

Estimación de la función de densidad.

Análisis de Datos 3.

Función de densidad

$$\begin{aligned} f(x) &= \lim_{h \rightarrow 0^+} \frac{P(x \leq X < x + h)}{h} \\ &= \lim_{h \rightarrow 0^+} \frac{P(x - h \leq X < x + h)}{2h} \end{aligned}$$

Estimación de la función de densidad.

Procedimientos para estimar la función de densidad.

Histograma

Teniendo en cuenta la siguiente aproximación para valores pequeños de h

$$\frac{P(x \leq X < x + h)}{h} \simeq \frac{n^{\circ}(x_i \in [x, x + h)) / n}{h}$$

el método consiste en agrupar los valores de la variable en intervalos del tipo $[x_0 - m_1 \cdot h, x_0 + m_2 \cdot h)$ y aplicar la aproximación anterior.

Para $m_1 = 0$ y $m_2 = 1$ la estimación viene dada por:

$$\hat{f}(x) = \frac{n^{\circ}(x_i \in [x, x + h))}{n h}$$

Estimación de la función de densidad.

Procedimientos para estimar la función de densidad.

Método ingenuo (naive)

La función de densidad se puede expresar como:

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

Por lo tanto un estimador directo de la densidad puede ser:

$$\begin{aligned}\hat{f}(x) &= \frac{n^{\circ}(x_i \in (x - h, x + h))}{2hn} = \\ &= \frac{1}{n} \sum_{i=1}^n \frac{1}{h} \cdot K\left(\frac{x - x_i}{h}\right) \quad \text{con} \quad K(x) = \begin{cases} \frac{1}{2} & \text{si } -1 < x < 1 \\ 0 & \text{si no} \end{cases}\end{aligned}$$

Estimación de la función de densidad.

Procedimientos para estimar la función de densidad.

Estimador núcleo

La estimación ingenua puede generalizarse permitiendo que K tome muchos más valores que 0 y $\frac{1}{2}$.

Para ello se pide que K verifique dos condiciones:

1. $K(x) \geq 0$
2. $\int_{-\infty}^{\infty} K(x) dx = 1$

K suele ser una función de densidad que cumple:

1. Es simétrica respecto al 0.
2. Creciente para valores negativos.
3. Decreciente para valores positivos.

Estimación de la función de densidad.

Procedimientos para estimar la función de densidad.

Estimador núcleo

El estimador núcleo de la densidad viene dado por:

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

Propiedades:

1. $\hat{f}(x) \geq 0$
2.
$$\begin{aligned} \int_{-\infty}^{\infty} \hat{f}(x) dx &= \int_{-\infty}^{\infty} \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) dx = \\ \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x - x_i}{h}\right) dx &= \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\infty} K(x) dx = \\ \frac{1}{n} \sum_{i=1}^n 1 &= \frac{n}{n} = 1 \end{aligned}$$

Estimación de la función de densidad.

Procedimientos para estimar la función de densidad.

Medidas de discrepancia de la estimación.

- Error cuadrático medio

$$\text{MSE}_x(\hat{f}) = \mathbb{E}([\hat{f}(x) - f(x)]^2) = [\mathbb{E}\hat{f}(x) - f(x)]^2 + \text{Var}[\hat{f}(x)]$$

- Error cuadrático medio integrado

$$\begin{aligned}\text{MISE}(\hat{f}) &= \mathbb{E} \int [\hat{f}(x) - f(x)]^2 dx \\ &= \int [\mathbb{E}\hat{f}(x) - f(x)]^2 dx + \int \text{Var}[\hat{f}(x)] dx\end{aligned}$$

Estimación de la función de densidad.

Procedimientos para estimar la función de densidad.

Núcleos más habituales.

- cuadrático, parabólico o de Epanechnikov

$$\frac{3}{4}(1 - x^2) \cdot \mathbb{1}_{[-1;+1]}(x)$$

- bicuadrático o cuártico (*biweight*)

$$\frac{15}{16}(1 - x^2)^2 \cdot \mathbb{1}_{[-1;+1]}(x)$$

- gaussiano

$$\frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

Estimación de la función de densidad.

Procedimientos para estimar la función de densidad.

Elección del parámetro de ventana.

El punto crítico en la aplicación de este procedimiento es la elección de un h adecuado al tamaño de la muestra y al tipo de distribución. Un procedimiento que funciona bien en el caso de distribuciones simétricas es considerar el siguiente criterio:

$$h_{opt} = 1,06 \sigma n^{-1/5} \quad \text{siendo } \sigma \text{ la desviación típica}$$

Una alternativa que protege de la influencia de los atípicos consiste en sustituir la desviación típica por una estimación más robusta de la dispersión:

$$h_{opt} = 0,79 \text{ RIC } n^{-1/5} \quad \text{con RIC} = \text{recorrido intercuartílico}$$

Estimación de la función de densidad.

Procedimientos para estimar la función de densidad.

Implementación en R.

- ▶ En R hay la función `density`.
- ▶ Por omisión, usa núcleo gaussiano.
- ▶ Por omisión, usa como ancho de banda h el valor

$$0,9 \times \min \left\{ \sigma; \frac{\text{RIC}}{1,34} \right\} \times n^{\frac{-1}{5}}$$

Estimación de la función de densidad.

Ejemplo en R.

```
> muestra <- rnorm (25)

> sort (round (muestra , 2))
[1] -1.77 -1.13 -0.84 -0.71 -0.59 -0.55 -0.37
[8] -0.33 -0.11  0.03  0.13  0.21  0.24  0.28
[15]  0.32  0.43  0.49  0.64  0.69  0.94  1.07
[22]  1.49  1.66  1.90  2.01

> densidad <- density (muestra)

> names (densidad)
[1] "x"          "y"          "bw"
[4] "n"          "call"       "data.name"
[7] "has.na"
```

Estimación de la función de densidad.

Ejemplo en R.

```
> densidad$n # tamaño muestral  
[1] 25
```

```
> length (densidad$x) # num. puntos calculados  
[1] 512
```

```
> densidad$x [c(100,200,300)] # abscisas  
[1] -1.7248983 -0.5460496 0.6327992
```

```
> densidad$y [c(100,200,300)] # ordenadas  
[1] 0.058970494 0.305151025 0.360550792
```

```
> # teóricamente:  
> dnorm (densidad$x [c(100,200,300)])  
[1] 0.090123382 0.343687113 0.326555306
```

Estimación de la función de densidad.

Ejemplo en R.

```
> 0.9 * sd(muestra) * length(muestra)^(-1/5)  
[1] 0.445286
```

```
> 0.9 * IQR(muestra)/1.34 * length(muestra)^(-1/5)  
[1] 0.3739862
```

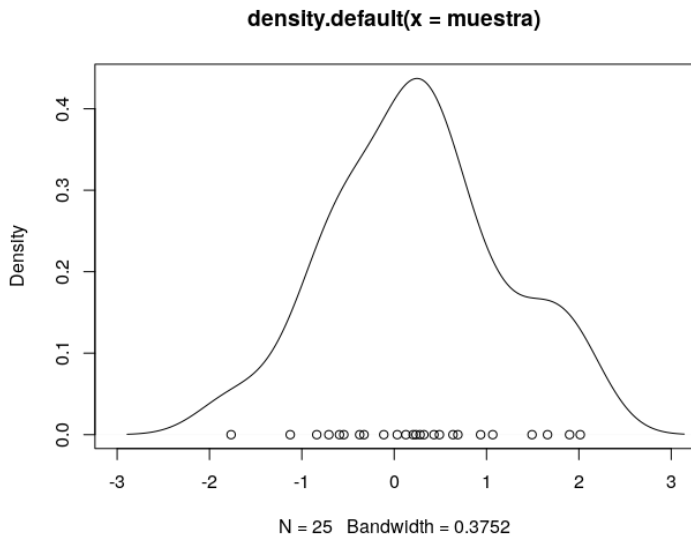
```
> densidad$bw  
[1] 0.3739862
```

```
> plot (densidad)
```

```
> points (cbind (muestra , 0))
```

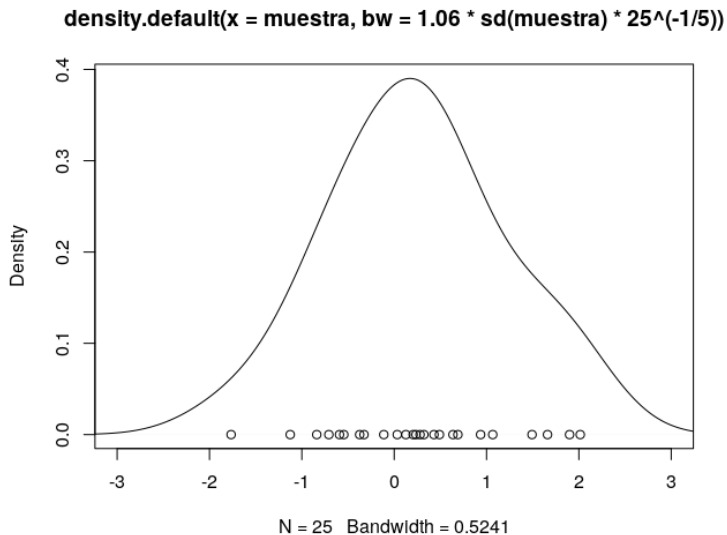
Estimación de la función de densidad.

Elección del parámetro de ventana.



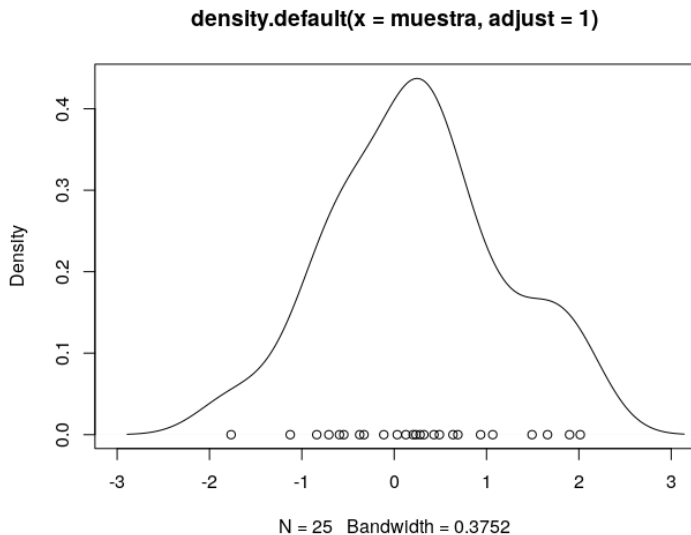
Estimación de la función de densidad.

Elección del parámetro de ventana.



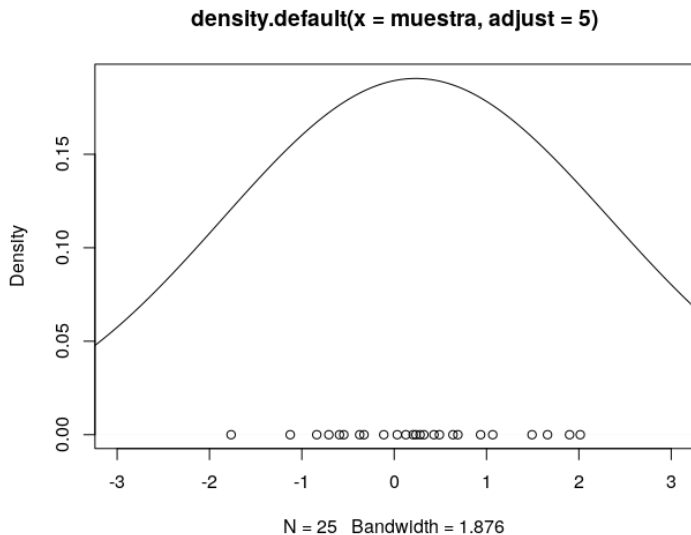
Estimación de la función de densidad.

Elección del parámetro de ventana.



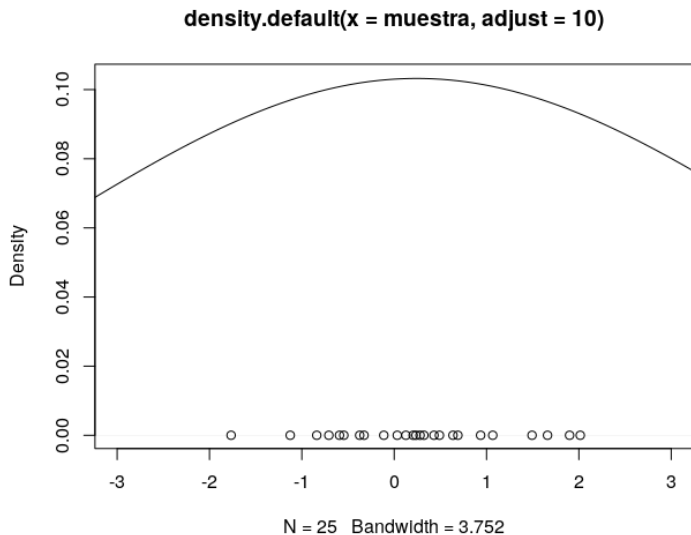
Estimación de la función de densidad.

Elección del parámetro de ventana.



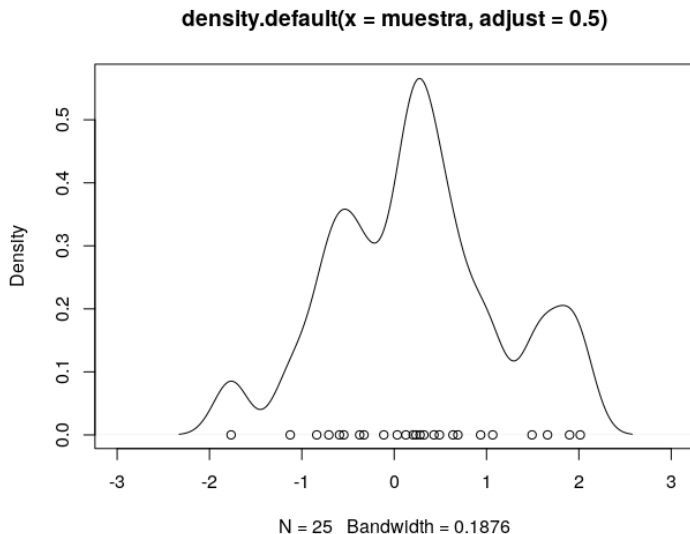
Estimación de la función de densidad.

Elección del parámetro de ventana.



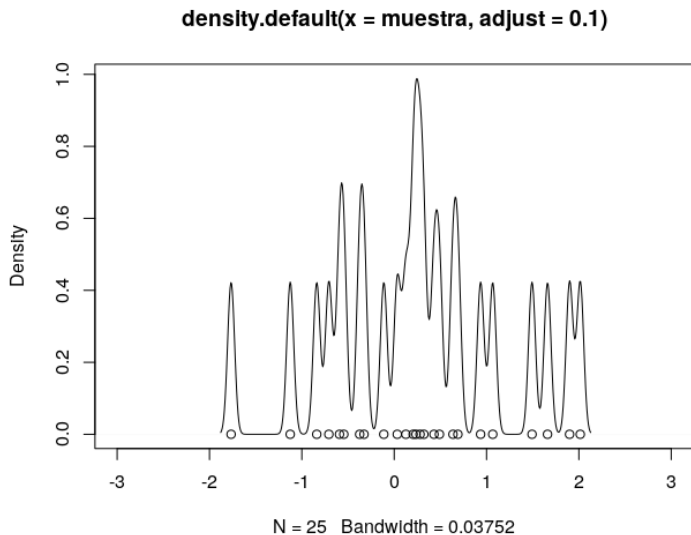
Estimación de la función de densidad.

Elección del parámetro de ventana.



Estimación de la función de densidad.

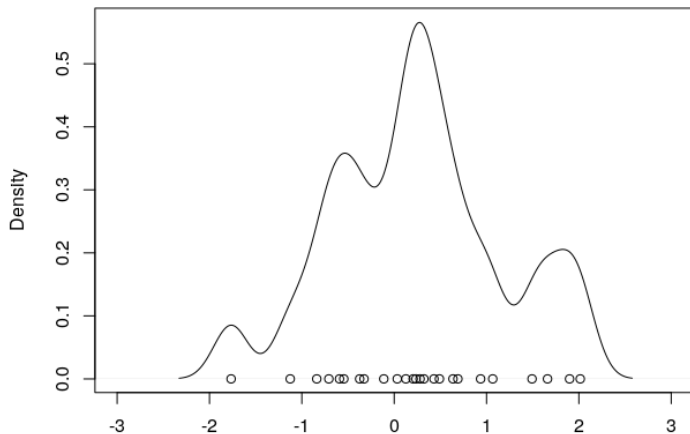
Elección del parámetro de ventana.



Estimación de la función de densidad.

Elección del núcleo.

`density.default(x = muestra, adjust = 0.5, kernel = "gaussian")`

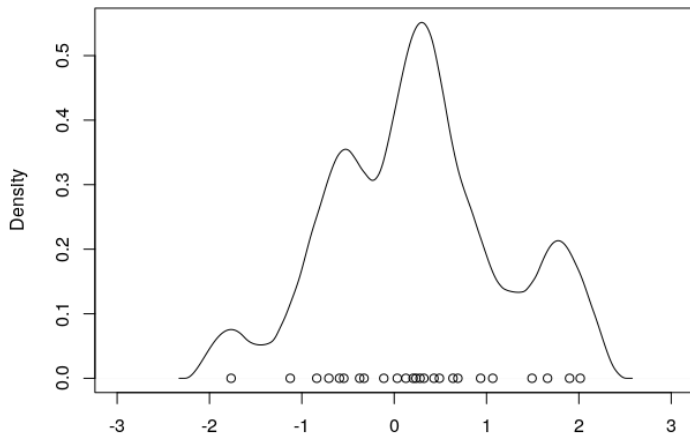


N = 25 Bandwidth = 0.1876

Estimación de la función de densidad.

Elección del núcleo.

density.default(x = muestra, adjust = 0.5, kernel = "biweight")

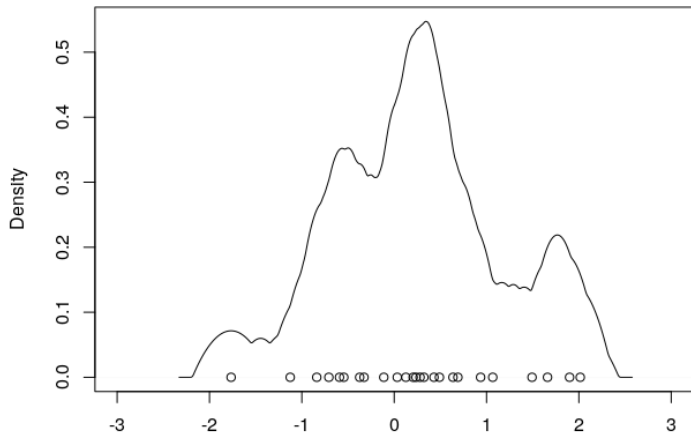


N = 25 Bandwidth = 0.1876

Estimación de la función de densidad.

Elección del núcleo.

```
density.default(x = muestra, adjust = 0.5, kernel = "epanech")
```

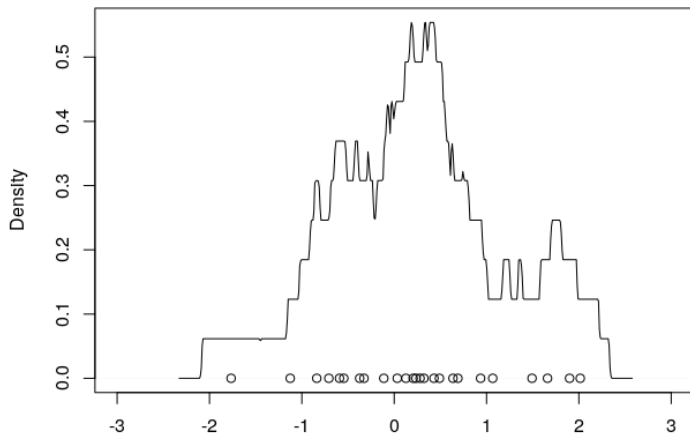


N = 25 Bandwidth = 0.1876

Estimación de la función de densidad.

Elección del núcleo.

`density.default(x = muestra, adjust = 0.5, kernel = "rectangular")`

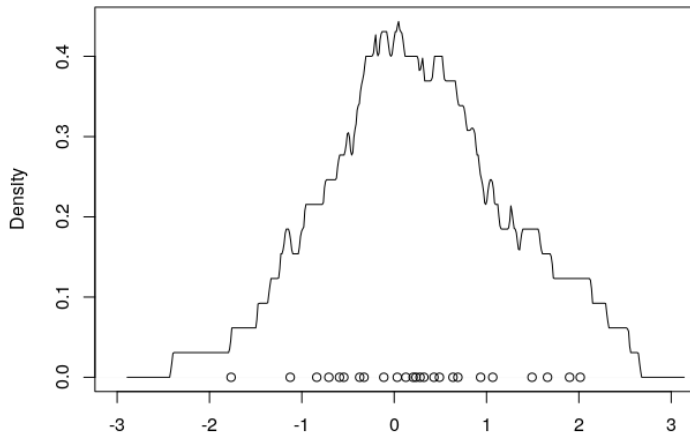


$N = 25$ Bandwidth = 0.1876

Estimación de la función de densidad.

Elección del núcleo.

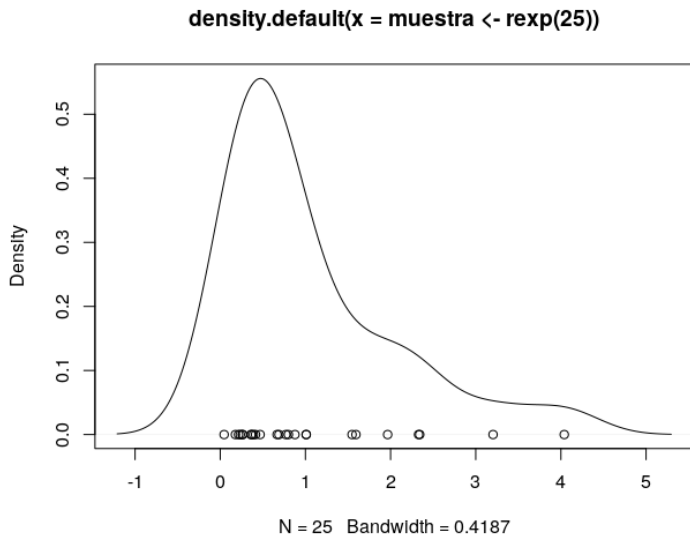
`density.default(x = muestra, adjust = 1, kernel = "rectangular")`



N = 25 Bandwidth = 0.3752

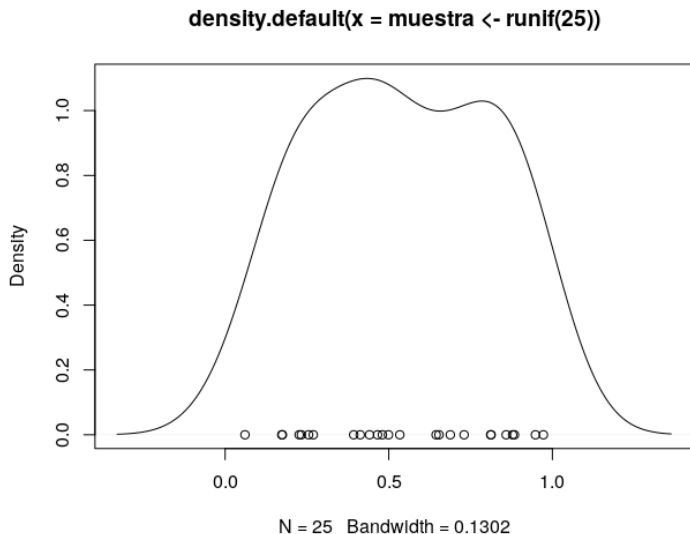
Estimación de la función de densidad.

Distribución asimétrica (exponencial).



Estimación de la función de densidad.

Densidad discontinua (uniforme).



Estimación de la función de densidad.

Ejercicios.

Tareas por resolver:

- ▶ Representar en una gráfica densidades teóricas y estimadas.
- ▶ Construir la función de densidad estimada.