Непараметрична статистика

Викладач: Майборода Ростислав Євгенович, кафедра теорії ймовірностей, статистики та актуарної математики.

e-mail: mre@univ.kiev.ua

Вступ

Курс "Непараметрична статистика" передбачає вивчення теоретичного матеріалу, оволодіння технікою статистичного аналізу за допомогою програмної системи R та виконання індивідуальних домашніх завдань, пов'язаних з реалізацією методів непараметричного статистичного аналізу в R.

Програмна система R є безкоштовним програмним забезпеченням, яке можна отримати за адресою

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

(для операційної системи Windows) або

http://cran.rstudio.com/

— для Windows, Linux, Mac OS.

Там же описана процедура встановлення R на комп'ютері.

Для знайомства з роботою у цій системі можна рекомендувати розділи 1 і 2 у книзі

Майборода Р.Є. Комп'ютерна статистика — професійний старт (2018).
http://probability.univ.kiev.ua/userfiles/mre/compsta1.pdf
або

Venables, W.N., Ripley, B.D. (2003) Modern Applied Statistics with S. Fourth Edition.

Для успішного проходження курсу необхідно звернутись до викладача за e-mail mre@univ.kiev.ua та отримати індивідуальні дані для виконання домашніх завдань. За результатами виконання завдань потрібно скласти звіти (на папері), у яких по кожній роботі мають бути: назва роботи, значення параметрів з якими виконувалась робота, тексти програм на R, графіки, числові результати обчислень та висновки, що вимагаються у відповідних завданнях роботи. На іспиті треба мати з собою звіти та програми в електронній формі.

Програма теоретичного курсу

1. Задача непараметричного оцінювання функції розподілу. Емпіричний розподіл та емпірична функція розподілу, її статистичні властивості (незміщеність, консистентність, асимптотична нормальність). Теорема Глівенка-Кантеллі. [Боровков, Глава 1, п.2, теоремы 1, 2, 2А, 3]

2. Аналіз тривалості життя. Поняття про цензуровані вибірки (випадкове цензурування з права). Оцінка Каплана-Мейера для функції розподілу за цензурованою вибіркою. Її асимптотичні властивості. Формула Грінвуда. Асимптотичний довірчий інтервал для функції розподілу. [Shao, section 5.1.2, Example 5.3; Гланц, глава 11]

3. Задача оцінювання щільності розподілу. Гістограма як оцінка щільності. Ядерні оцінки щільності, їх зміщення та дисперсія. Асимптотична нормальність ядерних оцінок щільності. [Боровков, Глава 1 п.10 Теорема 1; Hardle, Chapter 2, Chapter 3]

4. Асимптотичне зображення MISE для ядерних оцінок щільності. Вибір оптимального параметру згладжування та оптимального ядра. Правило Сільвермана та кросс-валідація. [Hardle, Chapter 3]

5.Задача класифікації. Поняття про баєсів та емпірично-баєсів класифікатор. Ймовірність помилки класифікації. Межа Джорфі. Метод найближчого сусіда. [Деврой, глава 10, п. 1, теорема 1.; п.5, теорема 5 (без доведення)]

6 Поняття статистичного тесту. Характеристики якості тестів: ймовірності помилок, потужність рівень значущості. Досягнутий рівень значущості як одна з можливих статистик тесту. Рангові тести. Тест для перевірки незалежності двох змінних на основі коефіцієнта кореляції Спірмена. [Боровков; Shao; Гланц, Глава 8.]

Рекомендації по виконанню лабораторних робіт.

Робота 1. Емпірична функція розподілу

Нехай спостерігається вибірка X з незалежних, однаково розподілених випадкових величин $X_1,\ldots,X_n$ з невідомою функцією розподілу F. Для оцінки F за X можна використовувати емпіричну функцію розподілу

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

$\hat{F}_n(x)$ є незміщеною оцінкою для $F(x)$ з дисперсією

$$\sigma^2_{\hat{F}_n(x)} = \frac{1}{n}F(x)(1-F(x)).$$

Це можна використати для побудови асимптотичного довірчого інтервалу для $F(x)$ при фіксованому x, аналогічно тому, як будуються асимптотичні довірчі інтервали для невідомих параметрів розподілу у параметричній статистиці (див. п. 8.5.1 в [4]).

Завдання лабораторної роботи.

0. Отримати індивідуальні значення параметрів F, x_0, $n_0,\ldots n_5$ у викладача.

1. Задати в R власну функцію, яка має виклик `Femp(x,Sample)` і підраховує значення емпіричної функції розподілу для вибірки `Sample` в точці `x`.

2. Згенерувати вибірку заданого обсягу $n = n_0$, з заданим розподілом F. Намалювати в R графіки справжньої функції розподілу та її оцінки на одному рисунку різним кольором. Інтервал значень аргумента підібрати так, щоб на рисунку можна було бачити зміну значень обох функцій в інтервалі від 0 до 1.

3. Написати програму в R, яка для довільного заданого обсягу вибірки n генерує $B = 1000$ вибірок з розподілом F, для кожної вибірки підраховує значення $\hat{F}_n(x_0)$ у заданій точці x_0. По отриманій вибірці з оцінок підраховуються вибіркові середнє і дисперсія та будується гістограма.

Цю програму застосувати для обчислень при $n = n_1$, n_2, n_3. Отримані результати порівняти з теоретичним математичним сподіванням та дисперсією оцінок. Зробити висновки про можливість застосування нормального закону для наближеного опису розподілу оцінок.

4. Реалізуйте розрахунок меж довірчого інтервалу для функції розподілу у заданій точці у вигляді функції в `R`. 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 п. 0, точок `x`, що дорівнюють квантилям рівня 1/3, 1/2 і 2/3 для розподілу F і обсягів вибірки $n = n_1, n_2, \ldots, n_5$. 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 α. 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.

Увага! При виконанні роботи не дозволяється використовувати оператори циклу та/або умовного переходу. Можна використовувати вбудовані функції `sum`, `mean`, `median`, `var` та ін. а також функції, що застосовують інші функції до масивів: `apply`, `sapply`.

Варіанти індивідуальних завдань:

(1) $F \sim N(1,1)$, $n_1 = 50$, $n_2 = 100$, $n_3 = 500$,

(2) $F \sim \chi^2(3)$, $n_1 = 100$, $n_2 = 500$, $n_3 = 1000$,

(3) $F \sim \mathrm{Exp}_{\lambda=1}$, $n_1 = 50$, $n_2 = 100$, $n_3 = 500$,

(4) $F \sim \chi^2(4)$, $n_1 = 250$, $n_2 = 500$, $n_3 = 1000$,

(5) F — T-Стьюдента з 6-ма ступенями вільності, $n_1 = 100$, $n_2 = 1000$, $n_3 = 5000$.

Література.

1. Боровков А.А. (1997) Математическая статистика.[Розділ 1, п. 2.]

2. Shao J. (1999) Mathematical Statistics.[Section 5.1.1]

Робота 2. Непараметрична оцінка щільності розподілу

Нехай $X = (X_1, \ldots, X_n)$ — вибірка з незалежних, однаково розподілених випадкових величин, що мають щільність розподілу відносно міри

Лебега - $f(x)$. Ця щільність невідома, задача полягає в її оцінці за вибіркою X.

Однією з можливих оцінок є гістограма відносних частот.

Інша — ядерна оцінка щільності. Для того, щоб її задати, потрібно обрати ядро $K(x)$ та параметр згладжування h. На роль ядра можна обрати, в принципі, будь-яку функцію, що сама є щільністю деякого ймовірнісного розподілу. Параметр згладжування має бути деяким додатнім числом.

Ядерна оцінка щільності визначається як

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

Вибір параметра згладжування h роблять з міркувань балансу між зміщенням (систематичною похибкою) оцінки та її розкидом. Збільшення h зменшує розкид та збільшує зміщення. Інтегральною характеристикою, що визначає якість оцінки є проінтегрована середньоквадратична похибка

$$\mathrm{MISE}(\hat{f}_n) = \int_{-\infty}^{+\infty} (\hat{f}_n(x) - f(x))^2 dx.$$

При виконанні певних умов гладкості на функцію f та умов $h = h_n \to 0$, $h_n n \to \infty$ при $n \to \infty$, має місце наступне асимптотичне зображення:

$$\mathrm{MISE}(\hat{f}_n) = \left(\frac{D^2 h_n^2}{2}\right)^2 \varphi + \frac{d^2}{n h_n}, \text{ при} n \to \infty,$$

де $d^2 = \int_{-\infty}^{+\infty} K^2(x)dx$, $D^2 = \int_{-\infty}^{+\infty} x^2 K(x)dx$, $\varphi = \int_{-\infty}^{+\infty} (f''(x))^2 dx$, $f''(x)$ — друга похідна функції $f(x)$.

Мінімум цього виразу досягається на

$$h_n = h_n^* = \left(\frac{d^2}{n D^4 \varphi}\right)^{1/5}.$$

Це значення будемо називати теоретичним оптимальним значенням параметра згладжування. Використати його на практиці неможливо, оскільки нам невідоме значення φ, що визначається за невідомою щільністю f. Якщо наближено вважати, що $f \sim N(a, S^2)$, то

$$\varphi = S^{-1/5} \frac{3}{8\sqrt{\pi}}.$$

Оцінити S^2 можна виправленою вибірковою дисперсією X. Підставивши отриману оцінку у формулу для φ замість S^2 отримуємо оцінку $\hat{\varphi}$, яку можна підставити у формулу для теоретичного оптимального значення h. Отримане в результаті h^S називають значенням параметра згладжування, обраним за правилом Сілвермана.

Метод кросс-валідації для вибору параметра згладжування описано у п. 3.3.2 [Härdle et al. 2003]. Для цього використовується функціонал кросс-валідації $CV(h)$. Оптимальне значення h обирається як точка мінімуму $CV(h)$.

Завдання лабораторної роботи

0. Отримати у викладача індивідуальні значення ядра K, модельованої щільності розподілу f, обсягу вибірки n_0.

1. Реалізувати у вигляді функції у системі R ядерну оцінку щільності розподілу із заданим ядром K. Параметри функції - аргумент щільності x, параметр згладжування h, вибірка X.

2. Згенерувати вибірку із заданою щільністю розподілу f заданого обсягу $n = n_0$. Оцінити щільність використовуючи ядерну оцінку щільності. Результат зобразити у вигляді графіків справжньої щільності та її оцінки на одному рисунку. Підібрати параметр згладжування за графіками "на око", так, щоб оцінка найближче відповідала оцінюваній функції.

3. Використати для вибору параметру згладжування метод Сілвермана, оптимальний теоретичний вибір та метод кросс-валідації. Для наближеної мінімізації функціоналу CV можна інтервал $[h^S/4, 4h^S]$ розбити рівномірно сіткою з багатьох (наприклад, 20) точок, підрахувати значення CV у всіх точках сітки і вибрати ту точку, у якій CV - найменше.

4. З усіма значеннями параметра згладжування, обраними у попередньому пункті, намалювати графіки відповідних оцінок функцій щільності разом зі справжнім значенням. Підрахувати функціонал

$$\mathrm{ISE}(\hat{f}_n) = \int_{-\infty}^{+\infty} (\hat{f}_n(x) - f(x))^2 dx.$$

Зробити висновок про те, який підхід до вибору параметра згладжування виявився кращим у Вашому випадку.

Значення параметрів для індивідуальних завдань.

(1) $f \sim N(1,1)$, K — трикутне ядро.

(2) $f \sim \chi^2(3)$, K — ядро Єпанєчнікова.

(3) $f \sim N(0,2)$ з мат. сподіванням 0, $\lambda = 1$, K — гауссове ядро.

(4) f — щільність бета-розподілу з $a = 2$, $b = 4$, K — гауссове ядро.

(5) f — щільність розподілу Фішера $F(4,10)$, K — ядро Єпанєчнікова.

Література

1. Боровков А.А. (1997) Математическая статистика.[Розділ 1, п. 10.]
2. Hardle, W., Muller, M., Sperlich, S., Werwatz A. (2004) Nonparametric and Semiparametric Models. [Chapter 3].

Робота 3. Перевірка залежності між двома змінними.

Один із способів перевірки статистичної залежності між двома змінними, що характеризують досліджувані об'єкти — використання коефіцієнтів кореляції. Якщо розподіл невідомий, доцільно використовувати рангові кореляції.

Нехай спостерігається n об'єктів, кожен з яких має дві числові характеристики (змінні) X_j та Y_j (для j-того об'єкта). Нехай всі значення X_j, $j = 1, \ldots, n$ — різні. Рангом j-того об'єкта відносно змінної X — R_j^X, називають кількість об'єктів у вибірці, для яких значення характеристики X не перевищує X_j. Таким чином, об'єкт з найменшим X має ранг 1, наступний по зростанню X — ранг 2, і т. д. Аналогічно визначаються ранги відносно Y — R_j^Y. Ранговий коефіцієнт кореляції Спірмена між X і Y можна визначити як

$$\rho(X,Y) = 1 - \frac{6\sum_{j=1}^{n}(R_j^X - R_j^Y)^2}{n^3 - n}.$$

Якщо розподіли X та Y неперервні і ці змінні незалежні між собою, то розподіл $\rho(X,Y)$ не залежить від розподілів X та Y. Крім того, у цьому випадку, при зростанні обсягу вибірки, $\rho(X,Y) \to 0$ за ймовірністю. Це дозволяє побудувати наступний тест для перевірки гіпотези H_0: X та Y є незалежними випадковими величинами, проти альтернативи H_1: X та Y корельовані (тобто граничне значення $\rho(X,Y)$ при $n \to \infty$ відрізняється від 0). Тест має вигляд:

Якщо $|\rho(X,Y)| \le C_\alpha$ — прийняти H_0, інакше — відхилити.

Поріг тесту C_α вибирають так, щоб ймовірність помилки 1-го роду для нього дорівнювала заданому рівню значущості α. Це еквівалентно умові

$$C_\alpha = \min\{C : \ \mathsf{P}_{H_0}\{|\rho(X,Y)| > C\} \le \alpha\}.$$

Інша форма того ж тесту використовує досягнутий рівень значущості (significance, p-level) p, який потрібно задати як функцію від $\rho(X,Y)$,

таким чином, щоб нерівність $|\rho(X,Y)| \le C_\alpha$ стала еквівалентною нерівності $p > \alpha$. Для цього обирають $p = 1 - F(|\rho(X,Y)|)$, де F — функція розподілу статистики $|\rho(X,Y)|$ при виконанні H_0. (Замість $\rho(X,Y)$ тут можна використовувати будь-яку еквівалентну статистику, наприклад T-статистику).

Для того, щоб реалізувати тест на практиці, можна скористатись двома підходами.

1. **Асимптотичний підхід.** Відомо, що при $n > 50$, розподіл T-статистики

$$T = \rho(X,Y)\sqrt{\frac{n-2}{1-\rho(X,Y)^2}}$$

добре наближається T-розподілом Стьюдента з $n-2$ ступенями вільності. Цей розподіл можна використовувати замість точного як при підрахунку порогу тесту за заданим α, так і для знаходження досягнутого рівня значущості.

2. **Імітаційне моделювання.** Якщо точність асимптотичного наближення недостатня, то можна скористатись наступним підходом. Згенерувати B незалежних вибірок обсягу n, у яких містяться пари (X_j, Y_j), $j = 1, \dots, n$ незалежних між собою однаково розподілених випадкових величин. (Розподіл цих величин неважливий, він може бути, наприклад, рівномірним на [0,1]). По кожній вибірці підрахувати коефіцієнт кореляції Спірмена між X та Y. За отриманою вибіркою з B коефіцієнтів кореляції підрахувати емпіричну функцію розподілу і використати її для апроксимації справжньої ф.р. F для коефіцієнта Спірмена незалежних спостережень.

Завдання роботи.

0. Отримати у викладача данні зі значеннями X та Y для аналізу.

1. Розробити власну функцію системи `R`, що реалізує підрахунок коефіцієнта кореляції Спірмена ρ.

2. На заданих даних підрахувати значення $\rho(X,Y)$ використовуючи функцію, розроблену у п.1 та функцію `cor`.

3. Підрахувати досягнутий рівень значущості тесту для перевірки незалежності X та Y, трьома способами — використовуючи (1) функцію `cor.test`, (2) асимптотичний підхід та (3) імітаційне моделювання. Зробити висновки.

Індивідуальні завдання для роботи.

Варіант 1.

	1	2	3	4	5	6	7	8	9	10
X	10	7	3	5	12	14	8	1	0	-3
Y	7	6	18	5	4	-2	-1	9	0	8

Варіант 2.

	1	2	3	4	5	6	7	8	9	10
X	-3	11	13	9	16	-5	5	15	16	12
Y	-2	-3	4	1	6	-1	10	5	17	15

Варіант 3.

	1	2	3	4	5	6	7	8	9	10
X	13	-5	10	5	-3	-1	0	-2	20	2
Y	-2	16	6	0	7	0	20	17	13	20

Варіант 4.

	1	2	3	4	5	6	7	8	9	10
X	10	5	4	12	9	8	1	3	2	-3
Y	-1	6	19	18	8	16	7	14	4	12

Варіант 5.

	1	2	3	4	5	6	7	8	9	10
X	-5	6	15	18	2	9	18	5	9	3
Y	3	1	6	11	10	20	12	4	7	2

Література 1. Гланц С.(1999) Медико-биологическая статистика.[Глава 8.]

2. Gibbons J.D., Chakraborti S.(2003) Nonparametric Statistical Inference

Литература

[1] Боровков А.А. (1997) Математическая статистика.

[2] Гланц С.(1999) Медико-биологическая статистика.

[3] Деврой Л., Дьерфи Л.(1988) Непараметрическое оценивание плотности.

[4] Майборода Р. Є. Комп'ютерна статистика/ 2017.- 405с. Режим доступу:

`http://probability.univ.kiev.ua/userfiles/mre/compsta.pdf`

[5] Gibbons Jean Dickinson, Chakraborti Subhabrata (2003) Nonparametric Statistical Inference, Fourth Edition (Statistics: a Series of Textbooks and Monographs)

[6] Hardle, W., Muller, M., Sperlich, S., Werwatz A. (2004) Nonparametric and Semiparametric Models.

[7] Shao J. Mathematical statistics.- Springer-Verlag: New York, 1998. - 530 p.

[8] Venables, W.N., Ripley, B.D. (2003) Modern Applied Statistics with S. Fourth Edition

[9] Larry Wasserman-All of Nonparametric Statistics.—Springer NY 2006.— 268.