Relación Datos – Muestra

En Probabilidad: problema típico

Una moneda tiene probabilidad de cara 0.3, si la lanzamos 10 veces

¿qué probabilidad tenemos de que ocurra esto?

¿Cuál es la probabilidad de cara?

Lanzamos 10 veces una moneda

¿Qué valor proponemos para la probabilidad de cara?

¿Cuál es la probabilidad de cara?

Lanzamos 10 veces una moneda

Las 10 observaciones podrían haber sido:

H H T T H T T T T T $\rightarrow 1\,1\,0\,0\,1\,0\,0\,0\,0\,0$

o bien

T T T T H T T T H H $\rightarrow 0\,0\,0\,0\,1\,0\,0\,0\,1\,1$

o bien

H T H T T H T T T T $\rightarrow 1\,0\,1\,0\,0\,1\,0\,0\,0\,0$

Modelo

Genéricamente los datos podemos representarlos como

$$x_1, x_2, \ldots, x_n$$

En nuestro caso son 1 y 0 y $n = 10$.

Modelo

Genéricamente los datos podemos representarlos como

$$x_1, x_2, \ldots, x_n$$

En nuestro caso son 1 y 0 y $n = 10$.

Propongamos un modelo para estos datos.

Llamaremos θ a la probabilidad de cara: magnitud desconocida

Muestra - Datos (Observaciones)

- Datos - Observaciones $x_1, \ldots, x_n$: Números.

  Datos-Observaciones: son los resultados obtenidos al realizar el "experimento"

Muestra - Datos (Observaciones)

- Datos - Observaciones $x_1, \ldots, x_n$: Números.

  Datos-Observaciones: son los resultados obtenidos al realizar el "experimento"

- Muestra $X_1, \ldots, X_n$: Variables aleatorias.

  Datos-Observaciones: son realizaciones de las variables aleatorias

Datos – Muestra

- Muestra aleatoria: $X_1, \cdots, X_n$ variables aleatorias i.i.d.

  i.i.d.: independientes e idénticamente distribuidas

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

¿Cómo estimamos la probabilidad de cara?

¿¿Propuestas??

¿Cómo estimamos la probabilidad de cara?

¿¿Propuestas??

$$\widehat{\theta}_n = \overline{X}_n$$

- El estimador es una variable aleatoria!!!
- $0.3 = \frac{3}{10}$ es la estimación que obtenemos con nuestros datos!!!

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

- Estimador: es una función de la muestra que permite aproximarnos al valor 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:
    - $\mathbb{P}_F(X_1 \leq 20)$
    - F
    - $\mathbb{E}_F(X_1)$
    - $\mathbb{V}_F(X_1)$
- ¿Cómo estimaríamos para cada uno de los *objetivos* planteados?

Ejemplo: Datos de Páncreas

Vayamos al Ejercicio 1 de los **Datos de Páncreas** de la Guía de TP.

Supongamos que nuestro objetivo es estudiar a los pacientes con diagnóstico 3:

¿Cómo estimaríamos la probabilidad de que tenga un valor del marcador $REG1B \leq 20$?

Ejemplo: Datos de Páncreas

Vayamos al Ejercicio 1 de los **Datos de Páncreas** de la Guía de TP.

Supongamos que nuestro objetivo es estudiar a los pacientes con diagnóstico 3:

¿Cómo estimaríamos la probabilidad de que tenga un valor del marcador $REG1B \leq 20$?
En R

```
REG1B_D3<-REG1B[diagnosis==3]
sum(REG1B_D3<= 20)/length(REG1B_D3)

0.1507538
```

Ejemplo: Datos de Páncreas

Vayamos al Ejercicio 1 de los **Datos de Páncreas** de la Guía de TP.

Supongamos que nuestro objetivo es estudiar a los pacientes con diagnóstico 3:

¿Cómo estimaríamos la probabilidad de que tenga un valor del marcador $REG1B \leq 20$?
En R

```
REG1B_D3<-REG1B[diagnosis==3]
sum(REG1B_D3<= 20)/length(REG1B_D3)
```

```
0.1507538
```

Observación: Estamos calculando la empírica.

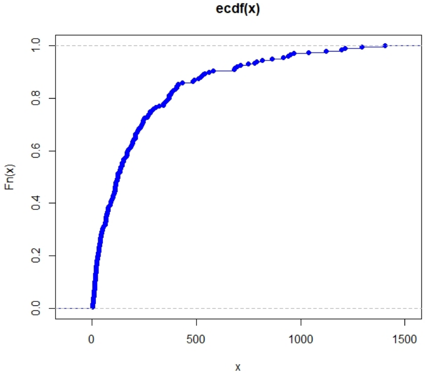
La empírica

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

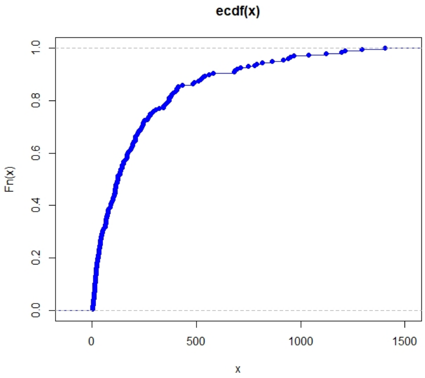
$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \le t\}}$$

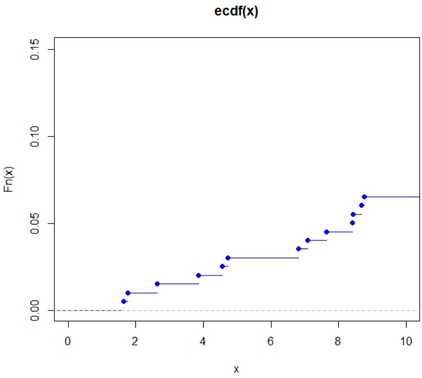
- $\widehat{F}_n(t)$ es una función aleatoria.
- $\widehat{F}_n(t)$ representa a una acumulada que da peso $1/n$ a $X_1, X_2, \ldots, X_n$.

La empírica



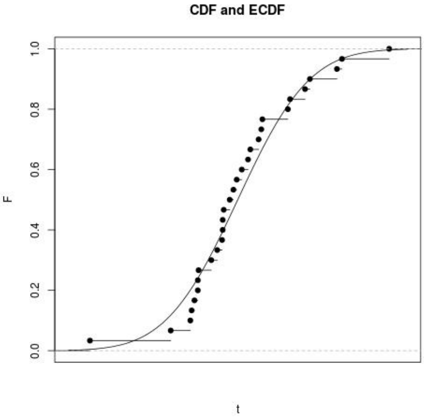
La empírica


ecdf(x)
Fn(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500

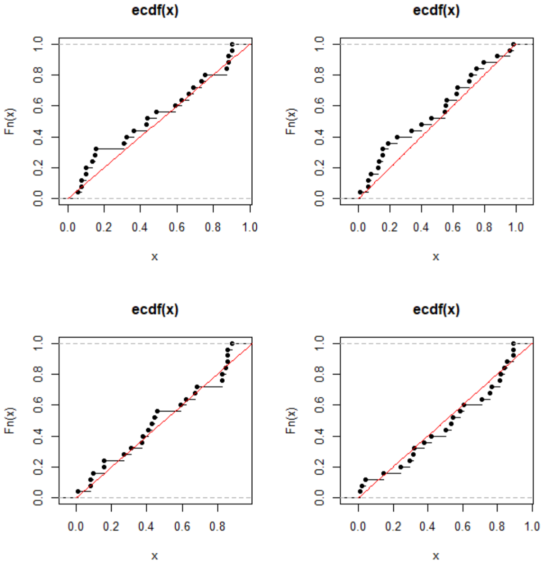

ecdf(x)
Fn(x)
x
0.00
0.05
0.10
0.15
0
2
4
6
8
10

Empírica: una realización

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



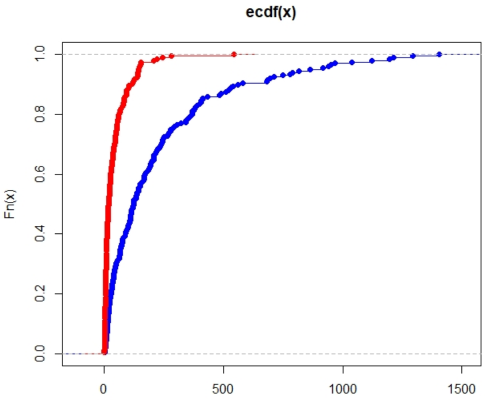
Datos simulados: $X_1, ..., X_{25}$ i.i.d., $X_i \sim \mathcal{U}(0,1)$



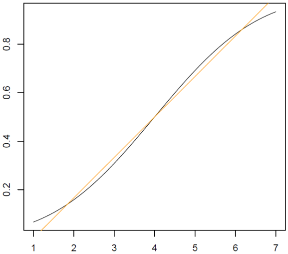
Datos de Páncreas

Comparemos los pacientes de control con los de diagnóstico 3:

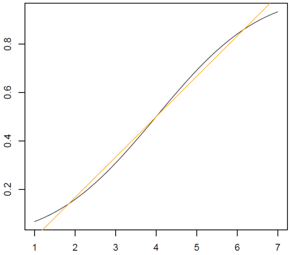
```
Fn1<-   ecdf(REG1B[diagnosis==1])
Fn3<-   ecdf(REG1B[diagnosis==3])
plot(Fn3,main="ecdf(x)",col="blue")
lines(Fn1,col="red")
```

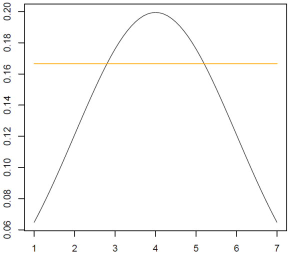


¿Cuál es cual?



¿Cuál es cual?





$$Y \sim N(4,2) \quad Z \sim U(1,7)$$

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"

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

Estimación No Paramétrica de la Densidad

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: **solo asumimos que 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: **solo asumimos que 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 $X_i \sim X$

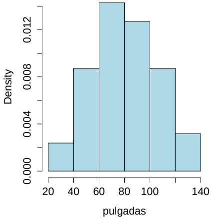
- Sea $\mathcal{C}_j$ una partición de intervalos o clases acotados (bins) disjuntos tales que:
$$\mathbb{R} = \cup_j \mathcal{C}_j$$

- 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) <u>ancho del bin</u>
ii) <u>punto inicial del primer bin</u>

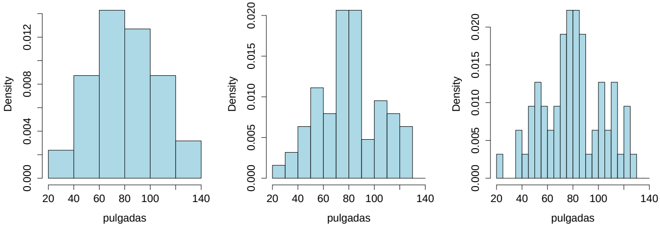
Ejemplo real

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



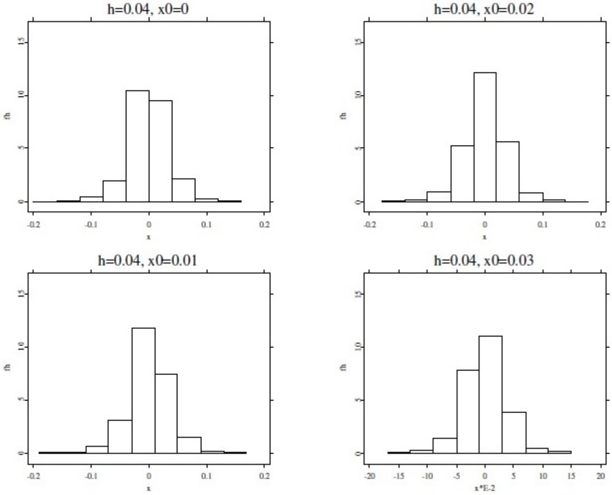
Ejemplo real

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



Histogramas con distinto punto inicial

Datos simulados

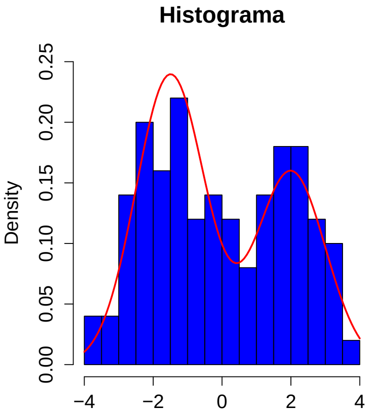


Desventajas del histograma

- 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.

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

¿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.

$$\mathbb{P}\left(X \in [x-h, x+h]\right) \;=\; \int\limits_{x-h}^{x+h} f(t)\, dt$$

Idea 1: Enfoque Frecuentista

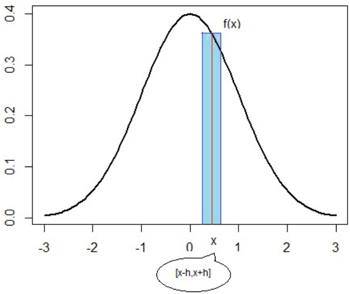
$X_1, \ldots, X_n$, i.i.d., donde $X_i \sim X$

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

$$\begin{aligned}
\mathbb{P}\left(X \in [x-h, x+h]\right) &= \int_{x-h}^{x+h} f(t)\, dt \\
\mathbb{P}\left(X \in [x-h, x+h]\right) &\approx \frac{\#\{X_i \in [x-h, x+h]\}}{n}
\end{aligned}$$

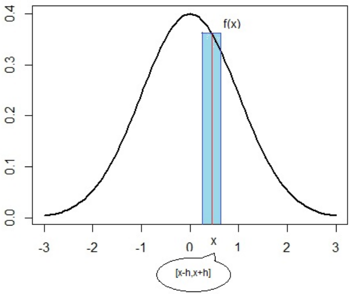
Idea 2: Enfoque analítico

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



Idea 2: Enfoque analítico

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



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

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 $X_i \sim X$

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

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

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

Hagamos algunas pruebas y pongamos manos a la obra en:
`https://glmconr2.shinyapps.io/app_regre2/`

Propuesta

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

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

Propuesta

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

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

Podemos reescribir esta expresión como:

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

- Estimador de Parzen

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

Propuesta

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

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

Podemos reescribir esta expresión como:

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

- Estimador de Parzen

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

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

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

Propuesta

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

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

Estimador de Parzen

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

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

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

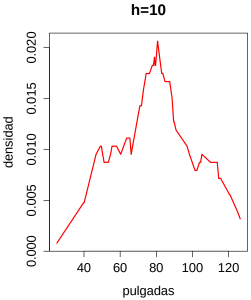
Probar que $\quad$ • $\widehat{f}_h(x) \geq 0$ $\quad$ • $\int \widehat{f}_h(x) dx = 1$

Juntando todo...

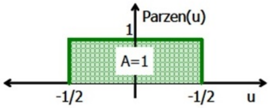
- $K(t) = \frac{1}{2}\mathcal{I}_{[-1,1]}(t) \quad \Rightarrow \quad \widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$
  - K : núcleo
  - h : ventana

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



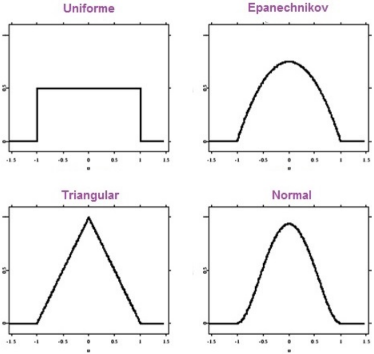
Núcleos



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

Núcleos



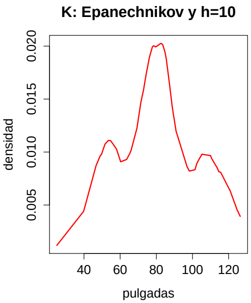
Volvamos al shiny a probar con el núcleo normal:

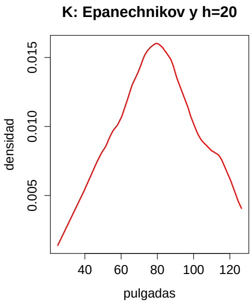
`https://glmconr2.shinyapps.io/app_regre2/`

Estimadores de núcleos (Rosenblatt-Parzen)

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

- 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





Dejamos una pregunta planteada....

¿Cómo elegimos h?

Interpretación del estimador de núcleos

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