lm(x~urban+I(urban^2),data=tbm1)
> summary(model2)
```

```
Call:
lm(formula = x ~ urban + I(urban^2), data = tbm1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.052625 -0.016260  0.003862  0.013059  0.049120

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.759e-01  1.640e-01  -1.683 0.110722
urban        2.319e-02  4.692e-03   4.942 0.000124 ***
I(urban^2)  -1.657e-04  3.289e-05  -5.039 0.000101 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02673 on
17 degrees of freedom
Multiple R-squared:  0.6035,
Adjusted R-squared:  0.5568
F-statistic: 12.94 on 2 and 17 DF,
p-value: 0.000385
```

```
> b<-model2$coefficients
> curve(b[1]+b[2]*x+b[3]*x^2,col="blue",add=T)
```

△

В результаті отримуємо прогнозну формулу

$$\widehat{\mathtt{x}} = -0.2759 + 0.02319\mathtt{urban} - 0.0001657(\mathtt{urban})^2.$$

На перший погляд здається, що коефіцієнт при $\mathtt{urban}^2$ малий і несуттєвий. Але крива регресії (синя парабола на рис. 10.18) виразно відрізняється від прямої лінії. Коефіцієнт детермінації моделі зріс порівняно з лінійною версією, більше, ніж у п'ять

разів[25] — до 0.6035. Тепер значущо відмінними від 0 є і коефіцієнт при `urban`(`p-level`=0.000124), і коефіцієнт при `urban`^{2} (`p-level`=0.000101).

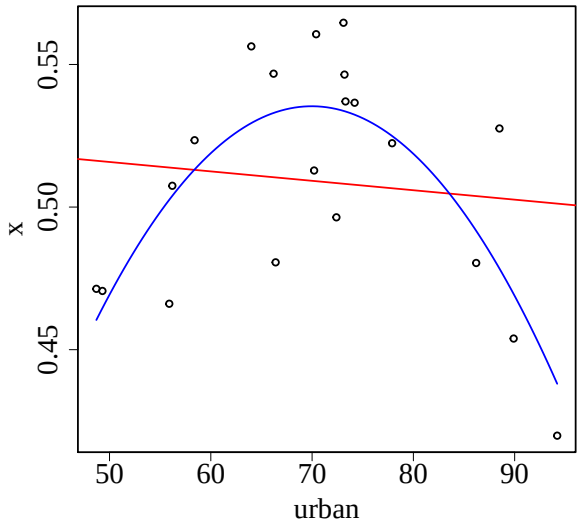

Рис. 10.18. Лінійна і квадратична регресія для даних про шорти

Цей приклад показує, що, не помітивши лінійної залежності між змінними, не варто зупинятись. Виявлення і аналіз нелінійностей можуть суттєво поліпшити прогноз на основі регресійної формули. ◀

[25] R^2 завжди зростає при включенні нових регресорів у регресійну формулу, але таке велике зростання при додаванні лише одного доданка — це дійсно виразний ефект.

10.3 Позначення у формулах для функції `lm()`

Досі розглядались лише приклади застосування формул `R` у викликах функцій. У цьому підрозділі систематично пояснюється, як складати формули специфікації моделей для функції `lm()`. Основні символи, що зустрічаються у таких формулах, наведені у табл. 10.1.

Таблиця 10.1. Символи формул для функції lm()

Символ	Приклад	Значення
`+`	`+X`	включити регресор у формулу
`-`	`-X`	вилучити регресор з формули
`:`	`A:B`	взаємодія: $A \times B$
`*`	`A*B`	включити ці змінні і їх взаємодію
`^`	`(A+B+C)^3`	включити змінні і всі взаємодії до 3-го порядку
`I`	`I(A+B)`	підрахувати вираз у дужках і використати як регресор
`1`	`-1`	регресія через 0 (без вільного члена)
`.`	`Y~ .`	включити у модель всі регресори фрейма даних

Спочатку розберемо на прикладах, як ці символи використовуються для опису специфікації моделі у випадку числових змінних, а потім з'ясуємо, які відмінності виникають при включенні до моделі змінних типу фактор.

Отже, якщо `Y`, `X`, `Z`, `V` — числові змінні, то формула `Y~X+Z+V` відповідає моделі

$$Y_j = b_0 + b_1 X_j + b_2 Z_j + b_3 V + \varepsilon_j$$

(тут і далі b_i позначають невідомі коефіцієнти регресії, що оцінюються, а ε_j — похибки регресії, які вважаються випадковими з нульовими математичними сподіваннями).

Позначення вигляду `X:Z:V` у формулах відповідають доданкам у регресійній моделі, що описують взаємодію регресорів. Для числових змінних ці доданки трактують як добутки відповідних регресорів: формула `Y\~X+Y:Z` відповідає моделі $Y_j = b_0 + b_1 X_j + b_2 Y_j Z_j + \varepsilon_j$.

Запис `Y~X*Z` позначає, що треба включити у модель і змінні `X`, `Z`, і їхню взаємодію[26]:

$$Y_j = b_0 + b_1 X_j + b_2 Z_j + b_3 X_j Z_j + \varepsilon_j.$$

Можна комбінувати символи, використовуючи дужки, наприклад, формула `Y~(X+Z)*V` відповідає моделі

$$Y_j = b_0 + b_1 X_j + b_2 Z_j + b_3 V_j + b_4 X_j V_j + b_5 Z_j V_j + \varepsilon_j.$$

Символ піднесення до ступеня — `^` або `**` позначає включення у модель всіх взаємодій відповідних змінних до даного ступеня: формула `Y~(X+Z+V)^3` відповідає моделі

$$Y_j = b_0 + b_1 X_j + b_2 Z_j + b_3 V_j +$$

$$+ b_4 X_j Z_j + b_5 Z_j V_j + b_6 X_j V_j + b_7 X_j Z_j V_j + \varepsilon_j.$$

Часто при звертанні до `lm()` у формулі вказують назви змінних з деякого фрейма, а у опції `data` назву самого фрейма. У цьому випадку у формулі можна використати символ `.` (крапка), щоб позначити всі змінні фрейма, крім тієї, яка обрана відгуком. Наприклад, якщо у фреймі `D` записані змінні `Y`, `X`, `Z` і `V`, то виклик функції `lm(Y~.,data=D)` відповідає моделі

$$Y_j = b_0 + b_1 X_j + b_2 Z_j + b_3 V_j + \varepsilon_j,$$

а виклик `lm(Y~.*V,data=D)` відповідає моделі

$$Y_j = b_0 + b_1 X_j + b_2 Z_j + b_3 V_j + b_4 X_j V_j + b_5 Z_j V_j + \varepsilon_j.$$

[26] Граматично правильно було б написати “і доданок, що описує їхню взаємодію”, але далі ми будемо скорочувати цю фразу до одного слова.

Інколи зручно вказати великий набір регресорів, а потім вилучити з нього зайві. Це робиться за допомогою символа "`-`" (мінус): формула `lm(Y~.-V,data=D)` відповідає моделі $Y_j = b_0 + b_1 X_j + b_2 Z_j + \varepsilon_j$.

Зокрема, якщо треба вилучити з моделі вільний член (коефіцієнт b_0), це можна зробити, вказавши у формулі `-1`: формула `Y~X+Z-1` відповідає $Y_j = b_1 X_j + b_2 Z_j + \varepsilon_j$.

При побудові специфікації моделі часто виникає потреба зробити перетворення початкових регресорів, наприклад, логарифмування. Це можна робити безпосередньо у формулі, наприклад, формула `Y~log(X)` відповідає $Y_j = b_0 + b_1 \log(X_j) + \varepsilon_j$.

Але при записі деяких перетворень може виявитись двозначність внаслідок відмінності семантики записів у формулах та у звичайних алгебраїчних перетвореннях. Скажімо, формула `Y~X^2` відповідає моделі $Y_j = b_0 + b_1 X_j$, оскільки "взаємодія" `X` з самим собою — це `X`. Тому, для того, щоб виконувати алгебраїчні перетворення всередині формул, використовують функцію `I()`, яка повертає значення свого аргументу (тотожна функція). Але це значення компілятор обчислює за звичайними алгебраїчними правилами: формула `Y~I(X^2)` відповідає моделі $Y_j = b_0 + b_1 (X_j)^2$. Так можна будувати формули для числових змінних.

Трактування **формул з участю факторів** суттєво відрізняється. Якщо змінна `A` має тип фактор, то вона може набувати значень лише з фіксованого скінченного набору рівнів. Позначимо можливі рівні фактора `A` через $a_1, \ldots, a_m$. Тоді формула `Y~A` задає модель

$$Y_j = b_1 + b_2 \mathbb{1}\{A_j = a_2\} + b_3 \mathbb{1}\{A_j = a_3\} + \cdots + b_M \mathbb{1}\{A_j = a_M\} + \varepsilon_j.$$

(Тут випадковою вважається лише похибка ε_j, A_j трактуються як невипадкові умови експерименту, в якому було отримано Y_j).

Цю модель можна трактувати так. Різним рівням фактора відповідають різні значення математичного сподівання $\mathsf{E}\, Y_j$ змінної Y_j. У моделі b_1 — це значення $\mathsf{E}\, Y_j$, якщо $A_j = a_1$. Коефіцієнт b_2 — це поправка, яку треба додати до b_1, щоб отримати $\mathsf{E}\, Y_j$, якщо $A_j = a_2$, тобто $\mathsf{E}\, Y_j = b_1 + b_2$.

І взагалі, при $i = 2, \ldots, M$,

$$\mathsf{E}\, Y_j = b_1 + b_i, \text{ якщо } A_j = a_i.$$

Таким чином, оцінюючи коефіцієнти b_i у цій моделі, ми оцінюємо математичні сподівання відгуку, які відповідають кожному можливому рівню фактора. Така дещо штучна схема дозволяє звести класичну модель однофакторного дисперсійного аналізу до моделі лінійної регресії.

Приклад 10.3.1. У прикладі 7.5.1 ми розглядали дані з фрейма `InsectSprays` про серію експериментів, де порівнюється дієвість різних інсектицидів. Відгуком є змінна `count` — кількість комах, що загинули у даному експерименті, а фактором — змінна `spray` — тип застосованого інсектициду.

Застосуємо функцію `lm()` для аналізу цих даних:

```
> summary(lm(count ~ spray, data = InsectSprays))

Call:
lm(formula = count ~ spray, data = InsectSprays)

Residuals:
   Min     1Q Median     3Q    Max
-8.333 -1.958 -0.500  1.667  9.333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.5000     1.1322  12.807  < 2e-16 ***
sprayB        0.8333     1.6011   0.520    0.604
sprayC      -12.4167     1.6011  -7.755 7.27e-11 ***
sprayD       -9.5833     1.6011  -5.985 9.82e-08 ***
sprayE      -11.0000     1.6011  -6.870 2.75e-09 ***
sprayF        2.1667     1.6011   1.353    0.181
---
```

```
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.922
on 66 degrees of freedom
Multiple R-squared:  0.7244,
Adjusted R-squared:  0.7036
F-statistic:  34.7 on 5 and 66 DF,
p-value: < 2.2e-16
```

△

З таблиці результатів видно, що оцінка математичного сподівання кількості комах `count`, що гинуть під дією інсектициду `A`[27] дорівнює 14.5. Оцінка математичного сподівання, що відповідає інсектициду `B` — на 0.8333 більше, тобто дорівнює $14.5 + 0.8333 = 15.3333$. Досягнутий рівень значущості в останньому стовпчику таблиці — $p = 0.604$ вказує на те, що відмінність дії інсектицидів `A` і `B`, незначуща. Дія інсектициду `C` відрізняється від дії `A` значущо — на -12.4167, тобто відповідне математичне сподівання дорівнює $14.5 - 12.4167 = 2.0833$ і т.д.

Як бачимо, для проведення такого однофакторного аналізу таблиця результатів `lm()` досить незручна. Для цього краще користуватись спеціальними функціями, що виконують дисперсійний аналіз, такими, як, наприклад, `aov()`. Цей приклад має на меті лише продемонструвати, як у формулах для `lm()` трактуються факторні змінні. ◀

Якщо змінні `A` і `V` є факторами з рівнями, відповідно a_m, $m = 1, \ldots, M$ і v_k, $k = 1, \ldots, K$, то формулі `Y~A+V` відповідатиме модель

$$Y_j = b_0 + \sum_{m=2}^{M} b_{1m} \mathbb{1}\{A_j = a_m\} + \sum_{k=2}^{K} b_{2k} \mathbb{1}\{V_j = v_m\} + \varepsilon_j$$

[27] `A` — перший з можливих рівнів фактора `sprey`, тому йому відповідає рядок `Intercept` у таблиці результатів.

— це те, що у дисперсійному аналізі називають двофакторною моделлю з перетином факторів без взаємодії.

Формулі `Y~A*V` відповідає двофакторна модель із взаємодією факторів, що має вигляд

$$Y_j = b_0 + \sum_{m=2}^{M} b_{m0}\mathbb{1}\{A_j = a_m\} + \sum_{k=2}^{K} b_{0k}\mathbb{1}\{\ V_j = v_k\}+$$

$$+ \sum_{m=2}^{M}\sum_{k=2}^{K} b_{mk}\mathbb{1}\{A_j = a_m,\ V_j = v_k\} + \varepsilon_j.$$

Такі моделі також краще досліджувати, використовуючи функцію `aov()`.

Використання `lm()` з факторними змінними доцільне, якщо треба визначити, як фактори взаємодіють з числовими змінними. Про це піде мова у наступному підрозділі.

10.4 Перевірка лінійних гіпотез. Тест Фішера

З'ясуємо, як за спостереженнями перевіряти гіпотези про коефіцієнти лінійної регресійної моделі. Для цього застосовується тест Фішера[28]. Спочатку ми визначимо клас гіпотез, які будуть перевірятись, далі опишемо формальну схему тесту, а потім розглянемо приклади. Один важливий приклад ми виділимо окремо: тести для перевірки наявності розшарування за деякою ознакою (такі, як тест Чоу).

Загальна лінійна гіпотеза. Обмежимось розглядом функціональних лінійних регресійних моделей з гауссовими похибками. Точніше, будемо вважати, що дані про кожен об'єкт спостереження містять значення змінної-відгуку Y_j та d змінних-регресорів

[28] Існує багато різних тестів, які називають тестами Фішера. Ті, що розглядаються у цьому підрозділі можна назвати тестами Фішера для перевірки лінійних гіпотез у лінійних регресійних моделях. Їх не треба плутати, наприклад, з точним тестом Фішера для таблиць 2×2 — це зовсім інший тест.

$X_j^1,\ldots,\ X_j^d$ (тут j — номер спостереження, $j = 1,\ldots,n$). Регресійна модель має вигляд

$$Y_j = \sum_{i=1}^{d} b_i X_j^i + \varepsilon_j, \tag{10.11}$$

де ε_j — похибки регресії, які вважаються незалежними гауссовими випадковими величинами з нульовим математичним сподіванням і однаковими дисперсіями σ^2. Регресори трактуються як невипадкові фіксовані числа. Коефіцієнти регресії b_i та σ^2 вважаються невідомими[29]. Вектор-стовпчик, складений з усіх коефіцієнтів регресії, будемо позначати $\mathbf{b} = (b_1,\ldots,b_d)^T$

Якщо на значення коефіцієнтів b_i не накладаються жодні обмеження, то модель (10.11) називають **необмеженою** (*unrestricted model*) і позначають літерою U. Таким чином,

гіпотеза (U): $\mathbf{b} \in \mathbb{R}^d$.

Гіпотези, які ми будемо перевіряти, полягають в тому, що для коефіцієнтів регресії виконуються обмеження, які задаються системою лінійних рівнянь. Нехай

$$\mathbf{L} = \begin{pmatrix} l_{11} & l_{12} & \ldots & l_{1d} \\ l_{21} & l_{22} & \ldots & l_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \ldots & l_{pd} \end{pmatrix}$$

— деяка фіксована (відома) матриця, а $\mathbf{c} = (c_1,\ldots,c_p)^T$ — фіксований вектор-стовпчик. Задамо **обмежену модель** (*restricted model*) як модель (10.11) за умови, що для вектора коефіцієнтів виконується обмеження

$$\mathbf{Lb} = \mathbf{c}. \tag{10.12}$$

[29] Регресійну модель з вільним членом $Y_j = b_0 + \sum_{i=1}^{m} b_i X_j^i + \varepsilon_j$ можна звести до вигляду (10.11), якщо формально ввести додатковий регресор $X_j^0 = 1$.

Це саме обмеження можна записати у вигляді системи p лінійних рівнянь:

$$\begin{cases} l_{11}b_1 + l_{12}b_2 + \cdots + l_{1d}b_d = c_1 \\ l_{21}b_1 + l_{22}b_2 + \cdots + l_{2d}b_d = c_2 \\ \qquad\qquad\qquad\qquad\qquad \cdots \\ l_{p1}b_1 + l_{p2}b_2 + \cdots + l_{pd}b_d = c_p \end{cases} \tag{10.13}$$

Надалі ми вважаємо, що $p \le d$, рядки матриці $\mathbf{L}$ лінійно незалежні, отже, (10.13) має хоча б один розв'язок і ні одне з рівнянь у цій системі не є зайвим (не може бути виражене через інші).

Гіпотезу про те, що у моделі (10.11) виконуються обмеження (10.13), будемо називати **обмеженою гіпотезою** і позначати (R). Оскільки обмеження (10.13) є лінійними, то цю гіпотезу також називають загальною лінійною гіпотезою.

Тест Фішера призначений для перевірки основної гіпотези H_0: виконано (R) проти альтернативи H_1: виконано (U), але обмеження (R) не виконуються. Саме такі гіпотези і розглядаються у цьому підрозділі.

Частковим прикладом лінійної гіпотези є гіпотеза про те, що коефіцієнт при даному регресорі дорівнює деякому фіксованому числу: $H_0 : \ b_i = c$. Приклади перевірки таких гіпотез ми розглядали у підрозд. 10.2 для випадку $c = 0$. Там для цього використовувався Т-тест Стьюдента, що є в даному випадку еквівалентним відповідному тесту Фішера.

Іншим прикладом лінійної гіпотези може бути припущення про те, що два регресори, скажімо, X^1 і X^2, впливають на відгук однаково: $H_0 : \ b_1 = b_2$.

Побудова тесту Фішера[30]**.** Нагадаємо, що оцінка методу найменших квадратів $\hat{\mathbf{b}}$ для коефіцієнтів регресії у моделі (U) визначається як те значення $\mathbf{b}$, при якому досягається мінімум функціоналу найменших квадратів

$$\hat{\mathbf{b}} = \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^d} J(\mathbf{b}),$$

[30]Теоретичне обґрунтування цього тесту див. [11], п. 2.4.

де мінімум береться за всіма можливими значеннями коефіцієнтів $\mathbf{b}$. Тут

$$J(\mathbf{b}) = \sum_{j=1}^{n}(Y_j - \sum_{i=1}^{d} b_i X_j^i)^2$$

— функціонал методу найменших квадратів (МНК).

Аналогічно можна ввести оцінки МНК в припущенні справедливості обмеженої моделі (R):

$$\hat{\mathbf{b}}^R = \underset{\mathbf{b}:\ \mathbf{Lb}=\mathbf{c}}{\operatorname{argmin}} J(\mathbf{b})$$

— мінімум береться лише за тими значеннями коефіцієнтів, які задовольняють (R).

Статистика тесту Фішера будується на порівнянні сум квадратів залишків прогнозу в обмеженій і необмеженій моделях.

Сума квадратів у необмеженій моделі:

$$SSU = J(\hat{\mathbf{b}}) = \sum_{j=1}^{n}(U_j)^2,$$

де

$$U_j = Y_j - \hat{Y}_j = Y_j - \sum_{i=1}^{d} \hat{b}_i X_j^i$$

— залишки прогнозу у необмеженій моделі.

Сума квадратів у обмеженій моделі:

$$SSR = J(\hat{\mathbf{b}}^R) = \sum_{j=1}^{n}(U_j^R)^2,$$

де

$$U_j^R = Y_j - \hat{Y}_j^R = Y_j - \sum_{i=1}^{d} \hat{b}_i^R X_j^i$$

— залишки прогнозу в необмеженій моделі.

Статистика тесту, яка називається **емпіричним F-відношенням Фішера**, визначається як

$$F_{emp} = \frac{(SSR - SSU)/p}{SSU/(n-d)},$$

де p — кількість рівнянь, що задають обмеження (R);

d — кількість коефіцієнтів регресії, що оцінюються у необмеженій моделі (U);

n — кількість всіх спостережень, за якими проводиться підгонка моделей.

Для того, щоб прийняти рішення, F_{emp} порівнюють з пороговим значенням, яке називається **F-теоретичне**:

$$F_{theor} = Q^{F(p,n-d)}(1-\alpha),$$

де α — заданий стандартний рівень значущості.

Сам **тест Фішера** має вигляд

Якщо $F_{emp} \le F_{theor}$ — прийняти (R), інакше — відхилити.

Досягнутий рівень значущості тесту —

$$p(F_{emp}) = 1 - G(F_{emp}),$$

де G — функція розподілу F-розподілу Фішера з p ступенями вільності чисельника і $n-d$ ступенями вільності знаменника.

Зауважимо, що (R) є гіпотезою, вкладеною в (U) у розумінні п. 9.3.2, тому для її перевірки можна використовувати тест відношення вірогідності з вибором порогу за асимптотичною формулою (9.9). Але, оскільки такий тест є асимптотичним, то він працюватиме адекватно лише на вибірках великого обсягу. Тест Фішера також є версією тесту відношення вірогідності, але він не асимптотичний і дозволяє робити висновки за вибірками будь-якого обсягу. При цьому для малих вибірок висновок тесту Фішера буде суттєво використовувати припущення про нормальність похибок дисперсії.

(На вибірках великого обсягу висновки тесту Фішера й асимптотичного тесту відношення вірогідності будуть майже збігатись).

Приклади застосування тесту Фішера. У `R` для перевірки гіпотез тестом Фішера можна використовувати функцію `linearHypothesis()` з бібліотеки `car`. Типове звертання до цієї функції має вигляд

```
linearHypothesis(model, hypothesis.matrix,
                 rhs, test),
```

де

`model` — результат підгонки необмеженої моделі, наприклад, функцією `lm()`;

`hypothesis.matrix` — матриця системи обмежень (матриця $\mathbf{L}$ в (10.12)). За умовчанням вважається, що $\mathbf{c} = 0$;

`rhs` — вектор правих частин системи обмежень (вектор $\mathbf{c}$ в (10.12));

`test` — параметр, що вказує, яку статистику буде використовувати тест: значення `"F"` відповідає F-відношенню Фішера, `"Chisq"` — статистиці відношення вірогідності (яка при великих обсягах вибірки має приблизно χ^2-розподіл). Якщо параметр `model` містить модель, підігнану функцією `lm()`, то стандартно вибирається F-відношення Фішера.

Замість того, щоб задавати значення $\mathbf{c}$ і $\mathbf{L}$ у параметрах `rhs` і `hypothesis.matrix`, можна у `hypothesis.matrix` записати рівняння обмежень у символьному вигляді. Як це робиться, показано у наступних прикладах.

Приклад 10.4.1. У прикладі 10.1.1 ми розглянули дані з файла `actors.csv` про вагу та зріст акторів. Нашою метою було визначити, яка залежність ваги від зросту найкраще відповідає уявленням про "людину гарної статури". В інтернеті можна знайти різні формули для такої залежності. Як приклад візьму дві: формулу Брока та індекс Кетле.

Формула Брока дуже проста: для того, щоб визначити IW — ідеальну вагу (у кілограмах), треба від зросту в сантиметрах (H) відняти 100 і помножити результат на 0.85 для жінок або на 0.9

для чоловіків:

$$IW = 0.9(H - 100) \text{ (для чоловіків).}$$

Враховуючи, що у наших даних вага (змінна `weight`) вимірюється у фунтах, вагу треба домножити на 2.2 (стільки фунтів у 1 кг). Отримуємо формулу Брока для прогнозу ідеальної ваги чоловіків у наших даних:

$$\widehat{\texttt{weight}} = 1.98\texttt{height} - 198.$$

Формула Кетле визначає "індекс ваги тіла", BMI[31]:

$$BMI = \frac{W}{h^2},$$

де W — вага людини у кілограмах, h — зріст у метрах. Стверджується, що нормальним є індекс Кетле у межах від 18.5 до 24.9. Для наших даних це дає таку формулу Кетле для прогнозу оптимальної ваги:

$$\widehat{\texttt{weight}} = 0.00022 \times K \times \texttt{height}^2,$$

де K — значення BMI, яке ми вважаємо оптимальним.

На рис. 10.19 зображена діаграма розсіювання зросту і ваги акторів-чоловіків з файла `actors.csv` та прямі, що відповідають прогнозам: зелена штрихова — формула Кетле з "середнім" індексом 22, червона штрихова — формула Брока для чоловіків, і дві лінії чорним пунктиром відповідають крайнім допустимим значенням BMI: нижня — 18.5, верхня — 24.9.

Синім кольором, суцільною лінією зображена пряма регресії, підігнана за даними методом найменших квадратів.

Як бачимо, "середній" прогноз Кетле практично збігається з прогнозом Брока[32]. Лінія регресії, підігнана за даними, помітно

[31] Body mass index, див. `en.wikipedia.org/wiki/Body_mass_index`.

[32] Як це може бути? Прогноз Кетле відповідає параболі, а формула Брока — прямій лінії! Справа у тім, що ми розглядаємо ці прогнози на такому вузенькому інтервалі, де кривина параболи майже непомітна. Це гарний приклад того, що багато нелінійних залежностей можна наближати лінійними, якщо обмежитись не дуже великими інтервалами зміни регресорів.

відрізняється від цього прогнозу. Чи свідчить це про те, що для голівудських акторів формулу Брока застосовувати не можна?

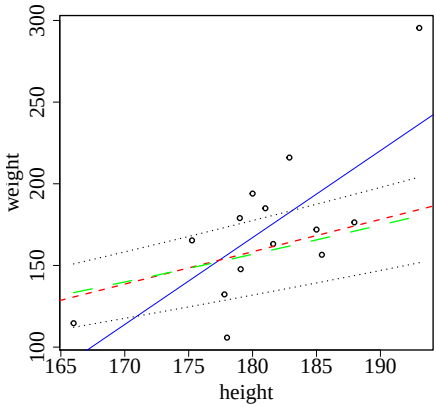

Рис. 10.19. Вага та зріст акторів-чоловіків з прогнозами за різними формулами

Даних у нас мало, тому обґрунтованість такого висновку сумнівна. Якщо придивитись до рис. 10.19, то можна побачити, що основна маса точок лежить навколо лінії прогнозу Брока, великі відхилення мають лише дві точки — крайня ліворуч і крайня праворуч. Саме вони повертають лінію МНК-регресії в бік від лінії Брока. Щоб переконатись, наскільки статистично значущим є цей ефект, застосуємо тест Фішера.

Гіпотеза, яку ми перевіряємо, полягає в тому, що дані описуються лінійною регресійною формулою

$$\texttt{weight}_j = b_0 + b_1 \texttt{height}_j + \varepsilon_j \text{ (необмежена гіпотеза)},$$

причому

$$b_0 = -198,\ b_1 = 1.98 \text{ (обмежена гіпотеза — формула Брока)}.$$

Обмеженій гіпотезі відповідає одинична матриця $\mathbf{L}$ і вектор правих частин $\mathbf{c} = (-198, 1.98)$. Застосуємо функцію
`linearHypothesis()`:

```
> actors<-read.csv2("c:/rem/term/actors.csv",header=T)
> actrm<-actors[actors$gender=="m",]
> linearHypothesis(lm(weight~height,data=actrm),
+ hypothesis.matrix=diag(c(1,1)),rhs=c(-198,1.98))

Linear hypothesis test

Hypothesis:
(Intercept) = - 198
height = 1.98

Model 1: restricted model
Model 2: weight ~ height

  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     14 21067
2     12 13403  2    7663.8 3.4307 0.06632 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Статистика тесту Фішера дорівнює $F_{emp} = 3.4307$, досягнутий рівень значущості — $p = 0.06632$. Оскільки $p > \alpha = 0.05$ приймаємо основну гіпотезу: незважаючи на помітну відмінність лінії регресії від лінії Брока, значущих відхилень не виявлено. (Тобто спостережувані відхилення можуть пояснюватись випадковими коливаннями, а не є особливістю ваги і зросту акторів Голлівуду).

Читачі можуть самі аналогічно перевірити, наскільки формулу Брока (Кетле) можна застосовувати до голівудських актрис.

Виклик функції `linearHypothesis()` можна було оформити у стилі, більш схожому на звичайний запис обмежених гіпотез:

```
linearHypothesis(lm(weight~height,data=actrm),
      c("(Intercept) = - 198","height = 1.98").
```

Тут у другому параметрі ми передаємо функції вектор з символьних рядків. Кожен рядок описує одне з рівнянь, що визначають обмежену гіпотезу. У цих рівняннях коефіцієнти при регресорах позначаються тими самими символами, якими у моделі позначались самі регресори. Коефіцієнт b_0 позначається як `(Intercept)`, а коефіцієнт при `height`,— позначається як `height`. Результат виконання цього виклику буде таким самим, як і у попереднього (перевірте). ◄

Приклад 10.4.2. У цьому прикладі розглядається цілком реальна задача, але дані, які ми будемо аналізувати умовні.

Нехай досліджується вплив різних добрив на інтенсивність цвітіння деяких квіткових рослин. В ході експерименту під час одного поливу рослин у воду додають добрива, а потім протягом місяця підраховують, скільки квіток виросло на рослині. Добрива містять сполуки фософору (P) та азоту (N). Перевіряються три типи добрив:

A — містить 100 діючих одиниць P у одній порції добрива;

B — містить 100 діючих одиниць N у одній порції;

C — містить 50 одиниць P і 50 одиниць N у одній порції.

Фірма, що випускає добриво типу C, стверджує, що, завдяки унікальній синергічній формулі, її продукт забезпечує значно вищу інтенсивність цвітіння, ніж при внесенні еквівалентної кількості P та N окремо. Нашою метою є перевірити це твердження.

Як звичайно (див. підрозд. 9.1) як основну — H_0 ми беремо гіпотезу, протилежну тій, яку перевіряємо, тобто вважаємо, що P та N у складі добрива C впливають на інтенсивність цвітіння так само, як у складі інших добрив.

Оскільки ми плануємо досліджувати вплив добрив лише у вузькому діапазоні можливих концентрацій, то можна сподіватись, що залежність відгуку (кількості квітів) від регресорів (концентрацій добрив) буде більш-менш добре описуватись лінійною моделлю.

Отже, будемо виходити з гіпотези H_0: середня інтенсивність

цвітіння залежить лінійно від кількості одиниць P та N, внесених у воду при поливі:

$$Y = b_0 + a_P P + a_N N + \varepsilon,$$

де Y — кількість квіток, що виросли на рослині протягом місяця після поливу добривами;

P — кількість фосфору, внесеного при поливі (у сотнях одиниць);

N — кількість азоту, внесеного при поливі (у сотнях однинць);

b_0 — середня базова інтенсивність цвітіння (при відсутності добрив);

a_P (a_P) — коефіцієнт, що вказує, наскільки зростає інтенсивність цвітіння при збільшенні вмісту фосфору (азоту) у воді на 100 одиниць.

Нехай для даної рослини при поливі було використано A порцій добрива A, B порцій B і C порцій C. Тоді $P = A + 0.5C$, $N = B + 0.5C$, отже,

$$Y = b_0 + a_P A + a_N B + 0.5(a_P + a_N)C + \varepsilon \text{ (при виконанні } H_0).$$

Якщо H_0 не виконується, то кожне добриво вносить свій внесок у інтенсивність цвітіння (необмежена модель):

$$Y = b_0 + b_A A + b_B B + b_C C + \varepsilon \text{ (у необмеженій моделі).}$$

Тут коефіцієнти загалом можуть бути будь-якими.

Зрозуміло, що регресійна формула для H_0 є частковим випадком загальної формули з такими коефіцієнтами:

$$b_A = a_A,\ b_B = a_B,\ b_C = 0.5(a_A + a_B).$$

Таким чином, гіпотеза H_0 відповідає такому обмеженню для коефіцієнтів необмеженої моделі:

$$b_A + b_B - 2b_C = 0.$$

Результати досліду з 15-ма рослинами (умовні) наведені у таб. 10.2

Таблиця 10.2. Результати досліду по впливу добрив на інтенсивність цвітіння квітів

N	1	2	3	4	5	6	7	8	9	10
A	0	1	0	0	2	0	0	1	1	0
B	0	0	1	0	0	2	0	1	0	1
C	0	0	0	1	0	0	2	0	1	1
Y	0	12	19	16	25	37	37	33	28	39

У дев'яти дослідах рослини поливали використовуючи по дві порції добрив у кожному досліді. При цьому перебирались всі можливі комбінації добрив. Крім того, перший експеримент був контрольним, тобто взагалі без добрив.

Проведемо підгонку необмеженої моделі використовуючи функцію `lm()`:

```
> library(car)
> A<-c(0,1,0,0,2,0,0,1,1,0)
> B<-c(0,0,1,0,0,2,0,1,0,1)
> C<-c(0,0,0,1,0,0,2,0,1,1)
> Y<-c(0,12,19,16,25,37,37,33,28,39)
> summary(lm(Y~A+B+C))

Call:
lm(formula = Y ~ A + B + C)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2667 -1.1833  0.2333  0.9833  1.7333

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7333     1.3064  -0.561    0.595
A            12.6667     0.9737  13.008 1.27e-05 ***
B            19.6667     0.9737  20.197 9.57e-07 ***
C            18.3333     0.9737  18.828 1.45e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.687 on
6 degrees of freedom
Multiple R-squared:  0.9884,
Adjusted R-squared:  0.9825
F-statistic: 169.8 on 3 and 6 DF,
p-value: 3.433e-06
```

△

Ми отримали такі оцінки для коефіцієнтів необмеженої моделі: $\hat{b}_A = 12.6667$, $\hat{b}_B = 19.6667$, $\hat{b}_C = 18.3333$. Величина $\hat{b}_A + \hat{b}_B - 2\hat{b}_C = -4.3332$ — помітно відрізняється від 0. Можна сказати, що у наших експериментах дві порції добрива C дали в середньому на 4 квітки більше, ніж порція A + порція B, тобто твердження виробників C статистично підтвердилось.

Але ми зробили дуже мало дослідів — всього 10. Чи є наш результат статистично значущим? Щоб перевірити це, проведемо перевірку гіпотези H_0, використовуючи тест Фішера:

▽

```
> linearHypothesis(lm(Y~A+B+C),"A+B-2*C=0")

Linear hypothesis test
Hypothesis:
A  + B - 2 C = 0

Model 1: restricted model
Model 2: Y ~ A + B + C
```

```
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1      7 35.844
2      6 17.067  1    18.778 6.6016 0.04237 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

△

Отримали досягнутий рівень значущості 0.04237. При стандартному $\alpha = 0.05$, слід прийняти альтернативу: відхилення від основної гіпотези значуще. Спостережуваний ефект (перевага С над А і В) настільки помітний, що його не можна ігнорувати, послався на випадкові коливання. ◀

Перевірка розшарування. Тест Чоу. У підрозд. 9.3.3 і 9.5 ми вже розглядали дво- та багатовибіркові задачі перевірки гіпотез. У таких задачах статистик має дані про об'єкти, які належать кільком різним популяціям, причому розподіл спостережуваних характеристик може бути різним у об'єктів з різних популяцій. Треба перевірити, чи дійсно ці розподіли є різними, чи навпаки, дані є однорідними (див. приклади 9.3.3, 9.5.2, 9.5.4). Тепер з'ясуємо, як такі перевірки можна проводити у контексті регресійного аналізу.

Нехай маємо дані про n об'єктів, причому для j-го об'єкта спостерігаються змінні $Y_j, X_j^1, \ldots, X_j^d$. Крім того, ми знаємо, що кожен з об'єктів належить одному з K класів (груп, популяцій). Номер класу, якому належить j-й об'єкт позначимо κ_j.

Будемо вважати, що залежність між змінною Y та змінними X^1,...,X^d описується моделлю лінійної регресії, в якій коефіцієнти можуть залежати від того, до якого класу належить об'єкт:

$$Y_j = \sum_{i=1}^{d} b_i^{\kappa_j} X_j^i + \varepsilon_j, \tag{10.14}$$

де b_i^k, $i = 1, \ldots, d$, $k = 1, \ldots, K$ — невідомі коефіцієнти регресії;

ε_j — похибки регресії, які ми будемо вважати незалежними при різних j гауссовими випадковими величинами з нульовим математичним сподіванням і дисперсією σ^2 (невідомою).

Таким чином, коефіцієнти регресії можуть бути різними для різних класів, а дисперсія похибок (у цій моделі) є однаковою для всіх спостережень.

Можна вважати, що дані складаються з K різних вибірок, причому кожна вибірка містить лише об'єкти з одного класу. А можна трактувати це трохи по-іншому. Номери класів κ_j можна розглядати як іще одну змінну, що характеризує об'єкт. Ми будемо називати її змінною розшарування: вона розділяє всю вибірку на окремі прошарки-класи. Якщо коефіцієнти регресії у (10.14) дійсно залежать від номера класу, вибірку будемо називати розшарованою за змінною (ознакою) κ. Якщо $b_i^1 = b_i^2 = \cdots = b_i^K$ для всіх $i = 1, \ldots, d$, будемо казати, що вибірка є однорідною (розшарування за κ немає).

У прикладних задачах κ не обов'язково має бути номером (тобто числом). Це може бути будь-яка назва класу. В R для задання розшарування природно використовувати змінні типу фактор.

Модель (10.14) можна записати у вигляді звичайної лінійної регресійної моделі, якщо ввести формальні регресори $X_j^{ik} = X_j^i \mathbb{1}\{\kappa_j = k\}$:

$$Y_j = \sum_{i=1}^{d} \sum_{k=1}^{K} b_i^k X_j^{ik} + \varepsilon_j. \tag{10.15}$$

Це дозволяє використати загальний тест Фішера для перевірки гіпотези про однорідність вибірки:

$$H_0 : \; b_i^1 = b_i^2 = \cdots = b_i^K \text{ для всіх } k = 1, \ldots, K, \; i = 1, \ldots, d,$$

проти альтернативи — наявність розшарування за змінною κ:

$$H_1 : \; \text{існують такі } i, k_1 \text{ та } k_2, \text{ що } b_i^{k_1} \neq b_i^{k_2}.$$

Гіпотеза H_1 є, вочевидь, лінійною гіпотезою для необмеженої моделі (10.15). Використовуючи загальну схему тесту Фішера, можна визначити

$$SSR = J^{all}(\hat{\mathbf{b}}^{all}),$$

де

$$J^{all}(\mathbf{b}) = \sum_{j=1}^{n}(Y_j - \sum_{i=1}^{d} b_i X_j^i)^2$$

— функціонал МНК у обмеженій моделі (тут коефіцієнти регресії однакові для всіх спостережень),

$$\mathbf{b}^{all} = \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^d} J^{all}(\mathbf{b})$$

— оцінки МНК для коефіцієнтів у обмеженій моделі. Тобто для отримання SSR треба підігнати лінійну модель за всіма наявними даними ігноруючи змінну розшарування і обчислити суму квадратів залишків у підігнаній моделі.

Далі, для необмеженої моделі, функціонал МНК можна записати як

$$J(\mathbf{b}) = \sum_{k=1}^{K} J^k(\mathbf{b}^k),$$

де

$$J^k(\mathbf{b}^k) = \sum_{j:\ \kappa_j=k} (Y_j - \sum_{i=1}^{d} b_i^k X_j^i)^2$$

— сума квадратів для спостережень, що належать k-му класу.

Відповідно, щоб знайти оцінки коефіцієнтів у необмеженій моделі, треба підігнати лінійну регресію окремо по кожній вибірці об'єктів k-го класу для $k = 1, \ldots, K$:

$$\hat{\mathbf{b}}^k = \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^d} J^k(\mathbf{b}).$$

Тепер

$$SSU = \sum_{k=1}^{K} J^k(\hat{\mathbf{b}}^k)$$

— суму квадратів залишків у необмеженій моделі можна отримати підсумуванням сум квадратів залишків моделей, підігнаних для кожного класу окремо.

Нарешті, можна скласти відношення Фішера, враховуючи, що кількість обмежень d, а кількість коефіцієнтів, що оцінюються у необмеженій моделі — Kd:

$$F_{emp} = \frac{(SSR - SSU)/d}{SSU/(n - Kd)}.$$

Відповідно,

$$F_{theor} = Q^{F(d,n-Kd)}(1 - \alpha).$$

Сформулюємо правило прийняття рішення **тестом Фішера для перевірки розшарування**:

> Якщо $F_{emp} \leq F_{theor}$ — вважати вибірку однорідною,
> $F_{emp} > F_{theor}$ — вважати вибірку розшарованою.

У випадку двох класів ($K = 2$) і простої лінійної регресії як моделі для кожного класу, цей тест називають **тестом Чоу**.

Приклад 10.4.3. У прикладі 5.1.1 ми розглядали дані з фрейма `faithful` про поведінку гейзера Old Faithful. У даних дві змінні — `eruptions` і `waiting`. На діаграмі розсіювання (рис. 5.1) ми побачили дві групи-кластера, на які досить чітко розділились спостереження за значеннями цих змінних. На цій діаграмі проведена лінія регресії, підігнаної за моделлю

$$\texttt{waiting}_j = b_0 + b_1 \texttt{eruptions}_j + \varepsilon_j.$$

За рисунком можна припустити, що залежність між `eruptions` і `waiting` в обох кластерах описується однією спільною формулою.

Спробуємо зробити підгонку по кожному кластеру окремо. Для цього спочатку утворимо нову змінну `clust`, яка буде набувати значень `FALSE`, якщо спостереження потрапляє до нижнього лівого кластера, і `TRUE`, якщо до правого верхнього. (Належність до правого кластера визначається за умовою `eruptions > 3.2`):

```
> # кольори для відображення кластерів:
> colpr<-c("blue","red")
> # символи для відображення кластерів:
> symb<-c(1,2)
> # створюємо змінну розшарування:
> clust<-faithful$eruptions>3.2
> plot(faithful$waiting~faithful$eruptions,
+      col=colpr[clust+1],pch=symb[clust+1])
> # підгонка обмеженої моделі:
> modelall<-lm(waiting~eruptions,data=faithful)
> abline(modelall)
> # модель для верхнього кластера:
> modelhigh<-lm(waiting~eruptions,
+               data=faithful[!clust,])
> abline(modelhigh,col=colpr[1],lty="dotted")
> # модель для нижнього кластера:
> modellow<-lm(waiting~eruptions,
+              data=faithful[clust,])
> abline(modellow,col=colpr[2],lty="dashed")
```

На рис. 10.20 нижній кластер відображені синіми кружечками, верхній — червоними трикутниками. Лінія регресії для нижнього кластера — синя пунктирна, для верхнього — червона штрихова. Лінія регресії підігнана за обома кластерами одразу — чорна суцільна.

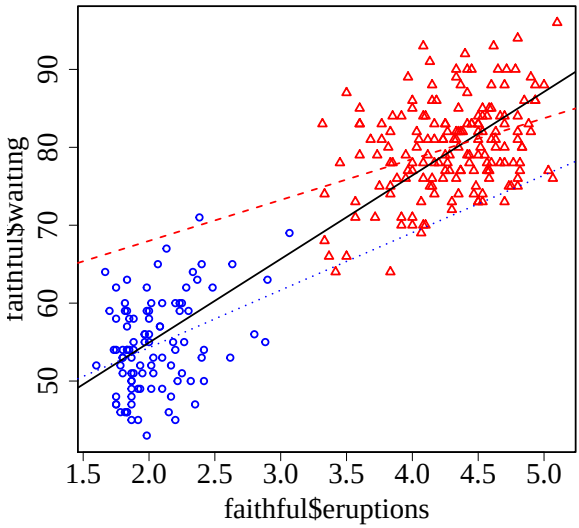

Рис. 10.20. Перевірка розшарування даних для гейзера Old Faithful

Як бачимо, лінії регресії, що відповідають окремим кластерам, помітно відрізняються одна від одної та від чорної прямої. Перевіримо чи є значущими ці відмінності, використовуючи тест Чоу:

```
> # Сума квадратів для обмеженої моделі:
> RestrSS<-sum((modelall$residuals)^2)
> # Сума квадратів для необмеженої моделі:
> UnrestSS<-sum((modellow$residuals)^2)+
+   sum((modelhigh$residuals)^2)
> # F емпіричне:
> Fchou<-(nrow(faithful)-4)*
+   (RestrSS-UnrestSS)/(2*UnrestSS)
> # Досягнутий рівень значущості:
> 1-pf(Fchou,2,nrow(faithful))
```

```
[1] 1.967714e-07
```

△

Отримали досягнутий рівень значущості 1.967714×10^{-07}, отже, гіпотезу про однорідність вибірки слід відхилити — розшарування є. ◀

10.5 Нелінійний МНК

У підрозд. 10.2 ми, зустрічаючись із нелінійними залежностями, намагались перетворити дані так, щоб з ними можна було працювати на основі лінійних моделей. Інколи це доцільно (саме такі приклади ми і розглянули). Часто краще працювати безпосередньо з нелінійними моделями.

Нехай треба підігнати регресійну модель вигляду

$$Y_j = g(X_j^1, \ldots, X_j^m; \mathbf{b}) + \varepsilon_j$$

за спостереженнями $Y_j, X_j^1, \ldots, X_j^m$, $j = 1, \ldots, n$, де

g — функція регресії, відома з точністю до вектора невідомих коефіцієнтів $\mathbf{b} \in \mathbb{R}^d$, ε_j — випадкові похибки регресії. Для цього можна використати функцію `nls()` (скорочення від nonlinear least squares — нелінійний МНК). Ця функція організована подібно до `lm()`, її можна викликати як

`nls(formula,data,start)`,

де `data` — фрейм даних, що використовується для підгонки,

`formula` — формула, яка задає специфікацію моделі,

`start` — набір початкових значень невідомих параметрів, які будуть уточнюватись підгонкою.

Розберемо детальніше параметри `formula` і `start`. Це зручно зробити, розглядаючи конкретні приклади.

Приклад 10.5.1. На відміну від формул для лінійних моделей, тепер нам треба явно вказувати вигляд функції регресії і невідомі параметри. Наприклад, специфікацію моделі

$$y_j = C \exp(a x_j) + \varepsilon_j \tag{10.16}$$

можна задати формулою:

`y~C1*exp(a1*x)`.

При цьому `x` та `y` повинні бути завантаженими у пам'ять векторами однакової довжини, а `C1` і `a1` — назвами змінних, які раніше не були визначені. (Саме завдяки цьому комп'ютер здогадується, які змінні позначають дані для підгонки, а які — невідомі параметри, що підлягають оцінці).

Відповідний виклик функції `nls()` можна оформити так:

```
> # генеруємо дані:
> set.seed(3)
> n<-100
> x<-runif(n,min=0,max=4)
> C0<-2.5 # справжнє C у модельованих даних
> a0<-0.5 # справжнє a у модельованих даних
> y<-C0*exp(a0*x)+rnorm(n)
> # виводимо діаграму розсіювання даних:
> plot(x,y)
> # підганяємо модель
> model<-nls(y~C1*exp(a1*x))
> # Записуємо оцінки коефіцієнтів:
> C<-coef(model)[1]
> a<-coef(model)[2]
> C; a # значення оцінок:

      C1
2.491098

       a1
0.5004129

> # рисуємо криву регресії:
> curve(C*exp(a*x),add=T,col="red")
```

Тут ми генеруємо дані, що відповідають моделі (10.16), з $a = 0.5$, $C = 2.5$ і похибками ε_j, що мають стандартний нормальний розподіл. Потім дані підганяються з використанням функції `nls()`. При цьому параметри `start` і `data` ми не вказували. Тому комп'ютер взяв ті `y` і `x`, які знайшов у робочій області, а все, що не знайшов, трактував як невідомі коефіцієнти, які треба підігнати.

В результаті підгонки отримали оцінки $\hat{C} = 2.491$ і $\hat{a} = 0.5004$, які є дуже близькими до справжніх значень параметрів, використаних при моделюванні. Діаграма розсіювання з підігнаною лінією регресії зображена на рис. 10.21.

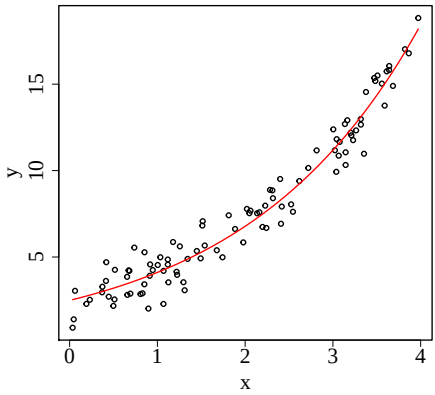

Рис. 10.21. Експоненційна регресійна модель

◀

Для того, щоб задати початкові наближення у параметрі `start`, їх треба вмістити в іменований список, в якому імена мають відповідати іменам параметрів, що підганяються. Скажімо, у прикладі 10.5.1 задати початкові наближення можна так[33]:

[33]Спробуйте це зробити і переконайтесь, що у цьому випадку отримується майже той самий результат, що і без задання початкових значень. Але так буває не завжди.

`nls(y~C1*exp(a1*x),start=list(a1=-1,C1=10)).`

При цьому комп'ютер вважатиме невідомими коефіцієнтами лише ті, для яких задано початкові значення. І тільки по них буде проводити підгонку. Якщо якийсь параметр пропустити у `start`, комп'ютер повідомить про помилку і шукати оцінки не буде.

Приклад 10.5.2. Для перевірки якості металевих виробів використовують ультразвукові локатори. Перед роботою їх треба калібрувати. Для цього виробляють спеціальні циліндричні алюмінієві тестові блоки. В нижній основі блоку висвердлюють заглиблення-порожнину. Сонар (ультразвуковий локатор) опромінює блок зверху. Товщина шару металу між дном заглиблення і верхньою поверхнею блоку називається *metal distance* (md) — це відстань, яку звук має пройти всередині металу, щоб досягти порожнини. Відбитий від дна порожнини звук утворює сигнал, що вловлюється локатором. Зафіксована локатором потужність сигналу називаэться ультразвуковим відгуком (*usr*).

Треба встановити відповідність між usr та md[34]. Модель цієї залежності, запропонована у стандарті США —

$$\text{usr} \approx \frac{\exp(-\beta_1 \text{md})}{\beta_2 + \beta_3 \text{md}}.$$

Тут β_1, β_2 і β_3 — невідомі коефіцієнти, які треба визначати при калібруванні сонара. Дані вимірювань, отримані у ході такого калібрування, містяться у наборі даних `Chwirut2` бібліотеки `NISTnls`. У змінній `Chwirut2$x` містяться значення md для блоків, що використовувались у дослідженні, а у змінній `Chwirut2$y` — відповідні значення usr, отримані при скануванні сонаром. Відповідна діаграма розсіювання зображена на рис. 10.22.

Якщо викликати `nls()`, не вказуючи початкові значення невідомих параметрів, наприклад, так:

[34] Цей приклад застосування техніки нелінійної регресії запозичений з книги [45]. Дані, прикладна задача і регресійна модель описані у національному стандарті США E 127 – 98 (1979 р.), присвяченому перевірці якості виробів з алюмінію за допомогою ультразвукової локації.

```
nls(y ~ exp(-beta1 * x)/(beta2 + beta3 *x),
    data = Chwirut2),
```

отримаємо повідомлення про те, що алгоритм наближеного пошуку точки мінімуму не зміг її знайти[35]. Так часто буває, якщо задавати початкові наближення навмання або доручити їх вибір комп'ютеру. Тому корисно вміти підбирати початкові наближення, виходячи з яких-небудь змістовних міркувань. Спробуємо зробити це для наших даних.

На діаграмі розсіювання поведінка даних нагадує експоненційно спадаючу функцію. Тому природно $\exp(-\beta_1 x)$ трактувати як головну частину функції регресії, а $1/(\beta_2+\beta_3 x)$ — як поправку до неї. Регресію $y \approx \exp(-\beta_1 x)$ легко лінеаризувати, прологарифмувавши відгук. Зробимо це і застосуємо лінійний МНК для оцінки β_1:

```
> library( NISTnls)
> lm(formula = log(y) ~ x, data = Chwirut2)


Call:
lm(formula = log(y) ~ x, data = Chwirut2)

Coefficients:
(Intercept)            x
     4.3698      -0.5051
```

Отримали оцінку для β_1 — β_1^{start} =0.5051. Щоб оцінити грубо β_2 і β_3, помітимо, що на діаграмі розсіювання значенню x= 1 відповідають точки, з координатою y, близькою до 60. Отже,

$$\frac{\exp(-0.5051 \times 1)}{\beta_2 + \beta_3 \times 1} \approx 60,$$

[35] Спробуйте!

звідки $\beta_2 + \beta_3 \approx \exp(-0.5051)/60 = 0.01005742$. Візьмемо початкові наближення $\beta_2^{start} = \beta_3^{start} = 0.005$.

Тепер можна знову спробувати викликати `nls()` з цими початковими наближеннями:

```
> # підключаємо бібліотеку з даними:
> library(NISTnls)
> # виводимо діаграму розсіювання даних:
> plot(y ~ x, data = Chwirut2,
+      xlab = "Metal distance",
+            ylab = "Ultrasonic response")
> # робимо підгонку нелінійним МНК:
> Chwirut2.m1 <- nls(y ~ exp(-beta1 * x)/
+                      (beta2 + beta3 *x),
+                 data = Chwirut2,
+                 start = list(beta1 = 0.5051,
+                              beta2 = 0.005,
+                              beta3 = 0.005)
+                                  )
> # результати підгонки:
> summary(Chwirut2.m1)


Formula: y ~ exp(-beta1 * x)/(beta2 + beta3 * x)

Parameters:
       Estimate Std. Error t value Pr(>|t|)
beta1 0.1665765  0.0383033   4.349 6.56e-05 ***
beta2 0.0051653  0.0006662   7.753 3.54e-10 ***
beta3 0.0121500  0.0015304   7.939 1.81e-10 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.172 on 51 degrees
of freedom

Number of iterations to convergence: 6
Achieved convergence tolerance: 6.706e-07
```

```
> # записуємо оцінки коефіцієнтів у вектор b:
> b<-coef(Chwirut2.m1)
> # рисуємо криву регресії на діаграмі розсіювання:
> curve(exp(-b[1] * x)/(b[2] + b[3] *x),
+ add = TRUE, col = "red")
```

△

Крива регресії на діаграмі розсіювання на рис. 10.22 має адекватний вигляд.

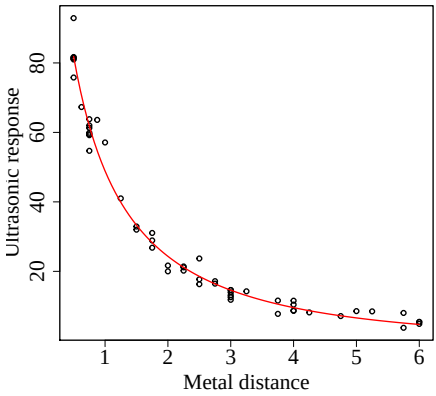

Рис. 10.22. Дані для калібрування сонара

У скрипті `coef()` — функція, яка при застосуванні до результату роботи `nls()` виділяє з нього оцінки коефіцієнтів регресії. Функція `summary()` за цими результатами друкує розширений звіт, подібний до того, який ми розглядали для функції `lm()` у прикладі 10.2.1.

В результаті застосування `nls()` отримали оцінки коефіцієнтів $\hat{\beta}_1 = 0.1665765$, $\hat{\beta}_2 = 0.0051653$, $\hat{\beta}_3 = 0.0121500$. Бачимо, що $\hat{\beta}_2$ виявилось схожим на наше початкове наближення, інші дві оцінки помітно змінились. Тест для перевірки значущості визнав всі коефіцієнти значущо відмінними від 0.

Для кращої перевірки якості підгонки доцільно також провести графічний аналіз залишків регресії. Це можна зробити, використовуючи, наприклад, такий скрипт:

```
> resid<-residuals(Chwirut2.m1) # виділення залишків
> pred<-predict(Chwirut2.m1) # виділення прогнозів
> # діаграма розсіювання прогноз-залишки:
> plot(resid~pred)
> # нормальна Q-Q діаграма залишків:
> qqnorm(resid)
> qqline(resid,col="red")
```

Результати зображені на рис. 10.23. Зверху — діаграма розсіювання прогноз-залишки для даної нелінійної моделі. З неї видно, що основна маса точок розкидана хаотично, яких-небудь неврахованих нами закономірностей, що можна використати для прогнозування, непомітно. У правій верхній частині діаграми є три точки, які можна інтерпретувати як невиразні викиди.

Знизу — Q-Q діаграма для перевірки нормальності похибок. Помітно, що основна маса точок добре вкладається на пряму лінію, що відповідає нормальному розподілу, але знову виділяються три викиди праворуч угорі. Можливо, доцільно розібратись з тим, чому у цих експериментах спостерігались такі помітні відхилення, але вони могли виникнути внаслідок випадкових коливань. ◄

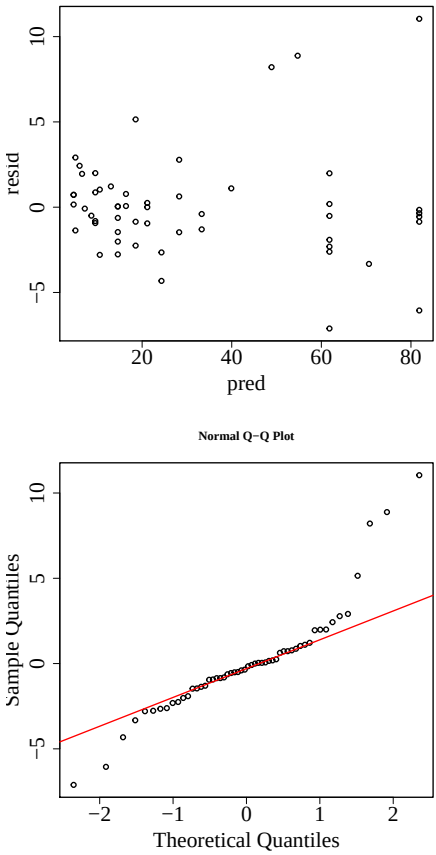

Рис. 10.23. Залишки калібрування сонара

Література

1. Василик О. І. Лекції з теорії і методів вибіркових обстежень / О. І. Василик, Т. О. Яковенко. – Київ : ВПЦ "Київський університет", 2010. – 208 с.

2. Боровков А. А. Теория вероятностей / А. А. Боровков. – Москва : Наука, 1986. – 432 с.

3. Боровков А. А. Математическая статистика / А. А. Боровков. – Москва : Наука, 1997. – 774 с.

4. Верещагин Н. К. Колмогоровская сложность и алгоритмическая случайность / Н. К. Верещагин, В. А. Успенский, А. А. Шень. – Москва : МЦНМО, 2013. – 576 с.

5. Гихман И. И. Теория вероятностей и математическая статистика / И. И. Гихман, А. В. Скороход, М. И. Ядренко. – Киев : Вища шк., 1979. – 408 с.

6. Дороговцев А. Я. Элементы общей теории меры и интеграла / А. Я. Дороговцев. – Киев : Факт, 2007. – 164 с.

7. Зализняк А. А. "Слово о полку Игореве": взгляд лингвиста / А. А. Зализняк. 3-е изд., доп. – Москва : "Рукописные памятники Древней Руси", 2006. – 480 с.

8. Зімін А. А. Слово о полку Игореве / А. А. Зімін. – Санкт-Петербург : "Дмитрий Буланин", 2006. – 516 с.

9. Карташов М. В. Імовірність, процеси, статистика / М. В. Карташов. – Київ : ВПЦ "Київський університет", 2007.

– 494 с.

10. Кнут Д. Э. Искусство программирования. Т. 2. Получисленные методы / Д. Э. Кнут. – Москва : Вильямс, 2001. – 824 с.

11. Майборода Р. Є. Регресія: лінійні моделі / Р. Є. Майборода. – Київ : ВПЦ "Київський університет", 2007. – 296 с.

12. Майборода Р. Є. Аналіз даних за допомогою пакета R 6 : навч. посібник : [Електронний ресурс] / Р. Є. Майборода, О. В. Сугакова. – Київ , 2015. – 65 с. Режим доступу : http://matphys.rpd.univ.kiev.ua/downloads/courses/mmatstat/Statistics_with_R.pdf.

13. Моделі регресії з похибками вимірювання та їх застосування до оцінювання радіаційних ризиків / С. В. Масюк, О. Г. Кукуш, С. В. Шкляр та ін. – Київ : ДІА, 2015. – 288 с.

14. Новиков С. П. Воспоминания об А.Н. Колмогорове / С. П. Новиков // УМН, 1988. – Т. 43. – Вып. 6(264). – С. 35-36.

15. Оленко А. Я. Комп'ютерна статистика : навч. посібник / А. Я. Оленко. – Київ : ВПЦ "Київський університет", 2007. – 174 с.

16. Радченко В. М. Теорія міри та інтеграла : навч. посібник. – Київ : ВПЦ "Київський університет 2012. – 144 с.

17. Себер Дж. Линейный регрессионный анализ / Дж. Себер. – Москва : Мир, 1980. – 456 с.

18. Турчин В. М. Теорія ймовірностей і математична статистика / В. М. Турчин. – Дніпропетровськ : ІМА-пресс, 2014. – 566 с.

19. Харченко М. А. Корреляционный анализ / М. А. Харченко. – Воронеж : ИПЦ Воронежского государственного университета, 2008. – 31 с.

20. Abedin J. Data Manipulation with R / J. Abedin. – Birmingham, UK : Packt Publishing, 2014. – 103 p.

21. Bloomfield Victor A. Using R for Numerical Analysis in Science and Engineering / Victor A. Bloomfield. – London : Chapman and Hall/CRC, 2014. – 359 p.

22. Cheng C.-L. Statistical Regression with Measurement Error: Ken-dall's Library of Statistics 6 / C.-L. Cheng, J. Van Ness. – London : Arnold, 1999. – 262 p.

23. Clarke R. D. An application of the Poisson distribution / R. D. Clarke // J. of the Institute of Actuaries. – Cambridge, 1946. – Vol. 72, № 3. – P. 481.

24. Cryer J. D. Time Series Analysis: With Applications in R / J. D. Cryer, K.-S. Chan. – New York : Springer, 2008. – 491 p.

25. Dalgaard P. Introductory Statistics With R / P. Dalgaard. – New York : Springer, 2008. – 364 p.

26. Davies P. L. Data Analysis and Approximate Models: Model Choice, Location-Scale, Analysis of Variance, Nonparametric Regression and Image Analysis / P. L. Davies. – London : Chapman and Hall/CRC, 2014. – 320 p.

27. Encyclopedia of Statistical Sciences / Ed. by Samuel Kotz et al. – Hoboken N.J. : Wiley & Sons, 2006. – Vol. 12. – 546 p.

28. Fruchterman T. M. J. Graph Drawing by Force-Directed Placement / T. M. J. Fruchterman, E. M. Reingold // Software: Practice & Ex-perience, 1991. – Vol. 21, № 11. – P. 1129–1164.

29. Gladwell M. Outliers: The Story of Success / M. Gladwell. – New York : Little, Brown and Co, 2008. – 304 p.

30. Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations / S. Greenland, S. J. Senn, K. J. Rothman et al. // Eur. J. Epidemiol., 2016. – Vol. 31. – № 4. – P. 337–350.

31. Hastie T. The Elements of Statistical Learning Data Mining, Infer-ence and Prediction / T. Hastie, R. Tibshirani, J. Friedman. – New York : Springer, 2013. – 764 p.

32. Ingraham C. Map: Real American men don't wear cargo shorts : [Electronic resource] : / C. Ingraham. – Access mode : www.washingtonpost.com/news/wonk/wp/2016/08/05/map-real-americans-wear-jorts/

33. An Introduction to Statistical Learning with Applications in R / G. James, D. Witten, T. Hastie, Robert Tibshirani. – New York : Springer, 2013. – 440 p.

34. Kahneman D. Thinking, Fast and Slow / D. Kahneman. – New York : Farrar, Straus and Giroux, 2011. – 499 p.

35. Keenan E. L. Josef Dobrovsky and the Origins of the Igor' Tale / E. L. Keenan. – Harward : Harvard University Press, 2003. – 541 p.

36. Koenker R. Quantile Regression / R. Koenker. – New York : Cam-bridge University Press, 2005. – 198 p.

37. Lawless J. F. Statistical Models and Methods for Lifetime Data / J. F. Lawless. – 2nd ed. – New York : John Wiley and Sons, 2003. – 580 p.

38. Liu J. Equity valuation using multiples / J. Liu, D. Nissim, J. Thomas // J. of Accounting Research, 2002. – Vol. 40, № 1. – P. 135–172.

39. Makino J. Lagged-Fibonacci random number generator on parallel computers / J. Makino // Parallel Computing, 1994. Vol. 20, Issue 9. – P. 1357–1367.

40. Mazon A. Le Slovo d'Igor / A. Mazon. – Paris : Droz, 1940. – 182 p.

41. Moser M. Sind der "Relativisator" mo und die Syntax anderer Enkli-tika als klare Beweise f?r die Authentizit?t des Igorlieds zu werten? // Studia Slavica, 2005. – Vol. 50, issue 3-4. – S. 267–282.

42. Nuzzo R. Scientific Method: Statistical Errors / R. Nuzzo // Nature, 2014. – Vol. 506. – P. 150–152.

43. Numerical Recipes in C: The Art of Scientific Computing / W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery. – New York : Cambridge University Press, 1992. – 949 p.

44. Ramsey F. The Statistical Sleuth: a Course in Methods of Data Analysis / F. Ramsey, D. W. Schafer. – Boston : Brooks/Cole, 2013. – 760 p.

45. Ritz C. Nonlinear Regression with R / C. Ritz, J. C. Streibig. – New York : Springer, 2008. – 144 p.

46. Shao J. Mathematical statistics / J. Shao. – New York : Springer, 1998. – 530 p.

47. Shumway R. H. Time Series Analysis and Its Applications with R Examples / R. H. Shumway, D. S. Stoffer. – New York : Springer, 2010. – 589 p.

48. Spector P. Data Manipulation with R / P. Spector. – New York : Springer, 2008. – 157 p.

49. Ugarte M. D. Probability and Statistics with R / M. D. Ugarte, A. F. Militino, A. T. Arnholt. – New York : Chapmen & Hall, 2016. – 700 p.

50. Venables W. N. Modern Applied Statistics with S / W. N. Venables, B. D. Ripley. – New York : Springer, 2002. – 504 p.

51. Wasserman L. All of Statistics. A Concise Course in Statistical In-ference / L. Wasserman. – New York : Springer, 2004. – 446 p.

52. Wasserman L. All of Nonparametric Statistics / L. Wasserman. – New York : Springer, 2006. – 272 p. L. All of Nonparametric Statistics. New York: Springer, 2006. 272 p.

53. Wasserstein R. L. The ASA's Statement on p-Values: Context, Process, and Purpose / R. L. Wasserstein, N. A. Lazar // The American Statistician, 2016. – Vol. 70, № 2. – P. 129–133.

54. Wickham H. Advanced R / H. Wickham. – New York : Chapman & Hall/CRC, 2015. – 466 p.

Додаток А

Векторна і матрична алгебра

Матрицею розміру (вимірності) $m \times n$ називають прямокутну таблицю чисел з m рядочків та n стовпчиків:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

Інший запис: $\mathbf{A} = (a_{ij})_{i=1,\ldots,m,j=1,\ldots,n}$. Індекси елементів можуть бути як верхніми, так і нижніми, наприклад:

$$(a_i^j)_{i=1,\ldots,m,j=1,\ldots,n}.$$

Ми не вживаємо тензорних позначень, тому розташування індексів вгорі або внизу визначається лише графічною зручністю і ніяк не характеризує структуру об'єкта.

Якщо $m = n$ матрицю називають квадратною матрицею вимірності m. Матрицю $m \times 1$ звуть m-вимірним вектором-стовпчиком,

а матрицю $1 \times n$ — n-вимірним вектором рядком. Запис

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 & \cdots & y_n \end{pmatrix}.$$

У цій книзі, там, де не обумовлено інше, вектори вважаються векторами-стовпчиками. Множину всіх векторів вимірності n позначають $\mathbb{R}^n$. (У деяких книгах множина векторів рядків вимірності m позначається $\mathbb{R}_m$).

Числа a_{ij} називають елементами (для векторів — координатами) матриці $\mathbf{A}$. Матрицю $\mathbf{A}^\top = (a_{ji})_{j=1,\ldots,n,i=1,\ldots,m}$ називають транспонованою матрицею $\mathbf{A}$. Операції додавання матриць однакової вимірності та множення на число виконуються поелементно: якщо $\mathbf{B} = (b_{ij})_{i=1,\ldots,m,j=1,\ldots,n}$, то

$$\mathbf{A} + c\mathbf{B} = (a_{ij} + cb_{ij})_{i=1,\ldots,m,j=1,\ldots,n}.$$

Множення матриць

$$\mathbf{A} = (a_{ij})_{i=1,\ldots,I,j=1,\ldots,J} \text{ та } \mathbf{B} = (b_{kl})_{k=1,\ldots,K,l=1,\ldots,L}$$

можливе лише при $J = K$. Результат множення — матриця $\mathbf{AB} = \mathbf{C} = (c_{il})_{i=1,\ldots,I,l=1,\ldots,L}$, де $c_{il} = \sum_{j=1}^{J} a_{ij}b_{jl}$. Елементи a_{ii} матриці $\mathbf{A}$ називаються діагональними. Квадратну матрицю $\mathbf{A}$, в якої всі недіагональні елементи дорівнюють 0, називають діагональною і позначають $\mathbf{A} = \operatorname{diag}(a_{11}, \ldots, a_{mm})$. Одиничною називають матрицю $\mathbf{E} = \operatorname{diag}(1, 1, \ldots, 1)$, нульовою — матрицю (вектор) $\mathbf{0}$, всі елементи якої — нулі. Матрицю $\mathbf{B}$ називають оберненою до $\mathbf{A}$, якщо $\mathbf{BA} = \mathbf{AB} = \mathbf{E}$. Позначення $\mathbf{B} = \mathbf{A}^{-1}$. (Зрозуміло, що обернена може бути лише у квадратної матриці). Якщо квадратна матриця не має оберненої, її називають виродженою, інакше — невиродженою. Для невироджених матриць обернена визначена однозначно.

Легко бачити, що для будь-яких матриць $\mathbf{A}, \mathbf{B}$, $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$, а отже, $(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1}$.

Квадратна матриця, для якої $\mathbf{A}^\top = \mathbf{A}$ (тобто $a_{ij} = a_{ji}$), називається симметричною.

Узагальнена обернена матриця. Для довільної $m \times n$ матриці $\mathbf{B}$ узагальненою оберненою називається будь-яка матриця $\mathbf{B}^-$, що задовольняє рівність $\mathbf{B}\mathbf{B}^-\mathbf{B} = \mathbf{B}$. Така матриця існує завжди (пор. [17], с. 78). Якщо, крім того, $\mathbf{B}^-$ задовольняє умови $\mathbf{B}^-\mathbf{B}\mathbf{B}^- = \mathbf{B}^-$, $(\mathbf{B}\mathbf{B}^-)^\top = \mathbf{B}\mathbf{B}^-$, $(\mathbf{B}^-\mathbf{B})^\top = \mathbf{B}^-\mathbf{B}$, то вона називається псевдооберненою, або матрицею Мура — Пенроуза. Псевдообернена матриця визначена однозначно.

Скалярний добуток і довжина. Якщо $\mathbf{x} = (x_1, \ldots, x_n)^\top$, $\mathbf{y} = (y_1, \ldots, y_n)^\top$ — n-вимірні вектори стовпчики, то їх (евклідів) скалярний добуток визначається як

$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle_E = \mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x} = \sum_{j=1}^{N} x_j y_j.$$

Довжина (евклідова норма) вектора $\mathbf{x}$,

$$\|\mathbf{x}\|_E = \|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_E} = \sqrt{\sum_{j=1}^{n} (x_j)^2}.$$

Якщо $\mathbf{x} \neq \mathbf{0}$, $\mathbf{y} \neq \mathbf{0}$ і $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, вектори $\mathbf{x}$ і $\mathbf{y}$ називають ортогональними.

Лінійна залежність. Якщо $\mathcal{V} = \{\mathbf{v}^k,\ k = 1, \ldots, K\}$ — набір m-вимірних векторів, c_k — числа (скаляри), то $\sum_{k=1}^{K} c_k \mathbf{v}^k$ називається лінійною комбінацією векторів з $\mathcal{V}$. Якщо існують такі c_k, не всі одночасно дорівнюють 0, що $\sum_{k=1}^{K} c_k \mathbf{v}^k = \mathbf{0}$, то кажуть, що система $\mathcal{V}$ є лінійно залежною. Якщо таких c_k не існує, $\mathcal{V}$ звуть лінійно незалежною системою векторів.

Рангом $\mathrm{rank}(\mathbf{A})$ матриці $\mathbf{A}$ називають найбільшу кількість лінійно незалежних стовпчиків $\mathbf{A}$ ($\mathrm{rank}(\mathbf{A})$ дорівнює також найбільшій кількості лінійно незалежних рядків $\mathbf{A}$).

Визначник $\det \mathbf{A}$ квадратної матриці $\mathbf{A}$ можна задати рекурентною формулою:

— для 1×1-матриці (числа) a, $\det a = a$;

— для $n \times n$-матриці $\mathbf{A}$, $\det \mathbf{A} = \sum_{i=1}^{n} (-1)^{i+1} \det \mathbf{A}_{i1}$, де $\det \mathbf{A}_{i1}$ — $(n-1) \times (n-1)$-матриця, утворена викреслюванням i-го рядка та 1-го стовпчика з $\mathbf{A}$.

Для $m \times m$-матриці $\mathbf{A}$, $\det \mathbf{A} \neq 0$ еквівалентно $\mathrm{rank}\mathbf{A} = m$ та існуванню оберненої матриці $\mathbf{A}^{-1}$.

Матриці та системи лінійних рівнянь. Систему лінійних рівнянь відносно невідомих змінних $x_1, \ldots, x_n$ вигляду

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \cdots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{cases} \tag{A.1}$$

можна записати у матричній формі $\mathbf{A}\mathbf{x} = \mathbf{b}$, де

$$\mathbf{A} = (a_{ij})_{i=1,\ldots,m, j=1,\ldots,n}$$

(матриця системи), $\mathbf{x} = (x_1, \ldots, x_n)^\top$ (вектор невідомих), $\mathbf{b} = (b_1, \ldots, b_m)^\top$ (вектор правих частин). Якщо $n = m$, ця система має єдиний розв'язок тоді і тільки тоді, коли $\det \mathbf{A} \neq 0$. У загальному випадку розглядають розширену матрицю системи

$$\tilde{\mathbf{A}} = \begin{pmatrix} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & b_m \end{pmatrix}.$$

Якщо $\mathrm{rank}(\mathbf{A}) \neq \mathrm{rank}(\tilde{\mathbf{A}})$, то система (А.1) не має розв'язків. Якщо $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\tilde{\mathbf{A}})$, то всі розв'язки системи мають вигляд $\mathbf{x} = \sum_{k=1}^{p} c_k \boldsymbol{\psi}_k + \mathbf{x}_0$, де $\mathbf{x}_0$ — деякий розв'язок (А.1), $p = n - \mathrm{rank}(\mathbf{A})$, $\boldsymbol{\psi}_k$, $k = 1, \ldots, p$ — система лінійно незалежних розв'язків рівняння $\mathbf{A}\boldsymbol{\psi} = \mathbf{0}$, c_k — довільні числа. У матричній формі це записується так: $\mathbf{x} = \boldsymbol{\Psi}\mathbf{c} + \mathbf{x}_0$, де $\boldsymbol{\Psi}$ — матриця, стовпчиками якої є $\boldsymbol{\psi}_k$, $k = 1, \ldots, p$, $\mathbf{c} = (c_1, \ldots, c_p)^\top$ — довільний вектор.

Квадратний корінь з матриці. Матриця $\mathbf{B}$ називається квадратним коренем з $\mathbf{A}$, якщо $\mathbf{BB} = \mathbf{A}$ (позначення $\mathbf{B} = \mathbf{A}^{1/2}$). Якщо $\mathbf{A}$ — симетрична і $\mathbf{A} \geq 0$, то існує такий квадратний корінь з неї $\mathbf{A}^{1/2}$, який також є симметричною додатно визначеною матрицею.

Ортогональність. Кутом між векторами $\mathbf{a}$ та $\mathbf{b}$ називають кут α, $0 \leq \alpha \leq \pi$, такий, що

$$\cos(\alpha) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|}.$$

Вектори $\mathbf{a}$ та $\mathbf{b}$ називають ортогональними (позначення $\mathbf{a} \perp \mathbf{b}$), якщо $\langle \mathbf{a}, \mathbf{b} \rangle = 0$.

Додаток Б

Відомості з теорії ймовірностей

Б.1 Випадкові величини та вектори

Повний курс теорії ймовірностей можна знайти у [9]. Тут ми лише нагадаємо деякі поняття, що використовуються у даній книзі, звертаючи основну увагу на їхнє статистичне (частотне) трактування. Таке трактування аж ніяк не вичерпує змісту ймовірнісних понять і потрібне більше для інтуїтивного розуміння цих понять. Зокрема, у класичній теорії ймовірностей основним поняттям є простір випадкових подій з ймовірністю (мірою) на ньому. Ми розглядаємо переважно події, пов'язані з якими-небуть випадковими величинами, і тому всі поняття теорії ймовірностей вводимо, виходячи саме з випадкових величин (векторів). Це приводить до трохи нестандартної форми визначення, скажімо, умовних математичних сподівань. Думаю, такий підхід зрозуміліший для неспеціаліста і водночас еквівалентний стандартному при правильному формальному введенні поняття випадкової величини.

Випадкова величина ξ — це така числова характеристика досліджуваного об'єкта, що для неї можна яким-небудь чином задати

ймовірність того, що $\xi < x$ для будь-яких $x \in \mathbb{R}$. Ця ймовірність, трактована як функція від x, $F_\xi(x) = F(x) = \mathsf{P}\{\xi < x\}$, називається **функцією розподілу** ξ. Звичайний статистичний підхід до визначення цієї ймовірності ґрунтується на припущенні, що дослідник може (хоча б суто гіпотетично) отримати послідовність (вибірку) значень $\xi_1, \ldots \xi_n$ характеристики ξ у різних, але однотипних, незалежних між собою об'єктів[1]. Величини ξ_j називають незалежними копіями ξ. У такому випадку при великих n:

$$\mathsf{P}\{\xi \in [a,b]\} \simeq \nu_n(\xi \in [a,b]),$$

де

$$\nu_n(\xi \in [a,b]) = \frac{1}{n}\#\{j : \xi_j \in [a,b]\}.$$

(Тут $\#(A)$ — кількість елементів множини A). Величину ν_n називають (відносною) частотою події $\xi \in [a,b]$ у вибірці ξ_j, $j = 1, \ldots, n$. Питання про те, чому $\nu_n(\xi \in [a,b])$ повинні бути приблизно однаковими для різних вибірок, належить до філософських основ теорії ймовірності і ми його тут не розглядаємо. Досвід застосування математичної статистики показує, що таке припущення у багатьох випадках узгоджується з реальними даними.

Якщо функція розподілу F_ξ задана, то можна визначити ймовірність події $g(\xi) < x$ для будь-якої вимірної[2] (за Лебегом) фун-

[1] "Незалежність" поки трактуємо у фізичному розумінні як відсутність впливу одних об'єктів на інші. Скажімо, таким об'єктами можуть бути новобранці до армії, а ξ_j — зріст j-го новобранця.

[2] Загальна теорія міри, вимірності та інтеграла викладена у [6, 16], а у застосуванні до теорії ймовірностей — у [2, 9]. Зауважимо, що будь-яка функція g є вимірною за Лебегом, якщо її можна обчислити з довільною заданою точністю за допомогою певного конструктивного алгоритму (напр., на комп'ютері). Тобто всі функції, які використовують нематематики з прикладною метою є вимірними. Множина A є вимірною, якщо вимірним є її індикатор, тобто функція

$$g(x) = \mathbb{1}\{x \in A\} = \begin{cases} 1 & \text{якщо } x \in A \\ 0 & \text{якщо } x \notin A \end{cases}$$

кції g і події $\xi \in A$ для будь-якої вимірної множини A. Таким чином, якщо ξ — випадкова величина, то $g(\xi)$ — також.

Щільністю розподілу ξ відносно деякої міри μ називають таку функцію $f_\xi(t)$, для якої при всіх $x \in \mathbb{R}$:

$$F_\xi(x) = \int_{-\infty}^{x} f_\xi(t)\mu(dt). \qquad \text{(Б.1)}$$

Якщо міра μ не вказана, мають на увазі міру Лебега. Випадкові величини, що мають щільність відносно міри Лебега називають аболютно неперервними.

Крім лебегової міри ми будемо використовувати лічильну міру у випадку, коли випадкова величина є дискретною, тобто ξ може набувати лише значень з зліченної множини $\mathcal{X} = \{x_1, x_2, \dots\}$.

Лічильною мірою на $\mathcal{X}$ називають міру, що визначається як

$$\mu(A) = \#\{\mathcal{X} \cap A\},$$

тобто міра множини A дорівнює кількості елементів $\mathcal{X}$, що належать A. Для лічильної міри μ на $\mathcal{X}$:

$$\int_A g(x)\mu(dx) = \sum_{x \in \mathcal{X} \cap A} g(x). \qquad \text{(Б.2)}$$

Тому для дискретних величин ξ щільність відносно лічильної міри це $f(x) = \mathsf{P}\{\xi = x\}$, якщо $x \in \mathcal{X}$. Можливі і щільності відносно інших мір, але ми їх тут не розглядаємо. Читачі, не обізнані з загальною теорією міри, "інтеграл по мірі Лебега" можуть розглядати як звичайний інтеграл Рімана "по dx", тобто $\int f(x)\mu(dx) = \int f(x)dx$, якщо μ — міра Лебега. Таким чином, $\int f(x)\mu(dx)$ у цій книзі можна розглядати як об'єднаний запис для рімановых інтегралів (коли розглядаються абсолютно неперервні випадкові величини) і сум (коли йдеться про дискретні випадкові величини).

Математичне сподівання випадкової величини можна обчислювати за формулою

$$\mathsf{E}\, g(\xi) = \int g(t)F_\xi(dt) = \int g(t)f_\xi(t)\mu(dt)$$

(якщо такий інтеграл існує). Зі статистичної точки зору

$$\mathsf{E}\, g(\xi) \simeq \frac{1}{n} \sum_{j=1}^{n} g(\xi_j)$$

при великих n, тобто математичне сподівання — це приблизно середнє значення незалежних копій випадкової величини, якщо таких копій взято достатньо багато.

Моментом k-го порядку випадкової величини ξ називають $m_k = \mathsf{E}(\xi)^k$. Дисперсія ξ — це $\mathsf{D}\,\xi = \mathsf{E}(\xi - \mathsf{E}\,\xi)^2 = m_2 - (m_1)^2$.

Квантилем рівня α розподілу випадкової величини ξ називають

$$Q^{\xi}(\alpha) = Q^{F_\xi}(\alpha) = \inf\{x : \ F_\xi(x) > \alpha\}.$$

Якщо функція розподілу F неперервна і має обернену, то $Q^F(\alpha) = F^{-1}(\alpha)$. **Медіаною** розподілу ξ називають $Q^{\xi}(1/2)$.

Розподіл випадкового вектора $\boldsymbol{\xi} = (\xi^1, \ldots, \xi^d)^\top$ задається **функцією розподілу** $F_{\boldsymbol{\xi}} : \mathbb{R}^d \to [0, 1]$:

$$F_{\boldsymbol{\xi}}(\mathbf{x}) = \mathsf{P}\{\boldsymbol{\xi} < \mathbf{x}\} = \mathsf{P}\{\xi^1 < x^1, \ldots, \xi^d < x^d\}.$$

Щільність розподілу випадкового вектора $f_{\boldsymbol{\xi}}$ відносно міри μ на $\mathbb{R}^d$ визначається формулою

$$F_{\boldsymbol{\xi}}(x) = \int_{\mathbf{t}<\mathbf{x}} f_{\boldsymbol{\xi}}(\mathbf{t})\mu(d\mathbf{t}).$$

(Якщо μ — міра Лебега, — це d-кратний інтеграл за $dt^1 dt^2 \ldots dt^d$, а якщо μ — лічильна міра, то звичайна сума).

Математичне сподівання від $g(\boldsymbol{\xi})$ де $g : \mathbb{R}^d \to \mathbb{R}$ можна обчислювати за формулою

$$\mathsf{E}\, g(\boldsymbol{\xi}) = \int g(\mathbf{t}) F_{\boldsymbol{\xi}}(d\mathbf{t}) = \int g(\mathbf{t}) f_{\boldsymbol{\xi}}(\mathbf{t})\mu(d\mathbf{t}).$$

Якщо g — векторнозначна функція, $g : \mathbb{R}^d \to \mathbb{R}^m$, то математичне сподівання від g — це вектор, складений з математичних сподівань окремих координат.

Коваріація випадкових величин ξ, η — це[3]

$$\operatorname{cov}(\xi,\eta) = \mathsf{E}\,\xi\eta - \mathsf{E}\,\xi\,\mathsf{E}\,\eta.$$

Коваріаційна матриця вектора $\boldsymbol{\xi}$:

$$\operatorname{cov}(\boldsymbol{\xi}) = (\operatorname{cov}(\xi^i,\xi^k))_{i,k=1,\dots,d} = \mathsf{E}\,\boldsymbol{\xi}\boldsymbol{\xi}^\top - \mathsf{E}\,\boldsymbol{\xi}(\mathsf{E}\,\boldsymbol{\xi})^\top.$$

Б.2 Умовні ймовірності та математичні сподівання. Незалежність

Ми розглядаємо всі випадкові величини та вектори "над одним ймовірнісним простором" Ω, тобто вважаємо, що коли, скажімо, ξ, η і ζ — випадкові величини, то (ξ,η,ζ) — випадковий вектор з певною (можливо, невідомою) функцією розподілу.

Умовну ймовірність $\mathsf{P}\{\xi\in A \mid \eta\in B\}$ того, що $\xi\in A$ за умови $\eta\in B$, можна уявляти собі так. Нехай ми маємо "довгу" вибірку (ξ_1,η_1), …, (ξ_n,η_n) з незалежних копій вектора (ξ,η). Відібравши з неї лише ті вектори, для яких $\eta_j\in B$, отримаємо "коротку", проріджену вибірку (ξ_{j_1},η_{j_1}),…,(ξ_{j_m},η_{j_m}). Тепер відносна частота $\xi\in A$ у прорідженій вибірці

$$\nu_m(\xi\in A\mid\eta\in B\} = \frac{1}{m}\#\{i:\ \xi_{j_i}\in A\}$$

є умовною відносною частотою події $\xi\in A$ у довгій вибірці, за умови $\eta\in B$. Подібно до звичайних частот умовні виявляють властивість стабільності, вони мало змінюються при переході від однієї вибірки до іншої, якщо обсяги цих вибірок великі. Умовна ймовірність — це те значення, навколо якого коливаються такі частоти:

$$\mathsf{P}\{\xi\in A\mid\eta\in B\} \simeq \nu_m\{\xi\in A\mid\eta\in B\}.$$

[3] Математичне сподівання $\mathsf{E}\,\xi\eta$ можна розглядати тільки коли ξ та η є компонентами деякого випадкового вектора $\boldsymbol{\xi}=(\xi,\eta)$. У цьому випадку можна задати функцію $g((x,y))=x\cdot y$ і покласти $\mathsf{E}\,\xi\eta = \mathsf{E}\,g(\xi,\eta)$.

Аналогічно, умовне математичне сподівання — це приблизно вибіркове середнє по прорідженій вибірці:

$$\mathsf{E}(g(\xi) \mid \eta \in B) \simeq \frac{1}{m} \sum_{i=1}^{m} g(\xi_{j_i}).$$

У випадку, коли $\mathsf{P}\{\eta \in B\} \neq 0$, цей підхід приводить до означень

$$\mathsf{P}(\xi \in A \mid \eta \in B) = \frac{\mathsf{P}\{\xi \in A, \eta \in B\}}{\mathsf{P}\{\eta \in B\}}$$

— **умовна ймовірність**,

$$F_\xi(x \mid \eta \in B) = \mathsf{P}\{\xi < x \mid \eta \in B\} \tag{Б.3}$$

— **умовна функція розподілу**.

Умовна щільність розподілу визначається як функція $f_\xi(x \mid \eta \in B)$, для якої

$$\mathsf{P}(\xi \in A \mid \eta \in B) = \int_A f_\xi(x \mid \eta \in B)\mu(dx) \tag{Б.4}$$

виконано для всіх вимірних множин A.

Умовне математичне сподівання

$$\mathsf{E}(g(\xi) \mid \eta \in B\} = \int g(x) f_\xi(x \mid \eta \in B)\mu(dx) = \\ = \int g(x) F_\xi(dx \mid \eta \in B). \tag{Б.5}$$

На жаль, ці означення не працюють, коли $\mathsf{P}\{\eta \in B\} = 0$. Зокрема, так неможливо визначити $\mathsf{P}\{\xi \in A \mid \eta = t\}$, якщо розподіл η абсолютно неперервний. В цьому випадку зручно використовувати "обернений" підхід до означення умовних характеристик, починаючи з **умовного математичного сподівання** : $\mathsf{E}(g(\xi) \mid \eta = t)$ це така функція $h(t)$, що для будь-якої вимірної множини A

$$\mathsf{E}\, g(\xi)\mathbb{1}\{\eta \in A\} = \mathsf{E}\, h(\eta)\mathbb{1}\{\eta \in A\}$$

(тобто $h(\eta)$ має право заміняти $g(\xi)$ при обчисленні математичних сподівань такого вигляду). Випадкову величину (вектор) $h(\eta)$ позначають також $\mathsf{E}(\xi \mid \eta)$.

Відповідно, **умовний розподіл** ξ:

$$\mathsf{P}(\xi \in A \mid \eta = t) = \mathsf{E}(\mathbb{1}\{\xi \in A\} \mid \eta = t),$$

а **умовна щільність** $f_\xi(x \mid \eta = t)$ визначається для умовного розподілу так само, як і раніше:

$$\mathsf{P}(\xi \in A \mid \eta = t) = \int_A f_\xi(x \mid \eta = t)\mu(dx).$$

Якщо $\mathsf{P}(\xi \in A \mid \eta = t) = p(A, t)$, то випадкову величину $p(A, \eta)$ позначають $\mathsf{P}(\xi \in A \mid \eta)$. Ці означення узгоджені з (Б.3) — (Б.5) в тому розумінні, що вони збігаються при $\mathsf{P}\{\eta = t\} > 0$.

Незалежність. Випадкові вектори ξ та η називають незалежними, якщо для будь-яких множин A та B:

$$\mathsf{P}\{\xi \in A \text{ і } \eta \in B\} = \mathsf{P}\{\xi \in A\}\,\mathsf{P}\{\eta \in B\}.$$

Якщо ξ та η незалежні, то $\mathsf{P}\{\xi \in A\} = \mathsf{P}(\xi \in A \mid \eta \in B)$ для будь-яких A та B.

Скорочення м.н. (майже напевно) позначає, що відповідна подія відбувається з ймовірністю 1, наприклад, $\xi < \eta$ м.н., це те ж саме, що $\mathsf{P}\{\xi < \eta\} = 1$.

Аналогічно можна розглядати умовні математичні сподівання для випадкових векторів.

Обчислення умовних математичних сподівань та ймовірностей. Якщо (ξ, η) — пара випадкових величин, що мають спільну щільність розподілу $f_{\xi\eta}(x, y)$ (відносно міри Лебега), то

$$f_\xi(x \mid \eta = y) = \frac{f_{\xi\eta}(x, y)}{f_\eta(y)},$$

де $f_\eta(y) = \int f_{\xi\eta}(x, y)dx$ — щільність розподілу η. Відповідно

$$\mathsf{E}(g(\xi) \mid \eta = y) = \int g(x) f_\xi(x \mid \eta = y)dx$$

(якщо цей інтеграл існує),

$$\mathsf{P}(\xi \in A \mid \eta = y) = \int_A f_\xi(x \mid \eta = y\}dx.$$

Ці твердження легко узагальнюються на умовні характеристики випадкових векторів та щільності відносно довільних мір.

Б.3 Багатовимірний нормальний розподіл

Кажуть, що випадковий вектор $\mathbf{X} = (\xi_1, \ldots, \xi_d)^\top$ з середнім $\mathbf{m} = (m_1, \ldots, m_d)^\top$ та коваріаційною матрицею $\mathbf{S} = (s_{ij})_{i,j=1,\ldots,d}$, є **нормальним** (гауссовим), якщо його характеристична функція

$$\mathsf{E}\exp\left(i\langle \mathbf{X}, \mathbf{u}\rangle\right) = \exp\left(-\frac{(\mathbf{m}-\mathbf{u})^\top \mathbf{S}(\mathbf{m}-\mathbf{u})}{2}\right).$$

(Тут $i = \sqrt{-1}$.)

Позначення — $\mathbf{X} \sim N(\mathbf{m}, \mathbf{S})$. Якщо матриця $\mathbf{S}$ — невироджена, то щільність розподілу $\mathbf{X}$ має вигляд

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \mathbf{S}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^\top \mathbf{S}^{-1}(\mathbf{x}-\mathbf{m})\right).$$

При цьому $\mathsf{E}\,\mathbf{X} = \mathbf{m}$, $\operatorname{cov}(\mathbf{X}) = \mathsf{E}(\mathbf{X}-\mathbf{m})(\mathbf{X}-\mathbf{m})^\top = \mathbf{S}$.

Зокрема, якщо $d = 1$, маємо **нормальну випадкову величину** $\xi \sim N(m, \sigma^2)$ з щільністю розподілу

$$f_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

Стандартною нормальною називається випадкова величина $N(0,1)$. Функцію розподілу $N(0,1)$ позначають $\Phi(x)$, щільність — $\varphi(x)$.

Якщо координати нормального вектора некорельовані (тобто $\operatorname{cov}(\xi_i, \xi_j) = 0$), то вони незалежні. Для інших розподілів це не завжди так.

Б.4 Збіжність та граничні теореми

Збіжність. Якщо для довільного $\varepsilon > 0$, $\mathsf{P}\{|\xi_n - \xi| > \varepsilon\} \to 0$ при $n \to \infty$, то кажуть, що $\xi_n \to \xi$ **за ймовірністю** (запис $\xi_n \xrightarrow{\mathsf{P}} \xi$).

Якщо $\mathsf{P}\{\lim_{n\to\infty} \xi_n = \xi\} = 1$, то кажуть, що $\xi_n \to \xi$ **майже напевне** (м.н.).

Якщо $\mathsf{E}(\xi_n - \xi)^2 \to 0$, то кажуть, що $\xi_n \to \xi$ **у середньому квадратичному** (с.к.).

Якщо для всіх неперервних, обмежених функцій g виконується збіжність $\mathsf{E}\, g(\xi_n) \to \mathsf{E}\, g(\xi)$, то кажуть, що ξ_n збігається до ξ **слабко** (запис $\xi_n \Rightarrow \xi$ або $\xi_n \xrightarrow{w} \xi$). Слабку збіжність можна визначити як збіжність розподілів: якщо $\xi_n \Rightarrow \xi$, η_n має той же розподіл, що і ξ_n а η — той же, що і ξ, то $\eta_n \Rightarrow \eta$. Тому, можливо, краще казати, що розподіл ξ_n збігається до розподілу ξ слабко, але ми часто будемо скорочувати у цій фразі слово "розподіл".

Випадкові вектори ξ_n у $\mathbb{R}^d$ **збігаються слабко** до ξ тоді і тільки тоді, коли $\mathsf{P}\{\xi_n < x\} = F_{\xi_n}(x) \to F_\xi(x) = \mathsf{P}\{\xi < x\}$ для всіх $x \in \mathbb{R}^d$ у яких функція $F_\xi(x)$ є неперервною.

Якщо $\xi_n \Rightarrow \xi$ і g — неперервна функція, то $g(\xi_n) \Rightarrow g(\xi)$ (**теорема неперервності**, [3], підрозд. 1.5).

Якщо $\xi_n \Rightarrow \xi$ і $\eta_n \to c$ за ймовірністю, де c — невипадкове число, то $\eta_n \xi_n \Rightarrow c\xi$) (**теорема Слуцького**, наслідок з теореми неперервності).

Закони великих чисел — це твердження про збіжність вибіркових середніх до математичних сподівань.

Нехай $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N$ — незалежні випадкові вектори у $\mathbb{R}^d$, $\bar{\boldsymbol{\xi}}_N = \frac{1}{N}\sum_{j=1}^N \boldsymbol{\xi}_j$.

Теорема (посилений закон великих чисел) ([5], с.148). *Якщо* $\boldsymbol{\xi}_j$ *однаково розподілені і є скінченним* $\mathsf{E}\,\boldsymbol{\xi}_j = \mathbf{a}$*, то* $\bar{\boldsymbol{\xi}}_N \to \mathbf{a}$ *м.н.*

Теорема (закон великих чисел Чебишева) ([5], с.136). *Якщо існують* $\mathsf{E}\,\boldsymbol{\xi}_j = \mathbf{a}$*, і* $\mathsf{E}\,\|\boldsymbol{\xi}_j\|^2 \le C < \infty$ *то* $\bar{\boldsymbol{\xi}}_N \to \mathbf{a}$ *за ймовірністю.*

Існують і інші, більш загальні формулювання закону великих чисел див. [5], [2].

Центральна гранична теорема. Центральну граничну теорему можна розглядати як характеризацію швидкості збіжності у законі великих чисел. Ми наведемо лише варіант цієї теореми для однаково розподілених доданків, більш загальні версії можна знайти, наприклад, у [5].

Теорема. *Якщо* $\boldsymbol{\xi}_j$ *— однаково розподілені вектори у* $\mathbb{R}^d$, $\mathsf{E}\,\boldsymbol{\xi}_j = \mathbf{a}$, $\operatorname{cov}(\boldsymbol{\xi}_j) = \mathbf{S}$, *то*

$$\frac{1}{\sqrt{N}} \sum_{j=1}^{N} (\boldsymbol{\xi}_j - \mathbf{a}) = \sqrt{N}(\bar{\boldsymbol{\xi}}_N - \mathbf{a}) \Rightarrow \boldsymbol{\eta},$$

де $\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{S})$.

Покажчик засобів R

:, 34
<-, 26
«-, 79, 258
?, 22
??, 22
???, 24
%/%, 31
%>%, 73
%%, 31
..., 81

abline(), 108
add, 99, 109
aov(), 464
apply(), 163
as., 51

barplot(), 95, 283
boxplot(), 287
break, 88

c(), 30
car, 467, 513, 543
cat(), 90
cbind(), 41
chisq.test(), 439, 455
clipboard, 69
col.names, 68
colMeans(), 160
colnames(), 45
contour(), 113
cor(), 175, 191, 201
cor.test(), 201, 202, 472
corrplot(), 179
curve(), 109
cut(), 40, 284, 473

data.frame(), 49
Deming(), 497
detach(), 21
diag(), 45
dimnames(), 44
dplyr, 74
draw.ellipse(), 358

edit(), 49
else, 85

file.choose(), 66

findInterval(), 123
fitdistr(), 360, 399
floating.pie(), 118
for, 89
function, 76

gather(), 71
getData(), 119
ginv(), 46
goodfit(), 447
grey(), 123
grid(), 475

header, 67
help(), 22
help.search(), 22
hist(), 264

if, 84
ifelse(), 85
in, 89
Inf, 32
install.packages(), 21
integrate(), 337
IQR(), 145
iris, 48
is., 50

knitr, 19

lapply(), 52
ldeaths, 95
legend(), 119
legend.text, 98
length(), 53
letters, 45
levels, 39
leveneTest, 467
library(), 21
linearHypotheses(), 543
lines(), 101
list(), 45
lm(), 492, 508
load(), 65
loglik, 399

mad(), 161
maptools, 116
MASS, 46, 360, 456
matplot(), 106
matrix, 43
max(), 53
mean(), 53, 132
median(), 139
MethComp, 497
min(), 53
mosaicplot(), 456

NA, 33
names(), 33
NaN, 32
ncol, 43
next, 88
nleqslv(), 296
nlm(), 314
nls(), 557
nrow, 43
NULL, 33
numeric(), 35

optim(), 314, 360
ordered, 41
outer(), 114

pairs(), 168
palette(), 123
persp(), 113
pie(), 99
pipe, 73
plot(), 100, 168
plot3d(), 112, 250
plotCI(), 348
plotrix, 118, 348, 358
pmax(), 53
pmin(), 53
points(), 101
polygon(), 410
ppoints(), 278
prod(), 53, 134

qgraph(), 185
qqline(), 276, 519
qqnorm(), 276, 519
qqplot(), 281
quantreg, 499

raster, 116
rbind(), 41
read.csv(), 68
read.csv2(), 69
read.table(), 66
rect(), 107
reliaR, 365
remove.packages(), 22
rep(), 35
repeat, 88
rgl, 112
round(), 135
row.names, 49, 67
rownames(), 45
rq(), 499

sapply(), 52
save(), 65
scatterplot3d(), 110
scatterplotMatrix(), 171
scatterpotMatrix(), 513
sd(), 144
segments(), 279
select(), 76
separate(), 72
seq(), 33
set.seed(), 261
slice(), 73
solve(), 46
sort(), 53, 139
sos, 23
sp, 116
stringsAsFactors, 49
subset(), 43, 76
sum(), 53
switch(), 86

t(), 46
t.test(), 413
table(), 455
tapply(), 164
text(), 105
tidyr, 70

`var()`, 144, 159
`var.test()`, 423
`vcd`, 447
`Vectorize()`, 83
`weightedMedian()`, 162
`while()`, 87
`write.csv()`, 68
`write.csv2()`, 69
`write.table()`, 67

Навчальне видання

МАЙБОРОДА Ростислав Євгенович

Комп'ютерна статистика

Підручник

Оригінал-макет виготовлено автором за допомогою видавничого пакета LATEX2ε

Формат 60x84$^{1/16}$. Ум. друк. арк. 45,8. Наклад 100. Зам. № 218-9010.
Гарнітура Times New Roman. Папір офсетний. Друк офсетний. Вид. № М12*.
Підписано до друку 11.04.19

Видавець і виготовлювач
ВПЦ "Київський університет"
б-р Т. Шевченка, 14, м. Київ, 01601
☎ (044) 239 32 22; (044) 239 31 72; тел./факс (044) 239 31 28
e-mail: vpc_div.chief@univ.net.ua; redaktor@univ.net.ua
http: vpc.univ.kiev.ua
Свідоцтво суб'єкта видавничої справи ДК № 1103 від 31.10.02