Estimación no parametrica de la densidad

Daniela Rodriguez

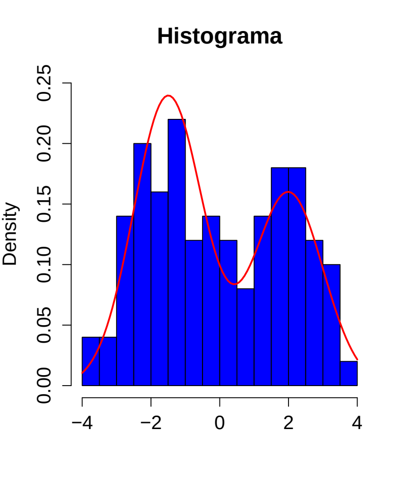
Estimador no paramétrico de la densidad

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

Estimador no paramétrico de la densidad

- X con densidad $f(x)$: queremos estimar $f(x)$
- $X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ 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

Ejemplo datos simulados



Estimación de la densidad: Histograma

Sea $X_1, \dots, X_n$ una muestra aleatoria con densidad desconocida f. El primer estimador de la densidad no parametrico que se estudio es el histograma.

Construcción:

- Dividir el rango en intervalos o bins con origen x_0 y ancho h

$$B_j = [x_0 + (j-1)h, x_0 + jh)$$

- Contar la cantidad de observaciones n_j en B_j
- Normalizar

$$f_j = \frac{n_j}{nh}$$

- Graficar barras de altura f_j y ancho h sobre cada B_j

Formalmente el histograma puede escribirse como

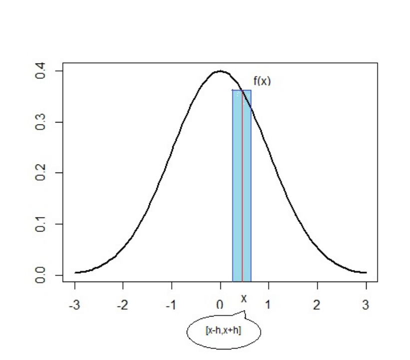
$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j} I(\mathbf{x}_i \in B_j) I(\mathbf{x} \in B_j)$$

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
- aproxima a la densidad en el punto medio de cada bin

ESTIMADOR DE DENSIDAD POR NUCLEOS

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



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

Juntando todo...

- Si h es pequeño y f continua en x,

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

- Por la LGN $X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

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

- Entonces,

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

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

Propuesta

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

Notemos que

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

- 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

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

Notemos que

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

- Estimador de Parzen

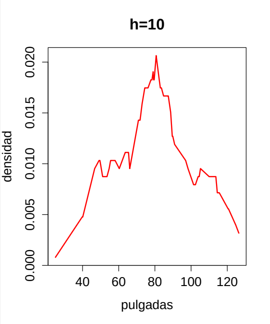
$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2}\mathcal{I}_{[-1,1]}\left(\frac{x-X_i}{h}\right)$$

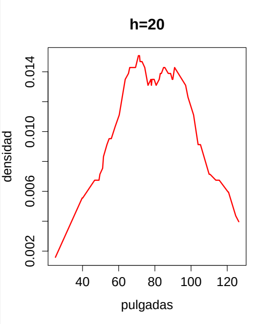
- si $K(t) = \frac{1}{2}\mathcal{I}_{[-1,1]}(t) \quad \Rightarrow$

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

Propuesta

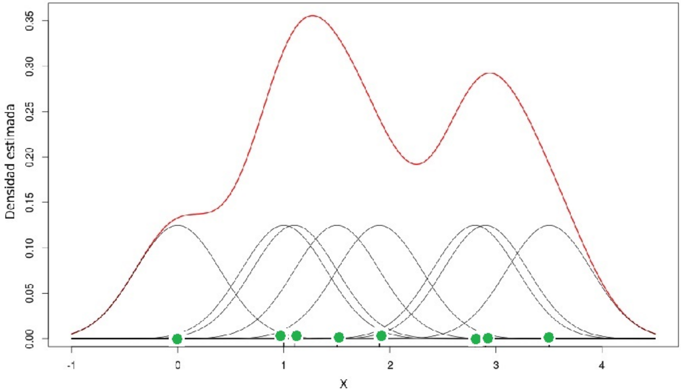
- $K(t) = \frac{1}{2}\mathcal{I}_{[-1,1]}(t) \quad \Rightarrow \quad \widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$
  - K : núcleo
  - h : ventana





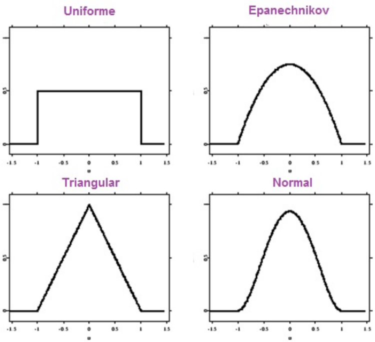
Núcleos

Interpretación del estimador de núcleos



Tipos de núcleos

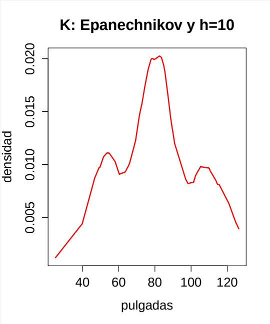
- Núcleo Rectangular: $K(t) = \frac{1}{2}\mathcal{I}_{[-1,1]}(t)$
- Núcleo Triangular: $K(t) = (1 - |t|)\mathcal{I}_{[-1,1]}(t)$
- Núcleo Gausssiano: $K(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}$
- Núcleo Epanechnikov: $K(t) = \frac{3}{4}(1 - t^2)\mathcal{I}_{[-1,1]}(t)$

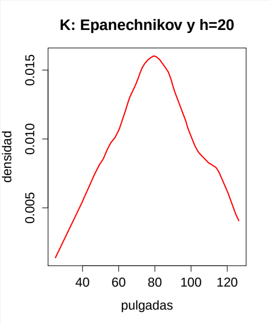


Estimadores de núcleos (Rosenblatt-Parzen)

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

- 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





PROPIEDADES

Notemos que si $\int_{+\infty}^{-\infty} K(x)dx = 1$ y $K \geq 0$, entonces el estimador $\widehat{f}$ es también una función de densidad.
Pues,

$$\int_{+\infty}^{-\infty} \widehat{f}(x)dx = \int_{+\infty}^{-\infty} \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{x - \mathbf{x}_i}{h}\right) dx =$$

$$\frac{1}{n} \sum_{i=1}^{n} \int_{+\infty}^{-\infty} \frac{1}{h} K\left(\frac{x - \mathbf{x}_i}{h}\right) dx = \frac{1}{n} \sum_{i=1}^{n} \int_{+\infty}^{-\infty} K(s)ds = 1$$

Propiedades: Sesgo

i) f es 2-veces derivable tal que $\int f^{''}(s)ds < \infty$.
ii) $\int K = 1$, $\int K(s)sds = 0$ y $\int K(s)s^2ds < \infty$.

Tenemos que

$$E\left[\widehat{f}(x)\right] = f(x) + \frac{h^2}{2}f^{''}(x)\mu_2(K) + o(h^2)$$

si $h \to 0$ para cada x. Donde $\mu_2(K) = \int s^2K(s)ds$

Propiedades: Sesgo

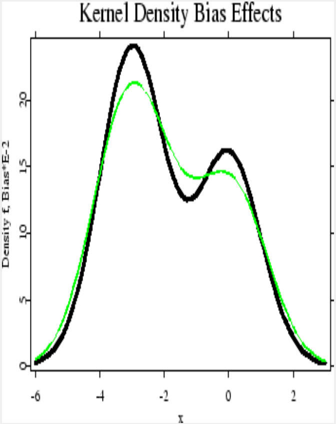
$$
\begin{aligned}
E\left[\widehat{f}(x)\right] &= E\Big(\frac{1}{hn}\sum_{i=1}^{n} K\Big(\frac{x-\mathbf{x}_i}{h}\Big)\Big) = \frac{1}{n}\sum_{i=1}^{n} E\Big(\frac{1}{h}K\Big(\frac{x-\mathbf{x}_i}{h}\Big)\Big) \\
&= E\Big(\frac{1}{h}K\Big(\frac{x-\mathbf{x}_i}{h}\Big)\Big) = \frac{1}{h}\int_{+\infty}^{-\infty} K\Big(\frac{x-u}{h}\Big)f(u)du \\
&= \int_{+\infty}^{-\infty} K(y)f(x-hy)dy \\
&= \int_{+\infty}^{-\infty} K(y)[f(x)+f'(x)hy+\frac{f''(x)}{2}y^2h^2+o(h^2)]dy \\
&= f(x)\int_{+\infty}^{-\infty} K(y) + f'(x)h\int_{+\infty}^{-\infty} K(y)ydy + h^2\frac{f''(x)}{2}\int_{+\infty}^{-\infty} K(y)y^2dy \\
&+ o(h^2)
\end{aligned}
$$

Propiedades: Sesgo

$$Sesgo\left(\widehat{f}(x)\right) = h^2\frac{f''(x)}{2}\int_{+\infty}^{-\infty} K(y)y^2dy + o(h^2)$$

Propiedades: Sesgo

Estimador de densidad (en verde) y verdadera densidad (en negro).



Propiedades: Varianza

Bajo las mismas hipótesis introducidas anteriormente, probaremos que la varianza del estimador es

$$Var(\widehat{f}(x)) = \frac{1}{nh}\|K\|^2 f(x) + o(\frac{1}{nh}).$$

Propiedades: Varianza

Sea $K_h(x) = \frac{1}{h}K(x/h)$, como las $\mathbf{x}_i$ son i.i.d

$$
\begin{aligned}
Var(\widehat{f}(x)) &= n^{-2}Var(\sum_{i=1}^{n} K_h(x - \mathbf{x}_i)) = n^{-2}\sum_{i=1}^{n} Var(K_h(x - \mathbf{x}_i)) \\
&= n^{-1}Var(K_h(x - \mathbf{x}_1)) \\
&= n^{-1}[E(K_h^2(x - \mathbf{x}_1)) - E^2(K_h(x - \mathbf{x}_1))] \\
&= n^{-1}[E(K_h^2(x - \mathbf{x}_1)) - (f(x) + o(h))^2]
\end{aligned}
$$

Usando los mismos argumentos que antes, es decir, cambio de variable y un desarrollo de Taylos tenemos que

$$E(K_h^2(x - \mathbf{x}_1)) = h^{-1}\int K^2(s)f(x - hs)ds = h^{-1}\|K\|_2^2 f(x) + o(h).$$

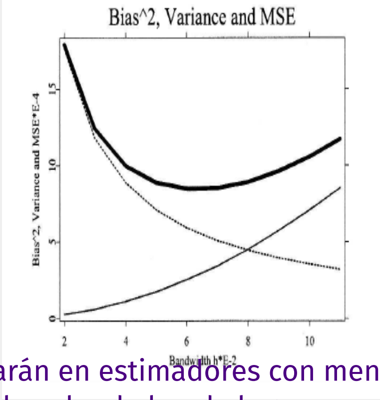
Propiedades: ECM

De esta forma hemos calculado el error cuadrático medio del estimador (*ECM*) para cada x,

$$ECM(\widehat{f}(x)) = h^4 \frac{(f''(x))^2}{4} \mu_2^2(K) + o(h^4) + \frac{1}{nh} \|K\|^2 f(x) + o(\frac{1}{nh})$$

Propiedades: ECM

Compromiso entre sesgo y varianza. En la figura: Sesgo al cuadrado (linea sólida); varianza (linea punteada) y error cuadrático medio (linea solida gruesa).



h pequeños derivarán en estimadores con menor sesgo mientras que al aumentar el ancho de banda lograremos disminuir la varianza.

Sesgo y Varianza de $\widehat{f}(x)$

- El sesgo es proporcional a $h^2 \Rightarrow$ elijamos h pequeña
- El sesgo depende de $f''(x)$ que mide la curvatura de f en x
- La varianza dismimuye a medida que nh crece
- Para disminuir la varianza necesitamos h o n grandes.

La elección de la ventana es crucial.

- Una ventana h pequeña dará un estimador muy rugoso, con muchos picos y difícil de interpretar
- una ventana h grande sobre-suaviza al estimador de la densidad y enmascara estructuras de los datos.

Propiedades: Consistencia

$$ECM(\widehat{f}(x)) = h^4 \frac{(f''(x))^2}{4} \mu_2^2(K) + o(h^4) + \frac{1}{nh} \|K\|^2 f(x) + o(\frac{1}{nh})$$

Si $h \to 0$ y $nh \to \infty$ tenemos que

$$\widehat{f}(x) \xrightarrow{p} f(x)$$

para cada x.

Intervalos y Bandas de confianza

Bajo ciertas condiciones de regularidad

- $h_n \to 0$
- $nh_n \to \infty$
- x tiene densidad f continua en x y dos veces diferenciable
- $K : \mathbb{R} \to \mathbb{R}$ es acotado, $\int K = 1$ y $\int u^2 K(u) > 0$ y con soporte compacto.

si $h_n = cn^{-1/5}$ entonces

$$\sqrt{nh}(\widehat{f}(x) - f(x)) \xrightarrow{\mathcal{D}} N(\frac{c^{5/2}}{2} f''(x)\mu_2(K), f(x)\|K\|^2).$$

Intervalos y Bandas de confianza

Luego resulta el siguiente intervalo de confianza de nivel aproximado $1 - \alpha$

$$\left[\widehat{f}(x) - \frac{h^2}{2} f''(x)\mu_2(K) - z_{\alpha/2}\sqrt{\frac{f(x)\|K\|^2}{nh}}, \widehat{f}(x) - \frac{h^2}{2} f''(x)\mu_2(K) + z_{\alpha/2}\sqrt{\frac{f(x)\|K\|^2}{nh}}\right]$$

si h es pequeña se puede despreciar el término que involucra a la derivada segunda y utilizar el siguiente intervalo

$$\left[\widehat{f}(x) - z_{\alpha/2}\sqrt{\frac{\widehat{f}(x)\|K\|^2}{nh}}, \widehat{f}(x) + z_{\alpha/2}\sqrt{\frac{\widehat{f}(x)\|K\|^2}{nh}}\right]$$

de lo contrario podemos estimar la derivada segunda, derivando un es de núcleos usando una ventana g.

Es importante notar que este intervalo es sólo para $f(x)$ y no para toda la densidad.

Para deducir bandas de confianza para toda la función es necesario emplear otras técnicas.

Bickel y Rosenblatt (1973) probaron el siguiente resultado: sea f una función de densidad definida sobre el $(0,1)$, $h_n = n^{-\delta} \in (1/5, 1/2)$, entonces para todo $x \in (0,1)$,

$$\lim_{n\to\infty} P\left(\widehat{f}(x) - \sqrt{\frac{\frac{\widehat{f}(x)\|K\|^2}{nh}}{\frac{z}{2\delta\log n} + d_n}} \le f(x) \le \widehat{f}(x) + \sqrt{\frac{\frac{\widehat{f}(x)\|K\|^2}{nh}}{\frac{z}{2\delta\log n} + d_n}}\right) = e^{-2e^{-z}}$$

donde $d_n = (2\delta\log n)^{1/2} + (2\delta\log n)^{-1/2}\log\left\{\frac{\|K'\|_2}{2\pi\|K\|_2}\right\}$.

Luego si $z \approx 3.663$, se tiene que $e^{-2e^{-z}} = 1 - 0.05$.