Práctico 4

Estimación de funciones de densidad de la probabilidad.

1. En el análisis del error global de tipo L_2 de un estimador histograma de densidad con ancho de intervalo constante, la expresión analítica del ancho de intervalo óptimo, con respecto al error cuadrático medio integrado asintótico (AMISE), involucra el término
$$R(f') := \int_{-\infty}^{\infty} |f'(x)|^2 dx\,.$$
En las reglas de referencias *normales*, tales como la de Scott y de Freedman-Diaconis, se utiliza la densidad f de una distribución normal con media μ y varianza σ^2. En este caso, mostrar que
$$R(f') = \frac{1}{4\sqrt{\pi}\sigma^3}\,.$$

2. Considere el conjunto de datos *Old Faithful Geyser* que consiste en la duración, en minutos, de 107 erupciones consecutivas del géiser Old Faithful (una fuente termal de la que brota agua caliente y vapor en el Parque Nacional de Yellowstone, Wyoming).
Dichos datos están disponibles en el paquete **R** *locfit*:
library(locfit)
data(package = "locfit")
edit(geyser)

   (a) Realizar un gráfico con seis paneles que muestre las estimaciones de densidad de la duración de las erupciones del géiser utilizando seis funciones núcleo disponibles en **R** y el ancho de ventana $h = 0.4$.
   ¿ Encuentra diferencias significativas entre las estimaciones obtenidas?

   (b) Realizar un gráfico con seis paneles que muestre las estimaciones de densidad de la duración de las erupciones del géiser utilizando los siguientes valores del ancho de ventana

$h = 0.05$; $h = 0.1$; $h = 0.2$; $h = 0.4$; $h = 0.6$; $h = 1$ (considere el núcleo Gaussiano).

¿Encuentra diferencias significativas entre los gráficos?

(c) Realizar un gráfico con paneles que muestre las estimaciones núcleo de la densidad de la duración de las erupciones del géiser *Old Faithful* utilizando los criterios de selección para el ancho de ventana disponibles en la función *density* de **R**.

3. (Densidad de Bart Simpson [3], pp.125-126)

Generar 1000 observaciones de una variable aleatoria con función de distribución acumulada

$$F(x) = \frac{1}{2}\Phi(x, 0, 1) + \frac{1}{10}\sum_{i=0}^{4} \Phi\left(x, \frac{i}{2} - 1, \frac{1}{10}\right),\ x \in \mathbb{R}\,,$$

donde $\Phi(x, \mu, \sigma)$ indica la función de distribución acumulada de una variable aleatoria normal con media μ y desviación estándar σ.

(a) Estimar la densidad por núcleos empleando distintas elecciones de la función núcleo K y del ancho de ventana h. Visualizar las estimaciones propuestas.

(b) ¿Cuál es la estimación núcleo que consideran más adecuada? Justificar la respuesta.

4. Sea

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

la estimación núcleo de la densidad f en el punto x con núcleo K y ancho de ventana constante h.

Probar que $AMISE\left\{\widehat{f}_h\right\} = \frac{1}{nh}R(K) + \frac{1}{4}h^4\left[\mu_2(K)\right]^2 R\left(f''\right)$,
donde $R(K) = \int [K(y)]^2 dy$, $\mu_2(K) = \int y^2 K(y) dy$, y
$R\left(f''\right) = \int \left[f''(y)\right]^2 dy$.

(a) Verificar que el ancho de ventana h que minimiza el $AMISE\left\{\widehat{f}_h\right\}$ es

$$h_{AMISE} = \left[\frac{R(K)}{\mu_2(K)^2 R(f'')}\right]^{1/5} n^{-1/5} .$$

(b) Si f es la densidad de una variable aleatoria $X \sim N(\mu, \sigma^2)$, calcular $R(f'')$ y verificar que h_{AMISE} con la regla de referencia normal y núcleo Gaussiano K está dado por

$$h_{AMISE} = 1.06\, \sigma\, n^{-1/5} .$$

5. Considere los datos publicados en ([2], pp.305-306) con la concentración de colesterol en plasma y la concentración de los triglicéridos en plasma (mg/dl) en 371 pacientes con síntoma de dolor en el pecho.
Hallar las estimaciones óptimas del ancho de ventana h de las estimaciones núcleo de las densidades de concentración, empleando los criterios de selección disponibles a través de la función *density* de **R**.

Bibliografía

[1] David W. Scott (2010). Averaged shifted histogram, *WIREs Comp Stat* 2, 160-164.

[2] David W. Scott (2015). Capítulo 5 *Average shifted histograms* del libro *Multivariate Density Estimation*, segunda edición, Wiley.

[3] Larry Wasserman (2006). *All of Nonparametric Statistics*, Springer.