Estimació no paramètrica de la densitat

Mètodes no paramètrics i de remostreig
Grau en Estadística UB-UPC

Planteig general i notació

- En tot el que segueix, suposarem que partim d'una **mostra** aleatòria simple x_1, x_2, ..., x_n,
- d'una variable aleatòria X **absolutament contínua** amb distribució F i densitat f.
- F_n designarà la corresponent **distribució empírica** o mostral
- $\hat{f}$ designarà una estimació de la funció de densitat f, f_n indicarà la densitat de F_n (prob. $1/n$ a cada valor observat x_i)

- Procediment de construcció:
  - Decidir "marques de classe":

$$b_0 < b_1 < \ldots < b_m, \text{ reals,}$$

$$b_0 < \min x_i \quad \text{i} \quad b_m \geq \max x_i$$

  - Intervals o classes:

$$B_j = \left(b_{j-1}, b_j\right], \; j = 1, \ldots, m$$

  - Freqüències:

$$n_j = \#\left\{x_i : x_i \in \left(b_{j-1}, b_j\right]\right\}, \quad f_j = n_j / n$$

Histograma: estimació d'*f* més coneguda i utilitzada

Histograma

- Estimació:

$$\hat{f}_H(x) = \begin{cases} f_j \Big/ \left(b_j - b_{j-1}\right) & \text{si} \quad x \in B_j \\ 0 \quad \text{si} \quad x \le b_0 \quad \text{o} \quad x > b_m \end{cases}$$

- Gràficament, representarem un rectangle damunt cada B_j, amb base $b_j - b_{j-1}$ i alçada $f_j / (b_j - b_{j-1})$

- Típicament s'agafa $b = b_j - b_{j-1}$ constant

$$\hat{f}_H(x) = f_j / b \quad \text{si} \quad x \in B_j$$

$$\hat{f}_H(x) = \sum_{j=1}^{m} f_j \frac{1}{b} I_{B_j}(x),$$

$$\text{on } I_{B_j}(x) = \begin{cases} 1 & \text{si } x \in B_j \\ 0 & \text{en cas contrari} \end{cases}$$

- Mixtura de distribucions uniformes sobre B_j, amb pesos f_j

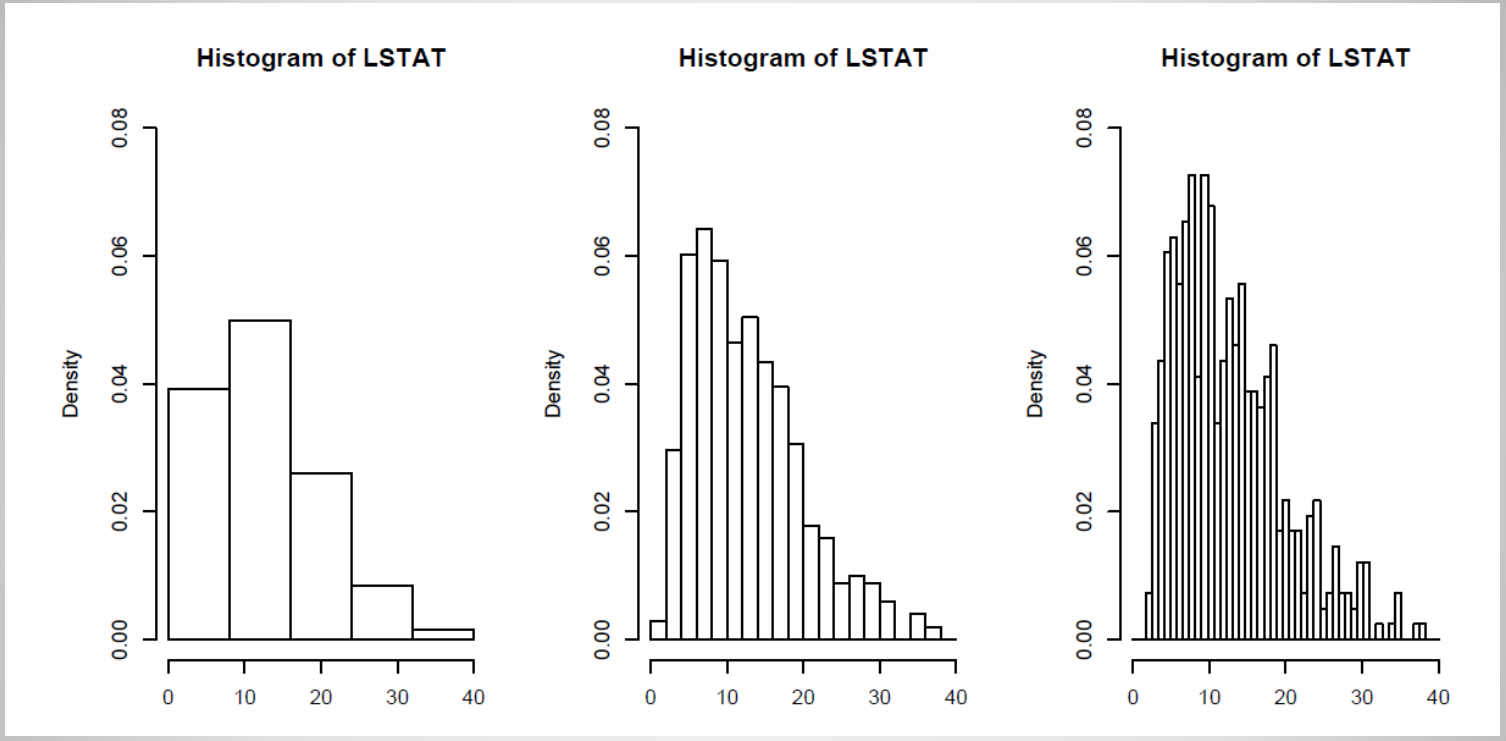
Histograma amb classes de llargada constant

- (Possible interès en generar remostres bootstrap a partir de densitat $\hat{f}_H$)
- Conseqüència directa de definició anterior, per generar $X^* \sim \hat{f}_H$:
  - 1) Triem a l'atzar una classe, amb probabilitats f_1 ,..., f_m . Suposem que hem triat B_j
  - 2) Generem un valor X^* amb distribució uniforme entre b_{j-1} i b_j

Generació aleatòria a partir d'histograma

Pros i contres de l'histograma

- ↑ Simplicitat. Facilitat d'interpretació
- ↓ Com estimació de la densitat, no és una funció contínua
- ↓ Constant a intervals
- ↓ Molt dependent de la tria de classes, en particular de l'**amplada** b i del punt d'**ancoratge**, b_0. (Observem la figura següent, per diferents valors de b...)



Les mateixes dades, diferent *b*

Biaix i variància de l'histograma

- Considerem un punt x entre b_0 i b_m
- La densitat $\hat{f}_H(x)$ es pot considerar un estimador de f(x)

$$\mathrm{E}\left\{\hat{f}_H(x)\right\} = f(x) + O(b)$$

$$\mathrm{var}\left\{\hat{f}_H(x)\right\} = \frac{f(x)}{nb} + O\left(\frac{1}{n}\right)$$

- Com més gran sigui b, més biaix però menys variància (i viceversa)

Consistència de l'histograma

$$\hat{f}_H(x) \xrightarrow{P} f(x)$$

$$n \to \infty \left(\text{i } b \to 0,\ nb \to \infty\right)$$

- (És a dir, consistència si b es fa cada cop més petit, però no massa de pressa)
- Propietats anteriors també tenen com a conseqüència que $\hat{f}_H$ estima millor f al centre de cada interval que a la perifèria

Elecció de *b* òptim

- Anteriors són propietats "locals" (per un *x* donat). Hi ha criteris globals integrant l'error quadràtic mitjà, MSE:

$$\mathrm{MSE}\left(\hat{f}_H(x)\right) = \left(\mathrm{E}\left\{\hat{f}_H(x)\right\} - f(x)\right)^2 + \mathrm{var}\left\{\hat{f}_H(x)\right\}$$

- Millor elecció de *b* segons AMISE (asymptotic mean integrated square error), valor que el minimitza:

$$b_{opt} = \left(\frac{n}{6}\int_{\mathrm{suport}(f)} \left(f'(x)\right)^2 dx\right)^{-1/3}$$

Elecció de *b* òptim

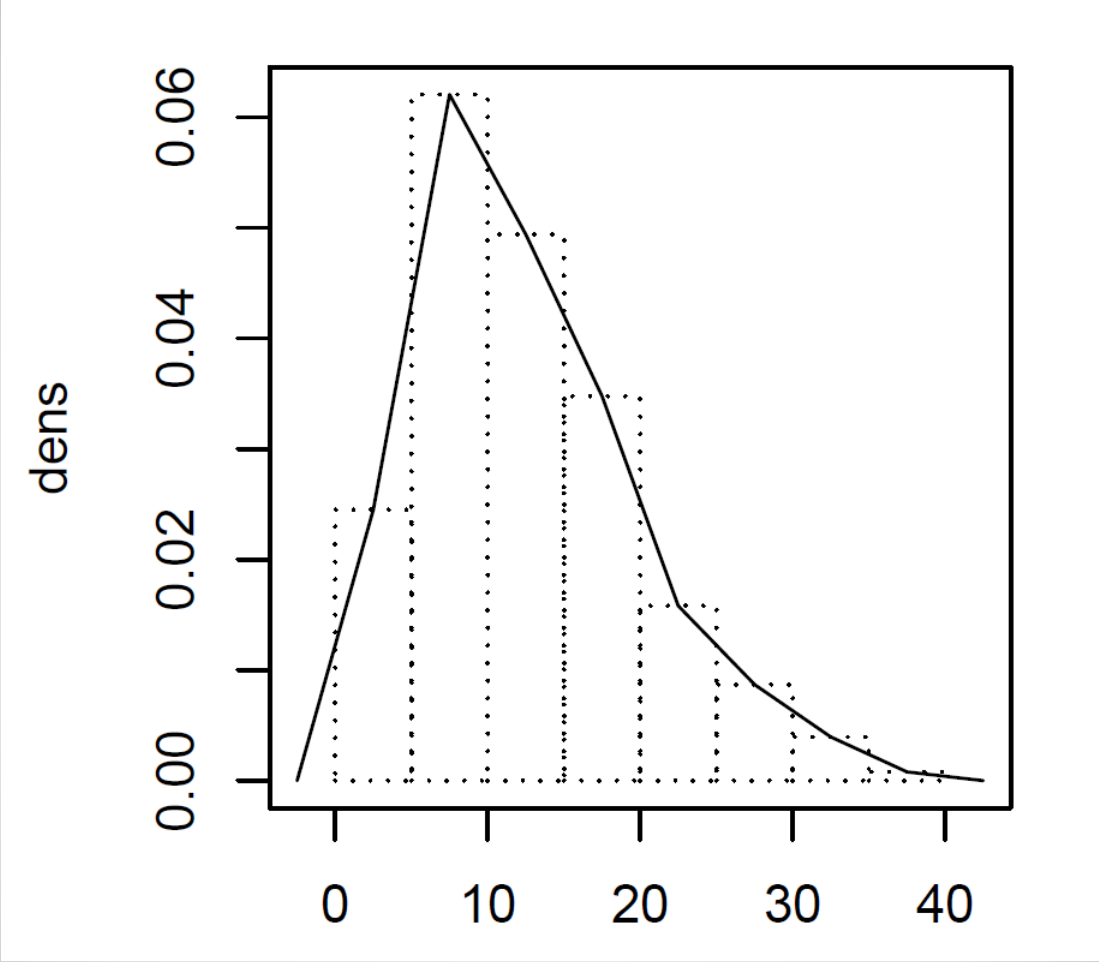
- b_{opt} tal com definit abans no serveix per a res, *f* és desconeguda...
- S'acostuma a calcular la integral agafant com a referència la densitat normal ("normal reference rule"), $N\left(\mu,\sigma^2\right)$
- Que dóna com a resultat:

$$b^*_{opt} = 3.491\,\sigma\,n^{-1/3}$$

- σ estimada com

$$\hat{\sigma} = \min\left\{S, \mathrm{IQR}/1.35\right\}$$

(S^2 variància mostral, IQR rang interquartílic)



Polígon de freqüències

Polígon de freqüències

- Interpolador lineal als punts centrals de cada classe de l'histograma
- Punt central de classe B_j: $c_j = b_{j-1} + b/2 = b_j - b/2$, $j = 1,\ldots, m$.
- (Més $c_0 = b_0 - b/2$, $c_{m+1} = b_m + b/2$)
- Millor expressada a partir de $C_j = (c_{j-1}, c_j]$, $j = 1,\ldots, m+1$

$$\hat{f}_{FP}(x) = \frac{f_{j-1}}{b} + \frac{f_j - f_{j-1}}{b^2}\left(x - c_{j-1}\right) \text{ si } x \in C_j$$

(i 0 en cas contrari)

Polígon de freqüències. Propietats bàsiques

- ↑ Millora la velocitat de convergència a la veritable densitat *f*
- Finestra òptima segons regla de referència a la normal $b^*_{opt} = 2.15\,\sigma\, n^{-1/5}$
- ↑ No presenta discontinuïtats
- ↓ Però no és derivable als punts c_j...
- ↓ Continua depenent del punt d'ancoratge, c_0, i d'amplada de finestra, *b*

- Simulació de $X^* \sim \hat{f}_{FP}$:
- Escollir, amb probabilitats $p_j = (f_{j-1} + f_j)/2$, $j = 2,...,m$, $p_1 = f_1/2$, $p_{m+1} = f_{m+1}/2$, un interval C_j
- Generar un valor T entre 0 i b segons densitat

$$f(t) = f_{j-1} \Big/ \left(bp_j\right) + \left(\left(f_j - f_{j-1}\right) \Big/ \left(b^2 p_j\right)\right) t$$

- $X^* = c_{j-1} + T$

Simulació de la densitat polígon de freqüències

Estimadors nucli de la densitat

- Recordem: histograma és millor estimació d'f al centre de cada interval de classe B_j
- **Histograma mòbil**: Fem que interval de llargada b estigui **centrat a cada punt x** on volem estimar f: $B_x = (x - h, x + h]$, $h = b/2$

$$f_x = \#\left\{x_i \in B_x\right\}/n$$

$$\hat{f}_{HM}(x) = \frac{f_x}{b} = \frac{1}{2}\frac{f_x}{h} = \frac{\#\left\{x_i \in B_x\right\}}{2hn}$$

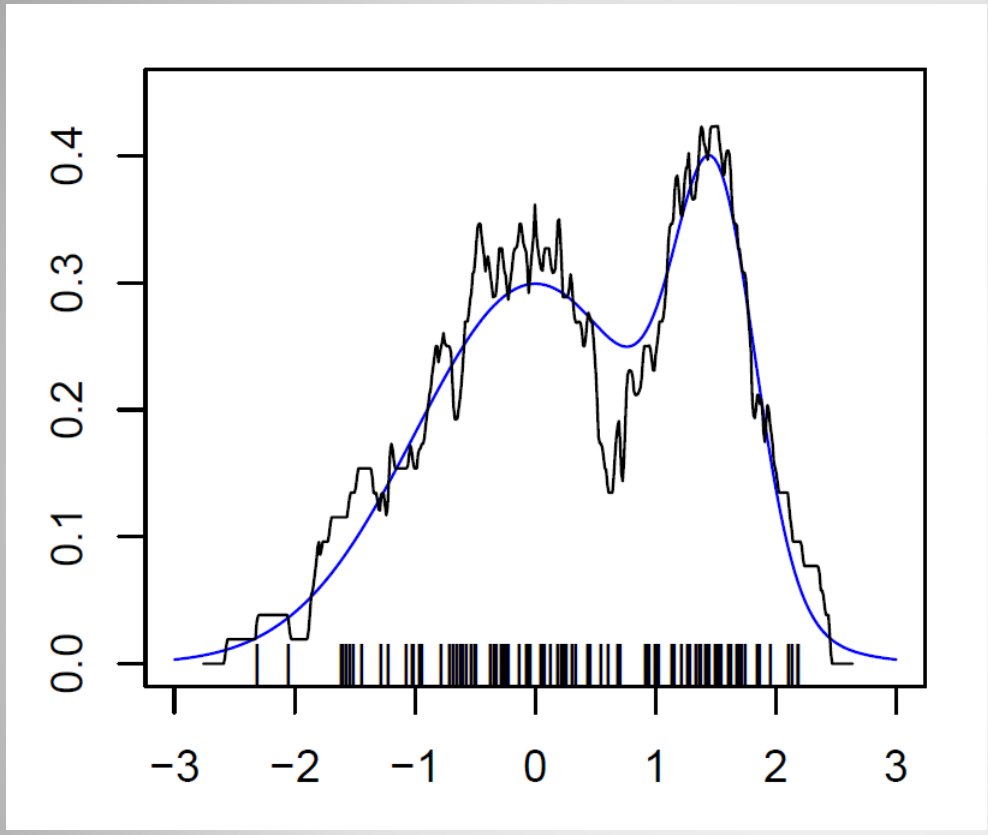
Estimadors nucli de la densitat

- Expressió alternativa:

$$\hat{f}_{HM}(x) = \frac{1}{nh}\sum_{i=1}^{n} \tfrac{1}{2} I_{B_x}(x_i)$$

$$= \frac{1}{nh}\sum_{i=1}^{n} \tfrac{1}{2} I_{(-1,+1]}\left(\frac{x - x_i}{h}\right)$$

- Variant *x* tindríem una estimació de la densitat. Si la representéssim gràficament, veuríem que no és suau



$$g(u) = \tfrac{1}{2} I_{(-1,+1]}(u)$$

densitat uniforme sobre $(-1,+1]$

Estimador nucli (uniforme)

Estimadors nucli

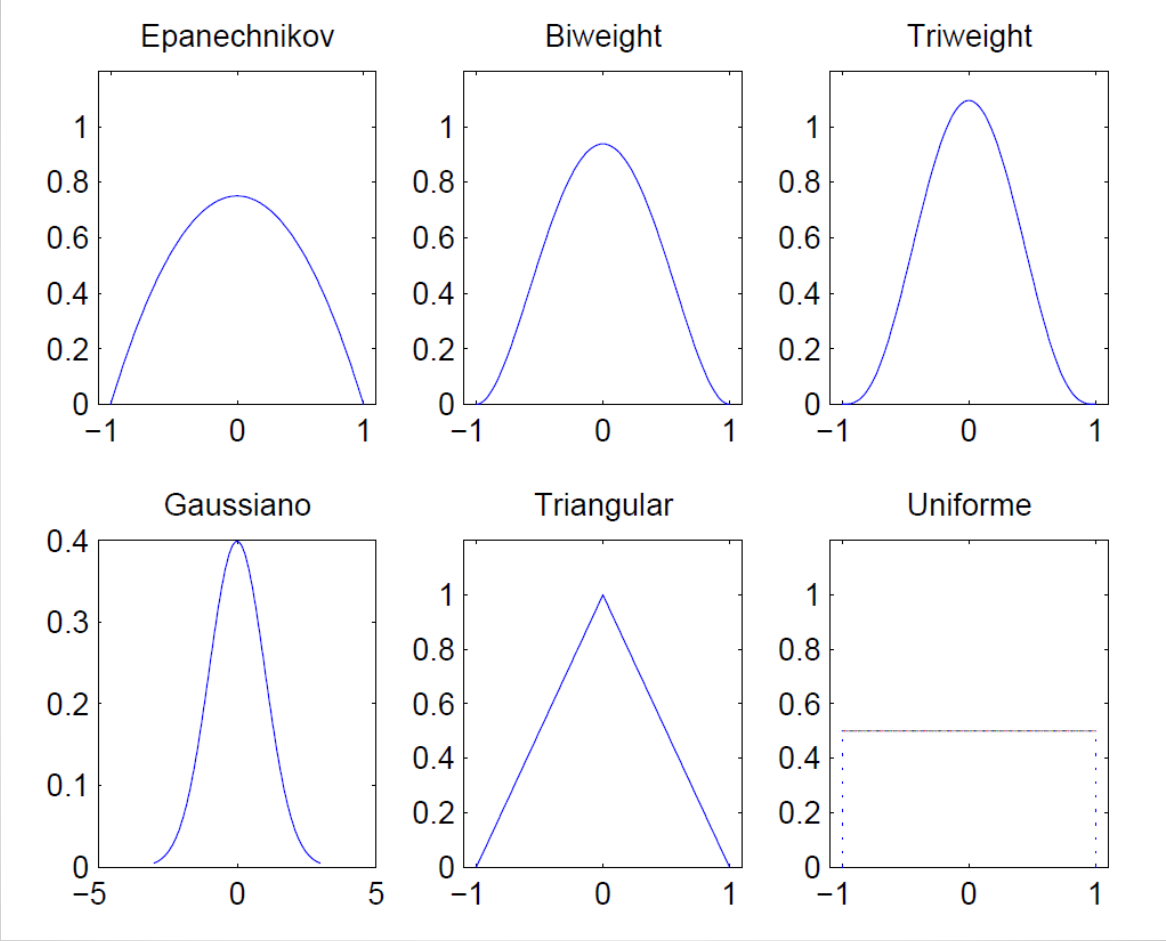
- Si al sumatori substituïm la densitat uniforme per una de més suau (un "nucli" K) i que sigui positiva en un interval més ampli...
- ...(per exemple, una normal)...
- ...tindrem un estimador nucli (o "kernel") pel nucli K amb paràmetre de suavitzat h:

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

Nuclis utilitzats habitualment

- Normalment s'utilitzen nuclis *K* basats en distribucions contínues, unimodals i simètriques

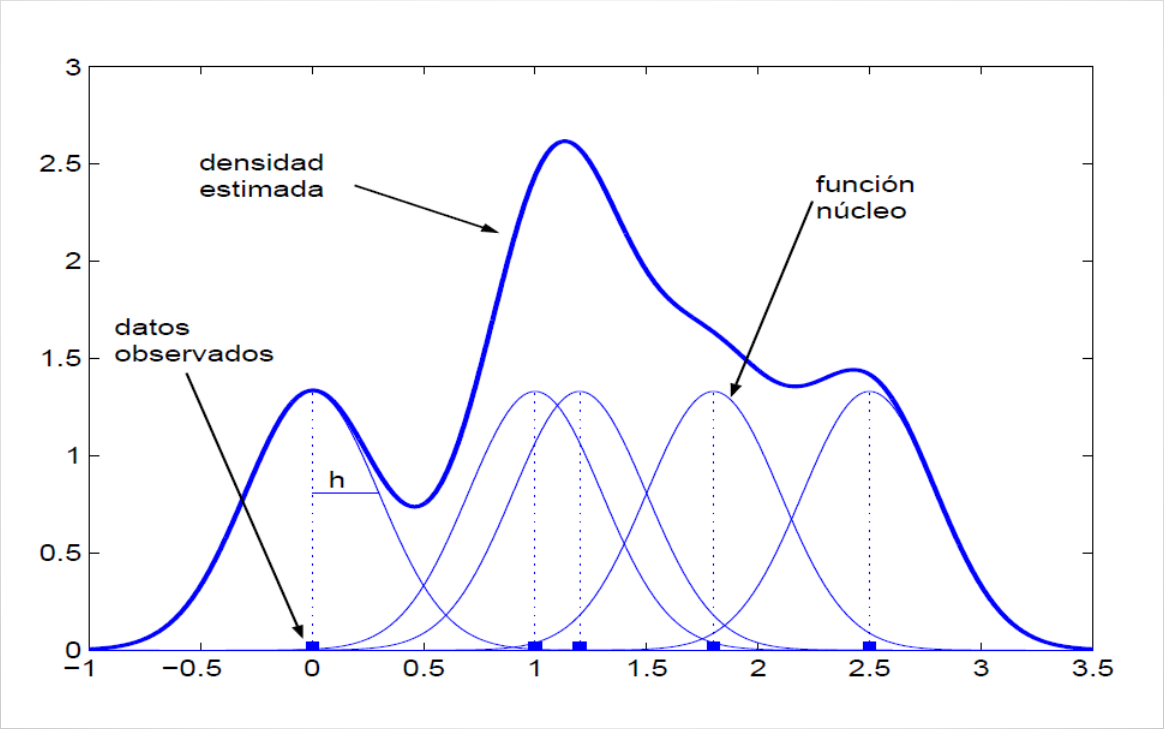
Kernel	K(u)
Uniform	$\frac{1}{2}I(\mid u \mid \leq 1)$
Triangle	$(1- \mid u \mid)\ \ I(\mid u \mid \leq 1)$
Epanechnikov	$\frac{3}{4}(1-u^2)I(\mid u \mid \leq 1)$
Quartic	$\frac{15}{16}(1-u^2)^2 I(\mid u \mid \leq 1)$
Triweight	$\frac{35}{32}(1-u^2)^3 I(\mid u \mid \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}}exp(-\frac{1}{2}u^2)$
Cosinus	$\frac{\pi}{4}cos(\frac{\pi}{2}u)I(\mid u \mid \leq 1)$



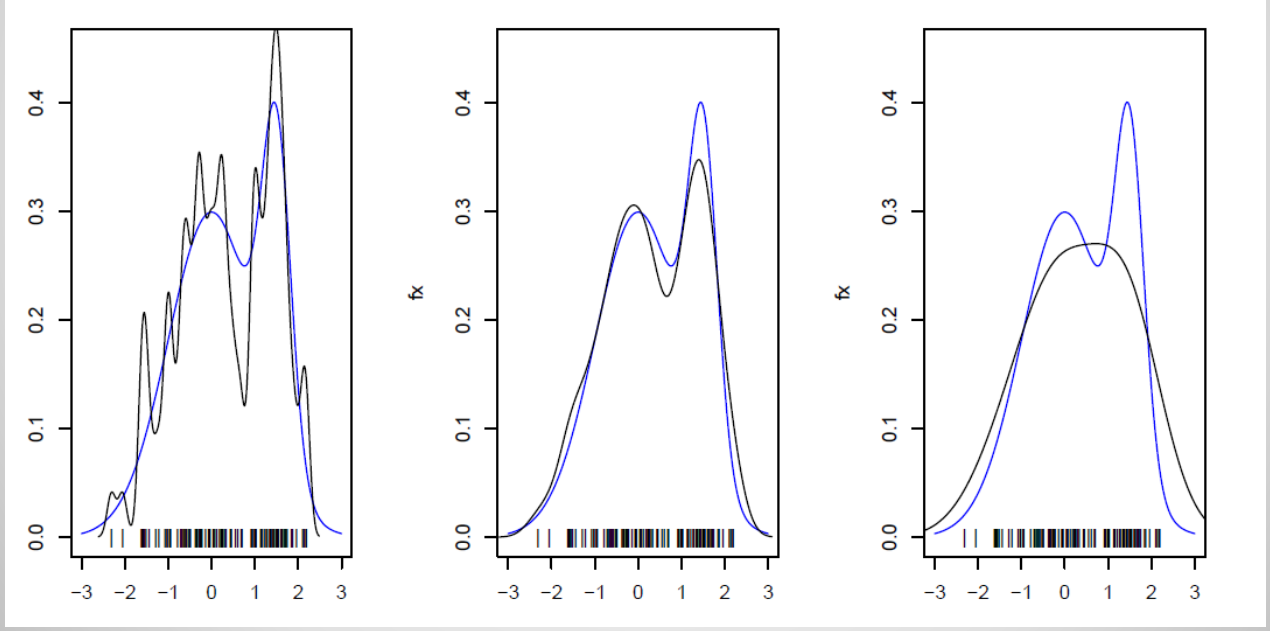
Nuclis utilitzats habitualment

Interpretació de l'estimació nucli

- Mixtura d'n densitats (amb pesos $1/n$) amb la mateixa forma que K però rescalades segons h i centrades a x_i



- *h* controla concentració del pes 1/*n* al voltant d'x_i: com més gran, més observacions llunyanes influiran en l'estimació de la densitat



Interpretació de l'estimació nucli

Principals propietats de l'estimació nucli de la densitat

- Biaix: augmenta en augmentar *h*
- Variància: disminueix en augmentar *h*
- Consistència:

$$\hat{f}_K(x) \xrightarrow{P} f(x)$$

- Finestra òptima: depèn de *K*, conegut, i de *f*, desconegut. Referència a normal:

K Gausià: $h^*_{opt} = 1.059\,\sigma\, n^{-\frac{1}{5}}$

En general: $h_{opt,K} = \left(\frac{R(K)}{\sigma_K^4 R(f'')} \right)^{-\frac{1}{5}} n^{-\frac{1}{5}}$

- Interpretació alternativa: Composició de n distribucions amb densitat $K(\varepsilon/h)/h$ i pesos $1/n$

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

- Experiment associat: primer triem una x_i amb probabilitat $1/n$ (distribució empírica); després obtenim un valor aleatori segons $K(\varepsilon/h)/h$ (dibuixeu l'arbre de probabilitats)

Relació amb la distribució empírica

- Generació de valors bootstrap segons densitat nucli:

1) Variable X' amb **distribució empírica** F_n

2) Variable aleatòria ε (pertorbació aleatòria) amb densitat $K(\varepsilon/h)/h$

- Simulació de $\hat{f}_K$: $X^* = X' + \varepsilon$
- Sovint designat com a ***bootstrap semiparamètric***

Conseqüència: aplicació a *bootstrap semiparamètric*