

Notas Curso de Estadística II

Maikol Solís Chacón y Luis Barboza Chinchilla

Actualizado el 30 March, 2022

Índice general

1. Introducción	5
2. Estimación no-paramétrica de densidades	7
2.1. Histograma	7
2.1.1. Construcción Estadística	7
2.1.2. Construcción probabilística	9
2.1.3. Propiedades estadísticas	9
2.1.4. Propiedades estadísticas	9
2.1.5. Sesgo	10
2.1.6. Varianza	11
2.1.7. Error cuadrático medio	11
2.1.8. Error cuadrático medio integrado	13
2.1.9. Ancho de banda óptimo para el histograma	13
2.2. Estimación de densidades basada en kernels.	16
2.2.1. Primera construcción	16
2.2.2. Otra construcción	17
2.2.3. Propiedades Estadísticas	21
2.2.4. Sesgo	22
2.2.5. Error cuadrático medio y Error cuadrático medio inte- grado	23

2.2.6.	Ancho de banda óptimo	24
2.2.6.1.	Referencia normal	25
2.2.6.2.	Validación Cruzada	27
2.2.7.	Intervalos de confianza para estimadores de densidad no paramétricos	29
2.3.	Laboratorio	30
2.3.1.	Efecto de distintos Kernels en la estimación	30
2.3.2.	Efecto del ancho de banda en la estimación	33
2.3.3.	Ancho de banda óptimo	37
2.3.4.	Validación cruzada	41
2.3.5.	Temas adicionales	42
2.4.	Ejercicios	47

Capítulo 1

Introducción

Estas son las notas de clase del curso CA0403: Estadística Actuarial II para el primer semestre del 2022.

Capítulo 2

Estimación no-paramétrica de densidades

2.1. Histograma

El histograma es una de las estructuras básicas en estadística y es una herramienta descriptiva que permite visualizar la distribución de los datos sin tener conocimiento previo de los mismos. En esta sección definiremos el histograma más como un estadístico que como una herramienta de visualización de datos.

2.1.1. Construcción Estadística

Suponga que X_1, X_2, \dots, X_n es una muestra independiente que proviene de una distribución desconocida f . En este caso no asumiremos que f tenga alguna forma particular, que permita definirla de manera paramétrica como en el curso anterior.

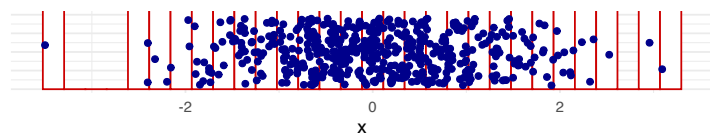
Construcción:

- Seleccione un origen x_0 y divida la línea real en *segmentos*.

$$B_j = [x_0 + (j - 1)h, x_0 + jh), \quad j \in \mathbb{Z}$$

- Cuente cuántas observaciones caen en el segmento B_j . Denótelo como n_j .

8 CAPÍTULO 2. ESTIMACIÓN NO-PARAMÉTRICA DE DENSIDADES



- Divida el número de observaciones en B_j por el tamaño de muestra n y el ancho de banda h de cada caja.

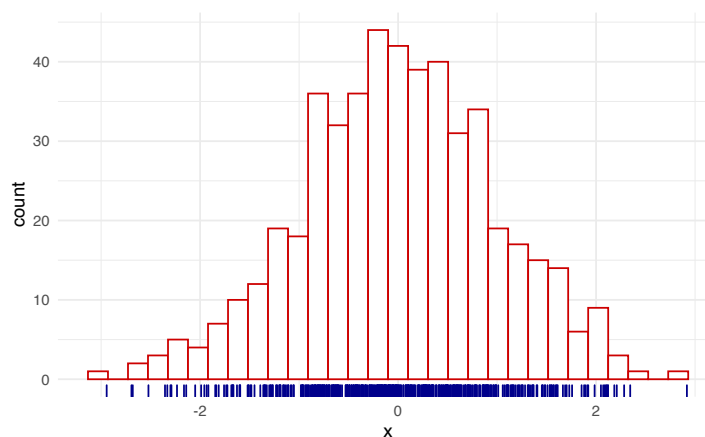
$$f_j = \frac{n_j}{nh}$$

De esta forma si se suma las áreas definidas por el histograma da un total de 1.

- Cuente la frecuencia por el tamaño de muestra n y el ancho de banda h .

$$f_j = \frac{n_j}{nh}$$

- Dibuje el histograma.



Formalmente el histograma es el

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n \sum_j I(X_i \in B_j) I(x \in B_j),$$

donde I es la indicadora.

2.1.2. Construcción probabilística

Denote $m_j = jh - h/2$ el centro del segmento,

$$\begin{aligned}\mathbb{P}\left(X \in \left[m_j - \frac{h}{2}, m_j + \frac{h}{2}\right]\right) &= \int_{m_j - \frac{h}{2}}^{m_j + \frac{h}{2}} f(u) du \\ &\approx f(m_j)h\end{aligned}$$

Otra forma de aproximarlos es:

$$\mathbb{P}\left(X \in \left[m_j - \frac{h}{2}, m_j + \frac{h}{2}\right]\right) \approx \frac{1}{n} \# \left\{ X \in \left[m_j - \frac{h}{2}, m_j + \frac{h}{2}\right] \right\}$$

Acomodando un poco la expresión

$$\hat{f}_h(m_j) = \frac{1}{nh} \# \left\{ X \in \left[m_j - \frac{h}{2}, m_j + \frac{h}{2}\right] \right\}$$

2.1.3. Propiedades estadísticas

Note que el estimador de histograma \hat{f}_h tiende a ser más suave conforme aumenta el ancho de banda h .

2.1.4. Propiedades estadísticas

Suponga que $x_0 = 0$ y que $x \in B_j$ es un punto fijo, entonces el estimador evaluado en x es:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n I(X_i \in B_j)$$

2.1.5. Sesgo

Para calcular el sesgo primero calculamos:

$$\begin{aligned}\mathbb{E} [\hat{f}_h(x)] &= \frac{1}{nh} \sum_{i=1}^n \mathbb{E} [I(X_i \in B_j)] \\ &= \frac{1}{nh} n \mathbb{E} [I(X_i \in B_j)]\end{aligned}$$

donde $I(X_i \in B_j)$ es una variable Bernoulli con valor esperado:

$$\mathbb{E} [I(X_i \in B_j)] = \mathbb{P}(I(X_i \in B_j) = 1) = \int_{(j-1)h}^{jh} f(u) du.$$

Entonces,

$$\mathbb{E} [f_h(x)] = \frac{1}{h} \int_{(j-1)h}^{jh} f(u) du$$

y por lo tanto el sesgo de $\hat{f}_h(x)$ es:

$$Sesgo(\hat{f}_h(x)) = \frac{1}{h} \int_{(j-1)h}^{jh} f(u) du - f(x)$$

Esto se puede aproximar usando Taylor alrededor del centro $m_j = jh - h/2$ de B_j de modo que $f(u) - f(x) \approx f'(m_j)(u - x)$.

$$Sesgo(\hat{f}_h(x)) = \frac{1}{h} \int_{(j-1)h}^{jh} [f(u) - f(x)] du \approx f'(m_j)(m_j - x)$$

Entonces se puede concluir que:

- $\hat{f}_h(x)$ es un estimador sesgado de $f(x)$.
- El sesgo tiende a ser cero cerca del punto medio de B_j .
- El sesgo es creciente con respecto a la pendiente de la verdadera densidad evaluada en el punto medio m_j .

2.1.6. Varianza

Dado que todos los X_i son i.i.d., entonces

$$\begin{aligned}\text{Var}(\hat{f}_h(x)) &= \text{Var}\left(\frac{1}{nh} \sum_{i=1}^n I(X_i \in B_j)\right) \\ &= \frac{1}{n^2 h^2} n \text{Var}(I(X_i \in B_j))\end{aligned}$$

La variable I es una bernoulli con parametro $\int_{(j-1)h}^h f(u)du$ por lo tanto su varianza es el

$$\text{Var}(\hat{f}_h(x)) = \frac{1}{nh^2} \left(\int_{(j-1)h}^h f(u)du \right) \left(1 - \int_{(j-1)h}^h f(u)du \right)$$

Ejercicio 2.1. Usando un desarrollo de Taylor como en la parte anterior, pruebe que:

$$\text{Var}(\hat{f}_h(x)) \approx \frac{1}{nh} f(x)$$

Consecuencias:

- La varianza del estimador es proporcional a $f(x)$.
- La varianza decrece si el ancho de banda h crece.

2.1.7. Error cuadrático medio

El error cuadrático medio del histograma es el

$$\text{MSE}(\hat{f}_h(x)) = \text{E}\left[\left(\hat{f}_h(x) - f(x)\right)^2\right] = \text{Sesgo}^2(\hat{f}_h(x)) + \text{Var}(\hat{f}_h(x)).$$

Ejercicio 2.2. ¿Pueden probar la segunda igualdad de la expresión anterior?

12CAPÍTULO 2. ESTIMACIÓN NO-PARAMÉTRICA DE DENSIDADES

Retomando los términos anteriores se puede comprobar que:

$$\text{MSE}(\hat{f}_h(x)) = \frac{1}{nh}f(x) + f' \left\{ \left(j - \frac{1}{2}\right)h \right\}^2 \left\{ \left(j - \frac{1}{2}\right)h - x \right\}^2 \quad (2.1)$$

$$+o(h) + o\left(\frac{1}{nh}\right) \quad (2.2)$$

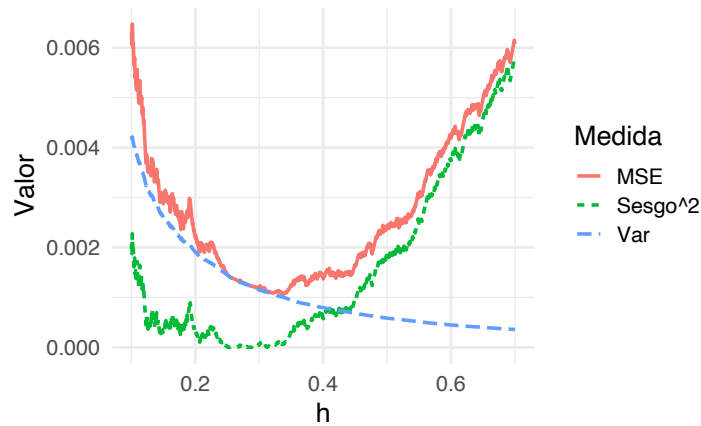
Nota: . Si $h \rightarrow 0$ y $nh \rightarrow \infty$ entonces $\text{MSE}(\hat{f}_h(x)) \rightarrow 0$. Es decir, conforme usamos más observaciones, pero el ancho de banda de banda no decrece tan rápido, entonces el error cuadrático medio converge a 0.

Como $\text{MSE}(\hat{f}_h(x)) \rightarrow 0$ (convergencia en \mathbb{L}^2) implica que $\hat{f}_h(x) \xrightarrow{\mathcal{P}} f(x)$, entonces \hat{f}_h es consistente. Además según la fórmula (2.2), concluimos lo siguiente:

- Si $h \rightarrow 0$, la varianza crece (converge a ∞) y el sesgo decrece (converge a $f'(0)x^2$).
- Si $h \rightarrow \infty$, la varianza decrece (hacia 0) y el sesgo crece (hacia ∞)

Ejercicio 2.3. Si $f \sim N(0, 1)$, aproxime los componentes de sesgo, varianza y MSE, y gráfíquelos para distintos valores de h .

Solución:



2.1.8. Error cuadrático medio integrado

Uno de los problemas con el $\text{MSE}(\hat{f}_h(x))$ es que depende de x y de la función de densidad f (desconocida). Integrando con respecto a x el MSE se logra resolver el primer problema:

$$\begin{aligned}\text{MISE}(\hat{f}_h) &= \text{E} \left[\int_{-\infty}^{\infty} \left\{ \hat{f}_h(x) - f(x) \right\}^2 dx \right] \\ &= \int_{-\infty}^{\infty} \text{E} \left[\left\{ \hat{f}_h(x) - f(x) \right\}^2 \right] dx \\ &= \int_{-\infty}^{\infty} \text{MSE}(\hat{f}_h(x)) dx\end{aligned}$$

Al MISE se le llama error cuadrático medio integrado. Además,

$$\begin{aligned}\text{MISE}(\hat{f}_h) &\approx \int_{-\infty}^{\infty} \frac{1}{nh} f(x) dx \\ &\quad + \int_{-\infty}^{\infty} \sum_j I(x \in B_j) \left\{ \left(j - \frac{1}{2} \right) h - x \right\}^2 \left[f' \left(\left\{ j - \frac{1}{2} \right\} h \right) \right]^2 dx \\ &= \frac{1}{nh} + \sum_j \left[f' \left(\left\{ j - \frac{1}{2} \right\} h \right) \right]^2 \int_{B_j} \left\{ \left(j - \frac{1}{2} \right) h - x \right\}^2 dx \\ &= \frac{1}{nh} + \frac{h^2}{12} \sum_j \left[f' \left(\left\{ j - \frac{1}{2} \right\} h \right) \right]^2 \\ &\approx \frac{1}{nh} + \frac{h^2}{12} \int \{ f'(x) \}^2 dx \\ &= \frac{1}{nh} + \frac{h^2}{12} \|f'\|_2^2\end{aligned}$$

la cual es una buena aproximación si $h \rightarrow 0$. A este último término se le llama MISE asintótico.

2.1.9. Ancho de banda óptimo para el histograma

El MISE tiene un comportamiento asintótico similar al observado en el MSE. La figura siguiente presenta el comportamiento de la varianza, sesgo y MISE para nuestro ejemplo anterior:

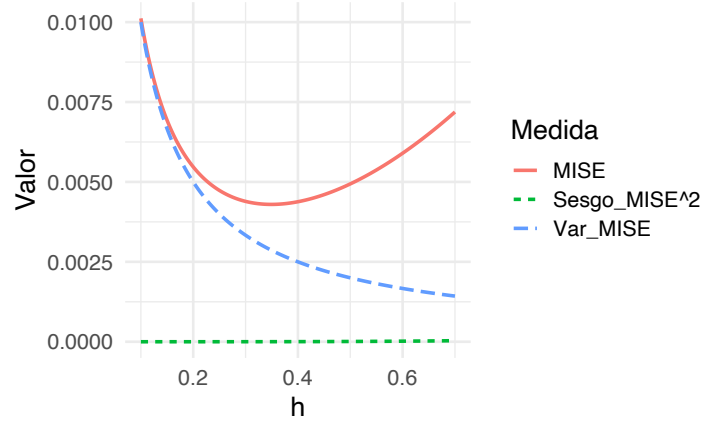
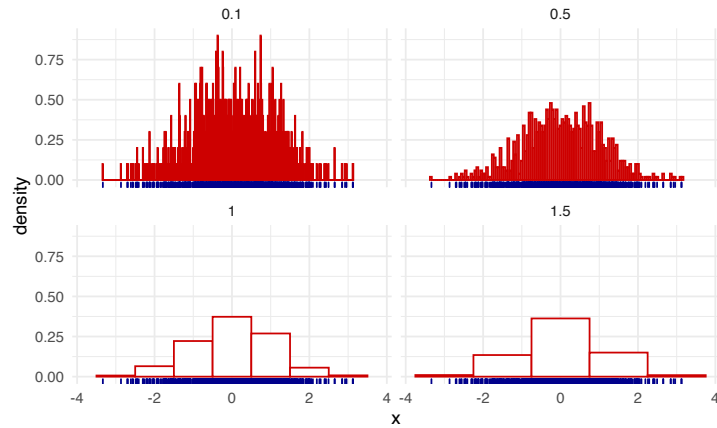


Figura 2.1:

Un problema frecuente en los histogramas es que la mala elección del parámetro h causa que estos no capturen toda la estructura de los datos. Por ejemplo, en el siguiente caso se muestra histogramas construidos a partir de 1000 números aleatorios según una $N(0, 1)$, bajo 4 distintas escogencias de ancho de banda.



Un criterio más preciso para seleccionar el ancho de banda es a través de la minimización del MISE:

$$\frac{\partial \text{MISE}(f_h)}{\partial h} = -\frac{1}{nh^2} + \frac{1}{6}h\|f'\|_2^2 = 0$$

lo implica que

$$h_{opt} = \left(\frac{6}{n \|f'\|_2^2} \right)^{1/3} = O(n^{-1/3}).$$

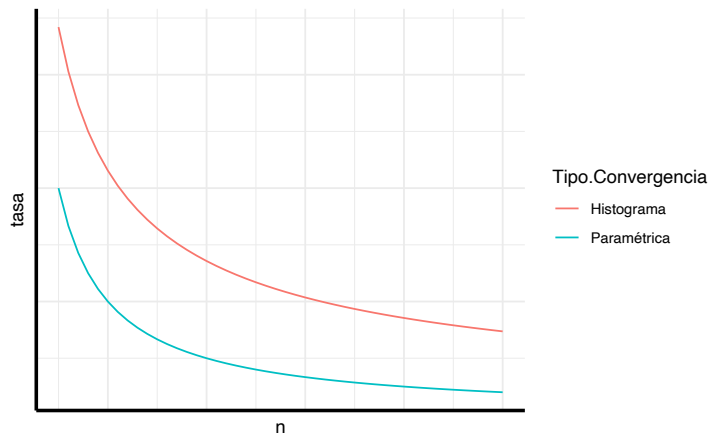
y por lo tanto

$$\text{MISE}(\hat{f}_h) = \frac{1}{n} \left(\frac{n \|f'\|_2^2}{6} \right)^{1/3}$$

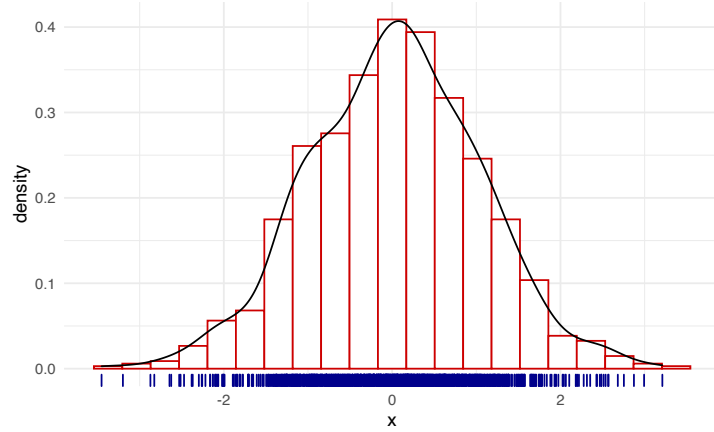
Nota: (Recuerde de Estadística I). Si $X_1, \dots, X_n \sim f_\theta$ i.i.d, con $\text{Var}(X) = \sigma^2$ y media θ , recuerde que el estimador $\hat{\theta}$ de θ tiene la característica que

$$\text{MSE}(\theta) = \text{Var}(\hat{\theta}) + \text{Sesgo}^2(\hat{\theta}) = \frac{\sigma^2}{n}$$

Según la nota anterior la tasas de convergencia del histograma es más lenta que la de un estimador paramétrico considerando la misma cantidad de datos, tal y como se ilustra en el siguiente gráfico:



Finalmente, podemos encontrar el valor óptimo del ancho de banda ($h = 0.3375$) del conjunto de datos en el ejemplo anterior.



Ejercicio 2.4. Verifique que en el caso normal estándar: $h_{opt} \approx 3,5n^{-1/3}$.

2.2. Estimación de densidades basada en kernels.

2.2.1. Primera construcción

Sea X_1, \dots, X_n variables aleatorias i.i.d. con distribución f en \mathbb{R} . La distribución de f es $F(x) = \int_{-\infty}^x f(t)dt$.

La distribución empírica de F es:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x).$$

Por la ley de los grandes números tenemos que $F_n(x) \xrightarrow{c.s.} F(x)$ para todo x en \mathbb{R} , conforme $n \rightarrow \infty$. Entonces, $F_n(x)$ es un estimador consistente de $F(x)$ para todo x in \mathbb{R} .

Nota: ¿Podríamos derivar F_n para encontrar el estimador \hat{f}_n ?

La respuesta es si (más o menos).

Suponga que $h > 0$ tenemos la aproximación

$$f(x) \approx \frac{F(x+h) - F(x-h)}{2h}.$$

Remplazando F por su estimador F_n , defina

$$\hat{f}_n^R(x) = \frac{F_n(x+h) - F_n(x-h)}{2h},$$

donde $\hat{f}_n^R(x)$ es el estimador de *Rosenblatt*.

Podemos describirlo de la forma,

$$\hat{f}_n^R(x) = \frac{1}{2nh} \sum_{i=1}^n I(x-h < X_i \leq x+h) = \frac{1}{nh} \sum_{i=1}^n K_0\left(\frac{X_i - x}{h}\right)$$

con $K_0(u) = \frac{1}{2}I(-1 < u \leq 1)$, lo cual es equivalente al caso del histograma.

2.2.2. Otra construcción

Con el histograma construimos una serie de segmentos fijo B_j y contabamos el número de datos que estaban **contenidos en** B_j

Nota: ¿Qué pasaría si cambiamos la palabra **contenidos** por **alrededor de “x”**?

Suponga que se tienen intervalos de longitud $2h$, es decir, intervalos de la forma $[x-h, x+h)$.

El estimador de histograma se escribe como

$$\hat{f}_h(x) = \frac{1}{2hn} \# \{X_i \in [x-h, x+h)\}.$$

Note que si definimos

$$K(u) = \frac{1}{2}I(|u| \leq 1)$$

con $u = \frac{x-x_i}{h}$, entonces parte del estimador de histograma se puede escribir como:

$$\frac{1}{2} \# \{X_i \in [x-h, x+h)\} = \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) = \sum_{i=1}^n \frac{1}{2} I\left(\left|\frac{x-x_i}{h}\right| \leq 1\right)$$

Finalmente se tendría que

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

18CAPÍTULO 2. ESTIMACIÓN NO-PARAMÉTRICA DE DENSIDADES

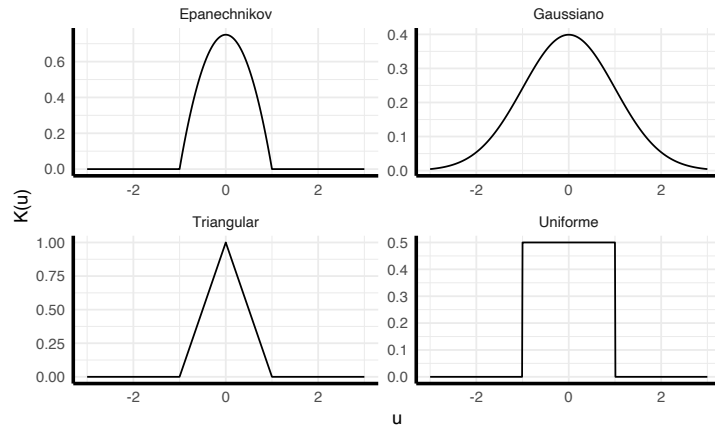
Nota: . ¿Qué pasaría si cambiaríamos la función K del histograma por una más general? Esto permitiría incluir la noción de “cercanía” de cada dato alrededor de x .

Esta función debería cumplir las siguientes características:

- $K(u) \geq 0$.
- $\int_{-\infty}^{\infty} K(u) du = 1$.
- $\int_{-\infty}^{\infty} uK(u) du = 0$.
- $\int_{-\infty}^{\infty} u^2 K(u) du < \infty$.

Por ejemplo:

- **Uniforme:** $\frac{1}{2}I(|u| \leq 1)$.
- **Triangular:** $(1 - |u|)I(|u| \leq 1)$.
- **Epanechnikov:** $\frac{3}{4}(1 - u^2)I(|u| \leq 1)$.
- **Gaussian:** $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$.



Entonces se tendría que la expresión general para un estimador por núcleos (kernel) de f :

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

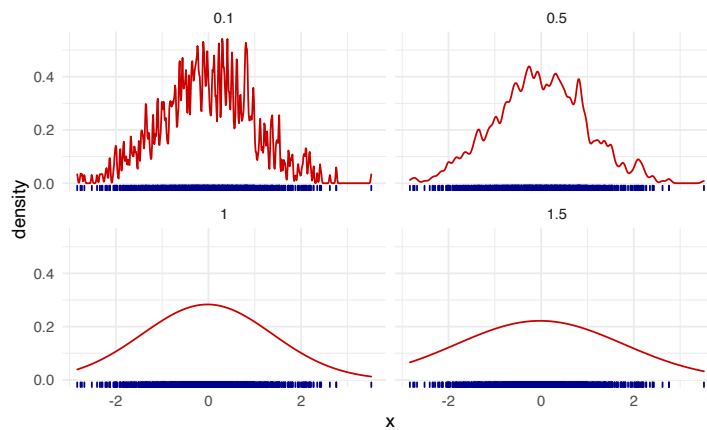
donde x_1, \dots, x_n es una muestra i.i.d. de f ,

$$K_h(\cdot) = \frac{1}{h}K(\cdot/h).$$

y K es un kernel según las 4 propiedades anteriores.

Nota: . ¿Qué pasaría si modificamos el ancho de banda h para un mismo kernel?

Nuevamente controlaríamos la suavidad del estimador a como se ilustra a continuación:

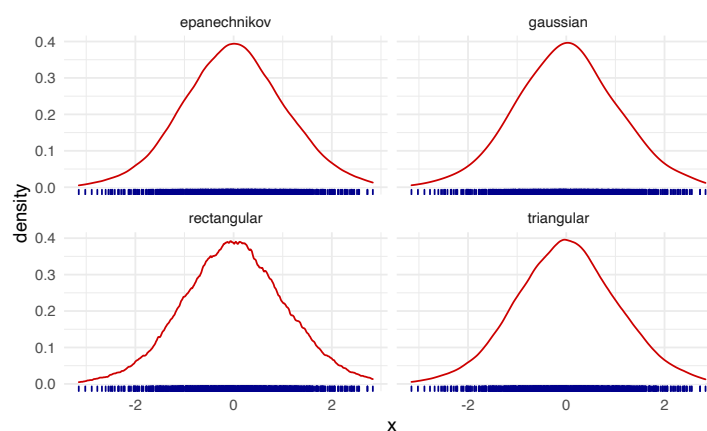


Inconveniente: no tenemos aún un criterio para un h óptimo.

Nota: . ¿Qué pasaría si modificamos el kernel para un mismo ancho de banda h ?

Usando 1000 números aleatorios según una normal estándar, con un ancho de banda fijo ($h = 0,3$) podemos ver que no hay diferencias muy marcadas entre los estimadores por kernel:

20CAPÍTULO 2. ESTIMACIÓN NO-PARAMÉTRICA DE DENSIDADES



Recordemos nuevamente la fórmula

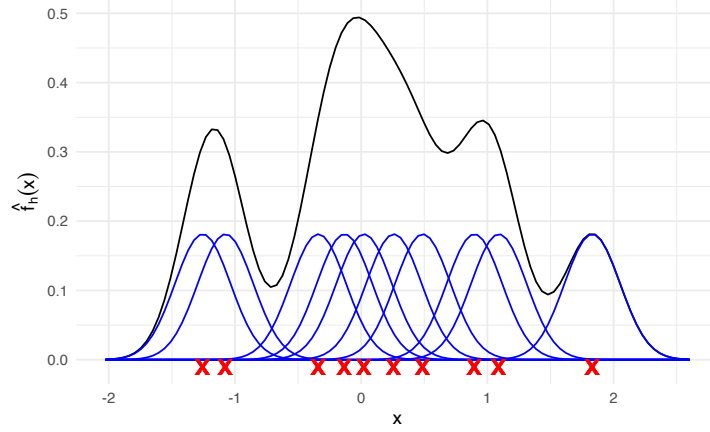
$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

Cada sumando de esta expresión es una función por si misma. Si la integramos se obtiene que

Cada sumando de esta expresión es una función de la variable x . Si la integramos se obtiene que

$$\frac{1}{nh} \int K\left(\frac{x - X_i}{h}\right) dx = \frac{1}{nh} \int K(u) h du = \frac{1}{n} \int K(u) du = \frac{1}{n}$$

En el siguiente gráfico se generan 10 puntos aleatorios según una normal estándar (rojo) y se grafica cada uno de los 10 componentes del estimador de la densidad usando kernels gaussianos (azul). El estimador resultante aparece en color negro. Note que cada uno de los 10 componentes tiene la misma área bajo la curva, la cual en este caso es 0.1.



2.2.3. Propiedades Estadísticas

Al igual que en el caso de histograma, también aplica lo siguiente:

$$\begin{aligned} \text{MSE}(\hat{f}_h(x)) &= \text{Var}(\hat{f}_h(x)) + \text{Sesgo}^2(\hat{f}_h(x)) \\ \text{MISE}(\hat{f}_h) &= \int \text{Var}(\hat{f}_h(x))dx + \int \text{Sesgo}^2(\hat{f}_h(x))dx \end{aligned}$$

donde

$$\text{Var}(\hat{f}_h(x)) = \mathbb{E} \left[\hat{f}_h(x) - \mathbb{E} \hat{f}_h(x) \right]^2 \text{ and } \text{Sesgo}(\hat{f}_h(x)) = \mathbb{E} [\hat{f}_h(x)] - f(x).$$

En el caso de la varianza:

$$\begin{aligned} \text{Var}(\hat{f}_h(x)) &= \text{Var} \left(\frac{1}{n} \sum_{i=1}^n K \left(\frac{x - X_i}{h} \right) \right) \\ &= \frac{1}{n^2 h^2} \sum_{i=1}^n \text{Var} \left(K \left(\frac{x - X_i}{h} \right) \right) \\ &= \frac{1}{n h^2} \text{Var} \left(K \left(\frac{x - X}{h} \right) \right) \\ &= \frac{1}{n h^2} \left\{ \mathbb{E} \left[K^2 \left(\frac{x - X}{h} \right) \right] - \left\{ \mathbb{E} \left[K \left(\frac{x - X}{h} \right) \right] \right\}^2 \right\}. \end{aligned}$$

Usando que:

$$\begin{aligned}
 \mathbb{E} \left[K^2 \left(\frac{x - X}{h} \right) \right] &= \int K^2 \left(\frac{x - s}{h} \right) f(s) ds \\
 &= h \int K^2(u) f(uh + x) du \\
 &= h \int K^2(u) \{f(x) + o(1)\} du \\
 &= h \left\{ \|K\|_2^2 f(x) + o(1) \right\}.
 \end{aligned}$$

$$\begin{aligned}
 \mathbb{E} \left[K \left(\frac{x - X}{h} \right) \right] &= \int K \left(\frac{x - s}{h} \right) f(s) ds \\
 &= h \int K(u) f(uh + x) du \\
 &= h \int K(u) \{f(x) + o(1)\} du \\
 &= h \{f(x) + o(1)\}.
 \end{aligned}$$

Por lo tanto se obtiene que

$$\text{Var} \left(\hat{f}_h(x) \right) = \frac{1}{nh} \|K\|_2^2 f(x) + o \left(\frac{1}{nh} \right), \text{ si } nh \rightarrow \infty.$$

2.2.4. Sesgo

Para el sesgo tenemos

$$\begin{aligned}
 \text{Sesgo} \left(\hat{f}_h(x) \right) &= \mathbb{E} \left[\hat{f}_h(x) \right] - f(x) \\
 &= \frac{1}{nh} \sum_{i=1}^n \mathbb{E} \left[K \left(\frac{x - X_i}{h} \right) \right] - f(x) \\
 &= \frac{1}{h} \mathbb{E} \left[K \left(\frac{x - X_1}{h} \right) \right] - f(x) \\
 &= \int \frac{1}{h} K \left(\frac{x - u}{h} \right) f(u) du - f(x)
 \end{aligned}$$

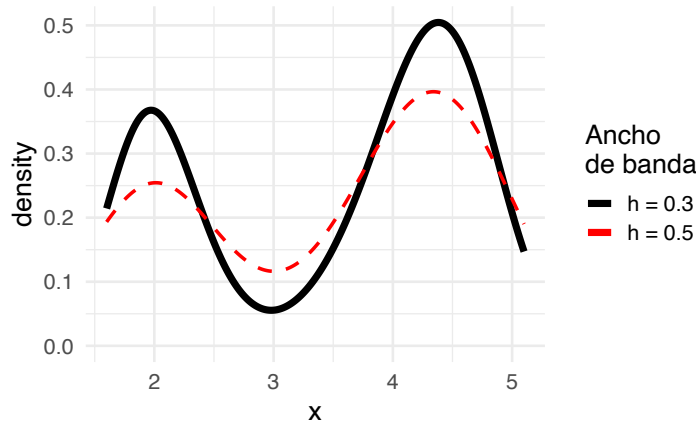
Ejercicio 2.5. Usando el cambio de variable $s = \frac{u-x}{h}$ y las propiedades del kernel pruebe que

$$\text{Sesgo}(\hat{f}_h(x)) = \frac{h^2}{2} f'' \mu_2(K) + o(h^2), \text{ si } h \rightarrow 0$$

donde $\mu_2 = \int s^2 K(s) ds$.

****Nota.**** En algunas pruebas más formales, se necesita además que f'' sea absolutamente continua y que $\int (f'''(x)) dx < \infty$.

En el siguiente gráfico se ilustra el estimador no paramétrico de la distribución de tiempos entre erupciones en la muy conocida tabla de datos *faithful*. El estimador se calcula bajo dos distintas escogencias de ancho de banda.



Nota: . Note como los cambios en el ancho de banda modifican la suavidad (sesgo) y el aplanamiento de la curva (varianza).

2.2.5. Error cuadrático medio y Error cuadrático medio integrado

El error cuadrático medio se escribe

$$\begin{aligned} \text{MSE}(\hat{f}_h(x)) &= \text{Sesgo}(\hat{f}_h(x))^2 + \text{Var}(\hat{f}_h(x)) \\ &= \frac{h^4}{4} (\mu_2(K) f''(x))^2 + \frac{1}{nh} \|K\|_2^2 f(x) + o(h^4) + o\left(\frac{1}{nh}\right). \end{aligned}$$

Y el error cuadrático medio integrado se escribe como,

$$\begin{aligned}\text{MISE}(\hat{f}_h) &= \int \text{MSE}(\hat{f}_h(x)) dx \\ &= \int \text{Sesgo}(\hat{f}_h(x))^2 + \text{Var}(\hat{f}_h(x)) dx \\ &= \frac{h^4}{4} \mu_2^2(K) \|f''(x)\|_2^2 + \frac{1}{nh} \|K\|_2^2 + o(h^4) + o\left(\frac{1}{nh}\right).\end{aligned}$$

Al igual que en el caso del histograma, el estimador por kernels es un estimador consistente de f si $h \rightarrow 0$ y $nh \rightarrow \infty$. Además el MISE depende directamente de f'' .

2.2.6. Ancho de banda óptimo

Minimizando el MISE con respecto a h obtenemos

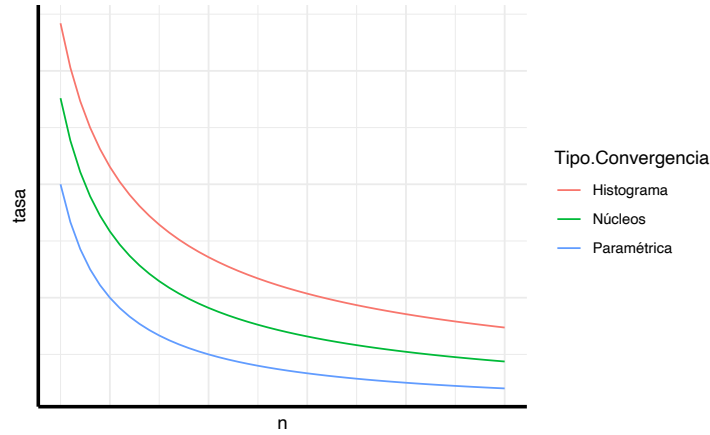
$$h_{opt} = \left(\frac{\|K\|_2^2}{\|f''\|_2^2 (\mu_2(K))^2 n} \right)^{1/5} = O(n^{-1/5}).$$

Nota: . De forma práctica, h_{opt} no es un estimador útil de h porque depende de $\|f''\|_2^2$ que es desconocido. Más adelante veremos otra forma de encontrar este estimador.

Evalutando h_{opt} en el MISE tenemos que

$$\text{MISE}(\hat{f}_h) = \frac{5}{4} (\|K\|_2^2)^{4/5} (\|f''\|_2^2 \mu_2(K))^{2/5} n^{-4/5} = O(n^{-4/5}).$$

y por lo tanto la tasa de convergencia del MISE a 0 es más rápida que para el caso del histograma:



Nota: . Como se comentó anteriormente, el principal inconveniente del ancho de banda:

$$h_{opt} = \left(\frac{\|K\|_2^2}{\|f''\|_2^2 (\mu_2(K))^2 n} \right)^{1/5} = O(n^{-1/5}).$$

es que depende de f'' .

A continuación se explica dos posibles métodos para determinar para aproximar el ancho de banda óptimo:

2.2.6.1. Referencia normal

Nota: . Este método es más efectivo si se conoce que la verdadera distribución es bastante suave, unimodal y simétrica. Más adelante veremos otro método para densidades más generales.

Asuma que f es normal distribuida y se utiliza un kernel K gaussiano. Entonces se tiene que

$$\begin{aligned} \hat{h}_{rn} &= \left(\frac{\|K\|_2^2}{\|f''\|_2^2 (\mu_2(K))^2 n} \right)^{1/5} = O(n^{-1/5}) \\ &= 1,06\hat{\sigma}n^{-1/5}. \end{aligned}$$

donde

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Ejercicio 2.6. Pruebe que la ecuación anterior es verdadera. Utilice el hecho de que:

$$\|f''\|_2^2 = \sigma^{-5} \int \phi''(x)^2 dx$$

donde ϕ es la función de densidad de una $N(0, 1)$.

Nota: . El principal inconveniente de \hat{h}_{rn} es su sensibilidad a los valores extremos:

Ejemplo 2.1. La varianza empírica de 1, 2, 3, 4, 5, es 2.5.

La varianza empírica de 1, 2, 3, 4, 5, 99, es 1538.

Para solucionar el problema anterior, se puede considerar una medida más robusta de variación, por ejemplo el rango intercuantil IQR:

$$\text{IQR}^X = Q_3^X - Q_1^X$$

donde Q_1^X y Q_3^X son el primer y tercer cuartil de un conjunto de datos X_1, \dots, X_n .

Con el supuesto que $X \sim \mathcal{N}(\mu, \sigma^2)$ entonces $Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$ y entonces:

$$\begin{aligned} \text{IQR} &= Q_3^X - Q_1^X \\ &= (\mu + \sigma Q_3^Z) - (\mu + \sigma Q_1^Z) \\ &= \sigma (Q_3^Z - Q_1^Z) \\ &\approx \sigma (0,67 - (-0,67)) \\ &= 1,34\sigma. \end{aligned}$$

Por lo tanto $\hat{\sigma} = \frac{\widehat{\text{IQR}}^X}{1,34}$

Podemos sustituir la varianza empírica de la fórmula inicial y tenemos

$$\hat{h}_{rn} = 1,06 \frac{\widehat{\text{IQR}}^X}{1,34} n^{-\frac{1}{5}} \approx 0,79 \widehat{\text{IQR}}^X n^{-\frac{1}{5}}$$

Combinando ambos estimadores, podemos obtener,

$$\hat{h}_{rn} = 1,06 \min \left\{ \frac{\widehat{\text{IQR}}^X}{1,34}, \hat{\sigma} \right\} n^{-\frac{1}{5}}$$

pero esta aproximación es conveniente bajo el escenario de que la densidad f sea similar a una densidad normal.

2.2.6.2. Validación Cruzada

Defina el *error cuadrático integrado* como

$$\begin{aligned} \text{ISE}(\hat{f}_h) &= \int \left(\hat{f}_h(x) - f(x) \right)^2 dx \\ &= \int \hat{f}_h^2(x) dx - 2 \int \hat{f}_h(x) f(x) dx + \int f^2(x) dx. \end{aligned}$$

Nota: . El MISE es el valor esperado del ISE.

Nuestro objetivo es minimizar el ISE con respecto a h .

Primero note que $\int f^2(x) dx$ NO DEPENDE de h . Podemos minimizar la expresión

$$\text{ISE}(\hat{f}_h) - \int f^2(x) dx = \int \hat{f}_h^2(x) dx - 2 \int \hat{f}_h(x) f(x) dx$$

Vamos a resolver esto en dos pasos partes

Integral $\int \hat{f}_h(x) f(x) dx$

Integral $\int \hat{f}_h(x) f(x) dx$

El término $\int \hat{f}_h(x) f(x) dx$ es el valor esperado de $E[\hat{f}_h(X)]$. Su estimador empírico sería:

$$E[\widehat{f_h(X)}] = \frac{1}{n} \sum_{i=1}^n \hat{f}_h(X_i) = \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n K\left(\frac{X_j - X_i}{h}\right).$$

28CAPÍTULO 2. ESTIMACIÓN NO-PARAMÉTRICA DE DENSIDADES

Nota: . El problema con esta expresión es que las observaciones que se usan para estimar la esperanza son las mismas que se usan para estimar $\hat{f}_h(x)$ (Se utilizan doble).

La solución es remover la $i^{\text{ésima}}$ observación de \hat{f}_h para cada i .

Redefiniendo el estimador anterior tenemos una estimación de $\int \hat{f}_h(x)f(x)dx$ a través de:

$$\frac{1}{n} \sum_{i=1}^n \hat{f}_{h,-i}(X_i),$$

donde (estimador *leave-one-out*)

$$\hat{f}_{h,-i}(x) = \frac{1}{(n-1)h} \sum_{\substack{j=1 \\ j \neq i}}^n K\left(\frac{x - X_j}{h}\right).$$

de esta forma nos aseguramos que las observaciones que se usan para calcular $\hat{f}_{h,-i}(x)$ son independientes de la observación que uno usa para definir el estimador de $E[\hat{f}_h(x)]$.

Siguiendo con el término $\int \hat{f}_h^2(x)dx$ note que este se puede reescribir como

$$\begin{aligned} \int \hat{f}_h^2(x)dx &= \int \left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \right)^2 dx \\ &= \frac{1}{n^2 h^2} \sum_{i=1}^n \sum_{j=1}^n \int K\left(\frac{x - X_i}{h}\right) K\left(\frac{x - X_j}{h}\right) dx \\ &= \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \int K(u) K\left(\frac{X_i - X_j}{h} - u\right) du \\ &= \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n K * K\left(\frac{X_i - X_j}{h}\right). \end{aligned}$$

donde $K * K$ es la convolución de K consigo misma.

Finalmente tenemos la función,

Finalmente definimos la función objetivo del criterio de validación cruzada como:

$$\text{CV}(h) = \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n K * K \left(\frac{X_i - X_j}{h} \right) - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K \left(\frac{X_i - X_j}{h} \right).$$

Nota: . Note que $\text{CV}(h)$ no depende de f o sus derivadas y además la función objetivo se adapta automáticamente a las características de la densidad f .

2.2.7. Intervalos de confianza para estimadores de densidad no paramétricos

Usando los resultados anteriores y asumiendo que $h = cn^{-\frac{1}{5}}$ entonces

$$n^{\frac{2}{5}} \left\{ \hat{f}_h(x) - f(x) \right\} \xrightarrow{\mathcal{L}} \mathcal{N} \left(\underbrace{\frac{c^2}{2} f'' \mu_2(K)}_{b_x}, \underbrace{\frac{1}{c} f(x) \|K\|_2^2}_{v_x} \right).$$

Si $z_{1-\frac{\alpha}{2}}$ es el cuantil $1 - \frac{\alpha}{2}$ de una distribución normal estándar, entonces

$$\begin{aligned} 1 - \alpha &\approx \mathbb{P} \left(b_x - z_{1-\frac{\alpha}{2}} v_x \leq n^{2/5} \left\{ \hat{f}_h(x) - f(x) \right\} \leq b_x + z_{1-\frac{\alpha}{2}} v_x \right) \\ &= \mathbb{P} \left(\hat{f}_h(x) - n^{-2/5} \left\{ b_x + z_{1-\frac{\alpha}{2}} v_x \right\} \right. \\ &\quad \left. \leq f(x) \leq \hat{f}_h(x) - n^{-2/5} \left\{ b_x - z_{1-\frac{\alpha}{2}} v_x \right\} \right) \end{aligned}$$

Esta expresión nos dice que con una probabilidad de $1 - \alpha$ se tiene que

$$\begin{aligned} &\left[\hat{f}_h(x) - \frac{h^2}{2} f''(x) \mu_2(K) - z_{1-\frac{\alpha}{2}} \sqrt{\frac{f(x) \|K\|_2^2}{nh}} \right. \\ &\quad \left. \hat{f}_h(x) - \frac{h^2}{2} f''(x) \mu_2(K) + z_{1-\frac{\alpha}{2}} \sqrt{\frac{f(x) \|K\|_2^2}{nh}} \right] \end{aligned}$$

Al igual que en los casos anteriores, este intervalo no es útil ya que depende de $f(x)$ y $f''(x)$.

Si h es pequeño relativamente a $n^{-\frac{1}{5}}$ entonces el segundo término $\frac{h^2}{2}f''(x)\mu_2(K)$ podría ser ignorado.

Si h es pequeño relativamente a $n^{-\frac{1}{5}}$ entonces el segundo término $\frac{h^2}{2}f''(x)\mu_2(K)$ podría ser ignorado.

Podemos reemplazar $f(x)$ por su estimador $\hat{f}_h(x)$. Entonces tendríamos una intervalo aplicable a nuestro caso:

$$\left[\hat{f}_h(x) - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{f}_h(x)\|K\|_2^2}{nh}}, \hat{f}_h(x) + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{f}_h(x)\|K\|_2^2}{nh}} \right]$$

Nota: . Este intervalo de confianza está definido para x fijo y no permite hacer inferencia sobre toda la función f . Una forma de determinar la banda de confianza de toda la función f es a través de la fórmula 3.52 en la página 62 de [HardleNonparametric2004].

2.3. Laboratorio

Comenzaremos con una librería bastante básica llamada KernSmooth.

2.3.1. Efecto de distintos Kernels en la estimación

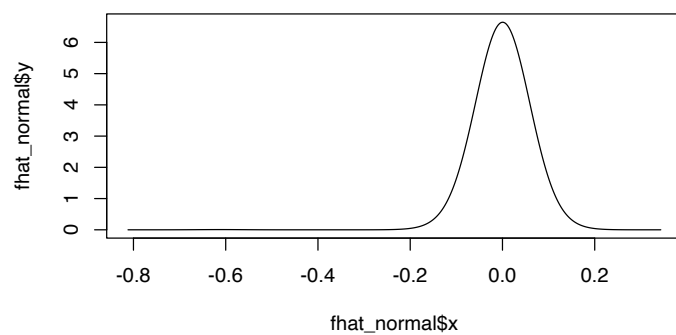
```
x <- read.csv("data/stockres.txt")
x <- unlist(x)
```

```
summary(x)
```

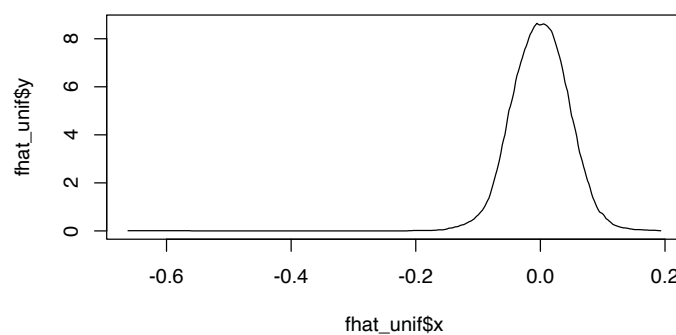
```
##           Min.        1st Qu.         Median          Mean        3rd Qu.          Max.
## -0.6118200 -0.0204085 -0.0010632 -0.0004988  0.0215999  0.1432286
```

```
library(KernSmooth)
```

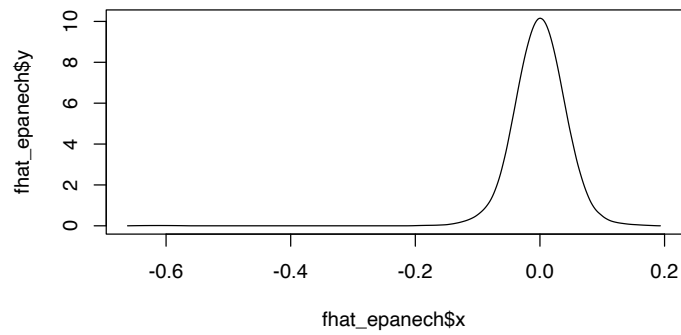
```
fhat_normal <- bkde(x, kernel = "normal", bandwidth = 0.05)
plot(fhat_normal, type = "l")
```



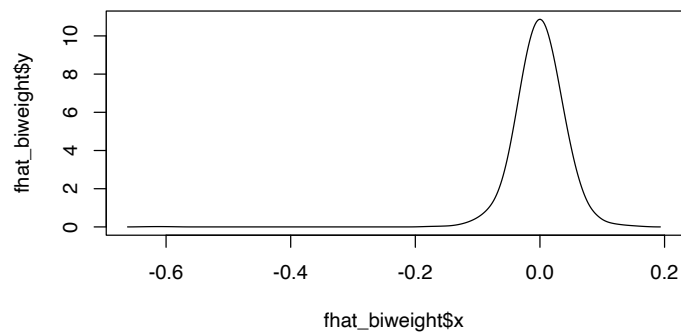
```
fhat_unif <- bkde(x, kernel = "box", bandwidth = 0.05)
plot(fhat_unif, type = "l")
```



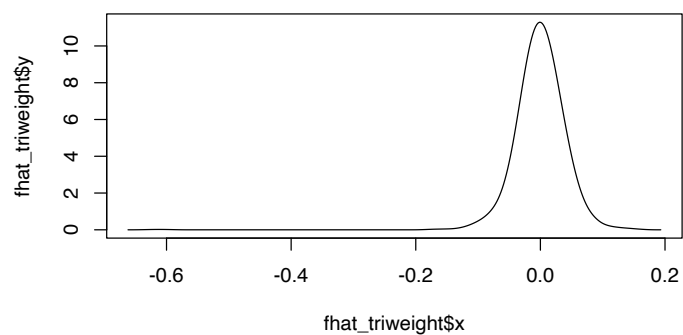
```
fhat_epanech <- bkde(x, kernel = "epanech", bandwidth = 0.05)
plot(fhat_epanech, type = "l")
```



```
fhat_biweight <- bkde(x, kernel = "biweight", bandwidth = 0.05)
plot(fhat_biweight, type = "l")
```



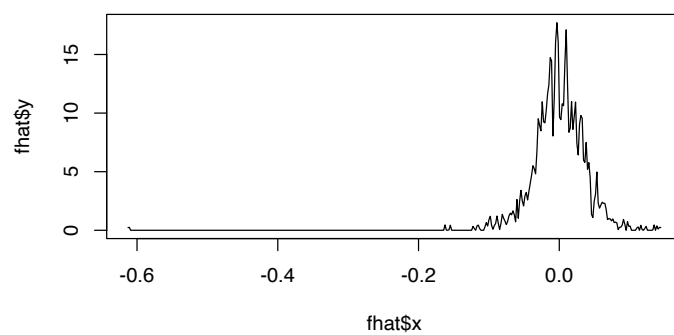
```
fhat_triweight <- bkde(x, kernel = "triweight", bandwidth = 0.05)
plot(fhat_triweight, type = "l")
```

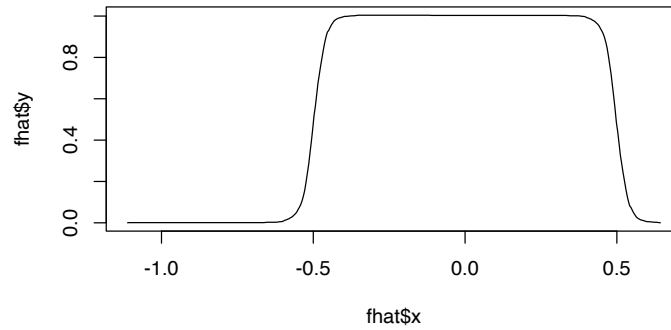
2.3.2. Efecto del ancho de banda en la estimación

**** Kernel uniforme ****

```
fhat <- bkde(x, kernel = "box", bandwidth = 0.001)
plot(fhat, type = "l")
```

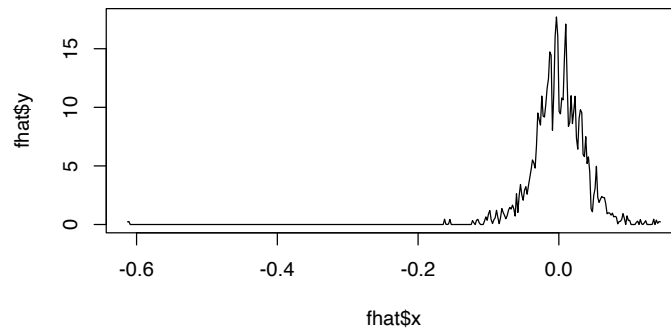


```
fhat <- bkde(x, kernel = "box", bandwidth = 0.5)
plot(fhat, type = "l")
```

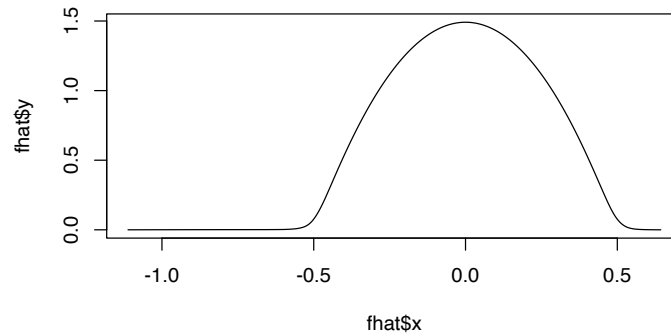


** Kernel Epanechnikov **

```
fhat <- bkde(x, kernel = "epa", bandwidth = 0.001)
plot(fhat, type = "l")
```



```
fhat <- bkde(x, kernel = "epa", bandwidth = 0.5)
plot(fhat, type = "l")
```



```
suppressMessages(library(tidyverse))
library(gganimate)

fani <- tibble()

for (b in seq(0.001, 0.02, length.out = 40)) {
  f <- bkde(x, kernel = "epa", bandwidth = b, gridsize = length(x))
  fani <- fani %>%
    bind_rows(tibble(xreal = sort(x), x = f$x,
                     y = f$y, bw = b))
}

ggplot(data = fani) + geom_line(aes(x, y), color = "blue") +
  labs(title = paste0("Ancho de banda = {closest_state}")) +
  transition_states(bw) + view_follow() + theme_minimal(base_size = 20)

# anim_save('manual_figure/bandwidth-animation.gif')
```

- Nota:* .
- Construya una variable llamada `u` que sea una secuencia de -0.15 a 0.15 con un paso de 0.01
 - Asigne `x` a los datos `stockrel` y calcule su media y varianza.
 - Usando la función `dnorm` construya los valores de la distribución de los datos usando la media y varianza calculada anteriormente. Asigne a esta variable `f\param`.
 - Defina un ancho de banda `h` en 0.02

36CAPÍTULO 2. ESTIMACIÓN NO-PARAMÉTRICA DE DENSIDADES

- Construya un histograma para estos datos con ancho de banda h . Llame a esta variable `f_hist`
- Usando el paquete `KernSmooth` y la función `bkde`, construya una función que calcule el estimador no paramétrico con un núcleo Epanechivok para un ancho de banda h . Llame a esta variable `f_epa`.
- Dibuje en el mismo gráfico la estimación paramétrica y no paramétrica.

```
x <- read.csv("data/stockres.txt")
x <- unlist(x)
# Eliminar nombres de las columnas
names(x) <- NULL

u <- seq(-0.15, 0.15, by = 0.01)

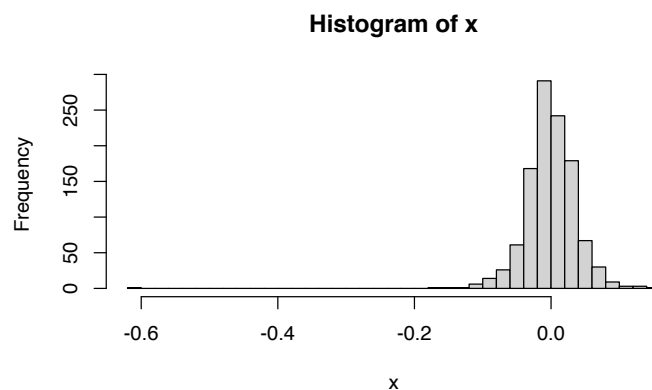
mu <- mean(x)
sigma <- sd(x)

f_param <- dnorm(u, mean = mu, sd = sigma)

h <- 0.02

n_bins <- floor(diff(range(x))/h)

f_hist <- hist(x, breaks = n_bins)
```

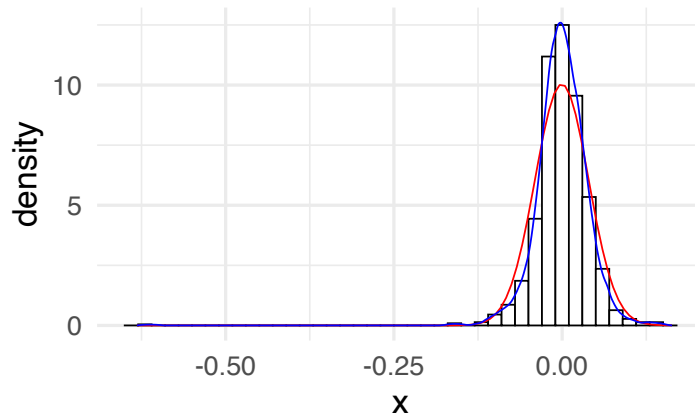


```
f_epa <- as.data.frame(bkde(x, kernel = "epa", bandwidth = h))

x_df <- data.frame(x)

library(ggplot2)

ggplot(x_df, aes(x)) + geom_histogram(aes(y = ..density..),
  binwidth = 0.02, col = "black", fill = "white") +
  stat_function(fun = dnorm, args = list(mean = mu,
    sd = sigma), color = "red") + geom_line(data = f_epa,
    aes(x, y), color = "blue") + theme_minimal(base_size = 20)
```



2.3.3. Ancho de banda óptimo

Usemos la regla de la normal o también conocida como Silverman. **Primero recuerde que en este caso se asume que $f(x)$ sigue una distribución normal.** En este caso, lo que se obtiene es que

$$\begin{aligned}\|f''\|_2^2 &= \sigma^{-5} \int \{\phi''\}^2 dx \\ &= \sigma^{-5} \frac{3}{8\sqrt{\pi}} \approx 0,212\sigma^{-5}\end{aligned}$$

38CAPÍTULO 2. ESTIMACIÓN NO-PARAMÉTRICA DE DENSIDADES

donde ϕ es la densidad de una normal estándar.

El estimador para σ es

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Y usando el cálculo realizado anteriormente, se obtiene que

$$h_{normal} = \left(\frac{4s^5}{3n} \right)^{1/5} \approx 1,06sn^{-1/5}.$$

Un estimador más robusto es

$$h_{normal} = 1,06 \min \left\{ s, \frac{IQR}{1,34} \right\} n^{-1/5}.$$

¿Por qué es $IQR/1,34$?

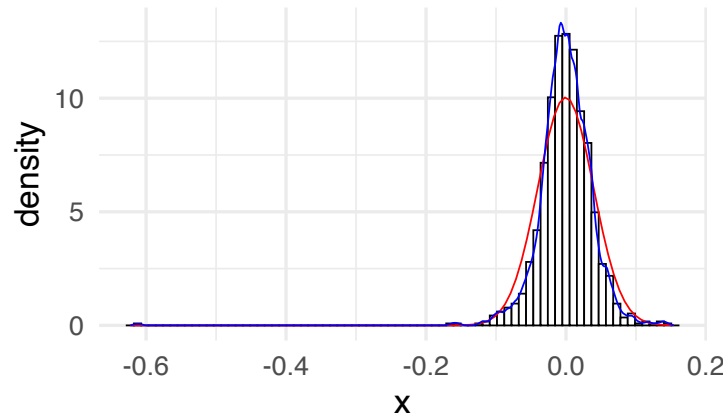
```
s <- sd(x)
n <- length(x)

h_normal <- 1.06 * s * n^(-1/5)

h <- h_normal

n_bins <- floor(diff(range(x))/h)
f_hist <- hist(x, breaks = n_bins, plot = FALSE)
f_epa <- as.data.frame(bkde(x, kernel = "epa", bandwidth = h))

ggplot(x_df, aes(x)) + geom_histogram(aes(y = ..density..),
  binwidth = h, col = "black", fill = "white") +
  stat_function(fun = dnorm, args = list(mean = mu,
    sd = sigma), color = "red") + geom_line(data = f_epa,
  aes(x, y), color = "blue") + theme_minimal(base_size = 20)
```

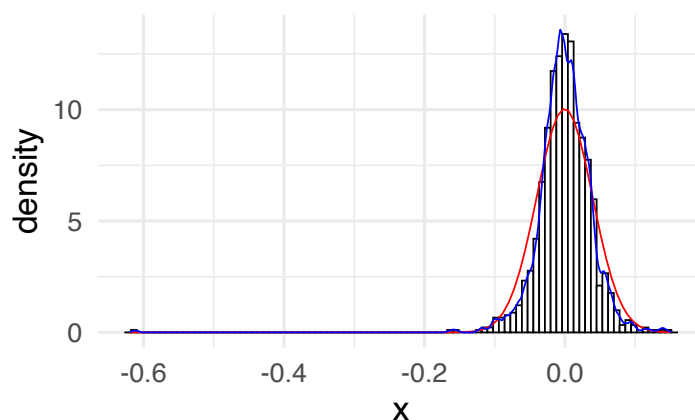


```
h_iqr <- 1.06 * min(s, IQR(x)/1.34) * n^(-1/5)

h <- h_iqr

n_bins <- floor(diff(range(x))/h)
f_hist <- hist(x, breaks = n_bins, plot = FALSE)
f_epa <- as.data.frame(bkde(x, kernel = "epa", bandwidth = h))

ggplot(x_df, aes(x)) + geom_histogram(aes(y = ..density..),
  binwidth = h, col = "black", fill = "white") +
  stat_function(fun = dnorm, args = list(mean = mu,
    sd = sigma), color = "red") + geom_line(data = f_epa,
    aes(x, y), color = "blue") + theme_minimal(base_size = 20)
```



Una librería más especializada es `np` (non-parametric).

```
library(np)

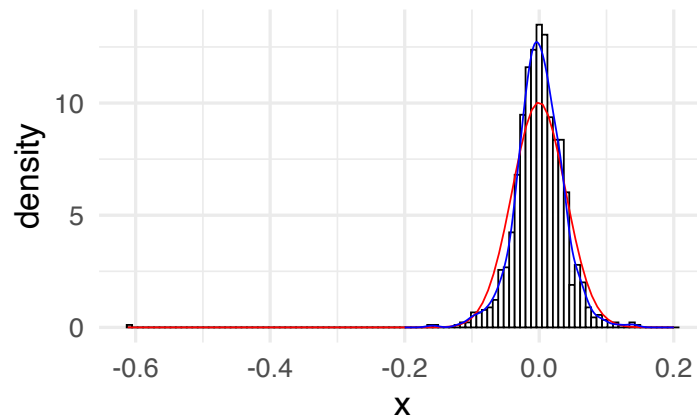
x.eval <- seq(-0.2, 0.2, length.out = 200)

h_normal_np <- npudensbw(dat = x, bwmethod = "normal-reference")

dens.ksum <- npksum(txdat = x, exdat = x.eval, bws = h_normal_np$bw)$ksum/(n *
  h_normal_np$bw[1])

dens.ksum.df <- data.frame(x = x.eval, y = dens.ksum)

ggplot(x_df, aes(x)) + geom_histogram(aes(y = ..density..),
  binwidth = h_normal_np$bw, col = "black", fill = "white") +
  stat_function(fun = dnorm, args = list(mean = mu,
    sd = sigma), color = "red") + geom_line(data = dens.ksum.df,
    aes(x, y), color = "blue") + theme_minimal(base_size = 20)
```

2.3.4. Validación cruzada

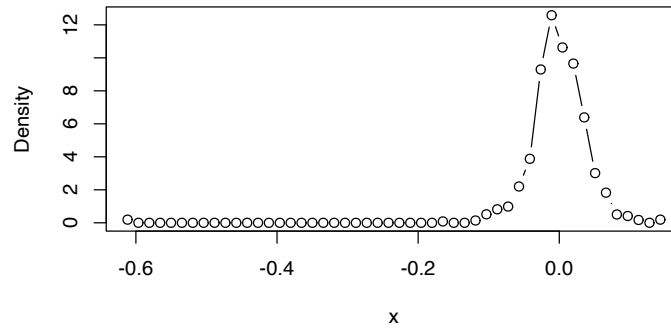
La forma que vimos en clase es la de validación cruzada por mínimos cuadrados “least-square cross validation” la cual se puede ejecutar con este comando.

```
h_cv_np_ls <- npudensbw(dat = x, bwmethod = "cv.ls",  
  ckertype = "epa", ckerorder = 2)
```

```
## Multistart 1 of 1 |Multistart 1 of 1 |Multistart 1 of 1 |Multistart 1 of 1 /Multis
```

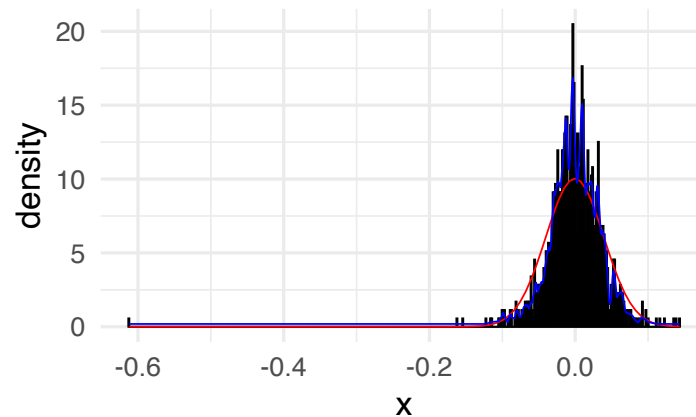
```
dens.np <- npudens(h_cv_np_ls)
```

```
plot(dens.np, type = "b")
```



```
dens.np.df <- data.frame(x = dens.np$eval[, 1], y = dens.np$dens)

ggplot(x_df, aes(x)) + geom_histogram(aes(y = ..density..),
  binwidth = h_cv_np_ls$bw, col = "black", fill = "white") +
  stat_function(fun = dnorm, args = list(mean = mu,
    sd = sigma), color = "red") + geom_line(data = dens.np.df,
    aes(x, y), color = "blue") + theme_minimal(base_size = 20)
```



2.3.5. Temas adicionales

**** Reducción del sesgo **** Como lo mencionamos en el texto, una forma de mejorar el sesgo en la estimación es suponer que la función de densidad es

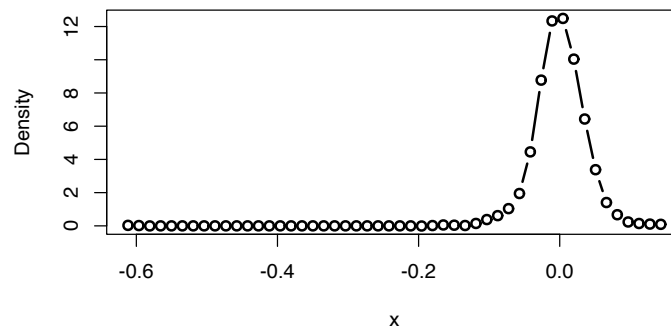
más veces diferenciable.

Esto se logra asumiendo que el Kernel es más veces diferenciable.

```
h_cv_np_ls <- npudensbw(dat = x, bwmethod = "cv.ls",
  ckertype = "epa", ckerorder = 4)
```

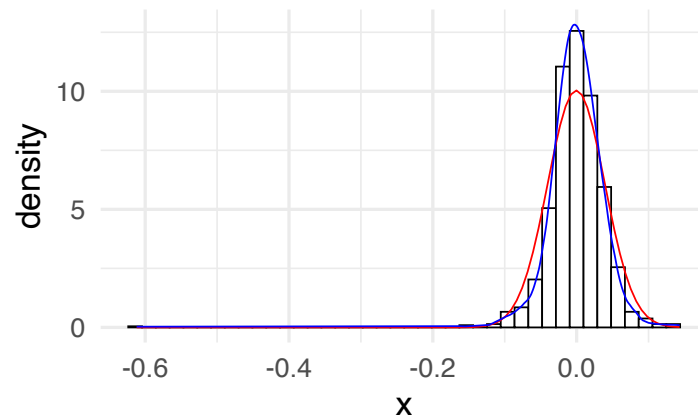
```
## Multistart 1 of 1 |Multistart 1 of 1 |Multistart 1 of 1 |Multistart 1 of 1 /Multis
```

```
dens.np <- npudens(h_cv_np_ls)
plot(dens.np, type = "b", lwd = 2)
```



```
dens.np.df <- data.frame(x = dens.np$eval[, 1], y = dens.np$dens)

ggplot(x_df, aes(x)) + geom_histogram(aes(y = ..density..),
  binwidth = h_cv_np_ls$bw, col = "black", fill = "white") +
  stat_function(fun = dnorm, args = list(mean = mu,
    sd = sigma), color = "red") + geom_line(data = dens.np.df,
  aes(x, y), color = "blue") + theme_minimal(base_size = 20)
```



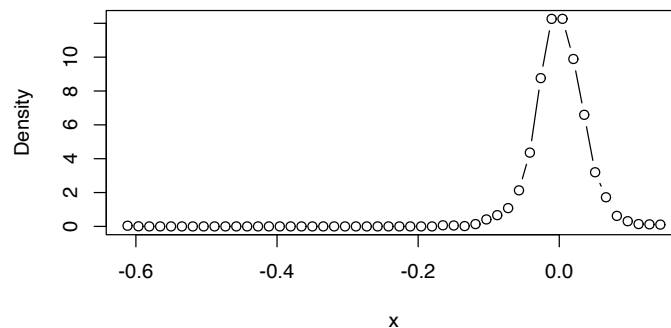
Otra forma de estimar el ancho de banda Otra forma de estimar ancho de bandas óptimos es usando máxima verosimilitud. Les dejo de tarea revisar la sección 1.1 del artículo de [Hall1987] para entender su estructura.

```
h_cv_np_ml <- npudensbw(dat = x, bwmethod = "cv.ml",
  ckertype = "epanechnikov")
```

```
## Multistart 1 of 1 |Multistart 1 of 1 |Multistart 1 of 1 |Multistart 1 of 1
```

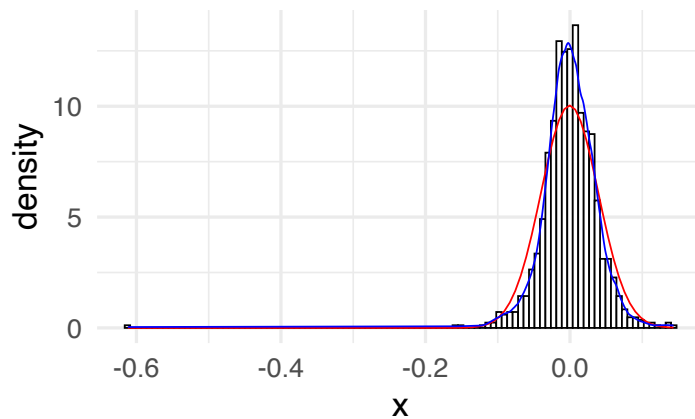
```
dens.np <- npudens(h_cv_np_ml)
```

```
plot(dens.np, type = "b")
```



```
dens.np.df <- data.frame(x = dens.np$eval[, 1], y = dens.np$dens)

ggplot(x_df, aes(x)) + geom_histogram(aes(y = ..density..),
  binwidth = h_cv_np_ml$bw, col = "black", fill = "white") +
  stat_function(fun = dnorm, args = list(mean = mu,
    sd = sigma), color = "red") + geom_line(data = dens.np.df,
    aes(x, y), color = "blue") + theme_minimal(base_size = 20)
```

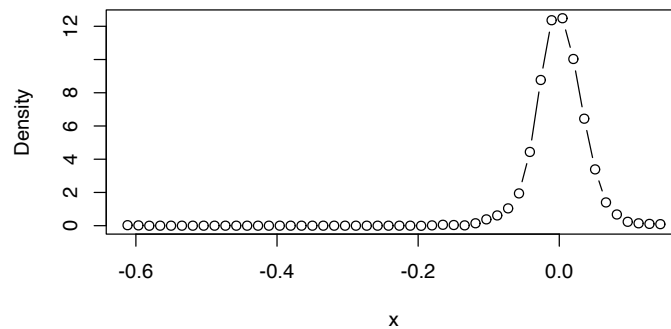


```
h_cv_np_ml <- npudensbw(dat = x, bwmethod = "cv.ml",
  ckertype = "epanechnikov", ckerorder = 4)
```

```
## Multistart 1 of 1 |Multistart 1 of 1 |Multistart 1 of 1 |Multistart 1 of 1 /Multis
```

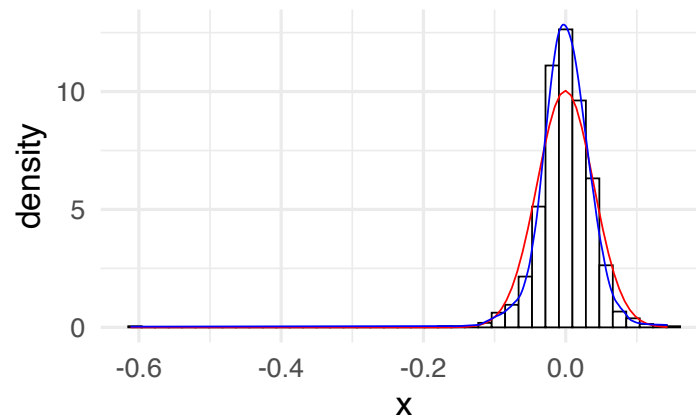
```
dens.np <- npudens(h_cv_np_ml)
```

```
plot(dens.np, type = "b")
```



```
dens.np.df <- data.frame(x = dens.np$eval[, 1], y = dens.np$dens)

ggplot(x_df, aes(x)) + geom_histogram(aes(y = ..density..),
  binwidth = h_cv_np_ml$bw, col = "black", fill = "white") +
  stat_function(fun = dnorm, args = list(mean = mu,
    sd = sigma), color = "red") + geom_line(data = dens.np.df,
    aes(x, y), color = "blue") + theme_minimal(base_size = 20)
```



```
fani <- tibble()

for (b in seq(0.001, 0.05, length.out = 40)) {
```

```

f <- npudens(tdat = x, ckertype = "epanechnikov",
             bandwidth.compute = FALSE, bws = b)
fani <- fani %>%
  bind_rows(tibble(xreal = sort(x), x = f$eval$x,
                   y = f$dens, bw = b))
}

ggplot(data = fani) + geom_line(aes(x, y), color = "blue") +
  labs(title = paste0("Ancho de banda = {closest_state}")) +
  theme_minimal(base_size = 20) + transition_states(bw) +
  view_follow()

# anim_save('manual_figure/bandwidth-animation-np.gif')

```

Ejercicio 2.7. Implementar el intervalo confianza visto en clase para estimadores de densidades por núcleos y visualizarlo de en ggplot.

Si se atreven: ¿Se podría hacer una versión animada de ese gráfico para visualizar el significado real de este el intervalo de confianza?

2.4. Ejercicios

Del libro de [HardleNonparametric2004] hagan los siguientes ejercicios

1. **Sección 2:** 1, 2, 3, 5, 7, 14
2. **Sección 3:** 4, 8, 10, 11, 16,