Notas del curso CA-403

2

Instrucciones de uso

Este es un archivo de LaTeXnormal, salvo que se le pueden agregar códigos de R.

Para hacerlo solo debe encerrar su código con lo siguiente comandos.

```
<<>>=
   Su código acá.
@
```

Algunas recomendaciones iniciales:

- Tratemos de ser ordenados con el texto y el código. Recuerden que esto será usado por ustedes en el examen.
- No usen comandos propios (`newcommand`) ya que eso solo haría más difícil que los compañeros puedan editar su trabajo.
- El documento es colaborativo, por lo que está bien editar o escribir "encima" de otro compañero, siempre y cuando esto sea para mejorar el texto.

Creo en el buen juicio de cada uno para hacer estas notas lo mejor posible.

Para entender mejor la estructura de estas notas, pueden revisar este enlace `https://www.overleaf.com/learn/latex/Knitr`

Capítulo 1

Estimación de densidades

1.1. Histograma

El histograma es una de las estructuras básicas en estadística. Básicamente con este objeto se puede visualizar la distribución de los datos sin tener conocimiento previo de los mismos.

1.1.1. Construcción Estadística

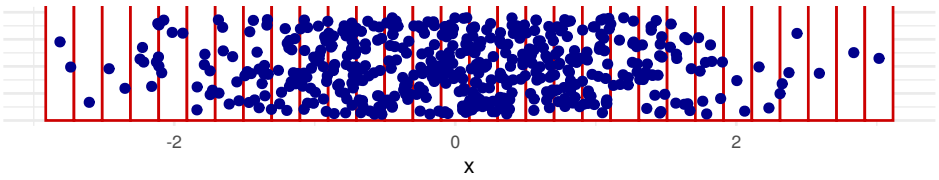
Suponga que $X_1, X_2, \ldots, X_n$ proviene de una distribución desconocida.

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

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

- Cuente cuántas observaciones caen en cada segmento. n_j.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

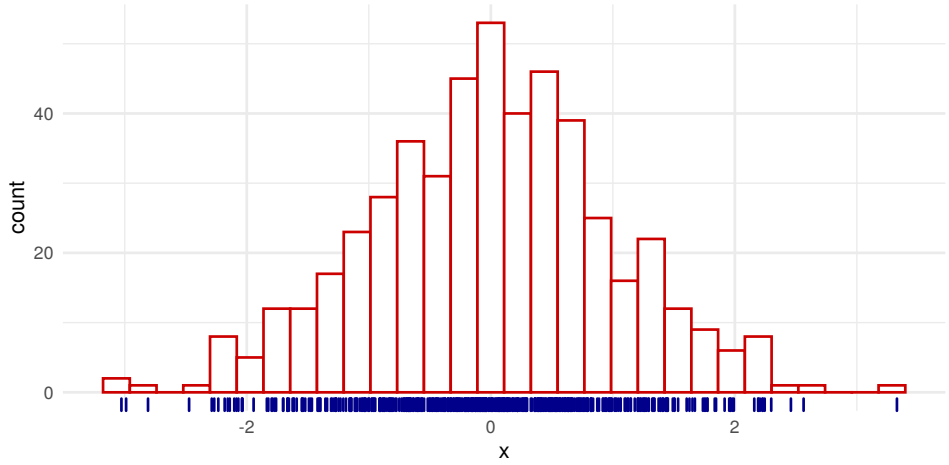


- Cuente la frecuencia por el tamaño de muestra n y el ancho de banda h.

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

- Dibuje el histograma.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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.

1.1.2. Construcción probabilistica

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

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

Esto se puede aproximar como

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

Acomodando un poco la expresión

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

1.1.3. Propiedades estadísticas

Suponga que $x_0 = 0$ y que $x \in B_j$ fijo, entonces

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

Sesgo

El cálculo del sesgo es el

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

$I(X_i \in B_j)$ es una indicadora con probabilidad de 1 de $\int_{(j-1)h}^{jh} f(u)du$ y 0 sino.

Entonces

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

Entonces,

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

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

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

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

Varianza

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

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

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

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

Tarea 1.1.1

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

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

1.1.4. Error cuadrático medio

El error cuadrático medio del histograma es el

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

Tarea 1.1.2

¿Pueden probar la segunda igualdad de la expresión anterior?

Solución 1.1.3

Prueba segunda igualdad:

$$\begin{aligned}
&\text{Sesgo}^2\left(\hat{f}_h(x)\right) + \text{Var}\left(\hat{f}_h(x)\right) = \\
&\left[E\left(\hat{f}_h(x)\right) - f(x)\right]^2 + E\left[\left(E\left(\hat{f}_h(x)\right) - \hat{f}_h(x)\right)^2\right] = \\
&E\left[\left[E\left(\hat{f}_h(x)\right) - f(x)\right]^2 + \left(E\left(\hat{f}_h(x)\right) - \hat{f}_h(x)\right)^2\right] \quad (*)
\end{aligned}$$

Ahora note que:

$$\begin{aligned}
&E\left[\left(E\left(\hat{f}_h(x)\right) - f(x)\right)\left(E\left(\hat{f}_h(x)\right) - \hat{f}_h(x)\right)\right] = \\
&E\left[E\left(\hat{f}_h(x)\right)^2\right] - E\left[E\left(\hat{f}_h(x)\right) \cdot \hat{f}_h(x)\right] - E\left[f(x) \cdot E\left(\hat{f}_h(x)\right)\right] + \\
&E\left[f(x) \cdot \hat{f}_h(x)\right] = \\
&E\left(\hat{f}_h(x)\right)^2 - E\left(\hat{f}_h(x)\right)^2 - E\left(\hat{f}_h(x)\right) \cdot E\left(f(x)\right) + E\left(f(x)\right) \cdot E\left(\hat{f}_h(x)\right) \\
&= 0
\end{aligned}$$

Entonces:

$$(*) = E\left[\left[E\left(\hat{f}_h(x)\right) - f(x)\right]^2 - 2\left(E\left(\hat{f}_h(x)\right) - f(x)\right)\left(E\left(\hat{f}_h(x)\right) - \hat{f}_h(x)\right) + \left(E\left(\hat{f}_h(x)\right) - \hat{f}_h(x)\right)^2\right] =$$
$$E\left[\left(E\left(\hat{f}_h(x)\right) - f(x) - E\left(\hat{f}_h(x)\right) + \hat{f}_h(x)\right)^2\right] =$$
$$E\left[\left(\hat{f}_h(x) - f(x)\right)^2\right]$$

□

Retomando los términos anteriores se tiene que

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

Nota 1.1.4

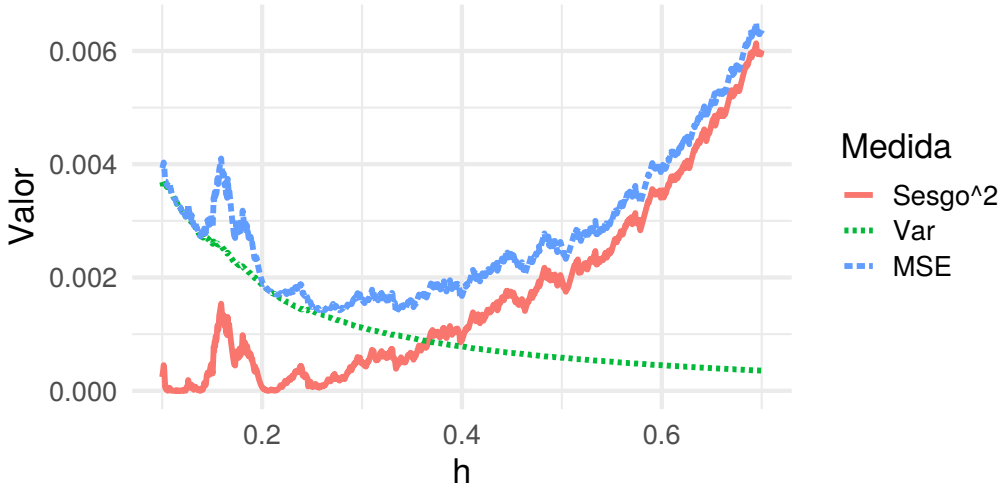
Si $h \to 0$ y $nh \to \infty$ entonces $\text{MSE}\left(\hat{f}_h(x)\right) \to 0$. Es decir, conforme usamos más observaciones, pero el ancho de banda de banda no decrece tan rápida, entonces el error cuadrático medio converge a 0. Esto indica que si $\text{MSE}\left(\hat{f}_h(x)\right) \to 0$ (convergencia en $\mathbb{L}^2$) implica que $\hat{f}_h(x) \xrightarrow{\mathcal{P}} f(x)$, por lo tanto $\hat{f}_h$ es consistente.

La fórmula anterior tiene la siguiente particularidad

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

- Si $h \to \infty$, la varianza decrece (hacia 0) y el sesgo crece (hacia ∞)

Note que la figura **??**



1.1.5. Error cuadrático medio integrado

El problema con el MSE $\left(\hat{f}_h(x)\right)$ es que depende completamente del punto escogido x.

La solución a esto es integrar el MSE.

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

Además,

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

1.1.6. Ancho de banda óptimo para el histograma

El MISE tiene el mismo comportamiento que el MSE. Figura 1.1 presenta el comportamiento de la varianza, sesgo y MISE para nuestro ejemplo.

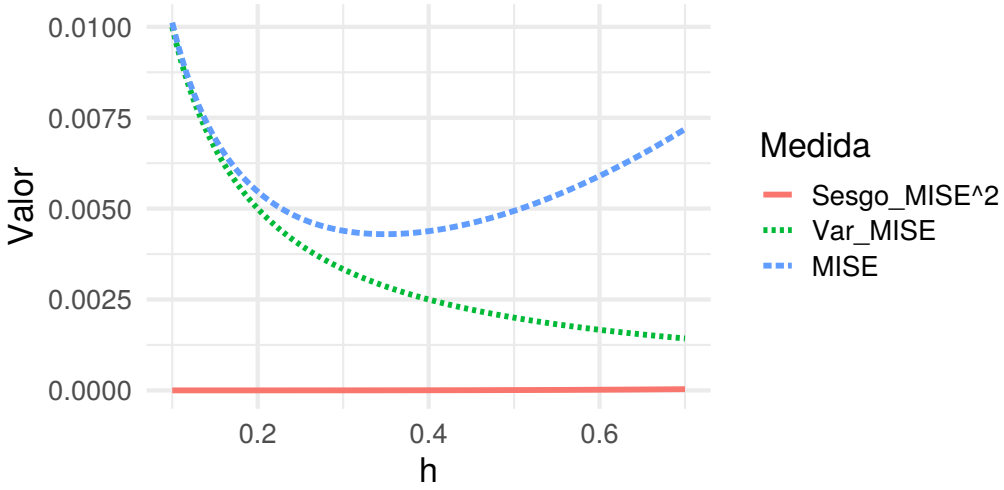
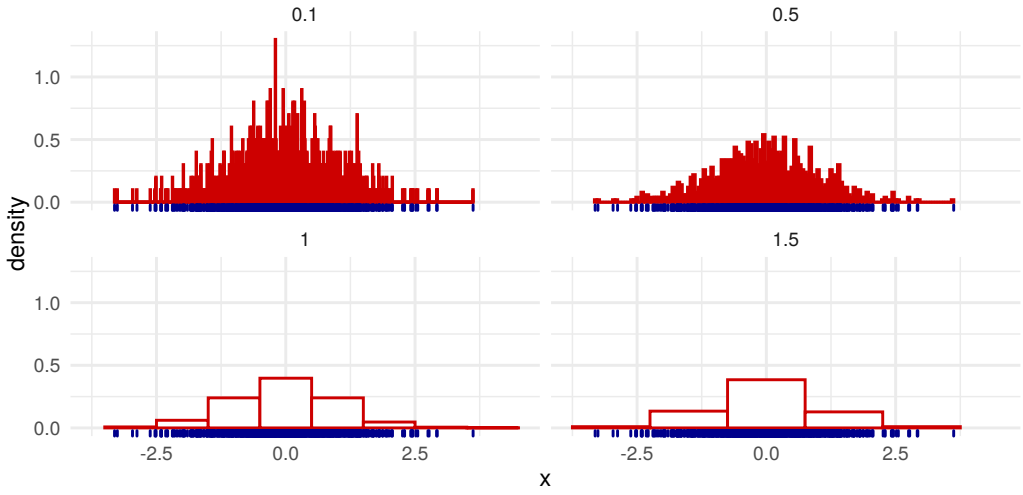


Figura 1.1:

La mala elección del parámetro h causa que el histograma no capture toda la estructura de los datos.



En este caso se puede simplemente minimizar el MISE de la forma usual,

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

implica que

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

y que por lo tanto

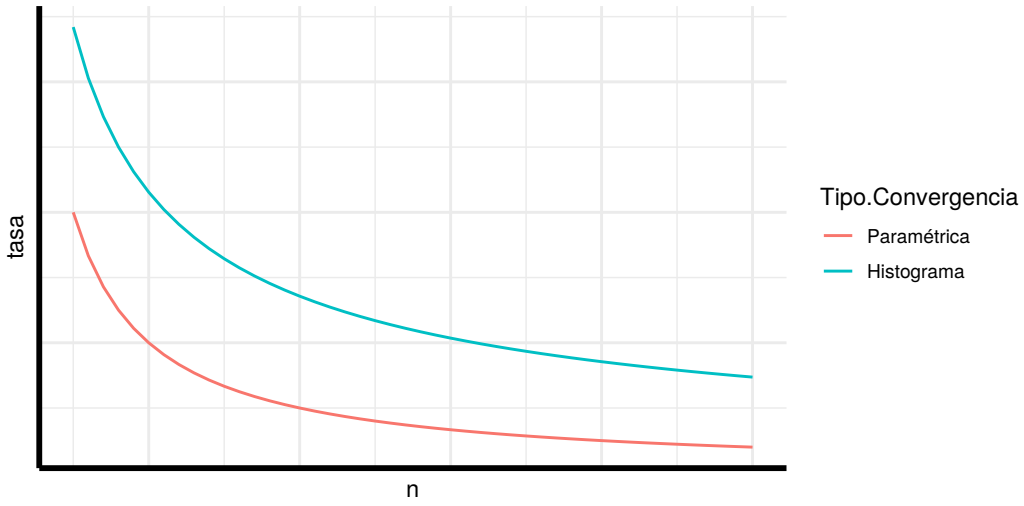
$$\text{MISE}(\hat{f}_h) = \frac{1}{n}\left(\frac{n\|f'\|_2^2}{6}\right)^{1/3}$$

Nota 1.1.5: Recuerde de Estadística I

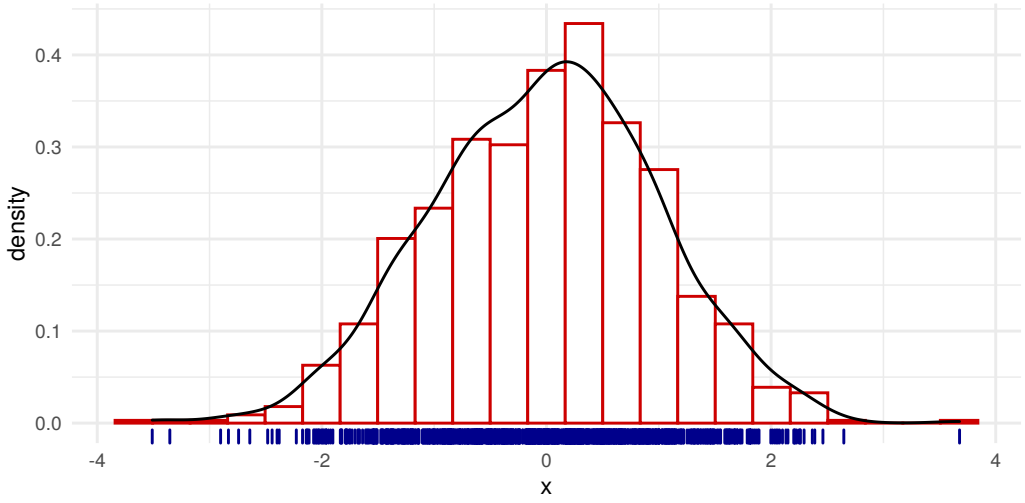
Si $X_1, \ldots, X_2 \sim f_\theta$ i.i.d, con $\text{Var}(X) = \sigma^2$, recuerde que el estimador $\hat{\theta}$ de θ tiene la característica que

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

Según la nota anterior la tasas de convergencia del histograma es más lenta que la de un estimador parámetrico considerando la misma cantidad de datos.



Finalmente, podemos encontrar el valor óptimo de esta datos dado por



$h = 0{,}334$

1.2. Estimación No-paramétrica de densidad

1.2.1. Primera construcción

Sea $X_1, \ldots, X_n$ variables aleatorias i.i.d. con distribución f en $\mathbb{R}$.

La distribución de f es $F(x) = \int_{-\infty}^{x} f(t)dt$.

Considere la distribución empírica como

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

Por la ley de los grandes números tenemos que $\hat{F}_n(x) \xrightarrow{c.s} F(x)$ para todo x en $\mathbb{R}$as $n \to \infty$. Entonces, $F_n(x)$ es consistente

para todo x in $\mathbb{R}$.

Pregunta 1.2.1

¿Podríamos derivar $\hat{F}_n$ para encontrar el estimar $\hat{f}_n$?

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 $\hat{F}_n$, 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 rescribirlo 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 cuál es equivalente al caso del histograma.

1.2.2. 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

Pregunta 1.2.2

¿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 histograma se escribe como

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

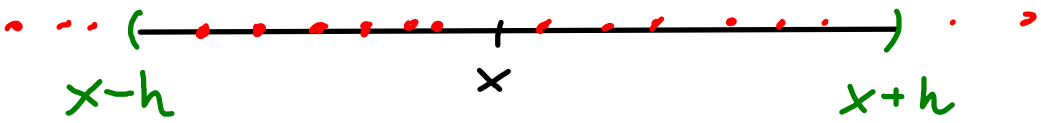
Ahora tratemos de modificar ligeramente esta expresión notando dos cosas

1.
$$\frac{1}{2} I\left(|u| \leq 1\right)$$
con $u = \frac{x-xi}{h}$

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

Finalmente se tiene que

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



Pregunta 1.2.3

¿Qué pasaría si cambiaríamos la función K del histograma por una más general?

Esta función debería cumplir las siguientes características

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

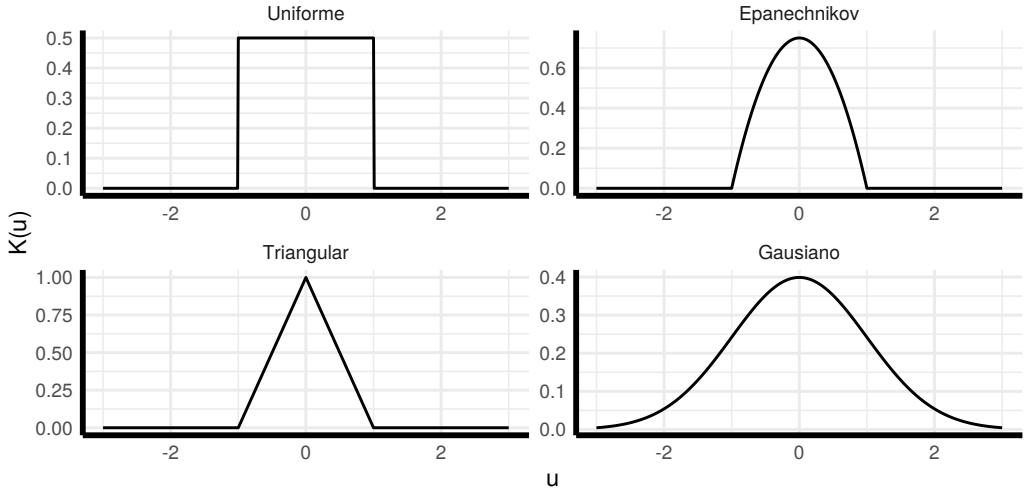
Por ejemplo:

Uniforme: $\frac{1}{2}I\left(|u| \leq 1\right)$.

Triangular: $(1 - |u|)I\left(|u| \leq 1\right)$.

Epanechnikov: $\frac{3}{4}(1 - u^2)I\left(|u| \leq 1\right)$.

Gausian: $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$.



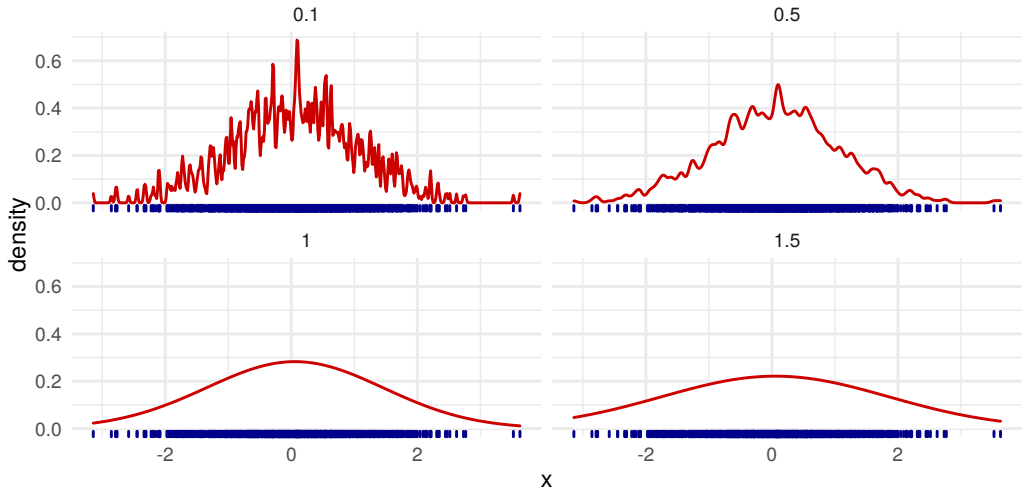
Entonces se tendría que la expresión general para un estimador por núcleos es

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

Pregunta 1.2.4

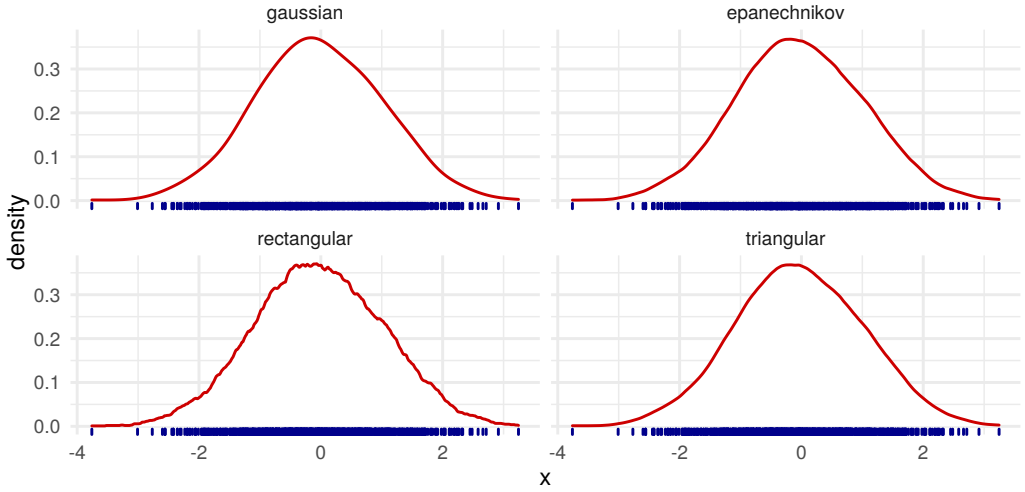
¿Qué pasaría si modificamos el ancho de banda h para un mismo kernel?

Nuevamente sería el ancho de banda ya que



Pregunta 1.2.5

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

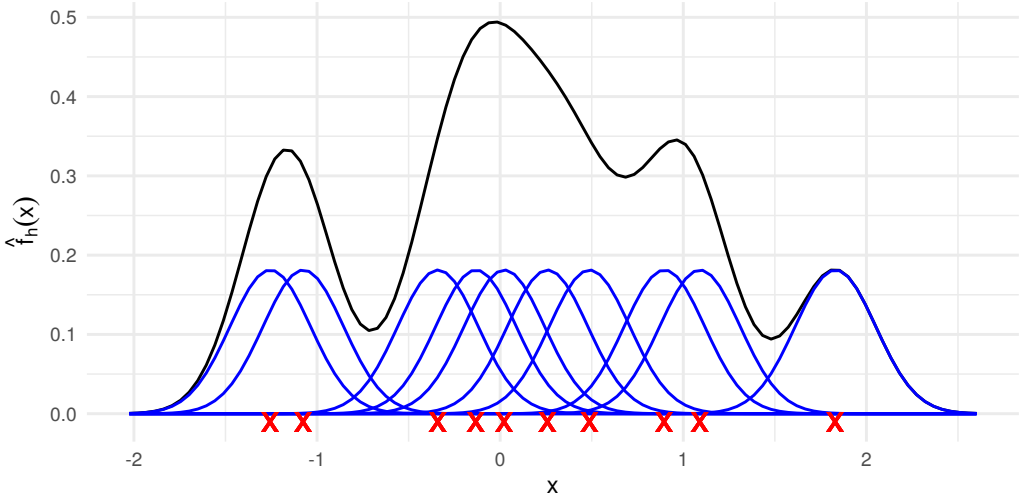


Recordemos nuevamente la fórmula

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

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

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



1.2.3. Propiedades Estadísticas

Pregunta 1.2.6

¿Podríamos imitar lo mismo que hicimos para el histograma?

Si. Las propiedades que vimos anteriormente son universales para estimadores.

Entonces:

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

donde

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

Varianza

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

Usando que:

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

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

Por lo tanto se obtiene que

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

Sesgo

Para el sesgo tenemos

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

Tarea 1.2.7

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

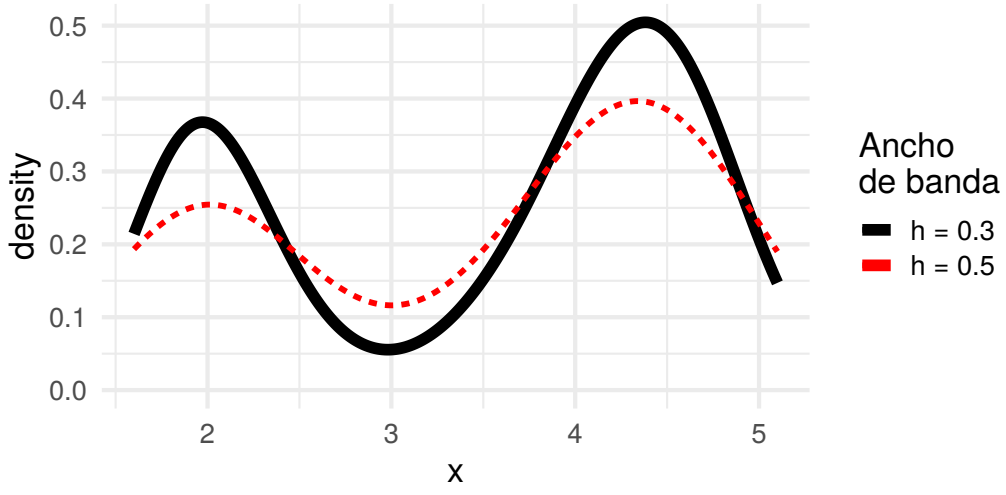
$$\text{Sesgo}\left(\hat{f}_h(x)\right) = \frac{h^2}{2} f'' \mu_2(K) + o(h^2), \text{ si } h \to 0$$

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 $\int (f'''(x))dx < \infty$.*

Solución 1.2.8

$$
\begin{aligned}
&\text{Sesgo}(\hat{f}_h(x)) \\
&= \int \frac{1}{h} K\left(\frac{x-u}{h}\right) f(u)du - f(x) \\
&= \frac{1}{h}\int hK(s)f(sh+x)ds - f(x) \\
&= \int K(s)\left[f(x) + f'(x)(sh+x-x) + \frac{f''(x)}{2}(sh+x-x)^2 + o(h^2)\right] - f(x) \\
&= \int K(s)f(x)ds + \int hf'(x)sK(s)ds + \int \frac{h^2}{2}f''(x)s^2K(s)ds + o(h^2) - f(x) \\
&= f(x) + 0 + \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2) - f(x) \\
&= \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2)
\end{aligned}
$$



Nota 1.2.9

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

Error cuadrático medio y Error cuadrático medio integrado

El error cuadrático medio se escribe

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

Y el error cuadrático medio integrado se escribe como,

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

Ancho de banda óptimo

Minimizando el MISE con respecto a h obtenemos

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

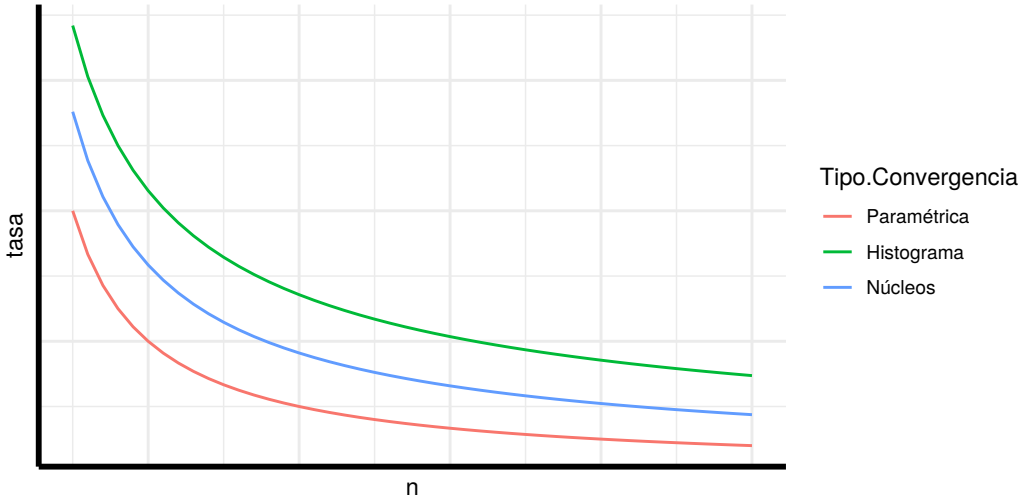
Nota 1.2.10

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.

Evaluando h_{opt} en el MISE tenemos que

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



Nota 1.2.11: Detalle técnico

Formalmente, es posible probar que si f es β veces continuamente diferenciable y $\int \left(f^{(\beta)}\right)^2 < \infty$, entonces se tiene que

$$h_{opt} = O\left(n^{-\frac{1}{2\beta+1}}\right).$$

Por lo tanto se podría aproximar a una tasa paramétrica de convergencia si β es grande.

1.2.4. Escogiendo el ancho de banda

Nota 1.2.12

La principal característica del ancho de banda

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

ES QUE ¡NO FUNCIONA!

Veremos dos métodos para determinar un h que funcione:

- Referencia normal.
- Validación cruzada.

Referencia normal

Cuidado 1.2.13

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 gausiano. Entonces se tiene que

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

donde

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

Tarea 1.2.14

Pruebe que la ecuación anterior es verdadera. Es decir, calcule $\|K\|_2^2$, $\|f''\|_2^2$ y $\mu_2(K)$

Nota 1.2.15

Un problema con $\hat{h}_{rn} = 1{,}06\hat{\sigma}n^{-1/5}$ es su sensibilidad a los valores extremos.

Ejemplo 1.2.16

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

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

El rango intercuantil IQR se define como

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

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

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

Entonces,

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

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

Podemos sustituir la varianza empírica de la fórmula inicial y tenemos

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

Combinando ambos estimadores, podemos obtener,

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

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 1.2.17

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

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

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

Cuidado 1.2.18

El problema con esta expresión es que las observaciones que se usan para estimar la esperanza son las misma 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 $\int \hat{f}_h(x)f(x)dx$ como

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

donde

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

Integral $\int \hat{f}_h^2(x)dx$

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

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

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

Finalmente tenemos la función,

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

Nota 1.2.19

Note que $\mathrm{CV}(h)$ no depende de f o sus derivadas.

Nota 1.2.20

Para efectos prácticos es mejor utilizar el criterio

$$CV(h) = \int \hat{f}_h^2(x)dx - \frac{2}{(n-1)h}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} K\left(\frac{X_i - X_j}{h}\right)$$

y luego calcular numéricamente la integral.

1.2.5. 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

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

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

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

Al igual que en los casos anteriores, este invtervalo 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.

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

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

Nota 1.2.21

Este intervalo de confianza solo funciona en cada punto particular de $f(x)$.

Existe una versión más general para determinar la banda de confianza de toda la función. Por favor revisar la página 62 de Härdle y col. [1].

1.2.6. Laboratorio

Comenzaremos con una librería bastante básica llamada `KernSmooth`.

Efecto de distintos Kernels en la estimación

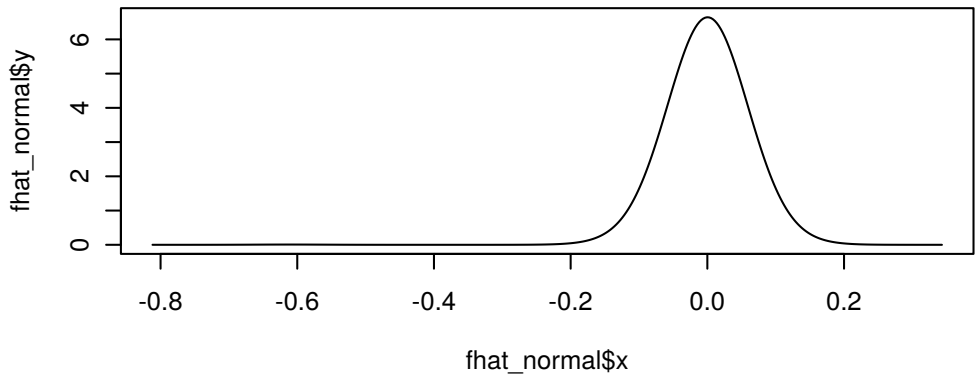
```
x <- read.csv("data/stockres.txt")
x <- x$STOCKRETURN
```

```
summary(x)
```

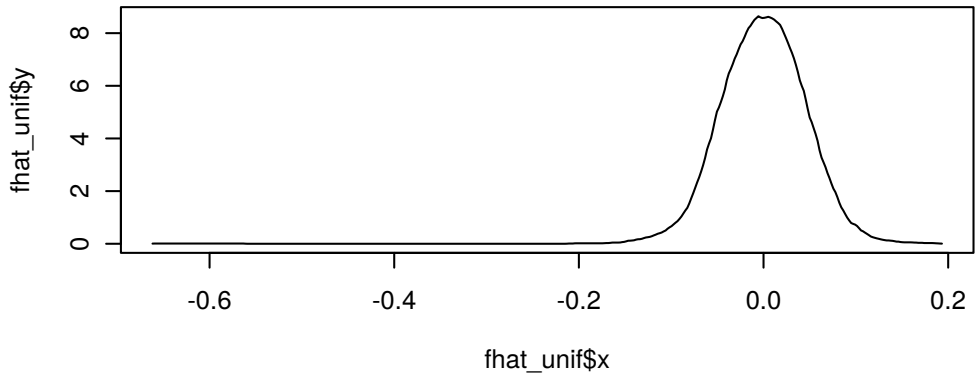
```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.6118200 -0.0204085 -0.0010632 -0.0004988  0.0215999  0.1432286
```

```
library(KernSmooth)
```

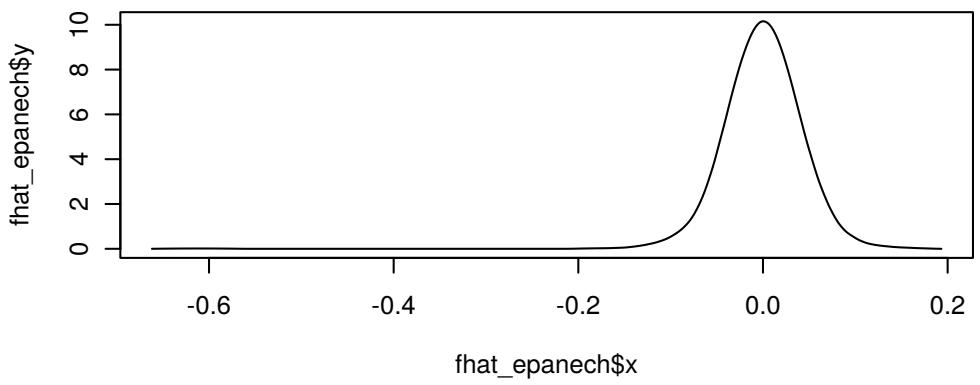
```
fhat_normal <- bkde(x, kernel = "normal", bandwidth = 0.05)
plot(fhat_normal, type = "l")
```



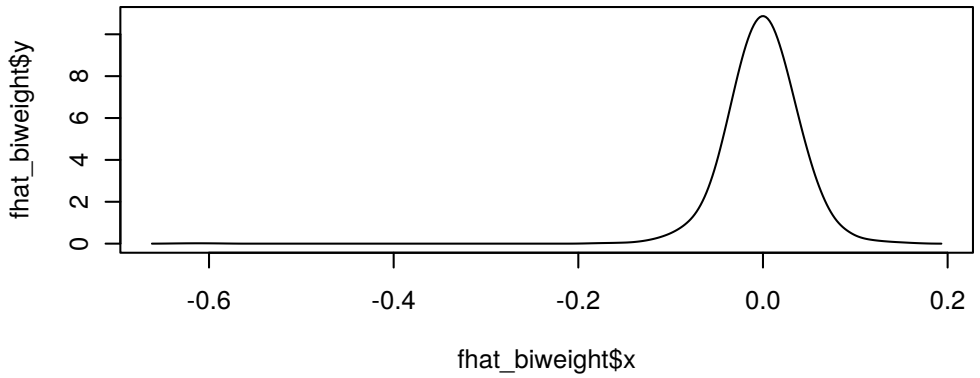

```
fhat_unif <- bkde(x, kernel = "box", bandwidth = 0.05)
plot(fhat_unif, type = "l")
```



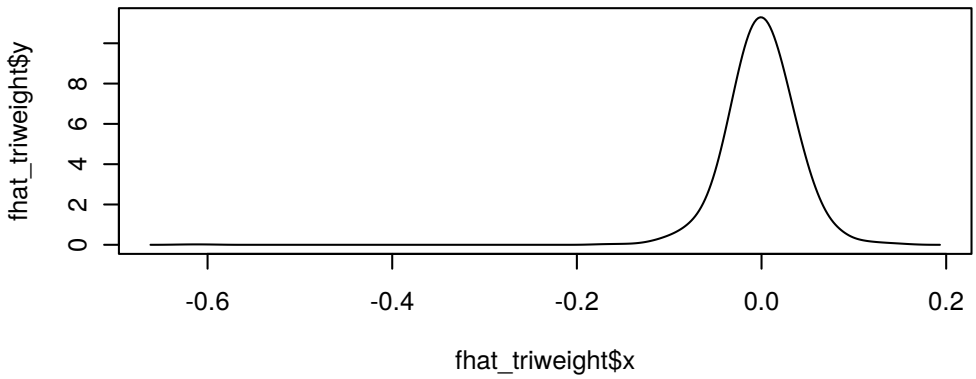

```
fhat_epanech <- bkde(x, kernel = "epanech", bandwidth = 0.05)
plot(fhat_epanech, type = "l")
```




```
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")
```

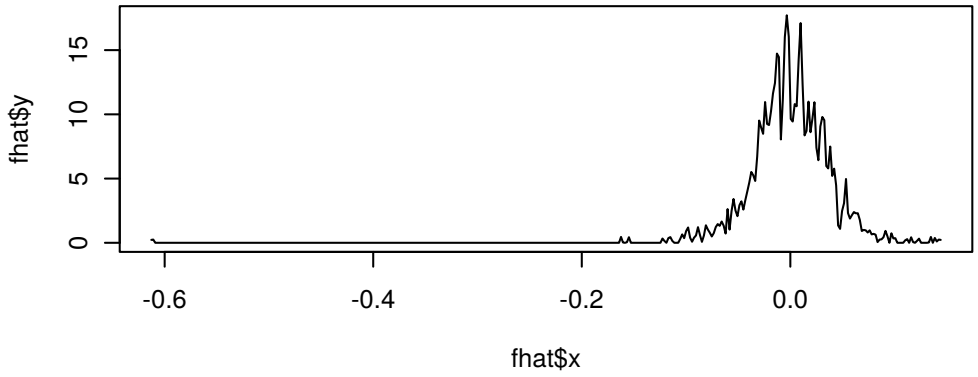


Efecto del ancho de banda en la estimación

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

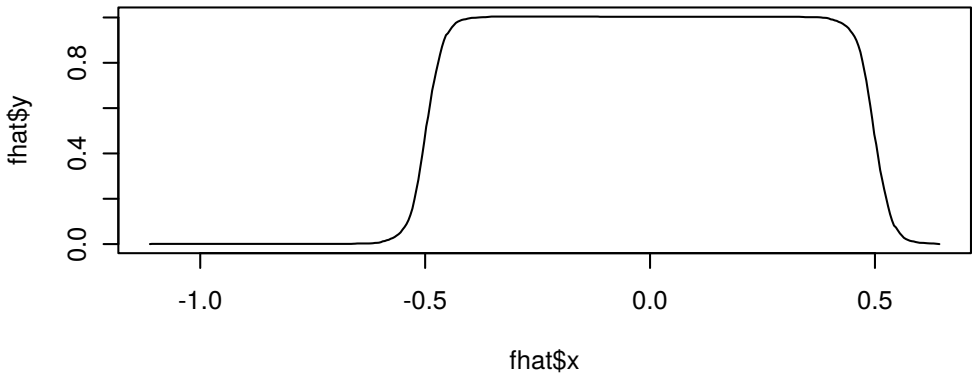
```
## Warning in bkde(x, kernel = "box", bandwidth = 0.001): Binning
grid too coarse for current (small) bandwidth: consider increasing
'gridsize'
```

```
plot(fhat, type = "l")
```



Kernel uniforme

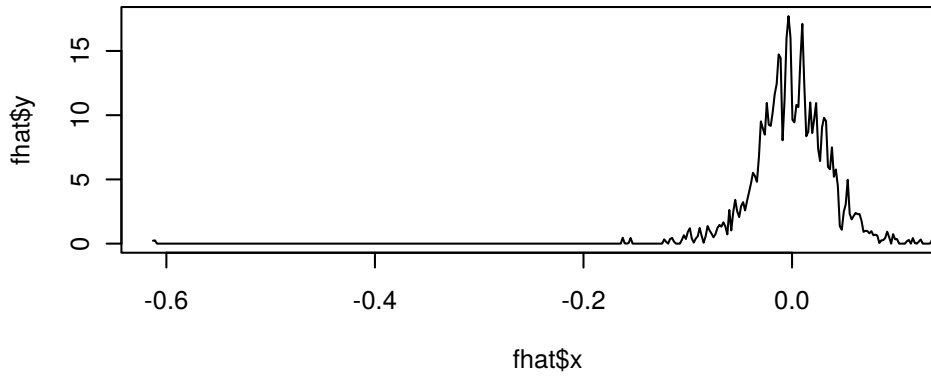
```
fhat <- bkde(x, kernel = "box", bandwidth = 0.5)
plot(fhat, type = "l")
```



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

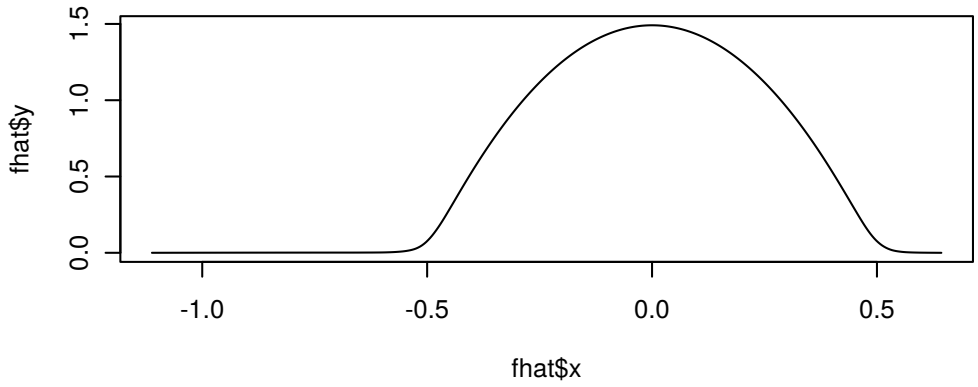
```
## Warning in bkde(x, kernel = "epa", bandwidth = 0.001): Binning
grid too coarse for current (small) bandwidth: consider increasing
'gridsize'
```

```
plot(fhat, type = "l")
```



Kernel Epanechnikov

```
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")
```

Pregunta 1.2.22

1. Construya una variable llamada ‘u‘ que sea una secuencia de -0.15 a 0.15 con un paso de 0.01
2. Asigne ‘x‘ a los datos ‘stockrel‘ y calcule su media y varianza.
3. 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‘.
4. Defina un ancho de banda ‘h‘ en 0.02
5. Construya un histograma para estos datos con ancho de banda

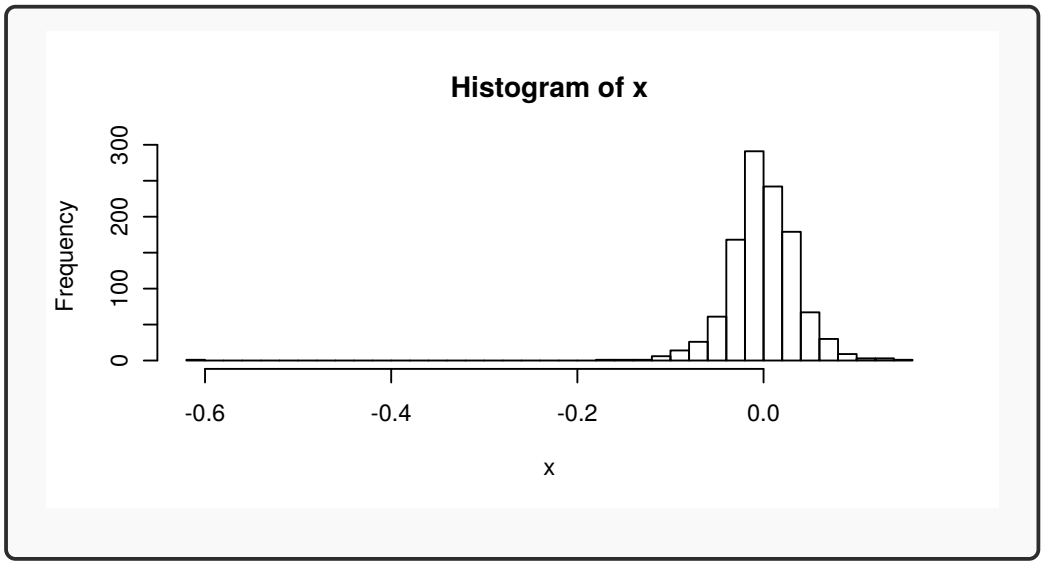
'h'. Llame a esta variable 'f_hist'

6. 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'.

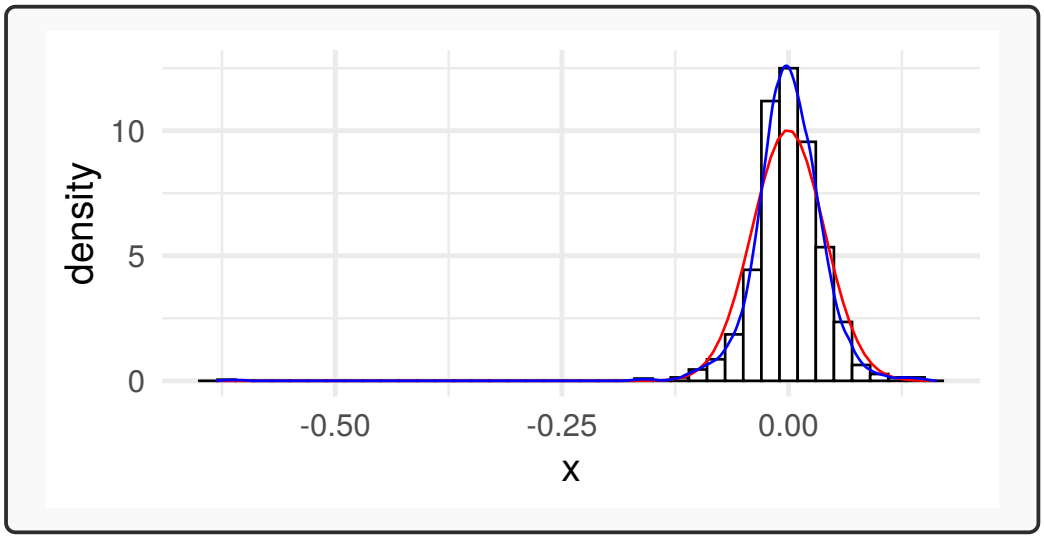
7. Dibuje en el mismo gráfico la estimación paramétrica y no paramétrica.

Solución 1.2.23

```
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)
```


Histogram of x
Frequency
0
100
200
300
-0.6
-0.4
-0.2
0.0
x

```r
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)
```



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

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

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

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

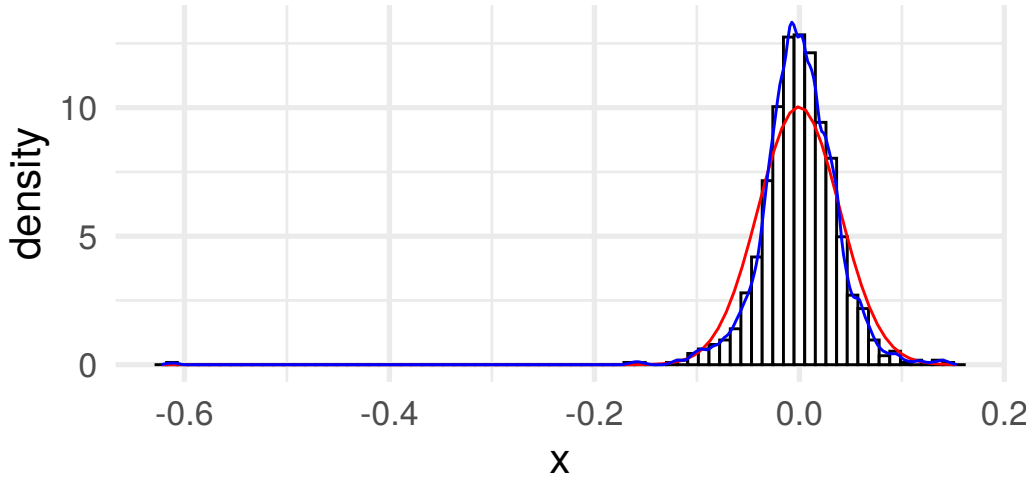
Un estimador más robusto es

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

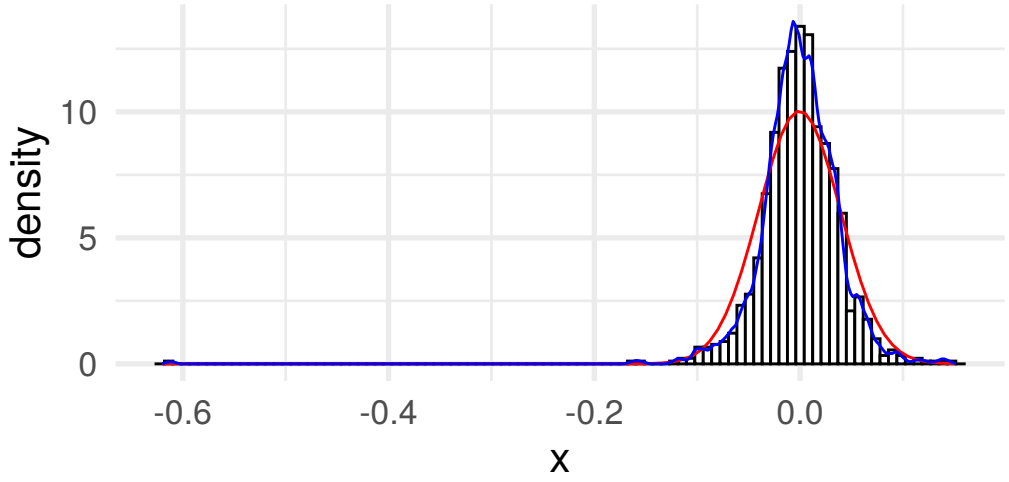
¿Por qué es $IQR/1{,}34$?

```
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)
```



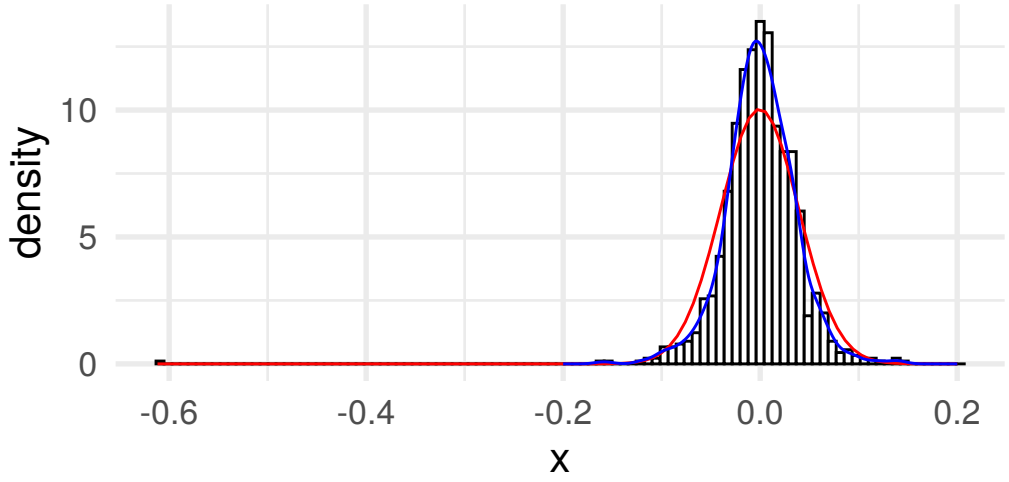
```r
h_iqr <- 1.06 * min(s, IQR(x) / 1.34) * n^(-1 / 5)

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)
```



Validación cruzada

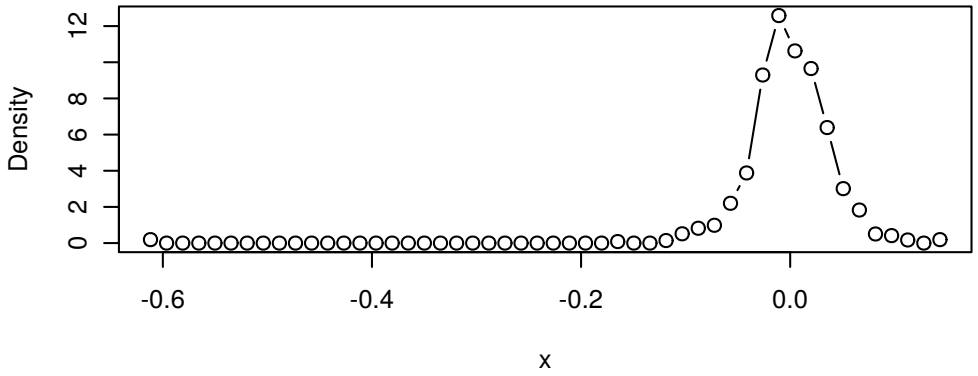
La forma que vimos en clase es la de validación cruzada por mínimos cuadrados"least-square cross validation" la cual se puede ejecutar con este comando.

```
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 /
Multistart 1 of 1 |
Multistart 1 of 1 |
```

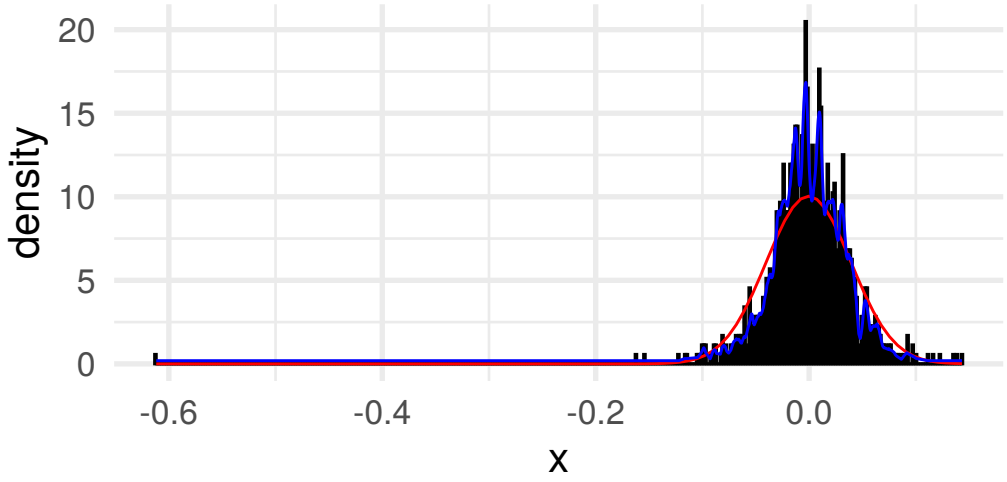
```
dens.np <- npudens(h_cv_np_ls)
```

```
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)
```



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.

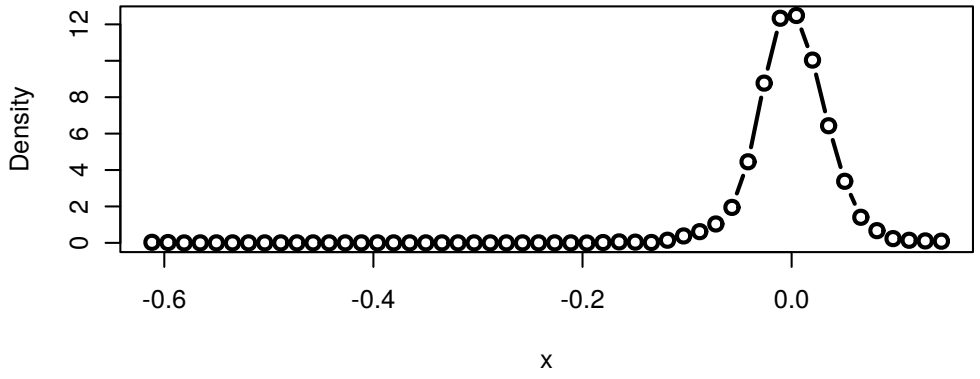
Esto se logra asumiendo que el Kernel es más veces diferenciable.

```
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 /
Multistart 1 of 1 |
Multistart 1 of 1 |
```

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

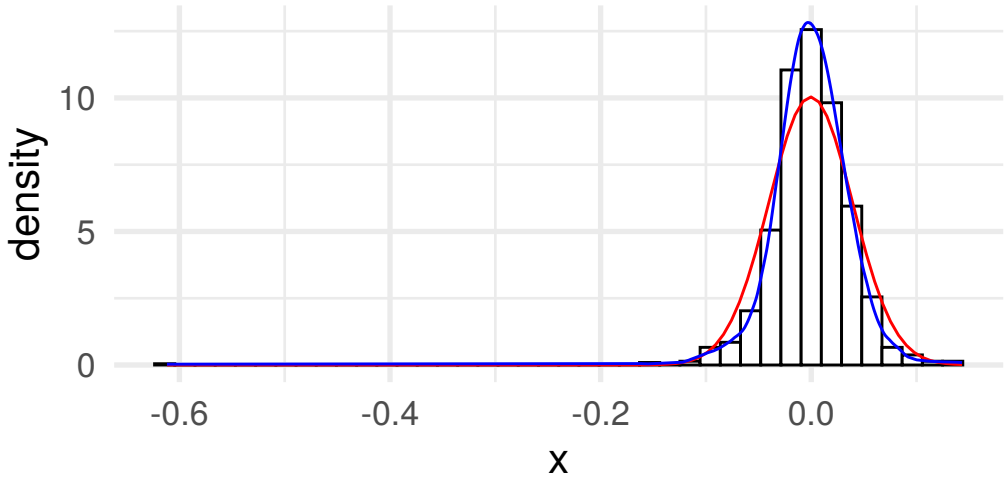
```
plot(dens.np, type = "b", lwd = 2)
```




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

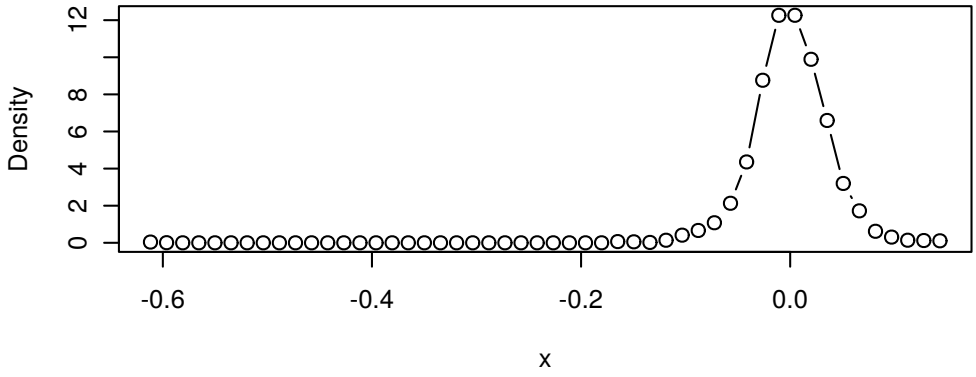
```
ggplot(x_df, aes(x)) +
```

```
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 Hall [2] para entender su estructura.

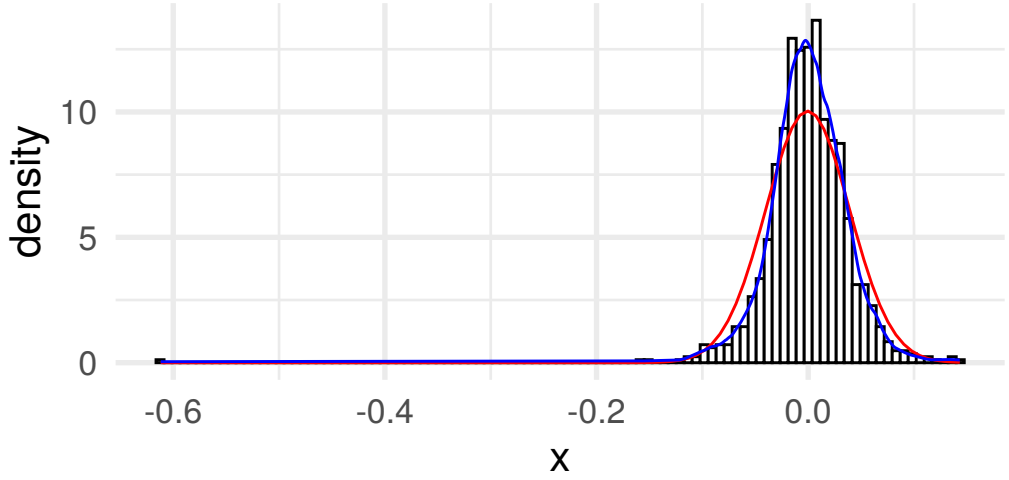
```
h_cv_np_ml <- npudensbw(
 dat = x,
 bwmethod = "cv.ml",
 ckertype = "epanechnikov"
)

##
Multistart 1 of 1 |
Multistart 1 of 1 |
Multistart 1 of 1 |
Multistart 1 of 1 /
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)
```

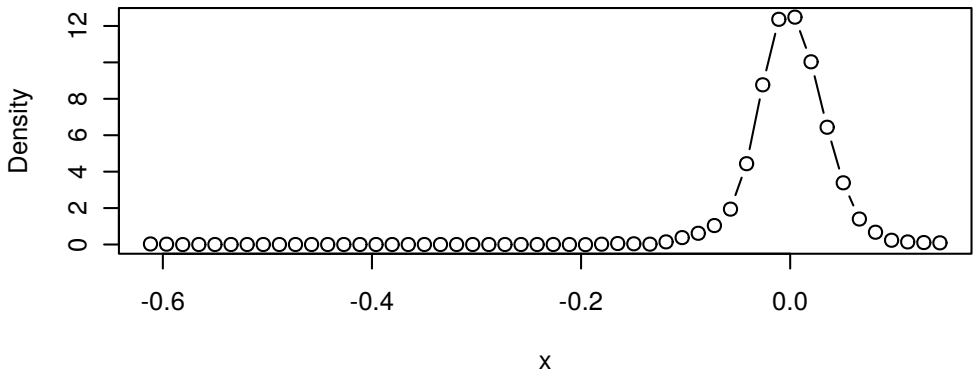



```
h_cv_np_ml <- npudensbw(
 dat = x,
 bwmethod = "cv.ml",
 ckertype = "epanechnikov",
 ckerorder = 4
)

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

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

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

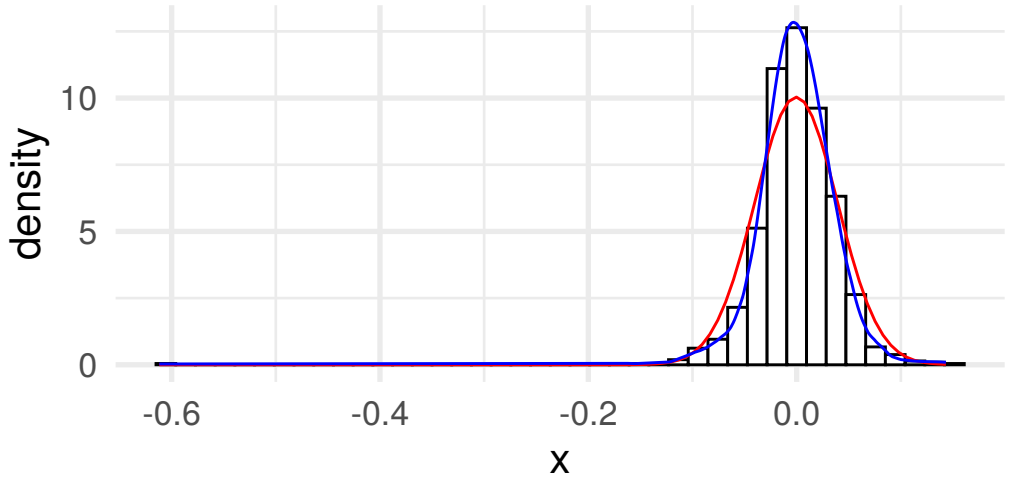
```
plot(dens.np, type = "b")
```




```
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)
```




```
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}")) +
 transition_states(bw) +
 view_follow() +
 theme_minimal(base_size = 20)

anim_save("manual_figure/bandwidth-animation-np.gif")
```

Tarea 1.2.24

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?

```
library(knitr)
```

Capítulo 2

Jacknife y Bootstrap

Suponga que se quiere estimar un intervalo de confianza para la media μ desconocida de un conjunto de datos $X_1, \ldots, X_n$ que tiene distribución $\mathcal{N}(\mu, \sigma^2)$.

Primero se conoce que

$$\sqrt{n}\left(\hat{\mu} - \mu\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right),$$

y esto nos permite escribir el intervalo de confianza como

$$\left[\hat{\mu} - \hat{\sigma} z_{1-\frac{\alpha}{2}}, \hat{\mu} + \hat{\sigma} z_{1-\frac{\alpha}{2}}\right]$$

donde $z_{1-\frac{\alpha}{2}}$ es el cuantil $1 - \frac{\alpha}{2}$ de una normal estándar.

La expresión anterior es posible ya que el supuesto es que la distribución de $\hat{\theta}$ es normal.

Pregunta 2.0.1

¿Qué pasaría si este supuesto es falso o al menos no conocemos la distribución de $\hat{\theta}$?

¿Cómo podemos encontrar ese intervalo de confianza?

Cuidado 2.0.2

Para una muestra fija, el estimador anterior $\hat{\mu}$ solamente un valor. No se conoce la distribución de $\hat{\mu}$. Lo único que se puede estimar son valores puntuales como la media, varianza, mediana, etc, pero no sabemos nada de su distribución.

2.0.1. Caso concreto

Suponga que tenemos la siguiente tabla de datos, que representa una muestra de tiempos y distancias de viajes en Atlanta.

Cargamos la base de la siguiente forma:

```
CommuteAtlanta <- read.csv2("data/CommuteAtlanta.csv")
kable(head(CommuteAtlanta))
```

City	Age	Distance	Time	Sex
Atlanta	19	10	15	M
Atlanta	55	45	60	M
Atlanta	48	12	45	M
Atlanta	45	4	10	F
Atlanta	48	15	30	F
Atlanta	43	33	60	M

Para este ejemplo tomaremos la variable `Time` que la llamaremos `x` para ser más breves. En este caso note que

```
x <- CommuteAtlanta$Time
```

```
mean(x)
```

```
## [1] 29.11
```

y su varianza es

```
(Tn <- var(x))
```

```
## [1] 429.2484
```

A partir de estos dos valores, ¿Cuál sería un intervalo de confianza para la media?

Note que esta pregunta es difícil ya que no tenemos ningún tipo de información adicional.

Las dos técnicas que veremos a continuación nos permitirán extraer *información adicional* de la muestra.

Nota 2.0.3

Para efectos de este capítulo, llamaremos $T_n = T(X_1, \ldots, X_n)$ al estadístico formado por la muestra de los X_i's.

2.1. Jacknife

Esta técnica fue propuesta por [3] y consiste en la siguiente observación.

Se puede probar que muchos de los estimadores tiene la propiedad que

$$\text{Sesgo}\,(T_n) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right) \tag{2.1}$$

para algún a and b.

Por ejemplo $\sigma^2 = \text{Var}\,(X_i)$ y sea $\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$. Entonces,

$$\mathbb{E}\left(\widehat{\sigma}_n^2\right) = \frac{n-1}{n}\sigma^2$$

por lo tanto

$$\text{Sesgo} = -\frac{\sigma^2}{n}$$

Por lo tanto en este caso $a = -\sigma^2$ y $b = 0$.

Defina $T_{(-i)}$ como el estimador T_n pero eliminando el i-ésimo término.

Es claro que en este contexto, se tiene que

$$\text{Sesgo}\left(T_{(-i)}\right) = \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{(n-1)^3}\right) \tag{2.2}$$

Laboratorio 2.1.1

Una forma fácil de construir los $T_{(-i)}$ es primero replicando la matriz de datos múltiple veces usando el producto de kronecker

```
n <- length(x)

jackdf <- kronecker(matrix(1, 1, n), x)


kable(jackdf[1:10, 1:10])
```

15	15	15	15	15	15	15	15	15	15
60	60	60	60	60	60	60	60	60	60
45	45	45	45	45	45	45	45	45	45
10	10	10	10	10	10	10	10	10	10
30	30	30	30	30	30	30	30	30	30
60	60	60	60	60	60	60	60	60	60
45	45	45	45	45	45	45	45	45	45
10	10	10	10	10	10	10	10	10	10
25	25	25	25	25	25	25	25	25	25
15	15	15	15	15	15	15	15	15	15

Y luego se elimina la diagonal

```
diag(jackdf) <- NA


kable(jackdf[1:10, 1:10])
```

NA	15	15	15	15	15	15	15	15	15
60	NA	60	60	60	60	60	60	60	60
45	45	NA	45	45	45	45	45	45	45
10	10	10	NA	10	10	10	10	10	10
30	30	30	30	NA	30	30	30	30	30
60	60	60	60	60	NA	60	60	60	60
45	45	45	45	45	45	NA	45	45	45
10	10	10	10	10	10	10	NA	10	10
25	25	25	25	25	25	25	25	NA	25
15	15	15	15	15	15	15	15	15	NA

Cada columna contiene toda la muestra excepto el i-ésimo elemento.

Solo basta estimar la media de cada columna:

```
T_i <- apply(jackdf, 2, var, na.rm=TRUE)


kable(T_i[1:10])
```

x
429.7098
428.1905
429.6023
429.3756
430.1087
428.1905
429.6023
429.3756
430.0764
429.7098

Definamos el sesgo *jackife* como

$$b_{jack} = (n-1)(\overline{T}_n - T_n)$$

donde

$$\overline{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_{(-i)}$$

Laboratorio 2.1.2

En nuestro caso tendríamos lo siguiente:

```
(bjack <- (n-1)*(mean(T_i) - Tn))

## [1] 0
```

es decir, que los `T_i` generan estimadores de `T_n` que contienen el mismo sesgo.

Observe que b_{jack} tiene la siguiente propiedad

$$\begin{aligned}
\mathbb{E}\left(b_{\text{jack}}\right) &= (n-1)\left(\mathbb{E}\left[\overline{T}_n\right] - \mathbb{E}\left[T_n\right]\right) \\
&= (n-1)\left(\mathbb{E}\left[\overline{T}_n\right] - \theta + \theta - \mathbb{E}\left[T_n\right]\right) \\
&= (n-1)\left(\text{Sesgo}\left(\overline{T}_n\right) - \text{Sesgo}\left(T_n\right)\right) \\
&= (n-1)\left[\left(\frac{1}{n-1} - \frac{1}{n}\right)a + \left(\frac{1}{(n-1)^2} - \frac{1}{n^2}\right)b + O\left(\frac{1}{n^3}\right)\right] \\
&= \frac{a}{n} + \frac{(2n-1)b}{n^2(n-1)} + O\left(\frac{1}{n^2}\right) \\
&= \text{Sesgo}\left(T_n\right) + O\left(\frac{1}{n^2}\right)
\end{aligned}$$

Nota 2.1.3

Es decir, en general, el estimador b_{jack} aproxima correctamente $\text{Sesgo}\left(T_n\right)$ hasta con un error del n^{-2}.

Podemos usar los T_i para generar muestras adicionales para estimar

el parámetro θ.

En este caso defina el siguiente estimador:

$$\widetilde{T}_i = nT_n - (n-1)T_{(-i)}.$$

Nota 2.1.4

A $\widetilde{T}_i$ se le llaman **pseudo-valores** y representa el aporte o peso que tiene la variable X_i para estimar T_n.

Tarea 2.1.5

Usado un cálculo similar para el b_{jack} pruebe que

$$\text{Sesgo}\left(T_{\text{jack}}\right) = -\frac{b}{n(n-1)} + O\left(\frac{1}{n^2}\right) = O\left(\frac{1}{n^2}\right).$$

¿Qué conclusión se obtiene de este cálculo?

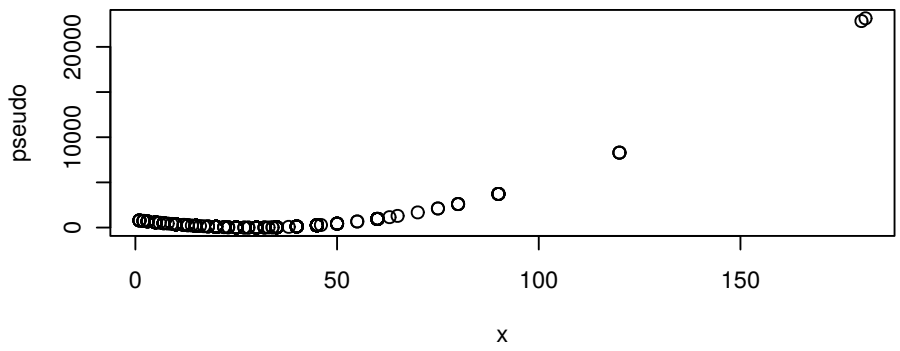
Laboratorio 2.1.6

Los pseudo-valores se estiman de forma directa como,

```
pseudo <- n * Tn - (n - 1) * T_i


pseudo[1:10]

##  [1] 199.02972209 957.16225222 252.64417993 365.79679037  -0.06666345
##  [6] 957.16225222 252.64417993 365.79679037  16.09799519 199.02972209
```

Lo importante acá es notar la similitud que tiene con los datos reales,

```
plot(x = x, y = pseudo)
```



Con estos pseudo-valores, es posible estimar la media y la varianza de T_n con sus respectivos estimadores:

$$T_{\text{jack}} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{T}_i$$

donde

$$v_{jack} = \frac{\sum_{i=1}^{n}\left(\widetilde{T}_i - \frac{1}{n}\sum_{i=1}^{n}\widetilde{T}_i\right)^2}{n(n-1)}.$$

Nota 2.1.7

Sin embargo, se puede demostrar fácilmente que se pueden usar pseudovalores para construir una prueba normal de hipótesis. Dado que cada pseudovalor es independiente e idénticamente distribuido

(iid), se deduce que su promedio se ajusta a una distribución normal a medida que el tamaño de la muestra aumenta. El promedio de los pseudovalores es solo T_{jack} y el valor esperado de ese promedio, debido a la construcción a la imparcialidad del estimador, es el parámetro bajo investigación, θ. Por lo tanto, tenemos que

$$\frac{\sqrt{n}\left(T_{jack} - \theta\right)}{\sqrt{v_{jack}}} \rightarrow N(0,1).$$

Laboratorio 2.1.8

```
(Tjack <- mean(pseudo))
```

```
## [1] 429.2484
```

```
(Vjack <- var(pseudo, na.rm = TRUE))
```

```
## [1] 2701991
```

```
(sdjack <- sqrt(Vjack))
```

```
## [1] 1643.774
```

```
(z <- qnorm(1 - 0.05 / 2))
```

```
## [1] 1.959964
```

```
c(Tjack - z * sdjack / sqrt(n),
Tjack + z * sdjack / sqrt(n))

## [1] 285.1679 573.3289
```

2.2. Bootstrap

Este método es un poco más sencillo de implementar que Jacknife y es igualmente de eficaz propuesto por [4].

Primero recordemos que estamos estimando una estadístico a partir de una muestra de modo que $T_n = g(X_1, \ldots, X_n)$ donde g es cualquier función (media, varianza, quantiles, etc).

Supongamos que conocemos la distribución real de los X's, llamada $F(x)$. Si uno quisiera estimar la varianza de X basta con hacer

$$\mathrm{Var}_F(T_n) = \frac{\sigma^2}{n} = \frac{\int x^2 dF(x) - \left(\int x dF(x)\right)^2}{n}$$

donde $\sigma^2 = \mathrm{Var}(X)$ y el subindice F es solo para indicar la dependencia con la distribución real.

Ahora dado que no tenemos la distribución real $F(x)$, una opción es encontrar un estimador de esta llamado $\hat{F}_n$.

La técnica de boostrap se basa en extraer muchas muestras iid de la distribución $\hat{F}_n$ de modo que se pueda conocer su varianza.

En simple pasos la técnica es

1. Seleccione $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$
2. Estime $T_n^* = g\left(X_1^*, \ldots, X_n^*\right)$
3. Repita los Pasos 1 y 2, B veces para obtener $T_{n,1}^*, \ldots, T_{n,B}^*$

4. Estime

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(T^*_{n,b} - \frac{1}{B}\sum_{r=1}^{B} T^*_{n,r}\right)^2$$

Por la ley de los grandes números tenemos que

$$v_{\text{boot}} \xrightarrow{\text{a.s.}} \mathbb{V}_{\widehat{F}_n}(T_n), \quad \text{si } B \to \infty. \tag{2.3}$$

además llamaremos,

$$\widehat{\text{se}}_{\text{boot}} = \sqrt{v_{\text{boot}}}$$

En pocas palabras lo que tenemos es que

$$\begin{aligned} \text{Mundo Real: } F &\quad \Longrightarrow X_1, \ldots, X_n \Longrightarrow \quad T_n = g(X_1, \ldots, X_n) \\ \text{Mundo Bootstrap: } \widehat{F}_n &\quad \Longrightarrow X^*_1, \ldots, X^*_n \Longrightarrow \quad T^*_n = g\left(X^*_1, \ldots, X^*_n\right) \end{aligned}$$

En términos de convergencia lo que se tiene es que

$$\text{Var}_F(T_n) \overbrace{\approx}^{O(1/\sqrt{n})} \text{Var}_{\widehat{F}_n}(T_n) \overbrace{\approx}^{O(1/\sqrt{B})} v_{boot}$$

Pregunta 2.2.1

¿Cómo extraemos una muestra de $\hat{F}_n$?

Recuerden que $\hat{F}_n$ asigna la probabilidad de $\frac{1}{n}$ a cada valor usado para

construirla.

Por lo tanto, todos los puntos originales $X_1, \ldots, X_n$ tienen probabilidad $\frac{1}{n}$ de ser escogidos, que resulta ser equivalente a un muestreo con remplazo n-veces.

Así que basta cambiar el punto 1. del algoritmo mencionando anteriormente con

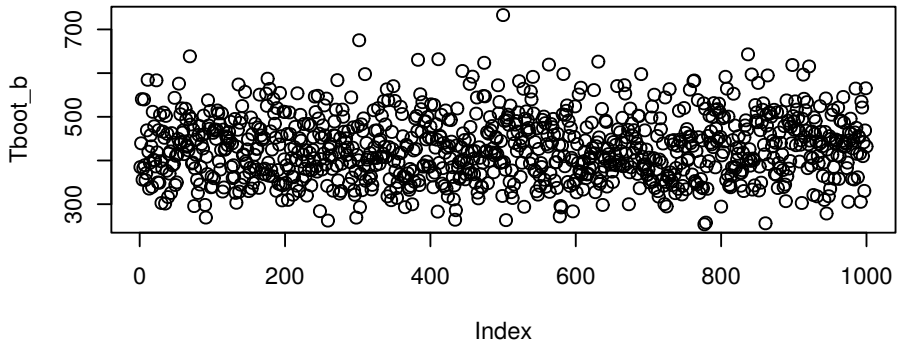
1. Seleccione una muestra con remplazo $X_1^*, \ldots, X_n^*$ de $X_1, \ldots, X_n$.

Laboratorio 2.2.2

En este ejemplo podemos tomar $B = 1000$ y construir esa cantidad de veces nuestro estimador.

```r
B <- 1000

Tboot_b <- NULL


for(b in 1:B) {

xb <- sample(x, size = n, replace = TRUE)

Tboot_b[b] <- var(xb)

}



Tboot_b[1:10]

##  [1] 384.8959 439.3992 540.4429 356.0363 382.5001 539.7398 389.2075 355.4033
##  [9] 371.0053 469.0120
```

```r
plot(Tboot_b)
```



Por supuesto podemos encontrar los estadísticos usuales para esta nueva muestra

```
(Tboot <- mean(Tboot_b))
```

```
## [1] 428.3197
```

```
(Vboot <- var(Tboot_b))
```

```
## [1] 5345.401
```

```
(sdboot <- sqrt(Vboot))
```

```
## [1] 73.11225
```

2.2.1. Intervalos de confianza

Intervalo Normal

Este es el más sencillo y se escribe como

$$T_n \pm z_{\alpha/2}\widehat{\text{Se}}_{\text{boot}} \tag{2.4}$$

Cuidado 2.2.3

Este intervalo solo funciona si la distribución de T_n es normal.

Laboratorio 2.2.4

El cálculo de este intervalo es

```
c(Tn - z * sdboot,
Tn + z * sdboot)

## [1] 285.9510 572.5458
```

Intervalo pivotal

Sea $\theta = T(F)$ y $\widehat{\theta}_n = T\left(\widehat{F}_n\right)$ y defina la cantidad pivotal $R_n = \widehat{\theta}_n - \theta$.

Sea $H(r)$ la función de distribución del pivote:

$$H(r) = \mathbb{P}_F\left(R_n \leq r\right).$$

Además considere $C_n^\star = (a, b)$ donde

$$a = \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \quad \text{y} \quad b = \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right).$$

Se sigue que

$$\begin{aligned}\mathbb{P}(a \leq \theta \leq b) &= \mathbb{P}\left(\widehat{\theta}_n - b \leq R_n \leq \widehat{\theta}_n - a\right) \\ &= H\left(\widehat{\theta}_n - a\right) - H\left(\widehat{\theta}_n - b\right) \\ &= H\left(H^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - H\left(H^{-1}\left(\frac{\alpha}{2}\right)\right) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha\end{aligned}$$

Nota 2.2.5

$C_n^\star = (a, b)$ es un intervalo de confianza al $1 - \alpha$ de confianza.

El problema es que este intervalo depende de H desconocido.

Para resolver este problema, se puede construir una versión *bootstrap* de H usando lo que sabemos hasta ahora.

$$\widehat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I\left(R_{n,b}^* \leq r\right)$$

donde $R_{n,b}^* = \widehat{\theta}_{n,b}^* - \widehat{\theta}_n$.

Sea r_β^* el cuantil muestral de tamaño β de $\left(R_{n,1}^*, \ldots, R_{n,B}^*\right)$ y sea θ_β^* el cuantil muestral de tamaño β de $\left(\theta_{n,1}^*, \ldots, \theta_{n,B}^*\right)$.

Nota 2.2.6

Según la notación anterior note que

$$r_\beta^* = \theta_\beta^* - \widehat{\theta}_n$$

Con estas observaciones It follows that an approximate $1-\alpha$ confidence interval is $C_n = (\widehat{a}, \widehat{b})$ where

$$\widehat{a} = \widehat{\theta}_n - \widehat{H}^{-1}\left(1-\frac{\alpha}{2}\right) = \widehat{\theta}_n - r^*_{1-\alpha/2} = \widehat{\theta}_n - \theta^*_{1-\alpha/2} + \widehat{\theta}_n = 2\widehat{\theta}_n - \theta^*_{1-\alpha/2}$$
$$\widehat{b} = \widehat{\theta}_n - \widehat{H}^{-1}\left(\frac{\alpha}{2}\right) = \widehat{\theta}_n - r^*_{\alpha/2} = \widehat{\theta}_n - \theta^*_{\alpha/2} + \widehat{\theta}_n = 2\widehat{\theta}_n - \theta^*_{\alpha/2}$$

Nota 2.2.7

El intervalo de confianza pivotal de tamaño $1-\alpha$ es

$$C_n = \left(2\widehat{\theta}_n - \widehat{\theta}^*_{((1-\alpha/2)B)}, 2\widehat{\theta}_n - \widehat{\theta}^*_{((\alpha/2)B)}\right)$$

Laboratorio 2.2.8

El intervalo anterior para un nivel de 95 % se estima de la siguiente forma

```
c(2 * Tn - quantile(Tboot_b, 1 - 0.05 / 2) ,
2 * Tn - quantile(Tboot_b, 0.05 / 2))

##     97.5%      2.5%
## 275.3768 558.6741
```

Intervalo pivotal studentizado

Una mejora del intervalo anterior sería normalizar los estimadores previamente

$$Z_n = \frac{T_n - \theta}{\widehat{\text{se}}_{\text{boot}}}.$$

Como θ es desconocido, entonces la versión a estimar es

$$Z^*_{n,b} = \frac{T^*_{n,b} - T_n}{\widehat{\text{se}}^*_b}$$

donde $\widehat{\text{se}}^*_b$ es un estimador del error estándar de $T^*_{n,b}$ no de T_n.

Cuidado 2.2.9

Esto requerira estimar la varianza de $T^*_{n,b}$ para cada b.

Con esto se puede obtener cantidades $Z^*_{n,1}, \ldots, Z^*_{n,B}$ que debería ser próximos a Z_n.

Sea z^*_α del α cuantíl de $Z^*_{n,1}, \ldots, Z^*_{n,B}$, entonces $\mathbb{P}(Z_n \leq z^*_\alpha) \approx \alpha$.

Define el intervalo

$$C_n = \left(T_n - z^*_{1-\alpha/2}\widehat{\text{se}}_{\text{boot}}, T_n - z^*_{\alpha/2}\widehat{\text{se}}_{\text{boot}}\right)$$

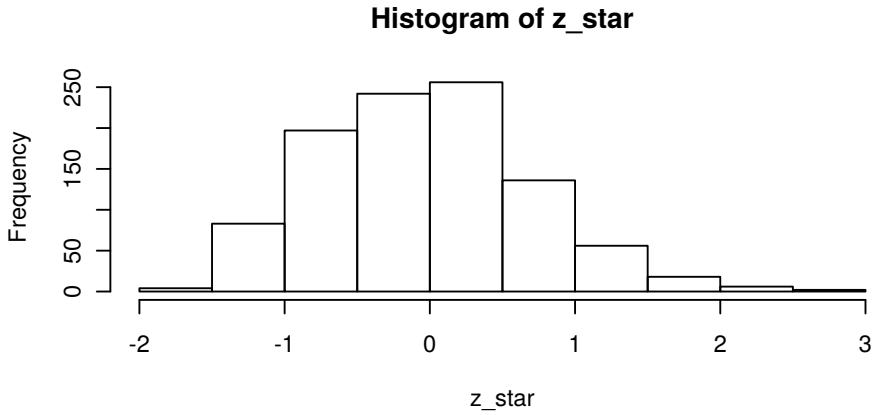
Justificado por el siguiente cálculo:

$$\begin{aligned}
\mathbb{P}(\theta \in C_n) &= \mathbb{P}\left(T_n - z^*_{1-\alpha/2}\widehat{\text{Se}}_{\text{boot}} \leq \theta \leq T_n - z^*_{\alpha/2}\widehat{\text{Se}}_{\text{boot}}\right) \\
&= \mathbb{P}\left(z^*_{\alpha/2} \leq \frac{T_n - \theta}{\text{se}_{\text{boot}}} \leq z^*_{1-\alpha/2}\right) \\
&= \mathbb{P}\left(z^*_{\alpha/2} \leq Z_n \leq z^*_{1-\alpha/2}\right) \\
&\approx 1 - \alpha
\end{aligned}$$

Laboratorio 2.2.10

Note que para este caso tenemos que hacer bootstrap para cada estimador bootstrap calculado.

```
  B <- 1000
  Tboot_b <- NULL
  Tboot_bm <- NULL
  sdboot_b <- NULL

  for (b in 1:B) {
xb <- sample(x, size = n, replace = TRUE)
Tboot_b[b] <- var(xb)
for (m in 1:B) {
  xbm <- sample(xb, size = n, replace = TRUE)
  Tboot_bm[b] <- var(xbm)
}
sdboot_b <- sd(Tboot_bm)
  }

  z_star <- (Tboot_b - Tn) / sdboot_b

  hist(z_star)
```




```
c(Tn - quantile(z_star, 1 - 0.05 / 2) *   sdboot,
Tn - quantile(z_star, 0.05 / 2) *   sdboot)

##     97.5%     2.5%
## 317.0387 519.9397
```

2.2.2. Resumiendo

Resumiendo todos lo métodos de cálculo de intervalos obtenemos

```
kable(data.frame(
  `Método` = c(
    "Jacknife",
    "Bootstrap Normal",
```

```
    "Bootstrap Pivotal",
    "Bootstrap Pivotal Estudentizado"
  ),
  Inferior = c(
    Tjack - z * sdjack / sqrt(n),
    Tn - z * sdboot,
    2 * Tn - quantile(Tboot_b, 1 - 0.05 / 2),
    Tn - quantile(z_star, 1 - 0.05 / 2) *   sdboot
  ),
  Superior = c(
     Tjack + z * sdjack / sqrt(n),
     Tn + z * sdboot,
      2 * Tn - quantile(Tboot_b, 0.05 / 2),
     Tn - quantile(z_star, 0.05 / 2) *   sdboot
  )
))
```

Método	Inferior	Superior
Jacknife	285.1679	573.3289
Bootstrap Normal	285.9510	572.5458
Bootstrap Pivotal	273.7930	554.8922
Bootstrap Pivotal Estudentizado	317.0387	519.9397

2.3. Ejercicios

1. Repita los ejercicios anteriores para calcular intervalos de confianza para la distancia promedio y la varianza del desplazamiento de las personas. Use los métodos de Jacknife y Bootstrap (con todos sus intervalos de confianza).

   Dada que la distancia es una medida que puede ser influenciada por distancias muy cortas o muy largas, se puede calcular el logaritmo de esta variable para eliminar la escala de la distancias.

2. Verifique que esta última variable se podría estimar paramétricamente con una distribución normal.

   Repita los cálculos anteriores tomando como cuantiles los de una normal con media 0 y varianza 1.

3. Compare los intervalos calculados y comente los resultados.

Bibliografía

[1] Wolfgang Härdle y col. *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, págs. xxviii+299. ISBN: 978-3-642-62076-8. DOI: 10.1007/978-3-642-17146-8. URL: http://link.springer.com/10.1007/978-3-642-17146-8.

[2] Peter Hall. «On Kullback-Leibler Loss and Density Estimation». En: *The Annals of Statistics* 15.4 (dic. de 1987), págs. 1491-1519. ISSN: 0090-5364. DOI: 10.1214/aos/1176350606. URL: http://projecteuclid.org/euclid.aos/1176350606.

[3] M. H. Quenouille. «Approximate Tests of Correlation in Time-Series». En: *Journal of the Royal Statistical Society: Series B (Methodological)* 11.1 (ene. de 1949), págs. 68-84. ISSN: 00359246. DOI: 10.1111/j.2517-6161.1949.tb00023.x. URL: http://doi.wiley.com/10.1111/j.2517-6161.1949.tb00023.x.

[4] B. Efron. «Bootstrap Methods: Another Look at the Jackknife». En: *The Annals of Statistics* 7.1 (ene. de 1979), págs. 1-26. ISSN: 0090-5364.

DOI: 10.1214/aos/1176344552. URL: http://projecteuclid.org/euclid.aos/1176344552.