ANÁLISIS DE EXTREMOS

Edición 2024

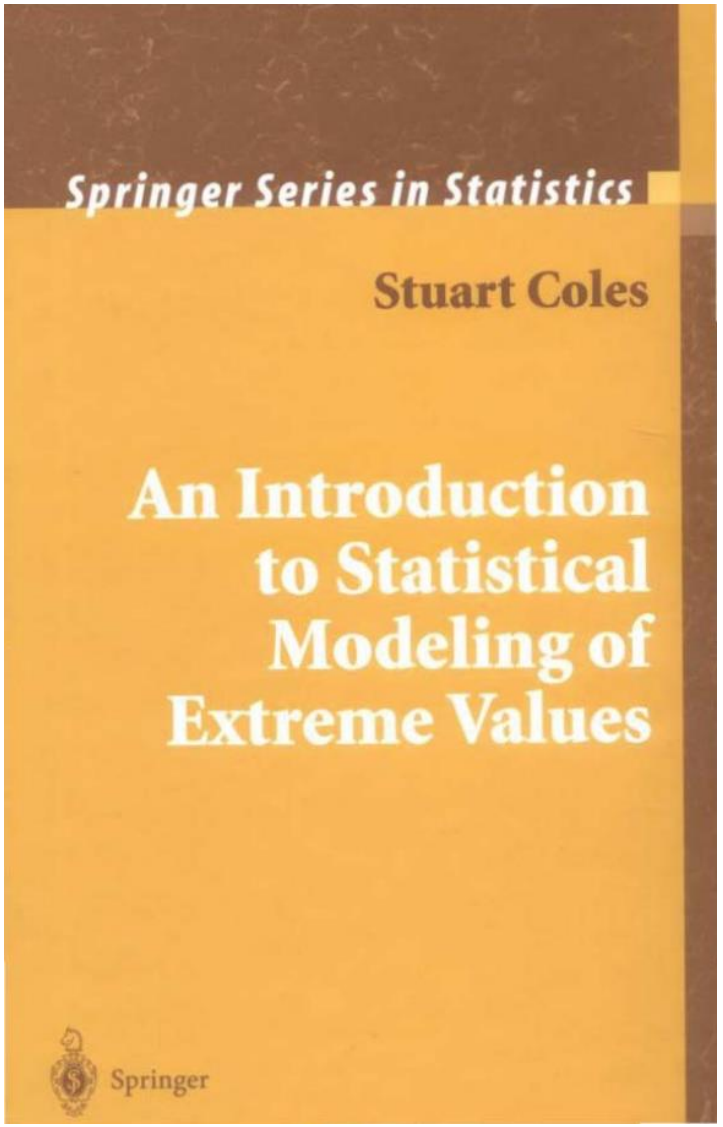
Rafael Terra (en base a notas de Sebastián Solari)

Instituto de Mecánica de los Fluidos e Ingeniería Ambiental (IMFIA)
Facultad de Ingeniería, Universidad de la República, Uruguay

rterra@fing.edu.uy

BIBLIOGRAFÍA

- Coles (2001) “An Introduction to Statistical Modeling of Extreme Values”
- Kottegoda & Rosso (2008) “Applied Statistics for Civil and Environmental Engineers” 2nd Edition
- Hosking & Wallis (1997) “Regional Frequency Analysis. An Approach Based on L-Moments”



CONTENIDO

- Ajuste de distribuciones
- Evaluación de ajustes
- Intervalos de confianza

- Selección del umbral para POT y GP

AJUSTE DE DISTRIBUCIONES

Se refiere a estimar los parámetros θ de una distribución de probabilidad a partir de los datos observados $\{x_i\}$, obteniendo una estimación de los parámetros $\hat{\theta}$.

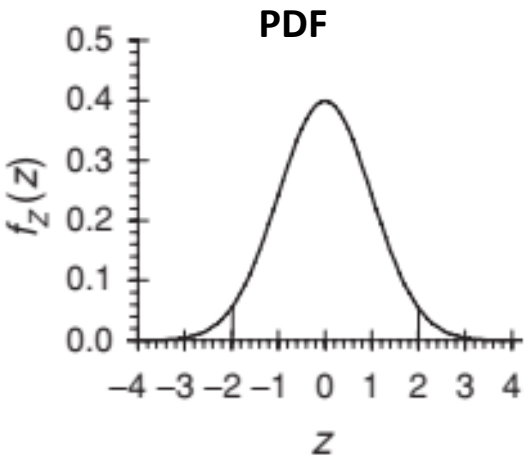
Inferencia estadística: se pretende decir algo respecto al comportamiento de la población (e.g. θ) a partir de la muestra (e.g. $\hat{\theta}$), en este caso asumiendo que la familia de distribuciones de probabilidad de la población es conocida.

Métodos de estimación de parámetros

- Igualdad de momentos
- Momentos L (L-Moments)
- Máxima verosimilitud
- …

MÉTODO DE LOS MOMENTOS

¿Qué son los momentos de una distribución?



La expresión general de los momentos para una variable X continua con densidad de probabilidad (PDF) dada por f_X es:

$$\mu_r^* = E[(X-a)^r] = \int_{-\infty}^{+\infty} (x-a)^r f_X(x)dx$$

MÉTODO DE LOS MOMENTOS

El **valor esperado de X** o **media** es el momento de orden uno de la distribución y es una medida de la tendencia central de la población:

$$\mu = E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx$$

Se puede estimar a partir de la muestra como:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

MÉTODO DE LOS MOMENTOS

El **varianza** es el momento centrado (respecto a la media) de orden dos de la distribución y es una medida de la dispersión de la población:

$$\sigma^2 = Var(X) = E[(X - E(X))^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 f_X(x) dx$$

Se puede estimar a partir de la muestra como:

$$\hat{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x - \hat{\mu})^2$$

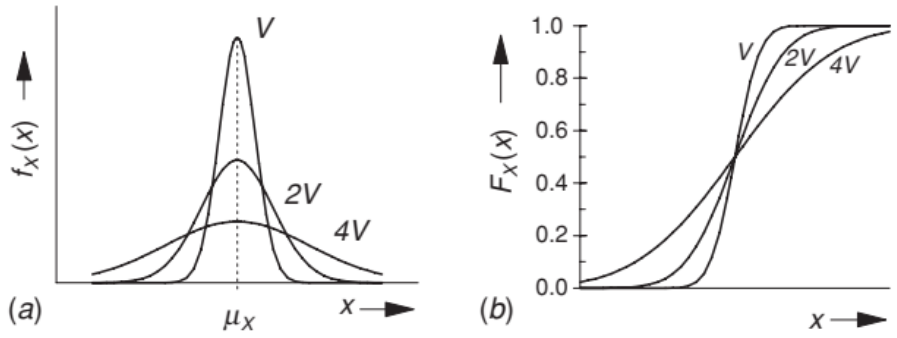


Fig. 3.2.2 Schematic diagrams of (*a*) symmetrical probability density functions and (*b*) cumulative distribution functions of three continuous variables X with different coefficients of variation V.

MÉTODO DE LOS MOMENTOS

El **asimetría** (skewness) es la relación entre los momentos (centrados respecto a la media) de orden tres y orden dos de la distribución y se interpreta como una medida de qué tan simétrica es la distribución respecto a la media:

$$\gamma_1 = \frac{E[(X - E(X))^3]}{\sqrt{\{E[(X - E(X))^2]\}^3}}$$

$$= \frac{\int_{-\infty}^{+\infty} (x - \mu)^3 f_X(x) dx}{\sqrt{\left\{\int_{-\infty}^{+\infty} (x - \mu)^2 f_X(x) dx\right\}^3}}$$

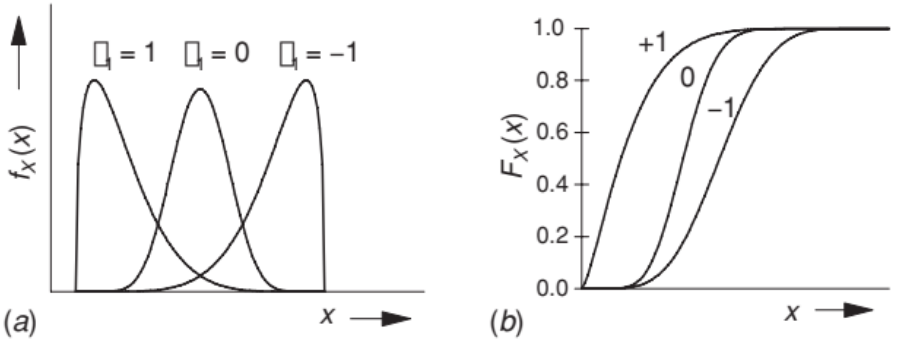
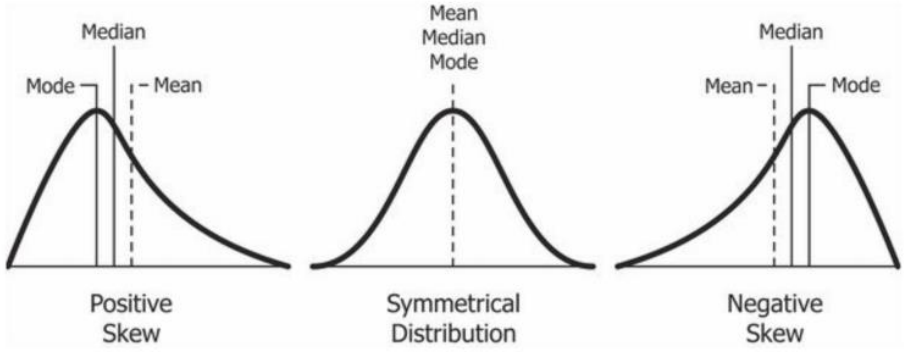


Fig. 3.2.3 Schematic diagrams of (*a*) the probability density functions and (*b*) the cumulative distribution functions of three continuous variables X with coefficients of skewness, $\gamma_1 = 1, 0,$ and -1.

Se puede estimar a partir de la muestra como:

$$\widehat{g_1} = \frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^3}{n\hat{s}^3}$$



MÉTODO DE LOS MOMENTOS

Se basa en igualar los momentos de la distribución, expresados en función de los parámetros de ésta, con los momentos muestrales estimados a partir de los datos. De esta forma queda definido un sistema de ecuaciones a partir del cual es posible despejar los parámetros de la distribución.

Será necesario usar tantos momentos como parámetros tenga la distribución. Típicamente para una distribución de dos parámetros se usan los momentos de orden uno y dos (media y varianza), mientras que para una distribución de tres parámetros será necesario incluir otro momento adicional (típicamente el momento de orden tres mediante la asimetría).

Es un método sencillo pero "poco preciso" ya que los momentos muestrales pueden diferir mucho de los de la población, en particular si la muestra es pequeña ($n < 30$ para los momentos de orden uno y dos, o $n < 100$ para momentos de orden tres o superior).

MÉTODO DE LOS MOMENTOS

Kottegoda&Rosso, 7.2

Para el caso de la Gumbel:

$$F_{X_{\max}}(x) = \exp[-e^{-(x-b)/\alpha}], \tag{7.2.17}$$

$$E[X_{\max}] = \mu = b + n_e\alpha, \tag{7.2.19}$$

n_e=0.5772..

and

$$\mathrm{Var}[X_{\max}] = \sigma^2 = \frac{\pi^2 a}{6}, \tag{7.2.20}$$

$$\alpha = \frac{\sqrt{6}}{\pi}\sigma, \tag{7.2.21}$$

and

$$b = \mu - n_e\alpha = \mu - \frac{n_e\sqrt{6}}{\pi}\sigma. \tag{7.2.22}$$

MÉTODO DE LOS MOMENTOS

Kottegoda&Rosso, 7.2

Para el caso de la GEV:

$$F_{X_{\max}}(x) = \exp\left\{-\left[1 - \frac{k(x-\varepsilon)}{\alpha}\right]^{1/k}\right\}, \tag{7.2.59}$$

$$E[X_{\max}] = \varepsilon + \frac{\alpha}{k}[1 - \Gamma(1+k)], \quad \text{for } k > -1, \tag{7.2.60}$$

and

$$\mathrm{Var}[X_{\max}] = \left(\frac{\alpha}{k}\right)^2 [\Gamma(1+2k) - \Gamma^2(1+k)], \quad \text{for } k > -0.5, \tag{7.2.61}$$

respectively. Therefore, the mean is not defined for $k < -1$, and the variance for $k < -1/2$. The coefficient of skewness is given by

$$\gamma_{1,X_{\max}} = \mathrm{sign}(k)\frac{-\Gamma(1+3k) + 3\Gamma(1+k)\Gamma(1+2k) - 2\Gamma^3(1+k)}{[\Gamma(1+2k) - \Gamma^2(1+k)]^{3/2}}, \quad \text{for } k > -1/3, \tag{7.2.62}$$

MÉTODO DE LOS MOMENTOS

Para el caso de la GEV:

Kottegoda&Rosso, 7.2

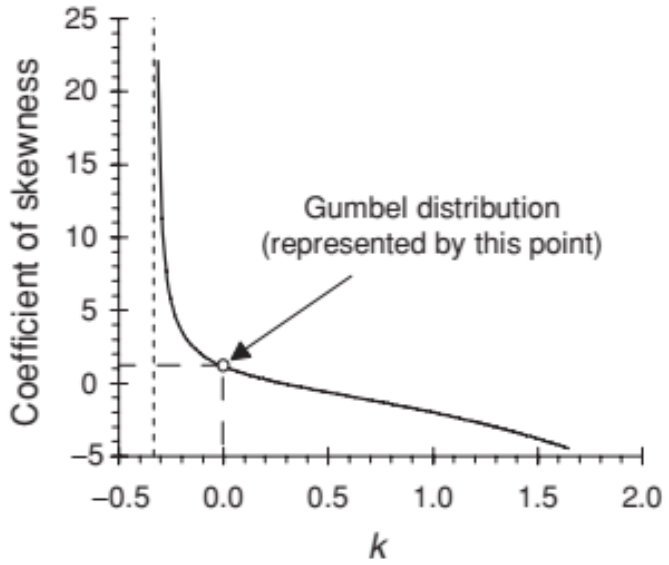
of the data. Since Eq. (7.2.62) indicates that k only depends on the coefficient of skewness for $k > -1/3$, one can solve this equation in k by substituting the sampling skewness coefficient. Then, from Eq. (7.2.61) the scale parameter is found as

$$\alpha = \sqrt{\frac{k^2\sigma^2}{\Gamma(1+2k) - \Gamma^2(1+k)}}, \tag{7.2.63}$$

where the sample variance is substituted for $\sigma^2 = \mathrm{Var}[X_{\max}]$. Finally, the location parameter is computed from

$$\varepsilon = \mu - \frac{\alpha}{k}[1 - \Gamma(1+k)],$$

where the sample mean is substituted for μ.



MÉTODO DE LOS MOMENTOS L (L-MOMENTS)

Los momentos L son otra forma de medir "centralidad", "dispersión", "asimetría", etc. de la población, pero **evitando elevar la variable a potencias mayores a 1**; de esta forma se logra que los momentos L obtenidos a partir de los datos (muestrales) no difieran tanto de los de la población, inclusive para conjunto de datos relativamente reducidos.

Los momentos L se construyen como combinaciones lineales de distintos estadísticos de orden (i.e. estadísticos que se obtienen a partir de la serie ordenada de menor a mayor; $X_{j:n}$ se refiere al dato que tiene posición j al ordenar la serie $\{X_1, \dots, X_i, \dots X_n\}$ de menor a mayor).

$\lambda_1 = E(X_{1:1})$ → medida de centralidad

$\lambda_2 = E(X_{2:2} - X_{1:2})$ → medida de dispersión

$\lambda_3 = E(X_{3:3} - 2X_{2:3} + X_{1:3})$

$\tau_3 = \dfrac{\lambda_3}{\lambda_2}$

} → medida de asimetría

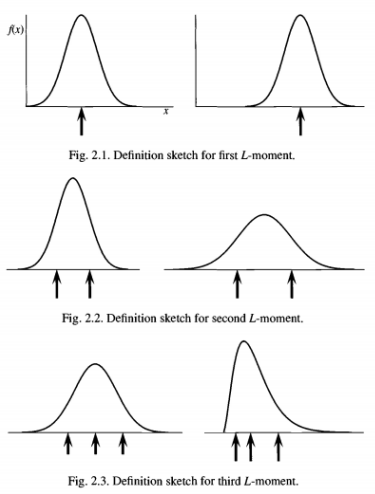


Fig. 2.1. Definition sketch for first *L*-moment.

Fig. 2.2. Definition sketch for second *L*-moment.

Fig. 2.3. Definition sketch for third *L*-moment.

MÉTODO DE LOS MOMENTOS L (L-MOMENTS)

Los momentos L se relacionan con los momentos ponderados por probabilidad (*probability weighted moments; PWM*), para cuyo cálculo también se evita elevar la variable a potencias mayores a uno:

$$\beta_r = \int_0^1 x(u)u^r du \quad \text{en donde} \quad u = F(x)$$

$$\lambda_1 = \beta_0$$
$$\lambda_2 = 2\beta_1 - \beta_0$$
$$\lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0$$

Luego, los momentos L muestrales se estiman de la siguiente forma:

$$b_0 = \frac{1}{n}\sum_{j=1}^{n} x_{j:n}$$
$$b_1 = \frac{1}{n}\sum_{j=2}^{n} \frac{j-1}{n-1} x_{j:n}$$
$$b_2 = \frac{1}{n}\sum_{j=3}^{n} \frac{(j-1)(j-2)}{(n-1)(n-2)} x_{j:n}$$

$$l_1 = b_0$$
$$l_2 = 2b_1 - b_0$$
$$l_3 = 6b_2 - 6b_1 + b_0$$
$$t_r = \frac{l_r}{l_2}$$

MÉTODO DE LOS MOMENTOS L (L-MOMENTS)

GEV:

Hoskins & Wallis, A.6

Parameters (3): ξ (location), α (scale), k (shape).

Range of x: $-\infty < x \le \xi + \alpha/k$ if $k > 0$; $-\infty < x < \infty$ if $k = 0$; $\xi + \alpha/k \le x < \infty$ if $k < 0$.

$$f(x) = \alpha^{-1} e^{-(1-k)y - e^{-y}}, \quad y = \begin{cases} -k^{-1}\log\{1 - k(x-\xi)/\alpha\}, & k \neq 0 \\ (x-\xi)/\alpha, & k = 0 \end{cases} \tag{A.42}$$

$$F(x) = e^{-e^{-y}} \tag{A.43}$$

$$x(F) = \begin{cases} \xi + \alpha\{1 - (-\log F)^k\}/k, & k \neq 0 \\ \xi - \alpha\log(-\log F), & k = 0 \end{cases} \tag{A.44}$$

OJO

$k=-\xi$

$\xi=\mu$

$\alpha=\sigma$

L-moments are defined for $k > -1$.

$$\lambda_1 = \xi + \alpha\{1 - \Gamma(1+k)\}/k \tag{A.50}$$

$$\lambda_2 = \alpha(1 - 2^{-k})\Gamma(1+k)/k \tag{A.51}$$

$$\tau_3 = 2(1 - 3^{-k})/(1 - 2^{-k}) - 3 \tag{A.52}$$

$$\tau_4 = \{5(1 - 4^{-k}) - 10(1 - 3^{-k}) + 6(1 - 2^{-k})\}/(1 - 2^{-k}) \tag{A.53}$$

MÉTODO DE LOS MOMENTOS L (L-MOMENTS)

Hoskins & Wallis, A.6

GEV:

L-moments are defined for $k > -1$.

$$\lambda_1 = \xi + \alpha\{1 - \Gamma(1+k)\}/k \tag{A.50}$$

$$\lambda_2 = \alpha(1 - 2^{-k})\Gamma(1+k)/k \tag{A.51}$$

$$\tau_3 = 2(1 - 3^{-k})/(1 - 2^{-k}) - 3 \tag{A.52}$$

$$\tau_4 = \{5(1 - 4^{-k}) - 10(1 - 3^{-k}) + 6(1 - 2^{-k})\}/(1 - 2^{-k}) \tag{A.53}$$

To estimate k, Eq. (A.52) must be solved for k. No explicit solution is possible, but the following approximation, given by Hosking et al. (1985b), has accuracy better than 9×10^{-4} for $-0.5 \le \tau_3 \le 0.5$:

$$k \approx 7.8590c + 2.9554c^2, \qquad c = \frac{2}{3 + \tau_3} - \frac{\log 2}{\log 3}. \tag{A.55}$$

The other parameters are then given by

$$\alpha = \frac{\lambda_2 k}{(1 - 2^{-k})\Gamma(1+k)}, \qquad \xi = \lambda_1 - \alpha\{1 - \Gamma(1+k)\}/k. \tag{A.56}$$

MÉTODO DE LOS MOMENTOS L (L-MOMENTS)

GP:

Hoskins & Wallis, A.5

Parameters (3): ξ (location), α (scale), k (shape).

Range of x: $\xi \le x \le \xi + \alpha/k$ if $k > 0$; $\xi \le x < \infty$ if $k \le 0$.

$$f(x) = \alpha^{-1} e^{-(1-k)y}, \quad y = \begin{cases} -k^{-1}\log\{1 - k(x-\xi)/\alpha\}, & k \neq 0 \\ (x-\xi)/\alpha, & k = 0 \end{cases} \tag{A.32}$$

$$F(x) = 1 - e^{-y} \tag{A.33}$$

$$x(F) = \begin{cases} \xi + \alpha\{1 - (1-F)^k\}/k, & k \neq 0 \\ \xi - \alpha\log(1-F), & k = 0 \end{cases} \tag{A.34}$$

Special cases: $k = 0$ is the exponential distribution; $k = 1$ is the uniform distribution on the interval $\xi \le x \le \xi + \alpha$.

OJO

$k=-\xi$

$\xi=\mu$

$\alpha=\sigma$

L-moments

L-moments are defined for $k > -1$.

$$\lambda_1 = \xi + \alpha/(1+k) \tag{A.35}$$

$$\lambda_2 = \alpha/\{(1+k)(2+k)\} \tag{A.36}$$

$$\tau_3 = (1-k)/(3+k) \tag{A.37}$$

$$\tau_4 = (1-k)(2-k)/\{(3+k)(4+k)\} \tag{A.38}$$

MÉTODO DE LOS MOMENTOS L (L-MOMENTS)

GP:

Hoskins & Wallis, A.5

If ξ is known, the two parameters α and k are given by

$$k = (\lambda_1 - \xi)/\lambda_2 - 2, \qquad \alpha = (1 + k)(\lambda_1 - \xi). \tag{A.40}$$

If ξ is unknown, the three parameters are given by

$$k = (1 - 3\tau_3)/(1 + \tau_3), \quad \alpha = (1 + k)(2 + k)\lambda_2, \quad \xi = \lambda_1 - (2 + k)\lambda_2. \tag{A.41}$$

MÁXIMA VEROSIMILITUD (ML)

Una de las formas más habituales y más aceptada de estimar los parámetros de una distribución es mediante el método de máxima verosimilitud (*maximum likelihood*), el cual se basa en determinar los valores de los parámetros de la distribución que maximizan la función de verosimilitud, calculada para la muestra disponible como:

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i; \theta)$$

En general resulta conveniente trabajar con el logaritmo de la función de verosimilitud (*log-likelihood function*). Dado que el logaritmo es una transformación monotónica (mantiene el orden de los valores originales), los valores de los parámetros que maximicen el logaritmo de la función de verosimilitud también maximizarán la función de verosimilitud:

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_X(x_i; \theta)$$

MÁXIMA VEROSIMILITUD (ML)

Coles, 3.3

GEV:

Under the assumption that $Z_1, \ldots, Z_m$ are independent variables having the GEV distribution, the log-likelihood for the GEV parameters when $\xi \neq 0$ is

$$\ell(\mu, \sigma, \xi) = -m \log \sigma - (1 + 1/\xi) \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right] - \sum_{i=1}^{m} \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right]^{-1/\xi}, \quad (3.7)$$

provided that

$$1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) > 0, \text{ for } i = 1, \ldots, m. \quad (3.8)$$

gevfit.m

MÁXIMA VEROSIMILITUD (ML)

Coles, 4.3

GP:

4.3.2 Parameter Estimation

Having determined a threshold, the parameters of the generalized Pareto distribution can be estimated by maximum likelihood. Suppose that the values $y_1, \ldots, y_k$ are the k excesses of a threshold u. For $\xi \neq 0$ the log-likelihood is derived from (4.2) as

$$\ell(\sigma, \xi) = -k \log \sigma - (1 + 1/\xi) \sum_{i=1}^{k} \log(1 + \xi y_i/\sigma), \qquad (4.10)$$

provided $(1 + \sigma^{-1}\xi y_i) > 0$ for $i = 1, \ldots, k$; otherwise, $\ell(\sigma, \xi) = -\infty$. In the case $\xi = 0$ the log-likelihood is obtained from (4.4) as

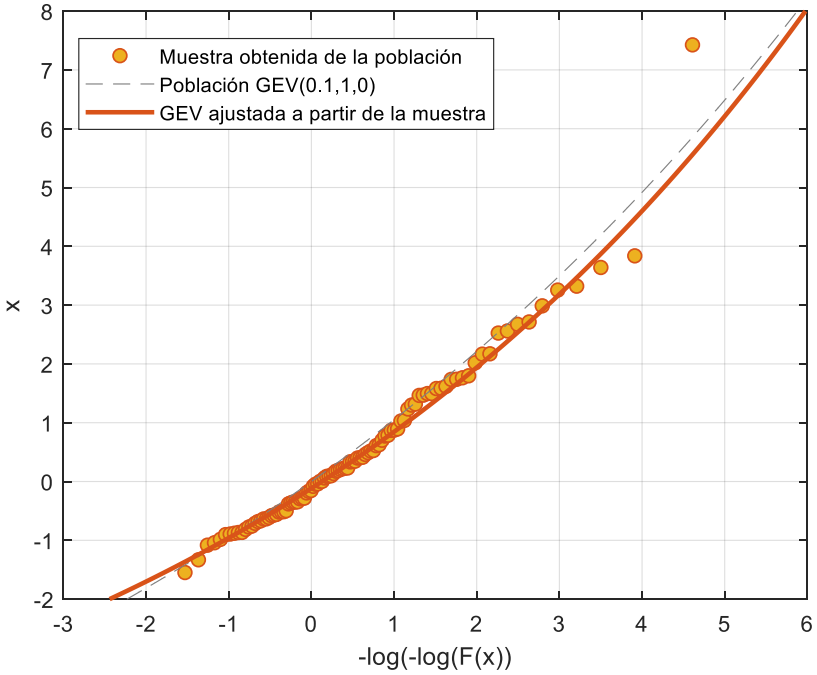
$$\ell(\sigma) = -k \log \sigma - \sigma^{-1} \sum_{i=1}^{k} y_i.$$

gpfit.m

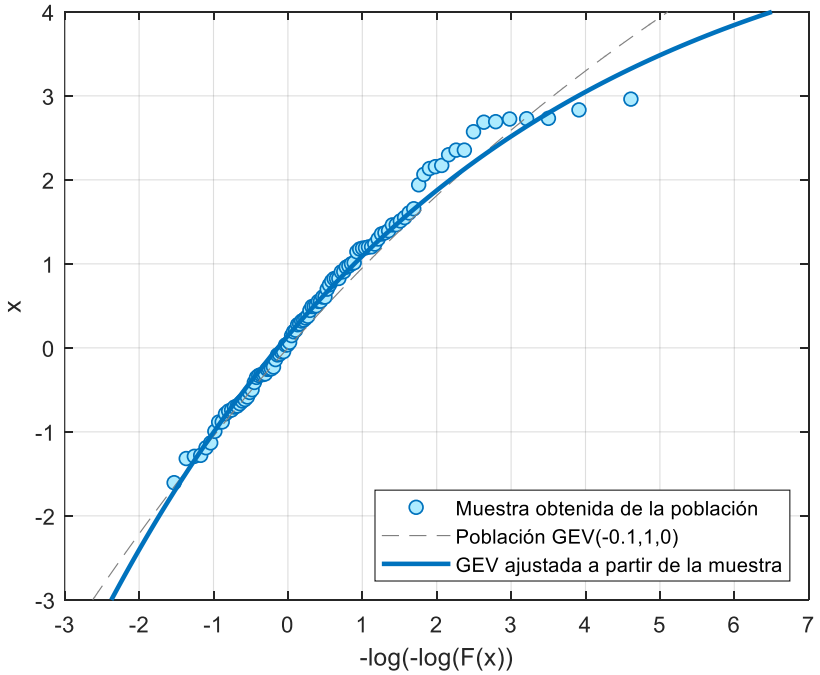
VERIFICACIÓN DE LA BONDAD DEL AJUSTE (GOODNESS OF FIT; GOF)

Cualitativo → métodos gráficos → Escala Gumbel

Tipo II o Fréchet

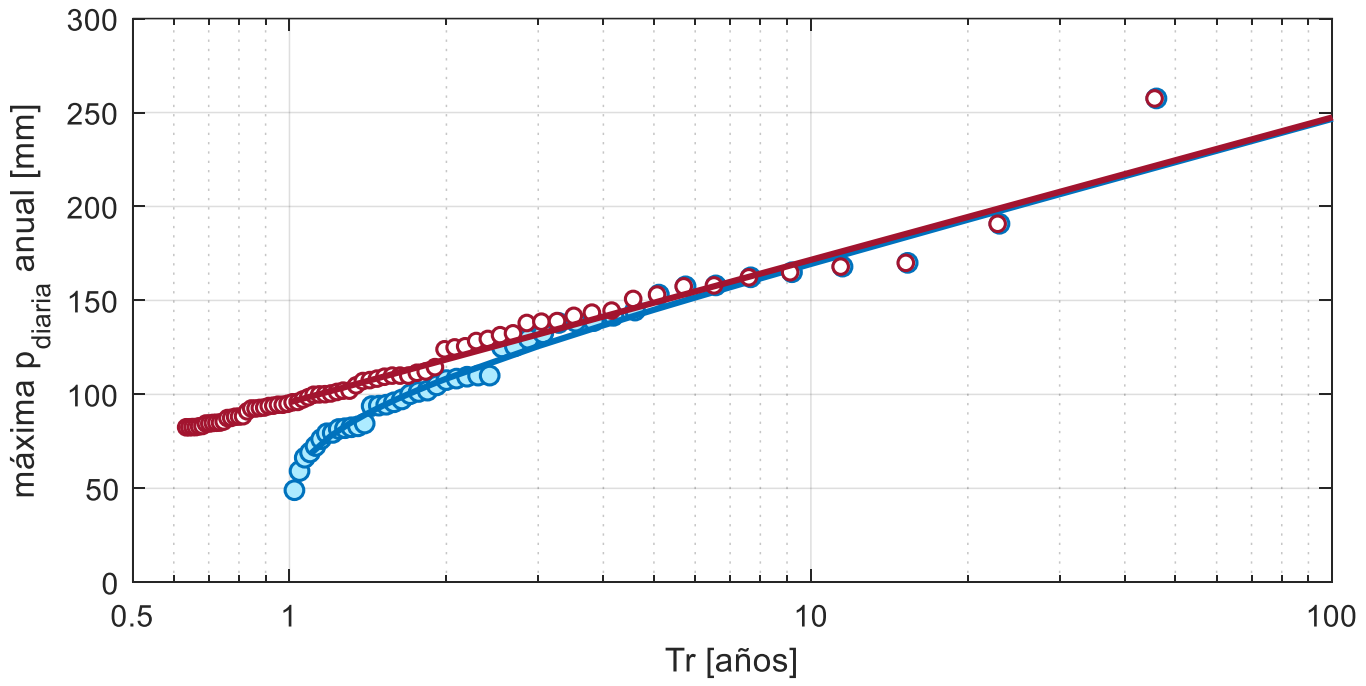


Tipo III o Weibull



VERIFICACIÓN DE LA BONDAD DEL AJUSTE (GOODNESS OF FIT; GOF)

Cualitativo → métodos gráficos → Tr-Intensidad



VERIFICACIÓN DE LA BONDAD DEL AJUSTE (GOODNESS OF FIT; GOF)

Cuantitativo → test de bondad de ajuste

Los test de bondad de ajuste (*goodness of fit tests*) son un subconjunto de los test de hipótesis (*hyphotesis tests*).

La lógica del test de hipótesis es la siguiente:

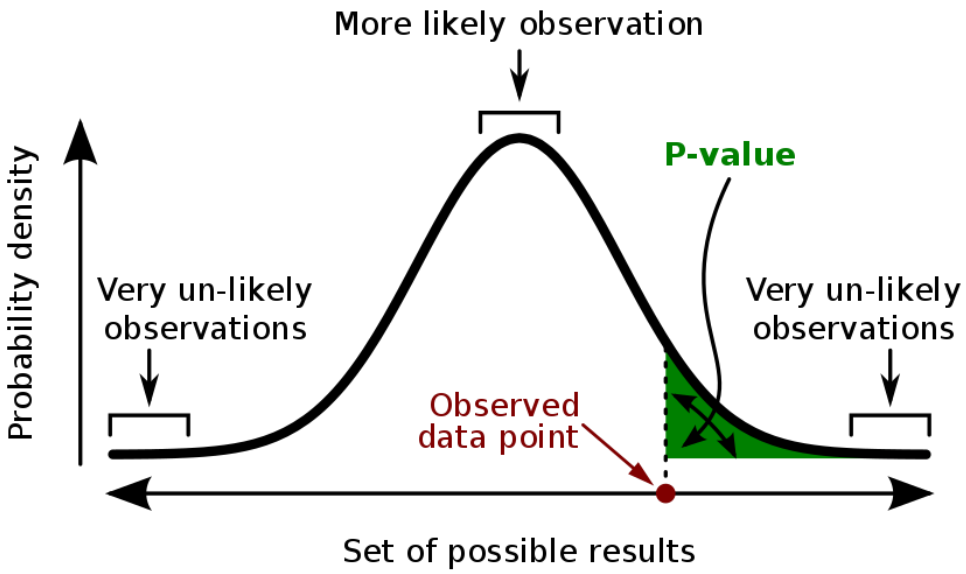
Definido un estadístico θ (i.e. algún valor que es posible calcular a partir de la muestra) y definida una hipótesis nula H_0, se determina cuál es el comportamiento del estadístico bajo la hipótesis nula, o sea, cuál debe ser la distribución de probabilidad del estadístico si la hipótesis nula es verdadera.

Luego se estima el estadístico del test a partir de la muestra ($\hat{\theta}$) y se compara el valor obtenido con los valores que sería esperable obtener si la hipótesis nula H_0 es correcta. En particular se determina la probabilidad de obtener un valor igual o superior al estadístico bajo la hipótesis nula (p-valor) y si esta probabilidad es menor o igual a un valor de referencia predefinido (nivel de significancia α), entonces se dice que hay evidencia significativa para rechazar la hipótesis nula.

VERIFICACIÓN DE LA BONDAD DEL AJUSTE (GOODNESS OF FIT; GOF)

Cuantitativo → test de bondad de ajuste

Los test de bondad de ajuste (*goodness of fit tests*) son un subconjunto de los test de hipótesis (*hyphotesis tests*).



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Verificación de la bondad del ajuste (goodness of fit; GoF)

- Chi-cuadrado → compara PDF
- Kolmogorov-Smirnov → compara CDF
- Anderson-Darling → compara CDF con más peso en las colas

Verificación de la bondad del ajuste (goodness of fit; GoF)

Chi-cuadrado

El estadístico utilizado es la sumatoria de la diferencia entre el número de datos observados y número de datos esperados en una serie de clases o rango de valores.

Para una muestra n suficientemente grande este estadístico sigue una distribución de probabilidad X^2 con $\nu = l - k - 1$, en donde l es el número de clases usado para calcular el estadístico y k el número de parámetros de la distribución estimados a partir de los datos.

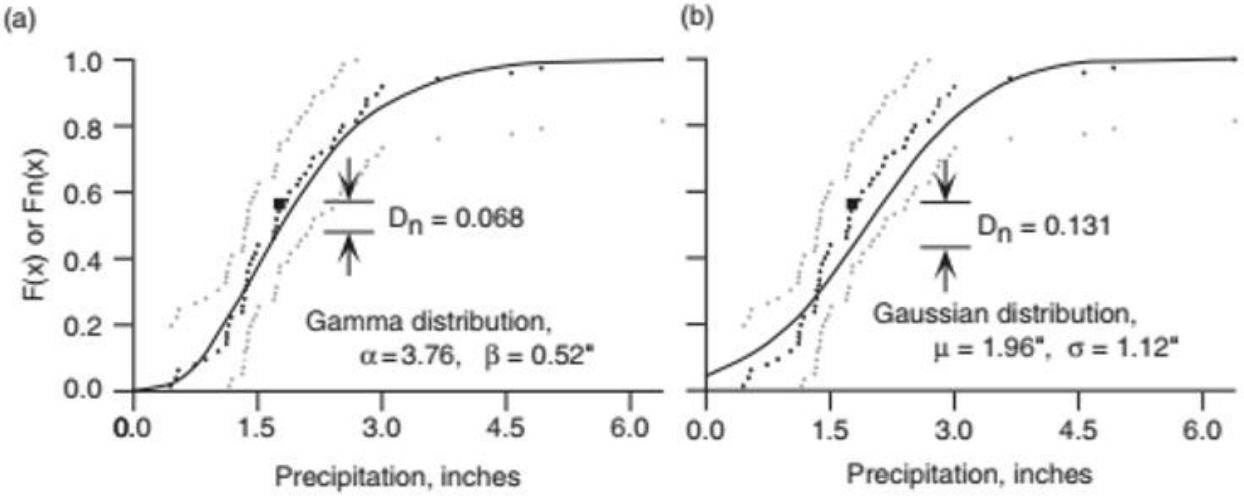
$$X^2 = \sum_{i=1}^{l} \frac{(O_i - E_i)^2}{E_i}.$$

VERIFICACIÓN DE LA BONDAD DEL AJUSTE (GOODNESS OF FIT; GOF)

Kolmogorov-Smirnov

A diferencia del test Chi-cuadrado, cuyo estadístico "mide" la diferencia entre la EPDF y la PDF de la H0, el test de Kolmogorov-Smirnov utiliza la distancia máxima, en términos de probabilidad acumulada entre la ECDF y la CDF de la hipótesis nula H0.

$$D_n = \sup |F_n(x) - F_0(x)$$



Atención!! El test K-S asume que los parámetros de la distribución de la hipótesis nula H0 no fueron estimados a partir de los datos de la muestra. Si este fuera el caso, para calcular el p-valor se debe recurrir al test de Lilliefors o bien estimar uno mismo la distribución del estadístico mediante simulación.

VERIFICACIÓN DE LA BONDAD DEL AJUSTE (GOODNESS OF FIT; GOF)

Anderson-Darling

El estadístico usado en el test de Anderson-Darling también “mide” la distancia entre la ECDF y la CDF correspondiente a la hipótesis nula, pero dando más peso a las colas de la distribución. Existen varias versiones:

than to the data in the central part of the distribution. In equation (4), z_i is the cumulative distribution function $F(x)$ evaluated at the order statistics $x_i; i=1,\ldots,n$ (i.e., $x_1 \leq \ldots \leq x_n$).

$$A^2 = -\frac{1}{n}\sum_{i=1}^{n}\left\{(2i-1)\left[\log(z_i)+\log\left(1-z_{(n+1-i)}\right)\right]\right\}-n \qquad (4)$$

$$A_R^2 = \frac{n}{2}-\sum_{i=1}^{n}\left[\left(2-\frac{2i-1}{n}\right)\log(1-z_i)+2z_i\right] \qquad (5)$$

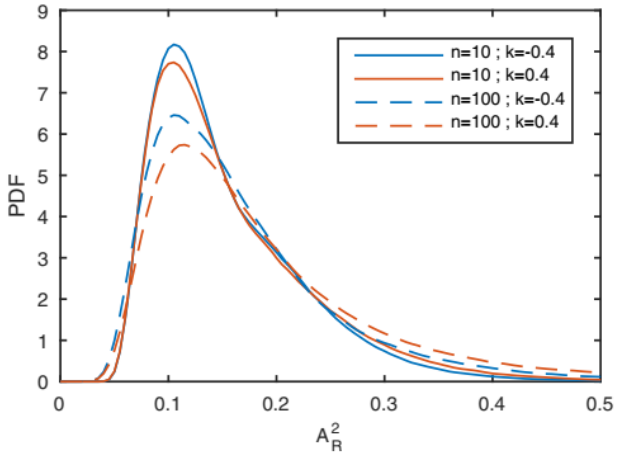


Figure A1. Example of the probability density function of the A_R^2 statistic obtained from the simulations for different values of the shape parameter k and sample length n.

La distribución del estadístico dependerá del número de datos de la muestra, de si estos datos se usan o no para estimar los parámetros de la distribución y del método usado en la estimación, Existen tablas que dan valores críticos del estadístico para algunos p-valores, pero son acotadas. Es habitual tener que recurrir simulaciones para estimar la distribución del estadístico en los casos que se estén estudiando.

INTERVALOS DE CONFIANZA

Dado un estadístico θ (por ejemplo, un parámetro de la distribución o el cuantil de 100 años de período de retorno calculado con la misma), cuya estimación a partir de la muestra es $\hat{\theta}$ y cuyo valor poblacional es θ_0 (desconocido), se busca definir un intervalo $[\theta_{inf}, \theta_{sup}]$ de confianza de $(1-\alpha)$ en el cual:

$$\Pr\left(\theta_{inf} \leq \theta_0 \leq \theta_{sup}\right) = 1 - \alpha$$

De esta forma se "sacrifica precisión" (ya no trabajamos con un único valor estimado sino con un rango de valores) pero se tiene información en cuanto a la incertidumbre de la estimación.

INTERVALOS DE CONFIANZA

- ML → Método Delta
  - Adecuado para los parámetros; inadecuado para cuantiles de alto período de retorno.

- Bootstrapping
  - General, cualquiera sea el método usado para el ajuste (estimación de los parámetros).
  - Basado en simulación.

INTERVALOS DE CONFIANZA

ML → Método Delta

Si los parámetros de la distribución se estimaron mediante máxima verosimilitud, entonces su distribución de probabilidad es aproximadamente normal, con media el valor estimado ($\hat{\theta}$) y varianza dada por la inversa de I_E (expected information matrix), la cual mide la curvatura esperada de la superficie de la función del logaritmo de la verosimilitud en el entorno del valor estimado.

$$\hat{\theta}_0 \dot{\sim} \text{MVN}_d(\theta_0, I_E(\theta_0)^{-1}),$$

Coles, 2.6.4

where

$$I_E(\theta) = \begin{bmatrix} e_{1,1}(\theta) & \cdots & & e_{1,d}(\theta) \\ \vdots & \ddots & e_{i,j}(\theta) & \vdots \\ & e_{j,i}(\theta) & \ddots & \\ e_{d,1}(\theta) & \cdots & & e_{d,d}(\theta) \end{bmatrix},$$

with

$$e_{i,j}(\theta) = E\left\{-\frac{\partial^2}{\partial\theta_i \partial\theta_j}\ell(\theta)\right\}.$$

Theorem 2.2 can be used to obtain approximate confidence intervals for individual components of $\theta_0 = (\theta_1, \ldots, \theta_d)$. Denoting an arbitrary term in the inverse of $I_E(\theta_0)$ by $\psi_{i,j}$, it follows from the properties of the multivariate normal distribution that, for large n,

$$\hat{\theta}_i \dot{\sim} N(\theta_i, \psi_{i,i}).$$

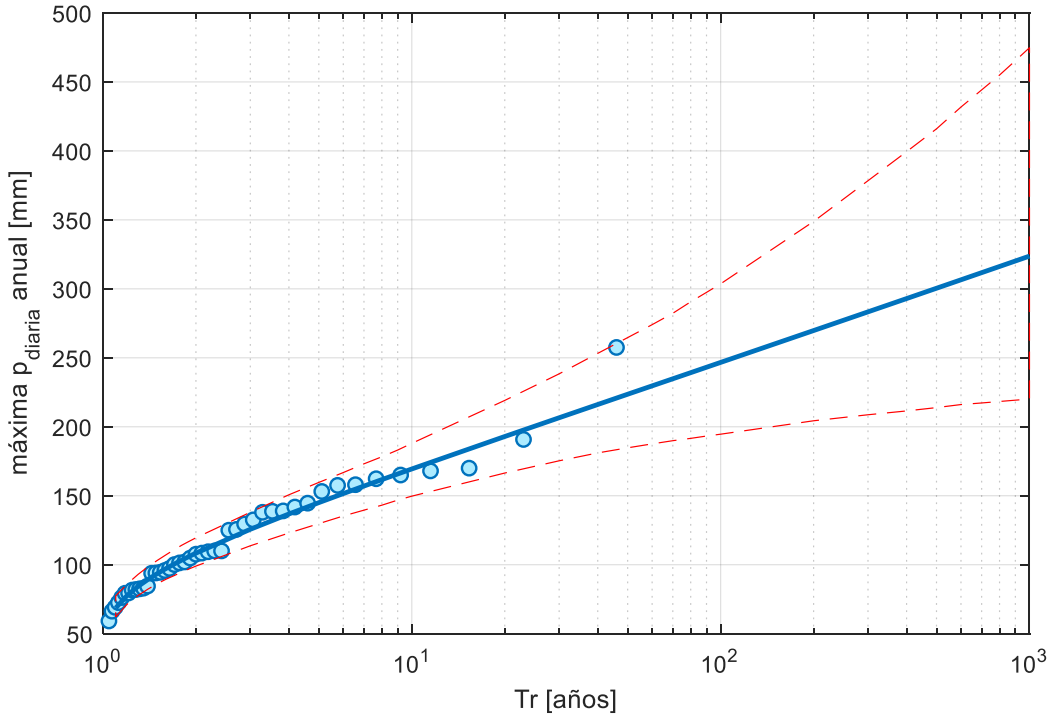
Hence, if $\psi_{i,i}$ were known, an approximate $(1-\alpha)$ confidence interval for θ_i would be

$$\hat{\theta}_i \pm z_{\frac{\alpha}{2}}\sqrt{\psi_{i,i}}, \qquad (2.10)$$

INTERVALOS DE CONFIANZA

ML → Método Delta

El principal inconveniente del método delta en el marco de la teoría de extremos es que no logra capturar la asimetría que existe en los intervalos de confianza de los cuantiles de alto período de retorno (i.e. al asumir normalidad los intervalos de confianza siempre son simétricos).



INTERVALOS DE CONFIANZA

Bootstrapping

Se basa en simular de forma aleatoria muchas realizaciones posibles de la muestra… a partir de la muestra.

Con cada una de las realizaciones se calcula el estadístico de interés, lo que permite construir una distribución empírica del mismo.

Luego se utiliza esta distribución para estimar los intervalos de confianza.

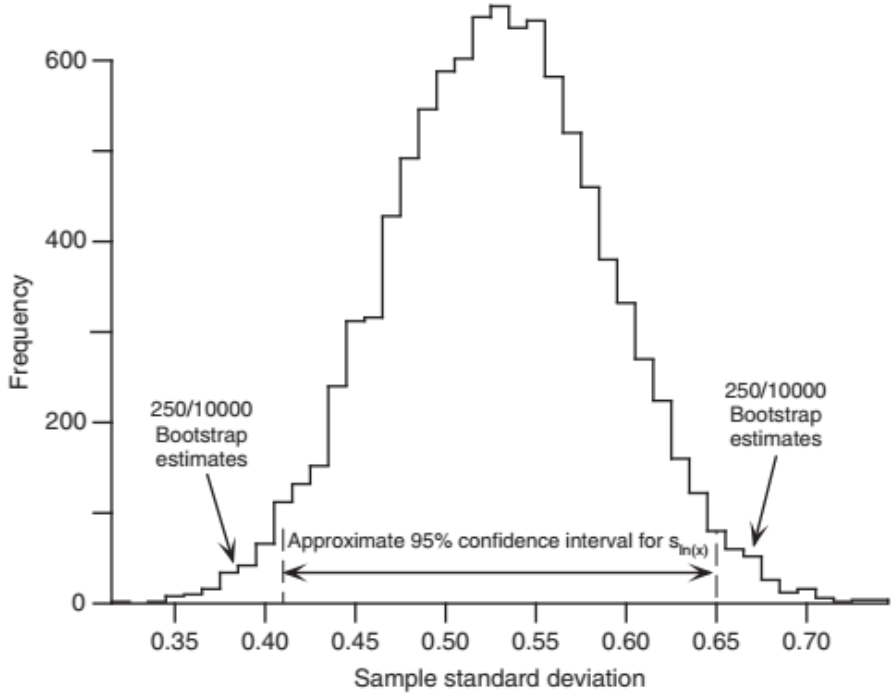


FIGURE 5.8 Histogram of $n_B = 10{,}000$ bootstrap estimates of the standard deviation of the logarithms of the 1933–1982 Ithaca January precipitation data. The sample standard deviation computed directly from the data is 0.537. The 95% confidence interval for the statistic, as estimated using the percentile method, is also shown.

INTERVALOS DE CONFIANZA

Bootstrapping paramétrico

(1) se simulan N series de igual duración que la serie original a partir de la distribución de probabilidad ajustada a la muestra original (e.g. GEV o GP)

(2) a cada serie se le ajusta la distribución de probabilidad objetivo (e.g. GEV o GP) y se calcula el cuantil de interés

(3) se estiman intervalos de confianza a partir de la muestra de cuantiles generada

Bootstrapping no paramétrico

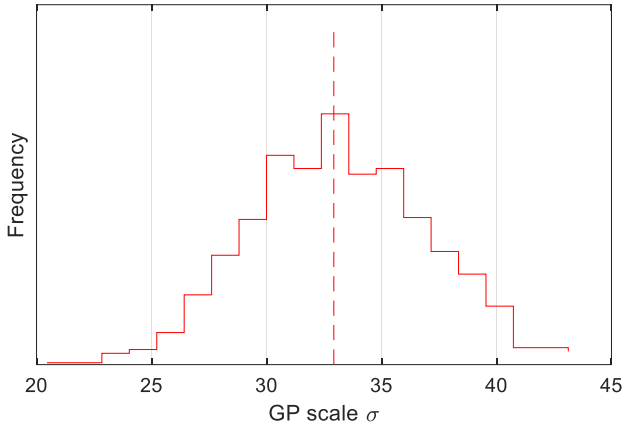
(1) se muestrean con repetición a partir de los datos originales N series de igual duración que la serie original

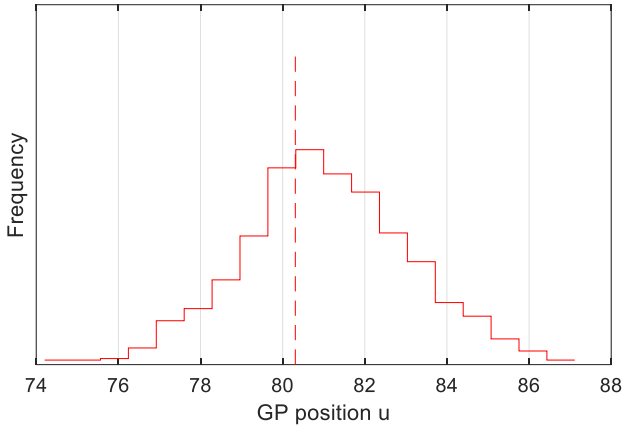
(2) a cada serie se le ajusta la distribución de probabilidad objetivo (e.g. GEV o GP) y se calcula el cuantil de interés

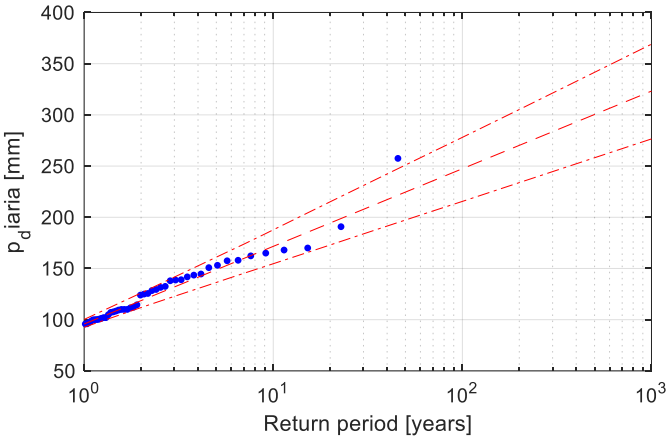
(3) se estiman intervalos de confianza a partir de la muestra de cuantiles generada

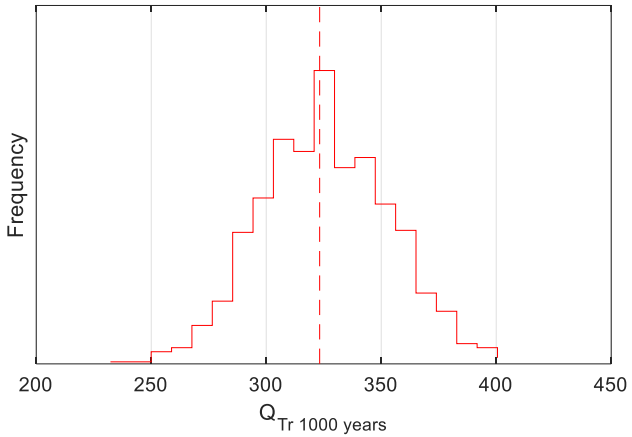
INTERVALOS DE CONFIANZA

Bootstrapping paramétrico









RESPECTO AL UMBRAL DE LA GP

- Físico vs. Estadístico (POT vs. GP)

- Métodos de selección del umbral → determina conjunto de datos.

Respecto al umbral de la GP

Primero seleccionar un umbral no demasiado alto pero que tenga sentido físico para determinar los picos a considerar (i.e. que los picos se consideren eventos extremos).

Luego determinar cuál es el umbral más adecuado para ajustar una GP (u otra distribución que se vaya a usar); es posible que este “umbral” estadístico sea mayor al anterior.

Existen varios métodos para seleccionar el “umbral estadístico”.

RESPECTO AL UMBRAL DE LA GP

Métodos basados en las propiedades de la GP:

Si una serie de datos proviene de una GP con umbral *u*, los datos mayores a *u*>u* también tendrán una distribución GP con:

(1) idéntico parámetro de forma

(2) con parámetro de escala lineal con umbral

$$\sigma^* = \sigma_u - \xi u$$

(3) Mean Residual Life (MRL) o Supervivencia

o Esperanza de las Excedencias

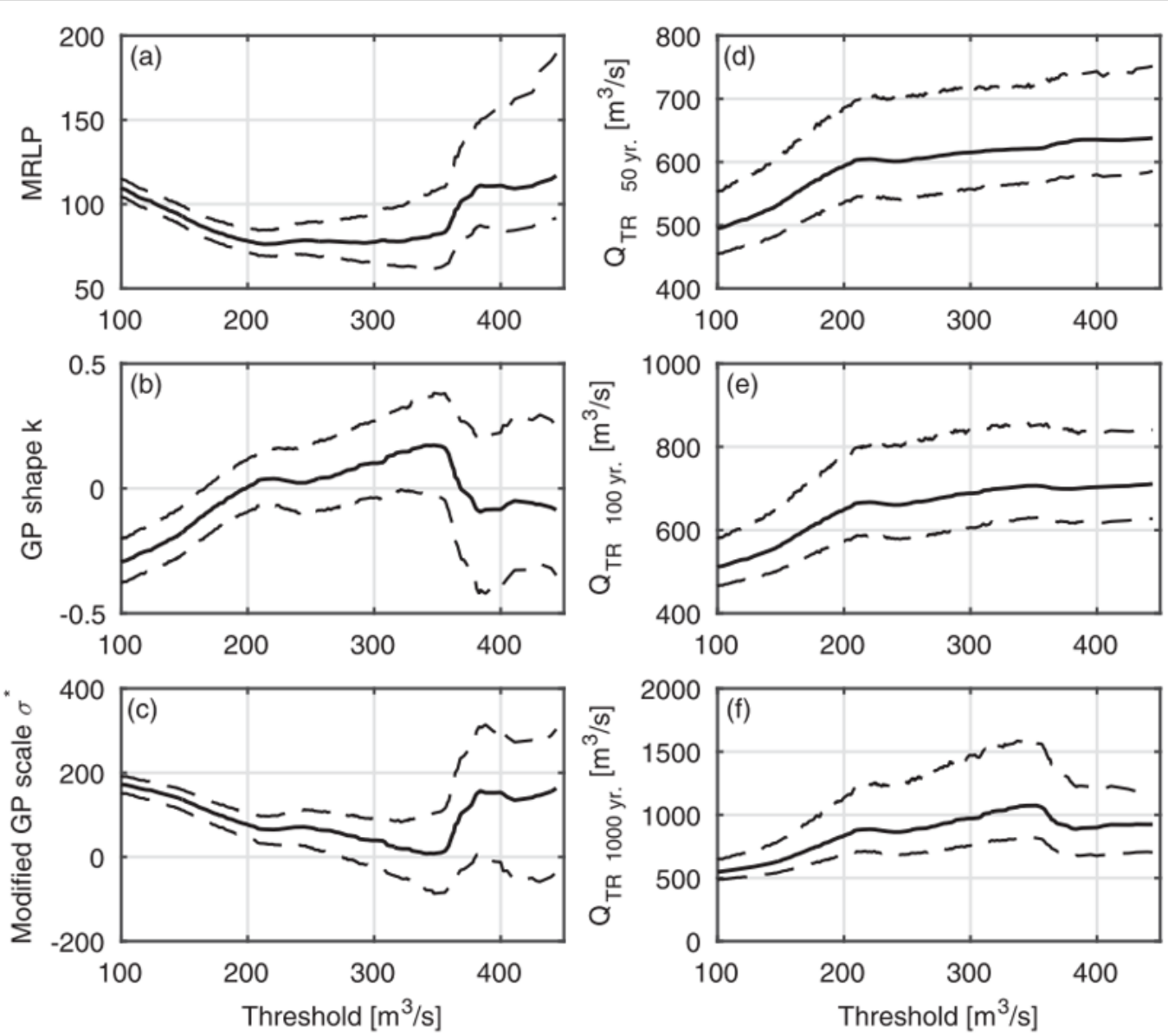
también es lineal con el umbral

RESPECTO AL UMBRAL DE LA GP

Recomendación:

Utilizar más de un método y, además, verificar que el número de eventos por año está en un rango razonable y la sensibilidad de los cuantiles objetivo (y sus intervalos de confianza) a la elección del umbral.

También graficar # de eventos !!



ANÁLISIS DE EXTREMOS

Edición 2024

Rafael Terra (en base a notas de Sebastián Solari)

Instituto de Mecánica de los Fluidos e Ingeniería Ambiental (IMFIA)
Facultad de Ingeniería, Universidad de la República, Uruguay

rterra@fing.edu.uy