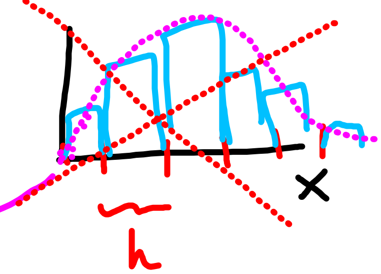
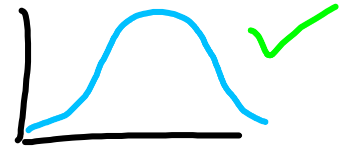
$f(x) \sim \hat{f}_n(x)$





2. Estimación No-paramétrica de densidad

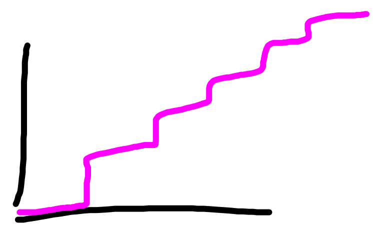
2.1. Primera construcción

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

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



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

[redacted]

Pregunta 2.1

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

La respuesta es si (más o menos).

h peq.

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

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

$$\hat{F}(x) = \frac{1}{n}\sum I(X_i < x)$$

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

Podemos rescribirlo de la forma,

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

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

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

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 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 $[x - h, x + h)$.

$\cancel{h}$

$\underbrace{x-h \quad x \quad x+h}_{2h}$

El histograma se escribe como

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

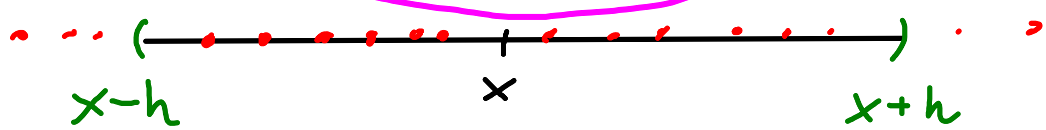
Ahora tratemos de modificar ligeramente esta expresión notando dos cosas

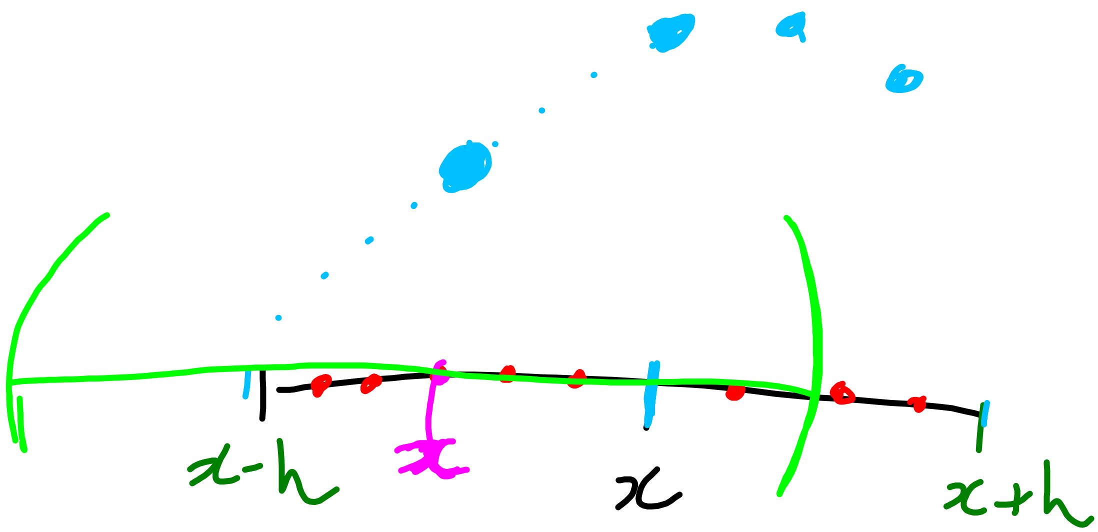
1.
$$K(u) = \frac{1}{2}I\left(|u| \leq 1\right)$$
con $u = \frac{x-xi}{h}$

2.
$$\frac{1}{2}\#\{X_i \in [x - h, x + h)\} = \underbrace{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)} = \sum_{i=1}^{n} \frac{1}{2} I\underbrace{\left(\left|\frac{x - x_i}{h}\right| \leq 1\right)}_{\text{\# de valores en el intervalo}}$$

Finalmente se tiene que

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





Pregunta 2.3

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

Esta función debería cumplir las siguientes características

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

K es una densidad

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

Simétrica

$\boxed{K(u) = K(-u)}$

Por ejemplo:

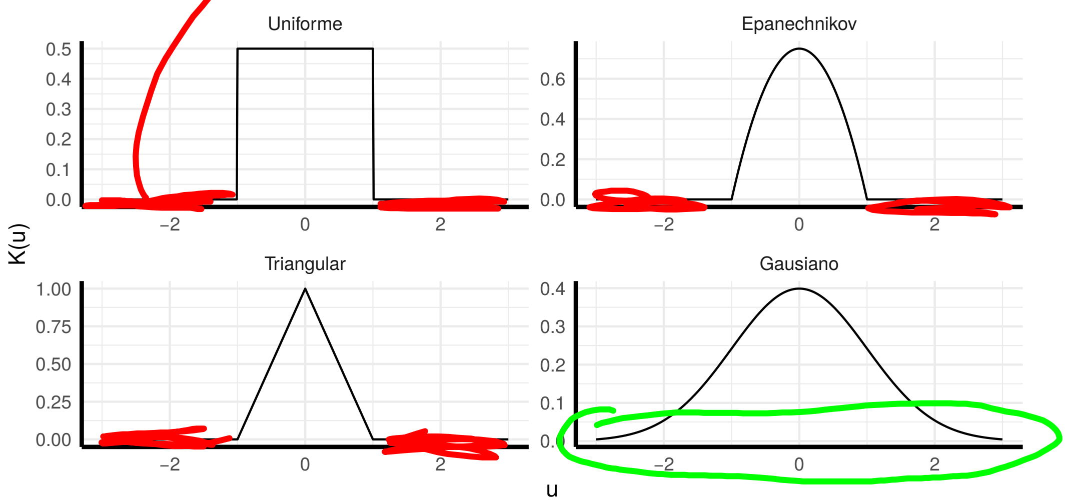
Uniforme: $\frac{1}{2}I\left(|u| \leq 1\right)$.

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

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

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

soporte = 0



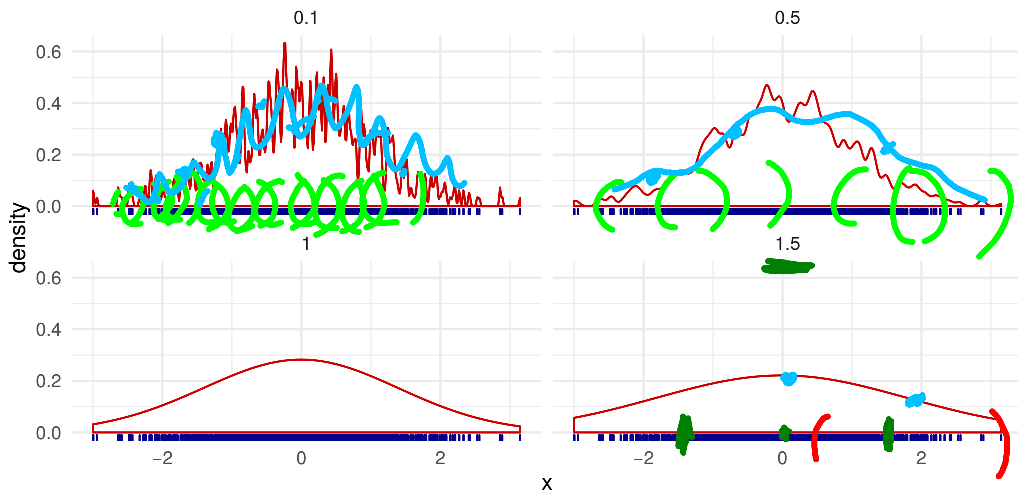
Entonces se tendría que la expresión general para un estimador por núcleos es

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

Pregunta 2.4

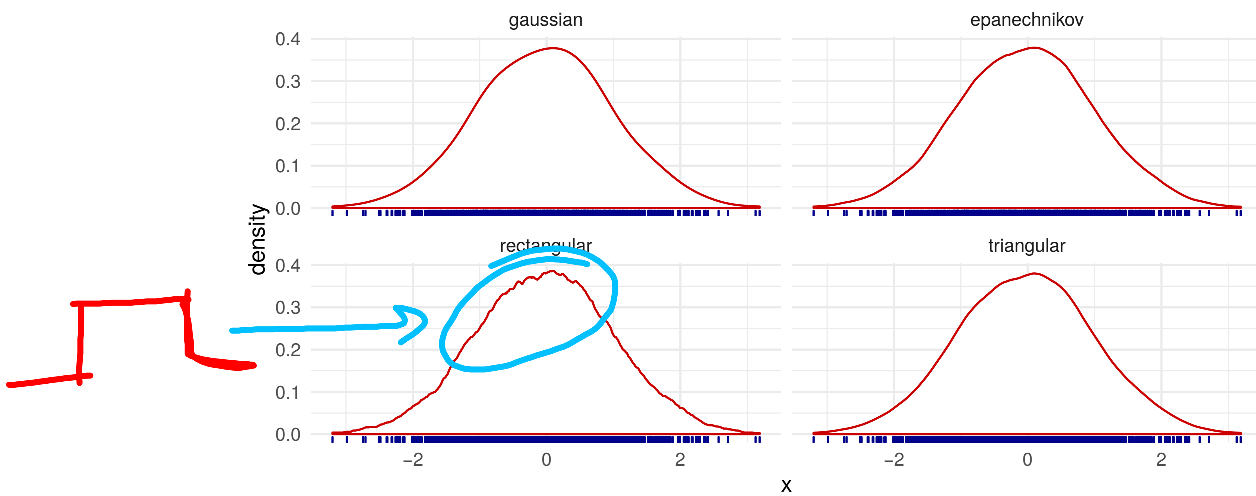
¿Qué pasaría si modificamos el ancho de banda h para un mismo kernel?

Nuevamente sería el ancho de banda ya que



Pregunta 2.5

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



Recordemos nuevamente la fórmula

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

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

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

$$= \frac{1}{n} \sum 1 \qquad u = \frac{x - X_i}{h}$$

$$= 1$$

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

$$\frac{1}{nh} K\left(\frac{x - X_i}{h}\right) \qquad \frac{1}{n} \sum K\left(\frac{x - X_i}{h}\right)$$