KDE와 히스토그램

정의

- $X_1, \dots, X_n$이 연속하는 pdf를 가지는 iid인 x의 확률표본이라고 하자.

1) 임의의 실현값 과, h>0이 주어졌을 때 임의의 구간 $(x_1 - \mathrm{h},\ x_1 + \mathrm{h})$에서

(1) 적분의 평균값 정리에 따르면 $f'(c) = \frac{\int_a^b f(t)dt}{b-a}$ 인 점 c가 반드시 존재한다.

(2) 따라서, $f(c) = \frac{\int_{x_1-\mathrm{h}}^{x_1+\mathrm{h}} f(t)dt}{2h}$ 인 점 c가 반드시 존재하며,

$\int_{x_1-\mathrm{h}}^{x_1+\mathrm{h}} f(x)\,dx = \mathrm{p}(x_1 - \mathrm{h} < x < x_1 + \mathrm{h}) = 2hf(c)$ 이다.

(3) 점 c를 추정하면

- $f(c) \approx f(x) = \frac{1}{2hn} I\{x\}$, 이 때 $I\{x\} = \begin{cases} 1, & x-h < x < x+h \\ 0, & else \end{cases}$
- 이를 확률밀도추정(KDE)라고 하고, $I\{x\}$ 를 커널함수라고 한다.

정의

- 히스토그램

1. m>0, 임의의 $a < \min(x_i)$인 구간을 다음과 같이 정의한다.

1) $\min(x_i)$ = 2, a = 1, h = 1, m = 7 일 때

(1) (a-h,a+h] , (a+h,a+3h] ,... , $(a + (2m-3)h, a + (2m-1)h]$

2) $f(x) = \frac{1}{2hn} I\{x\}$ 에 대하여

(1) $f(a + 2(i-1)h) = [\frac{1}{2hn} I\{-h < x < h\}, \frac{1}{2hn} I\{h < x < 3h\}, \ldots]$