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

**Р. Майборода**

# Непараметрична статистика

Навчальний посібник

*Рекомендовано до друку Науково-Методичною радою Київського національного університету імені Тараса Шевченка, протокол № 07-24 від 29 серпня 2024 р.*

Київ - 2024

УДК 519.22.35
ББК 22.172я73

Рецензенти:
**Ольга Василик** доктор фіз.-мат. наук, професор кафедри математичного аналізу та теорії ймовірностей НТУУ "КПІ імені Ігоря Сікорського"
**Ганна Сливка-Тилищак** доктор фіз.-мат. наук, завідувач кафедри теорії ймовірностей та математичного аналізу ДВНЗ "Ужгородський національний університет"
**Ростислав Ямненко** доктор фіз.-мат. наук, завідувач кафедри теорії ймовірностей, статистики та актуарної математики КНУ імені Тараса Шевченка

**Майборода Р.Є.** Непараметрична статистика. Навчальний посібник.

У навчальному посібнику розглядаються основні методи непараметричної статистики та їхнє застосування до розв'язання прикладних задач. Розглянуті задачі оцінки функцій розподілу, ймовірнісних характеристик розподілу, щільності, функції регресії. Показано, як будуються оцінки по кратних вибірках, цензурованих даних та даних, отриманих за зміщеними вибірковими процедурами. Розглянута асимптотична теорія оцінок, що дозволяє порівнювати різні оцінки та оптимально підбирати параметри налаштування. Як приклад застосування описаних непараметричних технік розглядаються алгоритми класифікації з вчителем. Окремий розділ присвячений непараметричним тестам для перевірки гіпотез про однорідність двох вибірок та про залежність двох змінних.

Матеріал, викладений у підручнику, відповідає курсу "Непараметрична статистика" магістерського рівня для студентів механіко-математичного факультету освітніх програм "Прикладна та теоретична статистика" та "Актуарна та фінансова математика". Він може бути корисний студентам і аспірантам фізико-математичних, економічних та медико-біологічних спеціальностей а також всім, хто цікавиться прикладною та теоретичною статистикою.

*Рекомендовано до друку вченою радою механіко-математичного факультету Київського національного університету імені Тараса Шевченка, протокол №14 від 2 травня 2024 р.*

© **Майборода Р.Є. 2024**

# Передмова

Ця книга написана за матеріалами різних курсів з непараметричної (і не тільки) статистики, які я читав у Київському національному університеті протягом більше двадцяти років. За цей час багато що змінилося у науці, у викладанні, та й взагалі — у житті. Найбільш помітні зміни вніс стрімкий розвиток комп'ютерної техніки. Потужність комп'ютерів неймовірно зросла. Від перших кроків машинного навчання ми дійшли до того, що дресування роботів стає чи не найбільш перспективною професією найближчого майбутнього. Свого часу статистика, і в першу чергу - непараметрична, чимало прислужилася розвитку алгоритмів машинного навчання. І сьогодні адекватна співпраця природних та штучних інтелектів неможлива без розуміння математичних основ аналізу випадкових даних. Вивчення непараметричної статистики — одна з чудових можливостей набути такого розуміння.

Непараметрична статистика весь час була і в центрі моїх наукових інтересів. Викладання часто переплiталося із науковим дослідженням. Інколи підготовка до лекції відкривала мені очі на шлях розв'язання чергової наукової проблеми. Часом результат з наукової статті виявлявся чудовим прикладом для розповіді студентам. Це захоплювало не менше ніж хороший пригодницький серіал. Мені хотілося б передати читачам хоча б частку цього захоплення.

Працюючи я завжди користався допомогою і підтримкою людей, які мене оточували. Хочу подякувати своїм вчителям — Михайлу Ядренку, Юрію Козаченку, Влерію Булдигіну за те, чому вони мене навчили у науці і не тільки. Своїм колегам, зокрема Юлії Мішурі і Олександру Кукушу, я вдячний за дискусії і поради. Дякую також рецензентам, Ользі Василик, Ганні Сливці-Тилищак та Ростиславу Ямненку, які доклали чимало зусиль щоб поліпшити книгу. І головне — студентам, прискіпливій увазі яких я завдячую наснагою до роботи.

Також і вам, шановний читачу, я буду вдячний, якщо читаючи цю книгу ви знайдете у ній щось цікаве та корисне для себе. А найбільше — якщо побачите як її можна покращити і поділитесь цим зі мною.

Бажаю успішного читання.

Дякую.

Ростислав Майборода

# Зміст

# Список позначень та скорочень

в.в. — випадкова величина;

м.н. — майже напевне;

ф.р. — функція розподілу;

$\mathrm{D}\,\xi$ — дисперсія в.в. $\xi$;

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

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

LDA — лінійний дискримінантний аналіз.

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

$\mathsf{MSE}(\hat{f}_n(x))$ — середньоквадратичний ризик оцінки $\hat{f}_n$ в точці $x$.

$\mathsf{MISE}(\hat{f}_n)$ — проінтегрований середньоквадратичний ризик оцінки $\hat{f}_n$.

$\mathbf{IQR}(X)$ — інтерквартильний розмах вибірки $X$;

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

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

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

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

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

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

QDA — квадратичний дискримінантний аналіз.

$\xi_n \xrightarrow{W} \xi$ — послідовність $\xi_n$ слабко збігається до $\xi$.

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

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

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

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

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

$\#A$ — кількість елементів множини $A$;

$n!$ — факторіал: $n! = n(n-1)\dots 1$;

$C_n^m$ — число комбінацій з $n$ елементів по $m$: $C_n^m = \frac{n}{m!(n-m)!}$

◀ — закінчення тексту прикладу.

# Вступ

Читаючи цю книгу корисно враховувати, що непараметрична статистика — це не якась окрема наука, а лише назва навчальної дисципліни. Справжня статистика неподільна, статистик має використовувати ті методи і технології аналізу даних, які найкраще відповідають поставленій задачі. Але для вивчення цих методів потрібна певна послідовність і групування за подібністю, інакще у розмаїтті інформації можна буде втонути. Тому ми і виділяємо ряд схожих між собою підходів під єдиною назвою, хоча і не триматимемось у викладанні строгої "непараметричності", а будемо захоплювати з сусідніх розділів все те, що допоможе розібратись по суті справи.

Отже, про що у нас піде мова. Аналізуючи випадкові явища, статистик намагається описувати їх за допомогою теоретичних ймовірнісних моделей. Для того, щоб такі моделі можна було застосовувати у реальному житті, вони повинні мати параметри, які будуть налаштовуватись для опису тих або інших даних. Є моделі, які можна задати одним — двома числовими параметрами. Скажімо, нормальний розподіл однозначно задається двома: математичним сподіванням і дисперсією. Таких числових параметрів може бути і більше, але все ж, порівняно невелике число, значно менше ніж обсяг даних, по яких ми збираємось підганяти модель. У цьому випадку модель називають параметричною.

Техніку підгонки параметричних моделей вивчає параметрична статистика. Але дуже часто буває, що у дослідника немає жодних теоретичних припущень про можливий розподіл даних. У такому випадку параметрична модель може виявитися занадто жорсткою — дані не вкладаються в неї, які б параметри ми ні підібрали. Для таких випадків доцільно використовувати більш гнучкі моделі, в яких невідомі параметри є функціями, або об'єктами ще більш складної природи. Так, дуже поширеною моделлю теоретичною даних є вибірка з однаково розподілених

випадкових величин, функція розподілу $F$ яких повністю невідома. В цьому випадку ми кажемо, що $F$ є невідомою непараметричною складовою моделі, а саму модель називаємо непараметричною. (Інколи, щоб об'єднати параметричну і непараметричну статистику в одному розгляді, непараметричну складову також називають "параметром" моделі, а числові параметри називають евклідовими).

Задачами оцінювання непараметричних складових, перевірки гіпотез про них, використання таких оцінок для прогнозування, класифікації та інших потреб займається непараметрична статистика. У цій книзі ми зупинимось на ряді основних таких задач.

У першому розділі розглядається задача оцінювання невідомої функції розподілу (ф.р.). Класичний варіант цієї задачі ми вже згадували — оцінка ф.р. за вибіркою з незалежних однаково розподілених випадкових величин, які мають цю ф.р. Класичною оцінкою в цьому випадку є емпірична ф.р. побудована за вибіркою. На практиці зустрічаються більш складні задачі, коли не всі величини, які нас цікавлять вдається повністю спостерігати (цензурована вибірка), або коли ф.р. у вибірці не відповідає ф.р. у популяції внаслідок використання зміщеної процедури відбору. У таких випадках оцінку треба модифікувати. Як це можна зробити, якими є статистичні властивості отриманих оцінок, як їх використовувати, наприклад, для побудови довірчих інтервалів — ми обговорюємо у першому розділі.

Другий розділ присвячений тому, як за оцінками ф.р. можна будувати оцінки для числових характеристик розподілу, таких, як математичні сподівання, дисперсії, квантилі...

У третьому розділі ми знайомимось з оцінками щільності розподілу. Формально щільність розподілу можна визначити як похідну[1] від функції розподілу, тобто на перший погляд здається, що тут можна було б використати якісь підходи з другого розділу. Але задача виявляється складнішою. Оцінки щільності суттєво відрізняються від оцінок математичних сподівань та інших числових характеристик, як за логікою побудови, так і за властивостями. Щоб правильно їх використовувати, потрібно ці властивості глибоко дослідити.

Таке дослідження проведено у четвертому розділі для одного виду оцінок щільності — ядерних. Тут ми розбираємо, як правильно налаштовувати ці оцінки, щоб забезпечити їхню оптимальну поведінку. Щоб

[1]Похідну Радона від розподілу по мірі Лебега, якщо бути пуристом.

технічні деталі не заважали бачити загальну логіку, акуратні доведення наводяться лише для одновимірних спостережень, а результати для багатовимірного випадку розбираються без доведень.

У п'ятому розділі з'ясовується для чого буває потрібно оцінювати щільності розподілу. Тут розглядаються задачі статистичної класифікації. Це як раз та область, в якій так стрімко нині розвиваються алгоритми машинного навчання. Ми розглядаємо баєсів підхід до задачі класифікації, який дозволяє вибрати найкращі в певному розумінні алгоритми. Щільності розподілу спостережуваних характеристик, за якими ми класифікуємо об'єкти, грають в цьому першу скрипку.

Шостий розділ присвячений непараметричним оцінкам функції регресії. Це також важлива для застосувань задача, тісно пов'язана із задачами прогнозування та класифікації. Підходи до неї дуже схожі на підходи до оцінювання щільності розподілу, тому ми не зупиняємося детально на дослідженні оцінок, а розглядаємо їхні модифікації, які виявляються найбільш вдалими при практичних застосуваннях.

Сьомий розділ розташований дещо осібно. Тут ми розглядаємо задачі перевірки непараметричних гіпотез та відповідні статистичні тести. Дві такі задачі, які найчастише зустрічаються на практиці — перевірка однорідності двох вибірок та перевірка залежності між двома змінними за однією вибіркою. Саме їм приділено основну увагу.

Наприкінці кожного розділу вміщені запитання, задачі для розв'язання ручкою на папері і завдання для виконання на комп'ютері. Звичайно, ці завдання не можуть охопити всю інформацію, вміщену у підручнику, але, якщо читач спробує подумати над ними, це допоможе йому зрозуміти як працюють описані тут технології. Комп'ютерні завдання сформульовано так, щоб їх можна було виконувати у будь-якій системі програмування загального призначення. Я рекомендую використовувати для цього мову статистичного програмування `R`, про яку можна дізнатись у [3].

Для успішного використання цієї книги читачу потрібні знання з теорії ймовірності та математичної статистики в обсязі загального бакалаврського курсу для математиків та статистиків. Якщо потрібно, прогалини можна заповнити, користуючись підручниками [1, 5]. З більш детальних курсів математичної статистики, в яких міститься багато інформації саме про непараметричні підходи до аналізу даних, можу рекомендувати [9, 26]. Підручник [28] може дати загальне уявлення про сучасну непараметричну статистику, хоча для більш глибокого розуміння математичної теорії, що лежить в її основі, краще користуватись

іншими книгами. Багато корисного можна знайти у книзі [23]. Тим, хто цікавиться застосуваннями статистики до аналізу медичних та біологічних даних, рекомендую [19].

# Розділ 1

# Оцінки функції розподілу

## 1.1 Оцінювання функції розподілу за кратною вибіркою

Перша задача непараметричної статистики, яку ми розглянемо детально у цьому підручнику — оцінювання функції розподілу (ф.р.) випадкової величини за кратною вибіркою. Слід відмітити, що сама функція розподілу у статистиці використовується порівняно не часто, але на основі розглянутих у цьому розділі оцінок ф.р. можна будувати оцінки для багатьох інших ймовірнісних характеристик випадкових випадкових величин. Крім того, на прикладі задачі оцінки ф.р. можна порівняно просто побачити, як працюють загальні методи оцінювання, котрі будуть потім застосовуватись у значно більш складних задачах.

### 1.1.1 Функції розподілу та емпіричні функції розподілу

Нехай $\xi$ — випадкова величина. Нагадаємо, що функцією розподілу $\xi$ називають функцію

$$F^{\xi}(x) = \mathsf{P}\{\xi < x\}.$$

Розподіл випадкової величини, тобто міра на наборі всіх вимірних множин, яка дорівнює ймовірності потрапити у цю множину, позначається

$$P^{\xi}(A) = \mathsf{P}(\xi \in A)$$

Далі у цій книзі, якщо для позначення якоїсь функції розподілу використано певну літеру, наприклад, $G$ то для відповідного розподілу може використовуватись та сама літера:

$$G(x) = \mathsf{P}\{\xi < x\} \text{ рівносильно } G(A) = \mathsf{P}\{\xi \in A\}.$$

Про яку саме функцію йдеться легко зрозуміти за її аргументом. Між цими функціями є очевидний зв'язок:

$$G(x) = G((-\infty, x)),\ G(A) = \int_A G(dx) = \int_{\mathbb{R}} \mathbb{1}\{x \in A\} G(dx),$$

де $\mathbb{1}\{B\}$ — індикатор події $B$:

$$\mathbb{1}\{B\} = \begin{cases} 1 & \text{якщо подія } B \text{ відбулась,} \\ 0 & \text{якщо подія } B \text{ не відбулась.} \end{cases}$$

Припустимо, що спостерігається кратна вибірка, тобто набір незалежних, однаково розподілених випадкових величин $\mathbf{X} = (\xi_1, \ldots, \xi_n)$, причому ф.р. $\xi_j - F$ є повністю невідомою і її потрібно оцінити за даними $\mathbf{X}$. Формально спільний розподіл всіх спостережуваних даних $\mathbf{X}$ можна задати як

$$P_F^{\mathbf{X}}(A) = \mathsf{P}\{\mathbf{X} \in A\} = \int_{x_1 \in \mathbb{R}} \cdots \int_{x_n \in \mathbb{R}} \mathbb{1}\{\mathbf{x} \in A\} F(dx_1) \ldots F(dx_n)$$

де $B \subseteq \mathbb{R}^n$, $\mathbf{x} = (x_1, \ldots, x_n)^T$.

Цей розподіл залежить від невідомого параметра $F$, який може були будь-якою функцією розподілу: $F \in \Theta$. Тут $\Theta$ — множина всіх неспадних, неперервних зліва функцій $G : \mathbb{R} \to [0, 1]$, таких, що $G(-\infty) = 0$, $G(+\infty) = 1$. Оскільки $F$ не можна описати використовуючи невелику кількість числових параметрів[1], це - непараметрична задача.

Щоб побудувати оцінку для $F(x)$ згадаємо, що це ймовірність події $\{\xi_j < x\}$. Отже її природною оцінкою є відповідна відносна частота таких подій у нашій вибірці:

$$\hat{F}_n(x) = \frac{\#\{j : \xi_j < x\}}{n} = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{\xi_j < x\}.$$

[1]Насправді можна, але так неадекватно, що краще цього не робити.

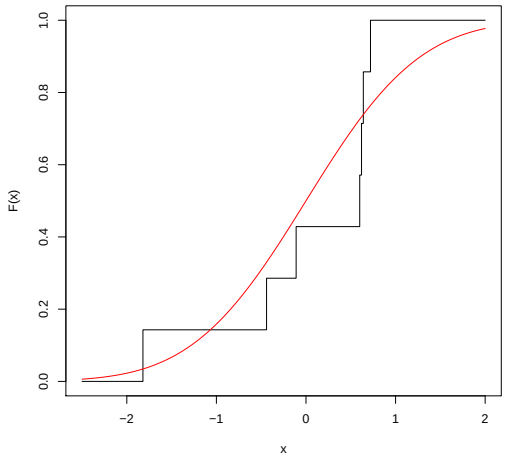



Рис. 1.1: Емпірична функція розподілу (чорна лінія) для семи спостережень та теоретична ф.р. (червона лінія)

(Символ $\#A$ позначає кількість елементів множини $A$).

Ця функція відповідає розподілу, який кожній вимірній множині $A \subseteq \mathbb{R}$ ставить у відповідність ймовірність, що дорівнює

$$\hat{F}_n(A) = \frac{1}{n}\#\{j : \xi_j \in A\}.$$

Якщо розглядати вибірку $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ як набір не випадкових фіксованих чисел, то $\hat{F}_n(A), A \subseteq \mathbb{R}$ це розподіл випадкової величини $\xi^*$, котра приймає значення $\xi_j$ з ймовірністю $w_j = 1/n$.

Наприклад, графік емпіричної функції розподілу $\hat{F}_n$ для вибірки з семи елементів $\mathbf{X} = (0.59, 0.71, -0.11, -0.45, 0.61, -1.82, 0.63)$ зображено на рис. 1.1 чорною лінією. Ми бачимо, що графік являє собою ступінчасту лінію зі стрибками у точках, які відповідають вибірковим значенням. Висота кожного стрибка дорівнює $1/n$. Так буде, якщо у вибірці всі спостережувані значення — різні. Якщо деякі з них однакові, висота стрибка буде дорівнювати кількості значень, що попадають у дану точку, діленій на $n$.

Насправді, вибірка $\mathbf{X}$ була отримана генерацією (псевдо)випадкових чисел зі стандартним нормальним розподілом (з округленням до сотих). Графік ф.р. $F$ стандартного нормального розподілу зображено на рис.

1.1 червоною лінією. Оскільки $\hat{F}_n$ ми розглядаємо як оцінку для $F$, можна сподіватись, що, при достатньо великій кількості спостережень, їхні графіки будуть близькими. Звичайно, число 7 не можна вважати "достатньо великою кількістю" і не варто сподіватись, що оцінка ф.р. по такій вибірці буде хорошою. Подивимось, як ця оцінка працюватиме при більших обсягах вибірки.

**Приклад 1.1.1.** На рис. 1.3 зображений графік емпіричної функції розподілу, підрахованої за 150 спостереженнями, які були згенеровані генератором псевдовипадкових чисел зі стандартним нормальним розподілом (чорна лінія). Графік ф.р. цього розподілу зображено червоною лінією. Графік емпіричної ф.р. в цілому відображає форму справжньої ф.р., хоча у деяких місцях досить помітно від неї відхиляється. Зрозуміло, що на інших даних відхилення можуть бути в іншому місці і в інший бік. Тому важливо мати змогу визначати, наскільки великі відхилення характерні для емпіричних ф.р. при даному обсязі вибірки.

◀

У наступному прикладі розглянуто застосування емпіричних ф.р. у реальній прикладній задачі статистичного аналізу медичних даних, хоча тут для відображення на рисунку також використані модельовані дані.

**Приклад 1.1.2.** Коли розробляють нові вакцини, їх перевіряють на ефективність. Для цього проводиться "подвійне сліпе" дослідження. Вибирається група добровольців і кожному з них роблять щеплення або справжньою вакциною, що досліджується, або плацебо[2]. Хто прищеплений вакциною, а хто плацебо, не знають ні самі добровольці, ні лікарі, що проводять дослідження. Ця інформація відома тільки спеціальному наглядовому комітету, який оголошує її лише після завершення певної

[2]При плацебо-щепленні у процедурі щеплення замість справжньої вакцини використовують препарат, який не захищає від інфекції, але викликає такі ж зовнішні ефекти, як і вакцина - подразнення у місці уколу, підвищення температури, головний біль... Це робиться для того, щоб поведінка добровольців, які отримали плацебо, не відрізнялась від поведінки тих, хто справді був вакцинований. Іноді на роль плацебо використовується вакцина від якоїсь іншої хвороби, яка не захищає від тієї інфекції, з якою повинна боротись досліджувана вакцина. Скажімо, при дослідженні вакцини від коронавірусу, можна на роль плацебо взяти вакцину від сезонного грипу. Зрозуміло, що добровольців до початку експерименту попереджають, що саме вони можуть отримати у якості плацебо.

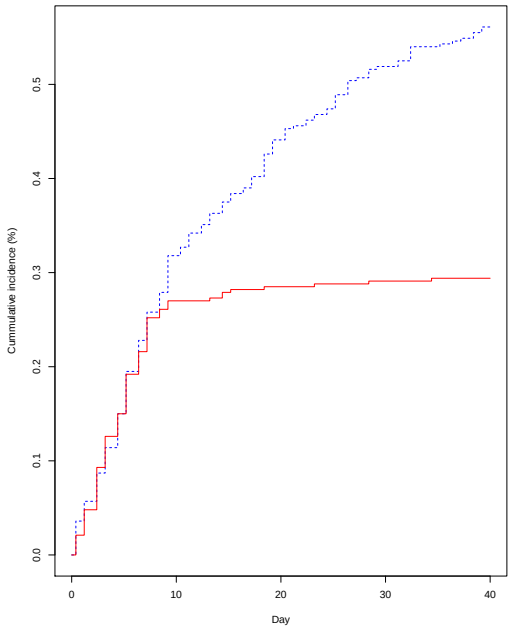



Рис. 1.2: Накопичений відсоток інфікованих. Червона лінія - вакциновані, синя лінія - контрольна вибірка.

стадії дослідження. Після проведення щеплення добровольці живуть звичайним для них життям, а якщо у них виявлена інфекція, від якої має захищати вакцина, це фіксується у протоколі дослідження.

Коли досягається певна кількість виявлених інфікованих (скажімо, 200 осіб) наглядовий комітет розкриває дані про те, хто був щеплений вакциною (ці особи складають основну вибірку) а хто отримав плацебо (контрольна вибірка). Для визначення ефективності вакцини проводять статистичний аналіз цих даних. Одним із способів їх графічного представлення є відображення накопиченої кількості інфікованих в обох вибірках на кожен день дослідження.

Приклад таких графіків можна побачити на рис. 1.2. Тут по горизонталі відкладено номер дня після щеплення, а по вертикалі відсоток добровольців, що були інфіковані за весь період від моменту щеплення до цього дня включно. (Відсоток задається від всього обсягу вибірки). Червона суцільна лінія відповідає вибірці тих, хто був вакцинований справ-

жньою вакциною, синя пунктирна — контрольній вибірці[3]. Помітно, що десь до десятого дня після щеплення кількість інфікованих зростає приблизно однаково для обох вибірок, а після цього для імунізованих вакциною зростання різко вповільнюється. У контрольній вибірці після десятого дня вповільнення не спостерігається. Це свідчить на користь того, що вакцина є ефективною і імунітет виробляється за десять днів після вакцинації.

Якщо позначити через $\xi_j$ — номер дня, в який був інфікований $j$-тий доброволець у контрольній вибірці, то синю криву можна вважати графіком емпіричної функції розподілу на інтервалі [0,40], побудованої за вибіркою $\xi_j$, $j = 1, \ldots, n$, де $n$ — обсяг контрольної вибірки. Єдиною особливістю є лише використання відсотків як одиниці вимірювання по вертикалі. Аналогічно — червона лінія є графіком емпіричної ф.р. для вибірки імунізованих.

Треба відмітити, що у цих даних є певні особливості. Деякі добровольці не будуть інфіковані ніколи, можливо, для них слід прийняти $\xi_j = +\infty$. Деякі добровольці можуть припинити участь у дослідженні, наприклад, зробивши собі щеплення іншою вакциною. У такому випадку вони випадають з поля нашого зору. При визначенні емпіричної ф.р. ми не враховували таких можливостей, вони будуть розглянуті у п. 1.2. ◀

Розглянемо тепер статистичні властивості $\hat{F}_n(x)$ як оцінки для $F(x)$.

Нагадаємо, що оцінка $\vartheta_n$ для невідомого параметра $\vartheta$ називається **незміщеною**, якщо $\mathsf{E}\,\vartheta_n = \vartheta$, тобто математичне сподівання оцінки дорівнює тому, що ми оцінюємо.

Легко бачити, що емпірична ф.р. кратної вибірки є незміщеною оцінкою теоретичної ф.р. одного спостереження з цієї вибірки:

$$\mathsf{E}\,\hat{F}_n(x) = \mathsf{E}\,\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{\xi_j < x\} = \frac{1}{n}\sum_{j=1}^{n} \mathsf{E}\,\mathbb{1}\{\xi_j < x\} = \frac{1}{n}\sum_{j=1}^{n} \mathsf{P}\{\xi_j < x\} = F(x).$$

Оцінка $\hat{\vartheta}_n$ називається **консистентною** (**строго консистентною**), якщо $\hat{\vartheta}_n \xrightarrow{P} \vartheta$ ($\hat{\vartheta}_n \to \vartheta \mod \mathsf{P}$) при $n \to \infty$. Тобто консистентна оцінка збігається до оцінюваного параметра при зростанні обсягу вибірки.

Оскільки у нашому випадку оцінюваний параметр є функцією, консистентність можна трактувати по різному. Можна розглядати "поточкову"

[3]Цей рисунок побудовано за умовними даними. Відповідний графік за справжніми даними для дослідження вакцини від covid-19 можна побачити у [24].

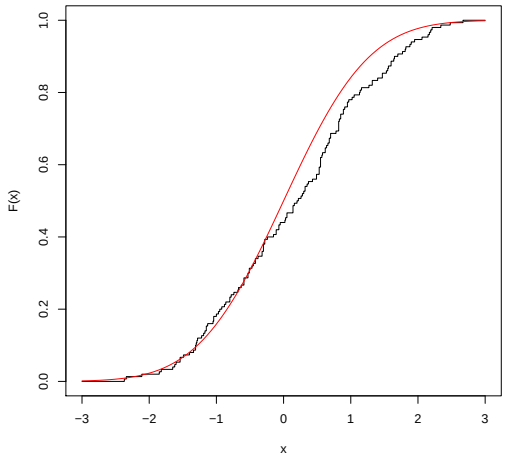



Рис. 1.3: Емпірична функція розподілу (чорна лінія) для 150 спостережень та теоретична ф.р. (червона лінія)

консистентність, тобто збіжність $\hat{F}_n(x)$ при фіксованому аргументі $x$ для всіх $x \in \mathbb{R}$. Або можна задати яку-небудь відстань у просторі функцій і розглядати збіжність як прямування до нуля відстані між $\hat{F}_n$ та $F$.

Зараз ми зупинимося на поточковій збіжності і, відповідно, поточковій консистентності. Зафіксуємо довільне $x \in \mathbb{R}$. Тоді $\eta_j = \mathbb{1}\{\xi_j < x\}$ це незалежні, однаково розподілені випадкові величини з математичним сподіванням $\mathsf{E}\,\eta_j = F(x)$. Тому, за підсиленим законом великих чисел,

$$\hat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \eta_j \to \mathsf{E}\,\eta_1 = F(x), \text{ при } n \to \infty \text{ м.н.} \tag{1.1}$$

Таким чином, емпірична ф.р. є поточково строго консистентною оцінкою справжньої ф.р. спостережень.

Крім поточкової збіжності інколи корисними бувають результати про рівномірну збіжність. Тут ми сформулюємо класичну теорему про рівномірну консистентність емпіричних функцій розподілу.

**Теорема 1.1.1.** ***(Глівенко — Кантеллі)***

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \to 0 \text{ } \textit{м.н. при } n \to \infty.$$

**Доведення.** Ми обмежимось розглядом випадку, коли $F$ — неперервна функція. Про доведення у загальному випадку див. [9], п.2 розділу 1.

Візьмемо довільне натуральне число $N$ і покладемо $\varepsilon_N = 1/N$. Оскільки $F$ — неперервна функція з множиною значень $[0,1]$, то існують $x_1, \ldots, x_{N-1}$, такі, що $F(x_i) = i\varepsilon_N$ для $i = 1, \ldots, N-1$. Покладемо формально $x_0 = -\infty$, $x_N = +\infty$, $F(x_0) = 0$, $F(x_N) = 1$. Тоді для $x \in [x_i, x_{i+1}]$, враховуючи, що функції $F$ і $\hat{F}_n$ неспадні, отримуємо

$$\hat{F}_n(x) - F(x) \le \hat{F}_n(x_{i+1}) - F(x_i) \le \hat{F}_n(x_{i+1}) - F(x_{i+1}) + \varepsilon_N$$

і

$$\hat{F}_n(x) - F(x) \ge \hat{F}_n(x_i) - F(x_{i+1}) \le \hat{F}_n(x_i) - F(x_i) - \varepsilon_N.$$

Тому

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \le \max_{i=0,\ldots,N} |\hat{F}_n(x_i) - F(x_i)| + \varepsilon_N.$$

За (1.1), $\hat{F}_n(x_i) - F(x_i) \to 0$ м.н. при $n \to \infty$. Тому існує подія $A_{i;N} \subseteq \Omega$ (тут $\Omega$ — простір елементарних подій, на якому визначені спостережувані дані), така, що $\mathsf{P}\{A_{i;n}\} = 1$ і для всіх $\omega \in A_{i;N}$ існує $n_{i;N}(\omega)$,

$$|\hat{F}_n(x_i) - F(x_i)| < \varepsilon_N,$$

для всіх $n > n_{i;N}(\omega)$.

Покладемо $A = \cap_{N=1,2\ldots} \cap_{i=0,\ldots,N} A_{i;n}$. Оскільки $A$ отримано перетином зілченної кількості подій що мають ймовірність 1, то $\mathsf{P}(A) = 1$. Покладемо $n_N(\omega) = \max_{i=0,\ldots,N} n_{i;N}(\omega)$. Тоді при виконанні події $A$,

$$\max_{i=0,\ldots,N} |\hat{F}_n(x_i) - F(x_i)| \le \varepsilon_N$$

для всіх $n > n_i(\omega)$, а отже і

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \le 2\varepsilon_N.$$

Оскільки $\varepsilon_N \to 0$ при $N \to \infty$, звідси випливає твердження теореми.

### 1.1.2 Асимптотична нормальність емпіричних ф.р. та довірчі інтервали

Для того, щоб акуратно використовувати оцінки, важливо не тільки знати, що вони збігаються до справжніх значень невідомого параметра, а й

вміти оцінювати швидкість такої збіжності. Це, зокрема, дає можливість будувати довірчі інтервали для параметрів та визначати обсяг вибірки, потрібний для досягнення заданої точності оцінювання.

Нагадаємо спочатку загальну техніку дослідження швидкості збіжності оцінок[4]. Нехай для невідомого (числового) параметра $\vartheta$ є консистентна оцінка $\hat{\vartheta}$ за даними $\mathbf{X}_n$. Консистентність $\hat{\vartheta}_n$ означає, що

$$\hat{\vartheta}_n \xrightarrow{P} \vartheta,$$

тобто із зростанням обсягу даних $n$ оцінка все ближче наближається до справжнього значення параметра. Зараз ми хочемо уточнити це твердження, щоб мати змогу сказати щось на зразок "при такому-то обсязі вибірки наша оцінка з такою-от ймовірністю буде відхилятись від справжнього значення не більше, ніж на стільки-то". Саме лише твердження про консистентність нам такої можливості не дає.

Тому ми розглянемо різницю $\hat{\vartheta}_n - \vartheta$ яка, зрозуміло, прямує до 0. Домножимо її на якусь (не випадкову) нормуючу послідовність $a_n$, що зростає до нескінченності:

$$\eta_n = a_n(\hat{\vartheta}_n - \vartheta).$$

При різних $a_n$ поведінка $\eta_n$ буде різною. Якщо обрати $a_n$, яка зростає повільно, то $(\hat{\vartheta}_n - \vartheta)$ буде прямувати до 0 швидше ніж $a_n$ до нескінченності і $\eta_n \to 0$. Якщо $a_n$ зростає швидко, то вона переважить прямування $(\hat{\vartheta}_n - \vartheta) \xrightarrow{P} 0$ і ми отримаємо $\eta_n \to \infty$. Посередині між цими крайнощами може знайтись така нормуюча послідовність $a_n$, що $\eta_n \to \eta_\infty$, де $\eta_\infty$ не є ні нулем, ні нескінченністю. У такому випадку $a_n$ називають правильною нормуючою послідовністю для оцінки $\hat{\vartheta}_n$ і кажуть, що збіжність $\hat{\vartheta}_n$ до $\vartheta$ має порядок $1/a_n$.

Точніше, ми казатимем, що $a_n$ - **правильна нормуюча послідовність** для $\hat{\vartheta}_n$, якщо

$$a_n(\hat{\vartheta}_n - \vartheta) \xrightarrow{W} \eta_n, \text{ при } n \to \infty,$$

де $\eta_\infty$ — деяка не нульова випадкова величина.

тут символ $\xrightarrow{W}$ позначає слабку збіжність випадкових величин (розподілів).

[4]Власне, не лише оцінок, а і будь-яких послідовностей.

Розподіл $\eta_\infty$ називають граничним розподілом (нормованої) оцінки $\hat{\vartheta}_n$.

Якщо $\eta_\infty$ є гауссовою випадковою величиною, то кажуть, що $\hat{\vartheta}_n$ є асимптотично нормальною оцінкою з нормуючою послідовністю $a_n$.

Тепер сформулюємо твердження про асимптотичну нормальність емпіричних функцій розподілу.

**Теорема 1.1.2.** *Нехай $\mathbf{X}_n = (\xi_1, \dots, \xi_n)$ — кратна вибірка з ф.р. $F$, $\hat{F}_n$ — емпірична ф.р., побудована за $\mathbf{X}_n$. Якщо $0 < F(x) < 1$, то*

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{W} N(0, \sigma^2(F(x))),$$

*де $\sigma^2(F(x)) = F(x)(1 - F(x))$.*

Таким чином, емпірична ф.р. є асимптотично нормальною з нормуючою послідовністю $\sqrt{n}$. Або можна сказати, що емпіричні ф.р. збігаються до справжніх ф.р. з порядком збіжності $1/\sqrt{n}$.

*Доведення* цієї теореми дуже просте. Помітимо, що

$$\eta_n = \sqrt{n}(\hat{F}_n(x) - F(x)) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \zeta_j,$$

де

$$\zeta_j = \mathbb{1}\{\xi_j < x\} - F(x).$$

Оскільки $\mathbf{X}_n$ — кратна вибірка з ф.р. $F$, то $\zeta_j$ — незалежні, однаково розподілені в.в., $\mathsf{E}\,\zeta_j = 0$, $\mathrm{D}\,\zeta_j = \sigma^2(F(x))$. Отже, за центральною граничною теоремою для незалежних, однаково розподілених доданків,

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \zeta_j \xrightarrow{W} N(0, \mathrm{D}\,\zeta_1).$$

Теорема доведена.

Подивимось, як можна скористатись цією теоремою для побудови довірчих інтервалів для значення функції розподілу у заданій точці.

Помітимо, що, оскільки $\sigma(t) = \sqrt{t(1-t)}$ є неперервною функцією, а $\hat{F}_n(x)$ — консистентна оцінка $F(x)$, то

$$\sigma(\hat{F}_n(x)) \xrightarrow{P} \sigma(F(x)).$$

Тому, враховуючи теорему 1.1.2 та лему Слуцького, отримуємо

$$\eta'_n = \frac{\sqrt{n}}{\sigma(\hat{F}_n(x))}(\hat{F}_n(x) - F(x)) \xrightarrow{W} \eta'_\infty \sim N(0,1). \tag{1.2}$$

Позначимо $\lambda_\alpha = Q^{N(0,1)}(1-\alpha)$. З (1.2) випливає, що

$$\lim_{n\to\infty} \mathsf{P}\{-\lambda_{\alpha/2} \le \eta'_n \le \lambda_{\alpha/2}\} = \mathsf{P}\{-\lambda_{\alpha/2} \le \eta'_\infty \le \lambda_{\alpha/2}\} = 1-\alpha.$$

Підставляючи сюди вираз для $\eta'_n$ з (1.2) і розв'язуючи нерівності відносно $F(x)$, отримуємо:

$$\lim_{n\to\infty} \mathsf{P}\{F_n^-(x) \le F(x) \le F_n^+(x)\} = 1-\alpha, \tag{1.3}$$

де

$$F_n^\pm(x) = \hat{F}_n(x) \pm \lambda_{\alpha/2}\sqrt{\frac{\sigma(\hat{F}_n(x))}{n}}.$$

Отже $F_n^-(x)$ і $F_n^+(x)$ — межі асимптотичного довірчого інтервалу рівня $\alpha$ для значення $F(x)$ за кратною вибіркою.

### 1.1.3 Метод емпіричної найбільшої вірогідності

Емпіричну функцію розподілу $\hat{F}_n(x)$ можна розглядати, як оцінку (узагальненого) методу моментів для невідомого параметра $F(x)$. Дійсно, ми підібрали моментну функцію $h_x(t) = \mathbb{1}\{t < x\}$ так, щоб $\mathsf{E}\, h_x(\xi_j) = F(x)$ і не роль оцінки взяли відповідний емпіричний момент:

$$\hat{h}_{x,n} = \frac{1}{n}\sum_{j=1}^{n} h_x(\xi_j) = \hat{F}_n(x).$$

(Див. п. 8.1 у [3] про узагальнений метод моментів для параметричного оцінювання).

Техніка методу моментів — потужний евристичний засіб для отримання нових оцінок. Але у параметричний статистиці більшою повагою користується метод найбільшої вірогідності. Зокрема, при виконанні певних умов регулярності, параметричні оцінки найбільшої вірогідності (ОНВ) є асимптотично найбільш точними (ефективними) оцінками. Як

ми побачимо далі, у непараметричному випадку ситуація складніша, хоча і у непараметричній статистиці ОНВ часто виявляються найбільш вдалим способом оцінювання.

У цьому підрозділі ми розберемось, як можна застосувати техніку найбільшої вірогідності у задачі оцінювання функцій розподілу за кратною вибіркою. Основні висновки я сформулюю одразу:

1. Безпосередньо метод найбільшої вірогідності, як він був сформульований для параметричного оцінювання (див. п. 8.3 у [3]) у цій задачі застосувати неможливо.

2. Можна створити певну модифікацію ОНВ, яка зветься оцінкою емпіричної найбільшої вірогідності, котра знову приводить до емпіричних ф.р. як оцінок для ф.р. у випадку кратної вибірки.

Навіщо ж тоді розбирати цю сумнівну модифікацію, якщо вона все одно не приводить ні до чого нового? Насправді ми потім побачимо, що емпіричний метод найбільшої вірогідності — дуже потужний спосіб отримання нетривіальних і якісних оцінок у більш складних задачах. А для того, щоб зрозуміти, як він працює, корисно почати з цього простого і порівняно легко зрозумілого випадку.

Пригадаємо, як визначаються звичайні оцінки методу найбільшої вірогідності. Нехай $\mathbf{X} \in \mathcal{X}$ — випадкові дані з розподілом $P_\vartheta(A) = \mathsf{P}\{\mathbf{X} \in A\}$, де $\vartheta \in \Theta$ — невідомий параметр, який потрібно оцінити. Припустимо, що існує така міра $\mu$ на просторі даних $\mathcal{X}$, відносно якої розподіли $P_\vartheta$ є абсолютно неперервними для всіх $\vartheta \in \Theta$. Це означає, що існує така функція $f^{\mathbf{X}}(x;\vartheta)$, для якої

$$P_\vartheta(A) = \int_A f^{\mathbf{X}}(x;\vartheta)\mu(dx)$$

для всіх вимірних $A \subseteq \mathcal{X}$. Функцію $f^{\mathbf{X}}(x;\vartheta)$ називають щільністю $\mathbf{X}$ відносно міри $\mu$ при значенні параметра $\vartheta$ (або у теорії міри — похідною Радона для $P_\vartheta$ по $\mu$).

Функцією вірогідності для параметра $\vartheta$ називають

$$L(\vartheta) = f^{\mathbf{X}}(\mathbf{X};\vartheta),$$

тобто це щільність розподілу даних в яку замість аргументу підставлені самі дані.

Оцінка методу найбільшої вірогідності $\hat\vartheta$ — це те значення невідомого

параметра, на якому досягається максимум функції вірогідності:

$$L(\hat{\vartheta}) = \sup_{t\in\Theta} L(t).$$

Повернемось до оцінювання ф.р. $F$ за кратною вибіркою $\mathbf{X} = (\xi_1, \dots, \xi_n)$. Оскільки $\xi_j$ — незалежні випадкові величини, щільність $f^{\mathbf{X}}(x; F)$ має розпадатись у добуток щільностей спостережень:

$$f(\mathbf{x}; F) = \prod_{j=1}^{n} f(x_j, F),$$

де $f$ — щільність розподілу одного спостереження $\xi_j$ відносно деякої міри $\mu_1$ на $\mathbb{R}$. (Тоді мірою $\mu$ буде $n$-кратний декартів добуток мір $\mu_1$ на $\mathcal{X} = \mathbb{R}^n$).

Щільність $f(x, F)$ має бути похідною Радона від розподілу $F$ по мірі $\mu_1$:

$$F(A) = \int_A f(x, F)\mu(dx).$$

Якщо обмежитись лише абсолютно неперервними розподілами, то на роль $\mu_1$ можна взяти міру Лебега. Якщо розглядати лише дискретні розподіли, зосереджені на фіксованій зліченній множині точок, на роль міри $\mu_1$ підійде лічильна міра на цій множині. Але не існує такої міри, відносно якої були б абсолютно неперервними всі взагалі ймовірнісні розподіли. Тому не можна визначити функцію вірогідності для $F$ у даній задачі оцінювання і, відповідно, не можна побудувати оцінку найбільшої вірогідності для $F$.

А тепер давайте побудуємо таку оцінку, але, звичайно, зробивши певні модифікації методу. По-перше, ми будемо розглядати не всі можливі функції розподілу $F$, а лише функції вигляду

$$F(x; \mathbf{w}) = \sum_{i=1}^{m} w_i \mathbb{1}\{x < x_i\},$$

де $\mathbb{X} = (x_1, \dots, x_m)$ — деякий скінченний набір набір чисел, $\mathbf{w} = (w_1, \dots, w_m)$ — набір вагових коефіцієнтів, такий, що

$$w_i \geq 0, \ \sum_{i=1}^{m} w_i = 1. \tag{1.4}$$

Таку ф.р. має випадкова величина $\eta$, що приймає лише значення з набору $\mathbb{X}$ з ймовірностями

$$\mathsf{P}\{\eta = x_i\} = w_i.$$

Для таких (дискретних) величин $\mathbb{X}$ інколи називають носієм $\eta$, а $\mathbf{w}$ — розподілом (рядом розподілу) $\eta$. В той же час, за нашою термінологією, розподіл $\eta$ — це міра на $\mathbb{R}$, що визначається як

$$F(A; \mathbf{w}) = \sum_{i:\ x_i \in A} w_i, \text{ для всіх вимірних } A \subseteq \mathbb{R}.$$

Міра $F(\cdot, \mathbf{w})$ має похідну Радона (щільність) відносно лічильної міри на $\mathbb{X}$

$$f(x; \mathbf{w}) = \begin{cases} w_1, & \text{якщо } x = x_1, \\ w_2, & \text{якщо } x = x_2, \\ & \dots \\ w_m, & \text{якщо } x = x_m, \\ 0 & \text{якщо } x \notin \mathbb{X}. \end{cases}$$

Отже, якщо вважати набір $\mathbb{X}$ фіксованим, а $\mathbf{w}$ — невідомим параметром, то функцію вірогідності для оцінки цього параметра по $\mathbf{X}$ можна записати, як

$$L(\mathbf{w}) = \prod_{j=1}^{n} f(\xi_j; \mathbf{w}).$$

Оцінкою найбільшої вірогідності $\hat{\mathbf{w}} = (\hat{w}_1, \dots, \hat{w}_m)$ для $\mathbf{w}$ має бути точка максимуму $L(\mathbf{w})$ по всіх наборах $\mathbf{w}$, що задовольняють (1.4).

Нехай для деякого $j$, $\xi_j \notin \mathbb{X}$, тобто вибраний нами носій розподілу включає у себе не всі точки спостережуваної вибірки. Тоді $L(\mathbf{w}) = 0$ для всіх $\mathbf{w}$. Користі з такої функції вірогідності немає жодної. Отже, як ні крути, а до носія потрібно включити всі спостережувані значення з вибірки. Що ж, включимо їх туди.

З іншого боку, припустимо, що у носій $\mathbb{X}$ потрапила точка $x_i$, яка не відповідає жодному спостереженню: $\xi_j \neq x_i$ для всіх $j$. Чому буде дорівнювати оцінка найбільшої вірогідності $\hat{w}_i$ для цієї точки? Припустимо, що $\hat{w}_i > 0$. Візьмемо точку $x_k$ з $\mathbb{X}$, таку, що $x_k = \xi_1$ і розглянемо набір $\mathbf{w}' = (w'_1, \dots, w'_m)$, такий, що

$$w'_i = 0,\ w'_k = \hat{w}_i + \hat{w}_k,\ w'_l = \hat{w}_l \text{ для } l \notin \{i, k\}.$$

Тоді

$$L(\mathbf{w}') > L(\hat{\mathbf{w}}),$$

оскільки всі співмножники у добутку, що задає $L(\mathbf{w}')$, такі самі, як у $L(\hat{\mathbf{w}})$, крім $w'_k$, котрий строго більше ніж $\hat{w}_k$. Але $\hat{\mathbf{w}}$ має бути точкою максимуму $L(\mathbf{w})$. Отже $\hat{w}_i = 0$, якщо точка $x_i$ не належить вибірці. А це означає, що немає рації включати таку точку у носій: все одно, для оцінки найбільшої вірогідності відповідна ймовірність потрапити у цю точку буде нульова.

Таким чином, на роль носія $\mathbb{X}$ слід обрати в точності вибірку $\mathbf{X}$. Ви помітили, що тут ми зробили нечесний викрутас? Начебто вирішили, що зафіксуємо носій і склали функцію вірогідності виходячи з того, що він фіксований. А потім стали вибирати носій в залежності від вибірки так, щоб ОНВ вийшла осмисленою. Так, саме тому те, що у нас вийде, буде не справжньою ОНВ, а оцінкою методу найбільшої емпіричної вірогідності (ОНЕВ).

Отже, ми вибрали $\mathbb{X} = \mathbf{X}$. Будемо вважати, що всі значення у $\mathbf{X}$ різні[5], тоді $n = m$ і кожне значення $w_i$ зустрічатиметься у функції вірогідності рівно один раз:

$$L(\mathbf{w}) = \prod_{j=1}^{n} w_j.$$

При якому $\mathbf{w}$ досягається максимум $L(\mathbf{w})$ за умови виконання (1.4)? Для розв'язання цієї задачі оптимізації зручніше спочатку перейти до логарифмічної функції вірогідності $l(\mathbf{w}) = \log L(\mathbf{w})$ і відмовитись від перевірки умов $w_i > 0$. Тобто ми шукатимем максимум

$$l(\mathbf{w}) = \sum_{j=1}^{n} \log(w_j)$$

за умови $\sum_{j=1}^{n} w_j = 1$.

Це класична задача умовної оптимізації з обмеженнями типу рівності, її можна розв'язати, використовуючи множники Лагранжа. Складемо функцію Лагранжа:

$$\mathcal{L}(\mathbf{w}) = \sum_{j=1}^{n} \log(w_j) + \lambda(\sum_{j=1}^{n} w_j - 1),$$

[5]Насправді, це чисто технічна умова, остаточний результат можна отримати і без неї.

(тут $\lambda$ — невизначений множник Лагранжа). Знайдемо стаціонарну точку функції Лагранжа, прирівнявши до 0 її похідні по $w_j$:

$$\frac{\partial L(\mathbf{w})}{\partial w_i} = \frac{1}{w_i} + \lambda = 0,\ i = 1 \ldots, n.$$

Звідси

$$w_i = -\frac{1}{\lambda}$$

і, з умови $\sum_{j=1}^{n} w_j = 1$ отримуємо

$$w_j = \hat{w}_j = \frac{1}{n}.$$

Легко переконатись, що ця стаціонарна точка відповідає саме максимуму функції $L(\mathbf{w})$. Отже $F(x; \hat{\mathbf{w}})$ є оцінкою методу найбільшої емпіричної вірогідності. Зрозуміло, що $F(x; \hat{\mathbf{w}})$ це звичайна емпірична функція розподілу, побудована за вибіркою $\mathbf{X}$. Хоча ми отримали цей результат для випадку, коли всі спостереження у вибірці різні, він залишається у силі і тоді, коли деякі значення можуть бути однакові:

$$F(x; \hat{\mathbf{w}}) = \hat{F}_n(x).$$

## 1.2 Аналіз виживання

### 1.2.1 Вибірка цензурована з права

У попередньому підрозділі ми вважали, що для кожного спостережуваного об'єкта відоме точне значення його характеристики, що досліджується. Але досить часто зустрічаються ситуації, коли для деяких об'єктів ми спостерігаємо не зовсім те, що нас цікавить. Відповідно, для дослідження таких даних застосовуються спеціальні статистичні техніки. Важливим прикладом такої техніки є аналіз виживання. Його методи застосовуються у багатьох предметних областях, зокрема у страховій справі та у аналізі надійності технічних приладів.

У класичній моделі аналізу виживання розглядаються величини, які називають "тривалостями життя" (людини) або "моментами відмови" (приладу)[6]. Їх трактують як випадкові величини, що можуть приймати лише

[6]Англійською мовою "аналіз виживання" — survival analysis, момент відмови — failure time

невід'ємні значення. У багатьох застосуваннях ці величини дійсно є довжинами проміжків часу між певними подіями у "житті" досліджуваного об'єкта. Наприклад, це може бути кількість годин, які прилад пропрацював до відмови, або тривалість часу між операцією, яку зробили пацієнту і моментом, коли у нього стався рецидив хвороби (інтервал ремісії). Ми будемо позначати досліджувану тривалість життя літерою $T$, коли описуємо теоретичну модель досліджуваного явища і $T_j$ коли йдеться про тривалість життя $j$-того досліджуваного об'єкта.

При зборі статистичних даних про такі характеристики часто виникає проблема, пов'язана з тим, що деякі спостережувані об'єкти зникають з поля зору дослідника раніше, ніж для них відбувається та подія, до якої має тривати $T$. Так, у прикладі 1.1.2 ми розглядали довжину інтервалу між щепленням добровольця і моментом, коли він буде інфікований певною хворобою. Якщо тривалість дослідження невелика, можна сподіватись, що практично всі добровольці, які отримали щеплення, будуть спостерігатись до його закінчення. Але, коли дослідження триває багато років, частина добровольців може вибути з нього у той чи інший час з причин, не пов'язаних із досліджуваною хворобою, наприклад — переїхати в іншу країну. Така подія зветься "цензуруванням". Момент цензурування[7] ми будемо позначати $C$ ($C_j$ — для $j$-того об'єкта). Тобто у цьому прикладі $C_j$ це час, який пройшов від моменту щеплення $j$-того добровольця до моменту, коли він випав з поля зору дослідника.

Таким чином, реально для досліджуваного об'єкта ми спостерігаємо або тривалість життя $T$, або момент $C$, коли об'єкт був відцензурований, причому для цензурованих об'єктів ми знаємо, що $T > C$. Цензурована вибірка складається зі значень $\xi_j = \min(T_j, C_j)$ (цензуроване спостереження) і $\delta_j = \mathbb{1}\{T_j \le C_j\}$ (індикатор відсутності цензурування у $j$-тому спостереженні), $j = 1, \dots, n$. За цим набором значень потрібно оцінити функцію розподілу моменту відмови $T$. Такі набори даних називають вибіркою з випадковим цензуруванням з права. У цій книзі ми не розглядаємо інших прикладів цензурування, тому будемо казати просто "цензурована вибірка".

Далі ми вважатимем, що всі випадкові величини $T_j$, $C_j$, $j = 1, \dots, n$ є незалежними в сукупності, моменти відмови $T_j$ мають (невідому) функцію розподілу $F(x) = \mathsf{P}\{T < x\}$, а цензори $C_j$ — функцію розподілу $G(x) = \mathsf{P}\{C < x\}$ (також невідому).

[7] Інколи момент цензурування називають просто "цензор".

### 1.2.2 Оцінка Каплана-Мейера

Щоб побудувати оцінку для $F$, введемо додаткові позначення. Позначимо $S(x) = \mathsf{P}\{T \geq x\} = 1 - F(x)$ — ймовірність дожити до моменту $x$ без відмови. Цю функцію називають функцією виживання[8].

Припустимо спочатку, що відмови можуть відбуватись лише у строго фіксовані моменти часу[9] $0 < t_1 < \cdots < t_K$. Покладемо $t_0 = 0$ (я буду вважати, що в момент $t_0$ відмова неможлива). Позначимо

$$q_k = \mathsf{P}\{T > t_k \mid T \geq t_k\}$$

— умовна ймовірність "пережити" момент $t_j$, за умови, що прилад дожив до цього моменту. Нехай $k(x)$ — найбільше $k$, таке, що $t_k < x$. Тоді

$$\begin{aligned} S(x) =& \mathsf{P}\{T \geq x\} = \mathsf{P}\{T > t_{k(x)}\} = \mathsf{P}\{T > t_{k(x)} \mid T \geq t_{k(x)}\}\,\mathsf{P}\{T \geq t_{k(x)}\} \\ =& q_{k(x)}\,\mathsf{P}\{T > t_{k(x)-1}\}. \end{aligned} \tag{1.5}$$

Остання рівність виконується тому, що на інтервалі між $t_{k(x)-1}$ і $t_{k(x)}$ відмова приладу неможлива, отже $\mathsf{P}\{T \geq t_{k(x)}\} = \mathsf{P}\{T > t_{k(x)-1}\}$. (Всі, хто пережив $t_{k-1}$ доживуть до $t_k$).

Повторюючи (1.5) рекурентно, отримуємо

$$S(x) = q_{k(x)}\,\mathsf{P}\{T > t_{k(x)-1}\} = q_{k(x)}q_{k(x)-1}\,\mathsf{P}\{T > t_{k(x)-2}\} = \cdots = \prod_{k:t_k<x} q_k. \tag{1.6}$$

Отже, для оцінки $S(x)$, а з ним і $F(x) = 1 - S(x)$, досить оцінити $q_k$, $k = 1, \ldots, K$.

Для оцінювання ймовірності $q_k$ скористаємось відповідною відносною частотою. Помітимо що $1 - q_k$ це ймовірність відмови в момент $t_k$, якщо прилад дожив до цього моменту. Тобто ця величина показує, наскільки

[8] У аналізі виживання функцією виживання часто називають ймовірність пережити момент $x$, тобто $\mathsf{P}\{T > x\}$, а функцію розподілу визначають як $\mathsf{P}\{T \leq x\}$. У цій книзі ми дотримуємось позначень які є більш прийнятими у теорії ймовірностей. Зрозуміло, що відмінність буде пов'язана лише з тим, з якої сторони ці функції є неперервними у точці розриву.

[9] На практиці можуть зустрічатись дані, для яких таке припущення є цілком природним. Наприклад, якщо відмова приладу може виникнути лише у момент його ввімкнення.

часто в середньому відбуваються відмови у момент $t_k$ серед приладів, що дожили до $t_k$. Підрахуємо

$$d_k = \#\{j : \ \xi_j = t_k, \delta_j = 1\}$$

— кількість приладів, що, як ми точно знаємо, відмовили у момент $t_k$, і

$$n_k = \#\{j : \ \xi_j \geq t_k\}$$

— кількість приладів, які "дожили" до моменту $t_k$.

Природною оцінкою для $1 - q_k$ буде $d_k/n_k$ — відносна частота відмов у $k$-тий момент серед приладів, що були у наявності на цей момент. Відповідно, оцінка для $q_k$ —

$$\hat{q}_k = 1 - \frac{d_k}{n_k},$$

а для $S(x)$ —

$$\hat{S}(x) = \prod_{k:t_k<x} \hat{q}_k = \prod_{k:t_k<x} \left(1 - \frac{d_k}{n_k}\right). \tag{1.7}$$

А тепер перейдемо до загального випадку, коли моменти відмови можуть бути довільними додатними числами. У цьому випадку ми включимо у набір $t_k$, $k = 1, \ldots, K$ всі значення $\xi_j$, яким відповідає $\delta_j = 1$ (тобто всі ті моменти часу, коли ми спостерігали реальні відмови, а не значення цензора). Ці значення $\xi_j$ розташуємо у порядку зростання — це і буде наш набір $t_k$, $k = 1, \ldots, K$. Оцінка для функції виживання визначається і в цьому випадку за (1.7), а оцінка для функції розподілу $F$ —

$$\hat{F}_n^{KM}(x) = 1 - \prod_{k:t_k<x} \left(1 - \frac{d_k}{n_k}\right). \tag{1.8}$$

Функцію $\hat{F}_n^{KM}(x)$ називають оцінкою Каплана-Мейера для функції розподілу за цензурованою вибіркою.

Зокрема, якщо припустити, що всі спостережувані значення $\xi_j$ є різними, то оцінку Каплана-Мейера можна записати у більш простому вигляді. Розглянемо цензуровану вибірку як набір пар $(\xi_j, \delta_j)$ і впорядкуємо ці пари по зростанню $\xi_j$. Будемо позначати отриману послідовність пар $(\xi_{[j]}, \tilde{\delta}_j)$, тобто $\xi_{[1]} < \xi_{[2]} < \cdots < \xi_{[n]}$ це варіаційний ряд для $\xi_j$, а $\tilde{\delta}_j$ — це

значення індикаторів відсутності цензурування $\delta_j$, переставлених разом з $\xi_j$ при впорядкуванні вибірки. Тоді

$$\hat{F}_n^{KM}(x) = 1 - \prod_{j:\xi_{[j]}<x} \left(1 - \frac{\tilde{\delta}_j}{n-j+1}\right). \tag{1.9}$$

Дійсно, оскільки ми вважаємо всі спостереження різними, то $d_j = \tilde{\delta}_j$ якщо покласти $t_j = \xi_{[j]}$. При цьому $n_j$ — кількість об'єктів, що "дожили" до моменту $t_j$ починається з $n$ для найменшого моменту $t_1$ і зменшується на 1 при переході від $j$ до $j+1$ — кожного разу з вибірки вибуває[10] один об'єкт. Отже $n_j = n - j + 1$. Це і дає формулу (1.9).

Е.Л. Каплан та П. Мейер отримали свою оцінку саме виходячи з міркувань, подібних до наведених вище. Звичайно, їх не можна назвати строгими, але вони досить прості і допомагають запам'ятати, як влаштована ця оцінка. З іншого боку, оцінку Каплана-Мейера можна також отримати, використовуючи метод емпіричної найбільшої вірогідності. Подивимося, як це робиться.

### 1.2.3 Оцінка Каплана-Мейера як оцінка найбільшої емпіричної вірогідності

Схема методу емпіричної найбільшої вірогідності у застосуванні до кратної вибірки описана у п. 1.1.3. Застосуємо його до нашої задачі оцінювання функції розподілу $F$ за вибіркою, цензурованою з права. Як і у п. 1.1.3 почнемо з того, що обмежимо розгляд допустимих функцій $F$ східчастими функціями, що можуть мати стрибки лише у точках з фіксованого набору точок $\mathbb{T}$ (носія ф.р.):

$$F(x) = F(x; \mathbf{w}) = \sum_{k=1}^{K} w_k \mathbb{1}\{t_k < x\}.$$

(Для відповідності нашим попереднім позначенням, тепер точки носія $F$ позначені $t_1 < t_2 < \ldots, t_K$). Тут $\mathbf{w} = (w_1, \ldots, w_K)$, $w_k \geq 0$, $\sum_{k=1}^{K} w_k = 1$. Якщо $T$ має ф.р. $F(x; \mathbf{w})$, то $w_k = \mathsf{P}\{T = t_k\}$ — ймовірність відмови (загибелі) у момент $t_k$.

[10] Або дає відмову, або відцензуровується, в усякому випадку, випадає з поля зору при збільшенні $x$.

Знайдемо тепер функцію вірогідності для оцінювання $F(x)$ за цензурованою вибіркою $(\xi_j, \delta_j)_{j=1}^n$. Як і раніше, функцію вірогідності ми трактуємо як щільність відносно лічильної міри на множині можливих значень даних (тобто ймовірність потрапити у задану точку), в яку замість аргументу щільності підставляються спостережувані дані. Оскільки спостереження незалежні при різних $j$, функція вірогідності за всією вибіркою буде дорівнювати добутку функцій вірогідності для окремих спостережень.

Отже, зафіксуємо значення невідомого параметра $\mathbf{w}$ і припустимо, що досліджувані моменти відмови $T_j$ мають ф.р. $F(x;\mathbf{w})$. Щільність їхнього розподілу відносно лічильної міри — $\mathsf{P}\{T_j = t_k\} = w_k$. Розподіл цензора ми позначили $G(x)$. Будемо вважати, що він також має (невідому нам) щільність відносно лічильної міри — $g(t_k) = g_k = \mathsf{P}\{C_j = t_k\}$. Ми спостерігаємо пару $(\xi_j, \delta_j)$, де $\xi_j \in \mathbb{T}$, $\delta_j \in \{0,1\}$. Отже щільність спостереження треба рахувати відносно лічильної міри на декартовому добутку $\mathbb{T} \times \{0,1\}$ і в точці $(t_k, z) \in \mathbb{T} \times \{0,1\}$ вона буде дорівнювати

$$f(t_k, z) = \mathsf{P}\{(\xi_j, \delta_j) = (t_k, z)\}.$$

(а в усіх точках, що не належать $\mathbb{T} \times \{0,1\}$ вона нульова). Розглянемо окремо випадки $z = 0$ і $z = 1$:

$$f(t_k, 1) = \mathsf{P}\{\min(T_j, C_j) = t_k, T_j \le C_j\} = \mathsf{P}\{T_j = t_k, C_j \ge t_k\}$$

$$= \mathsf{P}\{T_j = t_k\}\,\mathsf{P}\{C_j \ge t_k\} = w_k \sum_{i=k}^{K} g_i.$$

$$f(t_k, 0) = \mathsf{P}\{\min(T_j, C_j) = t_k, T_j > C_j\} = \mathsf{P}\{C_j = t_k, T_j > t_k\}$$

$$= \mathsf{P}\{C_j = t_k\}\,\mathsf{P}\{T_j > t_k\} = g_k \sum_{i=k+1}^{K} w_i.$$

(Тут ми скористались незалежністю $T_j$ і $C_j$).

Функція вірогідності за всіма даними матиме вигляд

$$L(\mathbf{w}) = \prod_{j=1}^{n} f(\xi_j, \delta_j).$$

Зрозуміло, що якщо для деякого $j$, $\xi_j \notin \mathbb{T}$, то $L(\mathbf{w}) = 0$ для всіх $\mathbf{w}$. Отже, щоб отримати змістовну оцінку, потрібно включити всі $\xi_j$ у набір $\mathbb{T}$.

Чи є сенс включати у $\mathbb{T}$ якісь інші точки крім $\xi_j$? Ті ж міркування, що у п. 1.1.3 показують, що між 0 і $\xi_{[n]} = \max_j \xi_j$ немає рації вставляти у носій точки крім вибіркових. Але, коли $\tilde{\delta}_n = 0$ (тобто найбільше спостереження у вибірці є значенням цензора) то це означає, що насправді було зафіксовано спостереження в якому $T_j$ є більшим, ніж $\xi_{[n]}$. Якщо ми не додамо у $\mathbb{T}$ якусь точку, в котрій могло б опинитись це значення $T_j$, то відповідна функція вірогідності буде тотожно рівною нулеві. Тому до набору $\mathbb{T}$ крім елементів вибірки додають іще одну умовну точку $t_{n+1} = +\infty$ (насправді це може бути будь-яке значення, більше ніж $\xi_{[n]}$, але ми не маємо інформації для того, щоб сказати яке саме).

Отже ми вибираємо $t_0 = 0$, $t_j = \xi_{[j]}$, $j = 1, \dots, n$, $t_{n+1} = +\infty$. Тоді можна записати

$$L(\mathbf{w}) = \tilde{L}(\mathbf{w}) Z_G,$$

де

$$\tilde{L}(\mathbf{w}) = \prod_{j=1}^{n} w_j^{\tilde{\delta}_j} \times \prod_{j=1}^{n} \left( \sum_{l=j+1}^{n} w_l \right)^{1-\tilde{\delta}_j},$$

а $Z_G$ це добуток усіх тих множників у $L(\mathbf{w})$, які залежать від функції розподілу цензора $Z_G$ і не залежать від $\mathbf{w}$. Зрозуміло, що точка максимуму $L(\mathbf{w})$ по $\mathbf{w}$ така сама, як і у $\tilde{L}(\mathbf{w})$, тому оцінку методу емпіричної найбільшої вірогідності можна шукати як точку максимуму $\tilde{L}(\mathbf{w})$. Максимум потрібно шукати з урахуванням обмежень: $w_j \geq 0$, $\sum_{j=1}^{n+1} w_j = 0$. Безпосередньо розв'язати таку задачу не просто.

Але вона спрощується, якщо зробити "репараметризацію", а саме задати функцію розподілу $F(x, \mathbf{w})$ в.в. $T$ використовуючи не безумовні ймовірності $w_k = \mathsf{P}\{T = t_k\}$ (ймовірність відмови у момент $t_k$), а умовні $q_k = \mathsf{P}\{T > t_k \mid T \geq t_k\}$ (ймовірність пережити $t_k$, якщо прилад не відмовив до моменту $t_k$).

Як ми встановили у (1.6),

$$\mathsf{P}\{T \geq t_k\} = \prod_{l=1}^{k-1} q_l, \tag{1.10}$$

отже

$$\begin{aligned} w_k = \mathsf{P}\{T = t_k\} &= \mathsf{P}\{T = t_k \mid T \geq t_k\} \cdot \Pr\{T \geq t_k\} \\ &= (1 - \mathsf{P}\{T > t_k \mid T \geq t_k\}) \cdot \Pr\{T \geq t_k\} = (1 - q_k) \prod_{l=1}^{k-1} q_l. \end{aligned} \tag{1.11}$$

Підставляючи цей вираз у формулу для $\tilde{L}(\mathbf{w})$, з урахуванням того, що $\sum_{l=j+1}^{n} w_l = \mathsf{P}\{T \geq t_{j+1}\} = \prod_{l=1}^{j} q_l$, отримуємо

$$\tilde{L}(\mathbf{w}) = \prod_{j=1}^{n} \left( (1-q_j) \prod_{l=1}^{j-1} q_l \right)^{\tilde{\delta}_j} \prod_{j=1}^{n} \left( \prod_{l=1}^{j} q_l \right)^{1-\tilde{\delta}_j}.$$

Цей вираз являє собою добуток співмножників вигляду $(1-q_k)$ або $q_k$, $k = 1, \ldots, n$. Підрахуємо, скільки разів входить у нього кожен співмножник.

З $(1-q_k)$ зовсім просто — він з'являється один раз, якщо $\tilde{\delta}_k = 1$, і не з'являється, якщо $\tilde{\delta}_k = 0$. Отже, кількість $(1-q_k)$ у добутку $\tilde{L}(\mathbf{w})$ дорівнює $\tilde{\delta}_k = d_k$, де $d_k = \#\{j : \xi_j = t_k, \delta_j = 1\}$ — кількість приладів, що відмовили у момент $t_k$.

Щоб підрахувати кількість $q_k$ у добутку, помітимо, що цей множник з'являється лише у тих випадках, коли $j$-тий прилад пережив $k$-тий момент можливої відмови. Таких приладів буде $n_j - d_j$, де $n_k = \#\{j : \xi_j \geq t_k\}$ — кількість приладів, що дожили до $t_k$.

Остаточно отримуємо

$$\tilde{L}(\mathbf{w}) = \prod_{k=1}^{n} (1-q_k)^{d_k} (q_k)^{n_k - d_k}.$$

Нам потрібно максимізувати цей вираз по $q_k$, $k = 1, \ldots, n$, причому на $q_k$ ми маємо лише обмеження $0 \leq q_k \leq 1$, тобто немає обмежень, які зв'язували б $q_k$ при різних $k$. Тому ми можемо максимізувати кожен множник

$$(1-q_k)^{d_k} (q_k)^{n_k - d_k}$$

по $q_k$ окремо. Легко бачити, що максимум цього виразу досягається при

$$\hat{q}_k = \frac{n_k - d_k}{n_k} = 1 - \frac{d_k}{n_k}.$$

Тепер, підставляючи $\hat{q}_k$ замість $q_k$ у формулу для $w_k$ (1.11) отримуємо оцінки методу емпіричної вірогідності $\hat{w}_k$ для $\hat{w}_k$. Враховуючи те, що формула (1.11) еквівалентна (1.10), маємо

$$F(x, \hat{\mathbf{w}}) = \hat{F}_n^{KM}(x) = 1 - \prod_{k: t_k < x} \left( 1 - \frac{d_k}{n_k} \right),$$

тобто оцінка емпіричної найбільшої вірогідності для функції розподілу за цензурованою вибіркою є оцінкою Каплана-Мейера.

Ми отримали це твердження за припущення, що всі спостереження $\xi_j$ у вибірці є різними. Але воно є правильним і у загальному випадку.

### 1.2.4 Асимптотика оцінок Каплана-Мейера

Для оцінок Каплана-Мейера можна отримати результати про асимптотичну поведінку, аналогічні тим, що ми розглядали у п. 1.1.2 для емпіричних функцій розподілу за кратною вибіркою. На жаль, доведення їх досить складне[11], тому тут ми їх сформулюємо без доведення. Позначення у теоремах цього підрозділу ті самі, що були введені у п. 1.2.1,1.2.2.

**Теорема 1.2.1.** *Нехай $G(x) < 1$. Тоді $\hat{F}_n^{KM}(x)$ — строго консистентна оцінка для $F(x)$.*

Умова $G(x) < 1$ є в цій теоремі є практично необхідною. Дійсно, якщо $G(x) = 1$, то це означає, що всі значення цензора $C_j$ є меншими ніж $x$. Оскільки спостереження $\xi_j = \min(T_j, C_j)$, то серед них не може бути значень $T_j$, які були б більшими або рівними $x$. Відповідно ми не можемо оцінити ймовірність такої події — не тільки оцінкою Каплана — Мейера, а й бідь-якою іншою оцінкою.

**Теорема 1.2.2.** *Нехай $G(x) < 1$, $0 < F(x) < 1$, $F$ і $G$ — неперервні функції. Тоді*

$$\sqrt{n}(\hat{F}_n^{KM}(x) - F(x)) \xrightarrow{W} N(0, \sigma_{KM}^2(x)),$$

*де*

$$\sigma_{KM}^2(x) = (1 - F(x))^2 \int_0^x \frac{(1 - G(t))F(dt)}{((1 - F(t))(1 - G(t)))^2}.$$

Ця теорема показує, що оцінка Каплана-Мейера є асимптотично нормальною з нормуючою послідовністю $\sqrt{n}$, тобто вона збігається до оцінюваної функції розподілу зі швидкістю $1/\sqrt{n}$, характерною для параметричних оцінок.

Щоб скористатись асимптотичною нормальністю для побудови асимптотичного довірчого інтервалу[12], потрібно мати оцінку для $\sigma_{KM}^2(x)$ за

[11] Див., наприклад, [7], підрозділ IV.3.1.

[12] Аналогічно тому, як ми це робили для кратних вибірок у п. 1.1.2.

спостереженнями. Таку оцінку можна отримати, оцінивши $F$ і $G$, наприклад, оцінками Каплана – Мейера і підставивши ці оцінки у формулу для $\sigma^2_{KM}(x)$ замість справжніх значень. Але є інша оцінка, яка дозволяє одразу оцінити $\sigma^2_{KM}(x)/n$, тобто величину, що приблизно відповідає дисперсії $\hat{F}^{KM}_n(x)$. Ця оцінка має вигляд

$$V_n(x) = (1 - \hat{F}^{KM}_n(x))^2 \sum_{k:\ t_k \le t} \frac{d_k}{n_k(n_k - d_k)}. \tag{1.12}$$

Формулу (1.12) називають формулою Грінвуда. Можна довести, що в умовах теореми 1.2.2,

$$nV_n(x) \xrightarrow{P} \sigma^2_{KM}(x).$$

Якщо у вибірці немає однакових значень $\xi_j$, то формулу Грінвуда можна записати у більш зручній формі:

$$V_n(x) = (1 - \hat{F}^{KM}_n(x))^2 \sum_{j:\ \xi_{[j]} \le x} \frac{\tilde{\delta}_j}{(n+1-j)(n+1-j-\tilde{\delta}_j)}. \tag{1.13}$$

З використанням оцінки Каплана — Мейера та формули Грінвуда можна побудувати асимптотичний довірчий інтервал для $F(x)$:

$$\lim_{n\to\infty} \mathsf{P}\{F(x) \in [F^{-KM}_n(x), F^{+KM}(x)]\} = 1 - \alpha,$$

де

$$F^{\pm KM}_n(x) = \hat{F}^{KM}_n(x) \pm \lambda_{\alpha/2}\sqrt{V_n(x)}. \tag{1.14}$$

**Приклад 1.2.1.** У прикладі 1.1.2 ми розглянули аналіз даних дослідження ефективності вакцини. Як було відмічено у цьому прикладі, реальні дані таких досліджень часто є цензурованими, оскільки частина добровольців у якийсь момент може відмовитись від продовження участі у експерименті, або випасти з поля зору дослідника з якихось причин. Відповідно, для такого добровольця ми можемо знати лише період часу від вакцинації, до того моменту, коли він вийшов з дослідження, і той факт, що до цього моменту він “не встигнув” захворіти.

Для оцінювання функцій розподілу довжини інтервалу $T$ між вакцинацією (або отриманням плацебо) та моментом виявлення захворювання у цьому випадку можна використовувати оцінку Каплана — Мейера. Величину $T$ можна трактувати як досліджуваний момент відмови, а цензором є момент виходу добровольця з дослідження. На рис. 1.4 зображені

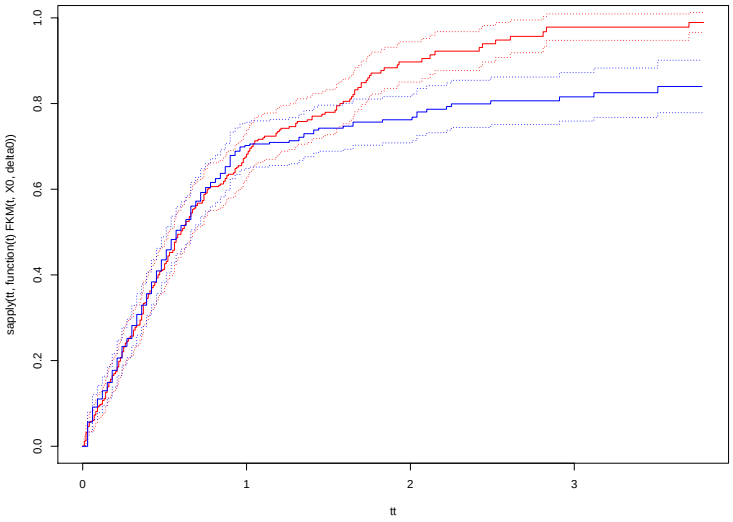



Рис. 1.4: Оцінки Каплана — Мейера для двох вибірок з довірчими смугами. Синя суцільна лінія - вакциновані, червона - контрольна вибірка. Межі довірчих смуг відмічені пунктирними лінями відповідних кольорів.

графіки оцінок Каплана — Мейера по двох модельованих наборах даних. (Можна умовно вважати, що червона лінія відповідає контрольній вибірці, а синя — добровольцям, вакцинованим досліджуваною вакциною. По горизонталі відкладено час від початку вакцинації в умовних одиницях, скажімо, у місяцях). Пунктирні лінії відповідних кольорів відповідають межам довірчих інтервалів, розрахованих за (1.14) для рівня значущості $\alpha = 0.05$.

На рисунку помітно, що до значення часу $t = 1$ оцінки по обох вибірках мало відрізняються одна від одної, а при більших $t$ починають розходитись. Але довірчі інтервали для справжніх функцій розподілу продовжують перетинатись, тобто при цих значеннях $t$ не можна впевнено казати, що ці функції дійсно є різними. Лише при $t \approx 1.7$ довірчі ітервали перестають перетинатись. Відповідно, можна сказати, що захисні властивості вакцини починають виявлятись лише через 1.7 місяця після вакцинації.

Варто відмітити, що побудовані тут довірчі інтервали не є одночасними[13], тому такий підхід до аналізу функцій розподілу не є строгим. Але він дає певне уявлення про те, наскільки можна покладатись на отримані оцінки. ◀

## 1.3 Оцінювання функцій розподілу за зміщеною вибіркою

### 1.3.1 Зміщена вибіркова процедура

У статистиці часто спостережувані об'єкти можна трактувати як вибрані з великої сукупності об'єктів, які підлягають дослідженню. Такий вибір здійснюється за певними правилами і називається вибірковою процедурою або процедурою відбору (sampling). Вся сукупність об'єктів що досліджуються називається генеральною сукупністю або популяцією (population), а відібраний набір спостережуваних об'єктів — вибіркою (sample). Скажімо, при проведенні дослідження ефективності вакцин, як у прикладі 1.1.2, вибіркою є відібрані добровольці, популяцією — все

[13]Про одночасні та неодночасні довірчі інтервали та їхнє використання для перевірки гіпотез див. [3], п. 9.4

населення країни, де проводяться дослідження[14], а вибірковою процедурою — схема відбору добровольців.

Інший приклад — популяція риб певного виду, що живуть у озері, вибірка — риби, що були виловлені для дослідження, а вибіркова процедура — спосіб вилову. Статистик, як правило, має справу з певною числовою або якісною характеристикою (змінною), яку мають всі досліджувані об'єкти і цікавиться властивостями розподілу цієї змінної у популяції. Скажімо, змінною може бути довжина рибини, а метою дослідження — опис розподілу довжин риб даного виду, що живуть у даному озері. При цьому висновки про розподіл змінної у популяції потрібно зробити на основі даних про значення змінної у об'єктів з вибірки — ми можемо виміряти довжини лише виловлених риб, а не тих, що плавають в озері.

Найкраще вибірка відтворює розподіл популяції при простому незалежному відборі, коли об'єкти з популяції мають однакові шанси потрапити до вибірки незалежно один від одного а також від значення досліджуваної змінної. Але не завжди вдається організувати вибіркову процедуру так, щоб забезпечити такий результат. Наприклад, при вилові риби рибальською мережею (неводом, ятером) з великими вічками, риба малої довжини має більше шансів вислизнути крізь вічко, ніж велика. Отже, якщо вибірка риб отримана ловом мережею, у ній відсоток довгих риб буде більшим, порівняно з їхнім відсотком у популяції, що живе в озері. Такі вибіркові процедури називають зміщеними.

Точніше, ми будемо називати вибіркову процедуру зміщеною, якщо для об'єкта з популяції ймовірність потрапити до вибірки залежить від значення тієї його змінної, яка досліджується. Існує багато різних зміщених вибіркових процедур, які можуть по різному "спотворювати" популяційний розподіл. Тут ми обмежимось розглядом простішого випадку, коли досліджувана популяція є дуже великою (практично нескінченною) і процедура організована так, що відбір одного об'єкта з неї практично не впливає на можливості відбору інших. Тому спостережувані значення досліджуваної змінної у відібраних об'єктів можна вважати незалежними між собою. Нехай $O$ — деякий об'єкт з популяції. Позначимо $s(O)$ — індикатор того, що цей об'єкт потрапить до вибірки при даний процедурі

[14] У таких дослідженнях можливі і інші інтерпретації того, що є популяцією. Наприклад, це може бути сукупність всіх людей певного віку, або всіх, хто потенційно може бути інфікований даною хворобою, тощо.

відбору:

$$s(O) = \begin{cases} 1 & \text{Якщо } O \text{ потрапив до вибірки,} \\ 0 & \text{Якщо } O \text{ не потрапив до вибірки.} \end{cases}$$

Якщо відбір є випадковим, то $s(O)$ — випадкова величина (випадкова функція від $O$). Позначимо $\xi(O)$ значення досліджуваної змінної $\xi$ у об'єкта $O$. Ми будемо вважати, що ймовірність для $O$ потрапити до вибірки залежить від значення $\xi(O)$ і її можна описати у такому вигляді:

$$\mathsf{P}\{s(O) = 1 \mid \xi(O) = t\} = cb(t), \tag{1.15}$$

де $w : \mathbb{R} \to [0, 1]$ — деяка функція, $c > 0$ — константа. Тобто умовна ймовірність того, що рибу вдасться піймати даним неводом, за умови, що її довжина дорівнює $t$ визначається (з точністю до константи) функцією $b(t)$. Функцію $b(t)$ будемо називати функцією, що реалізує зміщення (або функцією зміщення). У нашому прикладі це характеристика невода і, можливо, породи риб, що досліджується: чим більший розмір вічка невода, тим менше шансів, що риба даної довжини не зможе з нього втекти. Але для того, щоб бути впійманою, риба спочатку повинна потрапити до невода, а вже потім не зуміти з нього вибратись. Константа $c$ описує шанси рибини потрапити до невода. Чим більшим є озеро і популяція риб у озері, тим менший шанс для конкретної рибини опинитись у неводі і, відповідно, тим менше $c$.

Тепер визначимось з тим, що ми будемо оцінювати. Розглянемо $\xi^* = \xi(O)$ — значення досліджуваної змінної у об'єкта $O$, вибраного навмання з популяції, причому процедура відбору влаштована так, що всі об'єкти мають однакову ймовірність бути вибраними. Позначимо $F(x)$ — функцію розподілу $\xi^*$. Цю функцію будемо називати функцією розподілу популяції[15]. Якщо спостережувані об'єкти відбирались до вибірки зміщеною процедурою, то розподіл $\xi(O)$ для об'єкта з вибірки буде умов-

[15] Така складна конструкція може видатись дивною. Навіщо розглядати якусь умовну процедуру відбору, котру неможливо реалізувати на практиці, щоб характеризувати популяцію?

У науковій дисципліні, що зветься "методи вибіркових обстежень" часто використовують інший підхід. Популяція вважається скінченою, обсягу, скажімо $N$. Тоді всі значення досліджуваної величини $\xi(O)$ для всіх об'єктів з популяції можна формально перелічити $x_1, \ldots, x_N$. Визначимо формально популяційну функцію розподілу $F$ як емпіричну ф.р., пораховану за всією популяцією, тобто $F(x) = \frac{1}{n} \sum_{i=1}^{N} \mathbb{1}\{x_i < x\}$. Очевидно, що для такої "фіксованої" популяції це означення дає ту саму ф.р. $F$, яка була визначена у нас.

ним розподілом $\xi(O)$, за умови, що $s(O) = 1$, тобто що $O$ потрапив до вибірки. Функція розподілу для цього розподілу визначається як

$$G(x) = \mathsf{P}\{\xi(O) < x \mid s(O) = 1\}.$$

Таким чином, надалі ми будемо вважати, що зміщена вибірка складається з незалежних випадкових величин $\xi_1, \ldots, \xi_n$, які мають функцію розподілу $G$. За цією вибіркою потрібно оцінити популяційну функцію розподілу $F$.

Спочатку розберемося, як пов'язані між собою $F$ і $G$.

$$G(x) = \mathsf{P}\{\xi(O) < x \mid s(O) = 1\} = \frac{\mathsf{P}\{\xi(O) < x \text{ та } s(O) = 1\}}{\mathsf{P}\{s(O) = 1\}}.$$

Обчислимо спочатку чисельник у цьому виразі:

$$\begin{aligned}
\mathsf{P}\{\xi(O) < x \text{ та } s(O) = 1\} &= \mathbb{E}\mathbb{1}\{\xi(O) < x\}\mathbb{1}\{s(O) = 1\} \\
&= \mathbb{E}\left[\mathbb{E}[\mathbb{1}\{\xi(O) < x\}\mathbb{1}\{s(O) = 1\} \mid \xi(O)]\right] \\
&= \mathbb{E}\left[\mathbb{1}\{\xi(O) < x\}\,\mathsf{P}\{s(O) = 1 \mid \xi(O)\}\right] \\
&= \mathbb{E}\mathbb{1}\{\xi(O) < x\} \times cb(\xi(O)) = \int_{-\infty}^{x} cb(t)F(dt).
\end{aligned}$$

Для того, щоб обчислити знаменник, досить у цьому виразі покласти $x = +\infty$:

$$\mathsf{P}\{s(O) = 1\} = \mathsf{P}\{\xi(O) < +\infty \text{ та } s(O) = 1\} = \int_{-\infty}^{+\infty} cb(t)F(dt).$$

---

Підхід вибіркових обстежень є природним, якщо мова йде про дійсно фіксовану, незмінну популяцію. Наприклад при вибірковому контролю якості продукції, партія поставленого товару (скажімо, цегли) розглядається як "популяція". З великої партії, наприклад — вагону цегли, вибирають вибірку, скажімо, 20 цеглин і вимірюють якусь змінну, наприклад — вагу цеглини. Потрібно за значеннями змінної у вибірці зробити висновок про розподіл ваги цегли у всій партії. У таких дослідженнях зрозуміло, що обсяг популяції є сталим і значення досліджуваної змінної у всіх об'єктів також не змінюється. З рибами у ставку картина інша. Одні риби народжуються, інші гинуть, рибини ростуть, отже їхня довжина змінюється. Дослідника цікавить, як правило, не розподіл довжин риб у якийсь точно фіксований момент часу, а саме якою може виявитись довжина взятої навмання рибини з популяції. Звичайно, для того, щоб дослідження мало сенс, потрібно, щоб розподіл цієї довжини був достатньо стабільним, незважаючи на зміни у популяції.

Об'єднавши чисельник та знаменник, отримуємо:

$$G(x) = \frac{\int_{-\infty}^{x} cb(t)F(dt)}{\int_{-\infty}^{+\infty} cb(t)F(dt)} \tag{1.16}$$

$$= \frac{1}{R}\int_{-\infty}^{x} b(t)F(dt),$$

де

$$R = \int_{-\infty}^{\infty} b(t)F(dt).$$

Якщо $F(x)$ має щільність розподілу $f$, то диференціюючи (1.16), отримуємо, що щільність $G$ буде дорівнювати

$$g(x) = \frac{b(x)f(x)}{R}. \tag{1.17}$$

**Приклад 1.3.1.** Розглянемо статистичні дані, отримані за результатами вимірювання швидкості автомобілів, що їдуть по шосе. Вимірювання організоване наступними чином. На довгому шосе у фіксованій точці знаходиться дрожній патруль (інспектор) з радаром, який заміряє та фіксує швидкості всіх автомобілів, які їдуть повз нього по шосе у даному напрямку. Швидкість руху обмежена, скажімо, до 100 км/год. Вимірювання продовжуються з 13:00 по 13:30 за київським часом. Дані являють собою вибірку $v_1,\ldots, v_n$, де $v_j = v(O_j)$ — швидкість $j$-того спостережуваного автомобіля $O_j$. Будемо вважати, що водії не порушують правила дорожнього руху і всі швидкості не перевищують 100 км/год.

Припустимо, що нам потрібно оцінити за цими даними розподіл швидкостей автомобілів, які зазвичай рухаються по шосе у даний період часу[16]. Чи можна розглядати отриману інспектором вибірку, як незміщену? Якщо ми вважаємо популяцією всі автомобілі, що, в принципі, можуть їхати по шосе у цей час, то ця вибірка зміщена.

Дійсно, розглянемо автомобіль, який рухається по шосе зі швидкістю 10 км/год. Зрозуміло, що він потрапить у поле зору інспектора лише в тому випадку, коли о 13:00 він був від нього на відстані не більшій ніж 5

[16] Такі дані можуть бути цікавими, наприклад, для людини, яка вирішує де на шосе розташувати рекламу обідів у придорожньому ресторані. Чим більшими є звичайні швидкості автомобілів, тим далі від з'їзду до ресторану має бути реклама, щоб водій мав час її осмислити і підготуватись.

км (бо вже о 13:30 інспектор припинив спостереження і з більш далеких початкових положень на шосе автомобіль просто не встигне доїхати до нього). А автомобіль, що рухається зі швидкістю 100 км/год встигне відмітитись у інспектора навіть якщо на початку він був на відстані 50 км. Отже, більш швидкі автомобілі мають більше шансів потрапити до вибірки, ніж повільні. Автомобіль, який стоїть на шосе (зі швидкістю 0 км/год) взагалі не має шансів потрапити в поле зору інспектора і, відповідно, за нашою вибіркою ми не зможемо оцінити кількість таких автомобілів. Будемо вважати, шо автомобілі з нашої популяції можуть у момент 13:00 можуть з однаковою ймовірністю опинитись у будь-якій точці ділянки шосе на відстані від 0 до 50 км від інспектора. Розглянемо автомобіль $O$, який був на цій ділянці і позначимо $x(O)$ — відстань від нього до інспектора у момент 13:00. Інспектор побачить цей автомобіль, якщо $x(O) < v(O)/2$ (ділимо на два тому, що швидкість вимірюється у кілометрах на годину, а спостереження велись півгодини). За нашим припущенням $x(O)$ рівномірно розподілені на інтервалі $[0, 50]$ і незалежні від $v(O)$. Тоді ймовірність потрапити до вибірки для автомобіля $O$ з швидкістю $v(O) = t$ буде

$$\mathsf{P}\{s(O) = 1\} = c \times b(t),$$

де $c$ — ймовірність того, що $O$ опиниться на відрізку шосе на відстані меншій 50 км від інспектора в момент 13:00, а

$$b(t) = \mathsf{P}\{x(O) < t/2 \mid \text{за умови, що O на ділянці}\} = \frac{t}{100}$$

— ймовірність того, що такий автомобіль встигне доїхати до інспектора за час, коли той веде спостереження. Отже, $b(t)$ є в нашому випадку, функцією, що реалізує зміщення.

Наскільки помітним буде таке зміщення? Нехай, наприклад, $v(O)$ рівномірно розподілені на $[0, 100]$. (А оце припущення є неприродним і ми його розглядаємо лише для наочності). Тоді популяційна щільність швидкостей $f(x) = 1/100$ для $x \in [0, 100]$ і $f(x) = 0$ для всіх інших $x$. А вибіркова щільність

$$g(x) = \frac{b(x)f(x)}{\int_0^{100} b(t)f(t)dt} = \begin{cases} \frac{x}{5000} & \text{при } x \in [0, 100] \\ 0 & \text{при } x \notin [0, 100], \end{cases}$$

тобто рівномірний розподіл популяції таке зміщення перетворює на трикутний у вибірці.

Ефект, що спостерігається у цій задачі, називають **парадоксом інспектора**. Зрозуміло, що цей "парадокс" виявляється не тільки у дорожньому русі. Власне, спочатку ефект інспектора прийшлось враховувати при проведенні фізичних експериментів з космічними променями. Жорстке випромінення Сонця створює елементарні частинки високих енергій у високих шарах атмосфери Землі. Дякі з цих частинок досягають земної поверхні. Для того, щоб їх помітити, використовували фотопластини. На них при попаданні елементарної частинки утворюється темна пляма, розмір якої пропорційний кінетичній енергії частинки (квадрату швидкості). Оскільки час життя частинки обмежений, при фіксації виникає ефект інспектора: частинки з великою енергією мають більше шансів встигнути долетіти до пластини перш ніж розпастися[17]. ◀

У цьому прикладі функція, що реалізує зміщення, вибрана нами з теоретичних міркувань. У багатьох інших задачах визначити її теоретично не вдається. Виникає, відповідно, задача оцінювання функції $b(t)$ за якими-небудь спостережуваними даними. Наприклад, можна виходити з формули (1.17):

$$b(t) = R\frac{g(t)}{f(t)}.$$

. Якщо нам вдасться для деякої популяції якимось чином оцінити справжню популяційну щільність оцінкою $\hat{f}$ і за даними, отриманими певною вибірковою процедурою, оцінити вибіркову щільність $\hat{g}$, то на роль оцінки для $b(t)$ цієї процедури можна взяти

$$\hat{b}(t) = \frac{\hat{g}(t)}{\hat{f}(t)}.$$

(Оскільки у нас функція $b$ визначається з точністю до сталого множника, константу $R$ можна ігнорувати).

Зрозуміло, що, якщо у нас є хороша оцінка для $f$, то можна оцінити $F$ як інтеграл від $f$ і взагалі не клопотатись даними за зміщеною вибіркою. Але справа в тому, що $b(t)$ у таких задачах трактують як характеристику саме вибіркової процедури, яка не залежить (або мало залежить) від популяції, до якої застосовується ця процедура. Тому

[17] Цей ефект не треба плутати з поправками, які у обчислення спостережуваної тривалості життя рухомої частинки вносить теорія відносності.

$b(t)$, оцінену при застосуванні певної процедури до популяції, де ми знаємо $f$, можна потім використовувати для оцінки популяційних розподілів у інших популяціях, для яких $f$ невідоме. Наприклад, у задачі аналізу довжин риб, виловлених неводом, можна оцінити зміщуючу функцію неводу, перевіряючи його роботу у ставку, куди риби різних довжин були запущені дослідником і, отже, їх популяційна щільність розподілу $f$ — відома. А застосовувати цю функцію можна для врахування зміщення вибірок, отриманих цим самим неводом у озерах, де $f$ невідома.

Щоб реалізувати цю ідею потрібно вміти оцінювати щільності розподілу за спостереженнями. Цій задачі присвячено розділ 4.

### 1.3.2 Оцінка Горвіца — Томпсона

Отже, у нас є вибірка з н.о.р. випадкових величин $\mathbf{X} = (\xi_1, \ldots, \xi_n)$, які мають функцію розподілу

$$G(x) = \frac{1}{R} \int_{-\infty}^{x} b(t) F(dt),$$

де

$$R = \int_{-\infty}^{\infty} b(t) F(dt),$$

$F$ — невідома популяційна функція розподілу, $w : \mathbb{R} \to [0, 1]$ — відома функція, що реалізує зміщення.

Задача полягає в тому, щоб оцінити $F$ за вибіркою $\mathbf{X}$. Для цього ми застосуємо техніку оцінювання методом моментів. У цьому методі, якщо потрібно оцінити невідомий параметр $F$, ми вибираємо деяку моментну функцію $h(t)$ і розглядаємо відповідний емпіричний момент

$$\hat{h}_n = \frac{1}{n} \sum_{j=1}^{n} h(\xi_j).$$

За законом великих чисел, $\hat{h}_n \to \mathsf{E}_F\, h(\xi_1) = H(F)$ при $n \to \infty$. Щоб отримати оцінку, ми прирівнюємо теоретичний момент $H(F)$ та емпіричний і розв'язуємо отримане "моментне" рівняння відносно $F$.

У нашій задачі маємо

$$H(F) = \frac{1}{R} \int_{-\infty}^{+\infty} h(t) b(t) F(dt).$$

Якщо, як у п. 1.1.3, обрати $h(t) = h_x(t) = \mathbb{1}\{t < x\}$, отримаємо $\hat{h}_{x,n} = \hat{F}_n(x)$ — звичайна емпірична функція розподілу, підрахована за вибіркою $\mathbf{X}$. Для того, щоб знайти оцінку для популяційної ф.р., потрібно буде розв'язувати інтегральне рівняння відносно $F$:

$$\hat{F}_n(x) = \frac{1}{R}\int_{\infty}^{x} b(t)F(dt),$$

а це не так просто. Спробуємо підібрати більш зручну функцію $h$, так, щоб "вбити" вплив функції $b$ у цьому інтегралі. Якщо взяти

$$h_x(t) = \frac{\mathbb{1}\{t < x\}}{b(t)},$$

то ми отримаємо моментне рівняння

$$\hat{h}_{x,n} = \frac{1}{n}\sum_{j=1}^{n}\frac{\mathbb{1}\{\xi_j < x\}}{b(\xi_j)} = \frac{1}{R}\int_{-\infty}^{x} F(dx) = \frac{F(x)}{R},$$

отже наша оцінка, як розв'язок цього рівняння відносно $F$, може мати вигляд

$$\tilde{F}(x) = \tilde{R}\frac{1}{n}\sum_{j=1}^{n}\frac{\mathbb{1}\{\xi_j < x\}}{b(\xi_j)},$$

де $\tilde{R}$ — деяка оцінка для $R$. Якщо ми хочемо, щоб наша оцінка для ф.р. $F$ сама була деякою функцією розподілу, то потрібно, щоб $\tilde{F}(+\infty) = 1$, тобто

$$\tilde{R} = \frac{n}{\sum_{j=1}^{n}\frac{1}{b(\xi_j)}}.$$

Остаточно, збираючи оцінки, отримуємо

$$\hat{F}_n^{HT}(x) = \frac{\sum_{j=1}^{n}\frac{\mathbb{1}\{\xi_j<x\}}{b(\xi_j)}}{\sum_{j=1}^{n}\frac{1}{b(\xi_j)}}. \tag{1.18}$$

Функцію $\hat{F}_n^{HT}(x)$ називають оцінкою Горвіца — Томпсона для ф.р. за зміщеною вибіркою (Horvitz–Thompson estimator). Цю оцінку можна записати як навантажену емпіричну функцію розподілу, тобто

$$\hat{F}_n^{HT}(x) = F(x, \mathbf{w}^{HT}) = \sum_{j=1}^{n} w_j \mathbb{1}\{\xi_j < x\},$$

де $\mathbf{w}^{HT} = (w_1^{HT}, \ldots, w_n^{HT})$ — набір вагових коефіцієнтів Горвіца — Томпсона,

$$w_j^{HT} = \frac{1/b(\xi_j)}{\sum_{i=1}^n 1/b(\xi_i)}. \tag{1.19}$$

Коефіцієнти $w_j^{HT}$ називають ваговими коефіцієнтами Горвіца-Томпсона. Вони обернено пропорційні ймовірності того, що $j$-тий об'єкт потрапить до вибірки, причому коефіцієнт пропорційності вибрано з умови $\sum_{j=1}^n w_j^{HT} = 1$.

Евристичне пояснення такого вибору вагових коефіцієнтів полягає в тому, що чим меншою є для об'єктів з даними властивостями ймовірність потрапити до вибірки, тим більшу групу з популяції представляє такий об'єкт, якщо він справді опинився у вибірці. Для того, щоб це врахувати, ми надаємо йому більшої ваги. Умова $\sum_{j=1}^n w_j^{HT} = 1$ дозволяє нам трактувати отриману оцінку як ф.р. дискретної величини, що приймає значення $\xi_j$ з ймовірністю $w_j^{HT}$.

### 1.3.3 Асимптотична нормальність оцінок Горвіца — Томпсона

Зрозуміло, що якщо $b(\xi_j) = 0$ при деякому $j$, то оцінка Горвіца — Томпсона стає невизначеною. Але така ситуація може виникнути з ненульовою ймовірністю тільки тоді, коли обрана нами для оцінювання функція, що реалізує зміщення, не відповідає реальному зміщенню. Дійсно, з точністю до сталого множника, $b(t)$ — це ймовірність того, що об'єкт зі значенням $\xi = t$ потрапить до вибірки. Отже, якщо $b(t) = 0$, то значення $t$ не може опинитись у вибірці.

З іншого боку, якщо $b(t) = 0$ на деякому інтервалі $t \in [t_1, t_2]$, то зрозуміло, що консистентна оцінка функції розподілу $F(t)$ на цьому інтервалі неможлива: якщо риби з популяції, які мають довжини $t_0 \le \xi \le t_1$ взагалі не можуть потрапити до вибірки, то як ми оцінимо розподіл довжин у популяції на цьому інтервалі? Таким чином, щоб можна було будувати консистентні оцінки за зміщеною вибіркою, потрібно, щоб $b(t)$ не оберталась у 0 там, куди значення спостережуваної змінної у популяції можуть потрапити з ненульовою ймовірністю. Цю умову можна формально записати як:

$$\int_{-\infty}^{+\infty} \mathbb{1}\{b(t) = 0\} F(dt) = 0 \tag{1.20}$$

(Функція $b$ не дорівнює 0 майже напевно по мірі $F$).

**Теорема 1.3.1.** *Нехай виконується умова (1.20). Тоді, для будь-якого* $x \in \mathbb{R}$,

$$\hat{F}_n^{HT}(x) \to F(x) \text{ майже напевно при } n \to \infty.$$

Таким чином, оцінка Горвіца — Томпсона є сильно консистентною.

**Доведення.** Запишемо

$$\hat{F}_n^{HT}(x) = \frac{J_n^1}{J_n^2}, \text{ де } J_n^1 = \frac{1}{n}\sum_{j=1}^{n} \frac{\mathbb{1}\{\xi_j < x\}}{b(\xi_j)}, \; J_n^2 = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{b(\xi_j)}.$$

За підсиленим законом великих чисел

$$J_n^1 \to \mathsf{E}\,\frac{\mathbb{1}\{\xi_1 < x\}}{b(\xi_1)} = \frac{1}{R}\int_{-\infty}^{x} \frac{b(t)}{b(t)} F(dt) = F(x)/R,$$

$$J_n^2 \to \mathsf{E}\,\frac{1}{b(\xi_1)} = \frac{1}{R}\int_{-\infty}^{+\infty} F(dt) = 1/R \text{ м.н. при } n \to \infty.$$

(Зауважимо, що, оскільки $0 \le b(t) \le 1$, то ці математичні сподівання є скінченними, а умова (1.20) забезпечує $R > 0$).

Отже

$$\hat{F}_n^{HT}(x) \to \frac{F(x)/R}{1/R} = F(x) \text{ м.н., при } n \to \infty.$$

Теорема доведена.

Як показує наступна теорема, оцінка Горвіца — Томпсона є також асимптотично нормальною з нормуючою послідовністю $\sqrt{n}$. Позначимо

$$\sigma_{HT}^2(x) = R\int_{-\infty}^{+\infty} \frac{1}{b(t)} (\mathbb{1}\{t < x\} - F(x))^2 F(dx).$$

**Теорема 1.3.2.** *Нехай виконується умова (1.20). Тоді, для будь-якого* $x \in \mathbb{R}$, *такого, що* $\sigma_{HT}^2(x) > 0$, *виконано*

$$\sqrt{n}(\hat{F}^{HT}(x) - F(x)) \xrightarrow{W} N(0, \sigma_{HT}^2(x)) \text{ при } n \to \infty.$$

**Доведення.** Помітимо, що

$$\sqrt{n}(\hat{F}^{HT}(x) - F(x)) = \frac{1}{\frac{1}{n}\sum_{j=1}^{n} \frac{1}{b(\xi_j)}} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{b(\xi_j)} \left(\mathbb{1}\{\xi_j < x\} - F(x)\right). \tag{1.21}$$

Позначимо

$$\eta_j = \frac{1}{b(\xi_j)} \left( \mathbb{1}\{\xi_j < x\} - F(x) \right).$$

Зрозуміло, що $\eta_j$ — незалежні, однаково розподілені випадкові величини, $\mathsf{E}\,\eta_j = 0$,

$$\mathrm{D}\,\eta_j = \mathsf{E} \left( \frac{1}{b(\xi_j)} \left( \mathbb{1}\{\xi_j < x\} - F(x) \right) \right)^2$$

$$= \frac{1}{R} \int_{-\infty}^{+\infty} \left( \frac{\mathbb{1}\{t < x\} - F(x)}{b(t)} \right)^2 b(t) F(dt) = \frac{\sigma^2_{HT}(x)}{R^2}.$$

Тому, за центральною граничною теоремою,

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{b(\xi_j)} \left( \mathbb{1}\{\xi_j < x\} - F(x) \right) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \eta_j \xrightarrow{W} N\left( 0, \frac{\sigma^2_{HT}(x)}{R^2} \right).$$

За законом великих чисел

$$\frac{1}{\frac{1}{n} \sum_{j=1}^{n} \frac{1}{b(\xi_j)}} \to R.$$

Тому з (1.21), використовуючи лему Слуцького, отримуємо твердження теореми.

Теорема доведена.

### 1.3.4 Двовибіркова задача. Оцінка Варді

Досить часто зустрічаються задачі статистичного аналізу в яких статистик має справу з кількома наборами даних про одну характеристику досліджуваних об'єктів, отриманими за різних обставин. Такі задачі прийнято називати багатобірковими. У цьому підрозділі ми розглядаємо двовибіркову задачу, в якій одна вибірка отримана з популяції за допомогою незміщеної процедури, а друга — з тієї самої популяції, але зміщеною процедурою.

**Приклад 1.3.2.** Як приклад можна розглянути задачу оцінки розподілу певної фізіологічної характеристики $\xi$ (для простоти, хай це буде температура тіла) у людей на початку певної хвороби (тут можна вважати, що йдеться про covid19). У пацієнтів, що потрапили до лікаря з підозрою на хворобу, вимірюється $\xi$ і записується у набір даних. Якщо

підозра підтвердилась, це значення $\xi$ вноситься до вибірки, за якою будуть оцінювати розподіл цієї характеристики. Але потрапити до лікаря пацієнт може різними шляхами. Одні пацієнти приходять тому, що вони відчувають себе хворими ("самоплив", С) інші виявляються в результаті планового профілактичного обстеження ("профілактика", П). Фактично, дані утворюють дві вибірки - С і П, отримані різними вибірковими процедурами. Якщо П можна, з певними застереженнями, вважати незміщеною, то ймовірність потрапити до С, вочевидь, залежить від $\xi$: людина з високою температурою скоріше піде до лікаря, ніж із нормальною. ◀

Отже, нехай у нас є два набори даних про значення змінної $\xi$ у об'єктів з деякої популяції. Перший набір, обсягу $n_1$, отриманий незміщеною вибірковою процедурою, а другий, обсягу $n_2$ — зміщеною. Для зручності використання у формулах, складемо ці два набори в один, обсягу $n = n_1 + n_2$, розташувавши спочатку незміщений, а потім — зміщений. Таким чином, маємо вибірку $\xi_1,\ldots,\xi_{n_1},\xi_{n_1+1},\ldots,\xi_n$. Спостереження будемо вважати незалежними у сукупності, причому при $j = 1,\ldots,n_1$ вони мають функцію розподілу $F$, а при $j = n_1+1,\ldots,n$ — ф.р. $G$, де $G$ задано (1.16):

$$G(x) = \frac{1}{R}\int_{-\infty}^{x} b(t)F(dt), R = \int_{-\infty}^{\infty} b(t)F(dt).$$

Тут $F$ — невідома популяційна функція розподілу, яку потрібно оцінити на набором $\mathbf{X} = (\xi_j)_{j=1}^n$, $b$ — відома функція, що реалізує зміщення.

Для оцінки $F$ скористаємось методом емпіричної найбільшої вірогідності. Обмежимось розглядом функцій, вигляду

$$F(x) = F(x;\mathbf{w}) = \sum_{i=1}^{m} w_i \mathbb{1}\{x_i < x\},$$

де $x_k$ — точки з фіксованого набору, $w_k$ — вагові коефіцієнти, які потрібно оцінити. Складемо функцію вірогідності, використовуючи щільність спостережень відносно лічильної міри на наборі точок $x_k$. Для $j = 1,\ldots,n_1$, при $F(x) = F(x;\mathbf{w})$ щільність $\xi_j$ дорівнює

$$f(x;\mathbf{w}) = \begin{cases} w_1, & \text{якщо } x = x_1, \\ w_2, & \text{якщо } x = x_2, \\ \ldots & \\ w_m, & \text{якщо } x = x_m, \\ 0 & \text{якщо } x \notin \mathbb{X}. \end{cases}$$

При $j = n_1, \ldots, n$ щільність $\xi_j$, це щільність розподілу $G$, тобто

$$g(x, \mathbf{w}) = \frac{b(x)f(x; \mathbf{w})}{\int_{-\infty}^{\infty} b(t)F(dt; \mathbf{w})} = \frac{b(x)f(x; \mathbf{w})}{\sum_{i=1}^{m} b(x_i)w_i}.$$

Оскільки спостереження є незалежними, для отримання функції вірогідності потрібно підставити спостережувані значення у відповідні щільності і перемножити їх:

$$L(\mathbf{w}) = \prod_{j=1}^{n_1} f(\xi_j; \mathbf{w}) \prod_{j=n_1+1}^{n} g(\xi_j; \mathbf{w}) = \frac{\prod_{j=1}^{n} f(\xi_j; \mathbf{w}) \prod_{j=n_1+1}^{n} b(\xi_j)}{\left(\sum_{i=1}^{m} w_i b(x_i)\right)^{n_2}}.$$

Ті ж міркування, що і у п. 1.1.3, приводять до висновку, що на роль набору $x_i$ потрібно взяти ті і тільки ті значення, які є у вибірці $\mathbf{X}$. Розглянемо випадок, коли всі $\xi_j$ у вибірці є різними. Тоді у добутку $\prod_{j=1}^{n} f(\xi_j; \mathbf{w})$ кожне значення $w_j$ буде зустрічатись рівно один раз. Отже, отримуємо:

$$L(\mathbf{w}) = \frac{\prod_{j=1}^{n} w_j \prod_{j=n_1+1}^{n} b_j}{\left(\sum_{j=1}^{n} w_j w_j\right)^{n_2}},$$

де $b_j = b(\xi_j)$.

Для того, щоб отримати оцінку найбільшої вірогідності для $\mathbf{w}$, потрібно максимізувати $L(\mathbf{w})$ по $\mathbf{w}$ при виконанні обмежень:

$$\sum_{j=1}^{n} w_j = 1 \tag{1.22}$$

та $w_j \geq 0$. Забудемо на хвилину про невід'ємність $w_j$ і знайдемо точку максимуму $\log(L(\mathbf{w}))$ при обмеженні (1.22). Для цього скористаємось технікою множників Лагранжа. Складемо функцію Лагранжа:

$$\mathcal{L}(\mathbf{w}) = \sum_{j=1}^{n} \log(w_j) + \sum_{j=n_1+1}^{n} b_j - n_2 \log(\sum_{j=1}^{n} b_j w_j) + \lambda \sum_{j=1}^{n} w_j,$$

де $\lambda$ — невизначений множник Лагранжа. Позначимо

$$\bar{b} = \sum_{j=1}^{n} b_j w_j. \tag{1.23}$$

Візьмемо похідну від $\mathcal{L}(\mathbf{w})$ по $w_k$ і прирівняємо її до 0, щоб знайти стаціонарні точки функції Лагранжа:

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial w_k} = \frac{1}{w_k} - \frac{n_2 b_k}{\bar{b}} + \lambda = 0.$$

Отримуємо

$$w_k = \frac{\bar{b}}{n_2 b_k + c}, \tag{1.24}$$

де $c = -\bar{b}\lambda$. Оскільки невизначений множник $\lambda$ може приймати будь-які значення, то і $c$ також може бути будь-якою константою. Для знаходження $\bar{b}$ і $c$ можна скористатись рівняннями (1.22,1.23). Підставляючи (1.23) у (1.24), отримуємо рівняння відносно $c$:

$$\sum_{j=1}^{n} \frac{b_j}{c + n_2 b_j} = 1. \tag{1.25}$$

Помітимо, що при $c > 0$, кожен доданок у правій частині цього рівняння є монотонно спадною функцією від $c$. Отже і вся сума є спадною. При $c = 0$ права частина дорівнює $n/n_2 > 1$, а при $c \to \infty$ вона прямує до 0. Отже існує рівно один додатній корінь $\hat{c}$ рівняння (1.25) У роботі [27] показано, що хоча це рівняння може іще мати багато від'ємних коренів, $\hat{c}$ відповідає точці максимуму функції вірогідності. Якщо тепер підставити (1.24) у (1.22), отримуємо

$$\sum_{k=1}^{n} \frac{\bar{b}}{n_2 b_k + \hat{c}} = 1,$$

звідки

$$\bar{b} = \left( \sum_{k=1}^{n} \frac{1}{n_2 b_k + \hat{c}} \right)^{-1} \tag{1.26}$$

Вагові коефіцієнти $w_k^V = w_k$ визначені (1.24), де $\hat{c}$ є додатнім розв'язком (1.25), а $\bar{b}$ визначено (1.26), називають **коефіцієнтами Варді**, а навантажену емпіричну ф.р. $F(x, \mathbf{w}^V)$ з цими коефіцієнтами — оцінкою Варді для функції розподілу за об'єднанням зміщеної та незміщеної вибірок. Оцінка Варді є оцінкою методу емпіричної найбільшої вірогідності.

Помітимо, що, якщо $n_1 = 0$, тобто незміщеної частини даних не існує, з (1.25) отримуємо $\hat{c} = 0$ (бо $n = n_2$) і оцінка Варді перетворюється на

оцінку Горвіца — Томпсона. Таким чином, оцінка Горвіца — Томпсона є також і оцінкою найбільшої емпіричної вірогідності за зміщеною вибіркою.

У статті [18] доведена теорема про асимптотичну нормальність оцінок Варді. На жаль, це доведення і формула для асимптотичної дисперсії оцінки досить складні, тому ми їх тут не розглядатимем.

## 1.4 Запитання і задачі

**Запитання.**

1. Дайте означення емпіричної функції розподілу. Які статистичні властивості цієї функції ви знаєте?

2. Сформулюйте теорему Глівенко — Кантеллі. Чим твердження цієї теореми відрізняється від твердження про поточкову консистентність емпіричних функцій розподілу?

3. Як при побудові довірчого інтервалу для функції розподілу використовується твердження про асимптотичну нормальність емпіричних функцій розподілу?

4. Чому для побудови непараметричної оцінки функції розподілу неможливо використати звичайний метод найбільшої віроігдності?

5. Чим метод емпіричної найбільшої вірогідності відрізняється від звичайного методу найбільшої вірогідності?

6. Що таке вибірка, цензурована з права?

7. Дайте означення оцінки Каплана — Мейера для функції розподілу за цензурованою вибіркою у загальному випадку і у випадку, коли всі значення у вибірці — різні.

8. Поясніть, як було отримано функцію вірогідності для побудови оцінки методу емпіричної найбільшої вірогідності для функції розподілу за цензурованою вибіркою. Чому при побудові цієї оцінки не можна множину можливих точок стрибків функції розподілу не можна вибрати рівною множині всіх вибіркових значень, а потрібно додавати іще одну точку?

9. Запишіть формулу Грінвуда. Поясніть, що саме можна обчислити за цією формулою. Які є застосування формули Грінвуда?

10. Поясніть, що таке зміщена вибіркова процедура. Наведіть приклади. Поясніть, що таке функція, що реалізує зміщення і якою вона може бути у наведених вами прикладах.

11. Яким способом можна оцінити функцію, що реалізує зміщення? Що для цього потрібно?

12. Дайте означення оцінки Горвіца — Томпсона. Дайте евристичне пояснення використання вагових коефіцієнтів у цій оцінці.

13. Які теореми теорії ймовірностей використовуються при доведенні консистентності та асимптотичної нормальності оцінок Горвіца — Томпсона?

14. Яким методом отримана оцінка Варді? Опишіть отримання цієї оцінки.

15. За яких обставин доцільно використовувати оцінки Горвіца — Томпсона, а за яких — оцінки Варді?

**Задачі.**

1. Нехай вибірка обсягу $n = 100$ вибрана з експоненційного розподілу з інтенсивністю $\lambda = 1$. Чому буде дорівнювати дисперсія емпіричної функції розподілу, побудованої по цій вибірці, в точці $x = 0.5$?

2. За вибіркою $\mathbf{X} = (0.3, 2, -1.2, 3.1, 0.7)$ обчисліть значення емпіричної функції розподілу $\hat{F}_n(x)$ в точці $x = 2.5$. Нарисуйте на папері графік функції $\hat{F}_n(x)$ при $x \in [-2, 4]$.

3. У лабораторії працювали без зупинки шість однотипних приладів. Чотири з них перегоріли на 28, 41, 32 і 16му днях роботи відповідно. Два прилади були вимкнені у працездатному стані на 36 і 30му днях роботи, після чого розібрані для удосконалення. За цим набором даних побудуйте (на папері) графік оцінки Каплана — Мейера для функції розподілу тривалості роботи приладу від ввімкнення до перегоряння.

4. Вибірка отримана з використанням зміщеної вибіркової процедури з функцією, що реалізує зміщення, $b(x) = 1 - x$ при $x \in [0, 1]$. Функція розподілу спостережуваної величини у популяції $F(x) = x^2$ при $x \in [0, 1]$. Знайдіть функцію розподілу спостережуваної величини у об'єктів, що потрапили до вибірки.

5. Вибірка отримана з використанням зміщеної вибіркової процедури з функцією, що реалізує зміщення, $b(x) = x/10$ при $x \in [0, 10]$. Отримані такі вибіркові значення: 3, 6, 8, 9. Нарисуйте (на папері) графік оцінки Горвіца — Томпсона для функції розподілу у популяції за цією зміщеною вибіркою.

**Завдання для виконання на комп'ютері**

**Завдання 1. (Емпірична функція розподілу)**

Реалізуйте обчислення емпіричної функції розподілу за вибіркою у вигляді функції яка має специфікацію

`Femp(x,sample)`,

де

`x` — точка, в якій обчислюється емпірична функція розподілу,

`sample` — вибірка, по якій будується функція.

Значення функції `Femp(x,sample)` повинно дорівнювати значенню емпіричної функції розподілу у точці `x`.

(При написанні коду не використовуйте готові функції з бібліотек статистичних функцій. Дозволяється застосовувати лише функції загального призначення.)

Перевірте роботу розробленої Вами функції на модельованих даних. Для цього візьміть з таблиці індивідуальних варіантів до цього завдання розподіл $F$, що відповідає номеру Вашого варіанту. Для цього розподілу $F$ згенеруйте кратні вибірки обсягу

$$n = 10,\ 50,\ 100,\ 500,\ 1000$$

спостережень. За кожною з цих вибірок підрахуйте емпіричну функцію розподілу на інтервалі від квантиля $F$ рівня 0.01 до квантиля рівня 0.99. Для кожного обсягу вибірки нарисуйте на одному рисунку різними кольорами графіки емпіричної функції розподілу, підрахованої за вибіркою, і справжньої теоретичної функції розподілу $F$.

Порівняйте отримані графіки, зробіть висновки про те, чи підтверджують вони теорему Глівенко-Кантеллі.

Реалізуйте розрахунок меж довірчого інтервалу для функції розподілу у заданій точці у вигляді функції. Специфікація функції

`Fconf(x,sample,alpha)`,

де

`x` — довірчий інтервал будується для значення функції розподілу у цій точці,

`sample` — вибірка, по якій будується довірчий інтервал,

`alpha` — рівень значущості довірчого інтервалу.

Значенням функції `Fconf()` повинен бути вектор з двох елементів: нижньої і верхньої меж довірчого інтервалу.

3. Для розподілу $F$, обраного у завданні 1, точок `x`, що дорівнюють квантилям рівня 1/3, 1/2 і 2/3 для розподілу $F$ і обсягів вибірки

$$n = 10,\ 50,\ 100,\ 500,\ 1000$$

виконайте наступну роботу.

Для кожного варіанту $n$ згенеруйте $m = 10000$ псевдовипадкових вибірок обсягу $n$. По кожній вибірці знайдіть довірчий інтервал для значення функції розподілу у точці `x` з рівнем значущості $\alpha$. Підрахуйте відносну частоту попадання справжнього значення функції розподілу у отримані довірчі інтервали. Отримані частоти запишіть у табличку, в якій кожен рядочок відповідає одному значенню $n$, а кожен стовпчик — одному зі значень `x`.

Зробіть висновок про акуратність побудованих Вами довірчих інтервалів і доцільність їх використання при різних обсягах вибірки.

**Індивідуальні варіанти функції розподілу** $F$:

(1) $F \sim N(1,1)$,
(2) $F \sim \chi^2(3)$,
(3) $F$ — симетричний трикутний розподіл на [0,2],
(4) $F \sim \mathrm{Exp}_{\lambda=1}$,
(5) $F \sim \chi^2(4)$,
(6) $F$ — Т-Стьюдента з 6-ма ступенями вільності,
(7) $F$ — бета-розподіл з параметрами $\alpha = 2$, $\beta = 3$,
(8) $F$ — розподіл Лапласа з середнім 0 та інтенсивністю 1,
(9) $F \sim \mathrm{Exp}_{\lambda=0.5}$,
(10) $F$ — рівномірний розподіл на [0,1].

**Завдання 2. (Оцінка Каплана-Мейера)**

Реалізуйте оцінку Каплана-Мейера у вигляді функції. Функція повинна мати специфікацію

`FKM(x,sample,delta)`,

де

`x` — точка, в якій оцінюється невідома функція розподілу;

`sample` — набір спостережуваних значень;

`delta` — набір індикаторів того, що цензурування не відбулось (тобто відповідний елемент `sample` є значенням досліджуваної величини, а не цензора).

Значенням функції `FKM(x,sample,delta)` повинно бути значення оцінки Каплана-Мейера для функції розподілу у точці `x`, підрахованої за даними `sample,delta`.

За допомогою генераторів псевдовипадкових чисел згенеруйте цензуровану вибірку з розподілами досліджуваної величини $F$ і цензора $G$, вибраними з таблиці 1.1 відповідно до Вашого варіанту. Генерацію про-

| Варіант | $F$ | $G$ |
|---|---|---|
| 1 | $F \sim \chi^2(3)$ | $G \sim Exp_{\lambda=1/3}$ |
| 2 | $F \sim Exp_{\lambda=1}$ | $G \sim Exp_{\lambda=1/2}$ |
| 3 | $F$ — симетричний трикутний на [0,2] | $G$ рівномірний на [0,2] |
| 4 | $F$ — логнормальний $\sim \exp(\xi)$, $\xi \sim N(0,1)$ | $G \sim \chi^2(3)$ |
| 5 | $F$ — півнормальний $\sim \lvert\xi\rvert$, $\xi \sim N(0,1)$ | $G \sim \chi^2(2)$ |
| 6 | $F \sim Exp_{\lambda=1/3}$ | $G \sim \chi^2(3)$ |
| 7 | $F \sim Exp_{\lambda=1/2}$ | $G \sim Exp_{\lambda=1}$ |
| 8 | $F$ рівномірний на [0,2] | $G$— симетричний трикутний на [0,2] |
| 9 | $F \sim \chi^2(3)$ | $G$— логнормальний $\sim \exp(\xi)$, $\xi \sim N(0,1)$ |
| 10 | $F \sim \chi^2(2)$ | $G$ — півнормальний $\sim \lvert\xi\rvert$, $\xi \sim N(0,1)$ |

Табл. 1.1: Розподіли для цензурованої вибірки

ведіть для обсягів вибірки

$$n = 10,\ 50,\ 100,\ 500,\ 1000$$

По кожній з вибірок побудуйте графік відповідної оцінки Каплана-Мейера для функції розподілу $F$ разом з графіком справжньої теоретичної функції розподілу $F$.

Для кожного обсягу

$$n = 10,\ 50,\ 100,\ 500,\ 1000$$

згенеруйте по $m = 1000$ цензурованих вибірок з розподілом, що відповідає Вашому варіанту. По кожній вибірці оцініть функцію розподілу $F(x)$ в точці $x$, що дорівнює квантилю рівня $1/2$ для розподілу $F$. Таким чином, отримаєте вибірки з оцінок, по $m$ спостережень у кожній вибірці, що відповідає певному фіксованому значенню $n$.

Побудуйте гістограми по кожній з цих вибірок з оцінок і відобразіть на них нормовану щільність нормального розподілу з відповідними параметрами (див. п. 7.1 в [3]). Оцініть на око, для даних якого обсягу $n$ розподіл оцінок Каплана-Мейера добре описується нормальним наближенням.

Нехай $\hat{F}_n^{KM}(x)$ — оцінка Каплана-Мейера для функції розподілу $F(x)$. Розгляньте $\ln(1 - \hat{F}_n(x)^{KM})$ як оцінку для $\ln(1 - F(x))$. Проведіть дослідження, аналогічне попередньому, для перевірки того, наскільки добре описує нормальне наближення розподіл такої оцінки?

Використайте псевдовипадкові дані, згенеровані у попередньому завданні, для перевірки точності довірчих інтервалів.

Для цього, при фіксованому $n$ ($n = 10,\ 50,\ 100,\ 500,\ 1000$) по кожній з $m$ цензурованих вибірок побудуйте довірчі інтервали для $F(x)$ (як і раніше, $x$ — квантиль рівня $1/2$ для розподілу $F$) двома способами:

(1) використовуючи асимптотичну нормальність оцінок Каплана-Мейера,

(2) використовуючи асимптотичну нормальність $\ln(1 - \hat{F}_n(x)^{KM})$.

Для інтервалів кожного типу підрахуйте відносну частоту випадків, коли ці інтервали покривали справжнє значення $F(x)$.

Зробіть висновок, інтервали якого типу - (1) чи (2) виявились більш точними у Вашому випадку.

**Завдання 3. (Оцінювання функцій розподілу за зміщеною вибіркою)**

Для заданих функції розподілу $F$ та функції, що реалізує зміщення $b$, згенерувати незміщену вибірку обсягу $n_1 = 300$ та зміщену вибірку обсягу $n_2 = 300$, побудувати за ними чотири оцінки для $F$:

(1) емпірична функція розподілу, що підраховується за незміщеною вибіркою;

(2) оцінка Горвіца — Томпсона, що підраховується за зміщеною вибіркою;

(3) середнє значення оцінок (1) і (2), підрахованих в фіксованій точці $x$;

(4) оцінка Варді.

Вивести графіки цих функцій на одному рисунку разом із графіком справжньої функції $F$. Зробити попередні висновки щодо того, яка оцінка точніша (на око).

**Індивідуальні варіанти функції розподілу та зміщуючої функції**:

(1) $F \sim N(1,1)$, $b(t) = 1/(1+\exp(t-1)$,

(2) $F \sim \chi^2(3)$, $b(t) = 1/(t+1)$,

(3) $F$ — трикутний розподіл на [0,2], $b(t) = 1 - t/2$,

(4) $F \sim \mathrm{Exp}_{\lambda=1}$, $b(t) = (1+\cos(t))/2$,

(5) $F \sim \chi^2(4)$, $b(t) = 1 - 0.5/(1+x)$,

(6) $F$ — T-Стьюдента з 6-ма ступенями вільності, $b(t) = 1/(1+\exp(t-1)$,

(7) $F$ — бета-розподіл з параметрами $\alpha = 2$, $\beta = 3$, $b(t) = \cos(t)$.

(8) $F$ — розподіл Лапласа з середнім 0 та інтенсивністю 1, $b(t) = 1 - 0.5 * \exp(-t^2)$,

(9) $F \sim \mathrm{Exp}_{\lambda=0.5}$, $b(t) = 1/(t+1)$,

(10) $F$ — рівномірний розподіл на [0,1], $b(t) = 1 - t^2$.

# Розділ 2

# Оцінювання числових характеристик розподілу

Багато важливих числових характеристик розподілів випадкових спостережень можна виразити як функціонали від функції розподілу. Це, наприклад, математичне сподівання, дисперсія, медіана, інтерквартильний розмах. У розділі 1 ми розглянули кілька оцінок функції розподілу за даними різного вигляду — за вибірками кратними, цензурованими, зміщеними. Підставляючи отримані оцінки замість справжніх функцій розподілу у відповідні функціонали, можна отримувати оцінки для моментів, квантилів та інших характеристик розподілів. Така техніка оцінювання називається "метод підстановки" Цим ми займемося у даному розділі.

## 2.1 Оцінки моментів

Почнемо з, мабуть, найбільш популярної характеристики розподілу випадкових величин — математичного сподівання.

Нехай $\xi$ — випадкова величина. Тоді математичне сподівання виражається через її функцію розподілу як

$$\mu = \mathsf{E}\,\xi = \int_{-\infty}^{+\infty} xF(dx).$$

Нехай $\tilde{F}(x)$ — довільна оцінка для $F(x)$ за спостережуваними даними.

Тоді на роль оцінки для $\mathsf{E}\,\xi$ природно взяти

$$\tilde{\mu} = \int_{-\infty}^{+\infty} x\tilde{F}(dx).$$

Таку оцінку для $\mu$ називають оцінкою методу підстановки.

Оцінки, розглянуті у розділі 1, є функціями чистих стрибків, тобто їх можна записати у вигляді

$$\tilde{F}(x) = F(x; \mathbf{w}) = \sum_{j=1}^{m} w_j \mathbb{I}\{x_j < x\},$$

де $x_j$ — деякі фіксовані точки на числовій прямій (зазвичай це ті значення, які реально спостерігались у ході дослідження) а $w_j$ — фіксовані вагові коефіцієнти. У цьому випадку інтеграл по $\tilde{F}(x)$ перетворюється на суму:

$$\tilde{\mu} = \sum_{j=1}^{m} w_j x_j.$$

Якщо $w_j$ задовольняють умовам $w_j \geq 0$ і $\sum_{j=1}^{m} w_j = 1$, то їх можна розглядати як набір ймовірностей того, що деяка умовна випадкова величина $\tilde{\xi}$ приймає значення $x_j$:

$$\mathsf{P}\{\tilde{\xi} = x_j\} = w_j.$$

Тоді

$$\tilde{\mu} = \mathsf{E}\,\tilde{\xi}.$$

Наприклад, у випадку кратної вибірки $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ на роль $\tilde{F}(x; \mathbf{w})$ природно обрати звичайну емпіричну функцію розподілу. У випадку, коли всі спостереження є різними, точки стрибків є значеннями елементів вибірки: $\xi_j = x_j$, а вагові коефіцієнти — всі однакові: $w_j = 1/n$. Відповідно,

$$\tilde{\mu} = \hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} \xi_j,$$

тобто оцінкою для математичного сподівання є вибіркове середнє. Легко переконатись, що ця оцінка не зміниться і в тому випадку, коли серед вибіркових значень зустрічаються однакові.

У випадку цензурованої вибірки[1] $\mathbf{X} = (\xi_j, \delta_j)_{j=1}^N$ знову обмежимось випадком, коли всі спостережувані значення $\xi_j$ є різними. Тоді для оцінювання функції розподілу можна скористатись оцінкою Каплана — Мейера, що обчислюється за формулою (1.9). У цьому випадку моментами стрибків будуть вибіркові значення, впорядковані по зростанню $x_j = \xi_{[j]}$, а висоти стрибків визначаються як

$$w_j = \hat{w}_j^{KM} = \hat{F}_n^{KM}(x_j+) - \hat{F}_n^{KM}(x_j)$$

$$= \prod_{i:\ \xi_{[i]}<x_j} \left(1 - \frac{\tilde{\delta}_i}{n-i+1}\right) - \prod_{i:\ \xi_{[i]}\le x_j} \left(1 - \frac{\tilde{\delta}_i}{n-i+1}\right).$$

Помітимо, що при $\tilde{\delta}_j = 0$ цей стрибок дорівнює 0, тобто значення у вибірці, що відповідають цензурованим спостереженням, можна не включати у набір точок стрибків. Це може бути корисно при організації обчислень, але для теоретичної інтерпретації це не суттєво. Будемо називати $w_j^{KM}$ ваговими коефіцієнтами Каплана — Мейера.

Зауважимо, що при розгляді оцінок Каплана — Мейера у п. 1.2.3 ми вводили у набір точок стрибків оцінки формальну точку $x_{n+1} = +\infty$. Зрозуміло, що при оцінці середнього такий підхід не приведе до хорошого результату. Тому краще $x_{n+1}$ не включати у формулу для оцінки середнього, а її навантаження $w_{n+1}$ додати до навантаження точки $x_n = \max_j \xi_j$.

Оцінка математичного сподівання при використанні такого підходу матиме вигляд:

$$\tilde{\mu} = \hat{\mu}^{KM} = \int_0^\infty x\hat{F}^{KM}(dx) = \sum_{j=1}^n w_j^{KM}\xi_{[j]}.$$

Це — навантажене середнє з ваговими коефіцієнтами $w_j^{KM}$.

Якщо потрібно оцінити математичне сподівання за зміщеною вибіркою[2] $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ з функцією, що реалізує зміщення $b(t)$, то та ж техніка оцінювання приводить до

$$\hat{\mu}^{HT} = \int_{-\infty}^\infty xF^{HT}(dx) = \sum_{j=1}^n w_j^{HT}\xi_j$$

[1] Див. п. 1.2.1.

[2] див. п. 1.3.

$$= \frac{\sum_{j=1}^{n} \xi_j / b(\xi_j)}{\sum_{j=1}^{n} 1/b(\xi_j)}.$$

Тут $w_j^{HT}$ — вагові коефіцієнти Горвіца — Томпсона, див. (1.19).

У випадку, коли дані включають в себе і зміщену і незміщену вибірки, можна також скористатись ваговими коефіцієнтами Варді, див. (1.24).

Аналогічно, якщо потрібно оцінити функціональний момент з моментною функцією $h$, тобто

$$\bar{h} = \mathsf{E}\, h(\xi) = \int h(x) F(dx),$$

для цього можна використати відповідний навантажений емпіричний момент:

$$\tilde{h} = \int_{-\infty}^{\infty} h(x) F(dx; \mathbf{w}) = \sum_{j=1}^{m} w_j h(x_j) = \mathsf{E}\, h(\tilde{\xi}).$$

Багато характеристик розподілів, що використовуються на практиці, є функціями від різних теоретичних моментів. Зрозуміло, що такі характеристики можна оцінювати використовуючи відповідні функції від навантажених емпіричних моментів.

Наприклад, для дисперсії

$$\sigma^2 = \sigma_F^2 = \int_{-\infty}^{\infty} (x - \mu)^2 F(dx) = \mathsf{E}\, \xi^2 - (\mathsf{E}\, \xi)^2,$$

природною оцінкою буде

$$\tilde{\sigma}^2 = \int_{-\infty}^{\infty} (x - \tilde{\mu})^2 F(dx; \mathbf{w})$$

$$= \sum_{j=1}^{m} w_j (x_j - \tilde{\mu})^2 = \sum_{j=1}^{m} w_j (x_j)^2 - (\tilde{\mu})^2 = \mathrm{D}\, \tilde{\xi}.$$

В залежності від того, за якими даними відбувається оцінювання, на роль вагових коефіцієнтів можна використовувати константу $(1/n)$, коефіцієнти Каплана — Мейера, Горвіца — Томпсона або Варді, так само, як ми це робили для оцінювання математичного сподівання.

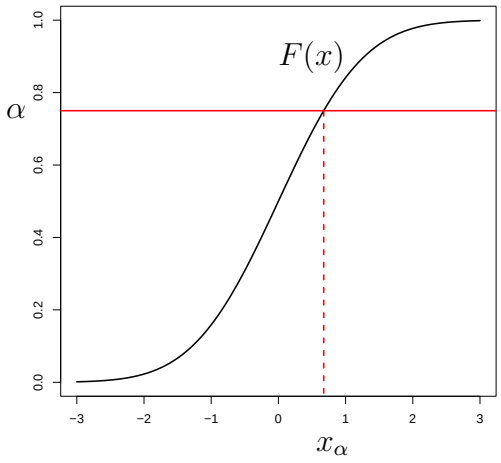



Рис. 2.1: Стандартна нормальна функція розподілу $F(x)$ (чорна лінія), рівень $\alpha = 3/4$ (червона лінія), квантиль $x_\alpha = Q^{N(0,1)}(\alpha)$.

## 2.2 Оцінювання квантилів розподілу

Крім моментів важливими характеристиками розподілу випадкової величини є квантилі. Зокрема, медіана та інтерквартильний розмах розподілу це характеристики, що визначаються через квантилі.

Почнемо з простішого означення квантиля довільного розподілу $F$. Кажуть, що $x_\alpha$ є квантилем рівня $\alpha \in (0,1)$ для розподілу $F$, якщо

$$F(x_\alpha) = \alpha.$$

Це число позначають $Q^F(\alpha) = x_\alpha$. (Див. рис. 2.1) Іншими словами, квантиль — це функція, обернена до функції розподілу: $Q^F(\alpha) = F^{-1}(\alpha)$. У термінах ймовірностей, якщо випадкова величина $\xi$ має розподіл $F$, то

$$F(x_\alpha) = \mathsf{P}\{\xi < x_\alpha\} = \alpha$$

— це також можна прийняти як означення квантиля випадкової величини $\xi$: $Q^\xi(\alpha) = Q^F(\alpha)$ — це таке число, що випадкова величина менша за це число з ймовірністю $\alpha$.

Медіана розподілу $F$ — це $\mathsf{med}(F) = Q^F(1/2)$. Квартилями називають квантилі рівнів $1/4$ (нижній квартиль), $1/2$ (медіана), $3/4$ (верхній

квартиль). Інтерквартильний розмах — різниця між верхнім і нижнім квартилями

$$\mathsf{IQR}(F) = Q^F(3/4) - Q^F(1/4).$$

Введене вище означення квантиля зручно використовувати, коли функція розподілу $F$ є неперервною і строго зростаючою для всіх $x$, для яких $0 < F(x) < 1$. Тоді обернена функція існує і квантилі визначаються однозначно. Ми обмежимося саме цим випадком.

Нехай ми маємо оцінку $\hat{F}(x)$ для $F(x)$. За логікою методу підстановки, можна спробувати визначити оцінку для квантиля $Q^F(\alpha)$ як розв'язок рівняння

$$\hat{F}(x) = \alpha \tag{2.1}$$

відносно $x$. Проблема полягає в тому, що навантажені емпіричні функції розподілу $F(x; \mathbf{w})$, які ми зазвичай використовуємо для оцінювання $F(x)$ не є ні неперервними, ні строго зростаючими. Це східчасті функції, тому для $\hat{F}(x) = F(x; \mathbf{w})$ рівняння (2.1) або не має коренів, або має їх нескінченно багато.

Тому, щоб отримати змістовну оцінку квантиля нам потрібно зробити оцінку функції розподілу більш акуратною. Для цього можна згладити $F(x; \mathbf{w})$ лінійною інтерполяцією, з'єднуючи середини стрибків прямими лініями (див. рис. 2.2).

Тут чорна східчаста лінія є графіком звичайної емпіричної функції розподілу по семи спостереженнях. Точки, що відповідають серединам стрибіків, відмічені червоними квадратиками. Червона ламана лінія отримана послідовним з'єднанням цих точок відрізками. Вона є графіком "згладженої" оцінки для справжньої функції розподілу даних. Позначимо цю згладжену оцінку $\check{F}_n(x)$.

Зрозуміло, що у такий можна побудувати згладжену оцінку $\check{F}(x; \mathbf{w})$ за будь-якою навантаженою емпіричною функцією розподілу $F(x; \mathbf{w})$

Оцінку $\hat{x}_\alpha = \hat{Q}^{\mathbf{X},\mathbf{w}}(\alpha)$ для квантиля рівня $\alpha$ тепер можна визначити як число, що задовольняє умову

$$\check{F}(x_\alpha; \mathbf{w}) = \alpha.$$

Тепер розберемося більш формально, як рахувати $\check{F}(x; \mathbf{w})$ і $= \hat{Q}^{\mathbf{X},\mathbf{w}}(\alpha)$. Для простоти обмежимось випадком, коли всі значення у вибірці $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ — різні. Переставимо їх у порядку зростання, щоб отримати варіаційний ряд:

$$\xi_{[1]} < \xi_{[2]} < \cdots < \xi_{[n]}.$$

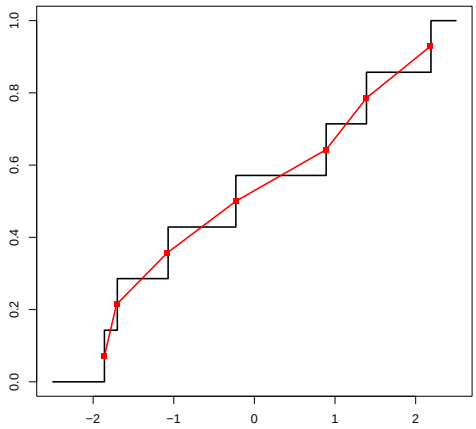



Рис. 2.2: Емпірична функція розподілу (чорна лінія) та її згладжування по серединах стрибків (червона лінія). Червоні квадрати відмічають середини стрибків.

Разом з $\xi_j$ будемо переставляти $w_j$. Позначимо $\tilde{w}_j$ — вагові коефіцієнти, переставлені у порядку зростання $\xi_j$. Тоді стрибки функції $F(x;\mathbf{w})$ відбуваються послідовно у точках $\xi_{[1]}\dots\xi_{[n]}$. Перед стрибком в точці $\xi_{[k]}$ функція $\hat{F}(x;\mathbf{w})$ набуває значення

$$\hat{F}(\xi_{[k]};\mathbf{w}) = \sum_{j=1}^{k-1} \tilde{w}_j,$$

в результаті стрибка — збільшується на 1. Отже координати точок середин стрибків (червоні квадрати на рис. 2.2) — $\xi_{[k]}$ по горизонталі, $W_k = \sum_{j=1}^{k-1} \tilde{w}_j + w_k/2$ по вертикалі.

Пряма, що з'єднує точки $(\xi_{[k]}, W_k)$ та $(\xi_{[k+1]}, W_{k+1})$, має рівняння

$$y = W_k + \frac{W_{k+1} - W_k}{\xi_{[k+1]} - \xi_{[k]}}(x - \xi_{[k]}),$$

Отже, для $x \in [\xi_{[k]}, \xi_{[k+1]}]$,

$$\check{F}(x;\mathbf{w}) = W_k + \frac{\tilde{w}_{k+1} + \tilde{w}_k}{2(\xi_{[k+1]} - \xi_{[k]})}(x - \xi_{[k]}). \tag{2.2}$$

Для того, щоб обчислити $\hat{Q}^{\mathbf{X},\mathbf{w}}(\alpha)$, потрібно знайти таке $k \in \{1, 2, \dots, k-1\}$, що $\alpha \in [W_k, W_{k+1}]$ і обчислити

$$\hat{Q}^{\mathbf{X},\mathbf{w}}(\alpha) = (\alpha - W_k)\frac{2(\xi_{[k+1]} - \xi_{[k]})}{\tilde{w}_{k+1} + \tilde{w}_k} + \xi_{[k]}.$$

Відмітимо, що не для всіх $\alpha$ можна задати оцінки квантилів у такий спосіб. При $\alpha < \tilde{w}_1$ або $\alpha > 1 - \tilde{w}_n/2$ ми не маємо другого кінця для інтерполяційного відрізку. Це може здатися недоліком методу, але насправді у даних просто немає інформації для визначення таких "хвостових" квантилів. Існують техніки які дозволяють будувати оцінки квантилів і у цьому випадку, але вони вимагають додаткових припущень щодо поведінки хвостів розподілу даних (див. [12]).

Зауважимо, що у випадку, коли $w_j = 1/n$, $j = 1, \dots, n$, а $\alpha = 1/2$, оцінка $\hat{x}_\alpha$ дорівнює вибірковій медіан

$$\mathsf{med}(\mathbf{X}) = \begin{cases} \xi_{[(n+1)/2]}, & \text{якщо } n \text{ непарне,} \\ \frac{1}{2}(\xi_{[n/2]} + \xi_{[n/2+1]}) & \text{якщо } n \text{ парне.} \end{cases} \tag{2.3}$$

**Асимптотична поведінка оцінок квантилів.** Для того, щоб користуватись такими оцінками, бажано мати результати щодо їхньої консистентності та асимптотичній нормальності. Тут ми обмежимось оцінками по кратній вибірці, побудованими за звичайними емпіричними функціями розподліу. Для оцінок квантилів за цензурованими спостереженнями чи за зміщеною вибіркою також можна отримати аналогічні результати.

Отже, нехай спостерігається кратна вибірка $\mathbf{X} = (\xi_1, \dots, \xi_n)$ з невідомою функцією розподілу $F$, а

$$\hat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{\xi_j < x\}$$

— звичайна емпірична функція розподілу. Згладжена емпірична функція $\check{F}(x)$ визначається за $\hat{F}_n(x)$ як описано вище.

Для деякого $\alpha \in (0, 1)$ розглянемо $x_\alpha = Q^F(\alpha)$. Оцінкою для $x_\alpha$ буде $\hat{x}_{\alpha;n} = \hat{Q}^{\mathbf{X}}(\alpha)$, що визначається як корінь рівняння

$$\check{F}(x) = \alpha.$$

**Теорема 2.2.1.** *1. Якщо $F$ — неперервна, строго зростаюча функція, то $\hat{x}_{\alpha;n} \to x_\alpha$.*

*2. Якщо, крім того, в деякому околі точки $x_\alpha$ існує неперервна похідна $f(x)$ функції $F(x)$ і $f(x_\alpha) > 0$, то*

$$\sqrt{n}(\hat{x}_{\alpha;n} - x_\alpha) \xrightarrow{W} N(0, \sigma^2), \tag{2.4}$$

*де*

$$\sigma^2 = \sigma^2(\alpha) = \frac{\alpha(1-\alpha)}{f^2(x_\alpha)}. \tag{2.5}$$

**Доведення. 1.** За теоремою 1.1.1 (Глівенко — Кантеллі),

$$\sup_x |\hat{F}_n(x) - F(x)| \to 0 \text{ м.н.}$$

За побудовою згладженої е.ф.р. $\check{F}_n(x)$,

$$\sup_x |\hat{F}_n(x) - \check{F}(x)| \to 0 \text{ м.н.}$$

Отже

$$\varepsilon_n = \sup_x |\check{F}_n(x) - F(x)| \to 0 \text{ м.н.}$$

Підставляючи сюди $x = \hat{x}_{\alpha;n}$ та враховуючи, що $\check{F}_n(\hat{x}_{\alpha;n}) = \alpha$, отримуємо

$$\alpha - \varepsilon_n = \check{F}(\hat{x}_{\alpha;n}) - \varepsilon_n \le F(\hat{x}_{\alpha;n}) \le \check{F}(\hat{x}_{\alpha;n}) + \varepsilon_n = \alpha + \varepsilon_n.$$

З умов п.1 теореми випливає, що функція $F^{-1}$, обернена до $F$ існує, є строго зростаючою і неперервною. Тому

$$F^{-1}(\alpha - \varepsilon_n) \le \hat{x}_{\alpha;n} \le F^{-1}(\alpha + \varepsilon_n).$$

Звідси, враховуючи неперервність $F^{-1}$ та прямування $\varepsilon_n$ до 0, отримуємо перше твердження теореми.

**2.** Доведемо тепер асимптотичну нормальність оцінок квантилів. З означення оцінки $\hat{x}_{\alpha;n}$ випливає, що, для будь-яких $x \in \mathbb{R}$, $\alpha \in (0,1)$,

$$\mathsf{P}\{\hat{x}_{\alpha;n} < x\} = \mathsf{P}\{\check{F}_n(x) > \alpha\}.$$

Фіксуємо довільне $\lambda \in \mathbb{R}$ і покладемо у цій рівності $x = x_\alpha + \lambda/\sqrt{n}$. Отримуємо

$$\mathsf{P}\{\sqrt{n}(\hat{x}_{\alpha;n} - x_\alpha) > \lambda\} = \mathsf{P}\{\check{F}_n(x_\alpha + \lambda/\sqrt{n}) - \alpha > 0\} = P_n(\lambda).$$

Нашою метою буде показати, що для буль-якого $\lambda \in \mathbb{R}$

$$P_n(\lambda) \to \Phi(\lambda/\sigma). \tag{2.6}$$

Це твердження еквівалентне (2.4).

Оцінка $\check{F}_n(x)$ визначається складно, аналізувати її важко. Але вона мало відрізняється від е.ф.р. $\hat{F}_n(x)$. Тому ми спочатку розглянемо асимптотику випадкової послідовності $\hat{F}_n(x_\alpha + \lambda/\sqrt{n}) - \alpha$, а потім скористаємось отриманим результатом для дослідження $P_n(\lambda)$. Запишемо

$$\hat{F}_n(x_\alpha + \lambda/\sqrt{n}) - \alpha = \frac{1}{n} S_n + Z_n,$$

де

$$S_n = \sum_{j=1}^{n} \eta_{j;n},$$

$$\eta_{j;n} = \mathbb{1}\{\xi_j < x_\alpha + \lambda/\sqrt{n}\} - \mathsf{E}\,\mathbb{1}\{\xi_j < x_\alpha + \lambda/\sqrt{n}\},$$

$$Z_n = F(x_\alpha + \lambda/\sqrt{n}) - \alpha.$$

Легко бачити, що $\mathsf{E}\,\eta_{j;n} = 0$,

$$\mathrm{D}\,\eta_{j;n} = F(x_\alpha + \lambda/\sqrt{n})(1 - F(x_\alpha + \lambda/\sqrt{n})) \to \alpha(1-\alpha), \text{ при } n \to \infty,$$

і $\mathsf{E}\,|\eta_{j;n}|^4 < 2^4$. Тому, за центральною граничною теоремою з умовою Ляпунова,

$$\frac{1}{\sqrt{n}} S_n \xrightarrow{W} N(0, \alpha(1-\alpha)). \tag{2.7}$$

Тепер придивимось до $Z_n$. За умовою теореми, $f(x) = \frac{d}{dx}F(x)$ є неперервною в околі $x_\alpha$, тому

$$Z_n = F(x_\alpha) + \frac{\lambda}{\sqrt{n}} f(\zeta_n) - \alpha = \frac{\lambda}{\sqrt{n}} f(x_\alpha) + \frac{\delta_n}{\sqrt{n}},$$

де $\zeta_n$ — проміжна точка між $x_\alpha$ і $x_\alpha + \lambda/\sqrt{n}$, $\delta_n \to 0$ при $n \to \infty$.

Враховуючи, що

$$\sup_x |\hat{F}_n(x) - \check{F}_n(x)| \le \frac{1}{n},$$

отримуємо

$$\frac{1}{n} S_n + \frac{\lambda}{\sqrt{n}} f(x_\alpha) - \frac{\delta_n'}{\sqrt{n}} \le \check{F}_n(x_\alpha + \lambda/\sqrt{n}) - \alpha \le \frac{1}{n} S_n + \frac{\lambda}{\sqrt{n}} f(x_\alpha) + \frac{\delta_n'}{\sqrt{n}}, \tag{2.8}$$

де $\delta'_n = |\delta_n| + 1/\sqrt{n} \to 0$.

Використовуючи (2.7), отримуємо

$$\tilde{P}_n(\lambda) = \mathsf{P}\left\{\frac{1}{n}S_n + \frac{\lambda}{\sqrt{n}}f(x_\alpha) > 0\right\} = \mathsf{P}\left\{-\frac{1}{\sqrt{n}}S_n < \lambda f(x_\alpha)\right\} \to \Phi(\lambda/\sigma),$$

для всіх $\lambda \in \mathbb{R}$.

Це майже (2.6), яке нам треба було довести, але не зовсім. Ймовірності $P_n(\lambda)$ і $\tilde{P}_n(\lambda)$ підраховуються для випадкових величин, що відрізняються лише на $\delta'_n \to 0$. Але те, що вони збігаються до тієї самої границі, варто довести акуратно. Зробимо це.

Зафіксуємо довільне $\varepsilon > 0$. Тоді, при достатньо великих $n$, $\delta'_n < \varepsilon f(x_\alpha)$ і, використовуючи нерівність праворуч у (2.8), отримуємо

$$P_n(\lambda) = \mathsf{P}\left\{\check{F}_n(x_\alpha + \lambda/\sqrt{n}) - \alpha > 0\right\} \le \mathsf{P}\left\{\frac{1}{n}S_n + \frac{\lambda}{\sqrt{n}}f(x_\alpha) + \frac{\varepsilon f(x_\alpha)}{\sqrt{n}} > 0\right\}$$

$$= \mathsf{P}\left\{-\frac{1}{\sqrt{n}}S_n < (\lambda+\varepsilon)f(x_\alpha)\right\} \to \Phi((\lambda+\varepsilon)/\sigma).$$

Отже,

$$\limsup_{n\to\infty} P_n(\lambda) \le \Phi((\lambda+\varepsilon)/\sigma).$$

Аналогічно, використовуючи нерівність ліворуч у (2.8), отримуємо

$$\liminf_{n\to\infty} P_n(\lambda) \ge \Phi((\lambda+\varepsilon)/\sigma).$$

Враховуючи неперервність $\Phi$ і довільність $\varepsilon$, отримуємо

$$\limsup_{n\to\infty} P_n(\lambda) = \liminf_{n\to\infty} P_n(\lambda) = \lim_{n\to\infty} P_n(\lambda) = \Phi(\lambda/\sigma).$$

Теорема доведена.

Цю теорему можна використати щоб побудувати довічий інтервал для справжнього квантиля $Q^F(p)$ за його оцінкою $\hat{Q}^{\mathbf{X}}(p)$. Дійсно, нехай $\hat{\sigma}^2_n(p)$ — деяка консистентна оцінка для $\sigma^2(p)$, визначеного (2.5). Покладемо

$$Q^{\mathbf{X}}_{\pm}(\alpha) = \hat{Q}^{\mathbf{X}}(p) \pm \frac{\lambda_{\alpha/2}\hat{\sigma}^2_n(p)}{\sqrt{n}}.$$

Тоді, аналогічно тому, як це зроблено у п. 1.1.2 для функцій розподілу, отримуємо

$$\mathsf{P}\{Q^F(p) \in [Q^{\mathbf{X}}_-(p), Q^{\mathbf{X}}_+(p)]\} \to 1-\alpha,$$

тобто $[Q^{\mathbf{X}}_{-}(p), Q^{\mathbf{X}}_{+}(p)]$ є асимптотичним довірчим інтервалом для квантиля $Q^F(p)$.

Для того, щоб на практиці скористатись цією ідеєю, потрібно мати оцінку$\hat{\sigma}^2_n(p)$ для $\sigma^2(p) = p(1-p)/f^(Q^F(p))$. Якби у нас була оцінка $\hat{f}(x)$ для щільності розподілу вибірки $f(x)$, ми могли б покласти

$$\hat{\sigma}^2_n(p) = \frac{p(1-p)}{\hat{f}(\hat{Q}^{\mathbf{X}}(p))}.$$

Різні оцінки для $f$ розглядаються у наступних розділах. Наприклад, на роль $\hat{f}_n$ можна взяти ядерну оцінку щільності (4.13) з параметром згладжування, вибраним за правилом Сілвермана (4.14).

## 2.3 Запитання і задачі

**Запитання.**

1. Опишіть оцінку для математичного сподівання за цензурованою вибіркою. Постарайтесь максимально спростити отриманий вираз.

2. Чому не можна використовувати звичайні вибіркові середні для оцінювання математичного сподівання за зміщеною вибіркою?

3. Запишіть у явному вигляді формулу для оцінки дисперсії у популяції за спостереженнями, отриманими зміщеною вибірковою процедурою, використовуючи вагові коефіцієнти Горвіца — Томпсона.

4. Що таке квантиль функції розподілу $F$ рівня $\alpha$? Дайте наочну інтерпретацію цього поняття.

5. Поясніть, як знайти оцінку для квантиля певного рівня $\alpha$ функції розподілу $F$, якщо у вас є для $F$ оцінка у вигляді навантаженої емпіричної функції розподілу.

**Задачі.**

1. Нехай вибірка отримана з популяції з використанням зміщеної вибіркової процедури. Функція, що реалізує зміщення — $b(x)$, розподіл спостережуваної величини у популяції — $F$. Для оцінки математичного сподівання $F$ використано вибіркове середнє спостережень. Знайдіть зміщення цієї оцінки.

2. Підрахуйте оцінку Горвіца — Томпсона для математичного сподівання у популяції за даними, описаними у задачі 5 з п. 1.4.

3. Вибірка отримана з нормального розподілу з математичним сподіванням 1 і дисперсією 4. Чому буде дорівнювати асимптотична дисперсія вибіркової медіани цієї вибірки?

4. Якщо дані є кратною вибіркою з нормального розподілу, то для оцінювання математичного сподівання можна використовувати дві оцінки: (1) вибіркове середнє і (2) вибіркову медіану. Порівняйте асимптотичні дисперсії цих оцінок — яка з них менше? Наскільки?

**Завдання для виконання на комп'ютері.**

**Завдання 1. (Оцінювання математичного сподівання за зміщеною вибіркою)**

(Продовження завдання 3 з підрозділу 1.4). Для заданих функції розподілу $F$ та функції, що реалізує зміщення $b$, згенерувати незміщену вибірку обсягу $n_1 = 300$ та зміщену вибірку обсягу $n_2 = 300$, побудувати за ними чотири оцінки для дисперсії $F$:

(1) вибіркова дисперсія, що підраховується за незміщеною вибіркою;

(2) навантажена вибіркова дисперсія з ваговими коефіцієнтами Горвіца — Томпсона, що підраховується за зміщеною вибіркою;

(3) середнє значення оцінок (1) і (2);

(4) навантажена вибіркова дисперсія з ваговими коефіцієнтами Варді.

Підрахувати оцінки (1)–(4) на згенерованій вибірці.

Повторити генерацію і обчислення оцінок $B = 1000$ разів, записуючи отримані значення у набори значень оцінок.

За отриманими наборами підрахувати дисперсії та зміщення оцінок (1)–(4).

Зробити висновок про те, яка з оцінок виявилась кращою у вашому випадку.

Значення $F$ і $b$ для вашого варіанту вибрати з таблиці завдання 3 з підрозділу 1.4.

**Завдання 2. (Асимптотика вибіркової медіани)**

У цьому завданні потрібно перевірити, при яких обсягах вибірки формула (2.5) дає хороше наближення для дисперсії вибіркової медіани. Для цього потрібно провести імітаційний експеримент. Для розподілу $F$ з вашого варіанту згенеруйте $B = 1000$ кратних вибірок. По кожній з цих вибірок обчисліть вибіркову медіану. Підрахуйте вибіркову дисперсію отриманих медіан, помножте її на $n$ і порівняйте з теоретичним асимптотичним значенням, обчисленим за (2.5). Цей експеримент проведіть для $n = 100, 500, 1000, 50000$. Зробіть висновки про можливість використання асимптотичної формули при різних $n$.

Значення розподілу $F$ для різних варіантів:

1. Нормальний розподіл з математичним сподіванням 1 і дисперсією 2.
2. Експоненційний розподіл з інтенсивністю $\lambda = 0.5$.
3. Хі-квадрат розподіл з чотирма ступенями вільності.
4. Розподіл Лапласа зі щільністю $f(x) = exp(-|x|)/2$.
5. Бета-розподіл з параметрами $\alpha = \beta = 3$.
6. Нормальний розподіл з математичним сподіванням -1 і дисперсією 1.
7. Експоненційний розподіл з інтенсивністю $\lambda = 2$.
8. Хі-квадрат розподіл з двома ступенями вільності.
9. Розподіл Лапласа зі щільністю $f(x) = exp(-|x-1|)/2$.
10. Бета-розподіл з параметрами $\alpha = \beta = 4$.

# Розділ 3

# Як можна оцінювати щільність розподілу?

Задача оцінювання щільності розподілу за спостереженнями відіграє велику роль як у теорії непараметричної статистики, так і у застосуваннях. Зокрема, багато широковживаних методів класифікації даних базуються на тих чи інших оцінках щільності (див. розділ 5). З іншого боку, на прикладі оцінок щільності зручно демонструвати різні математичні підходи, які використовуються у інших задачах непараметричного оцінювання, скажімо, у непараметричній регресії, або у спектральному аналізі випадкових процесів.

У цьому розділі ми познайомимось із основними проблемами, які виникають при побудові оцінок щільності і підходами до їх розв'язання. У розділі 4 більш детально розібрана теорія ядерних оцінок щільності, а у розділі 5 — застосування оцінок щільності до задачі класифікації.

## 3.1 Чи можна оцінити щільність стандартними методами? Оцінка Гренандера

Нехай $\mathbf{X} = (\xi_1, \dots, \xi_n)$ — кратна вибірка з щільністю розподілу одного спостереження $f(x)$, тобто

$$\mathsf{P}\{\xi_j \in A\} = \int_A f(x)dx$$

для будь-якої вимірної (борелевої) множини $A$. Ця щільність — невідома. Наше завдання — оцінити $f(x)$.

**Метод підстановки.** За логікою методу підстановки, яку ми використовували у розділі 2, можна спробувати записати $f(x)$ як деякий функціонал від функції розподілу $F$ одного спостереження і підставити у нього замість $F$ емпіричну функцію розподілу $\hat{F}_n$, порахрвану за вибіркою.

Якщо $f$ — неперервна функція, то $f(x) = \frac{d}{dx}F(x)$. Таким чином, наша оцінка повинна бути

$$\hat{f}(x) = \frac{d}{dx}F(x).$$

Але похідна емпіричної функції розподілу або дорівнює 0 (на інтервалах між стрибками), або не існує - у точках стрибка. Тому навряд чи варто сподіватись користі від такої оцінки.

Можна спробувати трактувати таку похідну як "узагальнену функцію". Тоді ця формула стане осмисленою, але, скажімо, значення функції $f(x)$ у конкретній точці $x$ за нею все одно оцінити не вдасться.

Можна замість емпіричної ф.р. підставити якусь згладжену оцінку, наприклад, ту функцію $\check{F}_n(x)$, яку ми використали для оцінювання квантилів, див. (2.2). Приклад реалізації цієї ідеї зображено на рис. 3.1. Ми бачимо, що оцінка дуже сильно коливається навколо справжньої щільності і, взагалі кажучи, зростання обсягу вибірки не поліпшує її поведінку, а, скоріше, погіршує. Це пов'язано з тим, що оцінка у кожній точці $x$ будується за двома спостереженнями, які опинились найближче до $x$ праворуч і ліворуч від неї. Значення оцінки обернено пропорційне відстані між цими спостереженнями. Тому, якщо два спостереження випадково потрапили поруч, оцінка на інтервалі між ними буде дуже великою — значно більшою, ніж справжнє значення. І це практично не виправляється наступними спостереженнями. Щоб отримати кращу оцінку, варто при згладжуванні емпіричної ф.р. враховувати не лише дві сусідні точки, а більше. Ті оцінки, які ми будуватимем у наступних підрозділах, будуть використовувати цю ідею.

**Метод найбільшої вірогідності.** Оскільки щільність $f$ однозначно задає розподіл спостережуваних даних $\mathbf{X}$, ми можемо спробувати використати для її оцінки метод найбільшої вірогідності. Зауважимо, що при цьому не виникає тих проблем, які у нас були при застосуванні цього методу до оцінювання функцій розподілу. Дійсно, оскільки ми одразу припускаємо існування щільності $f$ і спостереження у нашій вибірці не-

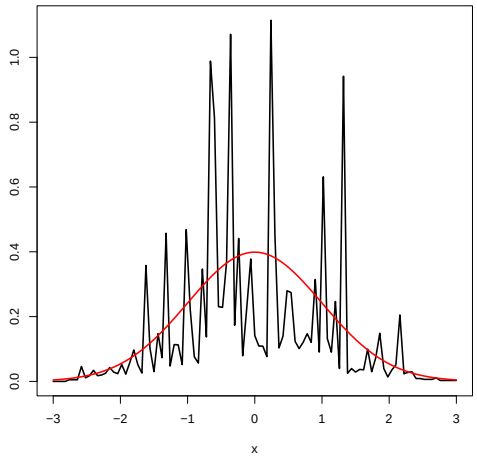



Рис. 3.1: Оцінка щільності за згладженою емпіричною ф.р. $\check{F}_n(x)$ (чорна лінія), щільність стандартного нормального розподілу (червона лінія).

залежні, однаково розподілені, то функція вірогідності для $f$ має вигляд:

$$L(f) = \prod_{j=1}^{n} f(\xi_j).$$

Оцінкою найбільшої вірогідності має бути така щільність розподілу, на якій досягається максимум $L$. Але, якщо не накладати додаткових умов на $f$, цей функціонал не обмежений зверху. Тому оцінки не існує.

Дійсно, припустимо для простоти, що всі $\xi_j$ у вибірці різні. Виберемо число $h$ менше, ніж половина найменшої відстані між сусідніми $\xi_j$. Покладемо

$$f_h(x) = \frac{1}{nh} \sum_{j=1}^{n} \mathbb{1}\{|x - \xi_j| < h/2\}.$$

Оскільки $f_h(x) \geq 0$ і $\int_{-\infty}^{+\infty} f_h(x)dx = 1$, то $f_h$ — щільність деякого розподілу на $\mathbb{R}$. При цьому

$$L(f_h) = \prod \frac{1}{nh} = \frac{1}{n^n h^n}.$$

При $h \to 0$, $L(f_h) \to \infty$, отже, яку б щільність розподілу $f^*$ ми ні обрали,

а при достатньо малому $h$ буде виконуватись

$$L(f^*) < L(f_h).$$

Отже у загальному випадку оцінка методу найбільшої вірогідності для щільності розподілу не існує.

**Оцінка Гренандера.** Однак, якщо вважати, що для щільності виконані певні додаткові умови, оцінку найбільшої вірогідності побудувати можна.

Припустимо, що спостережувані випадкові величини $\xi_j$ приймають лише додатні значення, а їхня щільність $f(x)$ є спадною (незростаючою) функцією на інтервалі $x \in [0, \infty)$. Тоді оцінку найбільшої вірогідності потрібно шукати, максимізуючи $L(f)$ по всіх $f$, що є спадними на додатній півосі. У. Гренандер [20] показав, що за цієї умови максимум функції вірогідності досягається на оцінці, яку ми зараз опишемо.

Розглянемо всі можливі неперервні, неспадні функції $G(x)$, $x \geq 0$, такі, що $G(0) = 0$, $\lim_{x\to+\infty} G(x) = 1$. Кожну таку функцію можна трактувати як функцію розподілу невід'ємної випадкової величини. Введемо додаткове обмеження на функції $G$. Будемо вимагати, щоб вони були "опуклими вгору", тобто для всіх $x_1, x_2 \in [0, \infty)$, $\gamma \in (0, 1)$,

$$G(\gamma x_1 + (1-\gamma)x_2) \geq \gamma G(x_1) + (1-\gamma)G(x_2).$$

Позначимо множину таких опуклих вгору функцій розподілу $\mathcal{G}$. Помітимо, що якщо $G \in \mathcal{G}$ має похідну $g$, то ця похідна буде незростаючою функцією. Тобто за нашою умовою, справжня функція розподілу, яку ми оцінюємо, $F \in \mathcal{G}$.

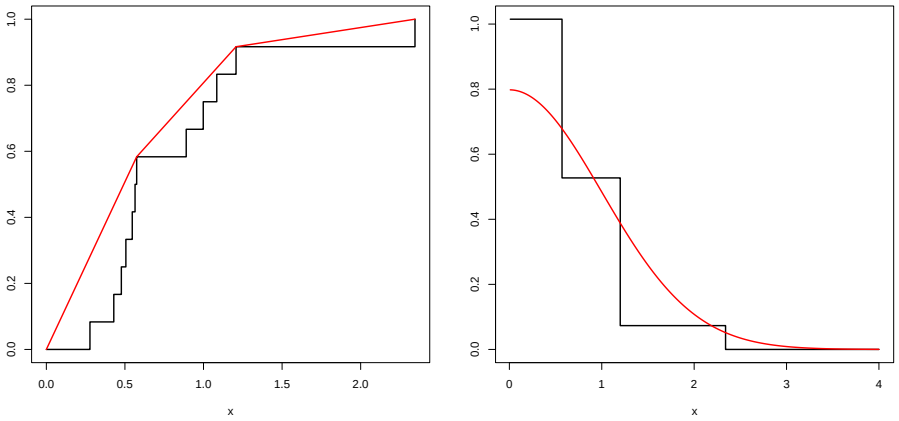
Нехай тепер $\hat{F}_n(x)$ — емпірична функція розподілу, побудована за кратною вибіркою з невід'ємних випадкових величин зі щільністю $F$. Розглянемо множину

$$\mathcal{G}(\hat{F}_n) = \{G \in \mathcal{G} \mid G(x) \geq \hat{F}_n(x) \text{ для всіх } x \geq 0\}.$$

Тобто $\mathcal{G}(\hat{F}_n)$ — це множина всіх опуклих функцій розподілу, графіки яких лежать вище, ніж графік $\hat{F}_n$. Виберемо серед цих функцій "найменшу" $\tilde{F}_n$. Тобто $\tilde{F}_n$ — це така функція, що

$$\tilde{F}_n(x) \leq G(x) \text{ для всіх } x \geq 0 \text{та всіх } G \in \mathcal{G}(\hat{F}_n).$$

Функцію $\tilde{F}_n$ називають (верхньою) опуклою огинаючою функції $\hat{F}_n$. Оскільки ми припускаємо, що справжня ф.р. $F$ що оцінюється, є опуклою, то




(a) (b)

Рис. 3.2: (a) Емпірична функція розподілу (чорна) та її опукла огинаюча (червона) (b) Оцінка Гренандера (чорна) та справжня оцінювана щільність (червона).

$\tilde{F}_n$ можна розглядати як виправлення (згладжування) $\hat{F}_n$ з урахуванням умови опуклості.

Оцінка для щільності розподілу $f(x)$ визначається тепер як похідна цієї виправленої оцінки для функції розподілу:

$$\tilde{f}_n(x) = \frac{d}{dx}\tilde{F}_n(x).$$

Цю $\tilde{f}_n(x)$ називають **оцінкою Гренандера**. На ній досягається максимум функції вірогідності при обмеженні лише спадними щільностями, отже вона є оцінкою найбільшої вірогідності.

Як обчислити $\tilde{f}_n(x)$? Уявіть собі, що ви взяли якусь функцію $G$ з $\mathcal{G}(\hat{F}_n)$ і "опускаєте" її вниз на графіку, щоб отримати огинаючу $\hat{F}_n$. Це опускання зупиниться тоді, коли графік $G$ зачепиться за точки на краю сходинок $\hat{F}_n$. Між цими точками графік витягнеться у прямі лінії. Деякі крайові точки прийдеться оминути, щоб функція залишалась опуклою вгору. Результат вийде приблизно таким, як на рис. 3.2 (a). Тут графік $\hat{F}_n$ зображений чорною лінією, а його опукла огинаюча - червоною. Точки, у яких вона дотикається до графіка $\hat{F}_n$, називають опорними точками. Оскільки графік $\tilde{F}_n(x)$ є ламаною лінією, то його похідна $\tilde{f}_n(x)$ є

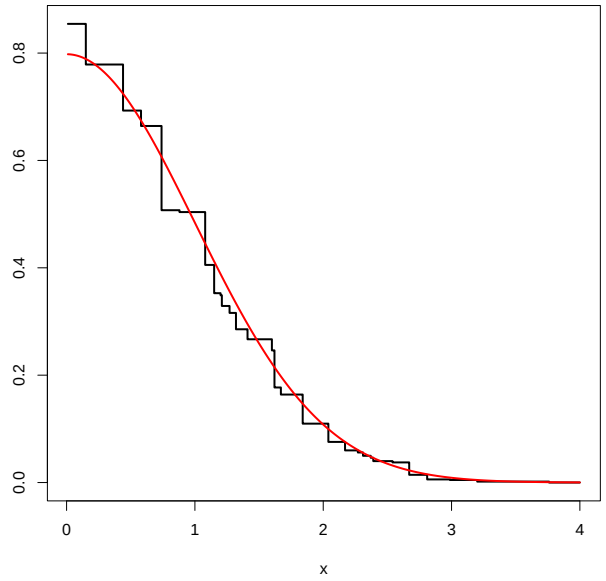



Рис. 3.3: Оцінка Гренандера по 1000 спостережень (чорна) та справжня оцінювана щільність (червона).

східчастою функцією. Її значення між опорними точками дорівнює кутовому коефіцієнту відповідної ланки ламаної. Графік $\tilde{f}_n(x)$ зображено на рис. 3.2 (b).

Для цього прикладу використана вибірка з 12 спостережень, згенерованих з півнормальним розподілом. Щільність цього розподілу зображена на рис. 3.2 (b) червоним кольором. Як і можна було сподіватись, точність оцінювання по 12 спостереженнях невелика, але загальна тенденція спадання щільності відображена відповідно. На рис. 3.3 Як бачимо, при зростанні обсягу вибірки точність оцінювання поліпшується. Як швидко оцінка наближається до оцінюваної функції? Про це каже наступна теорема з робіт [25] та [21].

**Теорема 3.1.1.** *Якщо $f'(x) < 0$ і $f'(x) = df(x)/dx$ — неперервна в околі точки $x$, то*

$$n^{1/3}(\tilde{f}_n(x) - f(x)) \xrightarrow{W} \left|\frac{1}{2}f(x)f'(x)\right|^{1/3} 2Z,$$

*де* $Z = \operatorname{argmin}_{t\in\mathbb{R}}(W(t) + t^2)$, $W(t)$ — *двосторонній Вінерів процес.*

Якщо ви не знаєте, що таке двосторонній Вінерів процес, не переймайтеся тим[1]. Для нас зараз важливо, що $Z$ — це випадкова величина, яка з ймовірністю 1 більша 0 і менша $+\infty$. При бажанні можна обчислити її функцію розподілу з довільною заданою точністю.

Таким чином, ми зустрілись з новою для нас ситуацією. У попередніх розділах розглянуті нами оцінки були асимптотично нормальними. Оцінка Гренандера такою не є, бо розподіл $Z$ не є гауссовим. І тепер у нас правильна нормуюча послідовність не $n^{1/2}$, як раніше, а $n^{1/3}$. Тобто швидкість збіжності гренандерових оцінок щільності помітно повільніша, ніж швидкість збіжності емпіричних функцій розподілу.

Далі ми побачимо, що ця особливість виникатиме і для інших оцінок щільності, які ми розглядатимем.

Оцінка Гренандера має досить обмежену область застосування, оскільки на у реальному світі припущення про монотонне спадання щільності розподілу спостережень виконується не часто. Але цю оцінку можна узагальнити, застосувавши аналогічний підхід до спостережень з унімодальною щільністю (тобто у випадку, коли $f$ має єдину точку максимуму). Про відповідні оцінки можна прочитати у [8].

## 3.2 Гістограмні та ядерні оцінки

Ми побачили у попередньому параграфі, що стандартні методи (метод підстановки, метод найбільшої вірогідності) або не дають хороших оцінок щільності, або вимагають суттєвих обмежень. Тим не менше, оцінити щільність розподілу можна і різних варіантів оцінок існує дуже багато.

**Гістограмна оцінка.** Мабуть найбільш поширеною оцінкою щільності розподілу за кратною вибіркою є гістограма відносних частот. Нагадаємо, як будується ця оцінка.

Нехай $\mathbf{X} = (\xi_1, \dots, \xi_n)$ — кратна вибірка з невідомою щільністю розподілу одного спостереження $f(x)$. Ми хочемо оцінити $f(x)$ на деякому

[1]Добре, для тих, кому це муляє. $W(t), t \in \mathbb{R}$, це гауссів випадковий процес з нульовим математичним сподіванням і коваіаційною функцією $\mathsf{cov}(W(t), W(s)) = \min(|t|, |s|)$, якщо $t$ і $s$ лежать по один бік від 0, тобто або $t > 0$, $s > 0$, або $t < 0$, $s < 0$. Якщо $t$ і $s$ лежать по різні боки від 0, то $\mathsf{cov}(W(t), W(s)) = 0$. При цьому $W(0) = 0$.

інтервалі[2] $[a, b] \in \mathbb{R}$. Для цього розбиваємо весь інтервал $[a, b]$ на $K$ підінтервалів $A_1, \ldots, A_K$ однакової довжини:

$$A_k = [t_{k-1}, t_k),\ k = 1, \ldots, K-1,\ A_K = [t_{K-1}, t_K],$$

де $t_k = a + ht$, $h = (b-a)/K$ — ширина одного підінтервалу.

Підраховуємо відносні частоти попадання спостережень $\xi_j$ у інтервали $A_k$:

$$\nu(A_k) = \nu_k = \frac{\#\{j :\ \xi_j \in A_k\}}{n}.$$

Для $x \in A_k$ визначаємо оцінку $f(x)$ як $\hat{f}_n^{hist}(x) = \nu_k/h$.

Таким чином, гістограмну оцінку щільності можна визначити як:

$$\hat{f}_n^{hist}(x) = \frac{1}{h}\sum_{k=1}^{K} \nu_k \mathbb{1}\{x \in A_k\}.$$

Графік гістограмної оцінки часто зображають у вигляді стовпчиків. Основою стовпчика є відповідний підінтервал розбиття, а його висота — значення гістограмної оцінки. Це, власне, і є гістограма (див. рис. 3.4).

Чому можна сподіватись, що гістограмна оцінка буде добре оцінювати справжню щільність?

При $n \to \infty$

$$\nu_k \to \mathsf{P}\{\xi_1 \in A_k\} = F(A_k) = \int_{t_{k-1}}^{t_k} f(t)dt \text{ м.н.}$$

за законом великих чисел, оскільки $\nu_k = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{\xi_j \in A_k\}$. Тому, при великих $n$,

$$\hat{f}_n^{hist}(x) \approx \frac{1}{t_k - t_{k-1}} \int_{t_{k-1}}^{t_k} f(t)dt \approx f(x)$$

для будь-якого $x \in A_k$, якщо $h = t_k - t_{k-1}$ — маленьке.

Отже, ми отримали $\hat{f}_n^{hist}(x) \approx f(x)$ при великих $n$ і малих $h$. Чим менше $h$, тим кращим буде наближення $\mathsf{E}\,\hat{f}_n^{hist} = F(A_k)/h \approx f(x)$, тобто тим менше зміщення оцінки.

[2]На практиці часто вибирають кінці інтервалу $a = \min_j \xi_j$, $b = \max_j \xi_j$, але це не обов'язково.

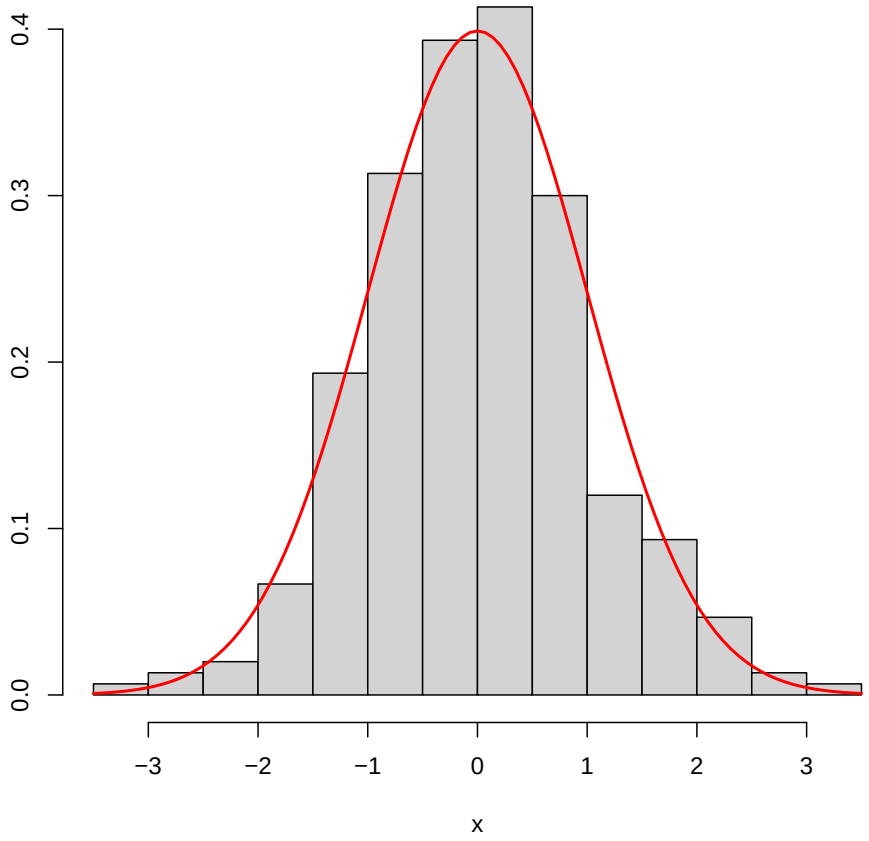
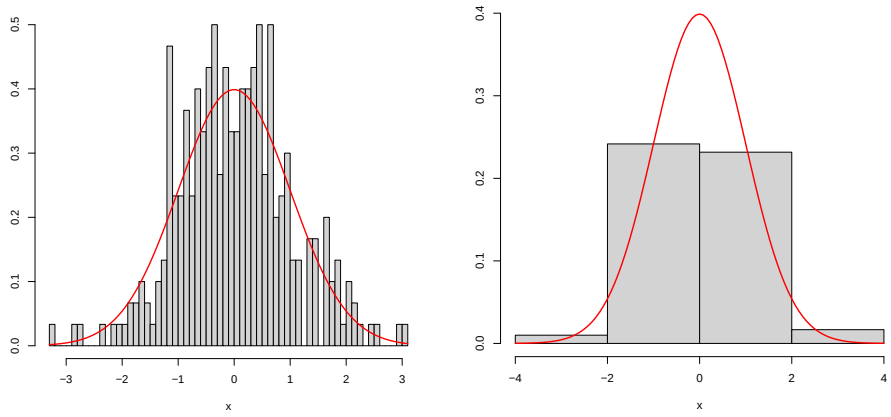



Рис. 3.4: Гістограма відносних частот за вибіркою з 300 спостережень з стандартного нормального розподілу (чорна) та справжня оцінювана щільність (червона).




(a) (b)

Рис. 3.5: (a) Гістограма з 50 підінтервалами розбиття — великий розкид (b) Гістограма з чотирма підінтервалами — велике зміщення. Справжня оцінювана щільність — стандартна нормальна (червона).

З іншого боку, зі зменшенням $h$ збільшується дисперсія, тобто розкид оцінки. Дійсно, для $x \in A_k$,

$$\mathrm{D}\,\hat{f}(x) = \mathrm{D}\,\frac{1}{nh}\sum_{j=1}^{n} \mathbb{1}\{\xi_j \in A_k\}$$

$$= \frac{1}{n^2h^2}\sum_{j=1}^{n} F(A_k)(1-F(A_k)) \approx \frac{f(x)h}{nh^2} = \frac{f(x)h}{nh}.$$

Отже, зменшуючи $h$ ми зменшуємо зміщення оцінки, але збільшуємо дисперсію. Що це означає для оцінки можна подивитись на рис. 3.5. Тут побудовані гістограми за тими самими даними, що і на рис. 3.4, але з іншою кількістю підінтервалів. На рис. 3.5 (a) кількість підінтервалів $K = 50$ занадто велика. В результаті стовпчики "стрибають", коливаючись приблизно навколо справжньої щільності. Тобто розкид оцінки великий, а зміщення мале (непомітне порівняно з розкидом). На рис. 3.5 (b) обрано дуже мало підінтервалів, $K = 4$. В результаті розкид стовпчиків практично непомітний, але значення оцінки далеко відхиляються від

справжньої щільності в околі піку. Можна сказати, що тепер гістограмна оцінка зрізала пік щільності.

Оскільки при зростанні обсягу вибірки $n$ дисперсія зменшується, ми можемо забезпечити консистентість оцінки, вибираючи $h = h_n$ в залежності від $n$ так, щоб $h_n \to 0$, але $h_n n \to \infty$ при $n \to \infty$. Ця рекомендація задає занадто широкі межі для вибору $h$. Наприклад, їй задовольняють $h_n = Cn^{-\beta}$ для будь-яких $C > 0$ та $0 < \beta < 1$. Зрозуміло, що для практичного вибору ширини підінтервалу $h$, або, що еквівалентно, кількості інтервалів розбиття $K$, потрібна формула без довільних констант.

Параметр $h$ та подібні йому називають "параметрами згладжування" оцінки. Це — параметр налаштування, ми вибираємо його самі, щоб забезпечити хорошу роботу оцінки. Параметри налаштування не слід плутати з невідомими параметрами розподілу, які існують незалежно від статистичних алгоритмів і потребують оцінювання.

У розділі 4 ми детальніше розглянемо питання оптимального вибору параметра згладжування, аналогічного $h$, але для ядерних оцінок щільності. Тут вкажемо лише практичні алгоритми, які дозволяють автоматично обирати кількість інтервалів розбиття $K$. Ці алгоритми дають не оптимальні оцінки, але їхні результати є достатньо хорошими для першої прикидки, особливо, коли гістограму відображають на екрані і статистик може уточнити вибір $K$, якщо бачить незадовільний результат.

Отже, за правилом Сторджеса (Sturges' formula),

$$K = \lceil \log_2 n \rceil + 1,$$

де $\lceil x \rceil$ позначає найменше ціле число, більше, або рівне $x$.

За правилом Піса (Rice rule),

$$K = \lceil 2n^{1/3} \rceil + 1.$$

Є й інші формули, жодна з них не є найкращою для всіх можливих випадків.

**Ядерні оцінки.** Очевидним недоліком гістограмної оцінки є те, що вона має розриви навіть тоді, коли оцінювана щільність є гладенькою. Причому положення точок розривів ніяк не пов'язане з даними, а задається самим статистиком.

Щоб усунути цей недолік, можна спробувати не фіксувати інтервали розбиття, а змусити інтервал, по якому підраховується кількість спостережень, рухатись разом із точкою в якій оцінюється щільність.

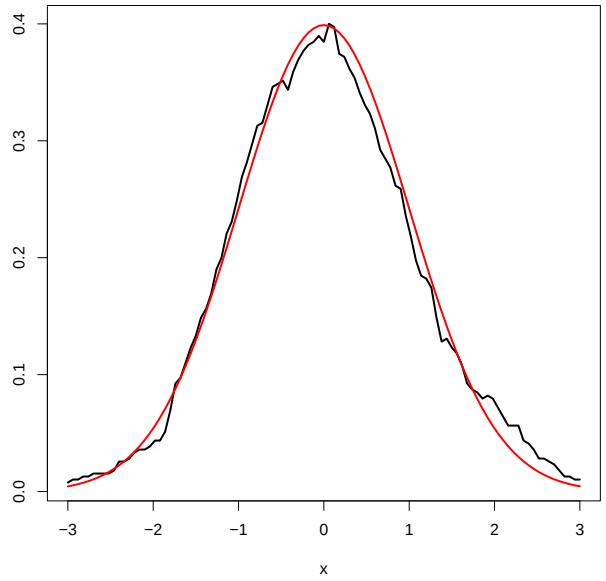



Рис. 3.6: Оцінка ковзаючого вікна за вибіркою з 300 спостережень з стандартного нормального розподілу (чорна) та справжня оцінювана щільність (червона). Ширина вікна $h = 1.3$.

Нехай щільність треба оцінити у точці $x$. Візьмемо інтервал ширини $h$ з серединою в точці $x$, підрахуємо відносну частоту попадань спостережень у цей інтервал —

$$\nu_n(x) = \frac{1}{n}\#\{j:\ \xi_j \in [x - h/2, x + h/2]\}.$$

Тепер на роль оцінки щільності можна взяти

$$\hat{f}_n^{rw} = \frac{\nu_n(x)}{h} = \frac{1}{nh}\sum_{j=1}^{n} \mathbb{1}\left\{|x - \xi_j| < \frac{h}{2}\right\}. \tag{3.1}$$

Цю оцінку щільності називають оцінкою ковзаючого вікна. Застосування цієї оцінки до даних, розглянутих вище, див. на рис. 3.6. Як бачимо, оцінка на око здається досить точною. Помітні випадкові коливання оцінки, але загалом вона добре відтворює форму оцінюваної щільності.

Хоча на рисунку розриви оцінки непомітні, насправді вони є, оскільки оцінка являє собою суму розривних функцій (індикаторів). Ці розриви маленькі і не так впадають в око, як на гістограмі, однак хотілося б отримати дійсно неперервну оцінку.

Для цього можна замінити індикатори у сумі в означенні оцінки (3.1) на які-небудь гладенькі функції, за своєю поведінкою близькі до індикаторів. Ця ідея приводить до власне ядерних оцінок.

Ядерною оцінкою щільності з ядром $K : \mathbb{R} \to \mathbb{R}$ і параметром згладжування $h > 0$ називають

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - \xi_j}{h}\right). \tag{3.2}$$

Якщо обрати у 3.2 ядро $K(x) = \mathbb{1}\{|x| < 1/2\}$, то ядерна оцінка перетворюється на оцінку ковзаючого вікна. Це ядро називають прямокутним. Інший популярний вибір — гауссове ядро, $K(x) = (2\pi)^{-1/2} \exp(-x^2/2)$, тобто на роль ядра обирають щільність стандартного гауссового розподілу. Зрозуміло, що ядерна оцінка з прямокутним ядром матиме розриви, а з гауссовим — буде неперервною і диференційовною.

Параметр згладжування $h$ у ядерній оцінці відіграє роль аналогічну ширині вікна у $f_n^{rw}$ та ширині підінтервалів у гістограмній оцінці. При великому $h$ зміщення оцінки буде великим а розкид (дисперсія) - малим. При зменшенні $h$ розкид зменшується, а дисперсія оцінки зростає. Для правильного вибору параметра згладжування потрібно вміти забезпечити баланс між зміщенням та розкидом.

Асимптотичному дослідженню поведінки ядерних оцінок щільності та можливим підходам до вибору ядра і параметра згладжування присвячено розділ 4.

## 3.3 Проекційні оцінки щільності

Для того, щоб обчислити яку-небудь оцінку з розглянутих у п. 3.2, потрібно знати всі елементи вибірки. Якщо оцінка обчислюється один раз у одній фіксованій точці, то, звичайно, без цього не обійтися. Але часто оцінки використовуються багаторазово, їх обчислюють у багатьох різних точках. При такому застосуванні кожного разу перебирати всі елементи вибірки для для знаходження чергового значення оцінки може бути

неефективно. А саме так ми робимо, коли обчислюємо, наприклад, ядерні оцінки щільності. Хотілося б мати алгоритм, який з великої вибірки виділяв би найбільш важливу інформацію для оцінювання щільності і зберігав її у невеликому обсязі пам'яті. Коли ж виникає потреба у обчисленні оцінки у конкретній точці, використовується лише ця стиснена інформація, а вибірку можна взагалі забути. У цьому підрозділі ми розглянемо техніку оцінювання, яка дозволяє таке використання.

Нехай, як і раніше, спостерігається кратна вибірка $\mathbf{X} = (\xi_1, \ldots, \xi_n)$, щільність елементів якої $f(x)$ потрібно оцінити. Ми будемо вважати, що для деякого інтервалу $S \subseteq \mathbb{R}$ (скінченного або нескінченного) $\int_S f(x)dx = 1$, тобто всі спостереження $\xi_j$ належать $S$ (з ймовірністю 1). $S$ називають носієм розподілу вибірки.

Щільність $f$ ми будемо трактувати як елемент функціонального простору $L_2(S)$, який складається з функцій $g : S \to \mathbb{R}$, таких, що

$$\int_S (g(x))^2 dx < \infty.$$

Як відомо, $L_2(S)$ є сепарабельним гільбертовим простором зі скалярним добутком

$$\langle f, g \rangle = \int_S f(x)g(x)dx.$$

Нормою функції $g$ у $L_2(S)$ називають

$$\|g\| = \sqrt{\langle g, g \rangle} = \sqrt{\int_S f^2(x)dx}.$$

Відмітимо, що не всі можливі щільності розподілу належать $L_2(S)$, тому припущення $f \in L_2(S)$ є додатковою умовою, яку ми накладаємо на невідому нам щільність. Ця умова виконується, зокрема, якщо $f$ — обмежена функція, тобто $f(x) < C_f < \infty$ для всіх $x \in S$. Дійсно, в цьому випадку

$$\int_S f^2(x)dx \leq \int_S C_f f(x)dx = C_f.$$

Оскільки $L_2(S)$ — сепарабельний простір, в ньому існує ортонормований базис, тобто набір функцій $v_i : S \to \mathbb{R}$, $i = 1, 2, \ldots$, таких, що

$$\|v_i\| = 1, \ \langle v_i, v_k \rangle = 0 \text{ при } i \neq k, \ i, k = 1, 2, \ldots$$

і будь-яку функцію $f \in L_2(S)$ можна розкласти за цим базисом:

$$f(x) = \sum_{i=1}^{\infty} a_i v_i(x). \tag{3.3}$$

Зафіксуємо деякий базис $v_i$, $i = 1, 2 \ldots$. Нехай $f$ у (3.3) — щільність розподілу, яку нам потрібно оцінити. Тоді задача зводиться до оцінки коефіцієнтів $a_i$. Ці коефіцієнти визначаються як

$$a_k = \langle f, v_k \rangle = \int_S v_k(x) f(x) dx = \mathsf{E}\, v_i(\xi_1).$$

Тобто, з ймовірнісної точки зору, нам потрібно оцінити функціональний момент за кратною вибіркою. Як і у п. 2.1, ми скористаємось для цього відповідним вибірковим моментом:

$$\hat{a}_k = \hat{a}_{k;n} = \frac{1}{n} \sum_{j=1}^{n} v_k(\xi_j).$$

Оскільки ряд $\sum_{k=1}^{\infty} a_k = \|f\|^2 < \infty$, то $a_k \to \infty$ при $k \to \infty$. Тому можна сподіватись, що для достатньо великих $k$ коефіцієнти $a_k$ можна замінити просто 0. В результаті отримуємо оцінку для $f$ вигляду

$$\hat{f}_n^{Pr}(x) = \sum_{j=1}^{M} \hat{a}_k v_k(x), \tag{3.4}$$

де $M$ — "параметр зрізання", ціле число, яке вибирають так, щоб сума відкинутих доданків для $k = M+1, M+2, \ldots$ була достатньо малою. (Як це зробити ми з'ясуємо пізніше).

Оцінки, визначені (3.4), називають проекційними оцінками щільності за базисом $v_i$, $i = 1, 2, \ldots$ Походження цієї назви зрозуміле. По суті, $\hat{f}_n^{Pr}(x)$ оцінює не саму щільність $f(x)$, а функцію

$$\bar{f}_M(x) = \sum_{i=1}^{M} a_i v_i(x),$$

яку можна розглядати як ортогональну проекцію функції $f$ на лінійний підпростір $V_M$ у $L_2(S)$, натягнутий на функції $v_1, \ldots, v_M$. Вимірність цього простору дорівнює $M$, тому цей параметр налаштування оцінки часто називають **вимірністю простору проекції**.

Наскільки вдалими будуть проекційні оцінки? Щоб розібратись у цьому, потрібно ввести певну характеризацію якості оцінок-функцій. Оскільки зараз ми трактуємо щільності та їх оцінки як елементи простору $L_2(S)$, природно характеризувати їхню близькість у термінах відстаней в $L_2(S)$. Популярною такою характеристикою є проінтегрований середньоквадратичний ризик (або середньоквадратична помилка, англійською мовою — mean integrated squared error, MISE). Якщо $\hat{f}_n(x)$ — довільна оцінка для функції $f(x)$, то проінтегрований середньоквадратичний ризик $\hat{f}_n$ визначається як

$$\mathsf{MISE}(\hat{f}_n) = \mathsf{E}_f \int_S (\hat{f}_n(x) - f(x))^2 dx = \mathsf{E}_f \, \|\hat{f}_n - f\|^2, \tag{3.5}$$

тут $\mathsf{E}_f$ позначає математичне сподівання в припущенні, що $f$ — справжнє значення невідомого параметру, тобто, у даному випадку — невідомої щільності розподілу.

Подивимось, як можна оцінити $\mathsf{MISE}(\hat{f}_n^{Pr})$. Помітимо, що

$$\mathsf{E}_f \, \hat{f}_n^{Pr}(x) = \mathsf{E} \sum_{k=1}^{M} \frac{1}{n} \sum_{j=1}^{n} v_k(\xi_j) v_k(x)$$

$$= \sum_{k=1}^{M} v_k(x) \, \mathsf{E} \, v_k(\xi_1) = \sum_{k=1}^{M} a_k v_k(x) = \bar{f}_M(x),$$

тобто $\hat{f}_n^{Pr}(x)$ є незміщеною оцінкою $\bar{f}_M(x)$. Тому

$$\mathsf{MISE}(\hat{f}_n) = \int_S \mathsf{E}_f[(\hat{f}_n^{Pr}(x) - \bar{f}_M(x)) + (\bar{f}_M(x) - f(x))]^2 dx$$

$$= \int_S \mathsf{E}_f(\hat{f}_n^{Pr}(x) - \bar{f}_M(x))^2 dx + \int_S (\bar{f}_M(x) - f(x))^2 dx$$

$$+2 \int_S \mathsf{E}_f(\hat{f}_n^{Pr}(x) - \bar{f}_M(x))(\bar{f}_M(x) - f(x)) dx.$$

Останній інтеграл дорівнює 0, оскільки $(\bar{f}_M(x) - f(x))$ — не випадкова функція, а $\mathsf{E}_f(\hat{f}_n(x) - \bar{f}_M(x)) = 0$.

Використовуючи означення $\hat{f}_n^{Pr}(x)$ і $\bar{f}_M(x)$, отримуємо

$$\int_S \mathsf{E}_f(\hat{f}_n^{Pr}(x) - \bar{f}_M(x))^2 dx = \int_S \mathsf{E}_f \left( \sum_{k=1}^{M} (\hat{a}_{k;n} - a_k) \right)^2 dx$$

$$= \int_S \mathsf{E}_f \left( \sum_{k=1}^{M} (\hat{a}_{k;n} - a_k) v_k(x) \right)^2 dx$$

$$= \int_S \mathsf{E}_f \sum_{k,l=1}^{M} (\hat{a}_{k;n} - a_k)(\hat{a}_{l;n} - a_l) v_k(x) v_l(x) dx$$

$$= \sum_{k,l=1}^{M} \mathsf{E}_f (\hat{a}_{k;n} - a_k)(\hat{a}_{l;n} - a_l) \langle v_k v_l \rangle$$

$$= \mathsf{E}_f \sum_{k=1}^{M} (\hat{a}_{k;n} - a_k)^2,$$

внаслідок ортонормованості $v_k$, $k = 1, 2, \ldots$.

Оскільки $\mathsf{E}_f \, \hat{a}_{k;n} = a_k$, то

$$\mathsf{E}_f \sum_{k=1}^{M} (\hat{a}_{k;n} - a_k)^2 = \sum_{k=1}^{M} \mathrm{D}_f \, \hat{a}_{k;n} = \sum_{k=1}^{M} \mathrm{D}_f \, \frac{1}{n} \sum_{j=1}^{n} v_k(\xi_j)$$

$$= \frac{1}{n} \sum_{k=1}^{M} \mathrm{D}_f \, v_k(\xi_1) \le \frac{1}{n} \sum_{k=1}^{M} \mathsf{E}_f (v_k(\xi_1))^2.$$

Крім того,

$$\int_S (\bar{f}_M(x) - f(x))^2 dx = \| \sum_{k=M+1}^{\infty} a_k v_k \|^2 = \sum_{k=M+1}^{\infty} (a_k)^2.$$

Об'єднуючи всі ці оцінки, отримуємо

$$\mathsf{MISE}(\hat{f}_n^{Pr}) \le \frac{1}{n} \sum_{k=1}^{M} \mathsf{E}_f (v_k(\xi_1))^2 + \sum_{k=M+1}^{\infty} (a_k)^2. \tag{3.6}$$

Оскільки за рівністю Парсеваля,

$$\sum_{k=1}^{M} (a_k)^2 = \|f\|^2 < \infty,$$

то, вибравши $M$ достатньо великим, можна зробити $\sum_{k=M+1}^{\infty} (a_k)^2$ як завгодно малою. При фіксованому $M$ перший доданок у правій частині (3.6)

прямує до нуля, коли обсяг вибірки прямує до нескінченності. Отже можна сподіватись, що, при правильному виборі $M$ в залежності від $n$, проекційні оцінки будуть консистентними. Як саме зробити такий правильний вибір залежить від конкретного вибору набору базисних функцій $v_k$.

**Приклад 3.3.1.** Нехай щільності розглядаються на множині $S = [-\pi, \pi]$. Розглянемо тригонометричні базисні функції:

$$v_0(x) = \frac{1}{\sqrt{2\pi}},\ v_{2k-1}(x) = \frac{\cos(kx)}{\sqrt{\pi}},\ v_{2k}(x) = \frac{\sin(kx)}{\sqrt{\pi}}, k = 1, 2, \dots$$

Легко бачити, що ці функції утворюють ортонормований базис в $L_2(S)$. Проекційна оцінка на основі цього базису матиме вигляд:

$$\hat{f}_n^{Trig}(x) = \frac{1}{2\pi} + \sum_{k=1}^{l} \hat{a}_{2k-1;n} \frac{\cos(kx)}{\sqrt{\pi}} + \sum_{k=1}^{l} \hat{a}_{2k;n} \frac{\sin(kx)}{\sqrt{\pi}}, \tag{3.7}$$

де $l$ — параметр налаштування, який визначає кількість використаних базисних функцій, він відповідає параметру $M$ у наших попередніх викладках: $M = 2l + 1$,

$$\hat{a}_{2k-1;n} = \frac{1}{n\sqrt{\pi}} \sum_{j=1}^{n} \cos(k\xi_j),\ \hat{a}_{2k;n} = \frac{1}{n\sqrt{\pi}} \sum_{j=1}^{n} \sin(k\xi_j). \tag{3.8}$$

Оцінимо $\mathsf{MISE}(\hat{f}_n^{Trig})$, використовуючи (3.6). Оскільки $|\sin(x)| \le 1$ і $|\cos(x)| < 1$ для всіх $x \in S$, то

$$\frac{1}{n} \sum_{k=1}^{M} \mathsf{E}_f(v_k(\xi_1))^2 \le \frac{M}{2\pi n}. \tag{3.9}$$

Оцінити $\sum_{k=M+1}^{\infty}(a_k)^2$ важче. Якщо не накладати додаткових умов на функцію $f$, то збіжність цього ряду може бути як завгодно повільною. Але чим більш гладенькою є функція $f$, тим швидшою буде збіжність. Для характеризації гладкості функцій скористаємось поняттям класів Гьольдера.

Класи Гьольдера визначаються порядком класу $\beta > 0$ і константою класу $L > 0$. Для $\beta$ вводиться однозначний розклад:

$$\beta = k + \alpha,$$

де $k = k(\beta) \geq 0$ ціле число, $\alpha = \alpha(\beta) \in (0, 1]$ (тобто $\alpha$ строго більше 0, але може дорівнювати 1).

**Означення.** Кажуть, що функція $f : \mathbb{R} \to \mathbb{R}$ належить класу Гьольдера на $\mathbb{R}$ порядку $\beta$ з константою $L$, якщо вона має $k$-ту похідну $f^{(k)}$ і для всіх $x_1, x_2 \in \mathbb{R}$

$$|f^{(k)}(x_1) - f^{(k)}(x_2)| \leq L|x_1 - x_2|^{\alpha}.$$

Позначення: $f \in \Sigma(\beta, L)$.

Параметр $\beta$ характеризує гладкість функцій, що належать цьому класу. Наприклад, $\beta = 1$ відповідає умові Ліпшиця. Функції, що мають обмежену другу похідну, належать класу Гьольдера порядку $\beta = 2$, тощо.

**Лема 3.3.1.** *Якщо $f \in \Sigma(\beta, L)$, то у тригонометричному базисі для деякого $\gamma < \infty$,*

$$\sum_{k=M+1}^{\infty} (a_k)^2 \leq \frac{\gamma L^2}{M^{2\beta}}.$$

(Доведення див. [13], лема 12 з п.5 розділу 12).

Припустимо, що оцінювана щільність належить класу Гьольдера $\Sigma(\beta, L)$ з відомим порядком $\beta$ і деякою константою $L$, яка може бути невідома. Об'єднуючи оцінки (3.6),(3.9) та лему 3.3.1, отримуємо

$$\mathsf{MISE}(\hat{f}_n^{Trig}) \leq \frac{M}{2\pi n} + \frac{\gamma L^2}{M^{2\beta}}.$$

Перший доданок у цьому виразі збільшується при зростанні $M$, другий — зменшується. Якщо ми хочемо вибирати $M$ в залежності від $n$ так, щоб обидва доданки мали приблизно однаковий порядок прямування до 0, то можна взяти

$$M = M_n = \lfloor \mu n^{\frac{1}{2\beta+1}} \rfloor,$$

(тут $\lfloor x \rfloor$ — найбільше ціле, що не перевищує $x$, $\mu$ — деяке фіксоване число). У цьому випадку маємо зверху оцінку для швидкості прямування до 0 проінтегрованого середньоквадратичного ризику:

$$\mathsf{MISE}(\hat{f}_n^{Trig}) \leq C n^{-\frac{2\beta}{2\beta+1}}, \tag{3.10}$$

де $C$ — деяка константа (залежна від $\mu$), котру можна визначити, якщо відома константа Гьольдера $L$ для оцінюваної щільності. На жаль, у більшості задач оцінки щільностей хороших обмежень на $L$ немає. Тому вибір оптимальний константи $\mu$ потребує додаткового аналізу.

Сформулюємо наш результат у вигляді теореми.

**Теорема 3.3.1.** *Нехай оцінювана щільність розподілу $f$ на $S = [0, 2\pi]$ належить класу Гольдера $\Sigma(\beta, L)$. Якщо покласти $l = l_n = \lfloor \lambda n^{\frac{1}{2\beta+1}} \rfloor$, де $\lambda$ — деяка константа, то для оцінки $\hat{f}_n^{Trig}$, заданої (3.7)–(3.8), для всіх $n$ виконано (3.10) з деяким $C < \infty$.*

Таким чином, чим більш гладенькою є оцінювана щільність (чим більше $\beta$), тим точніше ми можемо її оцінити. Скажімо, якщо ми припускаємо, що $f$ має першу неперервну похідну на $S$, то $f \in \Sigma(1, L)$ для деякого $L < \infty$, тобто, можна взяти $\beta = 1$. Вибравши $l \approx \lambda n^{1/3}$, ми, за теоремою 3.3.1, отримаємо $\mathsf{MISE}(\hat{f}_n^{Trig}) \le Cn^{-2/3}$. Якщо вважати, що $f$ має другу неперервну похідну, то з $\beta = 2$, отримуємо $l \approx \lambda n^{1/5}$ і, відповідно, $\mathsf{MISE}(\hat{f}_n^{Trig}) \le Cn^{-4/5}$.

Зрозуміло, що тригонометричний базис можна розглядати не тільки на інтервалі $[0, 2\pi]$, а і на будь-якому скінченному інтервалі. При цьому твердження теореми 3.3.1 зберігається — швидкість збіжність оцінок буде така сама, як на інтервалі $[0, 2\pi]$. Крім тригонометричних функцій для побудови базисів у просторі $L_2$ часто використовують системи ортогональних поліномів (наприклад, поліноми Лежандра або Ерміта). Для таких базисів також можна отримати теореми, аналогічні теоремі 3.3.1. ◀

Повернемось до питання, з якого починається цей підрозділ. Ми хотіли отримати оцінку щільності $f(x)$, для обчислення якої у різних точках $x$ не потрібно перебирати і враховувати наново всі $n$ спостережуваних значень $\xi_j$, $j = 1, ;n$. Проекційні оцінки задовольняють цю вимогу. Дійсно, для того, щоб підрахувати, скажімо, $\hat{f}_n^{Trig}(x)$ при різних значеннях $x$ за (3.7) – (3.8), треба лише один раз підрахувати коефіцієнти $\hat{a}_{k;n}$, $k = 1, \dots 2l$ за (3.8). Після цього значення спостережень $\xi_j$ можна взагалі забути, а для розрахунків $\hat{f}_n^{Trig}(x)$ для тих чи інших $x$ підставляти в (3.7) одні і ті ж вже обчислені коефіцієнти $\hat{a}_{k;n}$, $k = 1, \dots 2l$. При цьому коефіцієнтів використовується значно менше, ніж було початкових даних: ми маємо $n$ спостережень і порядку $n^{\frac{2\beta}{2\beta+1}}$ коефіцієнтів $\hat{a}_{k;n}$, $k = 1, \dots 2l$. Якщо $\beta = 1$, то мільйону елементів вибірки $n = 1000000$ відповідатиме лише сотня ($\sqrt[3]{1000000}$) коефіцієнтів.

Це називають "стисненням інформації": ми вибрали з даних великого обсягу лише ту інформацію, яка потрібна для побудови оцінки і можемо зберегти її для подальшого використання, а самі дані забути. (Звичайно, так можна робити лише тоді, коли дані потрібні тільки для отриман-

ня оцінки і не будуть використовуватись з якоюсь іншою метою). Якщо, наприклад, оцінка щільності використовується у деякому алгоритмі класифікації, подібному до описаних далі у п. 5.2, можливість стиснення інформації є важливою перевагою оцінки.

## 3.4 Запитання і задачі

**Запитання.**

1. Чому метод найбільшої вірогідності не дозволяє побудувати непараметричну оцінку щільності у загальному випадку? Чому це все ж вдається, коли будується оцінка Гренандера?

2. Поясніть, що таке "опукла огинаюча" графіка емпіричної функції розподілу, яка використовується у оцінці Гренандера.

3. Чи є оцінка Гренандера асимптотично нормальною?

4. Опишіть, як будується гістограмна оцінка щільності розподілу.

5. Як зміниться зміщення гістограмної оцінки щільності якщо збільшити кількість інтервалів розбиття? А як зміниться розкид (дисперсія) цієї оцінки?

6. Як вибирається кількість інтервалів розбиття при побудові гістограми за правилом Сторджеса?

7. Що таке ширина вікна у оцінці щільності за методом ковзаючого вікна?

8. Чому проекційні оцінки щільності називають "проекційними"?

9. В чому перевага проекційних оцінок над ядерними?

10. Яку характеристику ми використали для характеризації якості проекційних оцінок щільності?

11. Що таке клас Гьольдера $\Sigma(\beta, L)$? Яку властивість функцій, що належать $\Sigma(\beta, L)$ характеризує $\beta$.

12. Як пов'язана точність проекційної оцінки щільності з порядком класу Гьольдера $\beta$, якому насправді належить ця щільність? Чи збільшується точність із збільшенням $\beta$, чи зменшується, чи ні те, ні те?

**Задачі.** 1. Дана вибірка $\mathbf{X} = \{0.8, 0.6, 1.3, 0.0, 0.1, 0.3, 0.3, 0.1, 2.7, 0.0\}$. Побудуйте ручкою на папері графік емпіричної функції розподілу за цією вибіркою, знайдіть його опуклу огинаючу. Обчисліть значення оцінки Гренандера на всіх інтервалах, де вона відрізняється від 0 і нарисуйте її графік.

2. За вибіркою із задачі 1 побудуйте гістограму (графік гістограмної оцінки щільності) з $K = 3$ підінтервалами розбиття.

**Завдання для виконання на комп'ютері.**

**Завдання 1. (Оцінка Гренандера і гістограма)**

Напишіть функцію, що реалізує оцінку Гренандера. Згенеруйте вибірки обсягу $n = 50, 500, 1000$ спостережень з заданою щільністю $f$. Для кожної вибірки виведіть графік оцінки Гренандера, гістограму відносних частот і графік справжньої щільності $f$ на одному рисунку. Підберіть на око кількість підінтервалів розбиття для гістограм так, щоб вони якнайкраще оцінювали щільність. Опишіть результати: яка оцінка виявилась точнішою при яких обсягах вибірки?

Індивідуальні варіанти щільності $f$:

1. Щільність експоненційного розподілу з параметром $\lambda = 2$.

2. Щільність випадкової величини $|\xi|$, де $\xi$ — нормальна випадкова величина з математичним сподіванням 0 і дисперсією 1.

3. Щільність випадкової величини $\xi^2$, де $\xi$ — нормальна випадкова величина з математичним сподіванням 0 і дисперсією 1.

4. $f(x) = 2(1-x)\mathbb{1}\{x \in [0,1]\}$.

5. $f(x) = 3(1-x)^2\mathbb{1}\{x \in [0,1]\}$.

6. $f(x) = 3(1-x^2)\mathbb{1}\{x \in [0,1]\}/2$.

7. Щільність експоненційного розподілу з параметром $\lambda = 1/2$.

8. Щільність випадкової величини $|\xi|$, де $\xi$ — нормальна випадкова величина з математичним сподіванням 0 і дисперсією 2.

9. $f(x) = \cos(x)\mathbb{1}\{x \in [0, \pi/2]\}$.

10. Щільність випадкової величини $2\xi^2$, де $\xi$ — нормальна випадкова величина з математичним сподіванням 0 і дисперсією 1.

# Розділ 4

# Асиптототичне дослідження ядерних оцінок щільності

У цьому розділі ми розглядаємо задачу непараметричного оцінювання щільності розподілу $f$ за кратною вибіркою $X = (\xi_1, \ldots, \xi_n)$, тобто спостереження $\xi_j$ — незалежні, однаково розподілені випадкові величини[1]. Щільність розглядається відносно міри Лебега, тобто це функція $f$, така, що

$$\mathsf{P}\{\xi_j < x\} = \int_{-\infty}^{x} f(t)dt.$$

Функція $f$ вважається повністю невідомою, хоча іноді ми будемо накладати на неї додаткові умови, наприклад, вважати її неперервною, диференційовною, обмеженою, тощо.

## 4.1 Означення ядерних оцінок, їх зміщення та дисперсія. Консистентність

Ядерною оцінкою для щільності розподілу $f$ за кратною вибіркою $X = (\xi_1, \ldots, \xi_n)$ називають функцію

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - \xi_j}{h}\right). \tag{4.1}$$

[1]Далі у п. 4.7 ми також будемо розглядати оцінювання щільності розподілу за багатовимірними даними, тоді $\xi_j$ будуть випадковими векторами. Але поки що ми обмежуємось одновимірним випадком.

Тут $K : \mathbb{R} \to \mathbb{R}$ — функція, яку називають ядром (kernel) оцінки,
$h > 0$ — число, яке називають параметром згладжування (bandwidth).

Для того, щоб ядерна оцінка щільності сама була щільністю деякого розподілу, потрібно, щоб

$$\int_{-\infty}^{\infty} K(z)dz = 1$$

і

$$K(z) > 0 \text{ для всіх } z \in \mathbb{R}.$$

Виконання першої умови ми далі у цьому розділі будемо вимагати всюди, а від невід'ємності ядра інколи буває потрібно відмовитись.

Ядро і параметр згладжування є параметрами налаштування (tunning parameters), які ми обираємо самі, намагаючись зробити оцінку якомога точнішою. Для цього треба схарактеризувати точність оцінки у ймовірнісних термінах, що ми і зробимо.

Підрахуємо математичне сподівання $\hat{f}_n(x)$:

$$\mathsf{E}\,\hat{f}_n(x) = \frac{1}{h}\,\mathsf{E}\,K\left(\frac{x-\xi_j}{h}\right)$$

$$= \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-t}{h}\right) f(t)dt = \int_{-\infty}^{\infty} K(z)f(x-hz)dz,$$

(Тут ми зробили заміну змінної $z = (x-t)/h$).

Таким чином, **зміщення ядерної оцінки** має вигляд:

$$\mathsf{bias}(\hat{f}_n(x)) = \mathsf{E}\,\hat{f}_n(x) - f(x) = \int_{-\infty}^{\infty} K(z)(f(x-hz) - f(x))dz. \qquad (4.2)$$

Чим менше $h$, тим ближче $f(x-hz)$ до $f(x)$ і, отже, зміщення буде зменшуватись із зменшенням $h$.

Аналогічно можна оцінити **дисперсію** $\hat{f}_n(x)$:

$$\mathrm{D}\,\hat{f}_n(x) = \frac{1}{nh^2}\,\mathrm{D}\,K\left(\frac{x-\xi_j}{h}\right)$$

$$\leq \frac{1}{nh^2}\,\mathsf{E}\,K\left(\frac{x-\xi_j}{h}\right)^2 = \frac{1}{nh}\int_{-\infty}^{\infty}(K(z))^2 f(x-hz)dz.$$

Припустимо, що оцінювана щільність є обмеженою: $\sup_x f(x) \le C_f < \infty$ і

$$d^2 = \int_{-\infty}^{\infty} (K(z))^2 dz < \infty. \tag{4.3}$$

Тоді

$$\mathrm{D}\,\hat{f}_n(x) \le \frac{d^2 C_f}{nh}. \tag{4.4}$$

Права частина цієї нерівності збільшується при зменшенні $h$. Але вона також зменшується при зростанні обсягу вибірки $n$.

Таким чином, при зменшенні параметра згладжування $h$, зміщення оцінки зменшується, а розкид (дисперсія) зростає. Правильний вибір $h$ повинен забезпечувати баланс між зміщенням і дисперсією. При цьому, чим більше $n$ тим менше $h$ можна взяти, не роблячи дисперсію занадто великою.

Наступна теорема показує, як треба вибирати $h$ в залежності від $n$, щоб ядерна оцінка була консистентною.

**Теорема 4.1.1.** *Нехай виконуються наступні умови.*

*1. Оцінювана щільність $f$ є неперервною і обмеженою на $\mathbb{R}$.*

*2. $d^2 = \int_{-\infty}^{\infty} K^2(z)dz < \infty$.*

*3. $h = h_n \to 0$, $nh_n \to \infty$ при $n \to \infty$.*

*Тоді $\hat{f}_n(x) \xrightarrow{P} f(x)$ при $n \to \infty$.*

*Доведення.* Покажемо, що в умовах теореми, зміщення і дисперсія $\hat{f}_n(x)$ прямують до 0. Для дисперсії це безпосередньо випливає з (4.4) і умов 2 і 3 теореми.

Обмеженість $f$ означає, що існує константа $C_f$, така, що $f(x) < C_f$ для всіх $x \in \mathbb{R}$.

Для зміщення у формулі (4.2) з умови 1 теореми, отримуємо

$$K(z)(f(x - h_n z)) - f(x)) \to 0, \text{ при } n \to \infty \text{ для всіх } z \in \mathbb{R}.$$

Крім, того

$$|K(z)(f(x - h_n z)) - f(x))| \le 2C_f K(z)$$

і

$$\int_{-\infty}^{\infty} 2C_f K(z) < \infty.$$

За теоремою Лебега про мажоровану збіжність звідси випливає, що

$$\mathsf{bias}(\hat{f}_n(x)) = \int_{-\infty}^{\infty} K(z)(f(x-hz) - f(x))dz \to 0.$$

Отже, з урахуванням (4.4),

$$\mathsf{E}(\hat{f}_n(x) - f(x))^2 = \mathrm{D}\,\hat{f}_n(x) + [\mathsf{bias}(\hat{f}_n(x))]^2 \to 0,$$

тобто $\hat{f}_n(x) \to f(x)$ у середньому квадратичному. Зі збіжності у середньому квадратичному випливає збіжність за ймовірністю.

Теорема доведена.□

## 4.2 Оптимальна швидкість збіжності на класах Гьольдера

Теорема 4.1.1 показує, що оцінка збігається туди, куди потрібно - до оцінюваної функції. Але ця теорема не дає можливості оцінити швидкість цієї збіжності і вибрати таку послідовність $h_n$, для якої збіжність найшвидша. Для того, щоб це зробити, потрібно накласти умови на функцію $f$, яку ми оцінюємо. Такі умови зручно накладати у термінах класів Гьольдера. Нагадаємо означення цих класів.

Класи Гьольдера визначаються порядком класу $\beta > 0$ і константою класу $L > 0$. Для $\beta$ вводиться однозначний розклад:

$$\beta = k + \alpha,$$

де $k = k(\beta) \geq 0$ ціле число, $\alpha = \alpha(\beta) \in (0, 1]$ (тобто $\alpha$ строго більше 0, але може дорівнювати 1).

**Означення.** Кажуть, що функція $f : \mathbb{R} \to \mathbb{R}$ належить класу Гьольдера на $\mathbb{R}$ порядку $\beta$ з константою $L$, якщо вона має $k$-ту похідну $f^{(k)}$ і для всіх $x_1, x_2 \in \mathbb{R}$

$$|f^{(k)}(x_1) - f^{(k)}(x_2)| \leq L|x_1 - x_2|^{\alpha}.$$

Позначення: $f \in \Sigma(\beta, L)$.

Параметр $\beta$ характеризує гладкість функцій, що належать цьому класу. Наприклад, $\beta = 1$ відповідає умові Ліпшиця. Функції, що мають обмежену другу похідну, належать класу Гьольдера порядку $\beta = 2$, тощо.

Крім того, нам будуть потрібні додаткові умови на ядро оцінки.

**Означення.** Кажуть, що $K : \mathbb{R} \to \mathbb{R}$ — ядро $m$-того порядку, якщо:

$$\int_{-\infty}^{\infty} K(z)dz = 1,$$

$$\int_{-\infty}^{\infty} z^l K(z)dz = 0, \text{ для } l = 1, \dots, m-1,$$

$$\int_{-\infty}^{\infty} |z^m K(z)|dz < \infty.$$

**Теорема 4.2.1.** *Нехай для деяких $\beta$ і $L$ виконуються такі умови.*

*1. $f \in \Sigma(\beta, L)$.*

*2. $K$ є ядром $k(\beta)+1$ порядку.*

*3. $h = Hn^{-1/(2\beta+1)}$, де $H > 0$ — деяка константа.*

*Тоді для деякого $C < \infty$,*

$$\mathsf{E}(\hat{f}_n(x) - f(x))^2 \le Cn^{-2\beta/(2\beta+1)}. \tag{4.5}$$

**Зауваження.** 1. Для використання теореми не обов'язково знати $L$, але $\beta$ знати потрібно, щоб правильно обирати швидкість прямування $h$ до 0.

2. Величину $\mathsf{E}(\hat{f}_n(x) - f(x))^2$ називають **середньоквадратичним ризиком** оцінки $\hat{f}_n(x)$ у точці $x$ і позначають $\mathsf{MSE}(\hat{f}_n(x))$ (mean squared error). Ця величина характеризує відхилення оцінки від оцінюваної функції в точці $x$. Чим менше $\mathsf{MSE}(\hat{f}_n(x))$ тим краща оцінка. Як ми знаємо,

$$\mathsf{MSE}(\hat{f}_n(x)) = \mathrm{D}\,\hat{f}_n(x) + (\mathsf{bias}(f_n(x)))^2,$$

тобто характеризація якості оцінки у термінах середньоквадратичного ризику дозволяє врахувати як розкид, так і зміщення оцінки.

3. Швидкість збіжності у цій теоремі є оптимальною (непокращуваною) якщо не накладати додаткових умов на оцінювану щільність. Далі ми звідки це випливає і доведемо (4.5).

**Доведення.** Для оцінки зміщення за формулою (4.2) скористаємось розкладом $f(x - hz)$ за формулою Тейлора із залишковим доданком у формі Лагранжа:

$$f(x - hz) = f(x) + \sum_{l=1}^{k-1} \frac{(-zh)^l}{l!} f^{(l)}(x) + \frac{(-zh)^k}{k!} f^{(k)}(\zeta_z),$$

де $\zeta_z$ — деяка проміжна точка між $x$ і $x - hz$.

Підставляючи цей розклад у (4.2) з урахуванням того, що $K$ — ядро $k+1$-го порядку, отримуємо

$$\mathsf{bias}(\hat{f}_n(x)) = \frac{(-h)^k}{k!} \int_{-\infty}^{\infty} z^k K(z) f^{(k)}(\zeta_z) dz.$$

Знову, враховуючи, що $K$ — ядро $k+1$ порядку, маємо

$$\int_{-\infty}^{\infty} z^k K(z) f^{(k)}(x) dz = 0.$$

Тому

$$|\,\mathsf{bias}(\hat{f}_n(x))| \le \frac{(h)^k}{k!} \int_{-\infty}^{\infty} |z^k K(z)| \times |f^{(k)}(\zeta_z) - f^{(k)}(x)| dz \le \frac{(h)^k}{k!} \int_{-\infty}^{\infty} |z^k K(z)| \times L|\zeta_z - x|^\alpha dx$$

$$\le \frac{(h)^k}{k!} h^\alpha \int_{-\infty}^{\infty} |z|^\beta |K(z)| dz$$

(тут ми скористались нерівністю (4.5), розкладом $\beta = k + \alpha$ і тим, що $|\zeta_z - x| \le |hz|$).

Знову з умови 2 теореми отримуємо

$$\int_{-\infty}^{\infty} |z|^\beta |K(z)| dz < \infty.$$

Тому для деякої константи $C_b$, отримуємо

$$|\,\mathsf{bias}(\hat{f}_n(x))| \le C_b h^\beta.$$

Таким чином, з урахуванням (4.4), отримуємо

$$\mathsf{E}(\hat{f}_n(x) - f(x))^2 = \mathrm{D}\,\hat{f}_n(x) + [\mathsf{bias}(\hat{f}_n(x))]^2 \le \frac{d^2 C_f}{nh} + C_b^2 h^{2\beta}. \qquad (4.6)$$

Легко бачити, що мінімум по $h$ у правій частині (4.4) досягається при

$$h = H n^{-1/(2\beta+1)}, \; H = \left( \frac{d^2 C_f}{2\beta C_b^2} \right)^{1/(2\beta+1)}.$$

Підставивши цей вираз у (4.6), отримуємо (4.5).□

З теореми 4.2.1 випливає, що чим більш гладенькою ми вважаємо оцінювану щільність, тим точніше її вдається оцінити, використовуючи ядерні оцінки з правильно підібраним ядром та параметром згладжування.

Наприклад, якщо можна вважати, що щільність $f$ задовольняє умову Ліпшиця, то, поклавши $\beta = 1$, $h = Hn^{-1/3}$, отримуємо

$$\mathsf{MSE}(\hat{f}_n(x)) \leq Cn^{2/3}.$$

Для щільностей, які мають обмежену другу похідну, взявши $\beta = 2$, $h = Hn^{-1/5}$, і використовуючи ядро другого порядку, матимем

$$\mathsf{MSE}(\hat{f}_n(x)) \leq Cn^{4/5}.$$

Коли $\beta$ збільшується до нескінченності, порядок збіжності $\mathsf{MSE}$ наближається до $Cn^{-1}$ — це та швидкість збіжності $\mathsf{MSE}$, яка характерна для параметричних задач оцінювання.

Цікаво, що у для ядерних оцінок щільності ми отримали таку саму оптимальну швидкість збіжності, як і для проекційних у підрозділі 3.3. Насправді, можна показати, що жодна оцінка для щільності розподілу $f$ не може мати кращої швидкості збіжності ніж $\sim Cn^{2\beta/(2\beta+1)}$ для всіх $f \in \Sigma(\beta, L)$ (див. [16], підрозділ 5 розділу IV). Таким чином, і проекційні і ядерні оцінки є оптимальними за порядком швидкості збіжності на класах Гьольдера.

## 4.3 Асимптотична нормальність ядерних оцінок щільності

Теорема 4.2.1 дозволяє визначити оптимальний порядок збіжності для параметра згладжування, але для вибору константи $H$ потрібен більш акуратний аналіз асимптотичної поведінки ядерних оцінок. У цьому параграфі ми проведемо такий аналіз і покажемо, що ці оцінки за досить широких умов є асимптотично нормальними. Але саме трактування поняття асимптотичної нормальності оцінок при цьому зміниться порівняно з тим, яке у нас було для параметричного випадку.

Отже, ми розглядаємо ядерні оцінки щільності вигляду

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - \xi_j}{h}\right)$$

для справжньої щільності розподілу $f$ кратної вибірки $X = (\xi_1, \ldots, \xi_n)$.

Як і раніше,

$$d^2 = \int_{-\infty}^{\infty} K^2(z)dz.$$

**Теорема 4.3.1.** *Нехай виконуються наступні умови.*

*1. $f$ є неперервною обмеженою функцією.*

*2. $d^2 < \infty$.*

*3. $h = h_n \to 0$, $nh_n \to \infty$ при $n \to \infty$.*

*Тоді*

$$\sqrt{nh_n}(\hat{f}_n(x) - \mathsf{E}\,\hat{f}_n(x)) \xrightarrow{W} N(0, f(x)d^2) \text{ при } n \to \infty.$$

**Доведення** проведемо слідуючи [9], розділ 1, п.10, теорема 1. Розглянемо зображення

$$\zeta_n = \sqrt{nh_n}(\hat{f}_n(x) - \mathsf{E}\,\hat{f}_n(x)) = \sum_{j=1}^{n} \eta_j,$$

де

$$\eta_j = \frac{1}{\sqrt{nh}} \left( K\left(\frac{x - \xi_j}{h}\right) - \mathsf{E}\,K\left(\frac{x - \xi_j}{h}\right)\right). \tag{4.7}$$

Для дослідження асимптотичної поведінки $\sum_{j=1}^{n} \eta_j$ застосуємо наступну центральну граничну теорему у схемі серій з умовою Ліндеберга.

**Теорема 4.3.2.** *Розглянемо набір серій випадкових величин $\eta_{j;n}$, $j = 1, \ldots, n$, $n = 1, 2, \ldots$*

*Позначимо $\zeta_n = \sum_{j=1}^{n} \eta_{j;n}$.*

*Нехай виконуються наступні умови.*

*(i) При фіксованому $n$, $\eta_{j;n}$, $j = 1, \ldots, n$ — незалежні в сукупності випадкові величини.*

*(ii) $\mathsf{E}\,\eta_{j;n} = 0$, $\sigma_{j;n}^2 = \mathsf{E}(\eta_{j;n})^2 < \infty$ для всіх $j = 1, \ldots, n$, $n = 1, 2, \ldots$*

*(iii) Позначимо $B_n = \sum_{j=1}^{n} \sigma_{j;n}^2$. Для будь якого $C > 0$*

$$J(C) = \sum_{j=1}^{n} \mathsf{E}(\eta_{j;n})^2 \mathbb{1}\left\{\frac{|\eta_{j;n}|}{\sqrt{B_n}} > C\right\} \to \infty \text{ при } n \to \infty. \tag{4.8}$$

*Тоді*

$$\frac{\zeta_n}{\sqrt{B_n}} \xrightarrow{W} N(0, 1) \text{ при } n \to \infty.$$

(Умова (4.8) має назву умова Ліндеберга, а сума $J(C)$ — сума Ліндеберга).

Перевіримо виконання умов цієї теореми для $\eta_{j;n}$, визначених (4.7). Оскільки $\xi_j$ — незалежні при різних $j$, умова (i) виконана.

Те, що $\mathsf{E}\,\eta_{j;n} = 0$ випливає безпосередньо з (4.7). Позначимо $\bar{f}_n(x) = \mathsf{E}\,\hat{f}_n(x)$. Обчислимо

$$\sigma_{j;n}^2 = \mathrm{D}\,\eta_{j;n} = \frac{1}{nh}\,\mathsf{E}\left(K\left(\frac{x-\xi_j}{h}\right) - h\bar{f}_n(x)\right)^2$$

$$= \frac{1}{nh}\left[\int_{-\infty}^{+\infty} K^2\left(\frac{x-t}{h}\right) f(t)dt - (h\bar{f}_n(x))^2\right]$$

$$= \frac{1}{n}\int_{-\infty}^{+\infty} K^2(z)f(x-hz)dz - \frac{h}{n}(\bar{f}_n(x))^2.$$

За доведеним у теоремі 4.1.1, $\bar{f}_n(x) \to f(x)$ при $n \to \infty$.

Отже

$$B_n = \sum_{j=1}^{n}\sigma_{j;n}^2 = \int_{-\infty}^{+\infty} K^2(z)f(x-hz)dz - h(\bar{f}_n(x))^2 \to f(x)d^2. \qquad (4.9)$$

Перевіримо виконання умови Ліндеберга. Оскільки $\eta_{j;n}$ однаково розподілені при $j = 1, \ldots, n$,

$$J(C) = n\,\mathsf{E}\,\eta_{j;n}^2 \mathbb{1}\{|\eta_{j;n}|/\sqrt{B_n} > C\}.$$

Помітимо, що

$$\eta_{j;n}^2 = \frac{1}{nh}\left(K\left(\frac{x-\xi_j}{h}\right) - \mathsf{E}\,K\left(\frac{x-\xi_j}{h}\right)\right)^2$$

$$\leq \frac{2}{nh}\left(K\left(\frac{x-\xi_j}{h}\right)^2 + \mathsf{E}\,K\left(\frac{x-\xi_j}{h}\right)^2\right).$$

Враховуючи, що $B_n \to f(x)d^2 \in (0,\infty)$, для перевірки умови Ліндеберга досить показати, що $J'(C) \to 0$ при $n \to \infty$, де

$$J'(C) = \frac{2}{h}\,\mathsf{E}\,K\left(\frac{x-\xi_j}{h}\right)^2 \mathbb{1}\{|\eta_{j;n}| > C\} + \frac{2}{h}h^2(\bar{f}_n(x))^2\,\mathsf{E}\,\mathbb{1}\{|\eta_{j;n}| > C\}.$$

Другий доданок у правій частині цього виразу прямує до 0 при $n \to \infty$, оскільки $h = h_n \to 0$ і $\bar{f}_n(x) \to f(x)$.

Оцінимо перший доданок:

$$\frac{1}{h}\,\mathsf{E}\,K^2\left(\frac{x-\xi_j}{h}\right)\mathbb{1}\left\{\frac{1}{\sqrt{nh}}\left|K\left(\frac{x-\xi_j}{h}\right)-h\bar{f}_n(x)\right|>C\right\}$$

$$=\int_{-\infty}^{\infty}K^2(z)f(x-hz)\mathbb{1}\{|K(z)-h\bar{f}_n(x)|>\sqrt{nh}C\}dz\to 0,$$

за теоремою Лебега про мажоровану збіжність, оскільки, при $n \to \infty$, $nh \to \infty$ і, отже, індикатор під знаком інтегралу прямує до 0.

Таким чином, умова Ліндеберга виконана і за центрального граничною теоремою отримуємо

$$\zeta_n/\sqrt{B_n} \xrightarrow{W} N(0,1).$$

Оскільки $B_n \to f(x)d^2$, звідси випливає твердження теореми.

□

Зосередимось тепер на випадку, коли оцінювана щільність є двічі неперервно диференційовною функцією. (І, отже, належить $\Sigma(2, L)$ для деякого $L$). Тоді, аналогічно тому, як це було зроблено у п. 4.2, якщо на роль $K$ використовувати ядро другого порядку, то

$$\mathsf{E}\,\hat{f}_n(x)-f(x)=\int_{-\infty}^{\infty}K(z)\left(-hzf^{(1)}(x)+\frac{1}{2}h^2z^2f^{(2)}(\zeta_z)\right)dz.$$

Враховуючи, що для ядер другого порядку $\int_{-\infty}^{\infty} zK(z)dz = 0$, отримуємо

$$h^{-2}(\mathsf{E}\,\hat{f}_n(x)-f(x))\to\frac{1}{2}f^{(2)}(x)D, \tag{4.10}$$

де, як і раніше, $D = \int_{-\infty}^{\infty} z^2K(z)dz$.

Оптимальним за порядком збіжності вибором для параметра згладжування у даному випадку буде

$$h = Hn^{-1/5}.$$

З таким параметром згладжування з теореми 4.3.1 та (4.10) отримуємо

$$n^{2/5}(\hat{f}_n(x)-f(x))\xrightarrow{W}\zeta\sim N\left(\frac{1}{2}f^{(2)}(x)DH^2,\frac{1}{H}f(x)d^2\right). \tag{4.11}$$

Таким чином, твердження про асимптотичну нормальність ядерних оцінок, яке ми отримали, відрізняється від звичайних для параметричного випадку тим, що

1. Нормуюча послідовність для відхилення оцінки від оцінюваної величини дорівнює не $\sqrt{n}$, а $n^{2/5}$, тобто порядок збіжності повільніший, ніж у параметричному випадку. (Якщо оцінювана щільність має інший порядок гладкості, порядок збіжності теж може бути іншим).

2. Граничний нормальний розподіл має ненульове математичне сподівання. Це математичне сподівання можна назвати "асимптотичним зміщенням". У параметричному випадку асимптотично нормальні оцінки були асимптотично незміщеними.

Отже, порівнюючи різні оцінки щільності і вибираючи оптимальні параметри налаштування для них, потрібно враховувати не одну числову характеристику для оцінки (граничну дисперсію — коефіцієнт розсіювання), а три: порядок правильної нормуючої послідовності, асимптотичне зміщення та асимптотичну дисперсію.

## 4.4 Адаптивні ядерні оцінки. Правило Сільвермана

Досі для характеризації якості оцінок щільності ми використовували середньоквадратичний ризик

$$\mathsf{MSE}(\hat{f}_n(x)) = \mathsf{E}(\hat{f}_n(x) - f(x))^2.$$

Зрозуміло, що у різних точках $x$, в яких оцінюється щільність, він буде різним. Якщо застосовувати $\mathsf{MSE}(\hat{f}_n(x))$ для вибору параметра згладжування, то він також буде різним для різних $x$. Такий підхід (локальний вибір параметра згладжування) можливий, але для того, щоб він працював ефективно, потрібно мати багато спостережень.

Ми скористаємось іншим підходом, при якому використовують характеристику якості оцінки усереднену по всіх $x$ — **проінтегрований середньоквадратичний ризик**:

$$\mathsf{MISE}(\hat{f}_n) = \int_{-\infty}^{\infty} \mathsf{E}(\hat{f}_n(x) - f(x))^2 dx.$$

(mean integrated sqared error). Чим менше $\mathsf{MISE}(\hat{f}_n)$, тим кращою є оцінка в середньому, з урахуванням її поведінки для всіх $x \in \mathbb{R}$.

Обчислити $\mathsf{MISE}(\hat{f}_n)$ на практиці неможливо, оскільки оцінювана щільність $f(x)$ невідома. Але, використовуючи результати попередніх параграфів, можна отримати асимптотичні формули для $\mathsf{MISE}(\hat{f}_n)$ при $n \to \infty$.

Зробимо це неформально, вважаючи, що виконані такі умови

1. Оцінювана щільність $f$ є двічі неперервно диференційовною.

2. Ядро $K$ має другий порядок.

3. Параметр згладжування має оптимальний порядок збіжності: $h = Hn^{-1/5}$.

Тоді, виходячи з формули (4.11), можна сподіватись, що[2]

$$n^{4/5}\,\mathsf{E}(\hat{f}_n(x) - f(x))^2 \to \mathsf{E}(\zeta)^2 = \left(\frac{1}{2}f^{(2)}(x)DH^2\right)^2 + \frac{1}{H}f(x)d^2.$$

Проінтегруємо праву і ліву частини цієї формули по $x$ враховуючи, що $\int f(x)dx = 1$:

$$n^{4/5}\,\mathsf{MISE}(\hat{f}_n) \to \frac{1}{4}\varphi D^2H^4 + \frac{d^2}{H},$$

де

$$\varphi = \varphi(f) = \int_{-\infty}^{\infty}(f^{(2)}(x))^2dx.$$

(Знову, обґрунтування цього граничного переходу вимагає виконання додаткових умов, зокрема, $\varphi$ має бути скіченним).

Границя, до якої прямує $\mathsf{MISE}(\hat{f}_n)$ при правильному нормуванні, називається асимптотичним $\mathsf{MISE}$, або головною частиною $\mathsf{MISE}$ і позначається $\mathsf{aMISE}(\hat{f}_n)$.

Таким чином, ми (неформально) отримали наступній вираз для головної частини проінтегрованого середньоквадратичного ризику ядерної оцінки:

$$\mathsf{aMISE}(\hat{f}_n) = \frac{1}{4}\varphi D^2H^4 + \frac{d^2}{H}.$$

Чим менше $\mathsf{aMISE}$, тим кращою буде оцінка при великих обсягах вибірки. У цій формулі ми можемо вибирати параметр $H$ — це нормуючий множник, який визначає значення параметра згладжування — $h = Hn^{-1/5}$. Виберемо його так, щоб мінімізувати $\mathsf{aMISE}$.

[2] для акуратного обґрунтування цього граничного переходу потрібно ввести додаткові умови рівномірної інтегровності.

Взявши похідну по $H$ і прирівнявши її до 0, отримуємо:

$$H_{opt} = \left( \frac{d^2}{D^2 \varphi} \right)^{1/5}. \tag{4.12}$$

— це теоретично оптимальне значення нормуючої константи, якому відповідає **теоретично оптимальний параметр згладжування**:

$$h_{opt} = H_{opt} n^{-1/5} = \left( \frac{d^2}{nD^2 \varphi} \right)^{1/5}.$$

І знову за цією формулою не можна рахувати не знаючи $f$, бо від $f$ залежить $\varphi$.

Для того, щоб отримати реально працюючий алгоритм вибору параметра згладжування, використовують **адаптивний підхід**. А саме:

1. Спочатку оцінюють справжню щільність якою-небудь неточною "пілотною" оцінкою $\tilde{f}_n$ за даними.

2. Знаходять оцінку для $\varphi$, використовуючи пілотну оцінку:

$$\hat{\varphi} = \int_{-\infty}^{\infty} (\tilde{f}^{(2)}(x))^2 dx.$$

3. Обчислюють оцінку для оптимального значення параметра згладжування:

$$\hat{h} = \left( \frac{d^2}{nD^2 \hat{\varphi}} \right)^{1/5}.$$

У остаточній **адаптивній ядерній оцінці** на роль параметра згладжування використовується $\hat{h}$:

$$\hat{f}_n^{adapt}(x) = \frac{1}{n\hat{h}} \sum_{j=1}^{n} K\left( \frac{x - \xi_j}{\hat{h}} \right). \tag{4.13}$$

Слід мати на увазі, що у цій оцінці параметр згладжування залежить від даних і, отже, є випадковою величиною. Тому на неї теорія, розвинена у попередніх параграфах для не випадкових $h$, безпосередньо не переноситься. Але можна показати, що її асимптотика досить мало відрізняється від тієї, яку ми бачили для не випадкових $h$.

В залежності від вибору пілотної оцінки, розрізняють параметричну та непараметричну адаптацію.

**Параметрична адаптація. Правило Сілвермана.** Виберемо на роль пілотної оцінки щільність нормального розподілу з невідомими математичним сподіванням та дисперсією:

$$\psi(x; a, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right).$$

Обчисливши відповідний інтеграл, можна переконатись, що

$$\varphi(\psi(\cdot; a, \sigma^2)) = \frac{3}{8\sqrt{\pi}\sigma^5}.$$

Це дає оцінку для оптимального параметра згладжування

$$\hat{h} = \sqrt[5]{\frac{8\sqrt{\pi}d^2}{3D^2n}}\hat{\sigma}, \tag{4.14}$$

де $\hat{\sigma}$ — деяка оцінка для $\sigma$ за вибіркою.

Підставляючи у цю формулу замість $\hat{\sigma}$ традиційну оцінку

$$S_0 = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(\xi_j - \bar{\xi})^2},$$

отримуємо просте **правило Сілвермана**.

Враховуючи не робастність оцінки $S_0$, на практиці рекомендують використовувати поліпшене правило Сілвермана, в якому

$$\hat{\sigma} = \min(S_0, \mathrm{IR}/1.34),$$

де IR — інтерквартильний розмах вибірки.

Правило Сілвермана (в тому числі, і поліпшене) дає хороші значення параметра згладжування лише для оцінки щільностей, які не дуже сильно відрізняються від нормальних. Якщо оцінювана щільність дуже асиметрична, або має важкі хвости, або кілька максимумів, то параметр згладжування, розрахований за правилом Сілвермана, буде далеким від оптимального. У таких випадках краще застосовувати непараметричну адаптацію.

**Непараметрична адаптація.** На роль пілотної оцінки вибираємо також ядерну оцінку щільності

$$\tilde{f}_n(x) = \frac{1}{nh_{pilot}}\sum_{j=1}^{n} K\left(\frac{x-\xi_j}{h_{pilot}}\right)$$

з не зовсім оптимальним "пілотним" параметром згладжування $h_{pilot}$. Наприклад, можна вибрати його за правилом Сілвермана. Далі використовуємо загальну схему побудови адаптивних ядерних оцінок.

Непараметрична адаптація при великих обсягах вибірки дає значення параметра згладжування досить близькі до теоретично оптимальних. Але при порівняно малих обсягах вона може працювати гірше, ніж просте правило Сілвермана.

## 4.5 Крос-валідація для вибору параметра згладжування

Правило Сілвермана — порівняно простий алгоритм вибору параметра згладжування для ядерних оцінок щільності, який дає, хоча не оптимальні, але у багатьох випадках достатньо хороші значення. Якщо потрібно більш акуратне визначення параметра згладжування, використовують техніки, які потребують значного обсягу обчислень. Однією з них є крос-валідація.

Ми розглянемо версію цього метода, у якій метою є обрати таке $h$, що мінімізує квадрат відстані в $L_2$ між оцінкою та оцінюваною щільністю. Тобто теоретичним функціоналом якості буде

$$\mathsf{ISE}(h) = \mathsf{ISE}(\hat{f}_n) = \int_{-\infty}^{\infty} (\hat{f}_n(x) - f(x))^2 dx.$$

($\mathsf{ISE}$ розшифровується як integrated squared error). Зверніть увагу, що цей функціонал приймає випадкові значення і залежить від невідомої щільності $f$.

Ми хочемо знайти $h$, яке є достатньо близьким до точки мінімуму цього функціонала. Для цього зробимо перетворення:

$$\mathsf{ISE}(h) = \int_{-\infty}^{\infty} (\hat{f}_n(x))^2 dx - 2\int_{-\infty}^{\infty} \hat{f}_n(x) f(x) dx + \int_{-\infty}^{\infty} (f(x))^2 dx.$$

Третій доданок у цій формулі не залежить від $h$, отже його можна відкинути і це не вплине на положення точки мінімуму. Перший доданок не залежить від невідомої $f(x)$, його можна підрахувати на основі самих тільки спостережень.

У другий доданок входять і $h$ (через $\hat{f}_n(x)$) і невідома функція $f$. Його не можна ні обчислити, ні відкинути. Ми змушені оцінити його за спостереженнями. Як це зробити?

Якщо зафіксувати вибірку $X$, по якій побудована оцінка $\hat{f}_n(x)$, і розглянути іще одне нове спостереження $\xi_0$ з щільністю $f$, то

$$\mathsf{E}[\hat{f}_n(\xi_0)|\ X] = \int_{-\infty}^{\infty} \hat{f}_n(x) f(x) dx.$$

Тобто доданок, який нас цікавить, можна задати як умовне математичне сподівання при фіксованій вибірці від $\hat{f}_n(\xi_0)$, де $\xi_0$ — нове спостереження, яке не входить у ту вибірку, по якій побудована оцінка. Якби у нас було багато таких спостережень, ми могли б оцінити це математичне сподівання відповідним вибірковим середнім.

Оскільки таких нових спостережень у нас немає, скористаємось технікою крос-валідації. Розглянемо

$$\hat{f}_n^{i-}(x) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{x - \xi_j}{h}\right)$$

— це ядерна оцінка щільності, підрахована за всіма спостереженнями, крім $i$-того.

Тепер оцінимо $J(h) = \int_{-\infty}^{\infty} \hat{f}_n(x) f(x) dx$ за допомогою

$$\hat{J}(h) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_n^{i-}(\xi_i)$$

— тобто ми підставляємо в оцінку саме те спостереження, яке було викинуте при побудові цієї оцінки і потім усереднюємо отримані значення по всіх спостереженнях.

Комбінуючи цю оцінку з першим доданком у розкладі $\mathsf{ISE}$, отримуємо такий функціонал крос-валідації:

$$\mathsf{CV}(h) = \int_{-\infty}^{\infty} (\hat{f}_n(x))^2 dx - 2\hat{J}(h). \tag{4.15}$$

Значення $h$ при якому досягається мінімум $\mathsf{CV}(h)$:

$$\hat{h}_{\mathsf{CV}} = \operatorname{argmin}_{h>0} CV(h),$$

вибирається як оцінка для оптимального $h$, що мінімізує $\mathsf{ISE}(h)$.

Як показав Ч. Дж. Стоун,

$$\frac{\mathsf{ISE}(\hat{h}_{\mathsf{CV}})}{\min_h \mathsf{ISE}(h)} \to 1 \text{ м.н. при } n \to \infty.$$

Тобто вибираючи значення параметра згладжування за допомогою кросвалідації, ми, при великих обсягах вибірки, практично мінімізуємо $\mathsf{ISE}$ оцінки.

## 4.6 Оптимальний вибір ядра

Досі ми займались вибором параметра згладжування. Але ядерні оцінки щільності мають іще один параметр налаштування — ядро оцінки, $K(z)$. Нагадаю, що ядро повинно задовольняти умову

$$\int_{-\infty}^{\infty} K(z)dz = 1 \tag{4.16}$$

(інакше оцінка не буде консистентною). Бажано також, щоб $K(z) \geq 0$ для всіх $z \in \mathbb{R}$ — це гарантує невід'ємність оцінки. Якщо оцінювана щільність належить класу Гьольдера порядку 2, то, для забезпечення оптимальної швидкості збіжності, потрібно іще виконання умови

$$\int_{-\infty}^{\infty} zK(z)dz = 0. \tag{4.17}$$

Ця умова виконується, зокрема, для всіх ядер, що є парними функціями, тобто таких, що $K(z) = K(-z)$ для всіх $z$.

Ядер, що задовольняють ці умови, нескінченно багато. Найбільш популярні ядра:

1. Прямокутне: $K(z) = \mathbb{1}\{|z| < 1/2\}$,
2. Трикутне: $K(z) = (1 - |x|)\mathbb{1}\{|z| < 1\}$,
3. Гауссове: $K(z) = \exp(-x^2/2)/\sqrt{2\pi}$ — щільність стандартного гауссового розподілу,
4. Єпанєчнікова: $K(z) = \frac{3}{4}(1 - x^2)\mathbb{1}\{|x| < 1\}$.

Природно виникає питання, яке з цих ядер забезпечує найкращу точність оцінки. Ми знову обмежимось випадком, коли оцінювана щільність має другу похідну і параметр згладжування обирається як $h =$

$Hn^{-1/5}$. Тоді при використанні ядер другого порядку

$$\mathsf{aMISE}(H) = \mathsf{aMISE}(\hat{f}_n) = \frac{1}{4}\varphi D^2(K)H^2 + \frac{d^2(K)}{H}, \tag{4.18}$$

де

$$d^2(K) = \int_{-\infty}^{\infty} K^2(z)dz,$$
$$D(K) = \int_{-\infty}^{\infty} z^2 K(z)dz.$$

У цій формулі тільки $d^2(K)$ і $D(K)$ залежать від $K$.

Якщо вибрати теоретично оптимальне $H = H_{opt}$ за формулою (4.12) і підставити у (4.18), отримаємо

$$J(K) = \mathsf{aMISE}(H_{opt}) = \frac{5}{4}(d^8(K)D^2(K)\varphi)^{1/5}. \tag{4.19}$$

Це — найменше значення, якого може досягати aMISE ядерної оцінки при використанні ядра $K$. Чим воно менше, тим краще ядро.

Щоб знайти ядро що забезпечує найменше значення $J(K)$, потрібно розв'язати задачу умовної мінімізації функціонала

$$J_0(K) = d^4(K)D(K) \tag{4.20}$$

при виконанні умов (4.16) і (4.17).

Виявляється, що розв'язком цієї задачі є ядро Єпанєчнікова. Таким чином, саме воно є оптимальним для використання у ядерних оцінках коли оцінювана щільність має другу похідну.

Важливо також вміти порівнювати якість різних ядер. Скільки ми втратимо, якщо замість оптимального ядра Єпанєчнікова використаємо гауссове ядро? Для характеризації якості ядер використовують наступний підхід.

Нехай ми обрали деяке ядро $K$. Візьмемо довільне $\delta > 0$ і розглянемо ядро

$$K_\delta(z) = \frac{1}{\delta}K(x/\delta).$$

Тоді, якщо $\hat{f}_n(x; K, h)$ — ядерна оцінка з ядром $K$ і параметром згладжування $h$, то

$$\hat{f}(x; K_\delta, h) = \frac{1}{nh}\sum_{j=1}^{n} K_\delta\left(\frac{x-\xi_j}{h}\right) = \frac{1}{nh}\sum_{j=1}^{n}\frac{1}{\delta}K\left(\frac{x-\xi_j}{\delta h}\right)$$

$$= \frac{1}{n\delta h}\sum_{j=1}^{n} K\left(\frac{x-\xi_j}{\delta h}\right) = f(x; K, \delta h),$$

тобто, замість використання ядра $K_\delta$ і параметра згладжування $h$, можна взяти $K$ і $\delta h$ і отримати ту саму оцінку. Параметр $h$ ми обираємо з міркувань оптимальності оцінки, якщо ядру $K$ відповідало оптимальне значення $h = h_0$, ядру $K_\delta$ відповідатиме $\delta h_0$ і навпаки.

Тому ядра $K_\delta$ для всіх $\delta > 0$ можна вважати еквівалентними. Немає рації розглядати їх окремо, достатньо вибрати якогось одного представника з цього набору ядер. Такого представника зручно обрати так, щоб по можливості спростити формулу для aMISE (4.18). У цю формулу входять дві різних характеристики ядра: $d^2(K)$ і $D^2(K)$. Виберемо $\delta$ так, щоб виконувалось

$$d^2(K_\delta) = D^2(K_\delta).$$

Ця рівність виконується, якщо покласти

$$\delta = \delta_0 = \left(\frac{d^2(K)}{D^2(K)}\right)^{1/5}.$$

Величину $\delta_0$ називають канонічним параметром згладжування для ядра $K$, а ядро $K_{\delta_0}$ — канонічним ядром з класу ядер $K_\delta$, $\delta > 0$. Наприклад, для ядра Єпанєчнікова канонічний параметр згладжування дорівнює $\delta_0 = 15^{1/5}$, а канонічне ядро Єпанєчниікова має вигляд

$$K(z) = \frac{3}{4\times 15^{1/5}}\left(1 - \left(\frac{z}{15^{1/5}}\right)^2\right)\mathbb{1}\left(|z| < 15^{1/5}\right).$$

Тепер позначимо $T(K) = d^2(K_{\delta_0}) = D^2(K_{\delta_0})$ — спільне значення $d^2$ і $D^2$ для канонічного ядра, що відповідає $K$.

Тоді в (4.18) отримуємо

$$\mathsf{aMISE}(\hat{f}_n(\cdot; K_{\delta_0}, Hn^{-1/5}) = \left(\frac{1}{4}\varphi H^2 + \frac{1}{H}\right)T(K).$$

У цій формулі від вибору ядра залежить тільки $T(K)$. Чим менше $T(K)$, тим краще оцінка з даним ядром. Задача порівняння різних ядер зводиться до порівняння їх $T(K)$.

Виявляється, що для ядер, перерахованих на початку цього параграфу, значення $T(K)$ відрізняються досить мало (у межах 6%). Тому вибір ядра для оцінки щільності з цього набору не є принциповим у більшості задач оцінювання щільностей, які мають другу похідну.

## 4.7 Ядерна оцінка щільності за багатовимірними даними

Нехай тепер при кожному спостереженні ми отримуємо не одну числову характеристику об'єкта, а багато ($d$) характеристик. Скажімо, кожен об'єкт — людина, а її спостережувані характеристики - вага, зріст, обхват грудей, тощо. Таким чином, для $j$-того об'єкта ми маємо вектор спостережуваних характеристик $\xi_j = (\xi_j^1, \ldots, \xi_j^d)^T$ і наша вибірка $X = (\xi_1, \ldots, \xi_n)$ складається з цих векторів.

Ми вважаємо $\xi_j \in \mathbb{R}^d$ незалежними, однаково розподіленими векторами і хочемо оцінити щільність їх розподілу в $\mathbb{R}^d$, тобто таку функцію $f : \mathbb{R}^d \to \mathbb{R}$, що

$$\mathsf{P}\{\xi \in A\} = \int_A f(\mathbf{x}) d\mathbf{x} \text{ для всіх вимірних } A \in \mathbb{R}^d.$$

Зверніть увагу, що це, по суті, $d$-кратний інтеграл: $\mathbf{x} = (x^1, \ldots, x^d)^T$. Якщо у вектора $\xi_j$ є неперервно диференційовна функція розподілу $F$:

$$F(\mathbf{x}) = \mathsf{P}\{\xi_j < \mathbf{x}\} = \mathsf{P}\{\xi_j^1 < x^1, \ldots, \xi_j^d < x^d\},$$

то

$$f(\mathbf{x}) = \frac{\partial^d F(x^1, \ldots x^d)}{\partial x^1 \ldots \partial x^d},$$

тобто щільність $d$-вимірного вектора це $d$-кратна похідна по всіх змінних, які відповідають координатам вектора.

Для оцінювання $f$ можна скористатись ядерною оцінкою щільності:

$$\hat{f}_n(\mathbf{x}) = \frac{1}{(h)^d n} \sum_{j=1}^{n} K\left(\frac{1}{h}(\mathbf{x} - \xi_j)\right)$$

$$= \frac{1}{(h)^d n} \sum_{j=1}^{n} K\left(\frac{x^1 - \xi_j^1}{h}, \ldots, \frac{x^d - \xi_j^d}{h}\right).$$

Тут

$K : \mathbb{R}^d \to R$ — ядро оцінки,

$h > 0$ — параметр згладжування.

Ядро і параметр згладжування потрібно вибирати виходячи з якихось міркувань оптимальності оцінки, аналогічних тим, що були розглянуті у попередніх підрозділах для одновимірних щільностей. Для того,

щоб оцінка сама була щільністю якого-небудь ймовірнісного розподілу, потрібно, щоб

$$\int_{\mathbb{R}^d} K(\mathbf{z})d\mathbf{z} = 1$$

і $K(\mathbf{z}) \geq 0$ для всіх $\mathbf{z} \in \mathbb{R}^d$.

При таком підході до оцінювання ми використовуємо один спільний параметр згладжування для всіх координат вектора спостережень. Тобто, при оцінці, скажімо, щільності розподілу вектора, що складається з ваги і зросту людини, ми підганяємо $h$ одночасно і для ваги, і для зросту. Але ці характеристики вимірюються у неспівмірних одиницях (кілограми і сантиметри). Тому такий підхід може бути розумним лише коли всі координати вектора вимірювань є співмірними. У інших випадках краще обирати окремі параметри згладжування для кожної координати:

$$\hat{f}_n(\mathbf{x}) = \frac{1}{nh^1h^2\dots h^d}\sum_{j=1}^{n} K\left(\frac{x^1-\xi_j^1}{h^1}, \frac{x^1-\xi_j^1}{h^2}\dots, \frac{x^d-\xi_j^d}{h^d}\right). \tag{4.21}$$

де $h^i$ — параметр згладжування, обраний для $i$-тої координати вектора спостережень.

Можна іще більше узагальнити цей підхід, ввівши матричне нормування спостережень:

$$\hat{f}_n(\mathbf{x}) = \frac{1}{n\det(\mathbf{H})}\sum_{j=1}^{n} K\left(\mathbf{H}^{-1}(\mathbf{x}-\xi_j)\right), \tag{4.22}$$

де $\mathbf{H}$ — "матричний параметр згладжування" — матриця розміру $d \times d$, як правило, її обирають симетричною та додатновизначеною. Якщо на роль такого параметра взяти діагональну матрицю, отримаємо (4.21).

Для використання оцінок (4.21) і (4.22) потрібно вміти підбирати або $d$ значень параметрів згладжування, або $d \times d$ елементів матриці $\mathbf{H}$. Вже вибір одного такого параметра у попередніх підрозділах вимагав значних зусиль. Що й казати про $d$ або $d^2$ параметрів. Звичайно, тут можна запропонувати деякі розумні процедури, але вони, як правило, вимагають дуже великого обсягу даних. Тому на практиці часто застосовують компромісний варіант. А саме, обирають з яких-небудь міркувань нормуючу матрицю $\mathbf{S}$ а матричний параметр у (4.22) беруть рівним $\mathbf{H} = h\mathbf{S}$. Таким чином, задача зводиться до вибору одного параметра налаштування $h$.

У п. 3.6 [23] описана асимптотична теорія оцінок (4.22), з якої можна отримати асимптотичне зображення для проінтегрованої середньоквадратичної похибки:

$$\mathsf{MISE}(\hat{f}_n) = \int_{\mathbb{R}^d} \mathsf{E}(\hat{f}_n(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x}.$$

Для того, щоб описати це зображення, введемо ряд умов та позначень.

Припустимо, що оцінювана щільність $f$ є двічі неперервно диференційовною по всіх своїх аргументах. Позначимо $\mathcal{H}_f(\mathbf{x})$ матрицю других похідних (гессіан) для $f$. Ми будемо припускати, що

$$\int_{\mathbb{R}^d} (\mathcal{H}_f(\mathbf{x}))^2 d\mathbf{x} < \infty.$$

Зробимо також припущення про ядро:

$$\int_{\mathbb{R}^d} \mathbf{z} K(\mathbf{z}) d\mathbf{z} = 0,$$

(під інтегралом стоїть векторнозначна функція, тому у правій частині 0 це $d$-вимірний нульовий вектор),

$$\int_{\mathbb{R}^d} \mathbf{z}\mathbf{z}^T K(\mathbf{z}) d\mathbf{z} = \mu_2(K)\mathbb{E}$$

(тут $\mathbb{E}$ — одинична матриця розміру $d \times d$, $\mu_2(K)$ — число, що грає ту саму роль, яку величина $D(K)$ грає в одновимірному випадку).

Крім того, вважатимем, що

$$d^2(K) = \int_{\mathbb{R}^d} (K(\mathbf{z}))^2 d\mathbf{z} < \infty$$

Ці умови на ядро аналогічні вимозі, щоб мало другий порядок у одновимірному випадку.

За цих умов можна записати таке асимптотичне зображення:

$$\mathsf{MISE}(\hat{f}_n) \sim \frac{1}{4}\mu_2^2(K) \int [\mathrm{tr}(\mathbf{H}^T \mathcal{H}_f(\mathbf{x})\mathbf{H})]^2 d\mathbf{x} + \frac{d^2(K)}{n \det(\mathbf{H})} \tag{4.23}$$

при $n \to \infty$, $\mathbf{H} \to 0$.

Якщо тепер зафіксувати нормуючу матрицю $\mathbf{S}$ і покласти $\mathbf{H} = h\mathbf{S}$, отримуємо

$$\mathsf{MISE}(\hat{f}_n) \sim \frac{1}{4}\mu_2^2(K)h^4 \int [\mathrm{tr}(\mathbf{S}^T \mathcal{H}_f(\mathbf{x})\mathbf{S})]^2 d\mathbf{x} + \frac{d^2(K)}{n(h)^d \det(\mathbf{S})}. \qquad (4.24)$$

Мінімізуючи праву частину по $h$, отримуємо, що оптимальне $h$ має порядок збіжності

$$h_{opt} \sim Cn^{-1/(4+d)},$$

і цьому вибору $h$ відповідатиме порядок збіжності $\mathsf{MISE}$ —

$$\mathsf{MISE}(\hat{f}_n) \sim Cn^{-4/(4+d)}.$$

Помітимо, що, чим більша вимірність спостереження, тим гіршою є швидкість збіжності ядерної оцінки. При $d = 1$ маємо $\mathsf{MISE}(\hat{f}_n) = Cn^{-4/5}$ — це гірше, ніж $\sim n^{-1}$, як у параметричному випадку, але не набагато. Якщо $d = 10$, то $\mathsf{MISE}(\hat{f}_n) = Cn^{-4/14}$. Це вже значне погіршення, порівняно з параметричним випадком. Скажімо, якшо ви хочете поліпшити $\mathsf{MISE}$ оцінки у 10 разів, то у параметричному випадку вам потрібно збільшити об'єм вибірки також у 10 разів. Для одновимірної ядерної оцінки щільності вибірку прийдеться збільшити у 18 разів. А для десятивимірних спостережень — у 3162 рази. При $d > 10$ ядерні оцінки щільності стають практично непридатними навіть при оптимальному виборі параметрів налаштування і великих обсягах вибірки.

Слід відмітити, що таке катастрофічне погіршення якості оцінок щільності[3] властиве не тільки ядерним оцінкам, а й будь-яким іншим непараметричним оцінкам щільності. Швидкість збіжності $\sim Cn^{-4/(4+d)}$ є непокращуваною за порядком, якщо не накладати додаткових умов на ту щільність, яку ми оцінюємо.

Цей ефект має назву "прокляття багатовимірності" (curse of dimensionality). Непараметричні оцінки не має рації використовувати безпосередньо для аналізу багатовимірних даних. Тому коли для кожного досліджуваного об'єкта спостерігається багато змінних, що його характеризують, намагаються спочатку "знизити вимірність даних", тобто вибрати невелику кількість узагальнених характеристик, що описують цей об'єкт, а вже потім застосовувати непараметричні техніки до дослідження цих

[3]Це стосується і багатьох інших непараметричних задач оцінювання, наприклад, оцінок для функції регресії.

характеристик. Одна з найбільш популярних простих технік зниження вимірності — метод головних компонент. Замість того, щоб намагатись досліджувати розподіл, скажімо 20-вимірного вектора спостережуваних змінних досліджуваних об'єктів, розглядають значення 2-3 перших головних компонент для кожного об'єкта. Для цих компонент оцінюють сумісну щільність розподілу, наприклад, ядерною оцінкою. А висновки про початкові змінні об'єктів роблять виходячи з їхнього наближеного представлення за головними компонентами.

Щодо вибору ядра $K$ для багатовимірної ядерної оцінки, то для цього також можна використовувати асимптотичні розклади, але на практиці застосовують простіші підходи, конструюючи багатовимірні ядра з одновимірних. Найбільш популярних таких підходів є два.

Перший спосіб - мультиплікативний: якщо потрібне $d$-вимірне ядро $K : \mathbb{R}^d \to \mathbb{R}$, то беруть одновимірне ядро $K_1 : \mathbb{R} \to \mathbb{R}$ і задають

$$K(x^1, x^2, \ldots, x^d) = K_1(x^1) \times K_1(x^2) \times \cdots \times K_1(x^d).$$

(Якщо $K_1$ — щільність якогось одновимірного розподілу, то $K$ — щільність у $\mathbb{R}^d$).

Другий спосіб — радіальна симетризація. Ми беремо одновимірне ядро $K_1$ і задаємо $K$ як

$$K(\mathbf{x}) = C_K K_1(\|\mathbf{x}\|),$$

де константу $C_K$ вибирають так, щоб

$$\int_{\mathbb{R}^d} K(\mathbf{z}) d\mathbf{z} = 1.$$

Для грубого вибору параметрів згладжування у багатовимірному випадку існує декілька евристичних правил, побудованих на ідеї адаптивного оцінювання, аналогічного розглянутому у п. 4.4. Зокрема, якщо використовується оцінка (4.21), то можна обирати

$$\hat{h}_i = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-1/(d+4)} S_i,$$

(багатовимірне правило Сілвермана) або

$$\hat{h}_i = n^{-1/(d+4)} S_i \tag{4.25}$$

(правило Скота). Тут $S_i$ — оцінка для середьоквадратичного відхилення $i$-тої координати вектора спостережень за вибіркою. Її можна рахувати використовуючи вибіркову дисперсію, або комбінуючи вибіркову дисперсію з інетрквартильним розмахом, як описано у п. 4.4.

Якщо використовується оцінка (4.22), то для вибору матричного параметра згладжування $\mathbf{H}$ можна скористатись узагальненим правилом Скота:

$$\widehat{\mathbf{H}} = n^{-1/(d+4)}\hat{\mathbf{S}}^{1/2},$$

де $\mathbf{S}$ — оцінка для коваріаційної матриці спостережень. Це може бути вибіркова коваріаційна матриця, або яка-небудь робастна оцінка коваріаційної матриці.

Неважко також модифікувати техніку крос-валідації для вибору параметра згладжування у багатовимірному випадку (див. п. 3.6 в [23]).

## 4.8 Запитання і задачі

**Запитання.**

1. Дайте означення щільності розподілу випадкової величини і випадкового вектора.

2. Які умови потрібно накласти на ядро і параметр згладжування ядерної оцінки щільності, щоб ця оцінка сама була щільністю деякого ймовірнісного розподілу?

3. За якими формулами підраховуються зміщення та дисперсія ядерної оцінки щільності?

4. Як змінюються зміщення та дисперсія ядерної оцінки щільності при зменшенні параметра згладжування?

5. За яких умов ядерні оцінки щільності є консистентними?

6. Дайте означення ядра $m$-того порядку.

7. Як змінюється точність ядерних оцінок щільності при зростанні гладкості оцінюваної щільності — збільшується, зменшується?

8. Що таке $\mathsf{MSE}(\hat{f}_n(x))$? Як ця величина пов'язана зі зміщенням та дисперсією оцінки $\hat{f}_n(x)$?

9. Чим твердження про асимптотичну нормальність ядерних оцінок щільності відрізняється від твердження про асимптотичну нормальність емпіричних функцій розподілу (а також від аналогічних тверджень для параметричних задач оцінювання)?

10. З яких міркувань при оцінюванні двічі неперервно диференційовних щільностей розподілу, параметр згладжування вибирають пропорційним до $n^{-1/5}$, де $n$ — обсяг вибірки?

11. Що таке $\mathsf{MISE}(\hat{f}_n)$ і чим ця характеристика відрізняється від $\mathsf{aMISE}(\hat{f}_n)$?

12. З яких міркувань визначається теоретично оптимальне значення параметра згладжування для оцінювання двічі неперервно диференційовних щільностей розподілу? Чому це значення неможливо безпосередньо використовувати на практиці?

13. Поясніть, з яких міркувань отримано правило Сілвермана для визначення параметра згладжування. Які особливості щільностей розподілу роблять не ефективними правило Сілвермана при їхньому оцінюванні?

14. Опишіть ідею методу крос-валідації для вибору параметра згладжування. Які оптимальні властивості оцінки забезпечує цей метод?

15. Яке ядро є оптимальним для оцінювання щільностей розподілу, що мають неперервну другу похідну? За якою характеристикою якості можна порівнювати різні ядра?

16. Які проблеми виникають при оцінюванні щільностей багатовимірних розподілів? Які методи вибору параметрів згладжування можна використовувати у цій задачі?

**Задачі.**

1. Для щільностей розподілу

(а) експоненційного з параметром $\lambda = 1$;

(б) нормального;

(в) $f(x)\frac{3}{4}(1-x^2)\mathbb{1}\{x \in [-1,1]\}$

обчисліть теоретично оптимальне значення параметра згладжування при використанні ядерної оцінки з ядром Єпанєчнікова. Обчисліть дисперсію випадкової величини зі щільністю $f$. Це значення дисперсії підставте у формулу для визначення параметра згладжування за правилом Сілвермана замість відповідної оцінки і знайдіть "теоретичне значення параметра згладжування за правилом Сілвермана".

Обчисліть $\mathsf{aMISE}$ оцінки при використанні цих параметрів згладжування. Визначте, наскільки погіршується асимптотична точність оцінки при використанні правила Сілвермана, порівняно з теоретично оптимальним значенням.

2. Обчисліть характеристику $T(K)$ для ядер:

(а) Єпанєчнікова;

(б) гауссового;

(в) прямокутного;

(г) трикутного ($K(x) = (1 - |x|)\mathbb{1}\{x \in [-1, 1]\}$).

Порівняйте отримані значення, зробіть висновок про те, наскільки погіршується якість ядерної оцінки при використанні не оптимальних ядер.

**Завдання для виконання на комп'ютері.**

**Завдання 1. (Вибір параметра згладжування).**

Згенерувати вибірку з заданим розподілом і побудувати ядерну оцінку щільності використовуючи параметр згладжування

— обраний за правилом Сілвермана (простим і поліпшеним)

— обраний з використанням непараметричної оцінки $\varphi$

— обраний крос-валідацією.

Намалювати графіки отриманих оцінок разом із оцінюваною функцією щільності. Спробувати вручну підібрати параметр згладжування на око, так, щоб оцінка найточніше відповідала оцінюваній функції. Вивести також графік ядерної оцінки з теоретично оптимальним $h = h_{opt}$. Експеримент повторити з кількома вибірками з одним і тим самим розподілом.

Спробувати зробити висновки по таких питаннях:

— чи є $h_{opt}$ дійсно оптимальним значенням параметра згладжування, чи вдається знайти краще вибором на око?

— наскільки відрізняється від оптимального $h$ обране за правилом Сілвермана? Чи доцільно його використовувати, якщо оцінювана щільність подібна до розглядуваної у Вашому варіанті?

— наскільки непараметрична адаптація поліпшує оцінку порівняно з правилом Сілвермана? Чи варто застосовувати її?

— Чи можна вважати, що метод крос-валідації дає оцінки, близькі до оптимальних?

*Зауваження.* Варіанти відрізняються ядрами оцінок $K$ та щільностями $f$, які треба оцінити. У різних варіантах висновки щодо застосовності тих чи інших методів можуть бути різними.

Всі порівняння достатньо робити на око, але, за бажанням, можна порівнювати значення $L_2$ відстаней між оцінкою та оцінюваною функцією, усереднені по $\sim 100$ вибірках. (Це емпіричний аналог MISE).

**Значення параметрів для індивідуальних завдань**

(1) $f \sim N(1, 1)$, $K$ — трикутне ядро.

(2) $f \sim \chi^2(3)$, $K$ — ядро Єпанєчнікова.

(3) $f$ — розподіл Лапласа з мат. сподіванням 0, $\lambda = 1$, $K$ — гауссове ядро.

(4) $f$ — суміш розподілів $N(-1, 1)$, $N(1, 1)$ з ймовірністю змішування $p = 1/2$, $K$ — ядро Єпанєчнікова.

(5) $f$ — симетрична трикутна щільність на [-1,1], $K$ — ядро Єпанєчнікова.

(6) $f$ — щільність бета-розподілу з $a = 2$, $b = 4$, $K$ — гауссове ядро.

(7) $f$ — щільність розподілу Фішера $F(4, 10)$, $K$ — гауссове ядро.

(8) $f$ — щільність розподілу Стьюдента $T(5)$, $K$ — ядро Єпанєчнікова.

(9) $f$ — півнормальна щільність з $\sigma = 1$, $K$ — трикутне ядро.

(10) $f$ — щільність гамма-розподілу, $\alpha = 2$, $\lambda = 1$, $K$ — гауссове ядро.

**Завдання 2. (Довірчий інтервал для медіани розподілу).**

Наприкінці підрозділу 2.2 описано спосіб побудови довірчих інтервалів для квантилів розподілу вибірки за допомогою оцінок щільності. Реалізуйте знаходження такого довірчого інтервалу у вигляді функції, використовуючи ядерну оцінку щільності. Перевірте якість отриманого довірчого інтервалу для різних обсягів вибірки на модельованих даних за схемою, описаною у Завданні 1 з підрозділу 1.4. Для генерації псевдовипадкових даних використайте таблицю варіантів із завдання 2 підрозділу 2.3.

# Розділ 5

# Задачі класифікації

У цьому розділі розглядаються задачі класифікації спостережуваних об'єктів. Ми будемо вивчати статистичні методи "класифікації з вчителем", коли статистик має певний набір об'єктів, що вже були правильно класифіковані (навчаюча або навчальна вибірка) і задача полягає в тому, щоб класифікувати новий об'єкт, статистичні властивості якого подібні до властивостей об'єктів з навчальної вибірки.

## 5.1 Баєсова класифікація

Нехай спостерігаються об'єкти $O$, кожен з яких належить одній з $M$ різних популяцій (класів). Номер популяції, якій належить об'єкт, будемо позначати $\kappa(O)$. Цей номер нам невідомий, потрібно його встановити на основі спостережуваних характеристик об'єкта. Вектор цих характеристик позначимо $\xi = \xi(O)$. Далі ми, в основному, вважаємо, що $\xi = (\xi^1, \dots, \xi^d)$ — випадковий вектор в $\mathcal{X} \subseteq \mathbb{R}^d$, хоча це може бути також випадковий елемент будь-якого вимірного простору $\mathcal{X}$.

Нам відомо, що розподіл $\xi(O)$ залежить від $\kappa(O)$ тобто від того, до якого класу належить $O$. Ми хочемо навчитись найкращим чином за $\xi(O)$ визначати (оцінювати) $\kappa(O)$.

**Приклад 5.1.1.** Класичним прикладом застосування статистичних технік класифікації є медична (або технічна) діагностика. У задачі медичної діагностики об'єктами, які класифікуються, є пацієнти, а класами, на які вони розбиті, є хвороби (діагнози) якими пацієнти хворіють. Спостережувані характеристики — це ті дані, які має лікар про конкретного

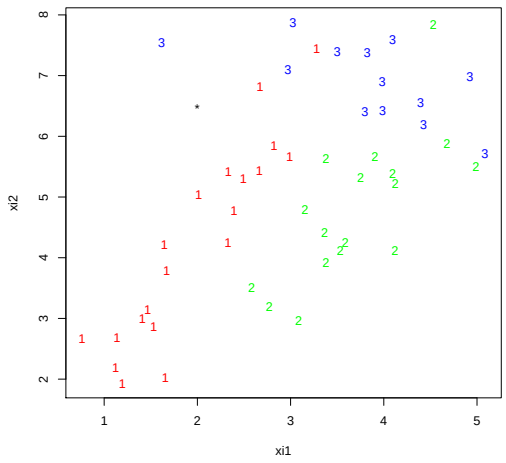



Рис. 5.1: Дані для класифікації

пацієнта при постановці діагнозу. Нехай, наприклад, пацієнт $O$, що прийшов до лікаря, скаржиться на симптоми, які дозволяють віднести його хворобу до одного з трьох видів — $M = 3$ (умовно назвемо ці хвороби "застуда" — $\kappa(O) = 1$, "грип" — $\kappa(O) = 2$ і "пневмонія" — $\kappa(O) = 3$). Лікар проводить вимірювання певних фізіологічних та біохімічних показників пацієнта і отримує вектор його спостережуваних характеристик. Для простоти відображення обмежимось двома числовими характеристиками, об'єднаними у двовимірний вектор $\xi(O) = \xi = (\xi^1, \xi^2)$. Умовно $\xi^1$ може бути виміряною температурою тіла, а $\xi^2$ — верхнім (систолічним) тиском. За цими характеристиками лікарю потрібно визначити, якою з трьох хвороб хворий пацієнт. (Тут ми припускаємо, що пацієнт не може мати двох хвороб одразу).

При постановці діагнозу лікар може враховувати статистичні дані про пацієнтів, які з подібними симптомами приходили до нього раніше. У цих пацієнтів $O_1, \ldots, O_n$ при звертанні до лікаря вимірювались ті ж характеристики $\xi(O_j) = \xi_j$ і за перебігом хвороби встановлений остаточний діагноз, який ми будемо вважати цілком правильним — це справжній клас, до якого належить пацієнт $\kappa_j = \kappa(O_j)$. Набір значень $\{(\xi_j, \kappa_j), j = 1, \ldots, n\}$ називають навчальною вибіркою (teaching sample).

Як приклад, розглянемо умовні дані, що відповідають цьому опису.

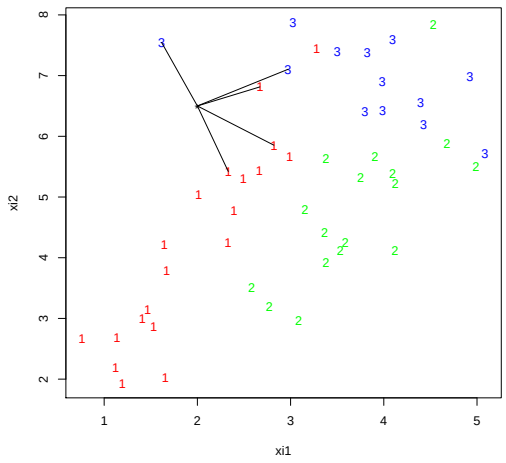



Рис. 5.2: Класифікація за методом $k = 5$ сусідів

На рисунку 5.1 зображена діаграма розсіювання даних, де точками (цифрами) зображені дані про 50 пацієнтів (об'єктів), що хворіли трьома хворобами. Координати точок відповідають першій та другій спостережуваним характеристикам, а цифри, поставлені у відповідній точці — номеру класу (хвороби) до якого насправді відноситься даний об'єкт. Це — наша навчальна вибірка.

Нехай тепер до нас прийшов новий об'єкт і потрібно його класифікувати. Значення спостережуваних характеристик об'єкта відобразимо на нашій діаграмі розсіювання зірочкою. До якого класу природно віднести цей об'єкт за його положенням на рисунку? Доцільно подивитись, яким класам належать навколишні об'єкти, для яких ми знаємо їх класифікацію і віднести новий до того класу, який поруч зустрічається найчастіше.

Формально це можна зробити, використовуючи алгоритм $k$ найближчих сусідів. За цим алгоритмом у навчальній вибірці знаходять $k$ об'єктів, у яких вектори спостережуваних характеристик лежать найближче до вектора характеристик нового об'єкта. Це — $k$ найближчих сусідів. Далі визначають, до яких класів належать найближчі сусіди і відносять новий об'єкт до того класу, який серед них зустрічається найчастіше. На рис. 5.2 відображена робота цього алгоритму для наших даних: п'ять найближчих сусідів з'єднані з новим об'єктом відрізками. Серед них найбільше (три) належать першому класу. Отже алгоритм відносить новий

об'єкт до першого класу.

Чи буде цей алгоритм працювати добре? При якому виборі $k$ він даватиме найкращі результати? Більш-менш зрозуміло, що у випадках, подібних до зображеного на рис. 5.2, годі сподіватись ідеально безпомилкової класифікації всіх можливих майбутніх об'єктів. Мова може йти лише про те, щоб помилки з'являлись не надто часто. Тому для відповіді на запитання про якість алгоритмів класифікації потрібна теоретична модель задачі і характеризація якості алгоритмів класифікації у термінах теорії ймовірностей. ◀

У цьому підрозділі ми розглянемо чисто теоретичну модель, у якій спостерігається один об'єкт $O$ з відомими ймовірнісними характеристиками. При **баєсовому підході** до задачі класифікації використовують такі характеристики:

$\pi_i = \mathsf{P}\{\kappa(O) = i\}$ — апріорна ймовірність того, що об'єкт належить $i$-тому класу;

$F_i(A) = \mathsf{P}\{\xi(O) \in A \,|\kappa(O) = i\}$ — розподіл спостережуваних характеристик об'єкта, що належить $i$-тому класу.

Ми будемо припускати, що існує деяка міра $\mu$, відносно якої всі розподіли $F_i$ мають щільності. Ці щільності позначимо $f_i$:

$$F_i(A) = \int_A f_i(\mathbf{x})\mu(d\mathbf{x}).$$

З математичної точки зору, будь-якому алгоритму класифікації відповідає функція, яка кожному можливому значенню спостережуваних характеристик об'єкта ставить у відповідність результат класифікації:

$$g : \mathcal{X} \to \{1, \ldots, M\}$$

— якщо вектор спостережуваних характеристик об'єкта $\xi(O) = \mathbf{x}$, то $g(\mathbf{x})$ — номер того класу, до якого даний алгоритм віносить цей об'єкт.

Функція $g$ називається **класифікатором**. Всюди далі ми вважаємо, що класифікатор є вимірною функцією на $\mathcal{X}$. Зараз класифікатор — це не випадкова функція.

Якість класифікатора будемо характеризувати ймовірністю помилки, яка можлива при його застосуванні. Цю ймовірність можна визначити так:

$$L(g) = \mathsf{P}\{g(\xi(O)) \neq \kappa(O)\} \tag{5.1}$$

— ймовірність того, що клас, визначений класифікатором не дорівнює справжньому класу, якому належить об'єкт.

Запишемо ймовірність помилки у термінах апріорних ймовірностей та розподілів спостережуваних характеристик різних класів.

$$L(g) = 1 - \mathsf{P}\{g(\xi(O)) = \kappa(O)\} = 1 - \sum_{m=1}^{M} \mathsf{P}\{g(\xi(O)) = \kappa(O) \text{ і } \kappa(O) = m\}$$

$$= 1 - \sum_{m=1}^{M} \mathsf{P}\{\kappa(O) = m\}\, \mathsf{P}\{g(\xi(O)) = m \mid \kappa(O) = m\}$$

$$= 1 - \sum_{m=1}^{M} \pi_m \int_{\mathbf{x}:\ g(\mathbf{x})=m} f_m(\mathbf{x})\mu(d\mathbf{x})$$

$$= 1 - \int_{\mathcal{X}} \sum_{m=1}^{M} \pi_m f_m(\mathbf{x}) \mathbb{1}\{g(\mathbf{x}) = m\}\mu(d\mathbf{x})$$

$$= 1 - \int_{\mathcal{X}} \pi_{g(\mathbf{x})} f_{g(\mathbf{x})}(\mathbf{x})\mu(d\mathbf{x}).$$

Отримали таку просту формулу для ймовірності помилки класифікатора:

$$L(g) = 1 - \int_{\mathcal{X}} \pi_{g(\mathbf{x})} f_{g(\mathbf{x})}(\mathbf{x})\mu(d\mathbf{x}). \tag{5.2}$$

Чим менше ця ймовірність тим кращим є класифікатор з точки зору баєсового підходу.

Якщо розглядається набір класифікаторів $\mathcal{G}$, то класифікатор $g^B \in \mathcal{G}$ називають **баєсовим в** $\mathcal{G}$, якщо

$$L(g^B) \le \inf_{g \in \mathcal{G}} L(g).$$

Тобто баєсів класифікатор — це класифікатор з найменшою ймовірністю помилки.

Розглянемо набір $\mathcal{G}$ всіх вимірних класифікаторів $g : \mathcal{X} \to \{1, \dots, M\}$. Знайдемо баєсів класифікатор у цьому наборі[1].

[1]Такі класифікатори просто називають баєсовими, не вказуючи явно у якому наборі.

Покладемо

$$g^B(\mathbf{x}) = \text{argmax}_{m=1,\dots,M}\, \pi_m f_m(\mathbf{x}). \tag{5.3}$$

Оскільки для будь-якого $\mathbf{x} \in \mathcal{X}$ і будь-якого $g$

$$\pi_{g(\mathbf{x})} f_{g(\mathbf{x})}(\mathbf{x}) \le \max_{m=1,\dots,M} \pi_m f_m(\mathbf{x}),$$

то

$$L(g) \ge 1 - \int_{\mathcal{X}} \max_{m=1,\dots,M} \pi_m f_m(x) \mu(d\mathbf{x}) = L(g^B),$$

тобто класифікатор $g^B$, визначений (5.3), є баєсовим у наборі всіх вимірних класифікаторів.

Можливий дещо інший погляд на класифікатори, визначені (5.3), пов'язаний з поняттям апостеріорної ймовірності класу.

Пригадаємо, що $\pi_m = \mathsf{P}\{\kappa(O) = m\}$ це апріорна ймовірність того, що спостережуваний об'єкт належить $m$-тому класу. Ця ймовірність ніяк не враховує інформацію про конкретний об'єкт $O$, а визначається лише за загальною інформацією про всю групу об'єктів, які в принципі можуть спостерігатись у даній ситуації. Скажімо, у задачі медичної діагностики, яку ми розглядали вище, $\pi_m$ це, грубо кажучи, частота поява пацієнтів хворих $m$-тою хворобою на прийомі у лікаря. Якщо лікарю відомо, що 50% пацієнтів приходять до нього із застудою (хвороба 1), то $\pi_1 = 0.5$. Інформація, за якою визначена ця ймовірність, відома до того, як лікар обстежить даного пацієнта $O$, відповідно і сама ймовірність називається апріорною (тобто, отриманою до дослідження об'єкта).

Після того, як об'єкт досліджений, ми отримуємо інформацію про значення його спостережуваних характеристик $\xi(O)$. Ця інформація змінює наші уявлення про можливість для $O$ належати певному класу. Скажімо, лікар, вимірявши температуру пацієнта, отримує $39^o$ і міркує: така температура не характерна для застуди, у мене може 3% пацієнтів з такою температурою виявились хворими на застуду. Отже **апостеріорна ймовірність**[2] того, що пацієнт хворий на застуду при температурі $39^o$, $p_1(39^o) = 0.03$.

Формально апостеріорна ймовірність це умовна ймовірність даного класу при фіксованому значенні спостережуваних характеристик об'єкта:

$$p_m(\mathbf{x}) = \mathsf{P}\{\kappa(O) = m \mid \xi(O) = \mathbf{x}\}. \tag{5.4}$$

[2]Тобто ймовірність, визначена після дослідження об'єкта з урахуванням індивідуальної інформації про нього.

Апостеріорні ймовірності можна обчислювати за формулою Баєса для щільностей:

$$p_m(\mathbf{x}) = \frac{\pi_m f_m(\mathbf{x})}{\sum_{i=1}^{M} \pi_i f_i(\mathbf{x})}. \tag{5.5}$$

(Якщо міра $\mu$ — лічильна, то це звичайна формула Баєса. У загальному випадку цю формулу можна отримати, використовуючи означення умовного математичного сподівання відносно сігма-алгебри, породженої $\xi(O)$).

Порівнюючи (5.3) з (5.5), бачимо, що

$$g^B(\mathbf{x}) = \operatorname{argmax}_{m=1,\dots,M} p_m(\mathbf{x}), \tag{5.6}$$

тобто баєсів класифікатор вибирає клас з максимальною апостеріорною ймовірністю помилки.

Інколи формулу (5.6) приймають як означення баєсового класифікатора.

## 5.2 Емпірично-баєсова класифікація

У підрозділі 5.1 ми розглянули побудову оптимального — баєсового класифікатора, для якої потрібно знати ймовірнісні характеристики об'єктів, що класифікуються. На практиці такі теоретичні характеристики, як правило, невідомі, але дослідник має навчальну вибірку, за якою будує класифікатор.

При емпірично-басовому підході для побудови класифікатора з використанням навчальної вибірки спочатку за вибіркою оцінюють щільності розподілу для різних класів та апріорні ймовірності, а потім класифікатор отримують, підставляючи ці оцінки замість справжніх значень у формулу (5.3).

Отже, нехай навчальна вибірка $\mathbf{X} = \{(\xi_j, \kappa_j), j = 1\dots, n\}$ містить дані про об'єкти $O_1, \dots, O_n$, а саме

$\xi_j = \xi(O_j)$ — значення спостережуваних характеристик $O_j$ і

$\kappa_j = \kappa(O_j)$ — номер класу (популяції) якій $O_j$ належить.

Якщо вважати, що новий об'єкт, який ми збираємось класифікувати, отримано у тих самих умовах, в яких отримувались об'єкти навчальної вибірки, то можна запропонувати такі оцінки для ймовірнісних характеристик задачі класифікації.

Оцінкою для апріорної ймовірності класу $\pi_m = \mathsf{P}\{\kappa(O) = m\}$ може бути відносна частота появ цього класу у навчальній вибірці:

$$\hat{\pi}_m = \frac{n_m}{n},\ n_m = \sum_{j=1}^{n} \mathbb{1}\{\kappa_j = m\}.$$

Оцінка для щільності розподілу характеристик об'єкта, що належить $m$-тому класу (в припущенні, що ці розподіли є абсолютно неперервними):

$$\hat{f}_m(\mathbf{x}) = \frac{1}{n_m h_m^1 h_m^2 \dots h_m^d} \sum_{j:\ \kappa_j = m} K\left(\frac{x^1 - \xi_j^1}{h_m^1}, \frac{x^2 - \xi_j^2}{h_m^1}, \dots \frac{x^d - \xi_j^d}{h_m^d}\right)$$

— це ядерна оцінка щільності, підрахована за тими спостереженнями у навчальній вибірці, які належать $m$-му класу. Тут $K$ — ядро, а $h_m^1, \dots,$ $h_m^d$ — параметри згладжування оцінки.

Підставляючи ці оцінки у формулу (5.3) для баєсового класифікатора замість справжніх значень, отримуємо

$$\hat{g}(\mathbf{x}) = \operatorname{argmax}_{m=1,\dots,M} \hat{\pi}_m \hat{f}_m(\mathbf{x}). \tag{5.7}$$

Це емпірично-баєсів класифікатор, побудований на основі ядерних оцінок щільності.

Взагалі кажучи, у формулі (5.7) можна використовувати не тільки ці, а й будь-які інші оцінки для апріорних ймовірностей та щільностей різних класів.

Наскільки хорошим є емпірично-баєсів класифікатор? Зрозуміло, що він не може бути кращим ніж баєсів, тобто мати меншу ймовірність помилки. Чим ближча ймовірність помилки емпірично-баєсового класифкатора до $L(g^B)$, тим він кращий. Перш ніж розглядати оцінки для цієї ймовірності, треба уточнити, що саме ми маємо на увазі.

Можна визначити характеристику якості емпірично-баєсового класифікатора $\hat{g}$ скориставшись формулою (5.2):

$$L(\hat{g}) = 1 - \int_{\mathcal{X}} \pi_{\hat{g}(\mathbf{x})} f_{\hat{g}(\mathbf{x})}(\mathbf{x}) \mu(d\mathbf{x}). \tag{5.8}$$

Помітимо, що ця величина є випадковою, оскільки емпірично-баєсів класифікатор $\hat{g}$ побудований за оцінками $\hat{f}_m(\mathbf{x})$, які, в свою чергу, залежать від навчальної вибірки $\mathbf{X}$, яку ми розглядаємо як випадкову.

То чи можна трактувати $L(\hat{g})$ як ймовірність помилки класифікатора? Так, якщо ми класифікуємо новий об'єкт, незалежний від навчальної вибірки, то $L(\hat{g})$ це умовна ймовірність помилки класифікатора $\hat{g}$ при фіксованій $\mathbf{X}$:

$$L(\hat{g}) = \mathsf{E}\left[\mathbb{1}\{\hat{g}(\xi(O)) \neq \kappa(O)\} \mid \mathbf{X}\right] = \mathsf{P}[\hat{g}(\xi(O)) \neq \kappa(O) \mid \mathbf{X}].$$

Це випадкова величина вимірна відносно $\sigma$-алгебри породженої $\mathbf{X}$, тобто вона залежить від навчальної вибірки, як і слід було сподіватись. Якщо для опису класифікатора потрібна невипадкова величина, то можна "усереднити" $L(\hat{g})$ по всіх можливих навчальних вибірках, тобто розглянути характеристику

$$\bar{L}(\hat{g}) = \mathsf{E}\, L(\hat{g}) = \mathsf{P}[\hat{g}(\xi(O)) \neq \kappa(O)].$$

Це вже безумовна ймовірність, котру можна підраховувати або оцінювати, не знаючи, якою виявиться конкретна навчальна вибірка, по котрій буде побудовано класифікатор, а виходячи лише з ймовірнісних характеристик задачі класифікації.

Отже розглянемо оцінки для відхилення помилки емпірично-баєсового класифікатора від найкращої можливої.

**Теорема 5.2.1.** *Якщо класифікатор $\hat{g}$ задано (5.7), то*

$$0 < L(\hat{g}) - L(g^B) \leq 2 \sum_{i=1}^{M} \int_{\mathcal{X}} |\hat{\pi}_i \hat{f}_i(\mathbf{x}) - \pi_i f_i(\mathbf{x})| \mu(d\mathbf{x}), \tag{5.9}$$

*де $g^B$ — баєсів класифікатор.*

(Нерівність (5.9) називають нерівністю Гйорфі).

**Доведення.** Нерівність в ліву сторону $0 < \ldots$ — очевидна, оскільки у $g^B$ ймовірність помилки найменша можлива. Доведемо нерівність праворуч.

Позначимо $z_i(\mathbf{x}) = \pi_i f_i(\mathbf{x})$, $\hat{z}_i(\mathbf{x}) = \hat{\pi}_i \hat{f}_i(\mathbf{x})$. Враховуючи (5.2) і (5.8), отримуємо

$$L(\hat{g}) - L(g^B) \leq \left| \int_{\mathcal{X}} \max_i z_i(\mathbf{x}) - z_{\hat{g}(\mathbf{x})}(\mathbf{x}) \mu(d\mathbf{x}) \right| \leq J_1 + J_2,$$

де

$$J_1 = \int_{\mathcal{X}} |\max_i z_i(\mathbf{x}) - \max_i \hat{z}_i(\mathbf{x})| \mu(d\mathbf{x}),$$

$$J_2 = \int_{\mathcal{X}} |\max_i \hat{z}_i(\mathbf{x}) - z_{\hat{g}(\mathbf{x})}(\mathbf{x})| \mu(d\mathbf{x}).$$

Помітимо, що

$$|\max_i z_i(\mathbf{x}) - \max_i \hat{z}_i(\mathbf{x})| \leq \max_i |z_i(\mathbf{x}) - \hat{z}_i(\mathbf{x})| \leq \sum_{i=1}^{M} |z_i(\mathbf{x}) - \hat{z}_i(\mathbf{x})|.$$

Отже

$$J_1 \leq \sum_{i=1}^{M} \int_{\mathcal{X}} |\hat{\pi}_i \hat{f}_i(\mathbf{x}) - \pi_i f_i(\mathbf{x})| \mu(d\mathbf{x}). \tag{5.10}$$

Крім того,

$$|\max_i \hat{z}_i(\mathbf{x}) - z_{\hat{g}(\mathbf{x})}(\mathbf{x})| = |\hat{z}_{\hat{g}(\mathbf{x})}(\mathbf{x}) - z_{\hat{g}(\mathbf{x})}(\mathbf{x})| \leq \sum_{j=1}^{M} |z_i(\mathbf{x}) - \hat{z}_i(\mathbf{x})|,$$

Тому

$$J_2 \leq \sum_{i=1}^{M} \int_{\mathcal{X}} |\hat{\pi}_i \hat{f}_i(\mathbf{x}) - \pi_i f_i(\mathbf{x})| \mu(d\mathbf{x}). \tag{5.11}$$

Об'єднуючи (5.10) і (5.10) отримуємо твердження теореми.

Взявши математичне сподівання від (5.9) отримуємо нерівність для безумовної ймовірності похибки:

$$0 < \bar{L}(\hat{g}) - L(g^B) \leq 2 \sum_{i=1}^{M} \int_{\mathcal{X}} \mathsf{E} |\hat{\pi}_i \hat{f}_i(\mathbf{x}) - \pi_i f_i(\mathbf{x})| \mu(d\mathbf{x}), \tag{5.12}$$

Таким чином, точність емпірично-баєсового класифікатора оцінюється у термінах відстані в $L_1$ між оцінкою щільності за навчальною вибіркою та справжньою щільністю (помноженими на апріорні ймовірності та їх оцінки). У підрозділі 4.4 ми досліджували збіжність $\mathsf{MISE}(\hat{f}_n)$ — тобто квадрату відстані в $L_2$. Якщо розподіли зосереджені на скінченних інтервалах, відстань в $L_1$ легко оцінити через відстань в $L_2$.

**Приклад.** Нехай спостерігається одна числова характеристика об'єкта, що приймає значення зі скінченного інтервалу, тобто $\mathcal{X} = [a, b]$, де $a, b$ скіченні числа. Припустимо, що всі щільності $f_i$ мають неперервну другу похідну, а їх оцінки $\hat{f}_{i,n}$ — це ядерні оцінки з параметром згладжування,

що має оптимальний порядок збіжності, а ядро є ядром другого порядку. Тоді, як ми розібрали у п. 4.4[3],

$$\mathsf{E}\int_a^b (\hat{f}_{i,n}(x) - f_i(x))^2 dx = \mathsf{MISE}(\hat{f}_{i,n}(x)) \le \frac{C}{n^{4/5}}.$$

Оцінимо:

$$\bar{L}(\hat{g}) - L(g^B) \le 2\sum_{i=1}^{M} \mathsf{E}\int_a^b |\hat{\pi}_i \hat{f}_{i,n}(x) - \pi_i f_i(x)| dx$$

$$\le 2\sum_{i=1}^{M} \mathsf{E}\int_a^b |\hat{\pi}_i - \pi_i| \hat{f}_i(x) dx + 2\sum_{i=1}^{M} \pi_i \,\mathsf{E}\int_a^b |\hat{f}_{i,n}(x) - f_i(x)| dx$$

$$\le 2\sum_{i=1}^{M} \sqrt{\mathsf{E}(\hat{\pi}_i - \pi_i)^2} + 2\sqrt{b-a}\sum_{i=1}^{M} \sqrt{\int_a^b \mathsf{E}(\hat{f}_{i,n}(x) - f_i(x))^2 dx}$$

$$\le \frac{C_1}{n^{1/2}} + \frac{C_2}{n^{2/5}}.$$

Отже отримуємо, що

$$\bar{L}(\hat{g}) - L(g^B) \le \frac{C}{n^{2/5}}.$$

Це досить груба оцінка, але вона показує, що ймовірність помилки емпірично-баєсового класифікатора у цьому випадку прямує до найменшої можливої коли обсяг вибірки прямує до 0.

## 5.3 Дискримінантний аналіз

Як приклад застосування емпірично-баєсового підходу, розглянемо випадок, коли щільності розподілів спостережуваних характеристик для об'єктів різних класів вважаються відомими з точністю до невідомих параметрів. Тобто зараз ми будемо використовувати параметричний підхід

[3]Точніше, наша ядерна оцінка за навчальною вибіркою для $i$-того класу побудована за $n_i$ спостереженнями. Величина $n_i$ — кількість об'єктів з $i$-того класу у навчальній вибірці є випадковою, отже сказане у п. 4.4 не можна безпосередньо переносити на даний випадок. Але $n_i/n \to \pi_i$, тому асимптотичний порядок збіжності MISE залишається тим самим.

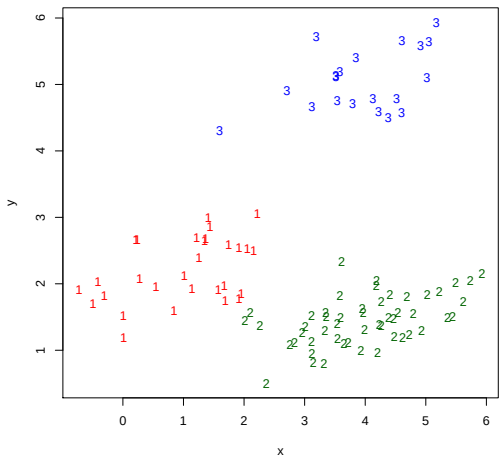



Рис. 5.3: Дані для дискримінантного аналізу

до задачі класифікації. Техніка, яка описана у цьому підрозділі, називається **лінійним дискримінантним аналізом, LDA**, і, не зважаючи на теоретичні обмеження, дуже широко застосовується на практиці. З лінійного дискримінантного аналізу традиційно починають спроби побудувати працюючий класифікатор. Якщо це вдається, то далі можна не йти.

Отже, нехай спостерігаються $d$-вимірні характеристики об'єктів $\xi = \xi(O) \in \mathbb{R}^d$, а їх розподіли для кожного класу є гауссовими:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \mathbf{S}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^T \mathbf{S}^{-1}(\mathbf{x}-\mu_i)\right), \ i = 1, \ldots, M,$$

де $\mu_i = (\mu_i^1, \ldots, \mu_i^d)^T$ — математичне сподівання характеристик об'єктів, що належать $i$-тому класу, $\mathbf{S}$ — коваріаційна матриця, яка вважається однаковою для всіх класів. Це "класична модель LDA".

Типовий приклад того, який вигляд можуть мати дані, що описуються такою моделлю, показано на рис. 5.3. Тут кожен об'єкт має двовимірний вектор спостережуваних характеристик $\xi = (x, y)$, номер класу, якому належить об'єкт, розміщується у точці з відповідними координатами. Хмари даних, що відповідають кожному класу, мають більш-менш еліпсоподібну форму і розрізняються тільки положенням центру, а не розкидом точок у різних напрямах.

Як і раніше, позначимо апріорні ймовірності класів $\pi_i$, $i = 1, \ldots, M$.

Визначимо, який вигляд матиме баєсів класифікатор $g^B$ у цьому випадку. Згідно з формулою (5.3), $g^B(\mathbf{x}) = k$, якщо

$$\pi_k f_k(\mathbf{x}) > \pi_i f_i(\mathbf{x}) \text{ для всіх } i \neq k.$$

Підставляючи у цю умову гауссові щільності классів, отримуємо:

$$\frac{\pi_k}{(2\pi)^{d/2}\sqrt{\det \mathbf{S}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_k)^T\mathbf{S}^{-1}(\mathbf{x}-\mu_k)\right)$$

$$> \frac{\pi_i}{(2\pi)^{d/2}\sqrt{\det \mathbf{S}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^T\mathbf{S}^{-1}(\mathbf{x}-\mu_i)\right).$$

Прологарифмувавши і звівши подібні доданки, маємо:

$$\log \pi_k + \mu_k^T\mathbf{S}^{-1}\mathbf{x} - \frac{1}{2}\mu_k^T\mathbf{S}^{-1}\mu_k > \log \pi_i + \mu_i^T\mathbf{S}^{-1}\mathbf{x} - \frac{1}{2}\mu_i^T\mathbf{S}^{-1}\mu_i.$$

Позначимо

$$D_i(\mathbf{x}) = b_i^0 + \mathbf{b}_i^T\mathbf{x},\ i = 1, \ldots, M,$$

де

$$b_i^0 = \log \pi_i - \frac{1}{2}\mu_i^T\mathbf{S}^{-1}\mu_i,\ \mathbf{b}_i = \mathbf{S}^{-1}\mu_i. \tag{5.13}$$

Функцію $D_i(\mathbf{x})$ називають **дискримінантною функцією** $i$-того класу.

Тепер баєсів класифікатор у даній задачі можна описати так:

$$g(\mathbf{x}) = k \text{ якщо } D_k(\mathbf{x}) > D_i(\mathbf{x}) \text{ для всіх } i \neq k. \tag{5.14}$$

Тобто класифікатор підраховує значення дискримінантних функцій на спостережуваних характеристиках об'єкта і відносить об'єкт до того класу, для якого дискримінантна функція найбільша. Якщо максимум досягається для двох класів одразу, об'єкт можна віднести до будь-якого з них, на ймовірність помилки класифікатора це не вплине.

Якщо математичні сподівання класів $\mu_i$ і спільна коваріаційна матриця $\mathbf{S}$ невідомі, то можна оцінити їх за навчальною вибіркою $(\xi_j, \kappa_j)$, $j = 1, \ldots, n$. Природною оцінкою $\mu_i$ є середнє значення спостережуваних характеристик по об'єктах, що належать $i$-тому класу:

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{j:\ \kappa_j = i} \xi_j,$$

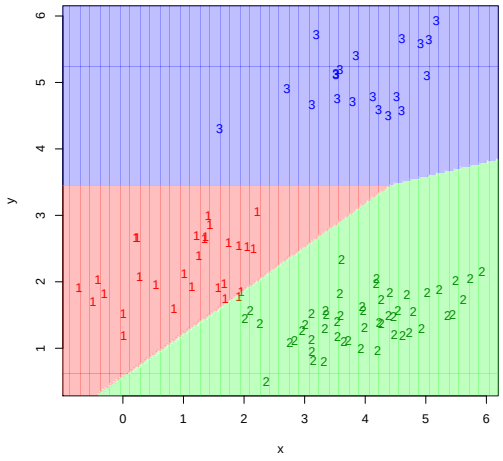



Рис. 5.4: Результат лінійного дискримінантного аналізу

(як і раніше, $n_i$ це кількість об'єктів з $i$-того класу у навчальній вибірці).

З коваріаційною матрицею трохи складніше, оскільки вона спільна для всіх класів. Тому її треба оцінювати по всій вибірці, а не окремо по кожному класу. Але при цьому віднімати від кожного спостереження треба середнє по тому класу, якому належить об'єкт:

$$\hat{\mathbf{S}} = \frac{1}{n-M}\sum_{j=1}^{n}(\xi_j - \hat{\mu}_{\kappa_j})(\xi_j - \hat{\mu}_{\kappa_j})^T.$$

(Завдяки множнику $n - M$ ця оцінка є незміщеною).

Якщо підставити оцінки $\hat{\mu}_i$ і $\hat{\mathbf{S}}$ у формулу (5.13) отримаємо емпіричні дискримінантні функції і, відповідно, емпірично-баєсів класифікатор на їх основі. Саме цим класифікатором користуються у практичних застосуваннях дискримінантного аналізу.

На практиці LDA інколи дає хороші результати і тоді, коли розподіл даних не описується класичною моделлю LDA, наприклад, коли він не є гауссовим. Але для цього потрібно, щоб дані в принципі можна було класифікувати класифікатором досить простої структури.

Наприклад, результат класифікації даних з рис. 5.3 за технікою лінійного дискримінантного аналізу відображено на рис. 5.4. Області значень спостережуваних характеристик, у яких класифікатор LDA відносить

об'єкт до певного класу, зафарбовані відповідним кольором. Помітно, що ці області обмежені прямими лініями[4] У загальному багатовимірному випадку ці області обмежені гіперплощинами виду $\mathbf{x}: \ D_i(\mathbf{x}) = D_k(\mathbf{x})$ і є опуклими багатогранними областями у $\mathbb{R}^d$. Зрозуміло, що коли "правильні" області на які розбиває простір спостережень справжній баєсів класифікатор не є опуклими, застосування LDA не може забезпечити хорошу якість класифікації.

## 5.4 Непараметрична класифікація багатовимірних даних

У попередніх параграфах ми розглянули приклади двох підходів до побудови емпірично-баєсових класифікаторів — непараметричний (у п.5.2) і параметричний (лінійний дискримінантний аналіз — у п. 5.3). Який з них кращий? Або можна поставити більш конкретне запитання — за яких обставин варто застосовувати той чи інший підхід?

Параметричні оцінки є зазвичай більш точними ніж непараметричні, якщо виконується та параметрична модель розподілу даних, на яких вони ґрунтуються. Це і зрозуміло — для побудови параметричних оцінок використовується значно більше апріорної інформації про розподіл даних. Якщо ця інформація є достовірною, вони будуть переважати непараметричні. Якщо ні — параметрична оцінка може бути зовсім хибною і її використання приведе до невдалого класифікатора. Непараметричні методи більш гнучкі, але для досягнення необхідної точності їм потрібний значно більший обсяг навчальної вибірки, ніж параметричним.

Наприклад, розглянемо техніку лінійного дискримінантного аналізу (LDA), описану у п. 5.3. При побудові класифікатора LDA використовувалось припущення, що спостережувані характеристики кожного класу мають гауссів розподіл, причому коваріаційні матриці однакові для всіх класів. Завдяки цим припущенням емпірично-баєсів класифікатор набуває особливо простої форми (5.14). Для його реалізації непотрібно оцінювати щільності різних класів, досить оцінити математичні сподівання для кожного класу і спільну коваріаційну матрицю. Це можна зробити

[4]Не так, щоб дуже помітно, але це недолік алгоритму розфарбовування, який був використаний у R для створення рисунка. Насправді лінії, що розділяють області різних классів — прямі.

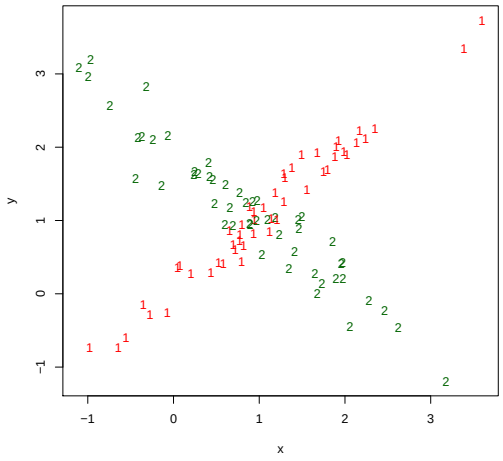



Рис. 5.5: Навчальна вибірка з двох класів

досить точно навіть для порівняно невеликої кількості спостережень, відповідно і побудований класифікатор LDA буде близьким до найкращого можливого (справжнього баєсового).

Але для того, щоб так вийшло, потрібно, щоб спостережувані дані мали розподіл коли не в точності гауссів, то хоч подібний до гауссового. Зокрема, LDA не буде працювати адекватно, якщо хмари даних для різних класів мають форму сильно відмінну від еліптичної. Більше того, класифікатор LDA буде невдалим і для випадку, коли для різних класів хмари даних еліптичні, але витягнуті у різних напрямках.

**Приклад 5.4.1.** Як приклад, розглянемо навчальну вибірку, відображену на діаграмі розсіювання рис. 5.5. Тут об'єкти з двох класів мають по дві спостережувані характеристики — `x` і `y`. Хмари даних частково перекриваються, але для більшості об'єктів їх розташування дозволяє досить очевидно визначити клас, якому вони належать. Природно сподіватись, що це зможе зробити і правильно побудований класифікатор.

Але LDA не може розв'язати цю задачу: на цих даних він будує класифікатор, котрий всі об'єкти відносить до другого класу[5]. Чому? По-

[5]Чому саме до другого? Оцінки щільностей для обох класів виходять практично однаковими, тому все залежить від кількостей спостережень кожного класу. На картинці двійок трохи більше ніж одиниць і це вирішує справу: якби ми завжди класи-

перше, середні значення характеристик для обох класів є однаковими, а саме на їх відмінності спирається LDA при побудові класифікатора. По-друге, спроба побудувати єдину оцінку для коваріаційної матриці для обох класів не може закінчитись вдало — їх матриці, вочевидь, відмінні, оскільки залежності між `x` і `y` добре виражені і цілком різні у першому і другому класах. ◀

Зрозуміло, що для побудови хорошого класифікатора за такими даними, потрібна техніка, що враховує відмінності залежностей між змінними і налаштовує алгоритм класифікації саме не ці відмінності.

### 5.4.1 Квадратичний дискримінантний аналіз

Перше, що спадає на думку, це використання більш загальної моделі дискримінантного аналізу, в якій коваріаційні матриці спостережуваних векторів вважаються різними для різних класів:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \mathbf{S}_i}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^T \mathbf{S}_i^{-1}(\mathbf{x}-\mu_i)\right),\ i = 1,\ldots,M,$$

де $\mu_i$ і $\mathbf{S}_i$ — вектор математичних сподівань і коваріаційна матриця для спостережень об'єкта, що належить $i$-тому класу. Проробивши для цієї моделі ті ж перетворення, що і у п. 5.3, легко бачити, що дискримінантні функції тепер будуть не лінійними, як у LDA, а квадратичними. Відповідно, у просторі спостережуваних ознак баєсів (і емпірично-баєсів) класифікатори будуть розділяти об'єкти різних класів не гіперплощинами, а поверхнями другого порядку. Тому така техніка класифікації має назву квадратичний дискримінантний аналіз. На даних з рис. 5.5 ця техніка приводить до класифікатора, робота якого відображена на рис. 5.6. Тут класифікатор розбиває площину на дві області, розділені кривою другого порядку (гіперболою). Об'єкти класифікуються в залежності від того, в яку область потраплять їх спостережувані характеристики. Видно, що навчальну вибірку класифікатор розділяє, в-основному, правильно. Помилки виникають за рахунок неправильної класифікації обєктів, що потрапили у центр рисунка, там, де хмари даних різних класів перетинаються. Зрозуміло, що такі помилки усунути не можна. Їх частота складає 0.1 — це оцінка для ймовірності помилки класифікатора.

---

фікували до першого класу, помилок було б більше.

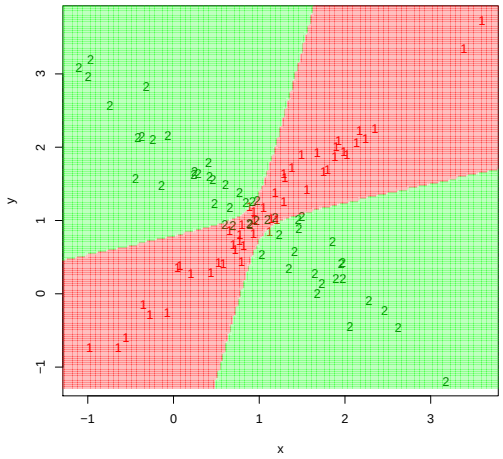



Рис. 5.6: Класифікатор квадратичного дискримінантного аналізу

Можна подивитись, який вигляд мають оцінки функцій $\pi_i f_i(\mathbf{x})$, за якими класифікатор побудований — див. рис. 5.7. Тут зображено лінії рівня для оцінок цих функцій за даними кожного класу. Помітно, що форма цих ліній в-основному відтворює форму відповідних хмар даних, отже не варто сподіватись помітно кращих результатів, якщо ми відмовимось від припущення про гауссовість.

У загальному випадку це не так: QDA не забезпечує оптимальність класифікації, якщо розподіли спостережуваних характеристик помітно відрізняються від гауссових. Подивимось, як на цих даних працюватиме емпірично-баєсів класифікатор з непараметричними ядерними оцінками щільності, описаний у п. 5.2. На рисунку 5.8 зображено лінії рівня для $\hat{\pi}_i \hat{f}_i(\mathbf{x})$, де $\hat{f}_i(\mathbf{x})$ — двовимірна ядерна оцінка з параметрами згладжування, обраними за правилом Скота (4.25). Порівняно з параметричними, вони вийшли сильно перезгладженими, оскільки параметр згладжування підбирався окремо по `x` і `y`, а досліджувана щільність "витягнута" вздовж лінії `y=x` і швидко змінюється у напрямку, перпендикулярному до цієї лінії. Тим не менше, класифікатор, отриманий в результаті підстановки цих оцінок у формулу для баєсового класифікатора, вийшов досить вдалим — див. рис. 5.9. Частота його помилок на навчальній вибірці — 0.17. Це помітно гірше, ніж у класифікатора QDA (який є близьким до оптимального), але значно краще ніж у LDA, який зовсім не дав собі ради з

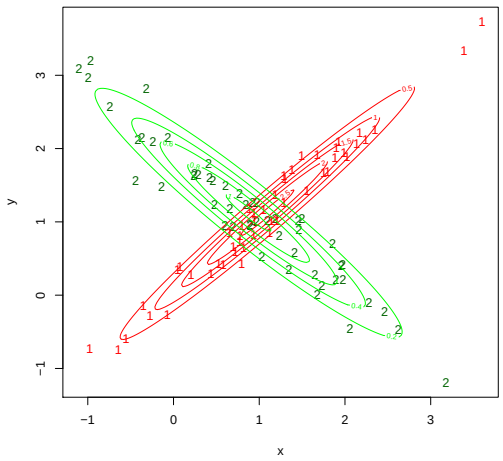



Рис. 5.7: Лінії рівня для оцінок $\pi_i f_i(\mathbf{x})$ отриманих за припущенням про гауссовість

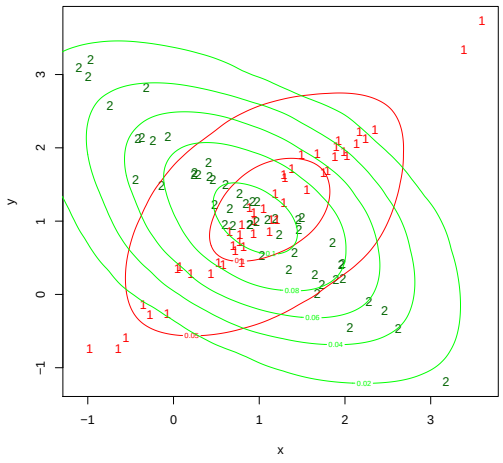



Рис. 5.8: Лінії рівня для двовимірних ядерних оцінок $\pi_i f_i(\mathbf{x})$

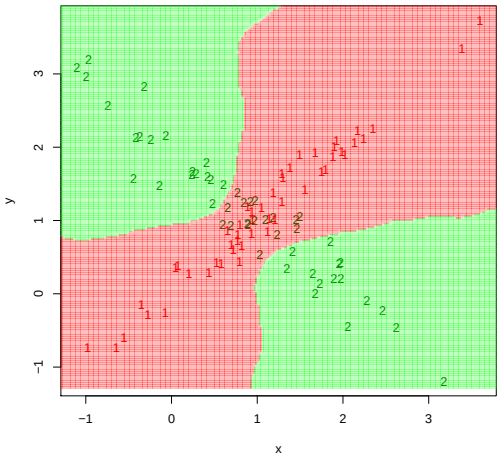



Рис. 5.9: Класифікатор на основі двовимірних ядерних оцінок

цим завданням.

При зростанні кількості спостережуваних характеристик $d$ (вимірності $\xi(O)$) якість класифікаторів, побудованих безпосередньо на $d$-вимірних оцінках щільності, різко погіршується. Це один з виявів "прокляття багатовимірності", яке обговорювалось у п. 4.7. Для хорошої підгонки такого класифікатора за 10-вимірними спостереженнями потрібні навчальні вибірки з велетенською кількістю спостережень. Тому намагаються звузити набір класифікаторів, серед яких шукають оптимальний так, щоб техніку непараметричного оцінювання можна було застосовувати лише для даних низької вимірності: одно- або дво-вимірних.

### 5.4.2 Наївний баєсів класифікатор

Найпростіший спосіб отримати для багатовимірних даних емпіричний баєсів класифікатор з оцінкою лише одновимірних щільностей — припустити, що координати вектора спостережень незалежні між собою для всіх класів. Тобто для $\mathbf{x} = (x^1, \ldots, x^d)^T$,

$$f_i(\mathbf{x}) = f_i^1(x^1) f_i^2(x^2) \ldots f_i^d(x^d),$$

де $f_i^k(x)$ — маргінальна щільність $k$-того елемента вектора спостережень для об'єктів з $i$-того класу.

Для оцінки $f_i^k(x)$ можна використати звичайну одновимірну ядерну оцінку:

$$\hat{f}_i^k(x) = \frac{1}{n_k h} \sum_{j:\ \kappa_j = i} K\left(\frac{x - \xi_j^k}{h}\right),$$

де, як і раніше, $K$ — ядро оцінки, $h$ — параметр згладжування, які можна вибирати методами, описаними у п. 4.4—4.6.

Підставляючи ці оцінки у формулу для емпірично-баєсового класифікатора, отримуємо так званий "наївний баєсів класифікатор", який відносить об'єкт з вектором характеристик $\xi = (\xi^1, \ldots, \xi^d)^T$ до $k$-того класу, якщо

$$\hat{\pi}_k \prod_{l=1}^{d} \hat{f}_k^l(\xi^l) > \hat{\pi}_i \prod_{l=1}^{d} \hat{f}_i^l(\xi^l) \text{ для всіх } i \neq k.$$

Використання такого класифікатора доцільне якщо окремі характеристики об'єктів слабко залежні між собою, або коли ця залежність є приблизно однаковою для об'єктів з усіх класів. Тобто такий класифікатор не може працювати добре у задачах, де відмінності розподілів характеристик об'єктів різних класів пов'язані з відмінностями у залежностях між цими характеристиками.

Зокрема, не слід сподіватись, що наївний баєсів класифікатор добре класифікуватиме об'єкти з рис. 5.5.

### 5.4.3 Класифікація з проекцією на оптимальний напрямок

Класифікатор LDA, побудований на використанні лінійних комбінацій початкових змінних (спостережуваних характеристик) наводить на думку використати такі лінійні комбінації для непараметричної класифікації. Ми розглянемо тут лише найпростішу реалізацію цієї ідеї.

Нехай класифікацію потрібно побудувати за $d$-вимірними спостережуваними характеристиками $\xi(O) = (\xi^1, \ldots, \xi^d)^T$. Виберемо у просторі $\mathbb{R}^d$ вектор одиничної довжини $\mathbf{u} = (u^1, \ldots, u^d)^T \in \mathbb{R}^d$, $\|\mathbf{u}\| = 1$. Розглянемо ортогональну проекцію вектора $\xi(O)$ на напрямок $\mathbf{u}$:

$$\xi^{\mathbf{u}}(O) = \mathbf{u}^T \xi(O) = \sum_{i=1}^{d} u^i \xi^i.$$

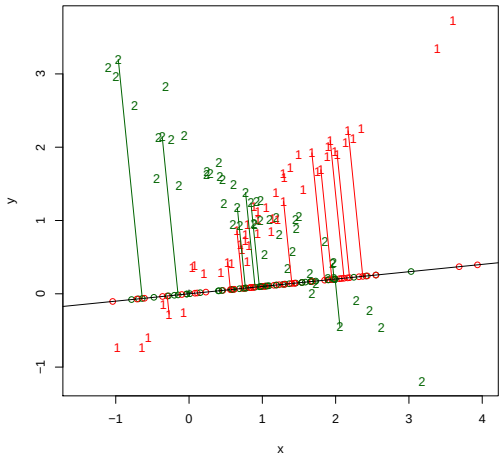



Рис. 5.10: Проекція на напрямок, невдалий для класифікації

Будемо класифікувати об'єкти $O$ за отриманими проекціями, використовуючи техніку непараметричної емпірично-баєсової класифікації, описану у п. 5.2. Для цього спроектуємо на $\mathbf{u}$ навчальну вибірку і підрахуємо ядерні оцінки щільності для проекцій:

$$\hat{f}_k(t;\mathbf{u}) = \frac{1}{n_k h} \sum_{j:\kappa_j=k} K\left(\frac{t-\xi_j^{\mathbf{u}}}{h}\right),$$

де $\xi_j^{\mathbf{u}} = \mathbf{u}^T\xi_j$. Тепер визначимо проекційний емпірично-баєсів класифікатор, що відповідає напрямку $\mathbf{u}$:

$$\hat{g}(\mathbf{x};\mathbf{u}) = \operatorname{argmax}_{k=1,\dots,M} \hat{\pi}_k \hat{f}_k(\mathbf{x};\mathbf{u}).$$

Наприклад, для даних, зображених на рис. 5.5, така проекція може мати вигляд, як на рис. 5.10. Тут чорна пряма відповідає обраному напрямку $\mathbf{u} = (\cos(0.1), \sin(0.1))^T$, а точки на ній — проекціям точок спостережень з навчальної вибірки. Легко зрозуміти, що цей напрямок проекції є невдалим для того, щоб робити класифікацію: точки, які відповідають об'єктам з різних класів, при проектуванні перемішались, розділити їх розумним способом по класах — неможливо. Це підтверджують і графіки оцінок щільності[6] за проекціями, зображені на рис. 5.11.

[6]Ядерні оцінки щільності з ядром Єпанєчнікова і параметрами згладжування, ви-

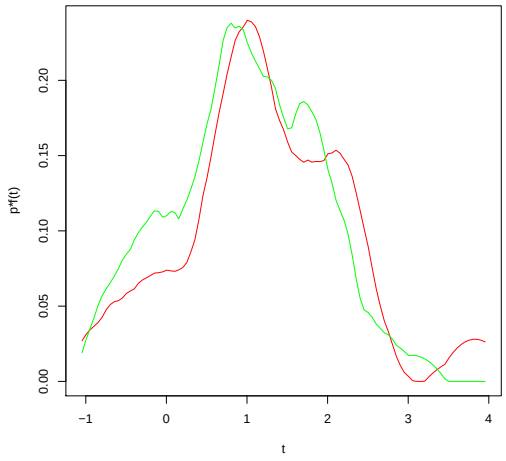



Рис. 5.11: $\hat{p}_i \hat{f}_i(x)$ для невдалої проекції

Таким чином, треба вміти правильно визначати оптимальний напрямок проекції. Один з найпростіших способів зробити це — використати як критерій якості напрямку частоту помилок відповідного класифікатора на навчальній вибірці:

$$\hat{L}(\mathbf{u}) = \#\{j : \ \hat{g}(\xi_j, \mathbf{u}) \neq \kappa_j\}/n.$$

На роль оптимального обирають той напрямок, при якому ця частота мінімальна —

$$\hat{\mathbf{u}} = \text{argmin}_{\mathbf{u}\in\mathbb{R}^d, \|\mathbf{u}\|=1} \hat{L}(\mathbf{u}).$$

Остаточно проекційний класифікатор визначається як:

$$g^P(\mathbf{x}) = \hat{g}(\mathbf{x}, \hat{\mathbf{u}}).$$

У випадку, коли $d = 2$, будь-який напрямок можна задати ортом як $\mathbf{u} = (\cos(\beta), \sin(\beta))^T$, де $\beta \in [0, \pi)$ — кут, який утворює $\mathbf{u}$ з додатнім напрямком горизонтальної осі[7]. Тому фактично, у цьому випадку досить

браними за правилом Сілвермана.

[7]Формально, треба розглянути іще $\beta \in [\pi, 2\pi)$, але проекції на ці орти будуть відрізнятись від розглянутих вище лише знаком — це не дасть нам кращих класифікаторів.

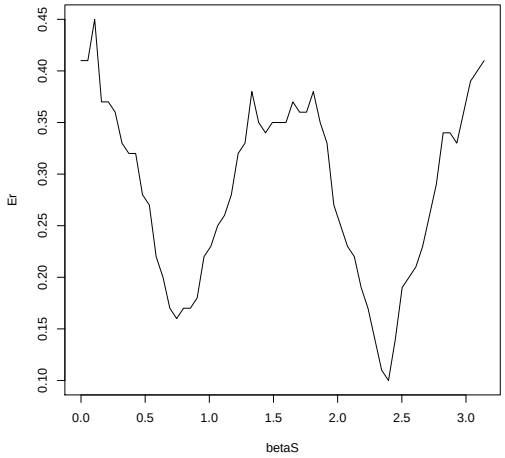



Рис. 5.12: Частота помилок класифікації як функція від $\beta$ — кута проекції

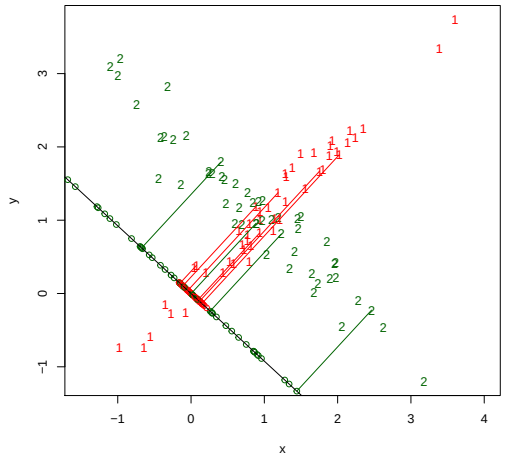



Рис. 5.13: Проекція на найкращий напрямок для класифікації

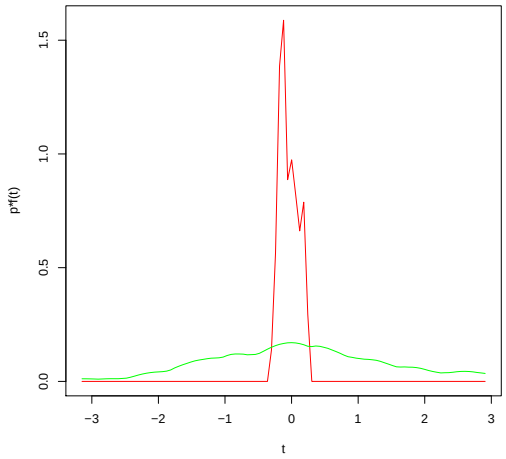



Рис. 5.14: $\hat{p}_i \hat{f}_i(x)$ для оптимальної проекції

вибрати оптимальне $\beta$. На рис. 5.12 для нашого прикладу зображено $\hat{L}(\mathbf{u})$ як функцію від $\beta$. Мінімум досягається при $\beta = 2.39613$. Це напрямок, приблизно паралельний напрямку найбільшого розкиду хмари даних другого класу, і найменшого — для першого класу (див. рис. 5.13). В результаті проекції спостережень другого класу розтягуються вздовж всієї прямої, на яку робиться проекція, а всі спостереження першого класу збираються у компактну групу точок посередині. Класифікатор може відносити до першого класу ті спостереження, що потрапляють у область, зайняту цією компактною групою, а всі інші — до другого класу. Саме так і працює емпірично-баєсів класифікатор у цьому випадку — див. рис. 5.14

При цьому частина точок з другого класу класифікується до першого. Акуратний підрахунок показує, що у цьому прикладі $\hat{L}(\hat{\mathbf{u}}) = 0.1$ — оптимальний проекційний класифікатор має таку ж частоту помилок на навчальній вибірці, як і QDA, який використовував апріорне припущення про гауссовість спостережень. І ця частота помилок значно менша, ніж у багатовимірного емпірчно-баєсового класифікатора.

При використанні цієї техніки для по-справжньому багатовимірних даних можуть виникати ситуації перепідгонки, аналогічні перепідгонці регресійних моделей у випадку нестрогої мультиколінеарності. Це яви-

ще виникає тому, що у критерію якості $\hat{L}(\mathbf{u})$ для перевірки роботи класифікатора використовується та ж вибірка, за якою класифікатор навчається. Щоб запобігти ефекту перепідгонки, можна використовувати техніку крос-валідації.

## 5.5 Запитання і задачі

**Запитання.**

1. Якими ймовірнісними характеристиками описується задача статистичної класифікації?

2. Що таке класифікатор? У який спосіб характеризують якість класифікатора при баєсовому підході до задачі класифікації?

3. Що таке баєсів класифікатор? Дайте означення і опишіть правило класифікації.

4. Дайте означення апостеріорної ймовірності належності об'єкта даному класу. Чим апостеріорні ймовірності відрізняються від апріорних? Схарактеризуйте баєсів класифікатор у термінах апостеріорних ймовірностей.

5. Чим емпірично-баєсів класифікатор відрізняється від справжнього баєсового?

6. Що таке навчальна вибірка? Як навчальна вибірка використовується при побудові емпірично-баєсового класифікатора?

7. Яка характеристика якості класифікатора оцінюється за допомогою нерівності Гйорфі?

8. Якими повинні бути розподіли спостережуваних характеристик класифікованих об'єктів для того, щоб можна було застосовувати до класифікації лінійний дискримінантний аналіз?

9. Чим модель квадратичного дискримінантного аналізу відрізняється від моделі лінійного дискримінантного аналізу?

10. Що таке "прокляття багатовимірності" загрожує при побудові емпірично-баєсових класифікаторів?

11. В чому полягає ідея наївного баєсового класифікатора? Які є обмеження для застосування таких класифікаторів?

12. Поясніть схему алгоритму класифікації з проекцією на оптимальний напрямок.

**Задачі.**

1. Розглядається задача класифікації об'єктів, кожен з яких належить одному з двох класів. Апріорні ймовірності класів $p_k$, $k = 1, 2$. Спостережувана характеристика об'єкта $\xi$ — випадкова величина з експоненційним розподілом, інтенсивність якого $\lambda_k$ залежить від номера класу $k$, якому належит об'єкт. Опишіть баєсів класифікатор для цієї задачі. Вкажіть області значень $\xi$, для яких класифікатор відносить об'єкт до $k$-того класу.

2. Класифікуються об'єкти, які належать одному з трьох класів $k = 1, 2, 3$. Спостережувана характеристика об'єкта — двовимірний випадковий вектор $\xi = (X, Y)$ з нормальним розподілом. Апріорні ймовірності однакові для всіх класів. Коваріаційна матриця $\xi$ є одиничною матрицею. Вектор математичних сподівань: для першого класу — (0,0), для другого — (5,0), для третього — (0,5). Опишіть баєсів класифікатор для цієї задачі класифікації. Нарисуйте на папері області значень $\xi$, для яких класифікатор відносить об'єкт до $k$-того класу.

**Завдання для виконання на комп'ютері.**

**Завдання 1.**

У наборі даних `wine`[8] містяться дані про вміст певних хімічних речовин у пробах вина з трьох виноградників. Кожен рядочок відповідає одній пробі, у стовпчику Site — номер виноградника, на якому було вироблене вино, далі у кожному стовпчику відповідна характеристика вина.

Потрібно побудувати класифікатор, який за двома заданими характеристиками (наприклад, "Alcogol"та "phenols") визначатиме, на якому виноградникy вироблено дане вино. Класифікатори будуються двома способами (1) емпірично-баєсовим класифікатором з оцінюванням щільностей розподілу різних класів за допомогою багатовимірних ядерних оцінок щільності і (2) класифікатором з проекцією на оптимальний напрямок, описаним у підрозділі 5.4.3. Побудувавши класифікатори потрібно підрахувати частоту їх помилок на навчальній вибірці та зробити висновок про те, який з класифікаторів варто рекомендувати. Бажано також зобразити рисунки, які показують роботу класифікаторів.

Різні варіанти відрізняються різними обраними парами спостережуваних характеристик. (Для різних таких пар висновки щодо вибору класифікатора можуть бути різними).

[8]Цей набір даних вперше був опублікований у [6]. Він розміщений у депозитарії за адресою `https://archive.ics.uci.edu/dataset/109/wine`. Під назвою `wine` він входить у пакет `HDclassif` статистичної системи програмування `R`.

За бажанням, студенти можуть виконувати цю роботу на власних даних, або на даних, отриманих з інтернету, якщо результати будуть цікавими.

**Пари характеристик для індивідуальної роботи**

(1) alcogol, phenols.
(2) ash, phenols.
(3) alcohol, magnesium.
(4) Proanthocyanins, magnesium.
(5) Flavanoids, "Alcalinity of ash".
(6) "phenols "Malic acid".
(7) NF, Flavanoids.
(8) phenols, "Hue".
(9) "Magnesium OD.
(10) "Proanthocyanins phenols.

**Завдання 2.**

Подумайте, як можна застосувати техніку крос-валідації для вибору параметра згладжування оцінок щільності при використанні їх для емпірично-баєсової класифікації. Реалізуйте свій алгоритм у вигляді програмного коду і перевірте його роботу на даних із завдання 1. Опишіть результати, зробіть висновок про доцільність застосування цього алгоритму.

# Розділ 6

# Непараметрична регресія

## 6.1 Основні непараметричні моделі регресійного аналізу

Техніки непараметричної підгонки регресійних моделей дуже широко використовуються у прикладному статистичному аналізі даних. Одна з найбільш поширених областей їх застосування — візуальний аналіз даних. Зокрема, графіки непараметричних оцінок функції регресії зазвичай відображають на діаграмах розсіювання даних при початковому ознайомленні з даними. За формою цих графіків обирають параметричні моделі для подальшого дослідження. Тому розуміння того, як правильно проводити таку непараметричну підгонку та інтерпретувати її результати, є важливою складовою знань та вмінь професіонала-статистика.

Ми обмежимось розглядом моделі регресії з одним регресором, тобто вважатимем, що для кожного досліджуваного об'єкта спостерігаються дві характеристики $X$ і $Y$. Позначимо $X_j$ і $Y_j$ значення цих характеристик для $j$-того спостереження, $j = 1, \ldots, n$. Залежність $Y$ від $X$ описується регресійною формулою

$$Y = g(X) + \epsilon, \tag{6.1}$$

де $g$ — функція регресії, $\varepsilon$ випадкова похибка регресії.

На відміну від параметричних моделей, де функція регресії вважається відомою з точністю до невеликої кількості параметрів, у непараметричній статистиці $g$ — повністю невідома функція. Задача полягає саме в оцінюванні цієї функції.

Розрізняють дві великі групи регресійних моделей: функціональні і структурні. У функціональних моделях значення регресорів $X_j$ вважаються невипадковими фіксованими числами. Досить часто ці числа розташовуються регулярно, скажімо, популярною є модель, в якій $X_j = j/n$. Такі моделі виникають, наприклад, коли досліджується залежність якоїсь характеристики від часу спостереження, причому, дослідника цікавить "повільна" тенденція цієї залежності, що розвиваються протягом всього періоду спостережень[1]. У такому випадку природно для вимірювання часу обрати одиницю що дорівнює довжині інтервалу спостережень. Якщо спостереження проводяться регулярно, через однакові проміжки часу, в результаті отримуємо саме $X_j = j/n$ і з (6.1),

$$Y_j = g(j/n) + \varepsilon_j. \tag{6.2}$$

Тут функція $g$ визначена на інтервалі $[0, 1]$.

На похибки $\varepsilon_j$ у таких моделях накладають умови, аналогічні тим, які застосовуються у параметричному регресійному аналізі, наприклад:

1. $\mathsf{E}\,\varepsilon_j = 0$ (відсутність систематичної похибки);
2. $\mathrm{D}\,\varepsilon_j = \sigma^2$ (гомоскедастичність);
3. $\mathrm{Cov}(\varepsilon_j, \varepsilon_k) = 0$ при $j \neq k$ (некорельованість похибок).

Умови 1-3 це звичайні умови $L_2$-регресії для функціональних моделей. Якщо вважати, що, крім того, похибки є сумісно гауссовими, отримуємо непараметричну функціональну регресію з гауссовими похибками.

У структурних моделях вважається, що $Z_j = (X_j, Y_j)$ — це незалежні, однаково розподілені вектори. У цьому випадку модель (6.1) вимагає уточнення. Дійсно, залежність між двома випадковими величинами $X$ і $Y$ найповніше описується у термінах їх спільного розподілу. Тому і функція $g$ повинна якось виражатись через цей розподіл, або його характеристики, скажімо, моменти. Насправді можливі різні інтерпретації того, що слід розуміти під функцією регресії у структурній моделі. Одна з найбільш поширених спирається на трактування регресії як прогнозу для відгуку.

Отже, розглянемо всі можливі прогнози $\hat{Y}$ для випадкової величини $Y$ на основі спостереження випадкової величини $X$. Всі такі прогнози можна задавати як вимірні функції від $X$, тобто $\hat{Y} = H(X)$. Виберемо на роль критерію якості прогнозу його середньоквадратичне відхилення від

[1] А не швидкоплинні коливання навколо цієї тенденції, хоча б і регулярні.

справжнього значення прогнозованої величини (середньоквадратичний ризик):

$$\mathsf{MSE}(\hat{Y}) = \mathsf{E}(Y - \hat{Y})^2.$$

Прогноз $\hat{Y}$ тим кращий, чим менше $\mathsf{MSE}(\hat{Y})$.

**Теорема 6.1.1.** *Нехай $g(x) = \mathsf{E}[Y \mid X = x]$ умовне математичне сподівання*[2] *$Y$ при фіксованому $X$. Тоді для будь-якого прогнозу $\tilde{Y} = H(X)$ для $Y$ за $X$*

$$\mathsf{MSE}(g(X)) \leq \mathsf{MSE}(\tilde{Y}).$$

Таким чином, умовне математичне сподівання $Y$ при фіксованому $X$ є найкращим прогнозом для $Y$ (у середньому квадратичному).

**Доведення.** Нехай $g(x) = \mathsf{E}[Y \mid X = x]$. Візьмемо довільну вимірну функцію $\tilde{Y} = H(X)$ і розглянемо

$$\mathsf{MSE}(\tilde{Y}) = \mathsf{E}(Y - H(X))^2 = \mathsf{E}((Y - g(X)) - (H(X) - g(X)))^2$$

$$= \mathsf{E}(Y - g(X))^2 + \mathsf{E}(H(X) - g(X))^2 + 2\,\mathsf{E}(Y - g(X))(H(X) - g(X))$$

$$= \mathsf{MSE}(g(X)) + \mathsf{E}(H(X) - g(X))^2 + 2\,\mathsf{E}(H(X) - g(X))\,\mathsf{E}[Y - g(X) \mid X]$$

$$= \mathsf{MSE}(g(X)) + \mathsf{E}(h(X) - g(X))^2 + 0 \geq \mathsf{MSE}(g(X)).$$

Теорема доведена.

Підкреслимо, що оптимальність умовного математичного сподівання як прогнозу доведена нами лише для квадратичної функції витрат. Якщо, наприклад, на роль критерію якості обрати середнє абсолютне відхилення:

$$\mathsf{E}\,|Y - \hat{Y}|,$$

то оптимальний прогноз буде забезпечувати медіана умовного розподілу $Y$ при фіксованому $X$.

## 6.2 Прості непараметричні оцінки функції регресії

Оцінки, які ми розглянемо у цьому підрозділі, можна використовувати, як для структурних, так і для функціональних регресійних моделей.

[2] $g(x) = \mathsf{E}[Y \mid X = x]$ позначає, що $\mathsf{E}[Y \mid X] = g(X)$

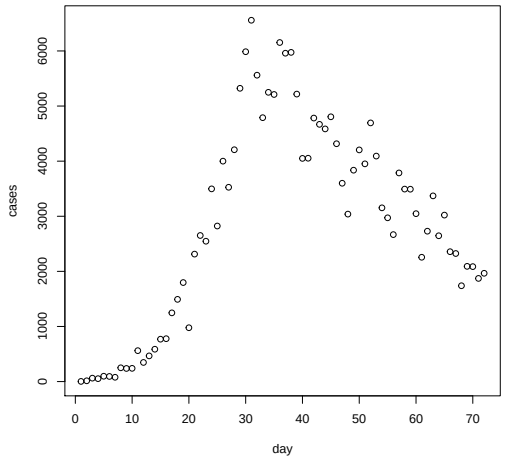



Рис. 6.1: Кількість виявлених інфікованих covid19 в Італії

Однак пояснити їх походження легше використовуючи структурний підхід. Ми так і зробимо, але ілюструвати роботу оцінок будемо на даних, для яких природним є використання структурного підходу — див. рис. 6.1 На цьому рисунку зображено дані про розвиток епідемії covid19 в Італії у 2019 році протягом перших 75 днів епідемії. По горизонталі відкладено номер дня від початку епідемії (від першого виявленого в Італії інфікованого), по вертикалі — кількість виявлених інфікованих осіб у цей день. На цій діаграмі розсіювання помітна загальна тенденція розвитку епідемії — спочатку практично експоненційний підйом, потім спад. Помітні також випадкові (а може і не випадкові) коливання кількості інфікованих у сусідні дні. Якщо потрібно усунути ці коливання і виразніше виділити загальну тенденцію, можна застосувати техніку непараметричної регресії.

Трактуючи функцію регресії $g(x)$ як умовне математичне сподівання $Y$ при фіксованому $X$, природно було б оцінювати її усереднюючи значення $Y$ для даного $X = x$. Але у наших даних кожному $X = 1, \ldots, n$ відповідає в точності одне $Y$, тому таке усереднення не усуне випадкові коливання. Ідея полягає в тому, щоб для фіксованого $x$ усереднювати значення $Y$ для всіх спостережень, для яких змінна $X$ потрапила близь-

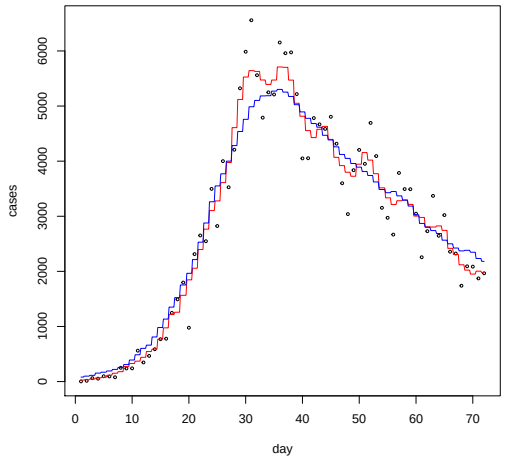



Рис. 6.2: Оцінки ковзаючого середнього для епідемії covid19 в Італії

ко до $x$. В результаті приходимо до такої оцінки:

$$\hat{g}^{mean}(x) = \frac{\sum_{j=1}^{n} Y_j \mathbb{1}\{|x - X_j| < h/2\}}{\sum_{j=1}^{n} \mathbb{1}\{|x - X_j| < h/2\}} \tag{6.3}$$

— це середнє значення $Y_j$ для всіх $j$, для яких $X_j \in [x-h/2, x+h/2]$. Відрізок $[x-h/2, x+h/2]$ називають "ковзаючим вікном", $h$ — шириною вікна, а оцінку (6.3) — оцінкою ковзаючого середнього. Ширина вікна для цієї оцінки є параметром налаштування. При зростанні $h$ розкид оцінки зменшується, тому що вона враховує більше спостережень, які потрапляють до вікна. Але при цьому оцінка сильніше згладжує, усереднює справжні коливання функції регресії, тобто збільшується її зміщення. Тому ширину вікна потрібно обирати виходячи з якихось міркувань оптимальності, балансу між зміщенням та дисперсією, подібно до того, як обирається параметр згладжування ядерної оцінки щільності (див. розділ 4).

Приклади графіка цієї оцінки для даних про епідемію в Італії з $h = 5$ (червона лінія) і $h = 15$ (блакитна лінія) зображені на рис. 6.2. Помітно, що при збільшенні ширини вікна оцінка сильніше згладжується, причому у неї збільшується відхилення від "загальної тенденції" в моменти швидких поворотів.

Щоб зробити оцінку неперервною функцією, можна замінити у ковзаючому середньому індикатори на гладенькі ядра, наприклад, на ядро

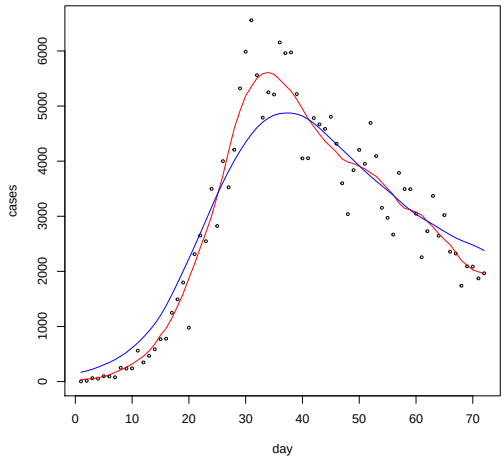



Рис. 6.3: Оцінки Надарая-Ватсона для епідемії covid19 в Італії

Єпанєчнікова. В результаті отримуємо оцінку Надарая-Ватсона:

$$\hat{g}^{NW}(x) = \frac{\sum_{j=1}^{n} Y_j K\left(\frac{x-X_j}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)}, \tag{6.4}$$

тут $K : \mathbb{R} \to \mathbb{R}$ — ядро оцінки, $h > 0$ — параметр згладжування. Приклади графіка оцінки Надарая-Ватсона з ядром Єпанєчнікова для даних про епідемію в Італії з $h = 5$ (червона лінія) і $h = 15$ (блакитна лінія) зображені на рис. 6.3. Як бачимо, у цій оцінці параметр $h$ грає ту ж роль, що і у оцінці ковзаючого середнього.

Інший варіант оцінки для функції регресії $g(x)$ отримуємо, якщо визначати $g(x)$ як медіану умовного розподілу $Y$ при фіксованому $X$. Ця оцінка зветься ковзаючою медіаною. У точці $x$ она визначається як медіана[3] набору $Y_j$ для всіх $j$, для яких $X_j \in [x - h/2, x + h/2]$. Таким чином,

$$\hat{g}^{median}(x) = \mathsf{med}\{Y_j |\ j \in \{1, \dots, n\} :\ |x - X_j| < h/2\}. \tag{6.5}$$

[3]Нагадаємо, що вибіркова медіана — це вибірковий квантиль, рівня 1/2, див. формулу (2.3) у п. 2.2).

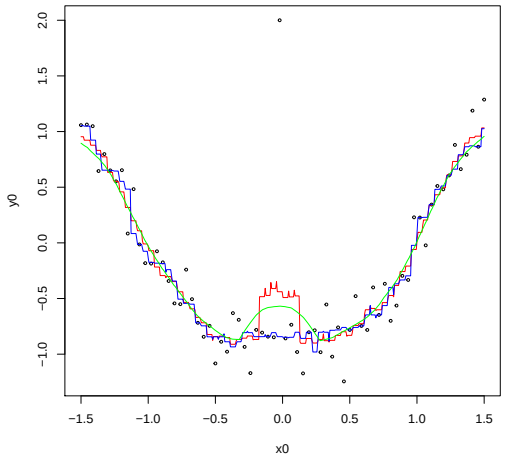



Рис. 6.4: Ковзаюча медіана (синя) не реагує на викид. Оцінка Надарая-Ватсона — зелена, ковзаюче середнє — червона

Відмітимо, що коли у формулі (6.1) похибки регресії $\varepsilon$ мають симетричний розподіл з нульовим математичними сподіванням, то ковзаюча медіана оцінює ту ж саму функцію регресії, що і ковзаюче середнє та оцінка Надарая-Ватсона. Якщо медіана похибки не дорівнює математичному сподіванню, то оцінювані функції будуть зміщені одна відносно одної.

Ковзаюча медіана є більш стійкою по відношенню до забруднень викидами, ніж ковзаюче середнє — див. рис. 6.4. Крім того, ковзаюча медіана зручна при оцінюванні функцій регресії, що мають помітні розриви, оскільки вона не згладжує їх так, як це роблять оцінки ковзаючого середнього та Надарая-Ватсона — див. рис. 6.5. З іншого боку, якщо похибки регресії є гауссовими, то оцінки ковзаючого середнього є більш точними ніж ковзаюча медіана для гладеньких функцій регресії. Для даних про епідемію в Італії, ковзаючі медіани зображені на рис. 6.6.

Порівняння ефективності різних технік непараметричного оцінювання функції регресії можливе із застосуванням асимптотичного підходу, аналогічного розглянутому у підрозділах 4.2-4.6 для ядерних оцінок щільності. Ми обмежимось лише одним прикладом такого асимптотичного результату для оцінок Надарая — Ватсона.

Спочатку опишемо акуратно теоретичну регресійну модель, в рамках якої буде досліджуватись асимптотична поведінка оцінок. Ми розгляда-

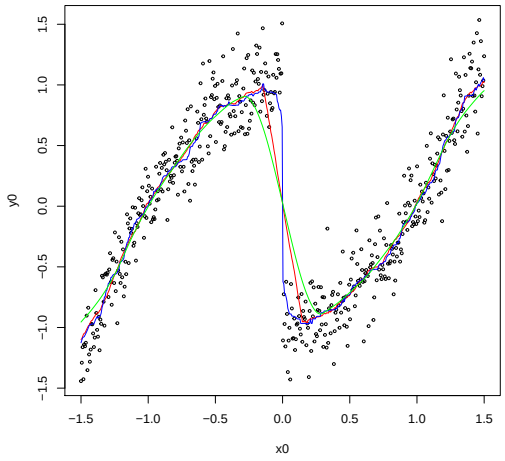



Рис. 6.5: Ковзаюча медіана (синя) не згладжує розрив. Оцінка Надарая-Ватсона — зелена, ковзаюче середнє — червона

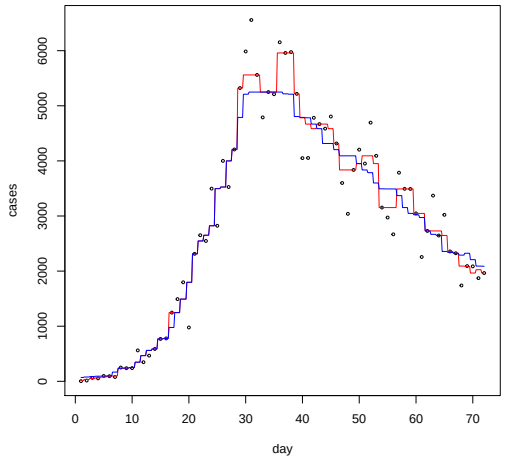



Рис. 6.6: Оцінки ковзаючих медіан для епідемії covid19 в Італії ($h = 5$ — червона, $h = 15$ — синя)

ємо вибірку з незалежних однаково розподілених двовимірних спостережень $(X_j, Y_j)$, $j = 1, \ldots, n$, які описуються непараметричною структурною регресійною моделлю:

$$Y_j = g(X_j) + \varepsilon_j, \;\; j = 1, \ldots, n,$$

де $g$ — невідома функція регресії, яку потрібно оцінити, $\varepsilon_j$ — похибки регресії.

Ми будемо вважати, що для похибок $\varepsilon_j$ виконуються наступні умови:

1. $\mathsf{E}\,\varepsilon_j = 0$,
2. $\mathrm{D}\,\varepsilon_j = \sigma^2 < \infty$,
3. Випадкові величини $X_j$ та $\varepsilon_j$ є незалежними між собою.

Будемо також вважати, що розподіл регресорів $X_j$ є неперервним і позначимо його щільність $f_X(x)$. Відносно функції регресії $g(x)$, будемо припускати, що вона є двічі неперервно диференційовною по $x$. Першу і другу похідні $g(x)$ позначимо $g'(x)$ і $g''(x)$.

Як і у розділі 4 будемо позначати

$$d^2(K) = \int_{-\infty}^{\infty} K^2(z)dz, \;\; D(K) = \int_{-\infty}^{\infty} z^2 K(z)dz,$$

вважаючи ці інтеграли скінченними. Будемо вважати, що $K$ — ядро другого порядку, отже, крім скінченності $d^2(K)$ і $D(K)$ потрібно також, щоб виконувались умови

$$\int_{-\infty}^{\infty} K(z)dz = 1, \;\; \int_{-\infty}^{\infty} zK(z)dz = 0.$$

Точність оцінки $\hat{g}_n(x)$ у фіксованій точці $x$ будемо характеризувати середньоквадратичним ризиком

$$\mathsf{MSE}(\hat{g}_n(x)) = \mathsf{E}(\hat{g}_n(x) - g(x))^2.$$

Наступна теорема (див. [23], теорема 4.3) описує поведінку середньоквадратичного ризику оцінки Надарая-Ватсона $\hat{g}^{NW}(x)$, визначеної формулою (6.4).

**Теорема 6.2.1.** *Нехай виконуються наступні умови:*

*1.* $\int_{-\infty}^{\infty} |K(z)|dz < \infty, \;\; \lim_{z\to\pm\infty} zK(z) = 0.$

*2.* $\mathsf{E}(X_j)^2 < \infty.$

*3. Функція $f_X$ є неперервно диференційовною в точці $x$ і $f_X(x) > 0$.*
*4. При $n \to \infty$, $h = h_n \to 0$ і $nh_n \to \infty$.*
*Тоді*

$$\mathsf{MSE}(\hat{g}^{NW}(x)) \approx \frac{1}{nh_n}\frac{\sigma^2 d^2(K)}{f_X(x)} + \frac{h^4}{4}\left\{g''(x) + 2\frac{g'(x)f'_X(x)}{f_X(x)}\right\}^2 (D(K))^2.$$

З цієї теореми легко бачити[4] що, у випадку двічі неперервно диференційовної функції регресії, для забезпечення оптимальної швидкості збіжності оцінок Надарая — Ватсона потрібно вибрати параметр згладжування $h_n = Hn^{-1/5}$, де константа $H$ визначається за даними за допомогою додаткових процедур. При цьому для середньоквадратичного ризику досягається швидкість збіжності порядку $\mathsf{MSE}(\hat{g}^{NW}(x)) = O(n^{-4/5})$.

Які можуть бути алгоритми визначення $H$? Це можна зробити використовуючи адаптивну техніку, подібну до тієї, яка була застосована у підрозділі 4.4 для вибору параметра згладжування оцінки щільності. Такі техніки обговорюються у підрозділі 4.2 книги [15]. Або можна скористатись технікою крос-валідації.

Як застосувати загальний підхід крос-валідації до оцінок функції регресії? Знову, ми продемонструємо це на прикладі оцінок Надарая — Ватсона, хоча так само можна підбирати параметр згладжування і для підгонки функції регресії іншими непараметричними та параметричними методами.

Для того, щоб перевірити, наскільки адекватно оцінка $\hat{g}_n^{NW}(x)$ прогнозує значення відгуку для $i$-того спостереження, не можна просто підставити значення регресора замість аргумента в оцінку і подивитись на відхилення прогнозу від справжнього значення відгуку: $Y_i - \hat{g}_n^{NW}(X_i)$ — тобто на залишок регресії. Дійсно, ми вже використали значення $(X_i, Y_i)$ для підгонки $\hat{g}_n^{NW}(x)$, тому ця оцінка "знає наперед" $Y_i$, а не прогнозує його. Щоб забезпечити "чистоту експерименту" для характеризації якості оцінювання ми спочатку знаходимо оцінку Надарая — Ватсона, підігнану за всіма спостереженнями, крім $i$-того:

$$\hat{g}_{-i}^{NW}(x) = \frac{\sum_{j\neq i} Y_j K\left(\frac{x-X_j}{h}\right)}{\sum_{j\neq i}^{n} K\left(\frac{x-X_j}{h}\right)}.$$

[4] Аналогічно тому, як це було зроблено у підрозділі 4.4 для ядерних оцінок щільності

Тепер порівнюємо прогноз для $Y_i$ на основі цієї оцінки зі справжнім значенням відгуку і усереднюємо квадрати відхилень по всіх можливих $i$. В результаті отримуємо функціонал крос-валідації:

$$\mathsf{CV}(h) = \sum_{i=1}^{n} (Y_i - \hat{g}_{-i}^{NW}(X_i))^2.$$

Чим менші відхилення прогнозу від справжніх значень відгуку, тим краща оцінка. Отже, можна вибирати параметр згладжування $h$ з умови мінімізації $\mathsf{CV}(h)$:

$$h^{CV} = \operatorname{argmin}_h \mathsf{CV}(h).$$

Крос-валідація та деякі інші підходи до вибору параметра згладжування у регресійних моделях розглядаються в підрозділі 4.3 книги [23].

## 6.3 Локально-лінійні і локально-поліноміальні оцінки функції регресії

Всі оцінки розглянуті у п. 6.2, мають важливий спільний недолік — вони добре оцінюють функцію регресії лише у точках $x$, поруч з якими є досить спостережуваних $X_j$ **по обидва боки**. Якщо всі (або переважна більшість) значення регресорів у вибірці розташовані по один бік від $x$, то у цій точці оцінка буде відхилятись від загальної тенденції у бік відповідних значень $Y$. Це називають **крайовим ефектом**. Приклад такого ефекту зображено на рис. 6.7 — тут точки спостережень лежать на прямій лінії. Оцінка ковзаючого середнього добре наближає цю лінію всередині інтервалу спостережуваних значень регресора, але на кінцях помітно відхиляється від неї.

Щоб усунути крайовий ефект розроблено багато підходів. Один з найбільш популярних — техніка локально-поліноміальної регресії. Ми розглянемо її у найпростішому випадку — локально лінійної регресії.

Нехай потрібно оцінити функцію регресії $g(x)$ у точці $x = x_0$ за спостереженнями $(X_j, Y_j)$, $j = 1, \ldots, n$, причому

$$Y_j = g(X_j) + \varepsilon_j.$$

Ідея методу полягає в тому, щоб в околі точки $x_0$ наблизити функцію $g$ прямою вигляду $y = b_0 + b_1(x - x_0)$, де коефіцієнти $b_0 = b_0(x_0)$ і $b_1 = b_1(x_0)$

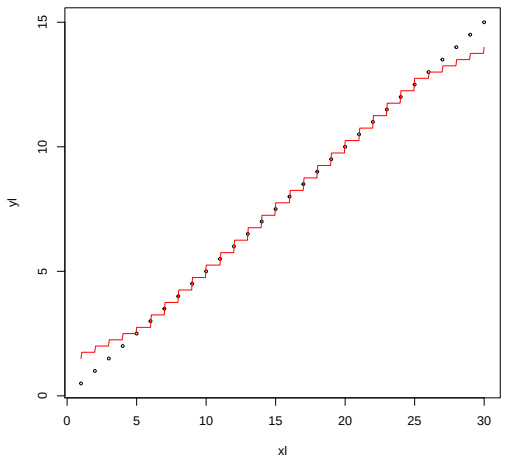



Рис. 6.7: Крайовий ефект: оцінка ковзаючого середнього

підганяються з використанням навантаженого методу найменших квадратів. При цьому вагові коефіцієнти функціоналу найменших квадратів підбирають так, щоб чим ближче розташовані значення регресора $X_j$ до $x_0$, тим більшою була б вага відповідного доданка у функціоналі:

$$J(b_0, b_1, x_0) = \sum_{j=1}^{n} w_j(x_0)(Y_j - b_0 - b_1(X_j - x_0))^2,$$

де

$$w_j(x_0) = K\left(\frac{x_0 - X_j}{h}\right),$$

тут

$K : \mathbb{R} \to \mathbb{R}$ — ядро, на роль якого вибирають парну функцію $K(z)$, що спадає до 0 при зростанні $|z|$,

$h$ — параметр згладжування оцінки.

Позначимо

$$(\hat{b}_0(x_0), \hat{b}_1(x_0)) = \operatorname{argmin}_{(b_0,b_1)\in\mathbb{R}^2} J(b_0, b_1, x_0)$$

— оцінка методу найменшик квадратів для коефіцієнтів прямої.

Тепер оцінка методу локально-лінійної регресії для $g(x_0)$ визначається як

$$\hat{g}^{ll}(x_0) = \hat{b}_0(x_0).$$

Запишемо цю оцінку у явному вигляді. Для знаходження мінімуму прирівняємо до нуля похідні $J$ по $b_0$ і $b_1$:

$$\begin{aligned} \sum_{j=1}^{n} w_j(x_0)(Y_j - b_0 - b_1(X_j - x_0)) &= 0, \\ \sum_{j=1}^{n} w_j(x_0)(Y_j - b_0 - b_1(X_j - x_0))(X_j - x_0) &= 0. \end{aligned} \tag{6.6}$$

Введемо такі позначення:

$$X'_j = X_j - x_0,\ m_w = \sum_{j=1}^{n} w_j(x_0),$$

$$m_x = \sum_{j=1}^{n} w_j(x_0)X'_j,\ m_y = \sum_{j=1}^{n} w_j(x_0)Y_j,$$

$$m_{xx} = \sum_{j=1}^{n} w_j(x_0)(X'_j)^2,\ m_{xy} = \sum_{j=1}^{n} w_j(x_0)X'_jY_j.$$

Тоді рівняння (6.6) зводяться до:

$$\begin{aligned} m_w b_0 + m_x b_1 &= m_y \\ m_x b_0 + m_{xx} b_1 &= m_{xy}. \end{aligned} \tag{6.7}$$

Розв'язуючи ці рівняння відносно $b_0$, $b_1$, знаходимо локально лінійну оцінку.

$$\hat{g}^{ll}(x_0) = \hat{b}_0(x_0) = \frac{m_y m_{xx} - m_x m_{xy}}{m_w m_{xx} - (m_x)^2}.$$

Помітимо, що коли точки $X_j$ в околі $x_0$ розташовані приблизно симетрично, а $K$ — парна функція, то $m_x \approx 0$, отже

$$\hat{g}^{ll}(x_0) \approx \frac{m_y}{m_w} = \frac{\sum_{j=1}^{n} Y_j K((x - X_j)/h)}{\sum_{j=1}^{n} K((x - X_j)/h)},$$

тобто у цьому випадку локально-лінійна оцінка мало відрізняється від оцінки Надарая-Ватсона. Але там, де симетрія $X_j$ порушується, локально лінійна оцінка виправляє поведінку оцінки Надарая-Ватсона.

Відмітимо, що оцінку Надарая-Ватсона у цьому контексті можна розглядати як "локально сталу" оцінку, тобто як результат мінімізації по $b_0$ навантаженого функціонала методу найменших квадратів:

$$J(b_0, x_0) = \sum_{j=1}^{n} w_j(x_0)(Y_j - b_0)^2.$$

Якщо потрібне більш акуратне згладження та усунення крайових ефектів, можна також використовувати локальну підгонку поліномами старших ступенів. Така техніка має назву локально поліноміальної регресії.

Асимптотичне дослідження локально-поліноміальних оцінок та алгоритми вибору параметра згладжування для них можна знайти у книзі [15], розділи 3 і 4.

## 6.4 Локальний вибір параметра згладжування. Алгоритм Lowess

У цьому підрозділі ми розглянемо алгоритм оцінювання функції регресії Lowess (locally weighted scatterplot smoothing)[5] Запропонований В. С. Клівлендом у роботі [11], він є нині одним з найбільш поширених у задачах згладжування експериментальних кривих. Часто його використовують за умовчанням при відображенні діаграм розсіювання, щоб показати залежність між змінними "в середньому". Зокрема, так зроблено у багатьох функціях R, які виводять діаграми розсіювання, наприклад, залишків регресії та ін.

У алгоритмі Lowess реалізовані три ідеї непараметричного регресійного аналізу:

— локально поліноміальна підгонка функції регресії;

— локальний вибір параметра згладжування за допомогою техніки найближчих сусідів (nearest neighbour);

— додаткова підгонка з уточненням вагових коефіцієнтів (reweighting) для забезпечення робастності.

Локально-поліноміальну регресію у її простішому (лінійному) варіанті ми розглянули у попередньому підрозділі. Подивимось, як вона виглядає у алгоритмі Lowess. Класична версія алгоритму використовує вагові

[5] Інколи версію цього алгоритму називають Loess (locally estimated scatterplot smoothing).

коефіцієнти

$$w_j(x_0) = K\left(\frac{x_0 - X_j}{h(x_0)}\right),$$

де $x_0$ — точка, у якій будується оцінка для функції регресії,

$h(x_0)$ — параметр згладжування, що обирається індивідуально для кожного $x_0$,

$K(z)$ — бікубічне ядро:

$$K(z) = \begin{cases} (1 - |z|^3)^3. & \text{якщо } |z| < 1; \\ 0 & \text{якщо } |z| \geq 1. \end{cases}$$

З такими ваговими коефіцієнтами будується навантажений функціонал методу найменших квадратів для підгонки кривої регересії поліномом ступеня $d$:

$$J(\mathbf{b}, x_0) = \sum_{j=1}^{n} w_j(x_0)(Y_j - b_0 - b_1 X_j - b_2 (X_j)^2 - \dots b_d (X_j)^d)^2.$$

Тут $\mathbf{b} = (b_0, \dots, b_d)$ — вектор невідомих коефіцієнтів, які підбирають так, щоб мінімізувати цей функціонал:

$$\hat{\mathbf{b}}(x_0) = (\hat{b}_0(x_0), \dots, \hat{b}_d(x_0)) = \operatorname{argmin}_{\mathbf{b} \in \mathbb{R}^{d+1}} J(\mathbf{b}, x_0).$$

Значення "прогнозу" для $Y$ в точці $x_0$ тепер визначається як

$$\hat{Y}(x_0) = \hat{b}_0(x_0) + \sum_{k=1}^{d} \hat{b}_k(x_0)(X_j)^k.$$

У звичайній локально поліноміальній регресії це була б остаточна оцінка для функції регресії в точці $x_0$, тобто для $g(x_0)$. Але не будемо поспішати.

Розглянемо тепер питання вибору параметра згладжування. Як і в усіх інших таких алгоритмах, його можна трактувати як "півширину вікна", через яке ми дивимось на вибірку. Оскільки вибране нами ядро обертається у 0 поза інтервалом $[-1, 1]$, на оцінку в точці $x_0$ впливають тільки ті спостереження, у яких $X_j \in [x_0 - h, x_0 + h]$. Якщо обрати ширину вікна маленькою, в нього потрапить мало спостережень і оцінка буде сильно коливатись. Якщо ширина вікна велика — усереднення буде йти по великому інтервалу і не дасть нам змоги побачити дрібні деталі

оцінюваної функції. Ми намагаємось обрати ширину балансуючи ці дві небезпеки. Якщо у різних областях можливих значень $x$ густина $X_j$, що потрапляють у них різна, то природно і $h$ у цих областях вибирати різне. Там, де багато $X_j$ потрапляють поруч одне з одним, краще взяти $h$ менше, щоб роздивитись дрібні деталі. Там, те $X_j$ мало, і вони розташовані далеко один від одного, вікно треба розтягнути. Подробиць тут все одно не побачиш, а випадкові коливання можуть замаскувати реальну тенденцію.

Ідея — обирати $h$ так, щоб у вікні завжди була одна й та сама кількість спостережень, наприклад — $r$. Для цього підрахуємо відстані від $x_0$ до всіх $X_j$

$$s_j = |x_0 - X_j|,$$

впорядкуємо їх в порядку зростання і покладемо $h$ рівним $r+1$ значенню $s_j$ у цьому впорядкованому наборі. Тобто $h$ — відстань від $x_0$ до $r+1$ найближчого сусіда у вибірці (по осі абсцис). Якщо у вибірці немає об'єктів з однаковими значеннями $X_j$, то у вікно кожного разу буде видно $r$ спостережень. ($r+1$ опиниться на межі вікна і не буде впливати на оцінку).

Залишається відкритим питання про вибір $r$. Ця величина знову відіграє роль параметра згладжування, але тепер уже "глобального": вона підбирається одна для всіх значень $x_0$. Часто $r$ визначають як частку вибірки $r = fn$. У стандартній версії функції `lowess()`, реалізованій у системі статистичного програмування R, за умовчанням встановлено $f = 2/3$, але це, звичайно, занадто велике значення, якщо хотіти по справжньому непараметричної оцінки. При збільшенні обсягу вибірки $f$ доцільно зменшувати.

Нарешті, розберемось із технікою, яка дозволяє зробити підгонку кривої регресії більш робастною (стійкою по відношенню до забруднень викидами). Логіка тут така. Отримавши оцінки для параметрів $\mathbf{b}(x_0)$ ми будуємо прогнози $\hat{Y}(x_0)$ для значень $x_0 = X_j$, тобто $\hat{Y}_j = \hat{Y}(X_j)$ і розраховуємо залишки

$$U_j = Y_j - \hat{Y}_j.$$

Великі значення залишків вказують на те, що дане спостереження є викидом, що може відхилити підігнану криву регресії від тієї справжньої середньої тенденції, яку нам потрібно побачити. Тому ми будуємо оцінки "наступного кроку", в яких знову використовується навантажений функціонал найменших квадратів, але тепер вагові коефіцієнти виправля-

ються так, щоб можливі викиди у них отримали меншу вагу. А саме, підраховується

$$\mu = \mathsf{med}\{|U_1|, \dots, |U_n|\}$$

і "поправка на викид" для $j$-го спостереження визначається як

$$\delta_j = B(U_j/(6\mu)),$$

де

$$B(z) = \begin{cases} (1 - |z|^2)^2. & \text{якщо } |z| < 1, \\ 0 & \text{якщо } |z| \geq 1 \end{cases}$$

— так зване "біквадтатне ядро".

Тепер оцінка наступного кроку для коефіцієнтів $\mathbf{b}$ визначається як точка мінімуму навантаженого функціонала найменших квадратів з ваговими коефіцієнтами

$$w'_j(x_0) = \delta_j w_j(x_0).$$

Отримані оцінки використовують для прогнозування так само, як оцінки першого кроку. Після цього можна знову повторити крок: обчислити залишки прогнозу, знайти поправки до вагових коефіцієнтів і розрахувати нові оцінки.

У стандартному алгоритмі за умовчанням кількість таких кроків "робастифікації" оцінок дорівнює трьом.

Оцінка для функції регресії, обчислена на останньому кроці є остаточною оцінкою, яка видається як результат роботи алгоритму.

Як працює алгоритм Lowess на даних з викидом, можна побачити на рис. 6.8.

## 6.5 Проекційні оцінки функції регресії

Оцінки функції регресії розглянуті вище, забезпечують хорошу точність оцінювання. Однак всі вони мають спільний недолік: для того, щоб мати змогу обчислити таку оцінку у заданій точці, потрібно пам'ятати всі дані спостережень $(X_j, Y_j)$, $j = 1, \dots, n$. Це незручно, коли даних багато, а оцінка має використовуватись багато разів. Тому для багатократного використання застосовують оцінки, у яких для обчислення достатньо пам'ятати порівняно невелику кількість параметрів. Один з популярних варіантів — проекційні оцінки.

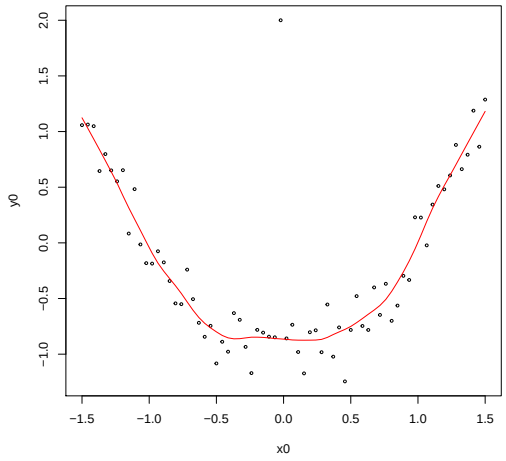



Рис. 6.8: Оцінка алгоритму Lowess на даних з викидом (червона лінія)

Розглянемо функцію регресії $g : [a_1, a_2] \to \mathbb{R}$ як елемент функціонального простору $L_2[a_1, a_2]$. Нехай у цьому просторі задано набір базисних функцій[6] $e_1$, $e_2$,... Тоді довільну функцію $g \in L_2[a_1, a_2]$ можна розкласти за цим базисом:

$$g(x) = \sum_{i=1}^{\infty} b_i e_i(x),$$

де $b_i$ — деякі числові коефіцієнти.

Ідея проекційної оцінки полягає в тому, щоб обрізати цю суму на деякому $i = M$ і розглядати отриману формулу як наближену параметричну модель для $g$:

$$g(x) \approx g^M(x) = \sum_{i=1}^{M} b_i e_i(x).$$

Функцію $g^M$ можна розглядати як ортогональну проекцію функції $g$ на лінійний підпростір $E_M$ у $L_2[a_1, a_2]$, натягнутий на функції $e_1, \ldots, e_M$.

Відповідно, "точна" регресійна модель

$$Y_j = g(X_j) + \varepsilon_j,$$

[6]Не обов'язково, щоб цей базис був ортонормованим.

замінюється наближеною:

$$Y_j = g^M(X_j) + \varepsilon_j = \sum_{i=1}^{M} b_i e_i(X_j) + \varepsilon_j,\ j = 1, \dots n.$$

Це — лінійна регресійна модель для відгуку $Y_j$ і регресорів $X_j^i = e_i(X_j)$. Відповідно, оцінки для $\mathbf{b} = (b_1, \dots, b_M)^T$ у цій моделі можна шукати методом найменших квадратів. Запишемо такі оцінки у явному вигляді. Позначимо

$$\mathbf{X} = \begin{pmatrix} e_1(X_1) & e_2(X_1) & \dots & e_M(X_1) \\ e_1(X_2) & e_2(X_2) & \dots & e_M(X_2) \\ \vdots & \vdots & \ddots & \vdots \\ e_1(X_n) & e_2(X_n) & \dots & e_M(X_n) \end{pmatrix}.$$

Тепер вектор оцінок $\hat{\mathbf{b}} = (\hat{b}_1, \dots, \hat{b}_M)^T$ для $\mathbf{b}$ можна записати як

$$\hat{\mathbf{b}} = \mathbf{A}^{-1}\mathbf{X}^T\mathbf{Y},$$

де $\mathbf{A} = \mathbf{X}^T\mathbf{X}$, $\mathbf{Y} = (Y_1, \dots, Y_n)^T$.

Оцінка для функції регресії $g$ має вигляд:

$$\hat{g}(x) = \sum_{i=1}^{M} \hat{b}_i e_i(x).$$

По суті, ця оцінка оцінює не $g(x)$, а $g^M(x)$. Оскільки $g^M(x)$ трактують як проекцію $g$ на $E_M$, то і оцінку $\hat{g}$ називають **проекційною оцінкою**.

Число $M$ — вимірність простору проекції, є у цій оцінці параметром налаштування, подібним до параметра згладжування $h$ у оцінці Надарая-Ватсона. При великих $M$ проекційна оцінка матиме велику дисперсію, тобто буде недозгладженою. При малих — велике зміщення, тобто буде перезгладженою. Для знаходження балансу можна використовувати різні методи, зокрема — крос-валідацію. Слід мати на увазі, що для забезпечення консистентності оцінки, при зростанні обсягу вибірки $n$ потрібно збільшувати вимірність проекції $M$. Типовий порядок оптимального вибору для вимірності простору проекції —

$$M = Cn^{1/5}.$$

Таким чином, для багатократного використання проекційної оцінки також потрібно зберігати у пам'яті багато ($M$) чисел. Але $M$ значно менше,

ніж $n$ — обсяг інформації, потрібний для обчислення оцінки Надарая-Ватсона. Інакше кажучи, проекційні оцінки дозволяють стиснення інформації, яка потрібна для їхнього обчислення.

Якщо значення регресора $X_j$ розташовані більш-менш рівномірно на інтервалі $[a_1, a_2]$, то зручно базисні функції $e_i$ обирати взаємно ортогональними в просторі $L_2[a_1, a_2]$. (Про системи ортогональних функцій у просторі $L_2[a_1, a_2]$ див. у підрозділі 3.3) Тоді матриця $\mathbf{A}$ буде близькою до діагональної (діагональні елементи значно більші ніж позадіагональні) і тому не виникатиме проблем з точністю обчислення оберненої матриці.

Зокрема, можна вибрати на роль базису поліноми Лежандра.

Стандартні поліноми Лежандра це система поліномів, ортогональних у просторі $L_2[-1, 1]$. Їх можна задати використовуючи рекурентну формулу:

$$P_0(x) = 1; P_1(x) = x;$$

$$P_{i+1} = \frac{2i+1}{i+1} x P_i(x) - \frac{i}{i+1} P_{i-1}(x), \text{ для } i = 2, 3, \ldots$$

Для того, щоб отримати систему поліномів, ортогональних на заданому відрізку $[a_1, a_2]$, досить зробити лінійне відображення області визначення $[a_1, a_2] \to [-1, 1]$:

$$e_i(x) = P_i\left(2\frac{x - a_1}{a_2 - a_1} - 1\right).$$

Можна також використовувати тригонометричні базиси, що складаються з функцій $1$, $\sin(kx)$, $\cos(kx)$, $k = 1, 2, \ldots$, див. підрозділ 3.3.

Але і поліноміальні і тригонометричні базиси мають спільний недолік: проекційні оцінки з такими базисами погано реагують на локальні коливання функції регресії — див. рис. 6.9–6.11, де зображено підгонку даних про епідемію covid19 в Італії за допомогою поліноміальної регресії з тригонометричним базисом, базисом Лежандра і з використанням алгоритму Lowess. Тому на практиці більш доцільно використовувати базиси, що дозволяють локальне налаштування оцінки, такі як сплайнові або вейвлет-базиси. Про сплайни та їхнє використання для підгонки регресійних залежностей див. [14], про вейвлети та їх застосування у статистиці — [22, 2].

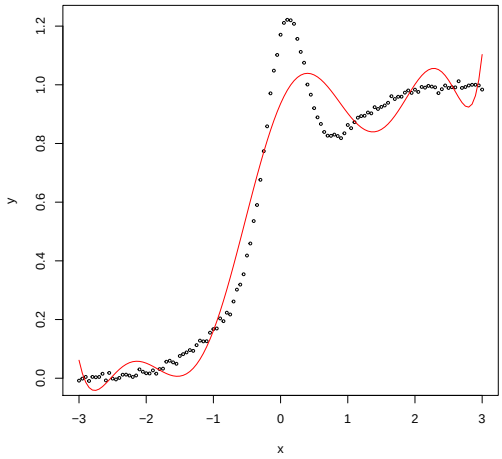



Рис. 6.9: Проекційна оцінка за базисом Лежандра ($M = 10$)

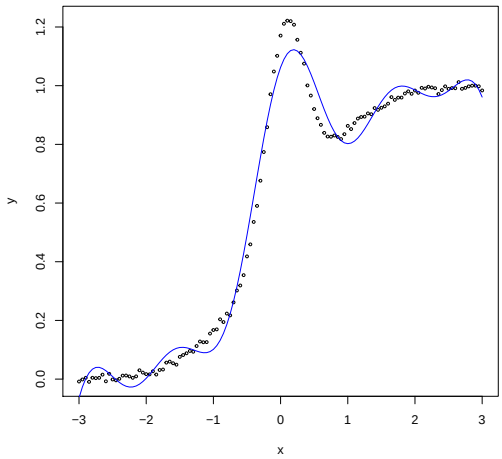



Рис. 6.10: Проекційна оцінка за тригонометричним базисом ($M = 11$)

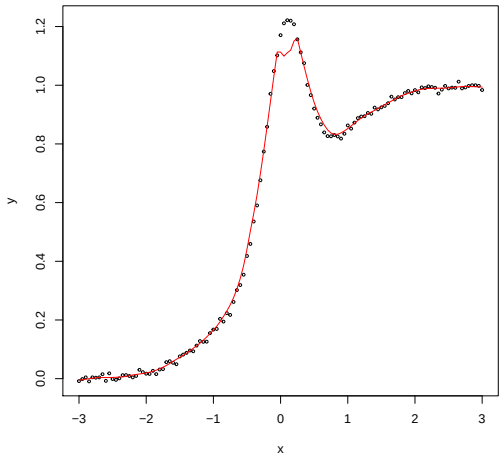



Рис. 6.11: Згладжування даних алгоритмом Lowess ($f = 0.1$)

## 6.6 Запитання і задачі

**Запитання.**

1. Що таке середньоквадратичний ризик прогнозу? Який прогноз є оптимальним виходячи з умови мінімізації середньоквадратичного ризику?

2. Як визначається оцінка ковзаючого вікна для функції регресії? Як залежать зміщення і розкид цієї оцінки від ширини вікна?

3. В яких випадках доцільно використовувати для оцінювання функції регресії метод ковзаючої медіани?

4. Що таке крайовий ефект при непараметричному оцінюванні функції регресії? Які оцінки виявляють крайовий ефект? Яку оцінку можна використати, щоб позбутись цього ефекту?

5. Чи можна використовувати техніку локально-лінійної регресії для підгонки розривних функцій регресії?

6. Поясніть, як у алгоритмі Lowess забезпечується стійкість оцінок по відношенню до забруднень викидами.

7. Які непараметричні оцінки функції регресії забезпечують можливість стиснення інформації?

**Задачі**

1. Отримайте явні формули для обчислення оцінки локально поліно-

міальної регресії ступеня 2 (локально-параболічна регресія).

2. Використовуючи теорему 6.2.1, виведіть формулу для теоретично-оптимального параметра згладжування оцінки Надарая — Ватсона у фіксованій точці $x$. Функції $g$ і $f$, а також дисперсію $\sigma^2$ вважайте відомими.

**Завдання 1.**

У цому завданні спочатку потрібно згенерувати, що складаються з пар чисел $(X_j, Y_j)$, які пов'язані регресійною залежністю:

$$Y_j = g(X_j) + \varepsilon_j, j = 1, \ldots, n$$

де $g$ — задана функція регресії, $\varepsilon_j$ — випадкова похибка регресії. Функція регресії вважається невідомою, її треба оцінити за даними. Для оцінювання використовуються: ковзаюче середнє, ковзаюча медіана, оцінка Надарая-Ватсона з ядром Єпанєчнікова, локально-лінійна оцінка та проекційні оцінки із заданими базисними функціями. Потрібно реалізувати ці оцінки у вигляді функцій та вивести їх графіки разом з даними для порівняння. Висновки про поведінку оцінок можна робити за цими графіками "на око".

Бажано подивитись, як змінюються оцінки при зміні допоміжних параметрів — таких, як ширина вікна, параметр згладжування, вимірність простору проекції.

Добре також ввести у вибірку викид і подивитись, як відреагують на нього різні оцінки.

**Індивідуальні завдання роботи.**

У завданні вказано розподіл змінної $X$ розподіл $\varepsilon$ та функцію регресії $g$. Обсяг вибірки $n = 300$.

Потрібно згенерувати дані відповідно до моделі регресії та оцінити функцію регресії, використовуючи оцінки

— у завданнях (1-5) оцінка ковзаючого середнього + ковзаюча медіана + локально-лінійна регресія з ядром Єпанєчнікова;

— у завданнях (6-10) ковзаюча медіана + оцінка Надарая-Ватсона з ядром Єпаначнікова + проекційна оцінка з базисом поліномів Лежандра.

Підбір параметра згладжування та вимірності простору проекції виконувати на око.

(1) $X \sim N(0, 2)$, $\varepsilon$ — рівномірний на [-0.3,0.3],

$$g(x) = \begin{cases} -x^2 + 1 & \text{при } x < 0 \\ x^2 - 1 & \text{при } x \geq 0 \end{cases}$$

(2) $X$ рівномірно розподілений на [0,3], $\varepsilon \sim N(0, 0.3)$,

$$g(x) = 3x(9 - 9x + 2x^2)$$

(3) $X$ рівномірно розподілений на [0,3], $\varepsilon$ рівномірно розподілений на [-1,1],

$$g(x) = -\frac{x}{7}(-242 + 805x - 742x^2 + 200x^3)$$

(4) $X \sim N(0, 1)$, $\varepsilon \sim N(0.5)$,

$$g(x) = 2|x|$$

(5) $X \sim N(0, 2)$, $\varepsilon$ — рівномірний на [-0.3,0.3],

$$g(x) = \begin{cases} x^2 - 1 & \text{при } x < 0 \\ x & \text{при } x \geq 0 \end{cases}$$

(6) $X \sim N(0, 2)$, $\varepsilon$ — рівномірний на [-0.3,0.3],

$$g(x) = \begin{cases} -x^2 + 1 & \text{при } x < 0 \\ x^2 - 1 & \text{при } x \geq 0 \end{cases}$$

(7) $X$ рівномірно розподілений на [0,3], $\varepsilon \sim N(0, 0.3)$,

$$g(x) = 3x(9 - 9x + 2x^2)$$

(8) $X$ рівномірно розподілений на [0,3], $\varepsilon$ рівномірно розподілений на [-1,1],

$$g(x) = -\frac{x}{7}(-242 + 805x - 742x^2 + 200x^3)$$

(9) $X \sim N(0, 1)$, $\varepsilon \sim N(0.5)$,

$$g(x) = 2|x|$$

(10) $X \sim N(0, 2)$, $\varepsilon$ — рівномірний на [-0.3,0.3],

$$g(x) = \begin{cases} x^2 - 1 & \text{при } x < 0 \\ x & \text{при } x \geq 0 \end{cases}$$

**Завдання 2.**

Реалізуйте алгоритм підбору параметра згладжування для оцінок Надарая — Ватсона та оцінки методу ковзаючих медіан з використанням техніки крос-валідації.

Для даних із завдання 1, проведіть оцінювання, використовуючи крос-валідацію для вибору параметра згладжування. Нарисуйте графіки отриманих оцінок, зробіть висновки про доцільність цього підходу та порівняйте точність оцінок Надарая — Ватсона і ковзаючої медіани.

# Розділ 7

# Непараметричні статистичні тести

## 7.1 Загальна теорія перевірки статистичних гіпотез

Перевірка статистичних гіпотез — один з основних розділів математичної статистики. Ми вже трохи обговорювали питання такої перевірки у прикладі 1.1.2. Але там це робилося візуально, за поведінкою довірчих інтервалів. У цьому розділі розглядаються спеціальні процедури перевірки непараметричних статистичних гіпотез.

Спочатку нагадаємо основні поняття та означення. Нехай спостережувані дані $\mathbf{X}$ являють собою випадковий елемент деякого вимірного простору даних $\mathcal{X}$ з розподілом $\mathsf{P}\{\mathbf{X} \in A\}$, який залежить від невідомого параметра:

$$\mathsf{P}\{\mathbf{X} \in A\} = \mathsf{P}_\theta\{A\} \text{ для всіх вимірих множин } A \subseteq \mathcal{X},$$

де $\vartheta \in \Theta$ — невідомий параметр, $\Theta$ — множина всіх можливих значень невідомого параметра (параметрична множина).

Статистичною гіпотезою називають припущення про те, що невідомий параметр $\vartheta$ належить деякій підмножині $\Theta_i \subset \Theta$. У класичній задачі перевірки гіпотез розглядають дві гіпотези $H_0 : \ \Theta_0$ і $H_1 : \ \Theta_0$, такі, що $\Theta_0 \cup \Theta_1 = \Theta$ і $\Theta_0 \cap \Theta_1 = \emptyset$. (Тобто гіпотези не можуть виконуватись одночасно і одна з них обов'язково виконується). Задача перевірки гіпотез

полягає в тому, щоб за даними $\mathbf{X}$ встановити, яка з гіпотез виконана — $H_0$, чи $H_1$.

Якщо множина можливих значень невідомого параметра $\Theta_i$ може бути інтерпретована як підмножина скінченновимірного векторного простору $\mathbb{R}^d$, то таку гіпотезу називають параметричною. Якщо це неможливо (наприклад, невідомий параметр — це функція розподілу, яка не задається скінченною кількістю числових параметрів) то гіпотезу називають непараметричною. Наприклад, розглянемо задачу перевірки гіпотези $H_0$ : розподіл кратної вибірки є нормальним, проти альтернативи $H_1$ : невідома функція розподілу вибірки може бути будь-якою крім нормальної. У цій задачі всі розподіли, що відповідають $H_0$, можна однозначно задати двома параметрами — математичним сподіванням та дисперсією. Це — параметрична гіпотеза. А для того, щоб задати $H_1$ треба вибрати практично довільну функцію розподілу $F$. Тому $H_1$ — непараметрична гіпотеза. Задача перевірки гіпотез вважається непараметричною, якщо хоча б одна з гіпотез $H_0$ або $H_1$ є непараметричною.

Для перевірки гіпотез використовують статистичні тести[1] — правила, які кожному можливому значенню даних $\mathbf{X}$ ставлять у відповідність гіпотезу, яку слід прийняти, якщо дані приймають це значення. Таким чином, з математичної точки зору, статистичний тест можна розглядати як функцію $\pi : \mathcal{X} \to \{0, 1\}$, котра значенням $x \in \mathcal{X}$ ставить у відповідність номер гіпотези, яку тест приймає, якщо $\mathbf{X} = x$.

Тести, призначені для перевірки гіпотез у непараметричних задачах, називають **непараметричними тестами**.

При класичному підході Неймана — Пірсона до задачі перевірки гіпотез, гіпотези $H_0$ і $H_1$ вважаються нерівноправними. Гіпотеза $H_0$ — основна (її називають "нульовою"). Цієї гіпотези потрібно дотримуватись доти, доки дані не переконають дослідника, що вона є хибною. Гіпотеза $H_1$ — альтернативна (її називають "альтернативою"), її приймають лише тоді, коли дані переконливо свідчать на її користь. Відповідно, для даного тесту $\pi$ можливі помилки двох родів (типів). Помилка першого роду полягає в тому, що виконана основна гіпотеза, а тест приймає альтернативу. Ймовірність цієї помилки позначають

$$\alpha_\theta(\pi) = \mathsf{P}_\theta\{\pi(\mathbf{X}) = 1\},\ \theta \in \Theta_0.$$

[1] Англійською мовою — test. В україномовній літературі зустрічається також назва "критерій".

Помилка другого роду виникає, якщо виконана альтернатива, а тест приймає основну гіпотезу. Її ймовірність:

$$\beta_\theta(\pi) = \mathsf{P}_\theta\{\pi(\mathbf{X}) = 0\},\ \theta \in \Theta_1.$$

Для того, щоб гарантувати малу ймовірність відхилення основної гіпотези, задають деяке мале значення $\alpha_0$ — стандартний рівень значущості, і розглядають лише тести $\pi$, у яких ймовірність помилки першого роду не перевищує $\alpha_0$ для всіх значень невідомого параметра, що відповідають альтернативі:

$$\sup_{\theta\in\Theta_0} \alpha_\theta(\pi) \le \alpha_0.$$

Такі тести називають тестами рівня $\alpha_0$. Ідея підходу Неймана — Пірсона полягає в тому, щоб обмежитись розглядом лише тестів певного заданого рівня $\alpha_0$ і шукати серед них той, у якого буде найменшою ймовірність помилки другого роду $\beta_\theta(\pi)$. Оскільки $\beta_\theta(\pi)$ є функцією від невідомого параметра $\vartheta \in \Theta_1$, такий тест існує далеко не завжди: при одних значеннях $\vartheta$ найменшою може бути ймовірність помилки одного тесту, при інших — іншого. Але, в усякому випадку, намагаються працювати з такими тестами, які забезпечують заданий рівень значущості і мають достатньо малу ймовірність помилки другого роду.

Досить часто статистичні тести мають вигляд

$$\pi(\mathbf{X}) = \mathbb{1}\{T(\mathbf{X}) > C\}, \tag{7.1}$$

де $T : \mathcal{X} \to \mathbb{R}$ — деяка вимірна функція від спостережуваних даних, яку називають статистикою тесту, а $C \in \mathbb{R}$ — число, що зветься порогом (або критичним значенням) тесту. (Також тест може мати форму $\pi(\mathbf{X}) = \mathbb{1}\{T(\mathbf{X}) < C\}$).

Для того, щоб забезпечити заданий стандартний рівень значущості $\alpha_0$, поріг тесту, заданого (7.1), обирають так, щоб

$$\alpha_\vartheta(\pi(\mathbf{X})) = \mathsf{P}_\vartheta\{T(\mathbf{X}) > C\} \le \alpha_0 \text{ для всіх } \vartheta \in \Theta_0. \tag{7.2}$$

Чим більшим обрати $C$, тим більшою буде ймовірність помилки другого роду — $\mathsf{P}_\vartheta\{T(\mathbf{X}) \le C\}$ для будь-якого $\vartheta \in \Theta_1$. Тому природно обирати на роль порогу найменше з тих $C$, при яких виконується (7.2).

Особливо просто вибирати поріг тесту, якщо розподіл статистики $T(\mathbf{X})$ при виконанні основної гіпотези не залежить від невідомого параметра $\vartheta$:

$$\mathsf{P}_\vartheta\{T(\mathbf{X}) < x\} = G(x) \text{ для всіх } x \in \mathbb{R}, \tag{7.3}$$

де $G$ — деяка неперервна функція розподілу. (Такі тести називають **тестами, незалежними від розподілу**). Тоді для

$$C = C_{\alpha_0} = Q^G(1 - \alpha_0), \tag{7.4}$$

отримуємо

$$\alpha_\vartheta(\pi(\mathbf{X})) = \mathsf{P}_\vartheta\{T(\mathbf{X}) > C\} = 1 - G(C_{\alpha_0}) = \alpha_0,$$

тобто $C_{\alpha_0}$, задане (7.4), є порогом, що забезпечує заданий рівень значущості $\alpha_0$ і дає найменшу ймовірність помилки другого роду для тестів, які використовують статистику $T(\mathbf{X})$.

Сучасні комп'ютерні програми як результат перевірки статистичної гіпотези зазвичай крім статистики використаного тесту $T(\mathbf{X})$ повідомляють досягнутий рівень значущості цього тесту. Нагадаємо, що це таке.

Насправді один і той самий тест можна реалізувати, використовуючи різні статистики і різні відповідні їм пороги. Дійсно, нехай у нас є тест (7.2). Якщо $h(x)$ — неперервна, строго зростаюча функція, то тест

$$\pi'(\mathbf{X}) = \mathsf{P}\{h(T(\mathbf{X})) > h(C)\},$$

буде приймати $H_0$ тоді і тільки тоді, коли її приймає $\pi(\mathbf{X})$. Тобто $\pi'$ і $\pi$ — це дві різні форми запису того самого тесту. Статистика $p = p(\mathbf{X})$, при використанні якої тест набуває вигляду

$$\pi(\mathbf{X}) = \mathbb{1}\{p(\mathbf{X}) < \alpha_0\},$$

де $\alpha_0$ — стандартний рівень значущості, називається **досягнутим рівнем значущості** (англ. p-level, significance). При виконанні (7.3), досягнутий рівень значущості для тесту (7.1) можна визначити як

$$p(\mathbf{X}) = 1 - G(T(\mathbf{X})).$$

Використання досягнутих рівнів значущості $p(\mathbf{X})$ для публікації результатів перевірки гіпотез зручно тим, що, знаючи $p(\mathbf{X})$, читач може сам, користуючись власним стандартним рівнем значущості $\alpha$, перевірити яку гіпотезу слід приймати при цьому $\alpha$ — основну, чи альтернативну.

## 7.2 Тести однорідності двох вибірок

### 7.2.1 Задача перевірки однорідності

У прикладній статистиці часто виникає потреба порівняти розподіл даних з двох різних популяцій за вибірками з кожної з них. Ми вже зустрічались з такою задачею у прикладі 1.1.2, де результати вакцинування порівнювались із контрольною вибіркою (для якої вакцина була замінена на плацебо). У цьому підрозділі ми розглянемо спеціальні непараметричні тести, які дозволяють проводити перевірку наявності чи відмінності розподілів за двома вибірками. Такі тести мають назву двовибіркових. Спочатку сформулюємо ймовірнісну модель даних.

Нехай спостерігаються дані, розбиті на дві вибірки: перша — $\mathbf{X} = (X_1, \ldots, X_m)$ і друга — $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_n)$. Спостереження в обох вибірках є незалежними в сукупності випадковими величинами, причому функція розподілу спостережень з першої вибірки $F$, а з другої — $G$. Ці функції невідомі. Можливо, що $F \equiv G$, тоді вибірки називають однорідними. Можливо, що розподіли вибірок різні, тоді їх називають неоднорідними. Задача полягає в тому, щоб побудувати тест для перевірки однорідності.

Основною ми будемо вважати гіпотезу про однорідність вибірок:

$$H_0 : \ F(x) = G(x) \text{ для всіх } x \in \mathbb{R}.$$

Щодо альтернативи, то тут можуть бути різні варіанти. Найбільш загальною альтернативою є припущення про те, що функції розподілу $F$ і $G$ відрізняються хоча б в одній точці:

$$H_1 : \ F(x) \neq G(x) \text{ для деякого } x \in \mathbb{R}.$$

Інколи можна припустити, що функції $F$ і $G$ відрізняються лише зсувом: $G(x) = F(x - \Delta)$ для деякого $\Delta \in \mathbb{R}$ і всіх $x \in \mathbb{R}$. Це припущення еквівалентне тому, що розподіл випадкових величин $\tilde{X}_j - \Delta$ такий самий, як величин $X_j$. Величина $\Delta$ у цьому випадку називається зсувом розподілу другої вибірки по відношенню до першої.

Якщо ми обмежуємося вибірками, котрі можуть відрізнятися лише зсувом $\Delta$, то гіпотеза про їхню однорідність зводиться до

$$H_0^{\Delta} : \Delta = 0,$$

а її альтернатива —

$$H_1^{\pm} : \Delta \neq 0.$$

Інколи дослідників цікавить лише можливе відхилення зсуву від 0 в той чи інший бік:

$$H_1^+ : \Delta > 0,$$

або

$$H_1^- : \Delta < 0.$$

Гіпотези $H_1^+$ і $H_1^-$ називають односторонніми, а гіпотезу $H_1^\pm$ — двосторонньою.

Відмітимо, що всі розглянуті тут гіпотези є непараметричними. Дійсно, щоб однозначно задати розподіл даних при виконанні, наприклад, $H_0^\Delta$ потрібно вказати функцію $F$, яку ми вважаємо цілком невідомою. Аналогічно і з іншими гіпотезами — розподіл даних при їхньому виконанні неможливо описати, задавши скінченну кількість числових параметрів. Відповідно і тести, які ми будемо розглядати, є непараметричними.

### 7.2.2 Медіанний тест однорідності

Почнемо з порівняно простого тесту, який називають **медіанним тестом** (тест Брауна — Муда). Спочатку ми визначимо його статистику. Об'єднаємо дві наші вибірки в одну, поклавши $X_{j+m} = \tilde{X}_j$ для $j = 1, \ldots, n$. Отримуємо набір незалежних випадкових величин з обох вибірок $\vec{\mathbf{X}} = (X_1, \ldots, X_N)$, де $N = m + n$, причому для $j = 1, \ldots, m$ розподіл $X_j$ є $F$, а для $j = m + 1, \ldots, N$ — $G$.

Надалі ми будемо припускати, що розподіли $F$ і $G$ є абсолютно неперервними. Зокрема, це означає, що всі спостереження $X_j$ можна вважати різними.

Позначимо через $\hat{\mu}$ вибіркову медіану об'єднаної вибірки $X_1,\ldots,X_N$. Позначимо через $S$ кількість тих елементів першої вибірки $X_1, \ldots, X_m$, які є меншими ніж $\hat{\mu}$. Якщо $e_j = \mathbb{1}\{X_j < \hat{\mu}\}$, то

$$S = S_m = \sum_{j=1}^{m} e_j. \tag{7.5}$$

Ця величина $S$ і буде основною статистикою, за допомогою якої ми будуватимем медіанний тест. Для того, щоб побудувати тест за загальною схемою, описаною у п. 7.1, потрібно знати розподіл статистики $S$ при виконанні основної гіпотези $H_0$ і розібратись, якою буде поведінка $S$ якщо виконана альтернатива.

Описати розподіл $S$ у випадку однорідності двох вибірок дуже просто. Позначимо $t$ — кількість $e_j$, $j = 1, \ldots, N$ які дорівнюють 1. Якщо $N$ — парне число, то $t = N/2$ (під медіаною і над медіаною лежить однакова кількість спостережень). Якщо $N$ — непарне, $t = (N-1)/2$ (Рівно для одного $j$ $X_j = \hat{\mu}$ і, для цього $j$, $e_j = 0$, а всі інші спостереження діляться медіаною навпіл).

**Теорема 7.2.1.** *Якщо виконана гіпотеза про однорідність обох вибірок, то*

$$p_k = \mathsf{P}\{S = k\} = \frac{C_m^k C_n^{t-k}}{C_N^t},$$

*якщо* $k = 0, 1, \ldots \min(t, m)$ *і* $k \geq t - n$. *Для всіх інших* $k$, $p_k = 0$.

*Доведення.* Розглянемо набір випадкових величин $\mathbf{e} = (e_1, e_2, \ldots, e_N)$. У цьому наборі рівно $t$ елементів приймають значення 1, а $N-t$ елементів — значення 0. При цьому, оскільки вибірка однорідна, всі такі набори значень є однаково можливими. Кількість таких наборів — $C_N^t$. Отже, ймовірність того, що $\mathbf{e}$ дорівнює даному допустимому набору нулів та одиниць є $p = 1/C_N^t$

Подія $S = k$ виконується тоді, коли у наборі $\mathbf{e}$ на перших $m$ місцях розташовано рівно $k$ одиниць, а на наступних $n$ місцях — $t - m$ одиниць. Всі інші місця займають нулі. Кількість таких можливих наборів значень $\mathbf{e}$ дорівнює $C_m^k C_n^{t-k}$. Помноживши це число на $p$ — ймовірність будь-якого конкретного набору, отримаємо $p_k$. □

Таким чином, при виконанні $H_0$, $S$ має гіпергеометричний розподіл, і, отже, тест на основі цієї статистики буде незалежним від розподілу. За відомими формулами для математичного сподівання та дисперсії гіпергеометричного розподілу (див. табл. 1.1 в [26]), отримуємо

$$\mathsf{E}\, S = \frac{mt}{N},\ \mathrm{D}\, S = \frac{mnt(N-t)}{N^2(N-1)}. \tag{7.6}$$

Отже, при виконанні основної гіпотези, статистика $S$ коливається навколо середнього $(mt)/N = m/2$ при парних $N$ і $m(N-1)/(2N) \approx m/2$ — при непарних. Причому розподіл цієї статистики не залежить від невідомої спільної функції розподілу обох вибірок.

А якою буде поведінка $S$, якщо виконана альтернатива? Точна відповідь на це залежить від невідомих параметрів — функцій $F$ і $G$. Але можна зробити певні евристичні міркування. Розглянемо, для прикладу,

односторонню альтернативу $H_1^+$ : розподіл другої вибірки зсунутий відносно розподілу першої із додатнім зсувом. Тоді значення спостережень у другій вибірці будуть частіше лежати вище значень з першої. Відповідно, спільна медіана $\hat{\mu}$ обох вибірок буде більшою, ніж медіана першої вибірки і ймовірність того, що спостереження з першої вибірки лежить нижче $\hat{\mu}$ буде більше $1/2$. Тому значення $S$ будуть коливатись вище ніж $m/2$. Тобто, в цьому випадку, значення $S$ при альтернативі будуть "в середньому" більші, ніж при виконанні основної гіпотези. Тому тест на основі $S$ має приймати $H_0$ при малих значеннях $S$ і відхиляти — при великих. Такий тест матиме вигляд:

$$\pi^+(S) = \mathbb{1}\{S > C\},$$

де $C$ — поріг, який потрібно обрати за заданим стандартним рівнем значущості.

Якщо розглядається альтернатива $H_1^-$, про наявність від'ємного зсуву, то на користь такої альтернативи свідчать малі значення статистики $S$ і тест матиме вигляд

$$\pi^-(S) = \mathbb{1}\{S < C\}.$$

Нарешті, якщо потрібно перевірити двосторонню гіпотезу $H_1^\pm$, то на користь альтернативи свідчитимуть відхилення $S$ від $Z = mt/N$ у будь-яку сторону. Відповідний тест буде мати вигляд

$$\pi^\pm(S) = \mathbb{1}\{|S - Z| > C\}.$$

Для того, щоб обрати поріг $C$, треба вміти визначати ймовірність помилки першого роду даного тесту $\alpha(\pi)$. (Оскільки розподіл $S$ при основній гіпотезі не залежить від невідомих параметрів, то і $\alpha(\pi)$ від них не залежить). Це неважко зробити, використовуючи ймовірності $p_k$, визначені у теоремі 7.2.1. Дійсно,

$$\alpha(\pi^+) = \mathsf{P}_{H_0}\{S > C\} = \sum_{k>C} p_k.$$

Отже, відповідно до загальної теорії з п. 7.1, на роль порогу $C_{\alpha_0}$ треба вибрати найменше ціле $C$, при якому $\sum_{k>C} p_k \le \alpha_0$, де $\alpha_0$ — заданий стандартний рівень значущості.

Аналогічно, для тесту $\pi^{\pm}(S)$,

$$\alpha(\pi^{\pm}) = \mathsf{P}_{H_0}\{|S - Z| > C\} = \sum_{k:\ |k-Z|>C} p_k,$$

і правильним порогом $C_{\alpha_0}$ буде найменше ціле $C$, для якого виконується $\sum_{k:\ |k-Z|>C} p_k < \alpha_0$.

Остаточно, медіанний тест для перевірки гіпотези $H_0$ проти $H_1^{\pm}$ має вигляд:

Якщо $|S - Z| \le C_{\alpha_0}$ — прийняти гіпотезу про однорідність вибірок,

якщо $|S - Z| > C_{\alpha_0}$ — прийняти гіпотезу про наявність зсуву.

Зрозуміло, що використовувати безпосередній підрахунок сум $p_k$ для знаходження порогів тестів доцільно лише при малих обсягах вибірок. Коли вибірки великі, такі підрахунки стають занадто обтяжливими. Але в цьому випадку можна скористатись асимптотичним наближенням на основі центральної граничної теореми.

Статистика $S$ є сумою випадкових величин $e_j$. Якби $e_j$ були незалежними випадковими величинами, для дослідження поведінки $S$ при великих обсягах вибірок можна було б скористатись звичайною центральною граничною теоремо. Але $e_j$ не є незалежними, тому що всі вони залежать від спільної медіани $\hat{\mu}$. Однак розподіл вектора $\mathbf{e}$ при виконанні $H_0$ є перестановочним: він не змінюється при будь-якій не випадковій перестановці елементів $\mathbf{e}$. Завдяки цьому можна застосувати спеціальну центральну граничну теорему для перестановочних випадкових векторів [10] і отримати наступне твердження.

**Теорема 7.2.2.** *Нехай вибірки однорідні, $m, n \to \infty$, так, що існує границя*

$$v = \lim m/(m+n), \ \textit{причому}\ 0 < v < 1. \tag{7.7}$$

*Тоді*

$$\frac{S - \mathsf{E}\,S}{\sqrt{\mathrm{D}\,S}} \xrightarrow{W} N(0,1).$$

Виходячи з цієї теореми, з урахуванням (7.6), для вибірок великого обсягу отримуємо

$$\alpha(\pi^{\pm}) = \mathsf{P}_{H_0}\{|S - Z| > C\} = \mathsf{P}_{H_0}\left\{\frac{|S - \mathsf{E}\,S|}{\sqrt{\mathrm{D}\,S}} > \frac{C}{\sqrt{\mathrm{D}\,S}}\right\}$$

$$\approx 2\left(1-\Phi\left(\frac{C}{\sqrt{\mathrm{D}\,S}}\right)\right).$$

Отже поріг $C = C_{\alpha_0}$, який відповідає умові $\alpha(\pi^{\pm}) = \alpha_0$, можна наближено обчислити як розв'язок рівняння

$$2\left(1-\Phi\left(\frac{C}{\sqrt{\mathrm{D}\,S}}\right)\right) = \alpha_0.$$

Знову враховуючи (7.6), отримуємо

$$C_{\alpha_0} \approx \lambda_{\alpha_0/2}\sqrt{\frac{mnt(N-t)}{N^2(N-1)}},$$

де, як і раніше $\lambda_{\alpha_0/2} = Q^{N(0,1)}(1-\alpha_0/2)$.

Цю формулу можна трохи спростити, враховуючи, що $t = N/2$, при парних $N$ і $t = (N-1)/2 \approx N/2$ при непарних великих $N$. Отже

$$C_{\alpha_0} \approx \frac{\lambda_{\alpha_0/2}}{2}\sqrt{\frac{mn}{m+n}}. \tag{7.8}$$

Аналогічно можна визначити порогові значення для односторонніх версій медіанного тесту.

Задавши статистику та поріг ми однозначно визначаємо тест. При цьому, за побудовою, ймовірність помилки першого роду у нього буде дорівнювати заданому стандартному рівню значущості. А якою буде ймовірність помилки другого роду? Відповісти на це питання значно важче, тому, що відповідь залежить від невідомих параметрів задачі. Для нашої двовибіркової моделі це функції розподілу першої ($F$) та другої ($G$) вибірок. Розподіл статистики $S$ при $H_0$ не залежав від цих параметрів. Тому і ймовірність помилки першого роду виходить сталою. При альтеративі таке неможливо: чим менша відмінність між $F$ і $G$, тим важче розпізнати неоднорідність і, відповідно, тим більшою буде ймовірність помилки другого роду $\beta_{F,G}(\pi)$. Неможливо вказати одну просту формулу для підрахунку цієї ймовірності, придатну в усіх випадках. Тому тут ми обмежимось грубою прикидкою поведінки $\beta_{F,G}(\pi)$ при необмеженому зростанні обсягу обох вибірок.

Мінімальною необхідною умовою придатності статистичного тесту $\pi$ вважається консистентність. Тест $\pi$ називають **консистентним**, якщо, при прямуванні обсягу даних до нескінченності, ймовірність помилки

другого роду тесту $\pi$ прямує до 0 для всіх значень невідомого параметра, які відповідають альтернативі.

З'ясуємо, за яких умов тест $\pi^{\pm}$ буде консистентним. За теоремою 1.1.1, якщо $n \to \infty$, $m \to \infty$, то

$$\sup_{x\in\mathbb{R}} |\hat{F}_m(x) - F(x)| \to 0 \text{ і } \sup_{x\in\mathbb{R}} |\hat{G}_m(x) - G(x)|, \tag{7.9}$$

де $\hat{F}_m$ і $\hat{G}_n$ — емпіричні функції розподілу для першої і другої вибірок відповідно. Позначимо $\hat{H}_{m,n}$ емпіричну ф.р. об'єднання першої і другої вибірки. Припустимо, що умова (7.7) виконана. Тоді з (7.9) легко отримати

$$\sup_{x\in\mathbb{R}} |\hat{H}_{m,n}(x) - H(x)| \to 0 \text{ при } n \to \infty, m \to \infty, \text{ м.н.}, \tag{7.10}$$

де $H(x) = vF(x) + (1-v)G(x)$ — функція розподілу, котру можна розглядати як суміш розподілів першої і другої вибірки.

Міркуючи як у 2.2.1, отримуємо, що вибіркова медіана об'єднаної вибірки $\mathsf{med}(\breve{\mathbf{X}}) \to \mathsf{med}(H)$. Тепер статистику медіанного тесту $S_m$, визначену (7.5), можна задати як

$$S_m = mF_m(\mathsf{med}(\breve{\mathbf{X}})),$$

і, враховуючи, (7.9), отримати

$$\frac{1}{m}S_m \to F(\mathsf{med}(H)) \text{ м.н.}$$

при виконанні (7.7). Таким чином, для статистики двостороннього медіанного тесту $\pi^{\pm}$ отримуємо (м.н при $m \to \infty$)

$$|S - Z| \sim m|F(\mathsf{med}(H)) - 1/2| \sim |F(\mathsf{med}(H)) - 1/2|vN.$$

А для порогу $C_{\alpha_0}$ за (7.8), маємо

$$C_{\alpha_0} \sim \frac{\lambda_{\alpha_0}}{2}\sqrt{\frac{v(1-v)N^2}{N}} = \frac{\lambda_{\alpha_0}}{2}\sqrt{v(1-v)N}.$$

Звідси випливає, що, за умови

$$F(\mathsf{med}(H)) \neq 1/2, \tag{7.11}$$

то при достатньо великих $N$, $|S-Z|$ стає більшим ніж $C_{\alpha_0}$ з ймовірністю, що прямує до 1. Тому

$$\beta(\pi^{\pm}) = \mathsf{P}_{H_1}\{|S-Z| < C_{\alpha_0}\} \to 0 \text{ при } N \to \infty,$$

при виконанні (7.11).

легко бачити, що умова (7.11) еквівалентна умові $\mathsf{med}(F) \neq \mathsf{med}(G)$. Отже, двосторонній медіанний тест однорідності є консистентним, якщо $\mathsf{med}(F) \not/ \mathsf{med}(G)$. Так само можна отримати консистентність одностороннього тесту $\pi^+$ за умови, що $\mathsf{med}(F) < \mathsf{med}(G)$ і тесту $\pi^-$ — за умови $\mathsf{med}(F) > \mathsf{med}(G)$.

Неформально можна сказати, що медіанні тести помічають відмінності медіани у двох вибірках і не помічають інших відмінностей, наприклад — змін розкиду. Зокрема, ці тести будуть консистентними, якщо при альтернативі розподіли вибірки відрізняються лише зсувом $\Delta$, адже у цьому випадку $\mathsf{med}(G) = \mathsf{med}(F) + \Delta$.

### 7.2.3 КС-тест

А чи можна побудувати тест однорідності двох вибірок так, щоб він помічав будь яку відмінність розподілів, тобто був би консистентним для будь-якої альтернативи вигляду $\exists x \in \mathbb{R}: \ F(x) \neq G(x)$? Так, щоб отримати статистику такого тесту, можна розглянути різницю емпіричних функцій розподілу двох вибірок $\hat{F}_m(x) - \hat{G}_n(x)$ і взяти яку-небудь функціональну норму від неї. Існує багато версій таких тестів, наприклад, використання норми $L_2$, приводить до тесту фон-Мізеса, а рівномірної норми — до КС-тесту (тест Смірнова, або Колмогорова — Смірнова).

Як приклад розглянемо двовибірковий КС-тест. Статистика цього тесту визначається як

$$D_{m,n} = \sup_{x\in\mathbb{R}} |\hat{F}_m(x) - \hat{G}_n(x)|,$$

де, як і раніше,

$$\hat{F}_m(x) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{X_j < x\}, \ \hat{G}_m(x) = \frac{1}{n}\sum_{j=1}^{m} \mathbb{1}\{\tilde{X}_j < x\}$$

— емпіричні функції розподілу для вибірок $\mathbf{X}$ та $\tilde{\mathbf{X}}$ відповідно.

Відмітимо, що $\hat{F}_m(x) - \hat{G}_n(x)$ є константою на інтервалах між послідовними значеннями порядкових статистик об'єднаного набору даних $\breve{\mathbf{X}}$. Тому замість супремуму по всіх $x \in \mathbb{R}$ досить використовувати максимум по значеннях з цього набору порядкових статистик.

Зрозуміло, що, при виконанні $H_0$, $\hat{F}_m(x) \to F(x)$ і $\hat{G}_n(x) \to F(x)$ рівномірно по $x$ м.н. за теоремою Глівенко — Кантеллі. Тому $D_{m,n}$ при великих обсягах вибірки повинно бути малим (прямувати до 0). Якщо виконана загальна альтернатива $H_1: \ F \not\equiv G$,

$$D_{m,n} \to \sup_{x\in\mathbb{R}} |F(x) - G(x)| \neq 0.$$

Отже тест повинен мати вигляд $\pi(\mathbf{X}) = \mathbb{1}\{D_{m,n} > C\}$, де поріг $C$ обирається за заданим рівнем значущості.

Для того, щоб знайти поріг тесту, потрібно визначити розподіл $D_{m,n}$ при виконанні основної гіпотези $H_0: \ F \equiv G$. Покажемо, що цей розподіл за досить широких умов не залежить від $F$ (тобто наш тест буде незалежним від розподілу).

Позначимо $a = \inf\{x: \ F(x) > 0\}$, $b = \sup\{x: \ F(x) < 1\}$ (можливо $a = -\infty$ і/або $b = +\infty$). Інтервал $(a, b)$ називають носієм функції розподілу $F$. Припустимо, що $F$ є неперервною і строго монотонною на $(a, b)$. Тоді у неї є обернена на цьому інтервалі — $F^{-1}$, тобто така функція, що $F^{-1}(F(x)) = x$ для всіх $x \in (a, b)$ і $F(F^{-1}(z)) = z$ для всіх $z \in (0, 1)$. Легко бачити, що, якщо випадкова велична $\xi$ має ф.р. $F$, то для випадкової величини $\eta = F(\xi)$ виконано

$$\mathsf{P}\{\eta < x\} = \mathsf{P}\{F(\xi) < x\} = \mathsf{P}\{F^{-1}(F(\xi)) < F^{-1}(x)\} = \mathsf{P}\{x < F^{-1}(x)\}$$

$$= F(F^{-1}(x)) = x,$$

для всіх $x \in (0, 1)$, тобто випадкова величина $\eta$ має рівномірний розподіл.

Позначимо $\eta_j = F(X_j)$, $j = 1, \ldots, m$ та $\tilde{\eta}_j = F(\tilde{X}_j)$, $j = 1, \ldots, n$. Тоді

$$D_{m,n} = \sup_{x\in\mathbb{R}} \left| \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{X_j < x\} - \frac{1}{n} \sum_{j=1}^{m} \mathbb{1}\{\tilde{X}_j < x\} \right|$$

$$= \sup_{z\in(0,1)} \left| \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{X_j < F^{-1}(z)\} - \frac{1}{n} \sum_{j=1}^{m} \mathbb{1}\{\tilde{X}_j < F^{-1}(z)\} \right|$$

$$= \sup_{z\in(0,1)} \left| \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{F(X_j) < z\} - \frac{1}{n}\sum_{j=1}^{m} \mathbb{1}\{F(\tilde{X}_j) < z\} \right|$$

$$= \sup_{z\in(0,1)} \left| \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{\eta_j < z\} - \frac{1}{n}\sum_{j=1}^{m} \mathbb{1}\{\tilde{\eta}_j < z\} \right|.$$

Але всі $\eta_j$, $\tilde{\eta}_j$ — незалежні, рівномірно розподілені на $[0,1]$ випадкові величини. Тому розподіл $D_{m,n}$ не залежить від $F$.

Використовуючи цю рівність можна підрахувати яким саме є розподіл $D_{m,n}$ при різних (невеликих) $m$ і $n$. Комбінаторна техніка таких розрахунків описана, наприклад, у підрозділі 6.3 [17]. Там само наведено і наступний асимптотичний результат.

**Теорема 7.2.3.** *Нехай вибірки однорідні, виконується умова (7.7) і функція розподілу спостережень є неперервної та строго монотонною на своєму носію. Тоді*

$$\lim_{m,n\to\infty} \mathsf{P}\left\{ \sqrt{\frac{mn}{n+m}} D_{m,n} < x \right\} = L(x),$$

*де*

$$L(x) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} \exp(-2i^2x^2).$$

Таким чином, для забезпечення заданого стандартного рівня значущості $\alpha_0$, при достатньо великих $m$ і $n$, поріг тесту можна обрати рівним

$$C_{\alpha_0} = L^{-1}(1-\alpha_0)\sqrt{\frac{n+m}{mn}}.$$

Осточно, КС-тест має вигляд:

Якщо $D_{m,n} \le C_{\alpha_0}$, прийняти гіпотезу про однорідність,

якщо $D_{m,n} > C_{\alpha_0}$, вважати вибірки неоднорідними.

Які альтернативи помічає цей тест? Нехай виконано (7.7). Якщо $F \not\equiv G$, то $D_{m,n} \to \sup_{m,n} |F(x) - G(x)| > 0$ м.н. А поріг $C_{\alpha_0}$ в цьому випадку прямує до 0. Тому ймовірність помилки другого роду КС-тесту $\mathsf{P}_{H_1}\{D_{m,n} < C_{\alpha_0}\} \to 0$. Отже КС-тест є консистентним проти всіх альтернатив.

## 7.3 Рангові тести.

### 7.3.1 Ранги та рангові статистики

Дуже багато популярних непараметричних тестів використовують статистики, які є функціями від рангів спостережень у вибірці. У цьому підрозділі ми спочатку введемо поняття рангу, а потім розглянемо деякі статистики і тести, які їх використовують.

Нехай вибірка складається з $n$ об'єктів і для кожного ($j$-того) об'єкта спостерігається значення числової змінної $X_j$, $j = 1, \dots n$. Спочатку припустимо, що всі значення $X_j$ у вибірці — різні. Тоді можна прийняти наступне означення рангу.

**Ранг** $R_j^X$ $j$-того об'єкта по відношенню до змінної $X$ це номер цього об'єкта у вибірці, впорядкованій по зростанню $X$.

Нехай, наприклад, спостерігаються такі значення змінної $X$ у вибірці з шести елементів:

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $X$ | 12 | -4.4 | 5.3 | 0 | -2 | 6 |

Переставимо об'єкти у вибірці так, щоб значення $X$ розташувались у порядку зростання. (Номери об'єктів переставляємо разом зі значеннями $X$):

| $j$ | 2 | 5 | 4 | 3 | 6 | 1 |
|---|---|---|---|---|---|---|
| $X$ | -4.4 | -2 | 0 | 5.3 | 6 | 12 |

Елемент з номером 5 опинився на другому місці, отже його ранг $R_5^X = 2$. Перший елемент опинився на шостому місці, тому $R_1^X = 6$. Аналогічно отримуємо всі інші значення рангів:

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $R_j^X$ | 6 | 1 | 4 | 3 | 2 | 5 |

Може так статися, що у вибірці присутні кілька елементів, які мають однакові значення змінної $X$. При перестановці у порядку зростання вони встануть на місця розташовані поруч, причому порядок їн на цих місцях може бути довільним. У такому випадку кажуть, що ці об'єкти мають зв'язані ранги (англ *tied ranks*). При стандартному способі обчислення рангів, ранг кожного такого об'єкта визначається як середнє значення тих рангів, які він міг би мати при всіх можливих перестановках в порядку зростання $X$.

Розглянемо, наприклад, таку вибірку.

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $X$ | -1 | 3 | -1 | 3 | -1 | 0 |

Переставляючи значення $X$ у порядку зростання, отримуємо таку послідовність:

-1, -1, -1, 0, 3, 3.

Елементи з номерами 1, 3, 5 у цій вибірці зв'язані (змінна $X$ у них всіх дорівнює $-1$ — це найменше значення у нашій вибірці). Тому при перестановці у порядку зростання кожен з них міг би опинитись на першому, другому або третьому місці. Середнє цих можливих рангів дорівнює двом. Тому всім цим об'єктам присвоюють ранг 2. Так само, зв'язаними є об'єкти номер два і чотири — кожен з них можна розташувати або на п'ятому, або на шостому місці. Відповідно їхній середній ранг $5.5 = (5+6)/2$. Нарешті, шостий елемент — не зв'язаний. Він однозначно знаходиться на четвертому місці, його ранг — 4:

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $R_j^X$ | 2 | 5.5 | 2 | 5.5 | 2 | 4 |

Для кожного об'єкта у вибірці може спостерігатись кілька змінних. У такому випадку його ранг за кожною змінною треба визначати окремо.

Відмітимо, що розглянуті у п. 7.2 статистики медіанного тесту $S$ і КС-тесту $D_{mn}$ є ранговими статистиками, тобто функціями від рангів у об'єднаній вибірці $\breve{\mathbf{X}} = (X_1, \ldots, X_N)$. Для статистики $S$ це очевидно. Дійсно, вона визначається як

$$S = \sum_{j=1}^{m} \mathbb{1}\{X_j < \mathsf{med}(\breve{\mathbf{X}})\} = \sum_{j=1}^{m} \mathbb{1}\{R_j^X < N/2\},$$

де $R_j^X$ — ранг $j$-того спостереження за $X$ у об'єднаній вибірці $\breve{\mathbf{X}}$.

Менш очевидно, що $D_{m,n}$ є ранговою статистикою. Але це стає зрозумілим, якщо помітити, що $D_{m,n}$ залежить лише від порядку об'єктів за зростанням $X$, а не від конкретних числових значень $X$ у вибірці.

В чому зручність рангових статистик для побудови непараметричних тестів? Значною мірою, вона полягає у наступній важливій властивості рангів.

Нехай $\mathbf{X} = (\xi_1, \ldots, \xi_n)$ — кратна вибірка, тобто набір незалежних, однаково розподілених випадкових величин, причому функція розподілу $F$ в.в. $\xi_j$ є неперервною. Тоді всі $\xi_j$ у вибірці з ймовірністю 1 є різними. Отже у наборі $R_j^\xi$ зв'язаних рангів не буде. Тому, при різних $j = 1, \ldots, n$, $R_j^\xi$ будуть приймати значення $1, 2, \ldots, n$ по одному разу кожне. Інакше

кажучи, набір $\mathcal{R} = (R_1^\xi, \ldots, R_n^\xi)$ є перестановкою $n$ чисел. Ця перестановка є випадковою, оскільки $\xi_j$ випадкові. Легко бачити, що всі такі перестановки є однаково можливими, тобто, якщо $r$ — довільна не випадкова перестановка $n$ чисел, то

$$\mathsf{P}\{\mathcal{R} = r\} = \frac{1}{n!}.$$

Це дає можливість підраховувати розподіл будь-якої рангової статистики як функції від $\mathcal{R}$, якщо ця статистика обчислюється за кратною вибіркою. У непараметричній статистиці часто основна гіпотеза відповідає припущенню про те, що дані утворюють кратну вибірку. Відповідно, ймовірність помилки першого роду та поріг тесту можна визначати виходячи саме з розподілу статистики на такій вибірці. Як це робиться ми бачили у п. 7.2 на прикладі медіанного тесту.

### 7.3.2 Тест Манна — Уітні — Вілкоксона

Повернемося до задачі перевірки однорідності двох вибірок, описаної у п. 7.2. Нехай, як і раніше, $\mathbf{X} = (X_1, \ldots, X_m)$ і $\mathbf{X}' = (X'_1, \ldots, X'_n)$ — дві незалежні між собою вибірки, що складаються з незалежних випадкових величин, причому у першій вибірці функція розподілу $F()$, а у другій $G = F(\cdot - \Delta)$, де $F$ — невідома неперервна функція розподілу, $\Delta \in \mathbb{R}$ — невідомий "зсув". Ми будемо перевіряти гіпотезу про однорідність цих вибірок — $H_0 :\ \Delta = 0$ проти загальної альтернативи $H_1^\pm : \Delta \neq 0$.

У 1945 році Ф. Вілкоксон запропонував застосувати для цього тест, що використовує наступну статистику[2]. Розглянемо об'єднану вибірку $\breve{\mathbf{X}} = (X_1, \ldots, X_m, X'_1, \ldots, X'_n)$. Нехай $R_j$ — ранг $j$-го об'єкта у цій об'єднаній вибірці за змінною $X$ $(X')$. Статистика Вілкоксона $W$ визначається як

$$W = \sum_{j=1}^{m} R_j,$$

тобто це сума всіх рангів, які у об'єднаній вибірці мають елементи з першої вибірки.

У об'єднаній вибірці ранги $R_j$ приймають всі значення $1, 2, \ldots, N = m+n$ по одному разу. Якщо виконана гіпотеза $H_0$, то всі можливі переста-

[2]Сама ця статистика розглядалася і до роботи Вілкоксона.

новки рангів однаково ймовірні. Тому в цьому випадку математичне сподівання одного рангу дорівнює середньому всіх рангів: $\mathsf{E}\, R_j = (N+1)/2$ і

$$Z = \mathsf{E}\, W = \frac{m(N+1)}{2} \qquad \text{(при виконанні } H_0\text{)}.$$

Якщо $\Delta > 0$, то елементи другої вибірки частіше будуть більшими за елементи першої і можна сподіватись, що $W$ буде менше ніж $Z$. При $\Delta < 0$, навпаки, статистика Вілкоксона буде відхилятись від $Z$ вгору. Таким чином, на роль статистики для перевірки $H_0$ проти $H_1^{\pm}$ природно взяти $|W - Z|$. Поріг тесту $C_{\alpha_0}$, що відповідає рівню значущості $\alpha_0$, потрібно знайти з умови:

$$C_{\alpha_0} \text{— найменше } C, \text{ для якого } \mathsf{P}_{H_0}\{|W - Z| > C\} \le \alpha_0.$$

Тест приймає $H_0$, якщо $|W - Z| \le C_{\alpha_0}$ і відхиляє у іншому випадку.

Як підрахувати $\mathsf{P}_{H_0}\{|W - Z| > C\}$? Переберемо всі можливі перестановки $\mathbf{r} = (r_1, \ldots, r_N)$ чисел $(1, 2, \ldots, N)$. Для кожного $\mathbf{r}$ обчислимо $W(\mathbf{r}) = \sum_{j=1}^{m} r_j$. Підрахуємо кількість тих $\mathbf{r}$, для яких $|W(\mathbf{r}) - Z| > C$. Нехай вона дорівнює $K_C$. Тоді

$$\mathsf{P}_{H_0}\{|W - Z| > C\} = \frac{K_C}{N!}.$$

Дійсно, для набору рангів $\mathcal{R} = (R_1, \ldots, R_N)$, при виконанні $H_0$, $\mathsf{P}\{\mathcal{R} = \mathbf{r}\} = 1/N!$, оскільки всі такі набори однаково ймовірні. Отже

$$\mathsf{P}_{H_0}\{|W - Z| > C\} = \sum_{\mathbf{r}:\ |W(\mathbf{r}) - Z| > C} \mathsf{P}\{\mathcal{R} = r\} = \frac{K_C}{N!}.$$

Зрозуміло, що перебирати всі $N!$ перестановок можна лише коли $N$ невелике число. Для великих обсягів вибірки використовують асимптотичні формули. Можна показати, що

$$D = \mathrm{D}\, W = \frac{1}{12} mn(N+1) \qquad \text{(при виконанні } H_0\text{)},$$

і при великих $m$ та $n$, розподіл $(W - Z)/\sqrt{D}$ є приблизно стандартним нормальним. Звідси, аналогічно тому, як це зроблено у п. 7.2.2, можна отримати формулу для порогу тесту.

У 1947 році Г. Манн та Д.Р. Вітні запропонували тест однорідності, який використовує статистику

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}\{X_i < Y_j\}.$$

Елементарними перетвореннями можна отримати

$$U = nm + \frac{n(n+1)}{2} - W,$$

тому тест на основі $U$-статистики Манна — Уітні є, фактично, варіантом тесту Вілкоксона, який дає ті самі результати, що і останній. Тому інколи ці тести об'єднують під назвою тест Манна — Уітні — Вілкоксона.

Який тест для перевірки однорідності є більш потужним (краще помічає альтернативу) — медіанний тест, чи тест Вілкоксона? У деяких книжках можна прочитати, що медіанний тест використовує менше інформації тест Вілкоксона. Справді, у медіанному тесті ми враховуємо лише те, чи лежить спостереження вище, чи нижче медіани. А у тесті Вілкоксона ранг показує, на якому саме місці опинилось спостереження. Начеб-то тут інформації більше, отже можна сподіватись, що тест Вілкоксона буде більш потужним.

Але в дійсності, для різних розподілів даних більш потужними виявляються різні тести. Одного тесту, який можна було б рекомендувати для всіх випадків, немає. Тому у різних прикладних застосуваннях використовуються різні тести, в залежності від того, який з них добре зарекомендував себе на практиці у даній області.

### 7.3.3 Перевірка незалежності двох змінних. Рангові коефіцієнти кореляції

Задача перевірки незалежності двох змінних — одна з найбільш поширених у роботі прикладного статистика. Нехай дані являють собою вибірку, у якій для кожного досліджуваного об'єкта спостерігаються дві змінні $X$ і $Y$. Потрібно перевірити, чи є залежність між цими змінними.

Таким чином вибірка складається з незалежних однаково розподілених векторів $(X_j, Y_j)$, $j = 1, \ldots, n$. Основною є гіпотеза $H_0$ про незалежність $X_j$ і $Y_j$. Загальна альтернатива полягає в тому, що між $X_j$ і $Y_j$

існує залежність. У випадку, коли припускається, що $(X_j, Y_j)$ — гауссів випадковий вектор, найкращим тестом для перевірки $H_0$ вважається тест на основі вибіркового коефіцієнта кореляції Пірсона. Нагадаємо, що коефіцієнт кореляції Пірсона між змінними $X$ і $Y$ визначається як

$$r(X,Y) = \frac{\sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y})}{\sqrt{\sum_{j=1}^n (X_j - \bar{X})^2 \sum_{j=1}^n (Y_j - \bar{Y})^2}},$$

де $\bar{X}$ і $\bar{Y}$ — вибіркові середні значення змінних $X$ і $Y$.

Але ми розглядаємо непараметричну задачу, в якій розподіл спостережень вважається повністю невідомим. У цьому випадку використовуються рангові коефіцієнти кореляції.

$\rho$ **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 кореляції Пірсона між рангами спостережень:

$$\rho(X.Y) = r(R^X, R^Y) = \frac{\sum_{j=1}^n (R_j^X - \bar{R}^X)(R_j^Y - \bar{R}^Y)}{\sqrt{\sum_{j=1}^n (R_j^X - \bar{R}^X)^2 \sum_{j=1}^n (R_j^Y - \bar{R}^Y)^2}}, \qquad (7.12)$$

де $\bar{R}^X$, $\bar{R}^Y$ — середні значення рангів по $X$ і $Y$ за всією вибіркою.

Якщо зв'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 зв'язаних рангів,

$$\rho(X.Y) = 1 - \frac{6\sum_{j=1}^n (R_j^X - R_j^Y)^2}{n(n^2-1)}. \qquad (7.13)$$

Ця формула більш популярна ніж (7.12), але вона не дає правильних результатів, якщо є зв'язані ранги. У цьому випадку для підрахунку $\rho$ слід користуватись (7.12).

$\tau$ **Кенделла** (англ. *Kendall's tau coefficient*). Коефіцієнт кореляції Спірмена $\rho$ між змінними $X$ і $Y$ порівнює ранги об'єкта за $X$ та за $Y$.

Чим більші різниці цих рангів, тим більше $\rho$. При обчисленні кореляції Кенделла використовується інший підхід — підраховують пари об'єктів, у котрих порядок по зростаню $X$ і $Y$ — однаковий. Якщо всі пари є такими — кореляція дорівнює 1. Якщо у всіх парах порядок по $X$ протилежний порядку по $Y$ — кореляція дорівнює -1. У проміжних випадках кореляція коливається між -1 і 1.

Припустимо, що зв'язаних рангів немає. Розглянемо всі можливі пари індексів $(i, j)$, $1 \leq i < j \leq n$.Назвемо пару $(i, j)$ узгодженою по змінних $X$ та $Y$, якщо

$$(X_i - X_j)(Y_i - Y_j) > 0.$$

В узгодженій парі елементи по порядку зростання $X$ розташовані так само, як і по порядку зростання $Y$. Якщо порядок по $X$ протилежний порядку по $Y$, тобто

$$(X_i - X_j)(Y_i - Y_j) < 0,$$

пара $(i, j)$ називається не узгодженою.

Нехай $n_+$ — кількість усіх узгоджених пар, $n_-$ — кількість всіх неузгоджених. Зрозуміло, що $n_+ + n_- = n(n-1)/2 = n_0$ — кількість всіх можливих пар індексів.

Кореляцією Кендалла між змінними $X$ і $Y$ (за відсутності зв'язаних рангів) називають

$$\tau(X, Y) = \frac{n_+ - n_-}{n_0},$$

При такому означенні $-1 \leq \tau(X, Y) \leq 1$.

У випадку наявності зв'язаних рангів існує декілька уточнень означення кореляції Кендалла. Одне з найбільш популярних — $\tau_b$ (читається "тау-бе Кендалла").

Припустимо, що у вибірці є зв'язані ранги. Позначимо $t_1, \ldots, t_k$ — кількості елементів у групах зі зв'язаними рангами по змінній $X$. (Тобто у вибірці присутні рівно $t_1$ елементів у яких $X$ приймає 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 не узгоджених пар (при цьому, якщо ранги по хоча б одній змінній у парі є зв'язаними, така пара не враховується ні серед узгоджених, ні серед не узгоджених), $n_0 = n(n-1)/2$.

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 $\pm 1$).

**Тести незалежності на основі рангових кореляцій.** При виконанні гіпотези $H_0$ про незалежність $X_j$ і $Y_j$ при великих обсягах вибірки рангові коефіцієнти кореляції є близькими до 0. Тому тести для $H_0$ на їхній основі полягають в тому, що абсолютна величина коефіцієнта кореляції порівнюється з пороговим значенням. Якщо кореляція виходить за поріг — гіпотеза про незалежність відкидається, інакше — приймається.

Якщо виконана $H_0$, розподіли кореляцій Спірмена і Кендалла не залежать від розподілу $X_j$ і $Y_j$, тому порогові значення можна розрахувати за стандартним рівнем значущості без додаткових припущень про ці розподіли (якщо вони неперервні). При малих обсягах вибірки використовують точні значення на основі комбінаторних розрахунків, при великих, користуються асимптотичними формулами. Наприклад, тест на основі коефіцієнта кореляції Спірмена з рівнем значущості $\alpha_0$ має вигляд:

Якщо $|\rho(X,Y)| \le C_{\alpha_0}$ приймається гіпотеза про незалежність $X$ і $Y$.

Якщо $|\rho(X,Y)| > C_{\alpha_0}$ — гіпотеза про незалежність відхиляється.

При $n > 30$ поріг тесту можна розраховувати за асимптотичною формулою

$$C_{\alpha_0} = \sqrt{\frac{f_{\alpha_0}/(n-2)}{1 + f_{\alpha_0}/(n-2)}},$$

де $f_{\alpha_0}$ — квантиль рівня $\alpha_0$ для $F$ розподілу Фішера з одним ступенем вільності чисельника і $n-2$ ступенями вільності знаменника. (Див. [17], формула (3.15)).

Аналогічно влаштований тест для перевірки гіпотези $H_0$ про незалежність змінних $X$ і $Y$ на основі коефіцієнта кореляції Кендалла. Він відхиляє гіпотезу про незалежність, якщо $|\tau(X,Y)|$ перевищує порогове значення. При великих обсягах вибірки це значення можна знайти, використовуючи асимптотичну нормальність $\tau(X,Y)$: якщо $H_0$ виконано,

то при $\to \infty$,

$$\frac{3\sqrt{n(n-1)}\tau(X,Y)}{\sqrt{2(2n+5)}} \xrightarrow{W} N(0,1),$$

(див. [17], с. 414).

Але що насправді перевіряють такі тести? Можна показати, що при зростанні $n \to \infty$

$$\rho(X,Y) \to 3(2\,\mathsf{P}[(X_2 - X_1)(Y_3 - Y_1) > 0] - 1) = \rho_\infty(X,Y),$$

і

$$\tau(X,Y) \to 2\,\mathsf{P}[(X_2 - X_1)(Y_2 - Y_1) > 0] - 1 = \tau_\infty(X,Y),$$

$(X_j, Y_j)$ — незалежні копії вектора $(X,Y)$.

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 справедливі і для $\tau$ Кендалла.

## 7.4 Запитання і задачі

**Запитання**

1. Що таке непараметрична статистична гіпотеза?
2. Чим відрізняються нульова гіпотеза і альтернатива?
3. Що таке стандартний рівень значущості? Як за стандартним рівнем значущості обрати поріг тесту?
4. Що таке однорідні і не однорідні вибірки?
5. Що таке вибірки, розподіли яких відрізняються лише зсувом?
6. Як визначається статистика медіанного тесту однорідності двох вибірок?
7. За яких умов розподіл статистики медіанного тесту однорідності є наближено нормальним?

8. Що таке ранг спостереження у вибірці за певною змінною?

9. Що таке зв'язані ранги, як визначають ранг спостереження за наявності зв'язків?

10. Чим коефіцієнт кореляції Спірмена відрізняється від коефіцієнта кореляції Пірсона?

11. Що таке узгоджені і неузгоджені пари, як за ними визначається коефіцієнт кореляції Кендала?

11. Як визначається поріг тесту незалежності двох змінних на основі кореляції Спірмена, якщо обсяг вибірки великий?

**Задачі**

1. Доведіть формули (7.6).

2. Знайдіть поріг для одностороннього медіанного тесту $\pi^+(S)$, який забезпечує заданий рівень значущості $\alpha_0$ при великих обсягах вибірки.

3. Опишіть у явному вигляді тест на основі статистики Вілкоксона для двосторонньої альтернативи. Знайдіть поріг цього тесту, який забезпечує заданий рівень значущості.

**Завдання для виконання на комп'ютері**

Опишіть явно тест для перевірки гіпотези про незалежність змінних $X$ і $Y$ на основі коефіцієнта кореляції Кендалла $\tau(X,Y)$. Реалізуйте цей тест та тест на основі коефіцієнта кореляції Спірмена у вигляді функцій. Перевірте роботу цих тестів на модельованих даних з $X$ що має розподіл $F$, $Y$ з розподілом $G$ ($F$ і $G$ вибирається відповідно до вашого варіанту). Обсяг модельованої вибірки — $n = 50, 100, 250$ спостережень.

Проведіть перевірку того, наскільки ймовірність помилки першого роду побудованих вами тестів відповідає номінальному рівню значущості (виберіть його $\alpha = 0.05$). Для цього згенеруйте $B = 10000$ вибірок, у яких $X$ та $Y$ — незалежні з розподілами $F$ і $G$ відповідно, для кожної з вибірок проведіть тестування і підрахуйте частоту помилок тестів. Якщо частота помилки на вашу думку значно відрізняється від $\alpha$, уточніть поріг тесту, використовуючи імітаційне моделювання.

Після цього проведіть перевірку потужності тестів. Для цього згенеруйте вибірки $\tilde{Y} = Y + \gamma X$, де $\gamma = 0.2$. Проведіть перевірку незалежності $X$ і $\tilde{Y}$ використовуючи тести на основі $\rho(X,\tilde{Y})$ і $\tau(X,\tilde{Y})$. Підрахуйте потужності тестів.

Опишіть результати ваших експериментів у вигляді таблиць та зробіть висновки про те, який тест виявився більш потужним у вашому випадку.

**Індивідуальні варіанти для моделювання:**

1. $F$ і $G$ — стандартний нормальний розподіл.
2. $F$ і $G$ — розподіл Лапласа зі щільністю $\exp(-|x|)/2$.
3. $F$ і $G$ — логістичний розподіл з функцією розподілу $1/(1+exp(x))$.
4. $F$ і $G$ — рівномірний розподіл на $[0,1]$.
5. $F$ — логістичний розподіл, $G$ — розподіл Лапласа.
6. $F$ — нормальний розподіл, $G$ — розподіл Лапласа.
7. $F$ — нормальний розподіл, $G$ — рівномірний розподіл.
8. $F$ — розподіл Лапласа, $G$ — рівномірний розподіл.
9. $F$ — нормальний розподіл, $G$ — логістичний розподіл.
10. $F$ — рівномірний розподіл, $G$ — логістичний розподіл.

# Література


[1] Карташов М.В. "Теорія ймовірностей і математична статистика". Київ, Видавничо-поліграфічний центр 'Київський університет', , - 2009

[2] Козаченко Ю. В. Лекції з вейвлет аналізу. – Київ: ТВіМС, 2004. – 147 с.

[3] Майборода Р. Є. Комп'ютерна статистика. ВПЦ "Київський університет", 589 р. - 2019.

[4] Майборода Р.Є. Непараметрична статистика. Рекомендації по виконанню індивідуальних робіт. Режим доступу:

`http://probability.univ.kiev.ua/userfiles/mre/nonparam_task.pdf`

[5] Турчин В.М. Теорія ймовірностей і математична статистика. Основні поняття, приклади, задачі.— Дніпропетровськ: ІМА-прес, 2014. — 556 с.

[6] Aeberhard S., Coomans D., de Vel O. Comparative analysis of statistical pattern recognition methods in high dimensional settings.– Pattern Recognition Volume 27, Issue 8, 1994, Pages 1065-1077

[7] Andersen P.K., Borgan, Ø, Gill R.D., Keiding N. Statistical Models Based on Counting Processes-Springer-Verlag New York (1993) 768p.

[8] Bickel P.J., Fan J. Some problems on the estimation of unimodal densities. – Statistica Sinica 6 (1999), 23-45.

[9] Borovkov A.A. Mathemanical Statistics/ Gordon and Breach 1998, 592p.

[10] Chernoff H., Teicher H. A central limit theorem for sums of interchangeable random variables.— Ann. Math. Statist., V29, 1958,p.118-130.

[11] Cleveland William S. Robust Locally Weighted Regression and Smoothing Scatterplots.— Journal of the American Statistical Association, Vol. 74, No. 368 (Dec., 1979), pp. 829-836.

[12] Dekkers, A.L.M.; de Haan, L. On the estimation of the extreme value index and large quantile estimation. Ann. Stat. 1989, 17, 1795–1832

[13] Devroye L., Györfi L. Nonparametric density estimation / The L1 View-John Wiley & Sons (1984)

[14]

[15] Fan J. , Gijbels I. Local Polynomial Modelling and its Applications (1996, Springer US) Green P.J., Silverman B.W. Nonparametric regression and generalized linear models.— Chapman & Hill, London, 1995, 182 p.

[16] Ibragimov I. A., Has'minskii R. Z. Statistical estimation, asymptotic theory.— Springer-Verlag, New York, 1981, 403p.

[17] Gibbons J.D., Chakraborti S. Nonparametric Statistical Inference.— Marcel Dekker, New York, 2003, 645p.

[18] Gill R. D., Vardi Y., Wellner J.A. Large sample theory of empirical distributions in biased sampling models. — Ann. Statist. 1988, V. 16, N. 3, p. 1069–1112.

[19] Glantz S.A. Primer of Biostatistics/ Mc Graw Hill 2012, 327p.

[20] Grenander, U. (1956). On the theory of mortality measurement. I. Skand. Aktuarietidskr. 39 70–96.

[21] Groeneboom, P. (1985). Estimating a monotone density. In Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983). Wadsworth Statist./Probab. Ser. 539–555. Wadsworth, Belmont, CA.

[22] Härdle W., Kerkyacharian G., Picard D., Tsybakov A. Wavelets, approximation and statistical applications. – New York: Springer, 1998. – 265 p.

[23] H ardle, W., Muller, M., Sperlich, S., Werwatz A. (2004) Nonparametric and Semiparametric Models.

[24] Fernando P. Polack, Stephen J. Thomas, Nicholas Kitchin, et al. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine.— N Engl J Med 2020; 383:2603-2615

[25] Prakasa Rao, B. L. S. (1969). Estimation of a unimodal density. Sankhyā Ser. A 31 23–36.

[26] Shao J. Mathematical statistics.- Springer-Verlag: New York, 1998. - 530 p.

[27] Vardi Y. Empirical distributions in selection bias models. — Ann. Statist. 1985, 13, 178-203.

[28] Larry Wasserman-All of Nonparametric Statistics.—Springer NY 2006.— 268.