Домашнее задание №2

Тема: непараметрическое оценивание, анализ экстремальных значений, доверительные интервалы

Дедлайн: 11 апреля 2024 г., 23:59

Задача 1 *(2.5 балла).*

Рассмотрим базу данных `RTdata`[1], доступную в пакете `mixtools` в `R` и содержащую данные о времени (в миллисекундах) реакции 197 детей при прохождении некоторой серии из 6 тестов. Для простоты рассмотрим только первый тест (первую колонку датафрейма). Основным вопросом является возможность выделения нескольких подгрупп, для которых время реакции имеет разное распределение.

1. Постройте гистограммы рассматриваемых данных с параметрами bandwidth, выбранными по правилу Стёржеса, Скотта и Фридмана-Дьякони. Опишите полученный результат.

2. Отобразите на одной картинке ядерные оценки плотности со всеми ядрами, доступными в `R`/`Python`, при фиксированном параметре bandwidth, а на другой — со всеми доступными bandwidth при фиксированном ядре.

3. Постройте (без вывода графиков) гистограммы рассматриваемых данных с $10, 11, \ldots, 20$ столбцами. Среди множества построенных на предыдущем шаге ядерных оценок и множества построенных гистограмм найдите наиболее близкие, то есть такие, для которых минимально значение

$$\frac{1}{M}\sum_{k=1}^{M}\left(\widehat{p}_n^H(x_k) - \widehat{p}_n^K(x_k)\right)^2,$$

где $\widehat{p}_n^H$ — гистограмма, $\widehat{p}_n^K$ — ядерная оценка, $\{x_k\}_{1\le k\le M}$ — точки, в которых доступна ядерная оценка. Отобразите полученную гистограмму и ядерную оценку на одном графике.

Сделайте вывод о возможности наличия различных подгрупп по времени реакции.

Задача 2 *(3 балла).* Пусть p — плотность нормального распределения со средним 0 и дисперсией σ^2. Рассматриваются две ядерные оценки плотности: с ядром Епанечникова и с ядром вида

$$K(x) = \sum_{m=0}^{2} L_m(0)L_m(x)e^{-x}\mathbb{1}\{x \ge 0\}, \tag{1}$$

[1]Загрузить данные в `R` можно командой `data(RTdata)`

где $L_0, L_1, \dots$ — полиномы Лагерра, определяемые как

$$L_n(x) = \frac{e^x}{n!} \frac{d^n}{dx^n}(x^n e^{-x}), \qquad n \in \{0, 1, \dots\}.$$

1. Докажите, что функция (1) является ядром порядка 2, то есть удовлетворяет соотношениям

$$\int_{\mathbb{R}} K(x)\,dx = 1, \qquad \int_{\mathbb{R}} x^j K(x)\,dx = 0, \qquad j \in \{1, 2\}.$$

2. Для ядерной оценки с ядром Епанечникова найдите значение bandwidth h, минимизирующее асимптотическую интегрированную среднеквадратическую ошибку (AMISE), то есть величину

$$\mathrm{AMISE}(\widehat{p}_n, h) = \frac{h^4}{4}\left(\int_{\mathbb{R}} (p''(x))^2\,dx\right)\left(\int_{\mathbb{R}} x^2 K(x)\,dx\right)^2 + \frac{1}{nh}\int_{\mathbb{R}} K^2(x)\,dx.$$

Покажите, что для найденного значения h_{opt} выполнено

$$\lim_{n\to\infty} n^{4/5}\,\mathrm{AMISE}(\widehat{p}_n, h_{opt}) = \frac{3^{4/5}}{5^{1/5}4}\left(\int_{\mathbb{R}} (p''(x))^2\,dx\right)^{1/5}.$$

3. Укажите хотя бы одно значение bandwidth h, для которого ядерная оценка с ядром (1) лучше (в смысле AMISE) ядерной оценки с ядром Епанечникова и любым bandwidth.

4. Зафиксируйте значение $\sigma > 0$ и симулируйте выборку объёма $n = 1000$ из нормального распределения со средним 0 и дисперсией σ^2. Для каждого из значений badwidth, взятых по решётке от 0.1 до 5 с шагом 0.1, и каждой из ядерных оценок рассчитайте эмпирический аналог MISE

$$\widehat{\mathrm{MISE}}(\widehat{p}_n, h) = \frac{1}{M}\sum_{k=1}^{M} \left(\widehat{p}_n(x_k) - p(x_k)\right)^2, \tag{2}$$

где $\{x_k\}_{1\le k\le M}$ — значения, выбранные по равномерной решётке от -3 до 3, $M = 1000$. Отобразите на одном графике зависимость ошибки (2) от bandwidth для обеих оценок. Для каждой из оценок определите значение bandwidth, соответствующее наименьшему значению ошибки, и постройте (также на одном графике) соответствующие оценки плотности и истинную плотность.

Задача 3 *(2.5 балла).*

1. Как известно, область максимального притяжения закона Гумбеля включает в

себя функции распределения F, для которых справедливо представление

$$\bar{F}(x) = c \exp\left\{-\int_y^x \frac{1}{a(u)}\,du\right\}, \quad y < x < x_F, \tag{3}$$

где $\bar{F}(x) = 1 - F(x)$ — хвост распределения, $x_F = \inf\{x \in \mathbb{R}\colon F(x) = 1\}$ — крайняя правая точка, $c > 0$ — константа, $a(x)$ — положительная абсолютно непрерывная функция, такая, что $\lim_{x \to x_F} a'(x) = 0$[2]. Основной сложностью в получении представления (3) для заданной функции F является определение вспомогательной функции a. Следующее утверждение даёт один из способов установления этой функции:

Утверждение 1. *Пусть F — функция распределения с крайней правой точкой $x_F \leq \infty$, и пусть существует такое $y < x_F$, что F дважды дифференцируема на (y, x_F), причём $F'(x) > 0$ и $F''(x) < 0$ для всех $x \in (y, x_F)$. Тогда F имеет представление (3) с функцией*

$$a(x) = \frac{\bar{F}(x)}{F'(x)}$$

тогда и только тогда, когда

$$\lim_{x \to x_F} \frac{\bar{F}(x) F''(x)}{(F'(x))^2} = -1.$$

Пусть F — функция распределения стандартного нормального закона.

(a) Рассмотрев предел
$$\lim_{x \to \infty} \frac{\bar{F}(x)}{x^{-1} F'(x)},$$
покажите, что $\bar{F}(x) \sim F'(x)/x$ для достаточно больших x[3].

(b) Используя результат предыдущего пункта, покажите, что для стандартного нормального закона выполнены условия Утверждения 1. Сделайте вывод о предельном распределении максимальных значений в данной модели.

2. Пусть G — функция распределения с плотностью

$$g(x) = \sqrt{\frac{2}{\pi}} e^{-x^2/2} \mathbb{1}\{x > 0\}. \tag{4}$$

(a) Определите, к какой области максимального притяжения принадлежит данное распределение.

[2] Функции F, имеющие представление (3), называются *функциями фон Мизеса*

[3] Данное соотношение известно как *Mill's ratio*

(b) Симулируйте выборку объёма $n = 1000$ из распределения с плотностью (4). Оцените параметр обобщённого распределения экстремальных значений с помощью оценок Хилла, Деккерса-Айнмаля-де Хаана и Пикандса и постройте графики зависимости данных оценок от количества k используемых порядковых статистик. Сравните результаты с полученными в пункте 2.

(c) Сгенерируйте 1000 выборок объёма $n = 1000$ из распределения с плотностью (4) и оцените максимальное значение по каждой из выборок. Постройте графики квантиль-квантиль (QQ-plot) для полученной выборки максимальных значений и выборок объёма $n = 1000$ из распределений Фреше, Вейбулла и Гумбеля. Интерпретируйте полученные результаты.

Задача 4 *(2 балла).*

1. Пусть $X_1, X_2, \ldots, X_n$ — независимые одинаково распределённые случайные величины с функцией распределения F. Обозначим

$$\theta := F(b) - F(a)$$

для некоторых $-\infty \leq a < b \leq \infty$. Для оценивания θ предлагается оценка

$$\widehat{\theta} := \widehat{F}_n(b) - \widehat{F}_n(a),$$

где $\widehat{F}_n(x)$ — эмпирическая функция распределения $X_1, X_2, \ldots, X_n$, то есть

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}\{X_k \leq x\}, \qquad \forall x \in \mathbb{R}.$$

Постройте асимптотический доверительный интервал уровня $1 - \alpha$, $\alpha \in (0, 1)$, для θ. Как точность (в смысле ширины) полученного интервала зависит от $\widehat{\theta}$? Сделайте вывод о зависимости минимального количества n наблюдений, необходимых для построения интервала с заданной точностью, от $\widehat{\theta}$.

2. Пусть $X_1, X_2, \ldots, X_n$ — выборка из распределения с плотностью

$$p(x) = e^{-(\theta - x)} \mathbb{1}\{x \leq \theta\}, \quad \theta \in \mathbb{R}. \tag{5}$$

Постройте точный и асимптотический доверительный интервалы уровня $1 - \alpha$, $\alpha \in (0, 1)$, для параметра θ.