Estimación No Paramétrica de la Densidad

Ana M. Bianco & Paula M. Spano

Introducción al Aprendizaje Estadístico

Muestra - Datos

- Muestra aleatoria: $X_1, \cdots, X_n$ variables aleatorias i.i.d. (independientes e idénticamente distribuidas)

- Datos u observaciones: $\mathbf{x} = x_1, \ldots, x_n$ constituyen una realización de cierta variable aleatoria.

Estadística

- Ingredientes: datos generados por un mecanismo aleatorio: por ej., tiramos una moneda al aires sucesivas veces.
- Objetivo: inferir *algo relacionado* con el mecanismo (aleatorio) que genera los datos, por ejemplo: ¿cuál es la probabilidad de obtener cara con cierta moneda?
- Mecanismo: Función de distribución.
  - Caso discreto: función de probabilidad puntual
  - Caso continuo: función de densidad
- Modus Operandi: hacer *alguna cuenta* con los datos para obtener un valor que *se parezca* al que queremos estimar.

Estadística

- Muestra: $(X_i)_{i\geq 1}$ i.i.d. $X_i \sim F$, $F \in \mathcal{F}$ familia de distribuciones posibles para nuestro problema

- Objetivo: inferir *algo relacionado* con el mecanismo que genera los datos:
  - $\mathbb{E}_F[X_1]$
  - $\mathbb{V}_F(X_1)$
  - $\mathbb{P}_F(X_1 \leq 40)$
  - F
- Proponer un estimador para cada uno de los *objetivos* planteados.

Vamos al TP2

Vayamos el Ejercicio **Análisis de datos Gamma-ray bursts: ítem 1**

La empírica

Sean $X_1, X_2, \ldots, X_n$ i.i.d., $X_i \sim F$. Definimos la función de distribución empírica como

$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq t\}}$$

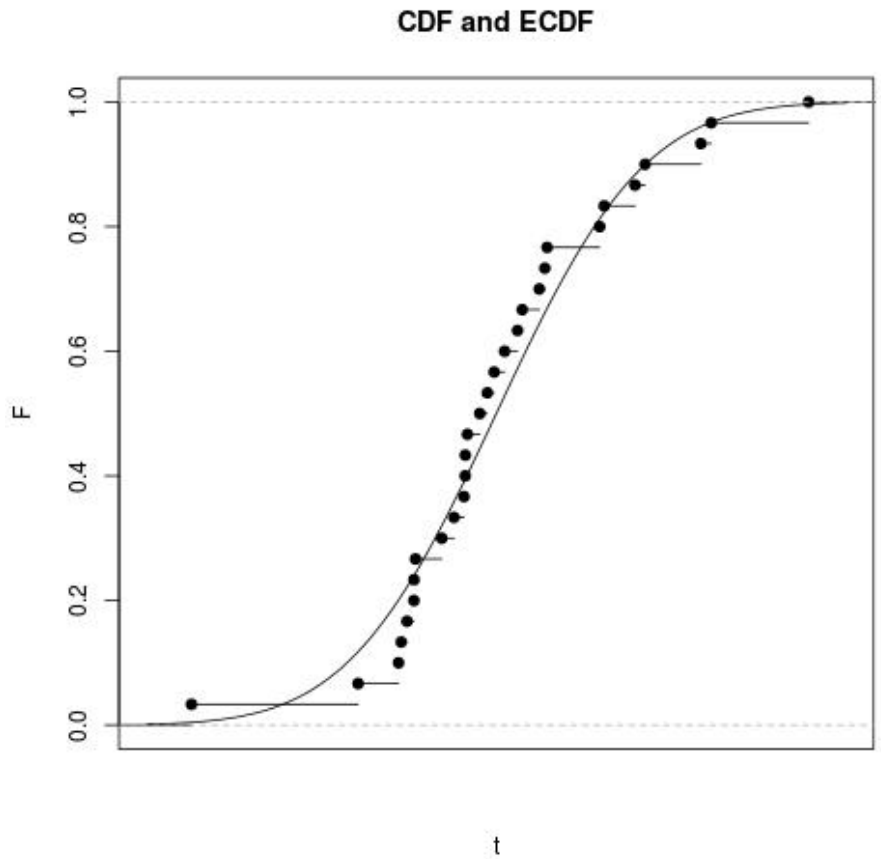
- $\widehat{F}_n(t)$ es una función aleatoria.
- $\widehat{F}_n(t)$ representa a una acumulada que da peso $1/n$ a $X_1, X_2, \ldots, X_n$.

Vamos al TP2

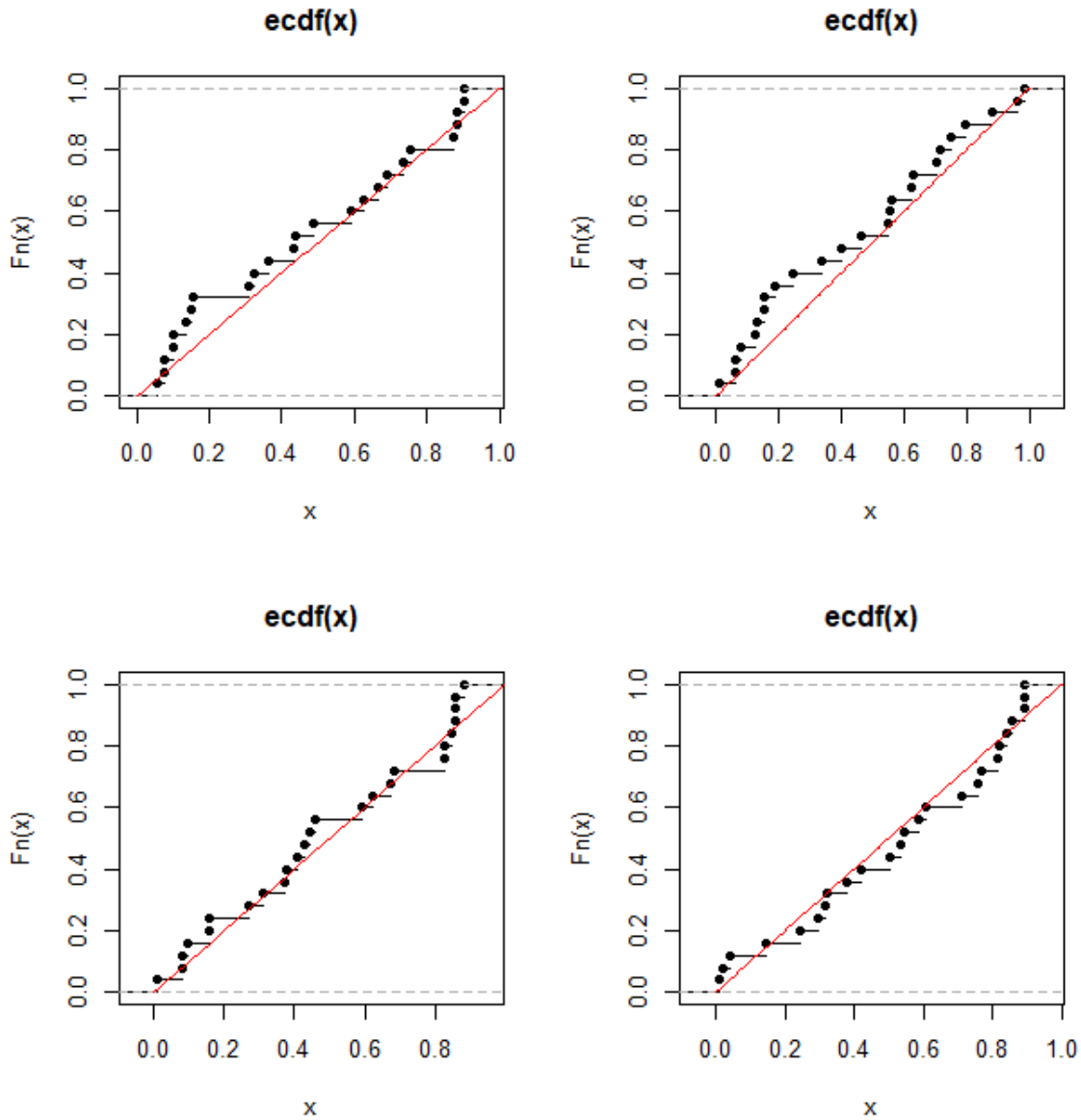
Vayamos el Ejercicio **Análisis de datos Gamma-ray bursts: ítems 2 y 3**

Empírica: una realización

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \leq t\}}$$



Datos simulados: $X_1, ..., X_{25}$ i.i.d., $X_i \sim \mathcal{U}(0,1)$



Enfoque No Paramétrico

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$.

$X \sim F$ v.a. continua con densidad $f(x)$

$\hat{F}_n =$ "la empírica"

$\widehat{f}(x) =?$

Enfoque No Paramétrico

- X con densidad $f(x)$: queremos estimar $f(x)$
- $X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$.
- Queremos estimar f sin asumir una determinada forma: **sólo asumimos que es f es suave**.

Enfoque No Paramétrico

- X con densidad $f(x)$: queremos estimar $f(x)$
- $X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$.
- Queremos estimar f sin asumir una determinada forma: **sólo asumimos que es f es suave**.
- La forma más sencilla: **Histograma**

Histograma

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

- Sea $\mathcal{C}_j$ una partición de intervalos o clases acotados (bins) disjuntos tales que:

$$\mathbb{R} = \cup_j \mathcal{C}_j$$

- Para cada $x \in \mathcal{C}_j$

$$\widehat{f}(x) = \frac{\#\{X_i : X_i \in \mathcal{C}_j\}}{n|\mathcal{C}_j|}$$

con $|\mathcal{C}_j|$ ancho del bin $\mathcal{C}_j$

Histograma

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

- Sea $\mathcal{C}_j$ una partición de intervalos o clases acotados (bins) disjuntos tales que:

$$\mathbb{R} = \cup_j \mathcal{C}_j$$

- Para cada $x \in \mathcal{C}_j$

$$\widehat{f}(x) = \frac{\#\{X_i : X_i \in \mathcal{C}_j\}}{n|\mathcal{C}_j|}$$

  con $|\mathcal{C}_j|$ ancho del bin $\mathcal{C}_j$

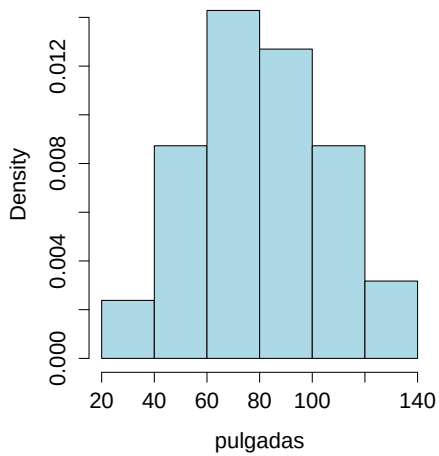
- El histograma requiere dos parámetros:
  i) ancho del bin
  ii) punto inicial del primer bin

Volvamos al TP2

Resolvemos el Ejercicio **Análisis de datos de Buffalo: items 4 y 5.**

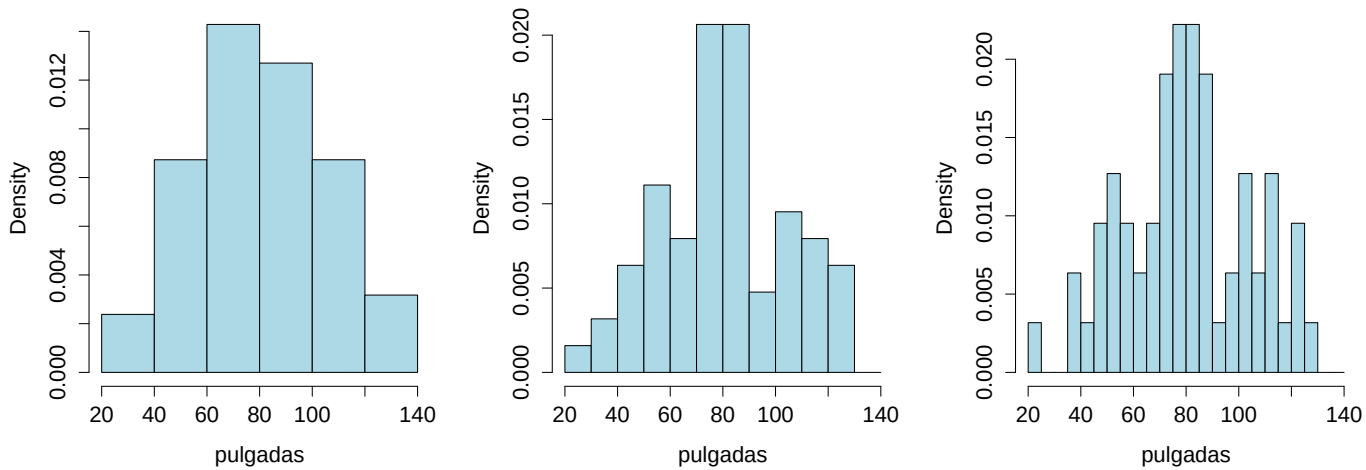
Ejemplo real

Caída de nieve anual en Buffalo (N. Y.) en inviernos entre 1910/11 to 1972/73.



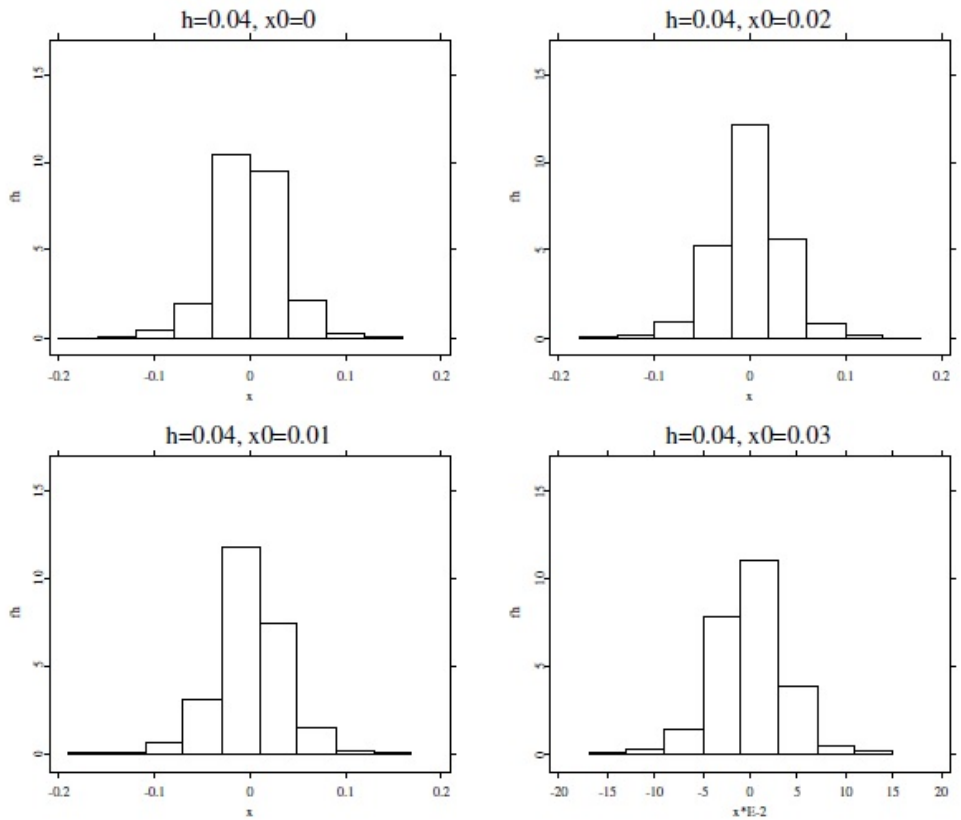
Ejemplo real

Caída de nieve anual en Buffalo (N. Y.) en inviernos entre 1910/11 to 1972/73.



Histogramas con distinto punto inicial

Datos simulados

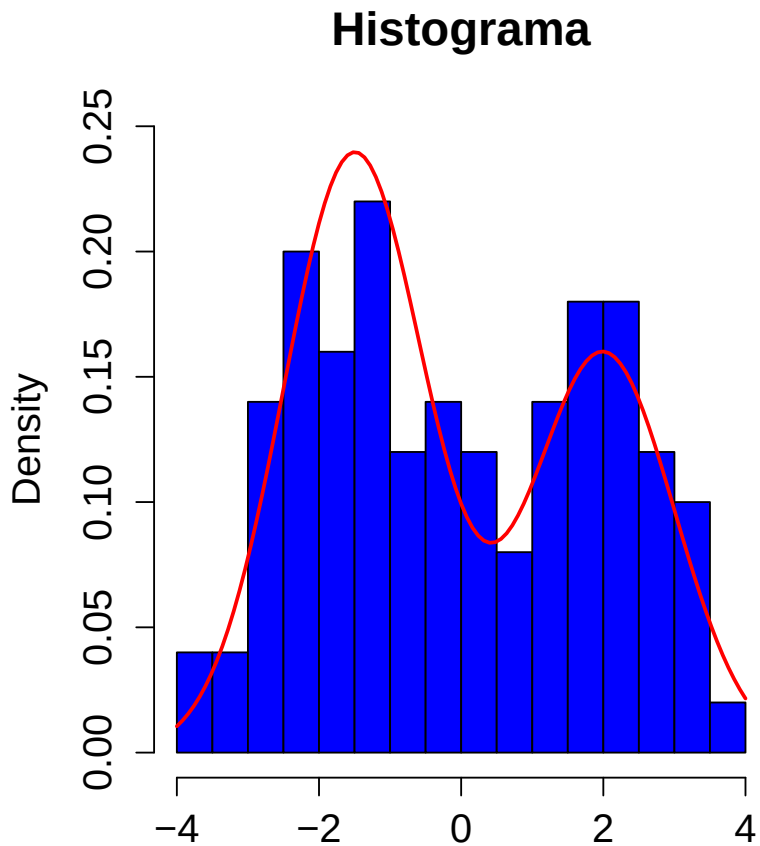


Desventajas del histograma

- el estimador de la densidad depende del punto inicial de los bins: para un número de bins fijo, la forma puede cambiar moviendo la ubicación de los bins
- la densidad estimada no es suave, es *escalonada* y esto no es propio de la densidad sino de la herramienta de estimación
- por estas razones, el histograma es usado sólo para visualización

Ejemplo: datos simulados

¿Podremos hacer algo mejor?



Busquemos otra idea...

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

- X con densidad $f(x)$: queremos estimar $f(x)$
- Queremos estimar f sin asumir una determinada forma: sólo asumimos que es f es suave.

Busquemos otra idea...

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

- X con densidad $f(x)$: queremos estimar $f(x)$
- Queremos estimar f sin asumir una determinada forma: sólo asumimos que es f es suave.

$$\mathbb{P}\left(X \in (x-h, x+h)\right) \quad = \quad \int_{x-h}^{x+h} f(t)\, dt$$

¿Cómo podemos aproximar esta probabilidad?

Idea 1: Enfoque Frecuentista

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

- X con densidad $f(x)$: queremos estimar $f(x)$
- Queremos estimar f sin asumir una determinada forma: sólo asumimos que es f es suave.

Idea 1: Enfoque Frecuentista

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

- X con densidad $f(x)$: queremos estimar $f(x)$
- Queremos estimar f sin asumir una determinada forma: sólo asumimos que es f es suave.

$$\mathbb{P}\left(X \in (x-h, x+h)\right) \quad = \quad \int_{x-h}^{x+h} f(t)\, dt$$

Idea 1: Enfoque Frecuentista

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

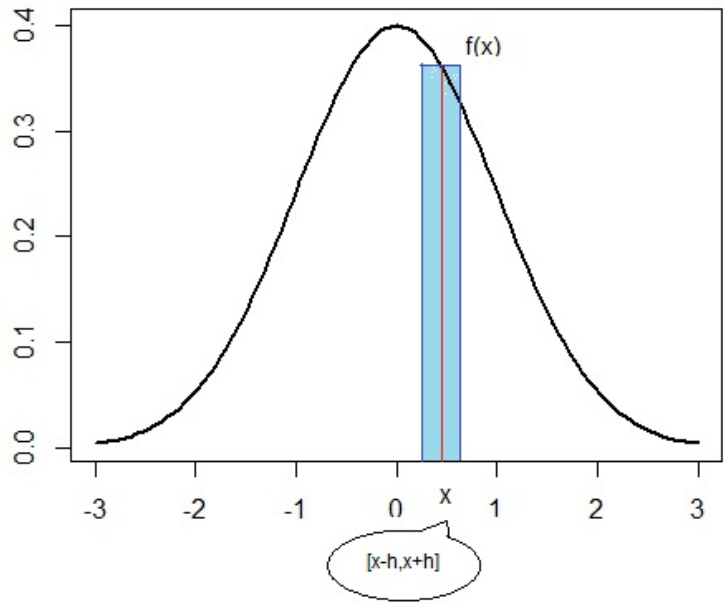
- X con densidad $f(x)$: queremos estimar $f(x)$
- Queremos estimar f sin asumir una determinada forma: sólo asumimos que es f es suave.

$$\mathbb{P}\left(X \in (x-h, x+h)\right) = \int_{x-h}^{x+h} f(t)\, dt$$

$$\mathbb{P}\left(X \in (x-h, x+h)\right) \approx \frac{\#\{X_i \in (x-h, x+h)\}}{n}$$

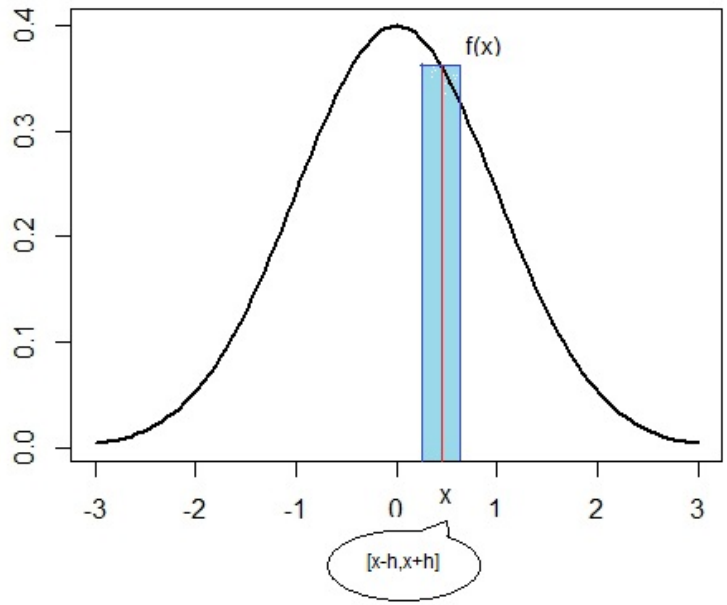
Idea 2: Enfoque analítico

- $\mathbb{P}(X \in (x-h, x+h)) = \int_{x-h}^{x+h} f(t)\,dt$
- Si h es pequeño y f continua en x,



Idea 2: Enfoque analítico

- $\mathbb{P}\left(X \in (x-h, x+h)\right) = \int_{x-h}^{x+h} f(t)\, dt$
- Si h es pequeño y f continua en x,



$$\int_{x-h}^{x+h} f(t)\, dt \quad \approx \quad 2hf(x)$$

Juntemos las dos ideas...

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

$$\mathbb{P}\left(X \in (x-h, x+h)\right) = \int_{x-h}^{x+h} f(t)\, dt$$

- $\mathbb{P}\left(X \in (x-h, x+h)\right) \approx \frac{\#\{X_i \in (x-h, x+h)\}}{n}$
- $\mathbb{P}\left(X \in (x-h, x+h)\right) \approx 2h\, f(x)$

Juntemos las dos ideas...

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

$$\mathbb{P}\left(X \in (x-h, x+h)\right) = \int_{x-h}^{x+h} f(t)\, dt$$

- $\mathbb{P}\left(X \in (x-h, x+h)\right) \approx \frac{\#\{X_i \in (x-h, x+h)\}}{n}$
- $\mathbb{P}\left(X \in (x-h, x+h)\right) \approx 2h\, f(x)$
- Entonces, podemos aproximar analíticamente

$$2h\ f(x) \approx \mathbb{P}\left(X \in (x-h, x+h)\right) \approx \frac{\#\{X_i \in (x-h, x+h)\}}{n}$$

Juntemos las dos ideas...

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

$$\mathbb{P}\left(X \in (x-h, x+h)\right) = \int_{x-h}^{x+h} f(t)\, dt$$

- $\mathbb{P}\left(X \in (x-h, x+h)\right) \approx \frac{\#\{X_i \in (x-h, x+h)\}}{n}$
- $\mathbb{P}\left(X \in (x-h, x+h)\right) \approx 2h\, f(x)$
- Entonces, podemos aproximar analíticamente

$$2h\ f(x) \approx \mathbb{P}\left(X \in (x-h, x+h)\right) \approx \frac{\#\{X_i \in (x-h, x+h)\}}{n}$$

$$f(x) \approx \frac{\#\{X_i \in (x-h, x+h)\}}{2\,h\,n}$$

Propuesta

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

$$\widehat{f}(x) = \frac{\#\{X_i \in (x-h, x+h)\}}{2h\,n}$$

Propuesta

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

$$\widehat{f}(x) = \frac{\#\{X_i \in (x-h, x+h)\}}{2h\,n}$$

Notemos que
- $\widehat{f}(x) \geq 0$
- $\int \widehat{f}(x)dx = 1$

Propuesta

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

$$\widehat{f}(x) = \frac{\#\{X_i \in (x-h, x+h)\}}{2h\,n}$$

Notemos que
- $\widehat{f}(x) \geq 0$
- $\int \widehat{f}(x)dx = 1$

Propuesta

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

$$\widehat{f}(x) = \frac{\#\{X_i \in (x-h, x+h)\}}{2h\,n}$$

Notemos que $\bullet\ \widehat{f}(x) \geq 0$ $\bullet\ \int \widehat{f}(x)dx = 1$

$$\widehat{f}(x) = \frac{1}{2h\,n} \sum_{i=1}^{n} \mathcal{I}_{(x-h,x+h)}(X_i)$$

- Estimador de Parzen

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} \mathcal{I}_{[-1,1]} \left(\frac{x - X_i}{h} \right)$$

Propuesta

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

$$\widehat{f}(x) = \frac{\#\{X_i \in (x-h, x+h)\}}{2h\,n}$$

Notemos que $\bullet\ \widehat{f}(x) \geq 0$ $\bullet\ \int \widehat{f}(x)dx = 1$

$$\widehat{f}(x) = \frac{1}{2h\,n} \sum_{i=1}^{n} \mathcal{I}_{(x-h,x+h)}(X_i)$$

- Estimador de Parzen

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} \mathcal{I}_{[-1,1]} \left(\frac{x - X_i}{h} \right)$$

- si $K(t) = \frac{1}{2}\mathcal{I}_{[-1,1]}(t) \quad \Rightarrow$

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

Propuesta

$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

$$\widehat{f}(x) = \frac{\#\{X_i \in (x-h, x+h)\}}{2h\,n}$$

Estimador de Parzen

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} \mathcal{I}_{[-1,1]} \left(\frac{x - X_i}{h} \right)$$

Si $K(t) = \frac{1}{2}\mathcal{I}_{[-1,1]}(t) \quad \Rightarrow$

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

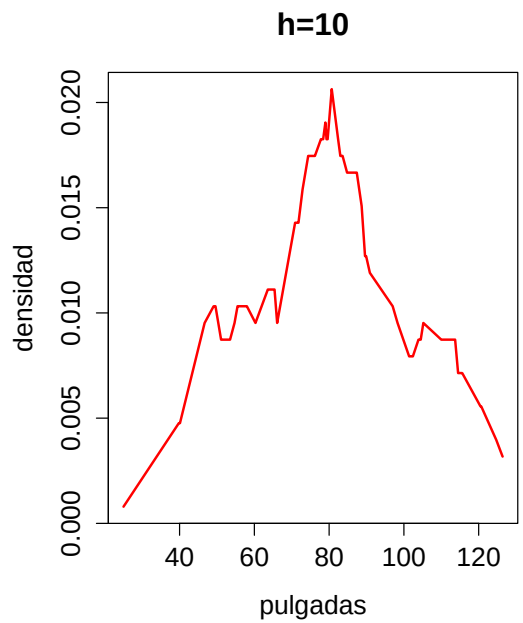
Juntando todo...

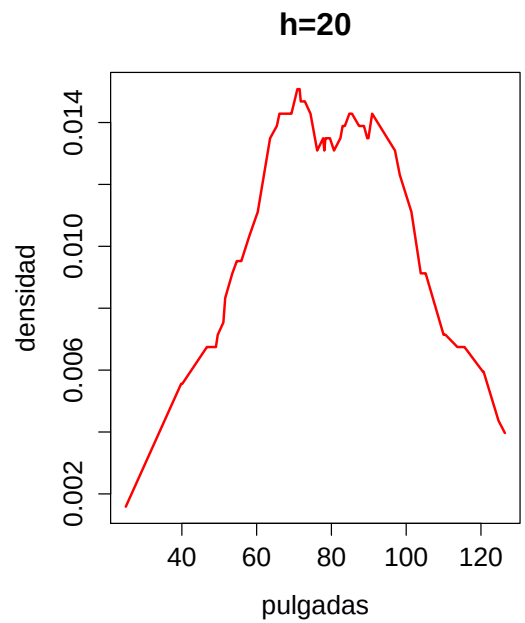
- $K(t) = \frac{1}{2}\mathcal{I}_{[-1,1]}(t) \quad \Rightarrow \quad \widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$
  - K : núcleo
  - h : ventana

Vayamos a resolver los ítems 6 a 10 del TP2.

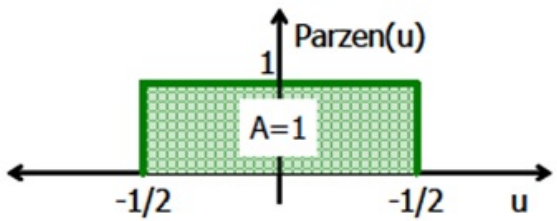
Juntando todo...

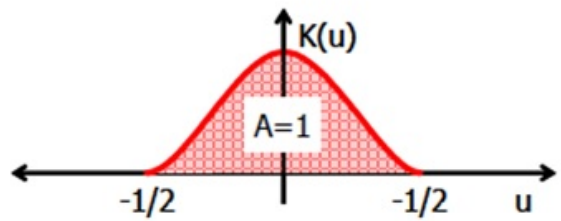
- $K(t) = \frac{1}{2}\mathcal{I}_{[-1,1]}(t) \quad \Rightarrow \quad \widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$
  - K : núcleo
  - h : ventana





Núcleos

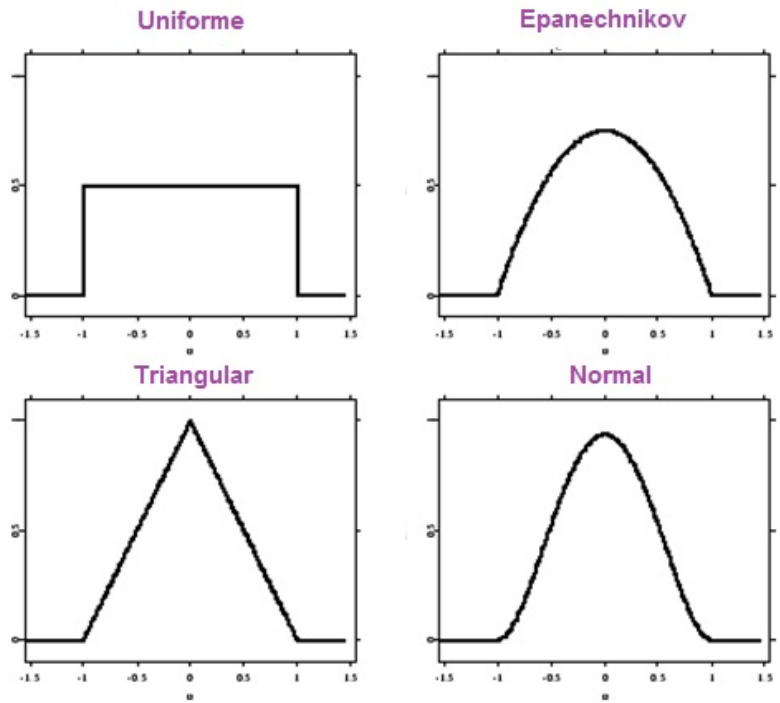




Tipos de núcleos

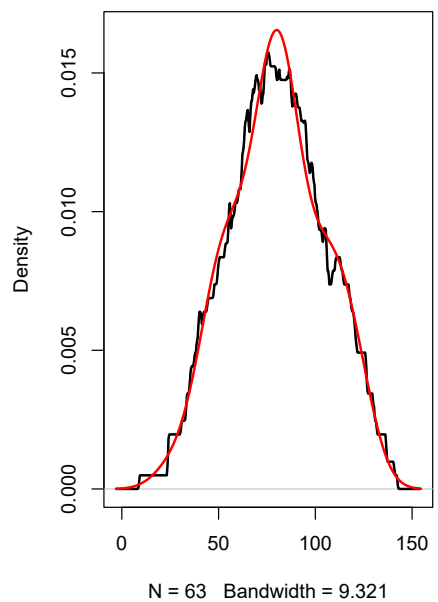
- Núcleo Rectangular: $K(t) = \frac{1}{2}\mathcal{I}_{[-1,1]}(t)$
- Núcleo Triangular: $K(t) = (1 - |t|)\mathcal{I}_{[-1,1]}(t)$
- Núcleo Gausssiano: $K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$
- Núcleo Epanechnikov: $K(t) = \frac{3}{4}(1 - t^2)\mathcal{I}_{[-1,1]}(t)$

Núcleos



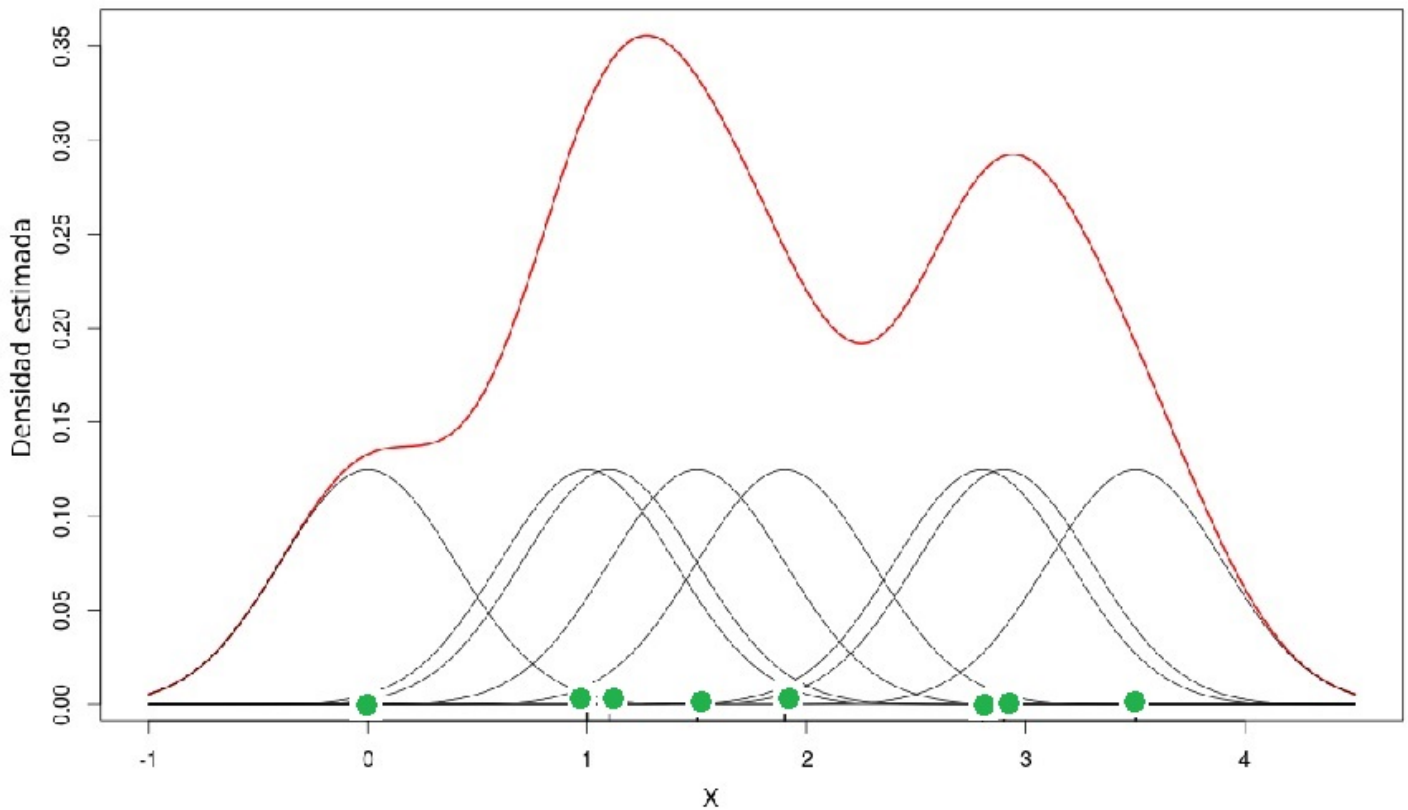
Comandos de R

```
nieve=scan()
126.4  82.4  78.1  51.1  90.9  76.2  104.5  ...

density(nieve,from=40,to=40,n=1,kernel="rectangular",bw=5)$y
[1]  0.003665716

pp.rec=density(nieve,kernel="rectangular",window=5)
pp.nor=density(nieve,kernel="gaussian",window=5)

plot(pp rec )
lines(pp.rec$x,pp.rec$y,type="l",col="black",lwd=2)
lines(pp.nor$x,pp.nor$y,type="l",col="red",lwd=2)
```



Interpretación del estimador de núcleos

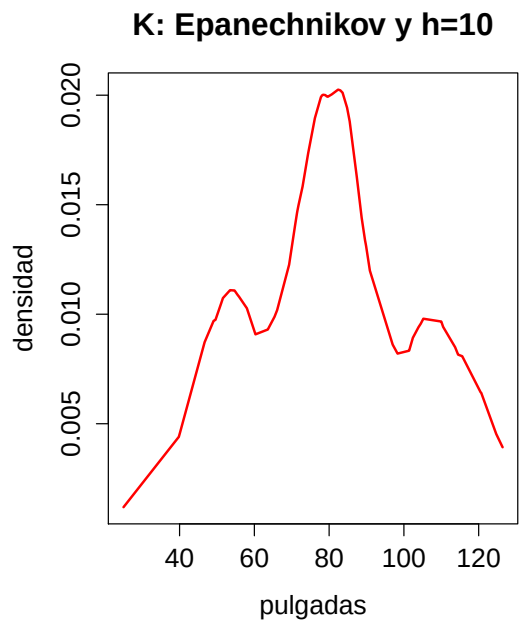
Fuente: Tesis de Lic. en Cs. Matem. de Sofía Ruiz, 2016.

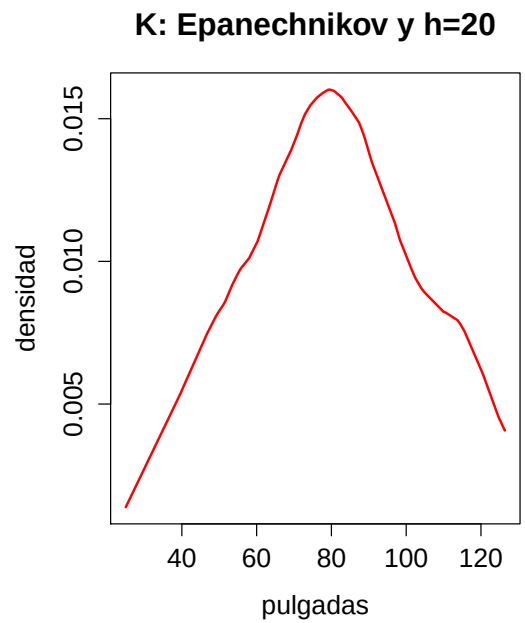


Estimadores de núcleos (Rosenblatt-Parzen)

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

- K núcleo: * $K \geq 0$ y * $\int K(x)dx = 1$.
- h: ventana o parámetro de suavizado
- Notemos que $\widehat{f}(x)$ depende de n, del núcleo K y de h





Visitemos

https://shinyserv.es/shiny/kde/