

# Introducción a la Teoría de Valores Extremos

## 2. Métodos Estadísticos

Joaquín Ortega Sánchez

Centro de Investigación en Matemáticas, CIMAT  
Guanajuato, Gto., Mexico

Octavo Congreso Latinoamericano de Sociedades de  
Estadística  
Montevideo, Octubre 2008

# Outline

Introducción

Estimación desde la DGVE

El Período de Retorno

Selección de Familias

Estimadores de Momentos Pesados por Probabilidad

Ejemplo

# Outline

## Introducción

Estimación desde la DGVE

El Período de Retorno

Selección de Familias

Estimadores de Momentos Pesados por Probabilidad

Ejemplo

# Introducción

Recordemos el siguiente resultado:

## Teorema

*Si existen sucesiones de constantes  $a_n > 0$ ,  $b_n \in \mathbb{R}$  tales que*

$$P(M_n - b_n)/a_n \leq z) \rightarrow G(z) \quad (n \rightarrow \infty)$$

*donde  $G$  es no-degenerada, entonces*

$$G(z) = \exp \left\{ - \left[ 1 + \xi \left( \frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \right\},$$

*definida en  $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$  con  $\mu, \xi \in \mathbb{R}$ ,  $\sigma > 0$ .*

# Introducción

El teorema anterior sugiere que para valores grandes de  $n$  podemos usar la familia DGVE como aproximación del máximo para sucesiones de v.a.i. Aparentemente tenemos que resolver la dificultad de que no conocemos las constantes de normalización. Sin embargo, tenemos

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \approx G(z)$$

para  $n$  grande, o equivalentemente

$$P(M_n \leq z) \approx G\left(\frac{z - b_n}{a_n}\right) = G^*(z)$$

donde  $G^*$  es otra distribución en la familia DGVE.

# Introducción

Es decir, si podemos aproximar la distribución de  $(M_n - b_n)/a_n$  por un miembro de la familia DGVE, también podemos aproximar la distribución de  $M_n$  por un miembro diferente de la DGVE.

Como de cualquier manera tenemos que estimar los parámetros de la distribución, es irrelevante que los parámetros de  $G$  sean distintos a los de  $G^*$ .

# Introducción

## Ajuste de Máximos Anuales o Método de Bloques

En esta clase nos vamos a concentrar en este método de estimación, que puede ser descrito de la siguiente manera.

Consideramos una colección de datos que agrupamos en bloques disjuntos de datos consecutivos y de igual longitud. Si el parámetro es el tiempo, cada conjunto contiene la información correspondiente a un período fijo de tiempo, digamos un año. En cada caso se escoge el período para compensar las variaciones internas.

# Introducción

Por lo tanto los datos originales son

$$\mathbf{X}^{(1)} = (X_1^{(1)}, \dots, X_s^{(1)})$$

$$\mathbf{X}^{(2)} = (X_1^{(2)}, \dots, X_s^{(2)})$$

$$\vdots \qquad \qquad \qquad \vdots$$

$$\mathbf{X}^{(n)} = (X_1^{(n)}, \dots, X_s^{(n)})$$

donde suponemos que los vectores  $(\mathbf{X}^{(i)})_{i=1}^n$  son iid, pero las componentes de cada vector  $\mathbf{X}^{(i)}$  pueden ser dependientes. El intervalo de tiempo  $s$  se escoge de modo que estas condiciones se satisfagan. La muestra iid para  $G_\theta$  sobre la cual se hará la inferencia es

$$M_i = \max(X_1^{(i)}, \dots, X_s^{(i)}), \quad i = 1, \dots, n. \quad (1)$$

# Introducción

Por ejemplo, en el caso de los datos de Maiquetía sólo tenemos la información de los máximos anuales. En este caso  $s = 365$  y tenemos información desde 1951 hasta 1998 ( $n = 48$ ).

Para los datos de nieve en Carolina del Norte tenemos información de las nevadas ocurridas durante el mes de enero desde 1948 hasta 2000. En este caso  $s = 31$ ,  $n = 53$ .

# Introducción

Para hacer inferencia sobre la distribución de  $M_i$  tenemos dos posibilidades

1. Hacer inferencia a partir de la Distribución Generalizada de Valores Extremos (DGVE)
2. Identificar el dominio de atracción y luego hacer inferencia a partir de la Distribución de Valores Extremos (DVE) correspondiente.

Vamos a explorar ambos caminos a continuación

# Outline

Introducción

**Estimación desde la DGVE**

El Período de Retorno

Selección de Familias

Estimadores de Momentos Pesados por Probabilidad

Ejemplo

## Estimación desde la DGVE

Sea  $g_\theta$  la densidad de la DGVE  $G_\theta$ . La función de verosimilitud basada en el vector de datos  $\mathbf{X} = (X_1, \dots, X_n)$  está dada por

$$L(\theta; \mathbf{X}) = \prod_{i=1}^n g_\theta(X_i) \mathbf{1}_{\{1+\xi(X_i-\mu)/\sigma > 0\}}.$$

Llamemos  $\ell(\theta; \mathbf{X}) = \log L(\theta; \mathbf{X})$  a la log verosimilitud. El estimador de máxima verosimilitud (EMV) para  $\theta$  es

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} \ell(\theta; \mathbf{X}),$$

es decir  $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$  maximiza  $\ell(\theta; \mathbf{X})$  sobre un espacio de parámetros adecuado  $\Theta$ .

## Estimación desde la DGVE

La log-verosimilitud para los parámetros de la DGVE cuando  $\xi \neq 0$  es

$$\begin{aligned} \ell(\xi, \mu, \sigma; \mathbf{X}) &= -m \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^m \log \left(1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right) \\ &\quad - \sum_{i=1}^m \left(1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right)^{-1/\xi} \end{aligned}$$

En el caso  $\xi = 0$  tenemos

$$\ell((0, \mu, \sigma); \mathbf{X}) = -n \log \sigma - \sum_{i=1}^n \exp \left\{ -\frac{X_i - \mu}{\sigma} \right\} - \sum_{i=1}^n \frac{X_i - \mu}{\sigma}.$$

## Estimación desde la DGVE

Diferenciando esta última función con respecto a  $\mu$  y  $\sigma$  e igualando a 0 obtenemos las siguientes ecuaciones en el caso Gumbel,

$$0 = n - \sum_{i=1}^n \exp \left\{ - \frac{X_i - \mu}{\sigma} \right\}$$

$$0 = n + \sum_{i=1}^n \frac{X_i - \mu}{\sigma} \left( \exp \left\{ - \frac{X_i - \mu}{\sigma} \right\} - 1 \right).$$

No hay una solución explícita de estas ecuaciones. La situación para  $G_\xi$  con  $\xi \neq 0$  es aún más complicada, de modo que se requieren procedimientos numéricos.

## Estimación desde la DGVE

Una dificultad con el uso del método de máxima verosimilitud son las condiciones de regularidad que se requieren para que las propiedades asintóticas usuales valgan. Estas condiciones no son satisfechas por la DGVE porque los extremos de las distribuciones son una función de los valores de los parámetros. Esta violación de las condiciones de regularidad habituales implica que los métodos de máxima verosimilitud no pueden aplicarse automáticamente.

## Estimación desde la DGVE

R. L. Smith estudió este problema en detalle y demostró que

- Cuando  $\xi > -0.5$  los EMV tienen las propiedades asintóticas usuales,
- Cuando  $-1 < \xi < -0.5$  es posible obtener los estimadores pero no tienen las propiedades usuales,
- Si  $\xi < -1$  es posible que no puedan obtenerse los estimadores.

## Estimación desde la DGVE

Bajo las restricciones mencionadas anteriormente los estimadores  $(\hat{\xi}, \hat{\mu}, \hat{\sigma})$  tienen aproximadamente una distribución normal multivariada de media  $(\xi, \mu, \sigma)$  y matriz de covarianza igual al inverso de la matriz de información observada, evaluada en los estimadores de máxima verosimilitud. Aunque esta matriz se puede calcular analíticamente, es más sencillo calcularla numéricamente.

A partir de la normalidad asintótica de los estimadores se pueden obtener intervalos de confianza aproximados.

## Intervalos de Verosimilitud-Confianza

Para comparar la plausibilidad asociada a los diferentes valores de  $\theta$  podemos usar la función de verosimilitud relativa, que se define como

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta}_n)}.$$

Al conjunto  $\{\theta : R(\theta) \geq I\}$  con  $I \in (0, 1)$  lo llamaremos *región de verosimilitud*, mientras que  $I$  es el *nivel de verosimilitud*.

Cuando la dimensión de  $\theta$  es igual a 1 estas regiones reciben el nombre de *intervalos de verosimilitud*.

## Intervalos de Verosimilitud-Confianza

Las regiones de verosimilitud se refieren sólo a la verosimilitud o plausibilidad relativa de los distintos valores del parámetro  $\theta$ , y no a la incertidumbre del intervalo. En algunos casos es posible aproximar la probabilidad de que estas regiones contengan el verdadero valor del parámetro. Una forma de hacer esto es considerando la función

$$D_n(\theta) = 2(\ell(\hat{\theta}_n) - \ell(\theta)) = -2 \log(R(\theta)),$$

porque bajo condiciones de regularidad estándar <sup>1</sup>

$$D_n(\theta) \xrightarrow{d} \chi_k^2, \quad n \rightarrow \infty.$$

---

<sup>1</sup>Serfling, R.J. Approximation Theorems of Mathematical Statistics, Wiley, 1980, p. 155

## Intervalos de Verosimilitud-Confianza

En consecuencia, una región de confianza de nivel aproximado  $(1 - \alpha)$  está dada por

$$C_\alpha = \{\theta : D(\theta) \leq c_\alpha\}, \quad (2)$$

donde  $c_\alpha$  es el cuantil  $1 - \alpha$  de la distribución  $\chi_d^2$ . Escribiendo (4) en términos de la verosimilitud relativa obtenemos

$$C_\alpha \{\theta : R(\theta) \geq \exp(-\frac{1}{2}c_\alpha)\}.$$

bajo las condiciones del resultado anterior tenemos que una región de verosimilitud con nivel  $\exp\{-c_\alpha/2\}$  tiene una probabilidad de cobertura aproximada de  $(1 - \alpha)$ .

## Intervalos de Verosimilitud-Confianza

En el siguiente cuadro presentamos algunos ejemplos para el caso  $d = 1$ .

Nivel de Verosimilitud	Nivel de Confianza
0.0362	0.99
0.1465	0.95
0.2585	0.90
0.7965	0.50

Tabla 2.1 Correspondencia entre niveles de verosimilitud y confianza usando la aproximación  $\chi^2$ .

## Verosimilitud Perfil

Consideremos un espacio de parámetros  $\Theta$  de dimensión  $d$  con elementos  $\theta = (\theta_1, \dots, \theta_d)$ . Supongamos que podemos dividir a  $\theta$  en dos componentes  $(\theta^{(1)}, \theta^{(2)})$  y estamos interesados únicamente en  $\theta^{(1)}$ , de modo que  $\theta^{(2)}$  son parámetros de estorbo. Para estimar  $\theta^{(1)}$  podemos usar la verosimilitud perfil, que se define como

$$L_p(\theta^{(1)}) = \max_{\theta^{(2)}|\theta^{(1)}} L(\theta^{(1)}, \theta^{(2)}),$$

donde  $L$  es la función de verosimilitud. La verosimilitud perfil se obtiene maximizando la función de verosimilitud evaluada en los elementos de  $\Theta$  con  $\theta^{(1)}$  fijo.

## Verosimilitud Perfil

Denotando por  $\hat{\theta}_n$  al EMV, podemos definir la verosimilitud relativa perfil para  $\theta^{(1)}$  como

$$R_p(\theta^{(1)}) = \frac{L_p(\theta^{(1)})}{L(\hat{\theta}_n)}.$$

Con esta función podemos construir los conjuntos  $\{\theta : R_p(\theta^{(1)}) \geq I\}$ , que son las regiones de verosimilitud para  $\theta^{(1)}$  de nivel  $I$ , a los cuales podemos asociar un nivel de confianza de manera similar a como hicimos anteriormente.

## Verosimilitud Perfil

La evaluación numérica de la verosimilitud perfil para cualquiera de los parámetros  $\xi$ ,  $\mu$  o  $\sigma$  es sencilla. Por ejemplo, para obtener la verosimilitud perfil de  $\xi$ , fijamos  $\xi = \xi_0$  y maximizamos la log-verosimilitud con respecto de los parámetros restantes. Repetimos esto para un rango de valores de  $\xi_0$ . Los valores correspondientes de la log-verosimilitud constituyen la log-verosimilitud perfil para  $\xi$ , a partir de la cual podemos obtener intervalos de confianza aproximados. A diferencia de los intervalos que se obtienen a partir del método delta, estos intervalos son generalmente asimétricos, reflejando la asimetría natural de la función de verosimilitud en estos casos.

# Outline

Introducción

Estimación desde la DGVE

**El Período de Retorno**

Selección de Familias

Estimadores de Momentos Pesados por Probabilidad

Ejemplo

## El Período de Retorno

Sea  $(X_j)$  una sucesión de v.a.i.i.d. con fd continua  $F$  y  $u$  un umbral dado. Consideramos la sucesión  $(\mathbf{1}_{\{X_j > u\}})$  de v.a.i.i.d. de Bernoulli con probabilidad de éxito  $p = \bar{F}(u)$ . En consecuencia, el instante del primer éxito

$$L(u) = \min\{i \geq 1 : X_i > u\},$$

es decir, el instante de la primera excedencia del umbral  $u$ , es una v.a. geométrica con distribución

$$P(L(u) = k) = (1 - p)^{k-1} p, \quad k = 1, 2, \dots$$

# El Período de Retorno

Observamos que las v.a.i.i.d.

$$L_1(u) = L(u), \quad L_{n+1}(u) = \min\{i > L_n(u) : X_i > u\}, \quad n \geq 1,$$

describen los intervalos de tiempo entre excedencias sucesivas de  $u$  por  $(X_n)$ . El período de retorno de los eventos  $\{X_i > u\}$  se define como  $E[L(u)] = p^{-1} = (\bar{F}(u))^{-1}$ , que crece a  $\infty$  cuando  $u \rightarrow \omega_F$ . Para facilitar la notación tomamos  $\omega_F = \infty$ . Todas las preguntas importantes relativas al período de retorno pueden responderse a través de las propiedades de la distribución geométrica.

## El Período de Retorno

Por ejemplo, si deseamos hallar el nivel correspondiente a un período de retorno de 100 años queremos que  $p^{-1} = 100$ , o sea que  $p = 0.01$  y en consecuencia  $F(u) = 0.99$ . Esto muestra que el nivel de retorno correspondiente a un período de 100 años es el cuantil 99 de la distribución  $F$ . Por esto, el problema de estimación de cuantiles, especialmente de cuantiles altos de las DVE y DGVE, tienen un papel central en el análisis de valores extremos.

# El Período de Retorno

Definimos

$$d_k = P(L(u) \leq k) = p \sum_{i=1}^k (1-p)^{i-1} = 1 - (1-p)^k, \quad k \in \mathbb{N}.$$

$d_k$  es la probabilidad de que haya al menos una excedencia de  $u$  antes del tiempo  $k$  (o en las próximas  $k$  observaciones). Esto nos da una relación 1-1 entre  $d_k$  y el período de retorno  $p^{-1}$ . La probabilidad de que haya una excedencia de  $u$  antes del período de retorno es

$$P(L(u) \leq E L(u)) = P(L(u) \leq [1/p]) = 1 - (1-p)^{[1/p]},$$

donde  $[x]$  denota la parte entera de  $x$ .

# El Período de Retorno

Para umbrales altos (para  $u \uparrow \infty$  y por lo tanto  $p \downarrow 0$ ) obtenemos

$$\begin{aligned}\lim_{u \uparrow \infty} P(L(u) \leq E L(u)) &= \lim_{p \downarrow 0} (1 - (1 - p)^{[1/p]}) \\ &= 1 - e^{-1} = 0.63212.\end{aligned}$$

Esto muestra que para umbrales altos la media del período de retorno  $L(u)$  es mayor que su mediana.

## El Período de Retorno

Se quiere asegurar una estructura sobre la base de que durará al menos 50 años con una probabilidad de falla que no supere el 10%. ¿Qué implica esto para el período de retorno? Con la notación anterior queremos

$$P(L(u) \leq 50) \leq 0.1.$$

Hemos supuesto tácitamente que el fallo de la estructura en cada año  $i$  puede modelarse a través de un evento  $\{X_i > u\}$ , donde  $X_i$  es, por ejemplo, un componente crítico de la estructura. Suponemos que las  $X_i$  son iid. y obtenemos

$$P(L(u) \leq 50) = 1 - (1 - p)^{50} = 0.1$$

de donde  $p = 0.002105$ , es decir  $E L(u) = 475$ . Hablamos en este caso de un evento de 475 años.

## El Período de Retorno

¿Qué implicación tiene un evento de  $t$ -años sobre el umbral correspondiente? Por definición para el umbral asociado  $u_t$ ,

$$t = E L(u_t) = \frac{1}{\bar{F}(u_t)}$$

y por lo tanto

$$u_t = Q(1 - t^{-1}) = F^{\leftarrow}(1 - t^{-1}).$$

donde  $Q$  es la función de cuantiles. En este ejemplo,  $u_{475} = Q(0.9979)$ . Esto nos lleva de nuevo al problema crucial de estimar cuantiles altos.

## Inferencia para el Período de Retorno

Regresemos al esquema de estimación clásico en el cual tenemos una muestra aleatoria simple  $X_1, \dots, X_n$  de  $G_\theta$ . En esta situación es fácil obtener un estimador de los cuantiles. Dado cualquier  $p \in (0, 1)$  definimos el  $p$ -cuantil por  $Q(p) = G_\theta^{\leftarrow}(p)$ . Un estimador natural para  $Q(p)$  basado en  $X_1, \dots, X_n$  es

$$\widehat{Q}(p) = G_{\widehat{\theta}}^{\leftarrow}(p),$$

donde  $\widehat{\theta}$  es un estimador de  $\theta$ . Teniendo en cuenta la definición de  $G_\theta$ , la función de cuantiles es

$$Q(p) = \begin{cases} \mu - \frac{\sigma}{\xi} (1 - (-\log p)^{-\xi}), & \text{para } \xi \neq 0, \\ \mu - \sigma \log(-\log p), & \text{para } \xi = 0. \end{cases}$$

## Inferencia para el Período de Retorno

Para  $0 < p < 1$  llamemos  $z_p$  el nivel de retorno asociado a un período de retorno  $1/p$ , es decir,  $G(z_p) = 1 - p$  y llamemos  $y_p = -\log(1 - p)$ . Con esta notación

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}(1 - y_p^{-\xi}), & \text{para } \xi \neq 0, \\ \mu - \sigma \log y_p, & \text{para } \xi = 0. \end{cases}$$

Sustituyendo los EMV de los parámetros de la DGVE obtenemos el EMV para  $z_p$ :

$$\hat{z}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}(1 - y_p^{-\hat{\xi}}), & \text{para } \xi \neq 0, \\ \hat{\mu} - \hat{\sigma} \log y_p, & \text{para } \xi = 0. \end{cases}$$

# Inferencia para el Período de Retorno

Para obtener intervalos de confianza para estos valores estimados podemos reparametrizar el modelo DGVE de modo que  $z_p$  sea uno de los parámetros del modelo. Para la reparametrización tenemos:

$$\mu = z_p + \frac{\sigma}{\xi} (1 - (-\log(1 - p))^{-\xi}),$$

y reemplazando esta expresión por  $\mu$  en la log-verosimilitud permite obtener la log-verosimilitud perfil para  $z_p$ , a partir de la cual podemos obtener intervalos de verosimilitud-confianza.

## Inferencia para el Período de Retorno

En cuanto al método delta tenemos que

$$\text{Var}(\hat{z}_p) \approx \nabla z_p^t V \nabla z_p, \quad (3)$$

donde  $V$  es la matriz de covarianza de  $(\hat{\xi}, \hat{\mu}, \hat{\sigma})$  y

$$\begin{aligned} \nabla z_p^t &= \left( \frac{\partial z_p}{\partial \xi}, \frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma} \right) \\ &= \left( \frac{\sigma}{\xi^2} (1 - y_p^{-\xi}) - \frac{\sigma}{\xi} y_p^{-\xi} \log y_p, 1, -\frac{1}{\xi} (1 - y_p^{-\xi}) \right) \end{aligned}$$

evaluado en  $(\hat{\xi}, \hat{\mu}, \hat{\sigma})$ . Usando la normalidad asintótica del estimador se pueden obtener intervalos de confianza aproximados.

## Inferencia para el Período de Retorno

Si  $\hat{\xi} < 0$  es posible hacer inferencia sobre el extremo derecho de la distribución, que es en efecto el valor de  $z_p$  correspondiente a  $p = 0$ . El estimador de máxima verosimilitud es

$$z_0 = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}$$

y (3) vale con

$$\nabla z_0^t = (\sigma \xi^{-2}, 1, -\xi^{-1})$$

La estimación correspondiente para la cola de la distribución  $\overline{G}_\theta(x)$ , para  $x$  en el dominio apropiado, corresponde a

$$\overline{G}_{\hat{\theta}}(x) = 1 - \exp \left\{ - \left( 1 + \hat{\xi} \frac{x - \hat{\mu}}{\hat{\sigma}} \right)^{-1/\hat{\xi}} \right\},$$

donde  $\hat{\theta} = (\hat{\xi}, \hat{\mu}, \hat{\sigma})$  es estimado por EMV.

# Outline

Introducción

Estimación desde la DGVE

El Período de Retorno

**Selección de Familias**

Estimadores de Momentos Pesados por Probabilidad

Ejemplo

## Selección de Familias

En este enfoque se utiliza la DGVE para obtener la verosimilitud perfil del parámetro de forma  $\xi$ , y a partir de ella se seleccionan 1 ó 2 modelos de las 3 DVE, según la forma de la verosimilitud perfil y tomando en cuenta la plausibilidad del valor 0 para  $\xi$ .

- Si 0 es poco plausible ( $R_p(0) < 0.15$ ) entonces la curva se concentra a uno de los dos lados del 0 y sólo una de las familias de DVE (Weibull o Fréchet) es razonable como modelo.
- Si 0 es un valor plausible para  $\xi$  ( $R_p(0) > 0.15$ ) entonces se seleccionan dos modelos: la familia que corresponde al estimador de máxima verosimilitud  $\hat{\xi}$  (Weibull si  $\hat{\xi} < 0$ , Fréchet si  $\hat{\xi} > 0$ ) y el modelo Gumbel.

## Selección de Familias

A pesar de que en el último caso las tres familias pueden tener alta plausibilidad, sólo seleccionamos dos porque la tercera siempre va a ser prácticamente indistinguible del modelo Gumbel. Para explicar esto con más detalle, supongamos que  $\hat{\xi} > 0$ , de modo que la familia más plausible es Fréchet. El mejor modelo Weibull para este caso, según la verosimilitud perfil, corresponde a un parámetro  $\xi$  que es prácticamente igual a 0, ya que las tres familias forman una colección continua de modelos y como la verosimilitud crece con  $\xi$ , el mayor valor de  $\xi$  con la restricción  $\xi < 0$  es el supremo de este conjunto, que es 0.

## Selección de Familias

Una vez seleccionadas una o dos familias, se procede a hacer el ajuste de los modelos correspondientes a cada caso y teniendo en cuenta la calidad del modelo y las razones de contexto que apoyen a uno u otro se selecciona uno de ellos. Si bien en general el estimador de máxima verosimilitud para el parámetro de forma  $\xi$  no cambia (salvo cuando seleccionamos el modelo Gumbel) los intervalos de verosimilitud-confianza si lo hacen, y con frecuencia (aunque no siempre) resultan más estrechos cuando se estima a partir de una DVE.

## Selección de Familias

Esto es particularmente cierto para la estimación de cuantiles. Como la familia DGVE incluye a todos los modelos, sean o no plausibles, la verosimilitud perfil para un nivel de retorno dado bajo este modelo incluye puntos que corresponden a modelos que no son plausibles de acuerdo a los datos. Al restringirse a una subfamilia DVE, muchos de estos modelos implausibles son eliminados y la curva refleja de manera más apropiada la información disponible.

## Selección de Familias

Otra observación importante es que la verosimilitud perfil con frecuencia es asimétrica, en algunos casos extremadamente asimétrica. Esto produce intervalos de verosimilitud confianza que son también asimétricos, reflejando así la información que proveen los datos. Esto contrasta con los intervalos que se obtienen por el método delta, que son siempre simétricos por provenir de una aproximación Gaussiana.

Finalmente, observamos que cuando la plausibilidad de 0 como valor de  $\xi$  es baja o nula, los resultados que se obtienen estimando desde la DGVE prácticamente coinciden con los que se obtienen seleccionando una familia DVE y estimando desde esta familia.

## Comparación de Intervalos

Tenemos tres maneras de obtener intervalos de confianza para cuantiles de interés:

- Estimar los parámetros por MV desde la DGVE y obtener el intervalo de confianza asintótico usando la normalidad asintótica del EMV.
- Estimar los parámetros por MV desde la DGVE y obtener el intervalo de verosimilitud-confianza usando la verosimilitud perfil para el cuantil.
- Estimar los parámetros por MV desde un submodelo (DVE) y obtener el intervalo de verosimilitud-confianza usando la verosimilitud perfil para el cuantil.

A continuación presentamos un estudio de simulación para comparar la calidad de estos tres procedimientos.

## Comparación de Intervalos

- Suponemos que las muestras provienen de una DVE (y no de una distribución en el dominio de atracción)
- Las muestras se simularon usando  $a = 1$ ,  $b = 1$  y

$$c \in \{-0.5, -0.4, -0.3, -0.2, -0.1, -0.05, 0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\},$$

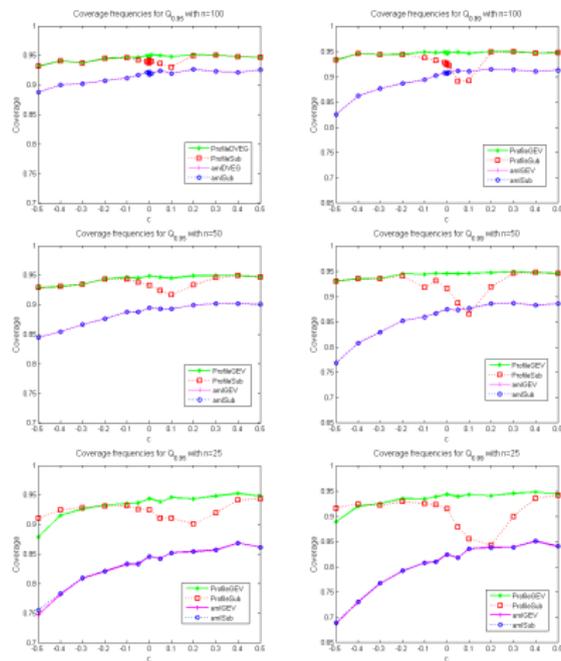
- Muestras de tamaño 25, 50 y 100. En este último caso consideramos también los siguientes valores de  $c$ :  $\pm 0.01$ ,  $\pm 0.001$ .
- En cada caso se simularon 10000 muestras usando Matlab para los cuantiles  $Q_{95}$  y  $Q_{99}$ .

# Comparación de Intervalos

## Analizamos

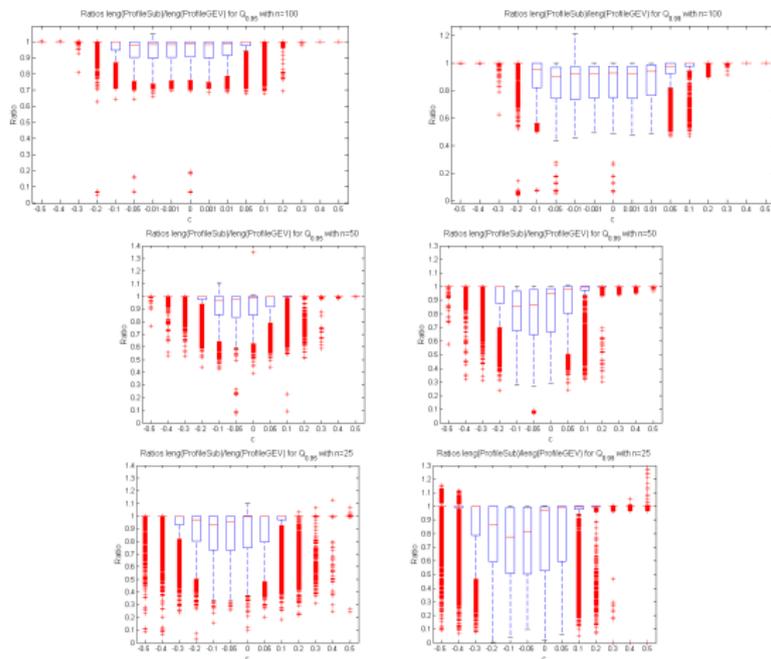
- Coberturas de los intervalos de confianza en cada caso.
- Longitud relativa de los intervalos obtenidos desde la DGVE y desde el submodelo.
- Longitud (en términos absolutos) para los intervalos de la DGVE

# Comparación de Intervalos



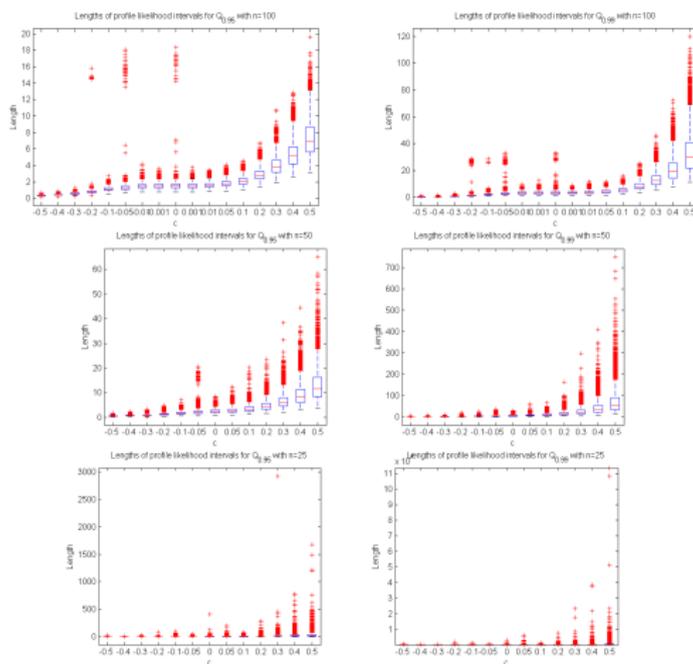
**Figure:** Coverage frequencies. The left column corresponds to  $Q_{95}$ , the right to  $Q_{99}$ . The first row corresponds to a sample size of 100, the middle row to sample size 50 and the bottom row to sample size 25.

# Comparación de Intervalos



**Figure:** Ratio of length of likelihood-confidence intervals for  $Q_{95}$  (left) and  $Q_{99}$  (right) for the submodel over length of intervals for the GEV, sample sizes 100 (top), 50 (middle) and 25 (bottom).

# Comparación de Intervalos



**Figure:** Length of profile likelihood-confidence intervals for  $Q_{95}$  (left) and  $Q_{99}$  (right), sample sizes  $n = 100$  (top),  $n = 50$ , middle and  $n = 25$  (bottom) for the GEV.

## Comparación de Intervalos

- Los intervalos asintóticos de MV tienen coberturas pobres. Los correspondientes a la DGVE muestran buen nivel y mucha estabilidad. Los de la DVE tienen niveles razonables o buenos pero no son tan estables.
- Los intervalos que provienen de la DVE tienen longitudes iguales o menores a los de la DGVE, y en algunos casos pueden ser considerablemente menores (especialmente alrededor de  $c = 0$ ).
- Para el mismo cuantil y tamaño de muestra, la longitud de los intervalos crece con  $c$ .

# Outline

Introducción

Estimación desde la DGVE

El Período de Retorno

Selección de Familias

**Estimadores de Momentos Pesados por Probabilidad**

Ejemplo

# Estimadores de Momentos Pesados por Probabilidad

El método de momentos para estimación de parámetros consiste en igualar los momentos del modelo basado en  $G_\theta$  con los correspondientes momentos empíricos basados en los datos. Sin embargo, en el caso de valores extremos estos estimadores son poco confiables. Resulta más útil la clase de los estimadores de momentos pesados por probabilidades, que se definen de la siguiente manera. Para una v.a.  $X \sim F$ ,

$$M_{p,r,s} = E[X^p(F(X))^r(1 - F(X))^s]$$

Si la función de cuantiles  $Q$  puede escribirse de manera explícita entonces

$$M_{p,r,s} = \int_0^1 Q(y)y^r(1 - y)^s dy$$

# Estimadores de Momentos Pesados por Probabilidad

Para nuestros propósitos definimos

$$w_r(\theta) = E(XG_\theta^r(X)) = M_{1,r,0}, \quad r \in \mathbb{N}_0, \quad (4)$$

donde  $G_\theta$  es la DGVE y  $X \sim G_\theta$  con  $\theta = (\xi, \mu, \sigma)$ . Recordemos que para  $\xi \geq 1$ ,  $\overline{G}_\theta$  es de variación regular con índice  $1/\xi$  y en consecuencia  $w_0 = \infty$ . Por lo tanto nos restringimos al caso  $\xi < 1$ . Definimos el análogo empírico de (4)

$$\widehat{w}_r(\theta) = \int_{-\infty}^{\infty} x G_\theta^r(x) dF_n(x), \quad r \in \mathbb{N}_0, \quad (5)$$

donde  $F_n$  es la fde correspondiente a los datos  $X_1, \dots, X_n$ .

# Estimadores de Momentos Pesados por Probabilidad

Para estimar  $\theta$  resolvemos las ecuaciones

$$w_r(\theta) = \widehat{w}_r(\theta), \quad r = 0, 1, 2.$$

A partir de (5) obtenemos

$$\widehat{w}_r(\theta) = \frac{1}{n} \sum_{j=1}^n X_{j:n} G_{\theta}^r(X_{j:n}), \quad r = 0, 1, 2. \quad (6)$$

Recordemos que

$$(G_{\theta}(X_{1:n}), \dots, G_{\theta}(X_{n:n})) \stackrel{d}{=} (U_{1:n}, \dots, U_{n:n}),$$

donde  $U_{1:n} \leq \dots \leq U_{n:n}$  son los estadísticos de orden de una sucesión iid  $U_1, \dots, U_n$  con distribución uniforme en  $(0, 1)$ . Con esta interpretación (6) se puede escribir como

$$\widehat{w}_r(\theta) = \frac{1}{n} \sum_{j=1}^n X_{j:n} U_{j:n}^r, \quad r = 0, 1, 2. \quad (7)$$

# Estimadores de Momentos Pesados por Probabilidad

Es claro que para  $r = 0$  el lado derecho es  $\bar{X}_n$ , la media muestral. Para calcular  $w_r(\theta)$  para  $r$  general observamos que

$$w_r(\theta) = \int_{-\infty}^{\infty} x G_{\theta}^r(x) dG_{\theta}(x) = \int_0^1 G_{\theta}^{\leftarrow}(y) y^r dy,$$

donde, para  $0 < y < 1$ ,

$$G_{\theta}^{\leftarrow}(y) = \begin{cases} \mu - \frac{\sigma}{\xi} (1 - (-\log y)^{-\xi}) & \text{si } \xi \neq 0, \\ \mu - \sigma \log(-\log y) & \text{si } \xi = 0. \end{cases}$$

# Estimadores de Momentos Pesados por Probabilidad

Para  $\xi < 1$  y  $\xi \neq 0$ , luego de algunos cálculos,

$$w_r(\theta) = \frac{1}{r+1} \left( \mu - \frac{\sigma}{\xi} (1 - \Gamma(1 - \xi)(1+r)^\xi) \right), \quad (8)$$

donde  $\Gamma$  es la función Gamma:  $\Gamma(t) = \int_0^\infty e^{-u} u^{t-1} du$ ,  $t > 0$ .  
Combinando (7) y (8) obtenemos un estimador de momentos pesado por probabilidades  $\hat{\theta}_n^{(1)}$ .

# Estimadores de Momentos Pesados por Probabilidad

A partir de (8) obtenemos

$$w_0(\theta) = \mu - \frac{\sigma}{\xi}(1 - \Gamma(1 - \xi)),$$

$$2w_1(\theta) - w_0(\theta) = \frac{\sigma}{\xi}\Gamma(1 - \xi)(2^\xi - 1),$$

$$3w_2(\theta) - w_0(\theta) = \frac{\sigma}{\xi}\Gamma(1 - \xi)(3^\xi - 1),$$

y en consecuencia

$$\frac{3w_2(\theta) - w_0(\theta)}{2w_1(\theta) - w_0(\theta)} = \frac{3^\xi - 1}{2^\xi - 1}.$$

# Estimadores de Momentos Pesados por Probabilidad

Usando los estimadores anteriores en esta ecuación obtenemos un estimador  $\hat{\xi}$  de  $\xi$ . Dado  $\hat{\xi}$ , los parámetros  $\mu$  y  $\sigma$  se estiman por

$$\hat{\sigma} = \frac{(2\hat{w}_1 - \hat{w}_0)\hat{\xi}}{\Gamma(1 - \hat{\xi})(2^{\hat{\xi}} - 1)},$$
$$\hat{\mu} = \hat{w}_0 + \frac{\hat{\sigma}}{\hat{\xi}}(1 - \Gamma(1 - \hat{\xi})),$$

donde  $\hat{w}_0$ ,  $\hat{w}_1$ ,  $\hat{w}_2$  son los momentos empíricos que hemos considerado anteriormente. El caso  $\xi = 0$  también puede ser considerado por este método.

# Estimadores de Momentos Pesados por Probabilidad

Es posible obtener otros estimadores reemplazando  $U_{j:n}^r$  en (7) por otro estadístico, como por ejemplo

$$E(U_{j:n}^r) = \frac{(n-j)(n-j-1)\cdots(n-j-r+1)}{(n-1)(n-2)\cdots(n-r)}, \quad r = 1, 2.$$

# Outline

Introducción

Estimación desde la DGVE

El Período de Retorno

Selección de Familias

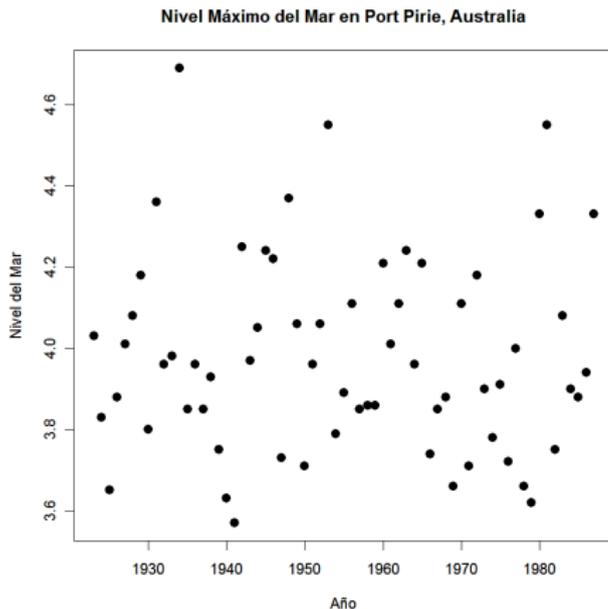
Estimadores de Momentos Pesados por Probabilidad

Ejemplo

# Ejemplo

## Niveles máximos del mar en Port Pirie

La gráfica muestra los máximos anuales del nivel del mar en Port Pirie, Australia.



## Ejemplo

El objetivo es obtener un estimador del máximo nivel que puede ocurrir en los próximos 10 ó 100 años.

A partir de la gráfica parece razonable asumir que el patrón de los niveles ha permanecido estable durante el período de observación, de modo que modelamos los datos como observaciones independientes de la DGVE.

Maximizando la log-verosimilitud de la DGVE obtenemos los estimadores

$$(\hat{\mu}, \hat{\sigma}, \hat{\xi}) = (3.87, 0.198, -0.05),$$

y el valor de la log-verosimilitud es 4.34.

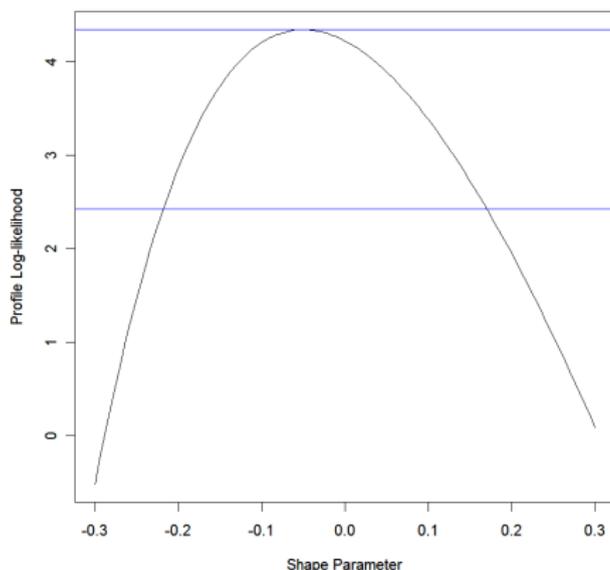
## Ejemplo

Los errores estándar son 0.028, 0.020 y 0.098 para  $\hat{\mu}$ ,  $\hat{\sigma}$  y  $\hat{\xi}$ , respectivamente. Combinando las estimaciones y los errores, los intervalos de confianza aproximados al 95% son [3.82, 3.93] para  $\mu$ , [0.158, 0.238] para  $\sigma$  y [-0.242, 0.142] para  $\xi$

Aún cuando el estimador de  $\hat{\xi}$  tiene un valor negativo, que corresponde a una distribución de Weibull, el intervalo de confianza abarca una porción importante de valores positivos, por lo que la evidencia no es determinante.

# Ejemplo

El intervalo de confianza obtenido a partir de la verosimilitud perfil es  $[-0.21, 0.17]$

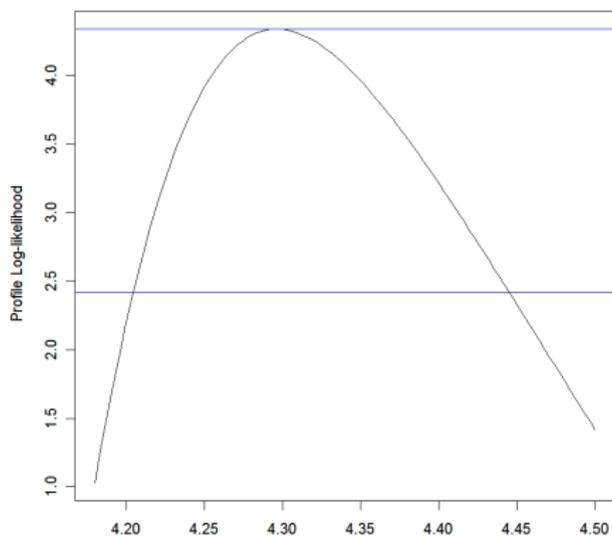


## Ejemplo

Estimación del nivel de retorno a 10 años: 4.30, con varianza 0.00303. Intervalo de confianza asintótico es

$$4.3 \pm 1.96 \times \sqrt{0.00303} = [4.19, 4.45]$$

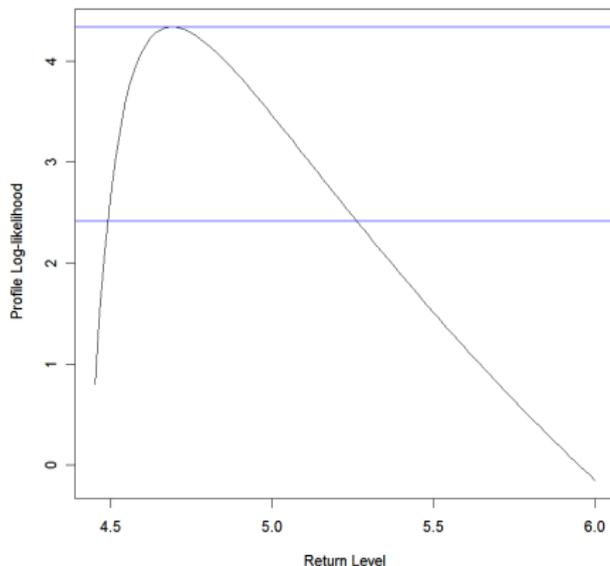
Intervalo usando la verosimilitud perfil: [4.21, 4.45]



## Ejemplo

Estimación del nivel de retorno a 100 años: 4.69 y el intervalo de confianza asintótico es [4.38, 5.00]

Intervalo usando la verosimilitud perfil: [4.50, 5.27]



# Ejemplo

## Gráficas de diagnóstico

