

Tema 3. Estimación de la función de densidad

Parte 1

Rosa M. Crujeiras
Alberto Rodríguez



Dpto. de Estadística e Investigación Operativa
Máster en Técnicas Estadísticas
Curso 2009-2010

Sea X una v.a. con distribución F continua, y función de densidad f (ambas desconocidas):

$$F'(x) = f(x), \quad F(x) = \int_{-\infty}^x f(u)du$$

y supongamos que tenemos X_1, \dots, X_n m.a.s. de X .

¿Cómo podemos estimar f a partir de la muestra?

Histograma:

Un histograma es una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados. (Wikipedia)

Histograma:

Dado un origen x_0 y un ancho $h > 0$, el histograma es una densidad constante en cada intervalo de la forma:

$$\{[x_0 + hm, x_0 + h(m + 1)), m \in \mathbb{Z}\}$$

¿Cuánto vale el estimador en cada uno de los intervalos?

Si partimos de una m.a.s. X_1, \dots, X_n , el estimador natural de la probabilidad en cada uno de los intervalos $[x_m, x_{m+1})$, con $x_m = x_0 + hm$ es:

$$\frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i \in [x_m, x_{m+1}))$$

Denotemos por $\hat{f}_{n,H}$ el histograma. Como es una densidad, la probabilidad del intervalo $[x_m, x_{m+1})$ se puede calcular como:

$$\int_{x_m}^{x_{m+1}} \hat{f}_{n,H}(u) du = f_m \cdot h$$

donde f_m es el valor de $\hat{f}_{n,H}$ en el intervalo $[x_m, x_{m+1})$ y h es el ancho del intervalo.

Igualando las dos expresiones se obtiene que:

$$f_m = \frac{1}{nh} \sum_{i=1}^n \mathbb{I}(X_i \in [x_m, x_{m+1}))$$

Los datos que vamos a presentar aquí fueron analizados en Azzalini y Bowman (1990) quienes registraron el tiempo (min) que dura una erupción del géiser “Old Faithful” que se encuentra en el parque nacional de Yellowstone (Wyoming, EEUU). Las medidas (299 erupciones en total) fueron tomadas entre el 1 y el 15 de Agosto de 1985.



Figure: El geysir 'Old Faithful' en plena erupción

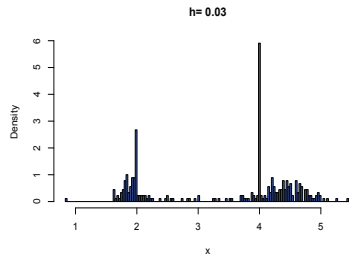
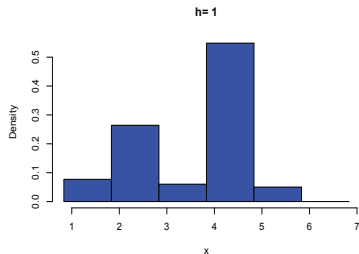
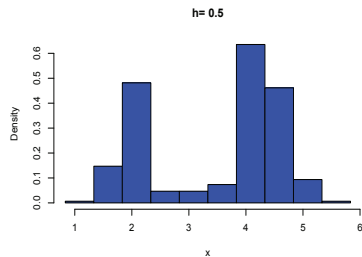
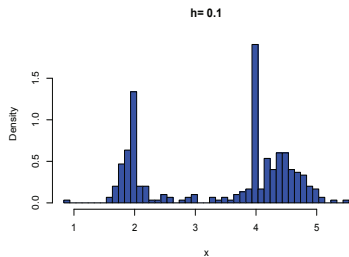


Figure: Histograma para 4 valores diferentes de h .

El histograma depende en exceso del punto x_0 :

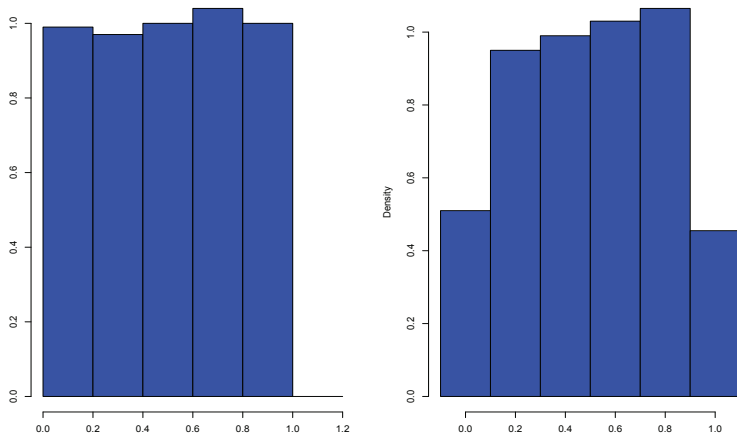


Figure: Histograma de una muestra de una población $U(0,1)$ para dos valores diferentes de x_0 .

Ejercicio (Histograma)

Sea X_1, \dots, X_n una m.a.s. de $X \sim U(0, 1)$. Consideremos un ancho $h = 0'5$ fijo y dos posibles valores para el origen x_0 :

a) $x_0 = 0$

b) $x_0 = -0'25$

Calcula el sesgo de $\hat{f}_{n,H}(0)$ como estimador de $f(0) = 1$ en las dos situaciones

El histograma se puede modificar para evitar la influencia del origen x_0 , de manera que para cada punto x se construye un intervalo de la forma $(x - h, x + h)$ (histograma móvil).

Recordemos la definición de función de densidad de una v.a. X en un punto x :

$$f(x) = \lim_{h \rightarrow 0} \frac{1}{2h} \mathbb{P}(x - h < X < x + h)$$

Un estimador natural, denominado *Estimador Naíve*, viene dado por:

$$\hat{f}_{n,N}(x) = \frac{1}{2nh} \sum_{i=1}^n \mathbb{I}(X_i \in (x - h, x + h)).$$

Esta formulación es equivalente a:

$$\hat{f}_{n,N}(x) = \frac{1}{2nh} \sum_{i=1}^n \mathbb{I}(x \in [X_i - h, X_i + h))$$

aunque esta expresión es menos práctica para hacer cálculos.

Ejercicio (Estimador Naive) Supongamos que f es continua en x .

- Prueba que si $h \rightarrow 0$ entonces $\mathbb{E}(\hat{f}_{n,N}(x)) \rightarrow f(x)$.
- Prueba que si $nh \rightarrow \infty$ entonces $\text{Var}(\hat{f}_{n,N}(x)) \rightarrow 0$

Ejercicio (Estimador Naive) Sea f la densidad de la exponencial de parámetro uno

$$f(x) = \begin{cases} e^{-x} & \text{si } x \geq 0 \\ 0 & \text{en otro caso.} \end{cases}$$

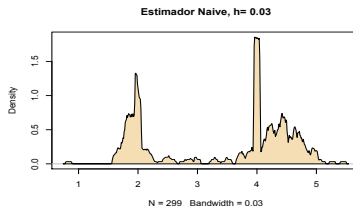
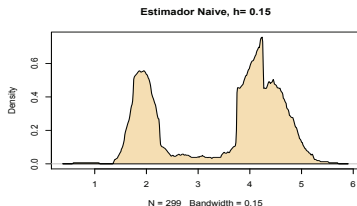
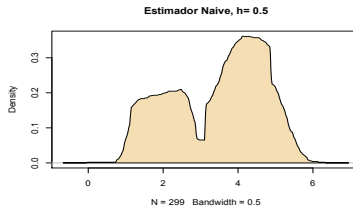
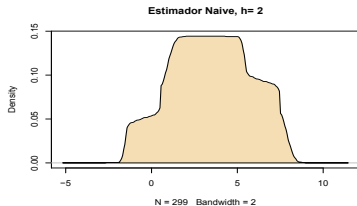
Prueba que: $\lim_{h \rightarrow 0} \mathbb{E}(\hat{f}_{n,N}(0)) = \frac{1}{2}$.

Ejercicio (Estimador Naive) Probar que si F es continua en un entorno de x entonces para h suficientemente pequeño

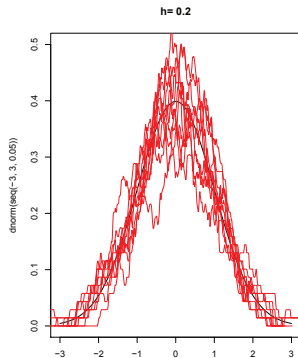
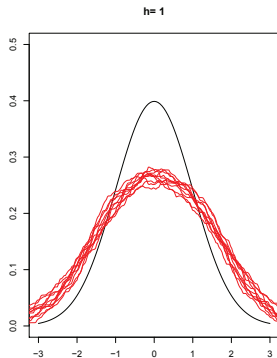
$$\mathbb{E} \left(\hat{f}_{n,N}(x) \right) = \frac{F(x+h) - F(x-h)}{2h}$$

$$\text{Var} \left(\hat{f}_{n,N}(x) \right) = \frac{F(x+h) - F(x-h)}{4nh^2} - \frac{(F(x+h) - F(x-h))^2}{4nh^2}$$

Estimador Naíve para cuatro valores diferentes de h para los datos del géiser “Old Faithful” presentados anteriormente. Nótese que el estimador es discontinuo en $X_i \pm h$.



El parámetro h juega un papel clave en el comportamiento del Estimador Naíve. A modo de ejemplo mostramos, para dos valores de h diferentes ($h = 1$ y $h = 0.2$), el Estimador Naíve construido a partir de 10 muestras diferentes de tamaño 200 de la normal estándar.



Para un punto x fijo, $\hat{f}_{n,N}(x)$ es una variable aleatoria. Para medir su calidad como estimador de $f(x)$ podemos utilizar el Error Cuadrático Medio (MSE):

$$MSE(x) = \mathbb{E}(\hat{f}_{n,N}(x) - f(x))^2,$$

que, como sabemos, se puede descomponer en sesgo al cuadrado más varianza

$$MSE(x) = (\mathbb{E}(\hat{f}_{n,N}(x)) - f(x))^2 + \text{Var}(\hat{f}_{n,N}(x)).$$

Supongamos que

- Existe la **derivada segunda** de f
- f'' es **continua**

Por el Teorema de Taylor, para $h > 0$ existe ξ_h en el intervalo $(x, x + h)$ y γ_h en $(x - h, x)$ verificando que

$$\begin{aligned} F(x + h) &= F(x) + hf(x) + \frac{h^2}{2}f'(x) + \frac{h^3}{3!}f''(\xi_h) \\ F(x - h) &= F(x) - hf(x) + \frac{h^2}{2}f'(x) - \frac{h^3}{3!}f''(\gamma_h), \end{aligned}$$

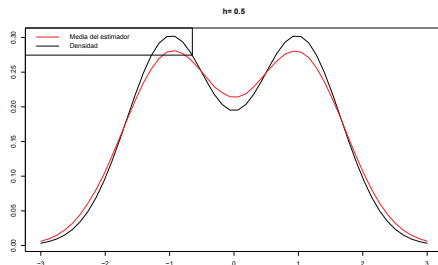
Así

$$\mathbb{E} \left(\hat{f}_{n,N}(x) \right) = \frac{F(x + h) - F(x - h)}{2h} = f(x) + \frac{h^2}{3!} \left(\frac{f''(\xi_h) + f''(\gamma_h)}{2} \right).$$

Si $h \rightarrow 0$ tendremos que

$$\mathbb{E} \left(\hat{f}_{n,N}(x) \right) = f(x) + \frac{h^2 f''(x)}{6} + o(h^2).$$

- En los mínimos de f el Estimador Naive tenderá a sobreestimar f porque $f'' > 0$.
- En los máximos de f el Estimador Naive tenderá a infraestimar f porque $f'' < 0$.



Ejercicio (Estimador Naive)

Demostrar que si $nh \rightarrow \infty$ entonces

$$\text{Var} \left(\hat{f}_{n,N}(x) \right) = \frac{f(x)}{2nh} + o((nh)^{-1})$$

Ejercicio (Estimador Naive)

Demostrar que si $h \rightarrow 0$ y $nh \rightarrow \infty$ entonces

$$MSE(x) = \frac{f(x)}{2nh} + \frac{h^4 (f''(x))^2}{36} + o(h^4 + (nh)^{-1})$$

Se define el Error Cuadrático Medio Asintótico en el punto x como

$$AMSE(x) = \frac{f(x)}{2nh} + \frac{h^4 (f''(x))^2}{36}$$

Ejercicio (Estimador Naive)

Demuestra que si $f''(x) \neq 0$ entonces el valor de h que minimiza el AMSE viene dado por la expresión

$$h_{AMSE}(x) = \left(\frac{9f(x)}{2n(f''(x))^2} \right)^{\frac{1}{5}}$$

Ejercicio (Estimador Naive)

Demuestra que si $f''(x) \neq 0$ entonces

$$\inf_{h>0} AMSE(x) = c(f(x))^{4/5} (f''(x))^{2/5} n^{-4/5},$$

donde c es una constante que no depende de x ni de n .

Un criterio de error global frecuentemente utilizado es el Error Cuadrático Medio Integrado

$$MISE(h) = \mathbb{E} \int (\hat{f}_{n,N}(x) - f(x))^2 dx,$$

que, intercambiando la esperanza con la integral, no es más que un promedio de los errores cuadráticos medios en cada punto

$$MISE(h) = \int MSE(x) dx.$$

El MISE es el criterio de error más utilizado. Sin embargo no es el único criterio de error empleado. Se puede emplear la distancia L_1 para medir la distancia entre $\hat{f}_{n,N}$ y f . Promediando esta distancia L_1 se obtiene el Error Absoluto Integrado Medio (MIAE)

$$MIAE(h) = \mathbb{E} \int |\hat{f}_{n,N}(x) - f(x)| dx.$$

Ejercicio (Estimador Naíve)

Sean X e Y dos variables aleatorias con funciones de densidad f y g . Para $a > 0$ sean f_a, g_a las densidades de aX y aY respectivamente. Probar que

$$\int |f_a(x) - g_a(x)| = \int |f(x) - g(x)| dx$$

¿Verifica esta propiedad la distancia L_2 entre densidades?

Si además de suponer que f'' existe y es continua suponemos que

$$R(f'') = \int (f''(x))^2 dx < \infty$$

entonces, integrando el $MSE(x)$, obtenemos la expresión asintótica del MISE del Estimador Naïve

$$MISE(h) = \frac{1}{2nh} + \frac{h^4}{36} R(f'') + o(h^4 + (nh)^{-1})$$

Se define el MISE asintótico como

$$AMISE(h) = \frac{1}{2nh} + \frac{h^4}{36} R(f'').$$

Ejercicio (Estimador Naive)

Prueba que el parámetro que minimiza el AMISE es

$$h_{AMISE} = \left(\frac{9}{2nR(f'')} \right)^{1/5},$$

y que

$$\inf_{h>0} AMISE(h) = \frac{5}{4} \left[\frac{R(f'')}{144} \right]^{1/5} n^{-4/5}$$

Ejercicio (Estimador Naive)

Sea X una variable con densidad f . Si $h_{AMISE,a,c}$, denota la ventana AMISE de $f_{a,c}$ donde $f_{a,c}$ es la densidad de $aX + c$ prueba

$$h_{AMISE,a,c} = ah_{AMISE}, \quad a > 0, \quad c \in \mathbb{R}$$

donde h_{AMISE} es la ventana AMISE de f .

Ejercicio (Estimador Naíve)

Prueba que si f es la densidad de la normal estándar entonces

$$R(f'') = \frac{3}{8\sqrt{\pi}}$$

Ejercicio (Estimador Naíve)

Prueba que si f es la densidad de la normal estándar entonces

$$h_{AMISE} = \left(\frac{12\sqrt{\pi}}{n} \right)^{1/5}$$

Ejercicio (Estimador Naíve)

Prueba que si f es la densidad de una normal con media cero y desviación típica σ entonces

$$h_{AMISE} = \left(\frac{12\sqrt{\pi}}{n} \right)^{1/5} \sigma$$

¿Cuál es el valor de h_{AMISE} si la media es μ ?

El estimador Naive viene dado por la expresión

$$\hat{f}_{n,N}(x) = \frac{\sum_{i=1}^n \mathbb{I}(x-h < X_i < x+h)}{2nh},$$

que se puede reescribir como

$$\hat{f}_{n,N}(x) = \frac{1}{nh} \sum_{i=1}^n \frac{1}{2} \mathbb{I}(x-h < X_i < x+h) = \frac{1}{nh} \sum_{i=1}^n \omega\left(\frac{x-X_i}{h}\right)$$

donde, ω es la densidad de la distribución uniforme en $(-1, 1)$. Por tanto la aportación de un dato X_i al Estimador Naive en el punto x viene determinado por el valor de

$$\omega\left(\frac{x-X_i}{h}\right),$$

que vale $1/2$ para todos los puntos en $(x-h, x+h)$ independientemente de su proximidad a x .

Si se reemplaza ω por una densidad K (denominada núcleo) con una única moda en cero y simétrica se obtiene el estimador tipo núcleo

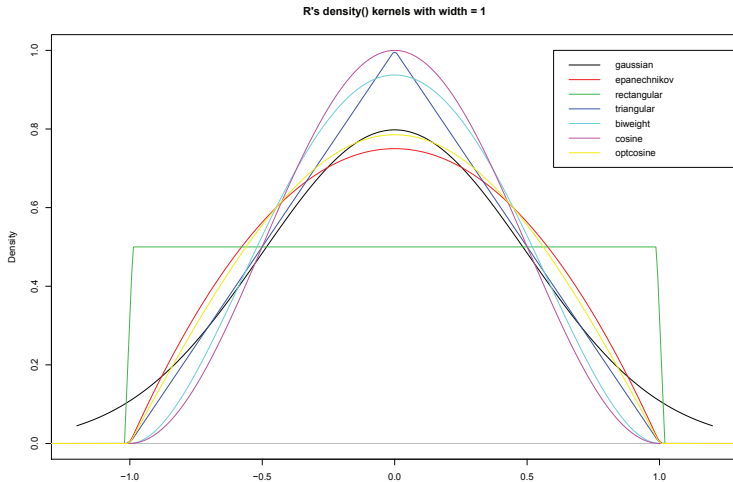
$$\hat{f}_{n,K}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i),$$

donde

$$K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$$

es la densidad de hX_K siendo X_K una variable con función de densidad K .

A continuación se muestran algunos núcleos utilizados en la práctica



El estimador tipo núcleo hereda las propiedades de suavidad del núcleo. Sin embargo la forma del núcleo no tiene un papel tan determinante como la ventana h .

