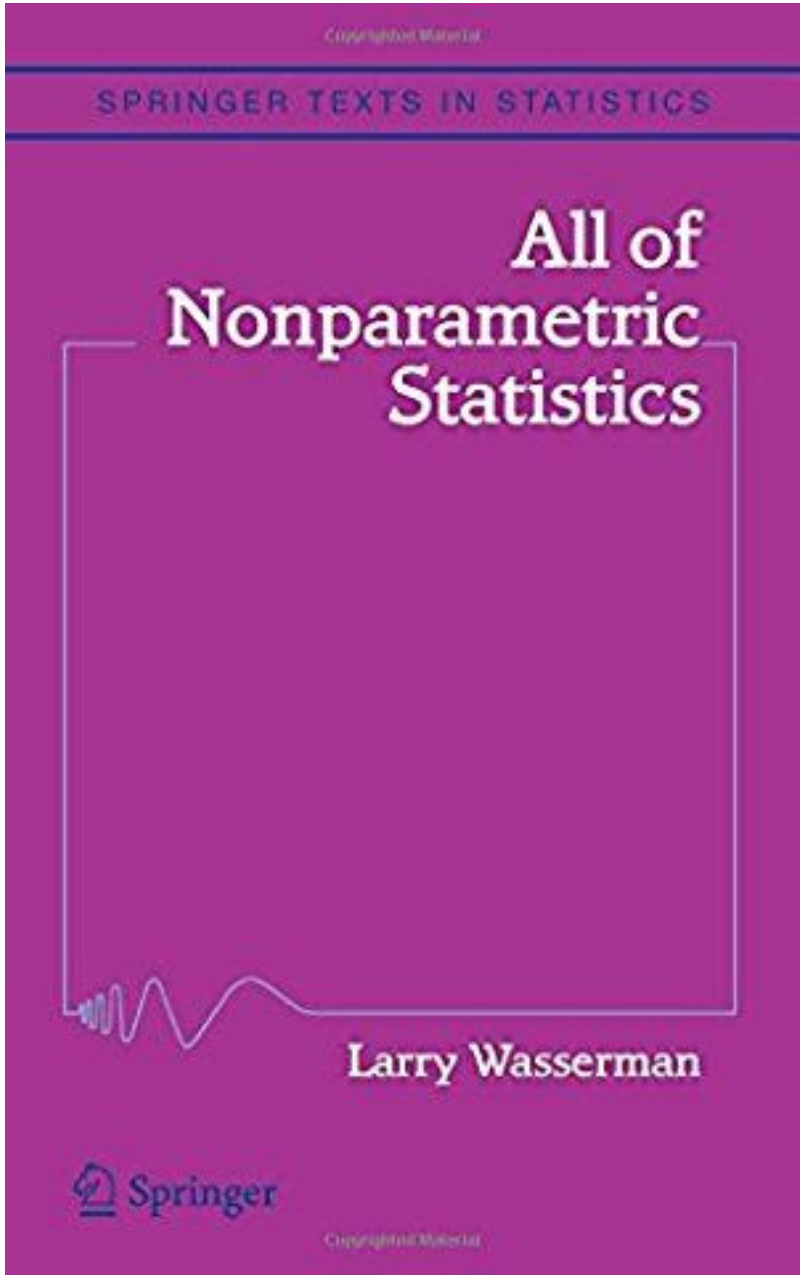
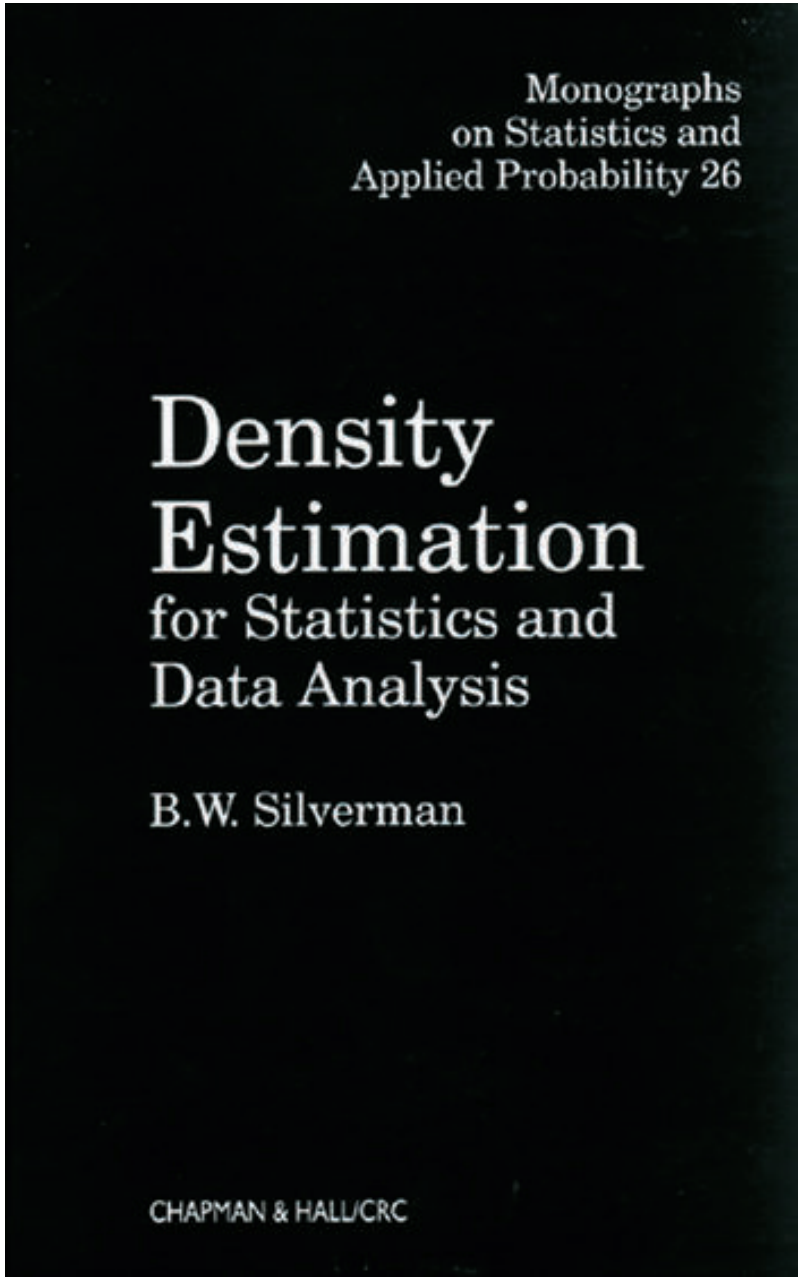
Inferencia Estadística

Tópicos inferencia no paramétrica

Gabriel Martos Venturini
gmartos@utdt.edu

UTDT


SPRINGER TEXTS IN STATISTICS
All of Nonparametric Statistics
Larry Wasserman
Springer


Monographs on Statistics and Applied Probability 26
Density Estimation
for Statistics and Data Analysis
B.W. Silverman
CHAPMAN & HALL/CRC

Agenda

- Objetivo: Construir métodos de inferencia basados en supuestos "débiles" sobre la distribución de la variable de interés en la población.

- Ejemplo: Asumimos para el modelo estadístico cierto grado de *suavidad* en f (o la cdf F) de la que se generan los datos.

  - $\underline{X} = \{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x)$.
  - En vez de asumir $f(x) \equiv f(x;\theta)$, le pedimos a la verdadera f pertenecer al conjunto de funciones $\mathcal{F}_c$ de variabilidad acotada:

$$f \in \mathcal{F}_c \equiv \left\{ f : \int (f''(x))^2 \mathrm{d}x < c \right\}.$$

- "*... a better name for nonparametric inference might be infinite–dimensional inference*" (L. Wasserman).

  - El modelo estadístico para los datos $f(x)$ pertenece al conjunto "$\mathcal{F}_c$" que no puede ser indexado mediante una cantidad finita de parámetros.

Example (Estimación de un funcional de f)

Asumimos que $\{X_1, \ldots, X_n, \} \overset{iid}{\sim} f$, siendo f una función de densidad continua. Queremos inferir (por ejemplo) $\theta_\nu = P(X \leq \nu) = \int_{-\infty}^{\nu} f(x)\mathrm{d}x$.
Si logramos construir un estimador $\widehat{f}(x)$, luego: $\widehat{\theta}_\nu = \int_{-\infty}^{\nu} \widehat{f}(x)\mathrm{d}x$.

Example (Regresión no paramétrica)

Asumimos que $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ es una muestra iid del modelo:

$$Y_i = f(X_i) + \varepsilon_i, \quad X_i \in [0,1], \ i = 1, \ldots, n,$$

donde $E(\varepsilon_i) = 0$ y $f : [0,1] \to \mathbb{R}$ es la **función de regresión** desconocida. No asumimos ninguna forma paramétrica específica para $E(Y|X) = f(X)$, sino que $f \in \mathcal{F}$ (un conjunto "grande" de funciones). El objetivo es construir un método "adecuado" para inferir $f(X)$ a partir de datos.

Agenda

1. Introducción

2. Estimación no paramétrica de la densidad
   - Distribución empírica e histogramas
   - Kernel density estimation

Agenda

1. Introducción

2. Estimación no paramétrica de la densidad
   - Distribución empírica e histogramas
   - Kernel density estimation

Distribución Empírica

Definition (Distribución empírica)

Sea $\{X_1, \ldots, X_n\} \overset{iid}{\sim} F$, definimos la función de distribución empírica como:

$$F_n(x_0) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x_0), \text{ para todo } x_0 \in \mathbb{R},$$

donde $\mathbb{I}$ es la función indicadora:

$$\mathbb{I}(X_i \leq x_0) = \begin{cases} 1 \text{ si } X_i \leq x_0, \\ 0 \text{ en otro caso.} \end{cases}$$

- Dada $\underline{x} = \{X_1 = x_1, \ldots, X_n = x_n\}$ asignamos $1/n$ de masa sobre cada punto muestral, de forma que: $F_n(x_0) \equiv F_n(x_0; \underline{x}) = \{\#x_i \leq x_0\}/n$.
- A una muestra en particular $\{X_1 = x_1, \ldots, X_n = x_n\}$ le corresponderá una estimación empírica de F diferente (ver diapositiva siguiente).

- Generamos 100 realizaciones de $\{X_1, \ldots, X_n\} \overset{iid}{\sim} N(0,1)$...

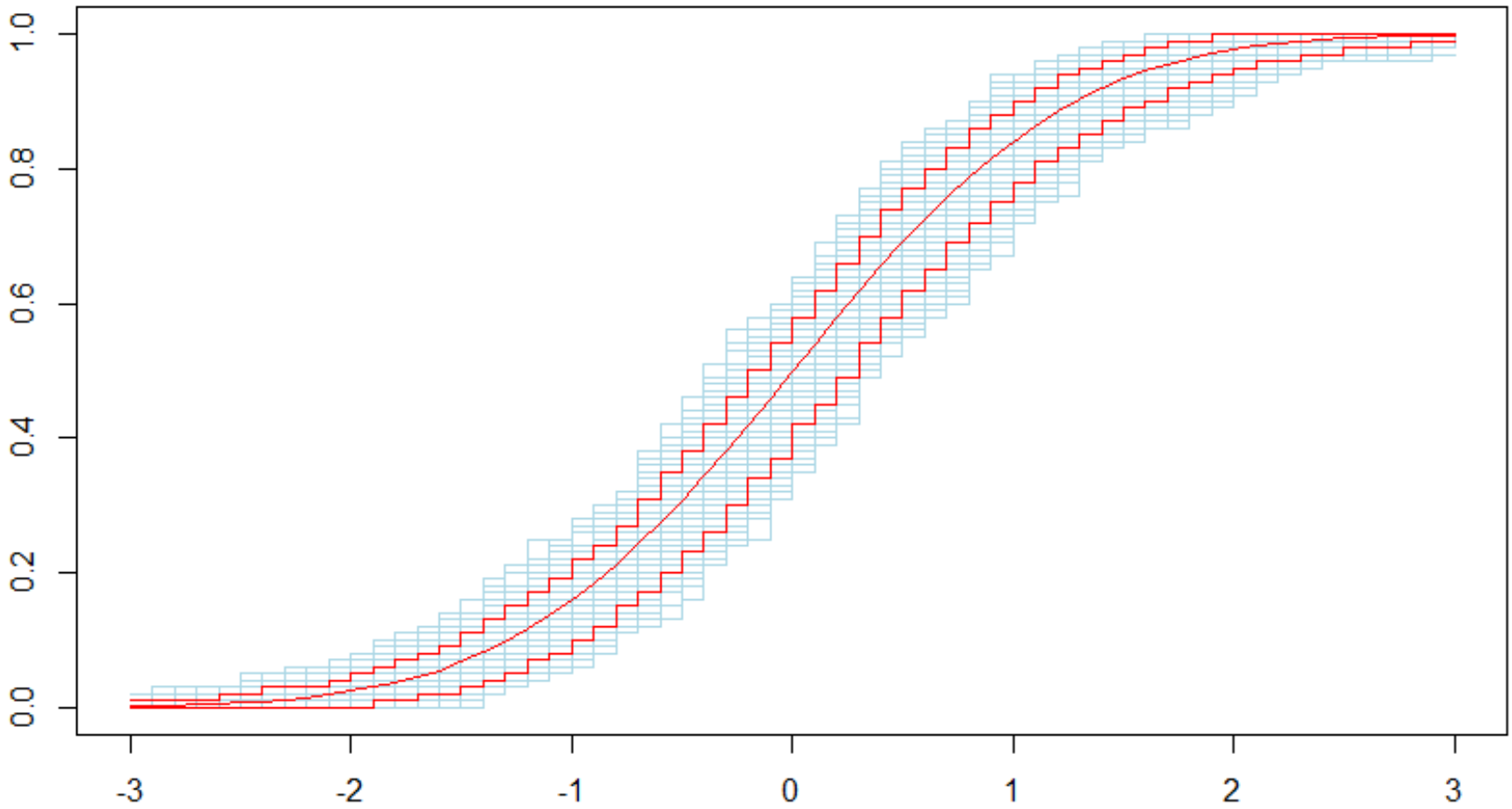


Figure: Verdadera F (rojo) y 100 estimaciones de F (intervalos empíricos).

- ... y computamos las 100 estimaciones de F relativas a cada muestra.

¿Qué propiedades tiene el estimador F_n?

- Si la verdadera F es continua, para cualquier $x \in \mathbb{R}$ fijo:

$$E(F_n(x)) = F(x), \text{ y } V(F_n(x)) = \frac{F(x)(1-F(x))}{n}.$$

- Consistencia: $\text{MSE}(F_n(x)) \overset{n\to\infty}{\longrightarrow} 0$ (¿porqué?)
- Teorema de Glivenko–Cantelli (consistencia uniforme):

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to_P 0.$$

- Dvoretzky–Kiefer–Wolfowitz (DKW) inequality: Para $\varepsilon > 0$

$$P\Big(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon\Big) \leq 2e^{-2n\varepsilon^2}.$$

- La probabilidad de que la máxima distancia entre F_n y F sea mayor a ε decrece exponencialmente con el tamaño de la muestra n.

Intervalos de confianza para F (DKW)

- Elegí α y considera $\varepsilon_n = \sqrt{\log(2/\alpha)/(2n)}$.
- Definí: $L(x) = \max\left\{F_n(x) - \varepsilon_n, 0\right\}$ y $U(x) = \min\left\{F_n(x) + \varepsilon_n, 1\right\}$.
- Luego se tiene que para toda $F(x)$ y $x \in \mathbb{R}$:

$$P(L(x) \leq F(x) \leq U(x)) \geq 1 - \alpha.$$

- $[L(x), U(x)]$ es un intervalos de confianza de nivel $1 - \alpha$ para $F(x)$.
- Dada una muestra concreta, podremos computar una estimación empírica de F y con ésta computaremos las funciones $l(x)$ y $u(x)$.
- Ilustración en `R`.

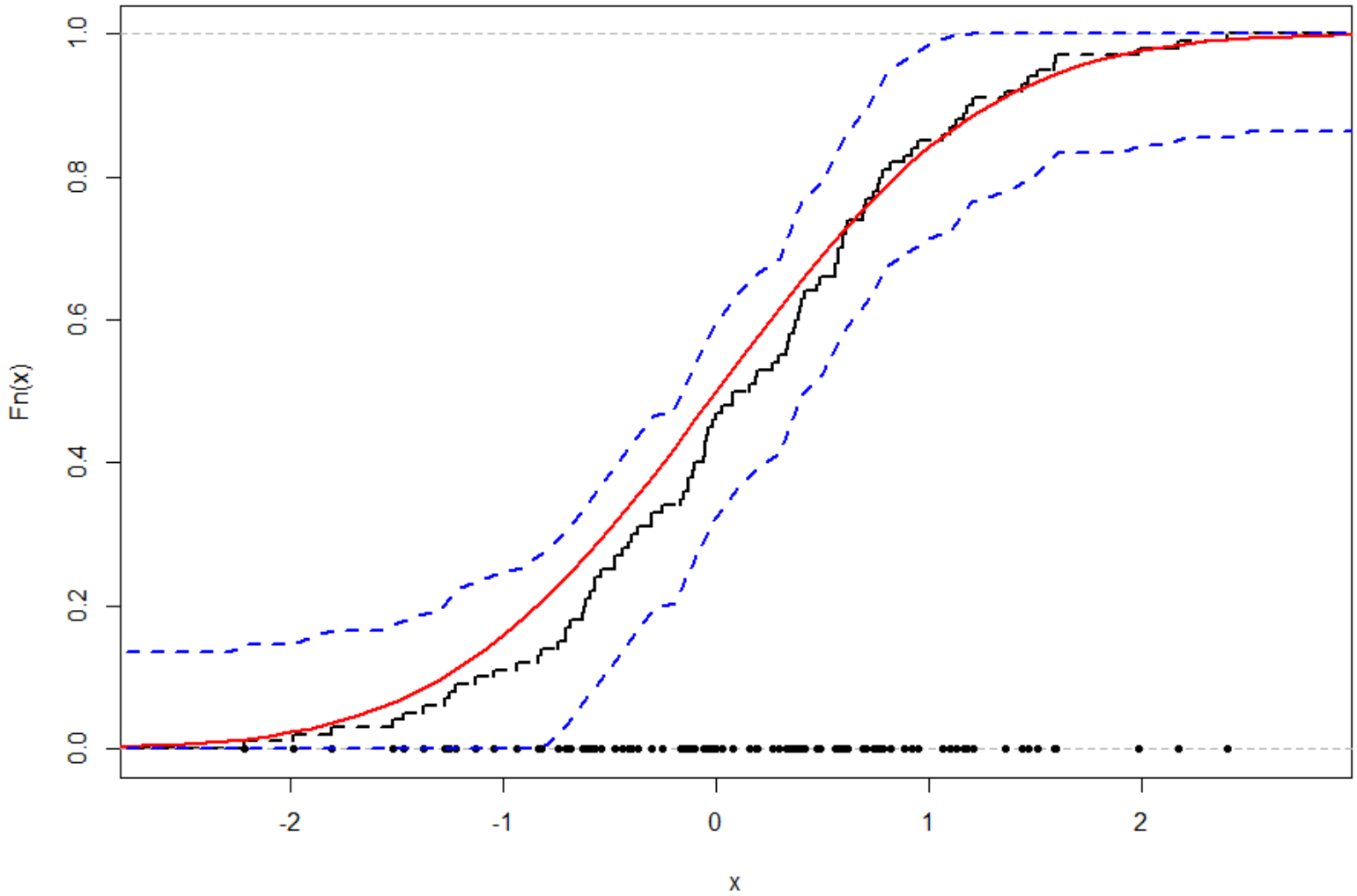


Figure: Simulamos 100 observaciones del modelo $N(\mu = 0, \sigma^2 = 1)$. Las líneas azules representan las bandas de confianza para F (`Distribución Empírica.R`).

- De la definición de distribución empírica, se tiene que:

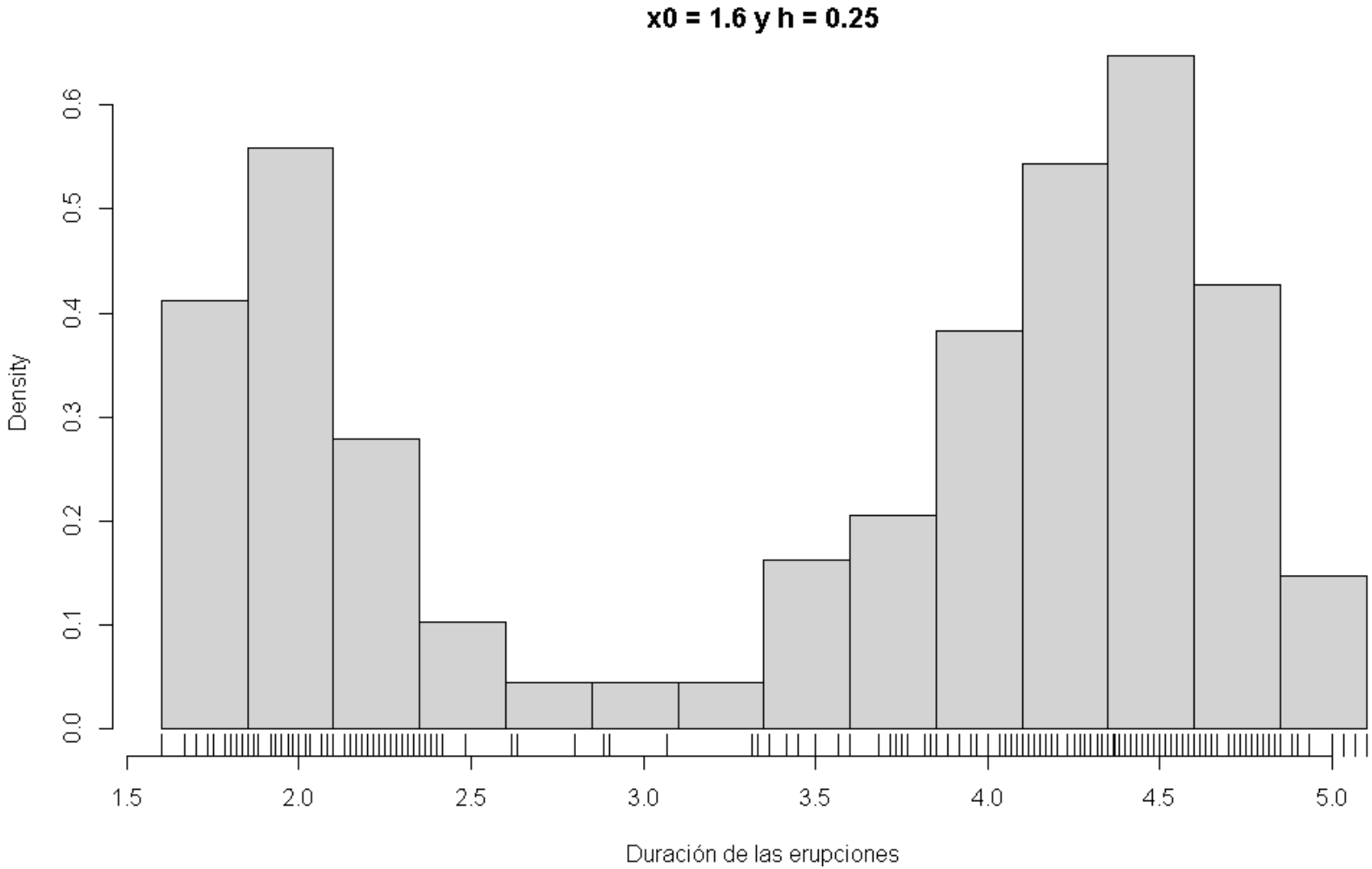
$$\widehat{f}_n(x) \equiv \widehat{f}(x; \underline{X}) = \frac{1}{n}\mathbb{I}_{x \in \underline{X}}(x).$$

  - Drawback: Discontinuidad.

- Histogramas: Particionamos el soporte de $f(x)$ en un subconjunto de intervalos (bins) de longitud h: $B_1 = [x_0, x_0 + h)$; $B_2 = [x_0 + h, x_0 + 2h)$, etc. Luego si $x \in B_k$, entonces:

$$\widehat{f}_n(x; x_0, h) \equiv \widehat{f}(x; \underline{X}, x_0, h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{x \in B_k}(X_i).$$

- Ejemplo en R.



- Los métodos de kernel suavizan los histogramas.

Agenda

- El método de estimación no paramétrico más utilizado en la práctica.
- Llamamos función de núcleo (o kernel) a una función $K : \mathbb{R} \to \mathbb{R}^+$ que verifica las siguientes 4 propiedades:

$$K(x) \geq 0, \int K(x)\mathrm{d}x = 1, \int xK(x)\mathrm{d}x = 0, \text{ y } \int x^2 K(x)\mathrm{d}x < \infty.$$

- Algunos ejemplos de núcleos:
  1. Kernel Gaussiano: $K(x) = \frac{1}{\sqrt{\pi}} e^{-x^2/2}$.
  2. Boxcar Kernel: $K(x) = \frac{1}{2}\mathbb{I}(|x| < 1)$.
  3. Epanechnikov Kernel: $K(x) = \frac{3}{4}(1 - x^2)\mathbb{I}(|x| < 1)$.
  4. Tricube Kernel: $K(x) = \frac{70}{81}(1 - |x|^3)^3\mathbb{I}(|x| < 1)$.
- Así lucen estos núcleos...

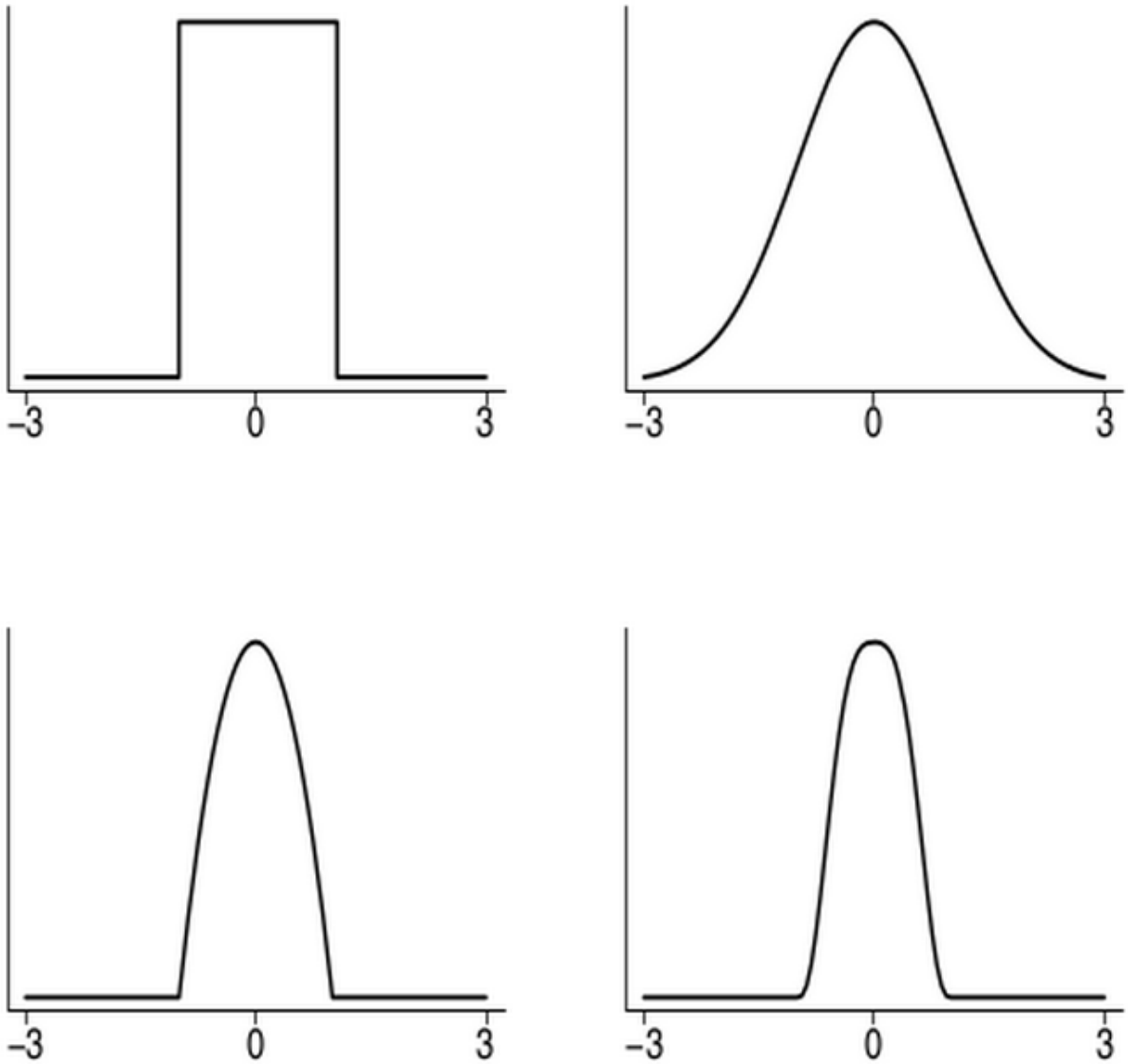


FIGURE 4.10. Examples of kernels: boxcar (top left), Gaussian (top right), Epanechnikov (bottom left), and tricube (bottom right).

- Dada $\underline{X} = \{X_1, \ldots, X_n\} \overset{iid}{\sim} f$, un núcleo K y un parámetro de banda $h > 0$, el estimador no paramétrico de la densidad se define como:

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

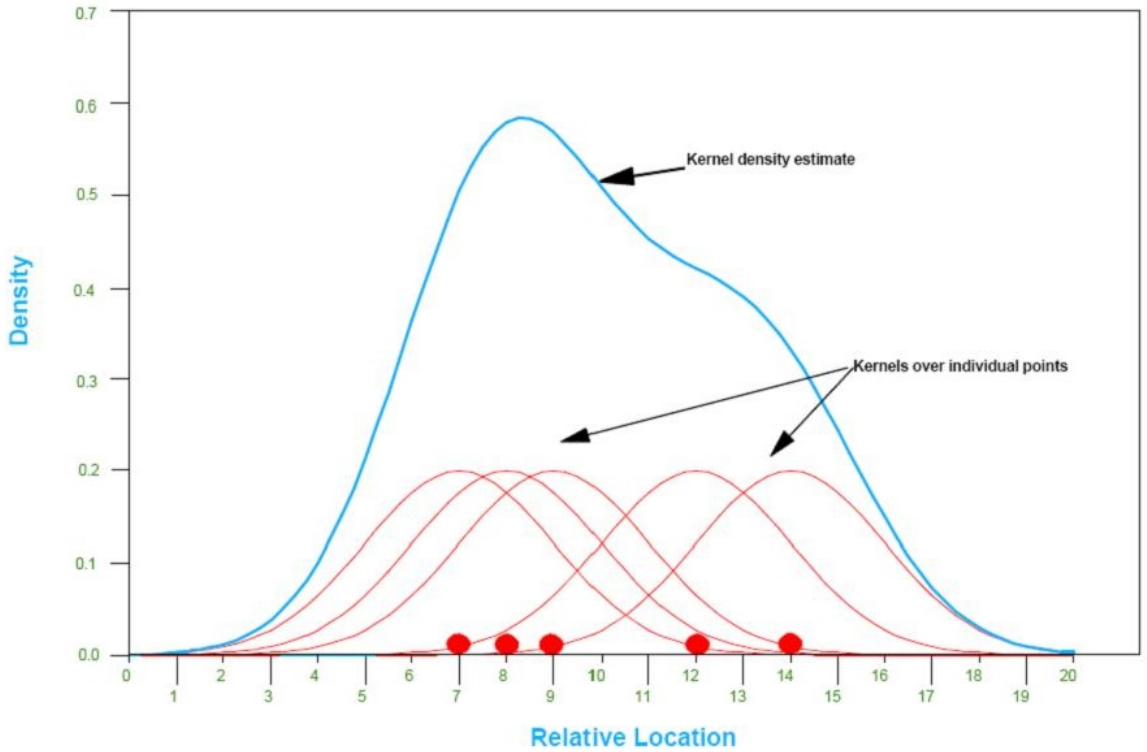
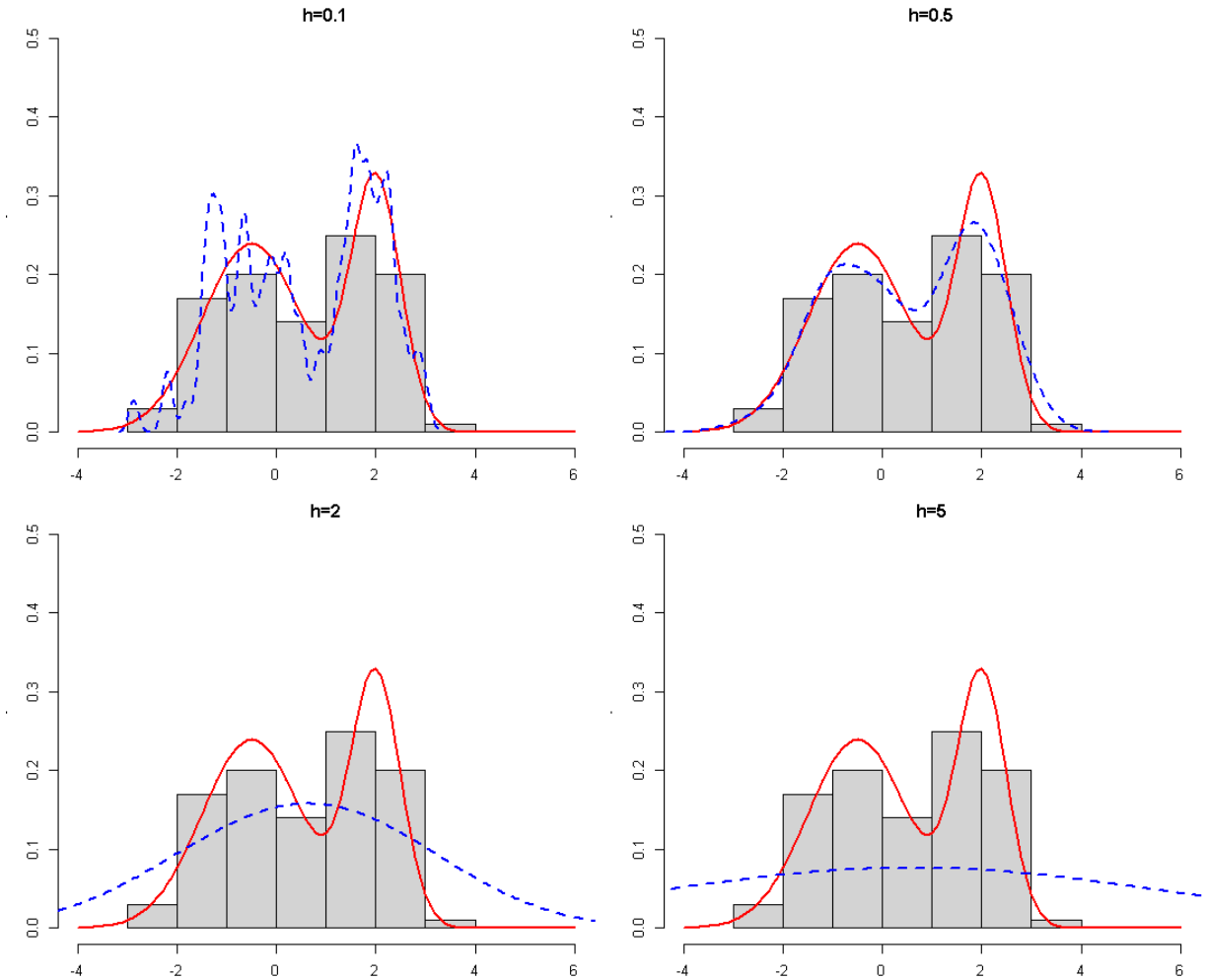


Figure: Notar que no importa como elijas K y h: $\widehat{f}_h(x) \geq 0$ y $\int \widehat{f}_h(x)\mathrm{d}x = 1$.

- La elección de K no resulta tan relevante. Elegir correctamente el valor de h nos permite evitar hacer under/oversmoothing.



Código: `Estimaciones basadas en núcleos.R`.

Consistencia

- Definamos el riesgo cuadrático del estimador en el punto $X = x$:

$$R_x(h) = E(f(x) - \widehat{f}_h(x))^2,$$

- y el riesgo a lo largo de todo el soporte como $R(h) = \int R_x(h)\mathrm{d}x$.
- Se puede demostrar que en general:

$$R(h) = \underbrace{c_{1,K} h^4 \int (f''(x))^2 \mathrm{d}x + O(h^6)}_{\text{Bias}^2(h)} + \underbrace{\frac{c_{2,K}}{nh} + O(1/n)}_{\text{Variance}(h)}.$$

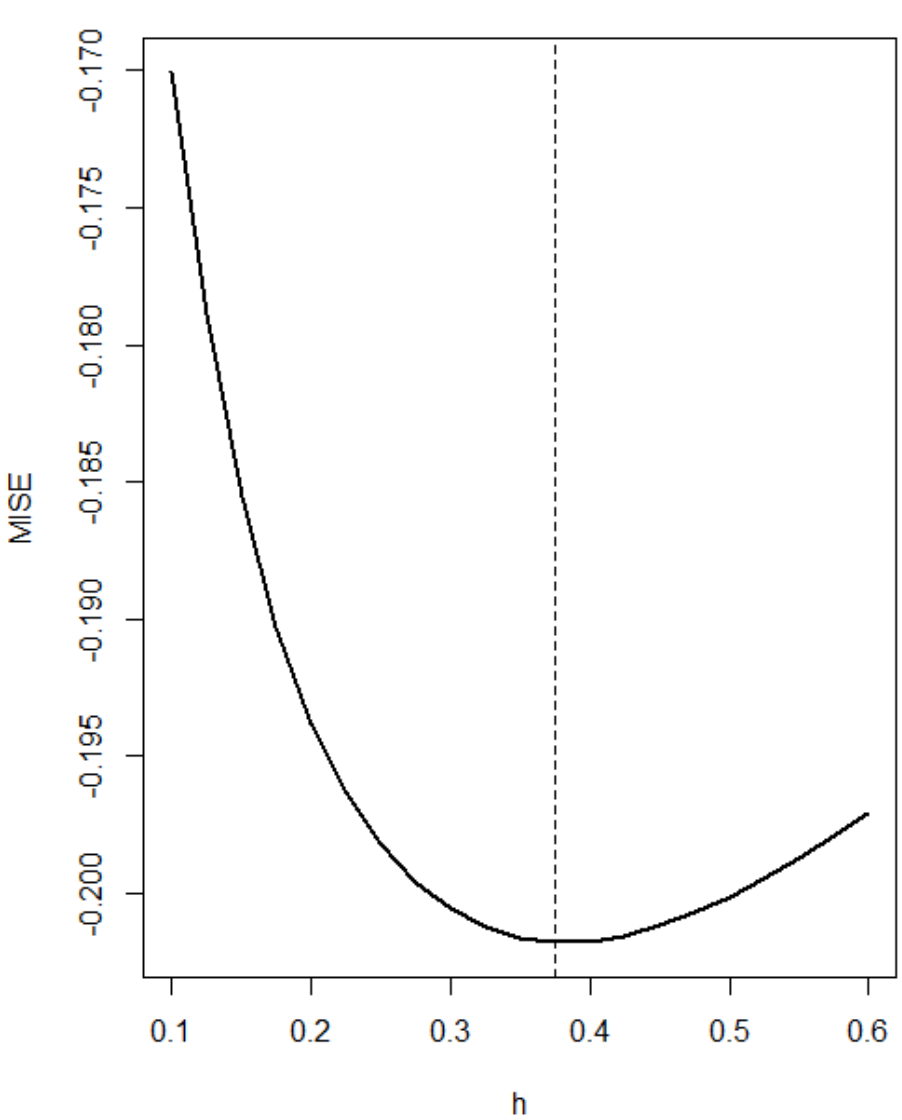
  - Con n fijo cuando $\downarrow h$, entonces $\downarrow$ Bias y $\uparrow$ Var.
  - Con n fijo cuando $\uparrow h$, entonces $\downarrow$ Var y $\uparrow$ Bias.
  - Si $h \to 0$ cuando $n \to \infty$ de tal forma que $nh \to \infty$ (h se va más despacio a 0 de lo que n se va a infinito), el estimador es consistente.

¿y cómo elijo h en la práctica? (LOO–CV)

- Bajo el esquema anterior: $h^* = (c_{2,K}/(c_{1,K}^2 \int (f''(x))^2))^{1/5}$.
- Como no conocemos f ni sus derivadas, aprendemos un h razonable a partir de los datos: Construimos una malla de posibles valores para h; y para cada valor de h estimamos el riesgo del estimador como:

$$\widehat{R}(h) = \int \widehat{f}_h^2(x)\mathrm{d}(x) - \frac{2}{n}\sum_{i=1}^{n} \widetilde{f}_{h,(-i)}(X_i).$$

- $\widetilde{f}_{h,(-i)}$ es el estimador de la densidad omitiendo X_i en la muestra.
- Existen expresiones aproximadas de $\widehat{R}(h)$ sin necesidad de recurrir a integrales numéricas (más detalles técnicos en ANPS § 6.3).
- Ilustración en `R`.



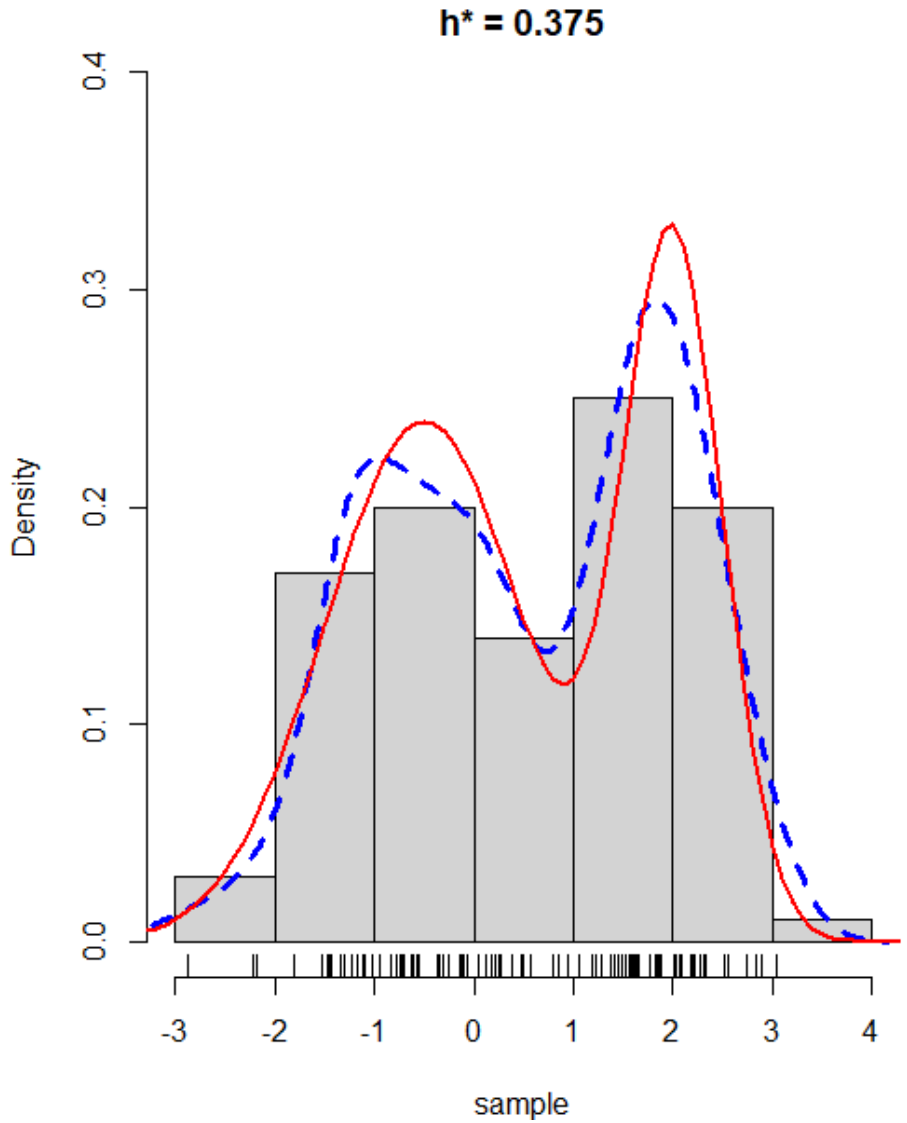


Figure: Estimación de $R(h)$ y de la densidad con $h^* = 0.375$.

Ejemplo: Distribución del ingreso (EPH)

- Estimación no paramétrica de la distribución del ingreso
- Una vez estimada $\widehat{f}_h$, podemos estimar cualquier parámetro que dependa de la verdadera f haciendo simplemente un plugg–in. Por ejemplo, estimamos $P(X \leq \nu)$ haciendo simplemente:

$$\widehat{P}(X \leq \nu) \equiv \int_0^{\nu} \widehat{f}_h(x)\mathrm{d}x.$$