Teoría de valores extremos: Una introduccón*

Fernando Velasco Luna †
Laboratorio de Investigación y Asesoría Estadística, LINAE
Facultad de Estadística e Informática
Universidad Veracruzana

Sergio Hernández González ‡
Laboratorio de Investigación y Asesoría Estadística, LINAE
Facultad de Estadística e Informática
Universidad Veracruzana

La teoría de valores extremos está relacionada con aspectos probabilíticos y estadísticos relacionados con valores muy altos o muy bajos en una sucesión de variables aleatorias. La teoría de valores extremos tiene aplicaciones en una gran variedad de áreas. Este artículo presenta una introducción a la teoría de valores extremos y sus áreas de aplicación.

Extreme value theory is concerned with probabilistic and statistical questions related to very high or very low values in sequences of random variables. Extreme value theory has applications in a variety of areas. This paper presents an introduction to extreme value theory and its areas of applications.

Palabras clave: Teoría de valores Extremos, Leyes de Límites, Máximo, Mínimo.
Keywords: Extreme value Theory, Limit Laws, Maximum, Minimum.

1. Introducción

La teoría de valores extremos se ha desarrollado rápidamente en las últimas dos décadas tanto desde el punto de vista metodológico como del de las aplicaciones. La mayoría de los estudios estadísticos tratan de la modelación del promedio de la distribución de la variable de interés, dicho promedio se estima a partir de la media muestral, por otra parte el teorema del límite central proporciona un valioso resultado relacionado con el comportamiento asintótico de la media muestral. En la teoría de los valores extremos el interés principal no está en el promedio, sino en los valores más bajos o más altos de la variable bajo estudio, es decir, el interés está en los eventos asociados a la cola de la distribución. Por ejemplo en estudios de oceanografía, es necesario estudiar el comportamiento de corrientes marinas extremas, en estadística ambiental es necesario analizar niveles altos de ozono en determinada región, en climatología es necesario conocer el comportamiento de velocidades extremas de huracanes, en ingenieria, el aumento del flujo de un río, en finanzas, un gran decremento o aumento del valor de una acción en el mercado, valores máximos

*Recibido el 15 de enero de 2007 y aceptado el 21 de febrero de 2007

†**Dirección postal**: Av. Xalapa Esq. Manuel Avila Camacho s/n. C.P. 91020. Xalapa Veracruz, México. **Correo electrónico**: `fvelasco@uv.mx`

‡**Dirección postal**: Av. Xalapa Esq. Manuel Avila Camacho s/n. C.P. 91020. Xalapa Veracruz, México. **Correo electrónico**: `sehernandez@uv.mx`

o mínimos en un portafolio de inversión, etc. Cualquier tipo de pregunta relacionada con las situaciones antes mencionadas tendrá que involucrar un manejo adecuado de la información con que se cuente y de una aplicación cuidadosa de herramientas analíticas para llevar acabo el proceso de análisis e inferencia. Este comportamiento de los valores más altos o más bajos es importante en varios campos los cuales incluyen, la ingeniería [7], la oceanografía [1, 21], el medioambiente [17, 20, 22], la Hidrología, [16], la climatología [5, 11, 23], las finanzas [10, 13], entre otros. En la actualidad se cuenta con software estadístico para el análisis de valores extremos. Un tratamiento y abundantes referencias acerca de la teoría de valores extremos se puede encontrar en [3, 6].

Un enfoque para la modelación de valores extremos es a partir de la distribución de Valores Extremos Generalizada. Esta distribución se ajusta a los valores máximos o mínimos de datos. Otro enfoque para el análisis de valores extremos es a partir del análisis de excedentes sobre umbrales. En este trabajo se trata el primer enfoque en el caso univariado. Cabe mencionar que la teoría para modelos para extremos multivariados se ha desarrollado con los trabajos de [4, 7, 18, 19]. Se tiene que mencionar que el análisis de datos extremos también ha sido llevado a cabo desde el punto de vista Bayesiano, ver, [2, 8, 9].

2. Preliminares

Sean $X_1, X_2, ..., X_n$ variables aleatorias independientes con función de distribución acumulativa común $F(x)$, es decir, $F(x) = P(X_i \leq x)$. La teoría de valores extremos estudia el comportamiento de los valores máximo y mínimo de $X_1, X_2, ..., X_n$. Definanse $Y_n = \max\{X_1, X_2, ..., X_n\}$, y $Y_1 = \min\{X_1, X_2, ..., X_n\}$. Los extremos son definidos como el máximo y el mínimo de las n variables aleatorias. ¿Cuál es la distribución de los valores extremos, es decir, de los valores máximo y mínimo Y_1 y Y_n?.

Teorema 1. Sean $X_1, X_2, ..., X_n$ variables aleatorias independientes con función de distribución común $F(x)$.

Definanse $Y_n = \max\{X_1, X_2, ..., X_n\}$, y $Y_1 = \min\{X_1, X_2, ..., X_n\}$. Entonces las distribuciones de Y_1 y Y_n están dadas por

$$L_n(x) = 1 - (1 - F(x))^n \tag{1}$$

y

$$H_n(x) = (F(x))^n \tag{2}$$

respectivamente.

Prueba. La distribución del máximo es $H_n(x) = (F(x))^n$, como se demuestra a continuación,

$$\begin{aligned} H_n(x) &= P(Y_n \leq x) \\ &= P\{X_1 \leq x, X_2 \leq x, ..., X_n \leq x\} \\ &= P\{X_1 \leq x\} P\{X_2 \leq x\} \cdots P\{X_n \leq x\} \\ &= \prod_{i=1}^{n} P\{X_i \leq x\} \\ &= (F(x))^n, \end{aligned}$$

y la distribución del mínimo por $L_n(x) = 1 - (1 - F(x))^n$:

$$\begin{aligned} L_n(x) &= P(Y_1 \leq x) = 1 - P(Y_1 > x) \\ &= 1 - P\{X_1 > x, X_2 > x, ..., X_n > x\} \\ &= 1 - P\{X_1 > x\} P\{X_2 > x\} \cdots P\{X_n > x\} \\ &= 1 - \prod_{i=1}^{n} P\{X_i > x\} \\ &= 1 - (1 - F(x))^n. \end{aligned}$$

■

Definición 1. Una sucesión de variables aleatorias $X_1, X_2, ...$ con función de distribución $F_1, F_2, ...$ respectivamente, se dice que converge en distribución a la variable aleatoria X, teniendo función de distribución F, denotado como $X_n \xrightarrow{d} X$, si

$$F_n(x) \longrightarrow F(x) \qquad \text{cuando } n \longrightarrow \infty$$

para todo punto de continuidad x de F.

Definición 2. Una variable aleatoria X tiene una distribución Normal con parámetros μ y σ^2, denotado por $X \sim N(\mu, \sigma^2)$, si su función de densidad tiene la forma

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)\right\}, \qquad -\infty < x < \infty$$

Si $X \sim N(\mu, \sigma^2)$, entonces la variable $\left(\frac{x-\mu}{\sigma}\right) \sim N(0,1)$, la cual se denomina distribución normal estandard. La variable $\left(\frac{x-\mu}{\sigma}\right)$ generalmente se denota por Z.

En la teoría estadística un resultado de suma importancia relacionado con la media muestral es el teorema central del límite el cual se menciona a continuación.

Teorema 2. Sea $X_1, X_2, ..., X_n$ una sucesión de tamaño n de variables independientes e identicamente distribuidas con función de distribución F cuya media esta dada por μ y con varianza $\sigma^2 > 0$. Entonces se tiene que

$$\frac{\left(\sum_{i=1}^{n} X_i - n\mu\right)}{\sigma/\sqrt{n}} \xrightarrow{d} Z$$

donde $Z \sim N(0,1)$.

3. Distribuciones asintóticas de valores extremos

Ya se conoce que las distribuciones del máximo y del mínimo están dadas por

$$L_n(x) = 1 - (1 - F(x))^n, \qquad H_n(x) = (F(x))^n. \tag{3}$$

En principio, el conocimiento de $F(x)$ es suficiente para determinar la distribución de Y_1 y de Y_n. Sin embargo, no siempre $L_n(x)$ y $H_n(x)$ tienen una forma simple. Además la función de distribución F es desconocida. Una posibilidad es usar técnicas estadísticas standard para estimar F de los datos observados, y entonces sustituir esta estimación en (3). Desafortunadamente, por el hecho de que en la obtención de $L_n(x)$ y $H_n(x)$, $F(x)$ se eleva a la n-ésima potencia, pequeñas discrepancias en la estimación de F conllevan a grandes discrepancias de $L_n(x)$ y $H_n(x)$.

Un enfoque alternativo es aceptar que F es desconocida, y estudiar la distribución asintótica de los extremos Y_1 y Y_n, cuando n es grande. Esto es similar a la práctica usual de aproximar la distribución de la media muestral para la distribución normal, lo cual esta justificado por el teorema central del límite. A este respecto, los resultados heuristicos fundamentales fueron descubiertos por Fisher y Tippett en 1928, [12], más tarde desde un punto de vista riguroso por Gnedenko en 1943, [14]. Gnedenko baso los fundamentos matemáticos de la teoría de los valores extremos en la clase de leyes límites de extremos. Los resultados asintóticos más importantes sobre el comportamiento del mínimo y del máximo de un conjunto de variables aleatorias. Las distribuciones $L_n(x)$ y $H_n(x)$, cuando $n \longrightarrow \infty$, son conocidas como las distribuciones asintóticas de los valores extremos.

El resultado más general nos dice que seleccionando adecuadamente sucesiones $\{a_n\}, \{b_n\}, \{A_n\}$ y $\{B_n\}$ las distribuciones límite de $(Y_1 - a_n)/b_n$ y de $(Y_n - A_n)/B_n$, pertenecen a la familia de distribuciones Generalizadas del valor Extremo.

Definanse

$$L(x) = \underset{n\longrightarrow\infty}{Lim} L_n(a_n + b_n x) = \underset{n\longrightarrow\infty}{Lim} P\left\{\frac{(Y_1 - a_n)}{b_n} \le x\right\} \tag{4}$$

y

$$H(x) = \underset{n\longrightarrow\infty}{Lim} H_n(A_n + B_n x) = \underset{n\longrightarrow\infty}{Lim} P\left\{\frac{(Y_1 - A_n)}{B_n} \le x\right\} \tag{5}$$

Resultados más específicos establecen que tanto las distribuciones límite L como H pueden tener solamente una de tres posibles formas, independientemente de la distribución original de las observaciones. El siguiente teorema establece lo anterior para el máximo Y_n.

Teorema 3 (Fisher-Tippett, Gnedenko) Sea $\{X_n\}$ una sucesión de variables independientes e identicamente distribuidas. Si existen sucesiones de constantes de normalización A_n en los reales, $B_n > 0$, constantes μ y $\sigma > 0$, y una función de distribución H no degenerada tal que

$$\underset{n\longrightarrow\infty}{Lim} P\left\{\frac{(Y_n - A_n)}{B_n} \le x\right\} = H(x), \tag{6}$$

para todo punto de continuidad x, entonces H pertenece a algunos de los siguientes tres tipos de distribución,

$$\text{Frechet}: \quad H^{(1)}(x) = \begin{cases} \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\}, & x > \mu, \xi > 0 \\ 0, & x \leq \mu, \xi > 0 \end{cases} \tag{7}$$

$$\text{Weibull}: \quad H^{(2)}(x) = \begin{cases} \exp\left\{-\left(-\frac{x-\mu}{\sigma}\right)^{1/\xi}\right\}, & x > \mu, \xi > 0 \\ 1, & x \geq \mu, \xi > 0 \end{cases} \tag{8}$$

$$\text{Gumbel}: \quad H^{(3)}(x) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}, \quad -\infty < x < \infty \tag{9}$$

La demostración del teorema anterior se basa en los resultados de Gnedenko y Kolmogorov que se encuentran en su trabajo de 1954 sobre teoremas límite para sumas de variables aleatorias independientes [15].

Con respecto a la distribución asintótica del valor mínimo Y_1, se tienen resultados análogos a los presentados hasta ahora para el valor máximo. Si ahora se cumple

$$\underset{n\longrightarrow\infty}{Lim} P\left\{\frac{(Y_1 - a_n)}{b_n} \leq x\right\} = L(x), \tag{10}$$

para algunas sucesiones de constantes de normalización a_n en los reales, $b_n > 0$, constantes μ y $\sigma > 0$, y una función de distribución L no degenerada, para todo punto de continuidad x, entonces L pertenece a alguno de los tres tipos de distribución,

$$\text{Frechet}: \quad L^{(1)}(x) = 1 - \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{1/\xi}\right\} \quad x > \mu, \xi > 0 \tag{11}$$

$$\text{Weibull}: \quad L^{(2)}(x) = 1 - \exp\left\{-\left(-\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\} \quad x \leq \mu, \xi > 0 \tag{12}$$

$$\text{Gumbel}: \quad L^{(3)}(x) = 1 - \exp\left\{-\exp\left(\frac{x-\mu}{\sigma}\right)\right\} \quad -\infty < x < \infty\,, \xi > 0 \tag{13}$$

4. Dominio de atracción

Dependiendo de cuál sea la distribución límite del máximo de una sucesión de variables aleatorias independientes con distribución común $F(x)$, se dice que F pertenece al dominio de atracción de $H^{(i)}$, $i = 1, 2, 3$, lo cual se denota con $F \in D\left(H^{(i)}\right)$. Es importante hacer notar que no es necesario conocer la forma exacta de F para determinar a que dominio de atracción pertenece. El comportamiento en la cola de $F(x)$ determina su dominio de atracción.

Con estos últimos resultados se tiene que si F es una distribución exponencial, normal o Weibull ($F(x) = 1 - \exp(-x^{\alpha}), x \geq 0, \alpha > 0$), entonces $F \in D\left(H^{(3)}\right)$, y si F es uniforme, $F \in D\left(H^{(2)}\right)$. También se cumple una propiedad de autocerradura, pues $H^{(i)} \in D\left(H^{(i)}\right)$. En cuanto a los valores extremos mínimos se tiene que si F es una distribución normal, $F \in D\left(L^{(3)}\right)$, y si F es exponencial, uniforme o Weibull, entonces $F \in D\left(L^{(1)}\right)$.

5. Familia de distribuciones del valor extremo generalizadas

Las tres distribuciones anteriores para el máximo pueden ser combinadas en una familia teniendo distribución de la forma

$$H(x) = \exp\left\{ -\left[1 + \xi \left(\frac{x-\mu}{\sigma} \right) \right]^{-1/\xi} \right\}, \tag{14}$$

definida en el conjunto $\{x : 1 + \xi(x-\mu)/\sigma > 0\}$, donde los parámetros satisfacen $-\infty < \mu < \infty$, $\sigma > 0$, y $-\infty < \xi < \infty$, esta es la familia de distribuciones del valor extremo generalizada para el máximo. El modelo tiene tres parámetros: un parámetro de localización, μ; un parámetro de escala, σ; y un parámetro de forma ξ. Las clase Fréchet corresponde al caso $\xi > 0$ y la clase Weibull corresponden al caso $\xi < 0$. Para $\xi = 0$, es definida como el limite de (14) cuando $\xi \longrightarrow 0$, correspondiente a la familia Gumbel con función de distribución

$$H(x) = \exp\left\{ -\exp\left[-\left(\frac{x-\mu}{\sigma} \right) \right] \right\}, \ -\infty < x < \infty. \tag{15}$$

La familia de distribuciones del valor extremo generalizada para el mínimo esta dada por:

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

definida en el conjunto $\{x : 1 - \xi(x-\mu)/\sigma > 0\}$.

Referencias

[1] Barao, M. I. and Tawn, J. A. "Extremal Analysis of short Series with Outliers: sea levels and athletics records". *Applied Statistics*, **48**(4), 1999, 469-487.

[2] Behrens, C. N., Lopes, H. F., and Gamerman, D. "Bayesian Analysis of Extreme Events with Threshold Estimation". *Statistical Modelling*, **4**, 2004, 227-244.

[3] Beirlant, J., Goegebeur, Y., Segers, J. y otros. *Statistics of extremes. Theory and Applications*. Wiley, London, 2004.

[4] Bortot, P., Coles, S., and Tawn, J. A. "The Multivariate Gaussian Tail Model: An Application to Oceanographic Data". *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **49**(1), 2000, 31-49.

[5] Casson, E, and Coles, S. "Simulation and Extremal Analysis of Hurricane Events". *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **49**(3), 2000, 227-245.

[6] Coles, S. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London, 2001.

[7] Coles, S., and Tawn, J. A. "Statistical Methods for Multivariate Extremes: An Application to Structural Design". *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **43**(1), 1994, 1-48.

[8] Coles, S., and Tawn, J. A. "A Bayesian Analysis of Extreme Rainfall Data". *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **45**(4), 1996, 463-478.

[9] Coles, S., and Pericchi, L. "Anticipating Catastrophes Through Extreme Value Modelling". *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **52**(4)**,** 2003, 405-416.

[10] Chavez-Demoulin. V, and Embrechts, P. "Smooth Extremal Models in Finance and Insurance". The *Journal of Risk and Insurance,* **71**(2)**,** 2004, 183-199.

[11] Davison, A. C., and Ramesh, N. I. "Local Likelihood Smoothing of Sample Extremes". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**(1)**,** 2000, 191-208.

[12] Fisher, R. A., and Tippett, L.H.C. "On the Estimation of the Frequency Distributions of the Largest or Smallest Member of a Sample". *Proceeding of the Cambridge Philosophical Society*, **24**, 1928, 180-190.

[13] Finkenstädt, B. and Rootzén, H. *Extreme Values in Finance, Telecommunications, and the Enviroment.* Chapman & Hall, Boca Raton, Florida, 2004.

[14] Gnedenko, B. V. "Sur la Distribution Limite du Terme Maximum d'une Série Aléatoire". *Annals of Mathematics*, **44**, 1943, 423-453.

[15] Gnedenko, B. V. and Kolmogorov, A. N. *Limit Distributions for Sums of Independent Random Variables.* Addison-Wesley, 1954 (Translated from Russian).

[16] Katz, R. W., Parlange, M. B., and Naveau, P. "Statistics of Extremes in Hydrology". *Advances in Water Resources*, **25**(8)**,** 2002, 1287-1304.

[17] Ledford, A. W., and Tawn, J. A. "Diagnostics for Dependence within Time Series Extremes". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(2)**,** 2003, 521-543.

[18] Nadarajah, S., Anderson, C. W., and Tawn, J. A. "Ordered Multivariate Extremes". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**(2)**,** 1998, 473-496.

[19] Nadarajah, S. "Simulation of Multivariate Extreme Values". *Journal of Statistical Computation and Simulation*, **62**(4)**,** 1999, 395-410.

[20] Pauli, F., and Coles, S. "Penalized Likelihood Inference in Extreme Value Analyses". *Journal of Applied Statistics*, **28**(5)**,** 2001, 547-560.

[21] Robinson, M. E. and Tawn, J. A. "Statistics for Extreme Sea Currents". *Applied Statistics*, **46**(2)**,** 1997, 183-205.

[22] Robinson, M. E. and Tawn, J. A. "Extremal Analysis of Processes Sampled at Different Frequencies". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**(1)**,** 2000, 117-135.

[23] Walshaw, D. "Modeling Extreme wind Speeds in Regions prone to Hurricanes". *Applied Statistics*, **49**(1)**,** 2000, 51-62.