

Ciencias de Datos con R: Fundamentos Estadísticos

Ana M. Bianco, Jemina García y Mariela Sued.

Trailer

Divergencia Kullback Leibler: Fórmula Variacional

- (Ω, \mathcal{F}) espacio de probabilidad
- μ, π dos probabilidades en (Ω, \mathcal{F}) .

$$D(\mu||\pi) = \sup_{h \in C_b(\Omega)} \int_{\Omega} h d\mu - \log \int_{\Omega} e^h d\pi$$

$$D(\mu||\pi) = \begin{cases} \int_{\Omega} \log \left(\frac{d\mu}{d\pi} \right) d\mu & \text{si } \mu \ll \pi, \\ \infty & \text{caso contrario.} \end{cases}$$

- $D(\mu||\pi) \geq 0$
- $D(\mu||\pi) = 0 \longleftrightarrow \mu = \pi$

Divergencia Kullback Leibler en \mathbb{R} : caso discreto

- F, G con puntuales f, g . $F \leftrightarrow f$, $G \leftrightarrow g$.
- $F \ll G$ si y solo si $f \ll g$
- $f \ll g \quad \leftrightarrow \quad g(x) = 0$ implica $f(x) = 0$

$$D(f||g) = \begin{cases} \sum_x \log\left(\frac{f(x)}{g(x)}\right) f(x) & \text{si } f \ll g, \\ \infty & \text{caso contrario.} \end{cases}$$

Divergencia Kullback Leibler en \mathbb{R} : caso continuo

- F, G con densidades f, g . $F \leftrightarrow f$, $G \leftrightarrow g$.
- $F \ll G$ si y solo si $f \ll g$
- $f \ll g \Leftrightarrow g(x) = 0$ implica $f(x) = 0$

$$D(f||g) = \begin{cases} \int_{-\infty}^{+\infty} \log\left(\frac{f(x)}{g(x)}\right) f(x) dx & \text{si } f \ll g, \\ \infty & \text{caso contrario.} \end{cases}$$

Mínima Divergencia

- F distribución en \mathbb{R}
- Familia de distribuciones $\{F(\cdot, \theta) , \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^k$.
- Busco el elemento de la familia mas cercano a F .
 - elemento en la familia: tengo que elegir $\theta = \theta(F)$
 - mas cercano: minimizar Divergencia.

$$D(F||F_{\theta(F)}) = \min_{\theta \in \Theta} D(F||F_\theta).$$

- $\theta(F) = \operatorname{argmin}_{\theta \in \Theta} D(F||F_\theta)$
- $D(F||F_{\theta(F)}) \leq D(F||F_\theta)$, para todo $\theta \in \Theta$.

Funcional: el mas cercano

- $\{F(\cdot, \theta) , \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^k$, familia de distribuciones.
- \mathcal{F} conjunto de distribuciones.

$$\begin{array}{ccc} \mathcal{F} & \rightarrow & \Theta \\ F & \rightarrow & \theta(F), \end{array}$$

$$D(F||F_{\theta(F)}) \leq D(F||F_\theta) , \quad \forall \theta \in \Theta$$

- Si $F \in \{F(\cdot, \theta) , \theta \in \Theta\}$,

$$F = F_{\theta(F)}.$$

Reformulacion del mas cercano

- $D(F||F_\theta) = \mathbb{E}_F[\log(f(X))] - \mathbb{E}_F[\log(f_\theta(X))]$
- Minimizar $D(F||F_\theta)$
- Maximizar $\mathbb{E}_F[\log(f_\theta(X))]$
- $\theta(F) = \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}_F[\log(f_\theta(X))]$
- Estimación: maximizar $\mathbb{E}_{\widehat{F}_n}[\log(f_\theta(X))]$

$$\widehat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}_{\widehat{F}_n}[\log(f_\theta(X))] = \operatorname{argmax}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log(f_\theta(X_i))$$

Consistencia de Máxima verosimilitud.

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}_{\hat{F}_n} [\log(f_\theta(X))] = \operatorname{argmax}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log(f_\theta(X_i))$$

- $D(F||F_\theta) = \mathbb{E}_F[\log(f(X))] - \mathbb{E}_F[\log(f_\theta(X))]$
- $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}_{\hat{F}_n} [\log(f_\theta(X))] \rightarrow \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}_F [\log(f_\theta(X))] = \operatorname{argmin}_{\theta \in \Theta} D(F||F_\theta) = \theta(F)$
- Si $F = F_{\theta_0}$, $\theta(F) = \theta_0$ y por consiguiente

$$\hat{\theta} \rightarrow \theta_0 .$$

Extremos

- Máximos

...extreme value theory (EVT), an approach that has historically been developed to treat phenomena in which extremes (maxima or minima) and not averages play the role of the protagonist.

naturephysics - ver acá
- Generalized Extreme Value Distributions (Gumbel, Fréchet y Weibull juntas)
- Distribución de Excesos
- Familia Pareto Generalizada
- Teorema de Pickands
- Return Levels (percentiles...)

Teorema de Fisher-Tippet (1928) y Gnedenko(1943).

$$(X_i)_{i \geq 1} \text{ iid , } M_n = \max\{X_1, \dots, X_n\}$$

Supongamos que existen b_n, a_n tales que

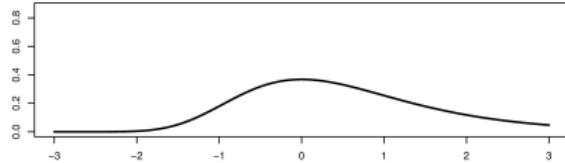
$$\frac{M_n - b_n}{a_n} \xrightarrow{\mathcal{D}} W \sim G$$

¿Puedo dar una caracterización de G ?

Sí, salvo parámetros de posición y escala. ($G(\sigma x + \mu)$)

Tipos de distribución para el máximo

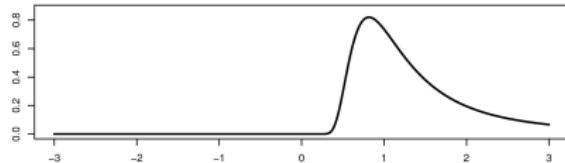
Tipo I



(Gumbel)

$$\text{Tipo I: } \Lambda(z) = \exp\{-e^{-z}\} \quad z \in \mathbb{R}$$

Tipo II



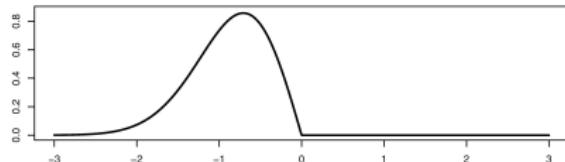
(Frechet)

$$\text{Tipo II: } \Phi_{\alpha}(z) = \begin{cases} 0 & \text{si } z < 0 \\ \exp\{-z^{-\alpha}\} & \text{si } z \geq 0. \end{cases}$$

(Weibull)

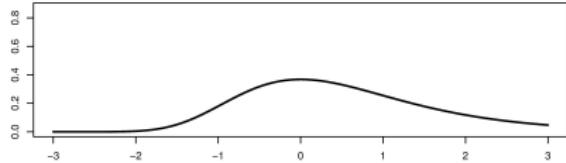
$$\text{Tipo III: } \Psi_{\alpha}(z) = \begin{cases} \exp(-|z|^{\alpha}) & \text{si } z < 0 \\ 1 & \text{si } z \geq 0. \end{cases}$$

Tipo III

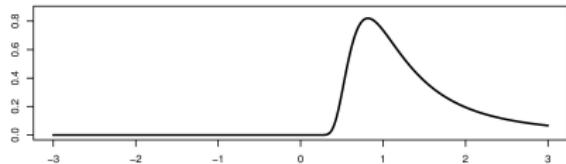


GEV (von Mises(1954), Jenkinson (1955))

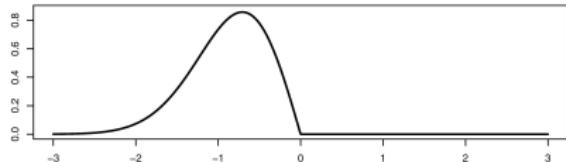
Tipo I



Tipo II



Tipo III



$$G_\gamma(x) = \exp\{-(1 + \gamma x)^{-1/\gamma}\}$$

para $1 + \gamma x > 0$

$$G_0(x) = \exp\{-\exp(-x)\}$$

Si $\gamma = 0$;

Dominio de atracción de F

$$X_i \sim F$$

Qué condiciones pedir sobre F para asegurar que existen $a_n > 0, b_n$ de manera que efectivamente suceda:

$$\frac{M_n - b_n}{a_n} \xrightarrow{\mathcal{D}} G_\gamma(z)$$

En tal caso, diremos que

$$F \in \mathcal{D}(G_\gamma)$$

Distribuciones Conocidas

Distribución	$F(x)$	$F \in D(G_\gamma)$
$N(\mu, \sigma^2)$	$\Phi\left(\frac{x-u}{\sigma}\right)$	G_0
$\mathcal{E}(\lambda)$	$F(x) = 1 - e^{-\lambda x}$	G_0
$\beta(a, b)$	$\int_0^x \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} w^{a-1} (1-w)^{b-1} dw$	$G_{-1/a}$
$U(a, b)$	$F(x) = \frac{x-a}{b-a}, x \in [a, b]$	G_{-1}
Cauchy (a, b)	$F(x) = \frac{1}{\pi} \arctan\left(\frac{x-a}{b}\right) + \frac{1}{2}$	G_1
$ T_n $	$F(x) = \int_{-\infty}^x 2 \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} (1+n^{-1}w^2)^{-(n+1)/2} dw$	$G_{1/n}$

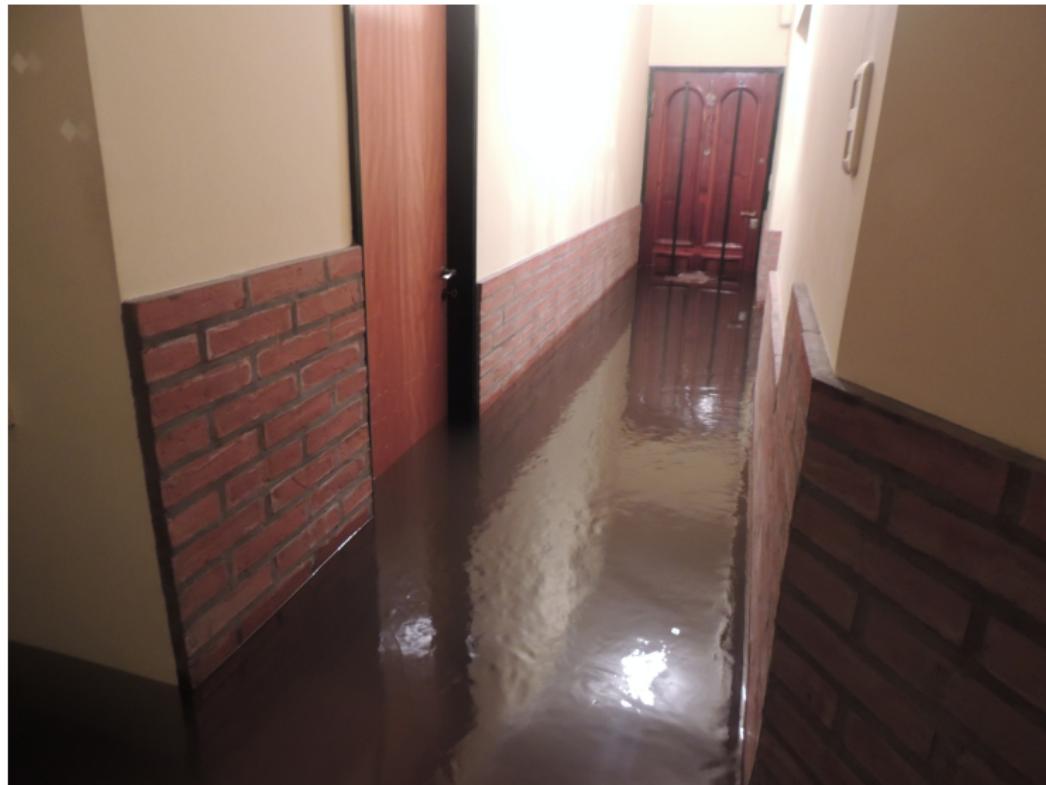
Table:

Distribución de excesos

- $X = \text{ml. de lluvia caídos en un día (de lluvia).}$
- Llevan caido 100 ml: $X > 100$



Distribución de excesos. $X > 100$



Distribución de excesos. $X > 100$



Distribución de excesos. $X > 100$

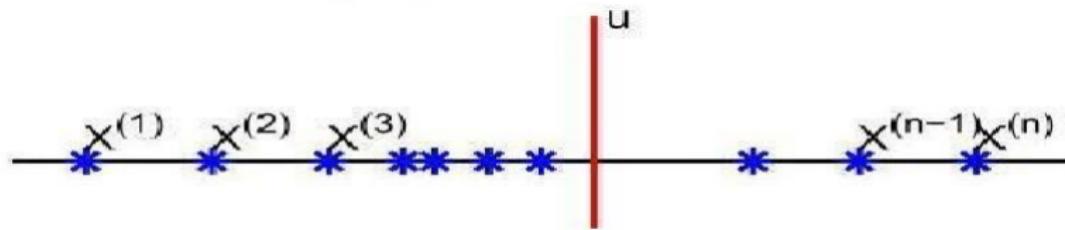


Distribución de excesos

- $X = \text{ml. de lluvia caídos en un día (de lluvia).}$
- Llevan caido 100 ml: $X > 100$
- $P(X > 160 | X > 100)$
- $P(X > 100 + y | X > 100) = 1 - P(X \leq 100 + y | X > 100)$
- $F_{100}(y) := P(X \leq 100 + y | X > 100), \text{ para } y \geq 0$
- $F_u(y) := P(X \leq u + y | X > u), \text{ para } y \geq 0$

Estadística para la Distribución de Excesos

- Queremos estimar $F_u(\cdot) := P(X \leq u + \cdot | X > u)$, para u grande.
- X_1, \dots, X_n muestra
- Poca muestra por encima de u :



Pickands(1975)

$$x^* = \inf\{x : F(x) = 1\}$$

Si $F \in D(G_\gamma) \Rightarrow$

$$\lim_{u \rightarrow x^*} \sup_{0 \leq y < \infty} |F_u(y) - H_{(\sigma(u), \gamma)}(y)| = 0$$

$$H_{(\sigma, \gamma)}(y) = 1 - \left(1 + \frac{\gamma}{\sigma}y\right)^{-1/\gamma}$$

Familia Pareto Generalizada

$$\mathcal{H} = \{H_\theta(y), \theta = (\sigma, \gamma) \in \mathbb{R}_{>0} \times \mathbb{R}\}$$

$$H_{\sigma,\gamma}(y) = 1 - \left(1 + \frac{y\gamma}{\sigma}\right)_+^{-1/\gamma}, \quad y > 0 \text{ (para } \gamma \neq 0)$$

$$H_{\sigma,0}(y) = 1 - e^{-\frac{y}{\sigma}}$$