Ciencias de Datos con R: Fundamentos Estadísticos

Ana M. Bianco, Jemina García y Mariela Sued.

Estimación No Paramétrica de la Densidad

Enfoque Paramétrico

- X v.a. continua con densidad $f(x)$: queremos estimar $f(x)$
- Muestra Aleatoria: $X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$.

- Familia paramétrica: asumimos que f pertenece a una familia determinada y que sólo desconocemos sus parámetros

$$f \in \mathcal{M} = \{f(\cdot, \theta), \theta \in \Theta\}.$$

Enfoque Paramétrico

- X v.a. continua con densidad $f(x)$: queremos estimar $f(x)$
- Muestra Aleatoria: $X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$.
- Familia paramétrica: asumimos que f pertenece a una familia determinada y que sólo desconocemos sus parámetros

$$f \in \mathcal{M} = \{f(\cdot, \theta), \theta \in \Theta\}.$$

- Plug–in:
  1) $\widehat{\theta}_n$ estimador de θ
  2) En particular, $\widehat{\theta}_n$ EMV de θ

Enfoque Paramétrico

- X v.a. continua con densidad $f(x)$: queremos estimar $f(x)$
- Muestra Aleatoria: $X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$.
- Familia paramétrica: asumimos que f pertenece a una familia determinada y que sólo desconocemos sus parámetros

$$f \in \mathcal{M} = \{f(\cdot, \theta), \theta \in \Theta\}.$$

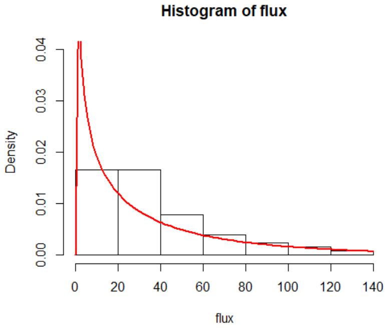
- Plug–in:
  1) $\widehat{\theta}_n$ estimador de θ
  2) En particular, $\widehat{\theta}_n$ EMV de θ

  $\Rightarrow \widehat{f} = f_{\widehat{\theta}}(x)$

- Así , por ejemplo:
  - $X \sim \mathcal{E}(\lambda)$, $\hat{f}(x) = f_{\widehat{\lambda}}(x)$.
  - $X \sim N(\mu, \sigma^2)$, $\hat{f}(x) = f_{\widehat{\mu}, \widehat{\sigma}^2}(x)$.

Ejemplo: Datos de Flux

```
hist(flux, freq=FALSE, ylim=c(0,0.04))
curve(dgamma(x, shape=alpha.MV, rate = lambda.MV), add=TRUE,
col="red", lwd=2, main="Histograma de Flux")
```



Enfoque

X v.a. continua con densidad $f(x)$

Paramétrico: $X \sim F_\theta$

$\hat{F}_\theta = \hat{F}_{\widehat{\theta}}$

$\widehat{f}(x) = f_{\widehat{\theta}}(x)$

Enfoque

X v.a. continua con densidad $f(x)$

Paramétrico: $X \sim F_\theta$

$\hat{F}_\theta = \hat{F}_{\widehat{\theta}}$

$\widehat{f}(x) = f_{\widehat{\theta}}(x)$

No Paramétrico: $X \sim F$

$\hat{F}_n =$ "la empírica"

$\widehat{f}(x) = ?$

Enfoque No Paramétrico

- 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.

Enfoque No Paramétrico

- 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**

Histograma

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

- Sea $\mathcal{A}_j$ una partición de intervalos o clases acotados (bins) disjuntos tales que:

$$\mathbb{R} = \cup_j \mathcal{A}_j$$

- Para cada $x \in \mathcal{A}_j$

$$\widehat{f}(x) = \frac{\#\{X_i : X_i \in \mathcal{A}_j\}}{n|\mathcal{A}_j|}$$

con $|\mathcal{A}_j|$ ancho del bin $\mathcal{A}_j$

Histograma

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

- Sea $\mathcal{A}_j$ una partición de intervalos o clases acotados (bins) disjuntos tales que:

$$\mathbb{R} = \cup_j \mathcal{A}_j$$

- Para cada $x \in \mathcal{A}_j$

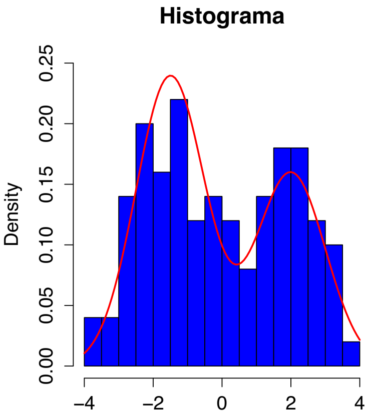
$$\widehat{f}(x) = \frac{\#\{X_i : X_i \in \mathcal{A}_j\}}{n|\mathcal{A}_j|}$$

  con $|\mathcal{A}_j|$ ancho del bin $\mathcal{A}_j$

- El histograma requiere dos parámetros:
  i) <u>ancho del bin</u>
  ii) <u>punto inicial del primer bin</u>

Vamos a las tareas de Clase: items 1 a 3.

Ejemplo: datos simulados



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

Busquemos otra idea...

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

- X con densidad $f(x)$: queremos estimar $f(x)$
- Queremos estimar f sin asumir una determinada forma: sólo asumimos que es f es suave.

Busquemos otra idea...

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

- X con densidad $f(x)$: queremos estimar $f(x)$
- Queremos estimar f sin asumir una determinada forma: sólo asumimos que es f es suave.
- **Idea frecuentista**: por la LGN

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

Busquemos otra idea...

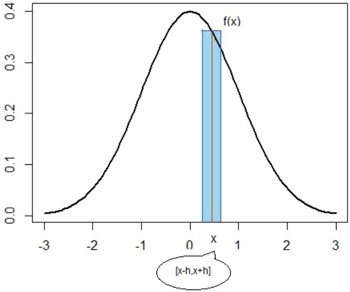
$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

- X con densidad $f(x)$: queremos estimar $f(x)$
- Queremos estimar f sin asumir una determinada forma: sólo asumimos que es f es suave.
- **Idea frecuentista**: por la LGN

$$\begin{aligned}
\mathbb{P}\left(X \in (x-h, x+h)\right) &\approx \frac{\#\{X_i \in (x-h, x+h)\}}{n} \\
\mathbb{P}\left(X \in (x-h, x+h)\right) &= \int_{x-h}^{x+h} f(t)\, dt
\end{aligned}$$

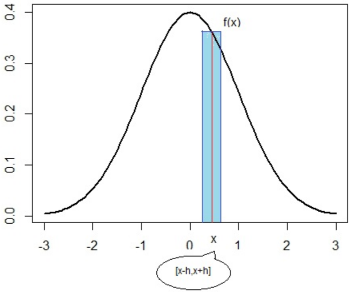
Aproximando analíticamente...

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



Aproximando analíticamente...

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



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

Juntando todo...

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

- $\mathbb{P}\left(X \in (x-h, x+h)\right) \approx \frac{\#\{X_i \in (x-h, x+h)\}}{n}$ por la LGN
- $\mathbb{P}\left(X \in (x-h, x+h)\right) = \int\limits_{x-h}^{x+h} f(t)\, dt$

Juntando todo...

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

- $\mathbb{P}\left(X \in (x-h, x+h)\right) \approx \frac{\#\{X_i \in (x-h, x+h)\}}{n}$ por la LGN
- $\mathbb{P}\left(X \in (x-h, x+h)\right) = \int\limits_{x-h}^{x+h} f(t)\, dt$
- Si h es pequeño y f continua en x,

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

Juntando todo...

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

- $\mathbb{P}\left(X \in (x-h, x+h)\right) \approx \frac{\#\{X_i \in (x-h, x+h)\}}{n}$ por la LGN
- $\mathbb{P}\left(X \in (x-h, x+h)\right) = \int\limits_{x-h}^{x+h} f(t)\, dt$
- Si h es pequeño y f continua en x,

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

- Entonces, podemos aproximar analíticamente

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

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

Juntando todo...

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

- $\mathbb{P}\left(X \in (x-h, x+h)\right) \approx \frac{\#\{X_i \in (x-h, x+h)\}}{n}$ por la LGN
- $\mathbb{P}\left(X \in (x-h, x+h)\right) = \int\limits_{x-h}^{x+h} f(t)\, dt$
- Si h es pequeño y f continua en x,

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

- Entonces, podemos aproximar analíticamente

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

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

Propuesta

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

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

Propuesta

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

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

Notemos que • $\widehat{f}(x) \geq 0$

Propuesta

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

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

Notemos que • $\widehat{f}(x) \geq 0$ • $\int \widehat{f}(x)dx = 1$

Propuesta

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

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

Notemos que • $\widehat{f}(x) \geq 0$ • $\int \widehat{f}(x)dx = 1$

- $X_i \in (x-h, x+h)$
- $x-h < X_i < x+h$
- $-h < X_i - x < h$
- $-1 < \dfrac{X_i - x}{h} < 1$
- $-1 < \dfrac{x - X_i}{h} < 1$

Propuesta

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

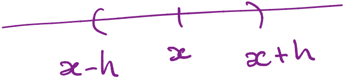
$$\widehat{f}(x) = \frac{\#\{X_i \in (x-h, x+h)\}}{2h\,n}$$

Notemos que $\bullet\, \widehat{f}(x) \geq 0$ $\quad \bullet \int \widehat{f}(x)dx = 1$

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

- 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

$X_1, \ldots, X_n$, i.i.d., $X_i \sim X$ donde $X_i \sim X$

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

Notemos que $\bullet \widehat{f}(x) \geq 0$ $\bullet \int \widehat{f}(x)dx = 1$

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

- Estimador de Parzen

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

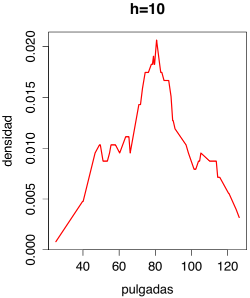
- si $K(t) = \frac{1}{2}\mathcal{I}_{[-1,1]}(t) \quad \Rightarrow$

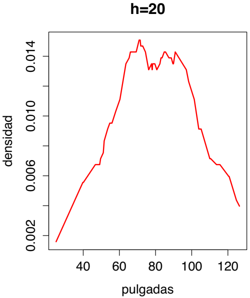
Ker = Núcleo

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

Juntando todo...

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





Vayamos a terminar las tareas de Clase: ítems 4 a 7.