Estadística No Paramétrica II-2023.
Taller Preparatorio Parcial 1.

Ramón Giraldo
Departamento de Estadística
Universidad Nacional de Colombia

1. Simule $n =$ ______ datos de $X \sim N\,(\mu =$ ______, $\sigma =$ ________) usando el siguiente código R:

```
n=
mu=
sigma=
set.seed(n)
datos=rnorm(n, mu, sigma)
```

Estimación histograma de la densidad:

a) Compruebe con los datos simulados que $f_H(x) = \frac{n_j}{2nb}$

b) Sea x el punto medio de un intervalo. Compruebe que

$$\hat{f}_H(x) = \frac{n_j}{2nb} = \frac{1}{nb}\sum_{i=1}^{n}\frac{1}{2}\mathrm{I}_{[-1,1]}\left(\frac{x_i - x}{b}\right)$$

$$\hat{f}_H(x) = \frac{n_j}{nb} = \frac{1}{nb}\sum_{i=1}^{n}\mathrm{I}_{[-\frac{1}{2},\frac{1}{2}]}\left(\frac{x_i - x}{b}\right)$$

c) Dibuje la curva de la densidad $f(x)$. Superponga al gráfico de la densidad la curva de la estimación histograma de la densidad $\hat{f}_H(x)$ (usando la la amplitud óptima) y la curva de $\mathbb{E}(\hat{f}_H(x)) = \frac{P_j}{b}$.

d) Dibuje la curva de $\mathbb{V}(\hat{f}_H(x))$

Estimación Kernel de la densidad:

- Emplee la simulación del punto 1 para dar respuesta a los literales (a hasta g) a continuación. En todos los casos estime la densidad entre el mínimo y el máximo de la muestra con 512 puntos (the number of equally spaced points at which the density is to be estimated).

a) El h de la regla de Silverman es __________.

b) Asumiedo un Kernel Gaussiano $MISE(\hat{f}_k) =$ __________.

c) Usando h de la regla de Silverman, el valor estimado de la densidad en el mínimo de la muestra es $\hat{f}(x) =$ __________. Un IC del 95 % para $\hat{f}(x)$, considerando despreciable el sesgo, en el mínimo de la muestra es [__________, __________]

d) El h de referencia a la normal es ________. Las estimaciones de la media y la varianza son, respectivamente, $\hat{\mu}$ __________ y $\hat{\sigma}$ __________. Con estos valores y asumiendo $x =$ el máximo de la muestra, encuentre:
a) $\hat{E}(\hat{f}_k(x)) =$ __________
b) $\hat{ee}(\hat{f}_k(x)) =$ __________
c) Las bandas de referencia a la normal son [__________, __________]

e) Dibuje las bandas de variabilidad para $E(\hat{f}_k(x))$

f) Dibuje las bandas de referencia a la normal $\left[\hat{E}(\hat{f}_k(x)) \pm 2{,}575\sqrt{\hat{V}(\hat{f}_k(x))}\right]$

g) Haga un test de normalidad usando la función `nise` de la librería `sm` de R.

h) Haga con datos simulados (bajo las hipótesis nula y alterna) un test de igualdad de distribuciones usando la función `sm.density.compare` de R.

i) Reproduzca, usando datos simulados, la Figura 1 del artículo de Sheather (2004).

2. Demuestre que $\mathbb{E}(\hat{f}_H(x)) = f(x) + O(b)$

3. Demuestre que $\mathbb{V}(\hat{f}_H(x)) = \frac{f(x)}{nb} + O(\frac{1}{n})$

4. Sea $X \sim N(\mu, \sigma^2)$. Compruebe que $R(f') = \int_{-\infty}^{\infty} [f'(x)]^2 = \frac{1}{4\sqrt{\pi}\sigma^3} dx$

5. Sea $X \sim N(\mu, \sigma^2)$. Compruebe que $R(f'') = \int_{-\infty}^{\infty} [f''(x)]^2 = \frac{3}{8\sigma^5\sqrt{\pi}}$.

6. Sea $\hat{f}_H(x)$ el estimador histograma de la densidad. Compruebe que la amplitud óptima es $b = 3{,}491\hat{\sigma}n^{-1/3}$

7. Sea $X \sim N(\mu, \sigma^2)$, $K(u)$ un Kernel Gaussiano con varianza σ_K^2, es decir $U \sim N(0, \sigma_K^2)$, y $\hat{f}_K(x)$ el estimador Kernel de la densidad con ancho de banda h. El error cuadrado medio integrado asintótico se define como $AMISE(\hat{f}) = \frac{R(K)}{nh} + \frac{\sigma_k^4 h^4 R(f'')}{4}$. Deduzca h_0 el ancho de banda de referencia a la normal. Siga los siguientes pasos:

   a) Encuentre h_0 que minimiza $AMISE(\hat{f})$
   b) Halle $R(K)$
   c) Encuentre $R(f'')$

d) Reemplace b) y c) en a) y determine el ancho de banda óptimo de referencia a la normal

Sea $\hat{f}_k(x)$ el estimador kernel kernel de la densidad.

8. Demuestre que $\mathbb{E}(\hat{f}_k(x)) \approx f(x) + \frac{f''(x)\sigma_k^2 h^2}{2}$, con $\sigma_k^2 = \int u^2 k(u) du$
9. Demuestre que $\mathbb{V}(\hat{f}_k(x)) \approx \frac{f(x)R(k)}{nh}$ con $R(k) = \int k^2(u) du$.
10. Compruebe que $\text{AMISE}(\hat{f}_k) = \frac{R(k)}{nh} + \sigma_k^4 h^4 R(f'')$
11. Compruebe que el ancho de banda óptimo es $h = \left(\frac{R(k)}{\sigma_k^4 R(f'')n}\right)^{\frac{1}{5}}$.
12. Usando la regla de referencia a la normal compruebe que $h_{\text{óptimo}} = 1{,}059\hat{\sigma}^{-1/5}$, con $\hat{\sigma}$ el mínimo de (RIC, s), con RIC el rango intercuartílico y s la desviación estándar muestral.
13. Sea $X \sim N(\mu, \sigma^2)$, $K(u)$ un Kernel Gaussiano con varianza h^2, es decir $U \sim N(0, h^2)$ y $\hat{f}_K(x)$ el estimador Kernel de la densidad. Demuestre que (ver Bowman and Azzalini, 1997, page 41)

$$\hat{\mathbb{V}}\left(\hat{f}_K(x)\right) = \frac{1}{n}\left[\Phi\left(0; 0, 2h^2\right)\Phi\left(x; \hat{\mu}, \hat{\sigma}^2 + \frac{h^2}{2}\right) - \Phi\left(x; \hat{\mu}, \hat{\sigma}^2 + h^2\right)^2\right]$$

14. Escriba un código R para probar la igualdad de dos densidades $f_X(x)$ y $f_Y(y)$ usando datos simulados bajo dos escenarios: (a) cuando es cierto que $f_X(x) = f_Y(y)$ y (b) cuando es falso que $f_X(x) = f_Y(y)$.
Incluya en el código las librerías requeridas. Verifique que el código escrito entregado no tiene errores (si muestra errores al ejecutarse no será calificado). Escriba comentarios para documentarlo. Fije una semilla para hacer la simulación y diga a partir de los datos simulados (tanto los del escenario (a) como los del (b)) si rechaza la hipótesis de igualdad de densidades y cuál es valor-p de la prueba en cada caso.