USO DE LA DISTRIBUCIÓN LÉVY PARA AJUSTAR DATOS CON MARCADA ASIMETRÍA Y VALORES EXTREMOS

Jessica Lizeth Martínez Naranjo*, Carlos Armando Alvear Rodríguez**, José Rafael Tovar Cueva*
*Universidad del Valle, Colombia.
**Universidad Santiago de Cali, Colombia.

ABSTRACT
In order to propose a statistical methodology that allows to model asymmetric data using the Lévy distribution, a simulation study is presented under nine different scenarios to evaluate the estimation of the parameters of the distribution in the two approaches of statistics (Classical and Bayesian). The probability distributions Log-Normal, Lévy and Lévy Standard were considered to model the behavior of two real data sets with positive asymmetry, finding that the Lévy distribution fitted well to the proposed data set, therefore the Lévy distribution can be considered as a candidate to adjust asymmetric data with the presence of extreme values.

KEYWORDS: Asymmetry, Lévy Distribution, Classical Statistics, Bayesian Statistics, Extreme Values.

MSC: 60E05

RESUMEN
Con el fin de proponer una metodología estadística que permita modelar datos asimétricos usando la distribución Lévy, se presenta un estudio de simulación bajo nueve escenarios diferentes para evaluar la estimación de los parámetros de la distribución en los dos enfoques de la estadística (Clásica y Bayesiana). Se consideraron las distribuciones de probabilidad Log-Normal, Lévy y Lévy Estándar para modelar el comportamiento de dos conjuntos de datos reales con asimetría positiva, encontrando que la distribución Lévy ajustó bien al conjunto de datos propuesto, por lo tanto se puede considerar la distribución Lévy como candidata para ajustar datos asimétricos con presencia de valores extremos.

PALABRAS CLAVE: Asimetría, Distribución Lévy, Estadística Clásica, Estadística Bayesiana, Valores Extremos.

1.. INTRODUCCIÓN

En diferentes situaciones prácticas se utilizan distribuciones de probabilidad para ajustar datos reales y frecuentemente se utiliza la distribución Normal como un modelo de referencia, sin embargo, en

*jessica.martinez.n@correounivalle.edu.co, carlos.alvear00@usc.edu.co,jose.r.tovar@correounivalle.edu.co

algunas ocasiones las variables analizadas distan de un comportamiento simétrico alrededor de la media y pueden estar influenciadas por la presencia de valores inusuales que tienen una baja probabilidad de ocurrir. Es por esta razón que el análisis y estudio de las distribuciones asimétricas cobra mayor sentido, puesto que para un investigador es de vital interés identificar la distribución de probabilidad a la cual se ajusta su variable de estudio, pues con base a esta distribución podrá realizar análisis posteriores e inferencias sobre la población. Así, el problema de caracterizar una distribución ha atraído recientemente la atención de muchos investigadores [11]. Por lo tanto, existe la necesidad de encontrar una distribución asimétrica que sea una buena alternativa frente a las distribuciones asimétricas tradicionales como lo son la Gamma, Weibull, Exponencial o Log-Normal y de esta forma poder modelar conjuntos de datos con presencia de valores extremos, situación que es muy común en campos como salud, finanzas y economía [1, 21, 17, 9, 20].

La distribución Lévy es una distribución de probabilidad continua para variables aleatorias no negativas, desarrollada por el matemático francés Paul Lévy en 1925 durante sus investigaciones del comportamiento de las variables aleatorias independientes. Esta distribución pertenece a la familia de distribuciones estables, las cuales han sido de gran interés porque permiten modelar la asimetría y colas arbitrariamente más grandes que la distribución Normal [5].

Las distribuciones estables son descritas por cuatro parámetros $(\alpha, \beta, \mu, \sigma)$. El parámetro α conocido como el carácter exponencial define la estabilidad o ancho de las colas, un valor grande de α implica colas delgadas $(0 < \alpha \leq 2)$. El parámetro β define la simetría de la distribución, el cual puede tomar valores entre $-1 \leq \beta \leq 1$. El parámetro de localización μ $(-\infty \leq \mu \leq \infty)$ define la posición o dominio de la distribución y el parámetro de escala σ $(\sigma > 0)$, define la dispersión o curtosis de la curva. Además, el dominio de todas las distribuciones estables está dado por $(-\infty, \infty)$, excepto en los siguientes casos: si $\alpha < 1$ y $\beta = 1$ entonces el dominio esta entre $(0, \infty)$; si $\alpha < 1$ y $\beta = -1$ el dominio está en $(-\infty, 0)$ [8].

Aunque en general no existe una forma cerrada simple para la función de densidad de probabilidad de una distribución estable, se conocen tres distribuciones de esta familia cuya función de densidad de probabilidad se puede expresar en forma cerrada, estas son: la distribución Normal, la distribución Cauchy y la distribución Lévy, para obtener estas distribuciones se deben tener ciertas combinaciones de los parámetros de estabilidad y simetría, las cuales se presentan en la tabla 1.

Distribución	α	β
Normal	2	0
Cauchy	1	0
Lévy	$\frac{1}{2}$	1

Tabla 1: Combinación de parámetros de estabilidad α y simetría β

Como se puede observar de las tres distribuciones de probabilidad que tienen una forma cerrada en la función de densidad de probabilidad, solo la distribución Lévy es asimétrica.

2.. DISTRIBUCIÓN LÉVY

De acuerdo con Ahsanullah y Nevzorov [2], la distribución Lévy de parámetros (μ, σ) tiene la siguiente función de densidad de probabilidad (f.d.p):

$$f(x;\mu,\sigma) = \left(\frac{\sigma}{2\pi}\right)^{1/2}(x-\mu)^{-3/2}exp\left(-\frac{\sigma}{2(x-\mu)}\right) \tag{2..1}$$

Donde $x \geq \mu$. Además μ y σ son los parámetros de localización y escala, respectivamente y satisfacen que $-\infty < \mu < \infty$ y $\sigma > 0$. Y la función de distribución acumulada (f.d.a) de la variable aleatoria X con una distribución Lévy con f.d.p (2..1) está dada por:

$$F(x;\mu,\sigma) = erfc\left[\left(\frac{\sigma}{2(x-\mu)}\right)^{1/2}\right] \tag{2..2}$$

Donde $erfc$ es el complemento de la función error, es decir, $erfc = 1 - erf(t)$ y la función error $erf(t)$ está dada por:

$$erf(t) = \frac{2}{\sqrt{\pi}}\int_0^t exp(-k^2)\,dk \tag{2..3}$$

Por lo tanto, la f.d.a para una variable aleatoria Lévy (μ, σ) se puede expresar como:

$$F(x;\mu,\sigma) = \frac{2}{\sqrt{\pi}}\int_t^\infty exp\left(-\frac{\sigma}{2(x-\mu)}\right)dx \tag{2..4}$$

El parámetro de localización μ tiene el efecto de desplazar la curva hacia la derecha una cantidad μ y cambia el dominio de la función en el intervalo $[\mu, \infty)$. El parámetro de escala σ tiene efecto sobre la curtosis de la curva, es decir, cuando el parámetro de escala es pequeño ($\sigma < 1$) la curva es más apuntada y con cola menos ancha (Leptocúrtica), pero a medida que el parámetro de escala aumenta, la curva es menos apuntada y con cola más ancha (Platicúrtica).

Por otra parte, la función característica de la distribución Lévy está dada por:

$$\phi_{\mu,\sigma}(k) = \int_\mu^\infty exp(ixk)\sqrt{\frac{\sigma}{2\pi}}(x-\mu)^{-3/2}exp\left(-\frac{\sigma}{2(x-\mu)}\right)dx = exp\left(i\mu k - \sqrt{-2i\sigma k}\right) \tag{2..5}$$

La función generatriz de momentos de una variable aleatoria X se define como $E[exp(kX)]$, $k \in \Re$, siempre que exista la esperanza. En este caso se debe tener en cuenta que una variable aleatoria proveniente de la distribución Lévy no tiene media ni varianza definida, por lo tanto, no existe ninguno de sus momentos puesto que la integral impropia correspondiente no es convergente:

$$E[exp(kX)] = \int_0^\infty exp\,(kx)\left(\frac{\sigma}{2\pi}\right)^{1/2}(x-\mu)^{-3/2}exp\left[-\frac{\sigma}{2(x-\mu)}\right]dx = \infty \tag{2..6}$$

De acuerdo con Achcar [1], un caso especial de la distribución Lévy se da cuando el parámetro de localización es igual a cero ($\mu = 0$), en cuyo caso se obtiene la denominada distribución Lévy Estándar, la cual solo está compuesta por un parámetro de escala σ.

3.. ESTIMACIÓN DE PARÁMETROS

3.1.. Método de Máxima Verosimilitud

La estimación de parámetros que caracterizan la distribución de probabilidad de una población es uno de los principales problemas de la inferencia Clásica y el método de máxima verosimilitud es una de las técnicas estadísticas empleadas para la solución de este problema de estimación [13].

Asumiendo una muestra aleatoria de tamaño n de datos $(X_1, ..., X_n)$ con distribución Lévy (μ, σ) y función de densidad de probabilidad dada en la ecuación (2..1), la función de verosimilitud y log-verosimilitud para μ y σ están dadas respectivamente por:

$$L(\sigma, \mu \mid x_i) = \left(\frac{\sigma}{2\pi}\right)^{n/2} \left[\prod_{i=1}^{n}(x_i - \mu)^{-3/2}\right] exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{\sigma}{x_i - \mu}\right) \tag{3..1}$$

$$\ell(\sigma, \mu \mid x_i) = \frac{n}{2}[log(\sigma) - log(2\pi)] - \frac{3}{2}\sum_{i=1}^{n} log(x_i - \mu) - \frac{1}{2}\sum_{i=1}^{n}\frac{\sigma}{x_i - \mu} \tag{3..2}$$

Entonces, para encontrar el estimador por máxima verosimilitud (EMV) para μ y σ se realizan las derivadas parciales de la función de log-verosimilitud dada en la ecuación (3..2) y se igualan a cero. Así para el parámetro μ se obtiene:

$$\begin{aligned} \frac{\partial \ell}{\partial \mu} &= \frac{3}{2}\frac{1}{\sum_{i=1}^{n}(x_i - \mu)} - \frac{\sigma}{2}\sum_{i=1}^{n}\frac{1}{(x_i - \mu)^2} = 0 \\ \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sum_{i=1}^{n}(x_i - \mu)} &= \frac{\sigma}{3} \end{aligned} \tag{3..3}$$

Debido a la complejidad para despejar μ de la ecuación (3..3) es necesario emplear métodos numéricos tales como Newton Raphson [23] o Nelder-Mead [16] para obtener el valor de $\hat{\mu}$. Mientras que, para el parámetro σ se obtiene:

$$\begin{aligned} \frac{\partial \ell}{\partial \sigma} &= \frac{n}{2\sigma} - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{(x_i - \mu)} = 0 \\ \frac{1}{\sigma} &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(x_i - \mu)} \end{aligned} \tag{3..4}$$

Por lo tanto, el estimador para σ por máxima verosimilitud está dado por:

$$\hat{\sigma} = \frac{n}{\sum_{i=1}^{n}(x_i - \hat{\mu})^{-1}} \tag{3..5}$$

En el caso específico de la distribución Lévy Estándar, se tiene que el estimador de máxima verosimilitud para el parámetro de escala σ está dado por la siguiente expresión:

$$\hat{\sigma} = \frac{n}{\sum_{i=1}^{n} x_i^{-1}} \tag{3..6}$$

3.2.. Método Bayesiano de Estimación

Para aplicar el teorema de Bayes al problema de inferencia de parámetros, se puede hacer una distinción de las variables en dos tipos. Por un lado, las variables conocidas (los datos experimentales) y por otro lado, las variables desconocidas, es decir, aquellas cuyos valores se quieren inferir mediante la aplicación del Teorema de Bayes [10]. Por lo tanto, en el enfoque Bayesiano los parámetros son vistos como variables aleatorias, a las cuales se les asigna una distribución a priori de probabilidad, con base en un comportamiento natural aleatorio [3]. En este caso, como no se cuenta con la información de un experto, se utilizarán distribuciones a priori no informativas. Cuando no se tiene ninguna información sobre los parámetros se toma una distribución inicial conjunta no informativa sobre ellos, en la que se supone que las distribuciones de cada parámetro son independientes, es decir:

$$\pi(\mu, \sigma) = \pi(\mu) \cdot \pi(\sigma) \tag{3..7}$$

Como μ es un parámetro de localización, se asume como distribución a priori no informativa una distribución Uniforme en el intervalo $[0, b]$, donde b es el mínimo de los datos de la muestra.

$$\pi(\mu) = \frac{1}{b} \tag{3..8}$$

Como σ es el parámetro de escala de la distribución Lévy, se asume una distribución a priori no informativa Gamma.

$$\pi(\sigma) = \frac{1}{\Gamma(\alpha)\beta} \left(\frac{\sigma}{\beta}\right)^{\alpha-1} exp\left(-\frac{\sigma}{\beta}\right) \tag{3..9}$$

Así, se tiene como distribución a priori conjunta para μ y σ:

$$\pi(\sigma, \mu) = \frac{1}{b\Gamma(\alpha)\beta^{\alpha}} \sigma^{\alpha-1} exp\left(-\frac{\sigma}{\beta}\right) \tag{3..10}$$

La distribución posterior está dada por:

$$f(\sigma, \mu \mid x_1 \cdots x_n) = \frac{L(\sigma, \mu \mid x_i)\pi(\sigma, \mu)}{\int_0^b \int_0^{\infty} L(\sigma, \mu \mid x_i)\pi(\sigma, \mu) d\sigma d\mu} \tag{3..11}$$

Donde $L(\sigma, \mu \mid x_i)$ es la función de verosimilitud y $\pi(\sigma, \mu)$ es la distribución a priori conjunta de los parámetros. Reemplazando estas funciones se obtiene que la distribución posterior es:

$$\begin{aligned} f(\sigma, \mu \mid x_1 \cdots x_n) &= \frac{\left(\frac{\sigma}{2\pi}\right)^{n/2} \prod_{i=1}^{n} (x_i - \mu)^{-3/2} \; exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{\sigma}{(x_i - \mu)}\right) \frac{1}{b\Gamma(\alpha)\beta^{\alpha}} \; \sigma^{\alpha-1} \; exp\left(-\frac{\sigma}{\beta}\right)}{\int_0^b \int_0^{\infty} \left(\frac{\sigma}{2\pi}\right)^{n/2} \prod_{i=1}^{n} (x_i - \mu)^{-3/2} \; exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{\sigma}{(x_i - \mu)}\right) \frac{1}{b\Gamma(\alpha)\beta^{\alpha}} \; \sigma^{\alpha-1} \; exp\left(-\frac{\sigma}{\beta}\right) d\sigma d\mu} \\ &= \frac{\sigma^{n/2+\alpha-1} exp\left(-\sigma\left[\sum_{i=1}^{n} \frac{1}{2(x_i - \mu)} + \frac{1}{\beta}\right]\right)}{\int_0^b \int_0^{\infty} \sigma^{n/2+\alpha-1} exp\left(-\sigma\left[\sum_{i=1}^{n} \frac{1}{2(x_i - \mu)} + \frac{1}{\beta}\right]\right) d\sigma d\mu} \end{aligned} \tag{3..12}$$

Debido a la complejidad de la expresión en la ecuación (3..12) es necesario emplear algún método de Cadenas de Markov Monte Carlo (MCMC) [14] para obtener la distribución posterior.

Por otra parte, si se considera la distribución Lévy Estándar, la expresión en la ecuación (3..12) se reduce a:

$$f(\sigma \mid x_1 \cdots x_n) = \frac{\sigma^{n/2+\alpha-1}\ exp\left[-\sigma\left(\frac{1}{2\sum_{i=1}^{n} x_i} + \frac{1}{\beta}\right)\right]}{\int_0^\infty \sigma^{n/2+\alpha-1}\ exp\left[-\sigma\left(\frac{1}{2\sum_{i=1}^{n} x_i} + \frac{1}{\beta}\right)\right] d\sigma} \tag{3..13}$$

Donde nuevamente debido a la complejidad de la expresión en la ecuación (3..13) es necesario emplear algún método MCMC para obtener la distribución posterior.

Método Empírico de Bayes

El método empírico de Bayes estima la distribución a priori $\pi(\theta)$ a partir de los datos. A este método se le puede dar una interpretación frecuencial, así este método puede ser esencialmente no Bayesiano, en el sentido de no involucrar probabilidades subjetivas [18].

En términos de distribuciones se trata de determinar una distribución posterior (ecuación (3..11)), donde $L(\theta \mid x)$ es la función de verosimilitud evaluada en θ y $\pi(\theta)$ es la distribución a priori, en el caso de Bayes empírico los parámetros desconocidos de $\pi(\theta)$ que se estimarán para poder trabajar la ecuación (3..11) de la misma manera que el método de Bayes.

Para poder ilustrar este método se consideran las variables aleatorias X_i $(i = 1, 2, \ldots, n)$, donde $X_i \sim Levy(\mu, \sigma)$, además de acuerdo con la ecuación (3..9), la distribución a priori para el parámetro de escala de la distribución Lévy se distribuye Gamma, es decir, $\sigma \sim Gamma(\alpha, \beta)$, donde α y β son desconocidos. Al asumir estos parámetros como desconocidos es necesario estimarlos a partir de los datos de la siguiente manera:

- Se considera una muestra aleatoria inicial equivalente al 10 % de la población.
- Se realiza la técnica de Bootstrap (Bootstrap no paramétrico) [7], para obtener mil muestras diferentes a partir de la muestra aleatoria inicial.
- A cada una de las muestras generadas se estiman los parámetros de la distribución Lévy mediante el método de máxima verosimilitud.
- Al vector de estimaciones del parámetro de escala, es decir, $\hat{\sigma} = (\hat{\sigma_1}, \hat{\sigma_2}, ..., \hat{\sigma}_{1000})$ se le calcula la media y la varianza muestral, es decir, $\bar{X}(\hat{\sigma}) = a$ y $S^2(\hat{\sigma}) = b$.
- Al tener los valores de la media y varianza del vector de estimaciones (a, b), se igualan al valor esperado y varianza esperada de la distribución Gamma, los cuales están dados por: $E(x) = \alpha\beta$ y $V(x) = \alpha\beta^2$, donde se genera un sistema de ecuaciones para encontrar los valores de α y β.

De este modo se obtiene la estimación de los hiperparámetros para la distribución a priori utilizando información proveniente de los datos.

4.. ESTUDIO DE SIMULACIÓN

Con el propósito de conocer el comportamiento de las estimaciones de los parámetros de localización y escala de la distribución Lévy, se realizó un análisis de simulación mediante el software estadístico R [19], donde se tomó 1000 muestras provenientes de la distribución Lévy(μ_j, σ_k), con $j = 1, 2, 3$ y $k = 1, 2, 3$, simulados mediante la librería rmutil [22]. En cada una de las muestras se estimó los valores de los parámetros μ y σ empleando la librería maxLik [12] para realizar las estimaciones mediante el método Clásico y la librería MHadaptative [6] para realizar las estimaciones mediante el método Bayesiano.

Para realizar este análisis se utilizó dos tamaños de muestra ($n = 30$ y $n = 500$) y se establecieron los siguientes valores para los parámetros de la distribución Lévy: para la localización μ (0.5, 2 y 10) y para la escala σ (0.7, 2 y 10) y a partir de cada una de las diferentes combinaciones entre tamaños de muestra y parámetros, se realizó las respectivas estimaciones de ambos parámetros. Con el fin de ver el desempeño de los estimadores en la estimación Clásica se calcularon algunos indicadores como el valor esperado, el sesgo y la raíz cuadrada del error cuadrático ($\sqrt{ECM}$). Mientras que en la estimación Bayesiana se calculó la media posterior de cada uno de los parámetros y sus correspondientes regiones de credibilidad al 95 %.

4.1.. Aproximación Clásica

La estimación Clásica se realiza mediante el método de máxima verosimilitud, empleando los métodos numéricos Newton-Raphson que emplea información de las derivadas de la función objetivo (log-verosimilitud) y Nelder-Mead el cual es un algoritmo que encuentra un mínimo de la función objetivo de una o varias variables sin diferenciación.

De acuerdo con los resultados obtenidos, se encontró que cuando el tamaño de muestra es pequeño ($n = 30$) se presentan los valores más altos para la raíz del error cuadrático medio (tablas 2 y 3), además se encontró que los mayores valores registrados de ($\sqrt{ECM}$) se presentan en los escenarios de simulación donde el parámetro de escala (σ) toma el valor más alto. Destacando que, cuando el tamaño de muestra es grande ($n = 500$) se presentan las mejores estimaciones de los parámetros, puesto que se presentan los menores sesgos y los menores valores para el ($\sqrt{ECM}$).

Se debe resaltar que, en todas las estimaciones realizadas, el parámetro de localización μ se sobreestima, mientras que el parámetro de escala σ se subestima. Otro aspecto que es importante de resaltar es que en las estimaciones realizadas cuando el tamaño de muestra es igual a 30 y se emplea el método Newton-Raphson (ver tabla 2) se presenta el mayor sesgo, mientras que, bajo el mismo escenario de simulación, pero empleando el método Nelder-Mead (ver tabla 3) se observan sesgos menores.

Tabla 2: Estimación Clásica (Newton-Raphson, n=30)

		$\mathbf{E}(\hat{\mu})$	**Sesgo**	**ECM**	$\mathbf{E}(\hat{\sigma})$	**Sesgo**	**ECM**
μ=0.5	σ=0.7	0.5350	0.0350	0.0059	0.6704	-0.0295	0.0812
	σ=2	0.5943	0.0943	0.0504	1.9655	-0.0344	0.6166
	σ=10	1.0458	0.5458	1.1721	9.4449	-0.5550	12.0546
μ=2	σ=0.7	2.0350	0.0350	0.0062	0.6718	-0.0281	0.0751
	σ=2	2.0841	0.0841	0.0520	1.9738	-0.0261	0.7749
	σ=10	2.5194	0.5194	1.3551	9.5960	-0.4039	13.5357
μ=10	σ=0.7	10.0334	0.0334	0.0058	0.6732	-0.0267	0.0706
	σ=2	10.0939	0.0939	0.0513	1.9649	-0.0350	0.6111
	σ=10	10.4435	0.4435	1.1941	9.7787	-0.2212	15.9954

Tabla 3: Estimación Clásica (Nelder-Mead, n=30)

		$\mathbf{E}(\hat{\mu})$	**Sesgo**	**ECM**	$\mathbf{E}(\hat{\sigma})$	**Sesgo**	**ECM**
μ=0.5	σ=0.7	0.5340	0.0340	0.0062	0.6720	-0.0279	0.0676
	σ=2	0.5865	0.0865	0.0480	1.9739	-0.0260	0.6312
	σ=10	0.6720	0.1720	1.9670	9.4656	-0.5343	10.0678
μ=2	σ=0.7	2.0326	0.0326	0.0058	0.6820	-0.0179	0.0727
	σ=2	2.1063	0.1063	0.0547	1.8782	-0.1217	0.5967
	σ=10	2.0510	0.0510	0.0665	9.9353	-0.0646	1.2484
μ=10	σ=0.7	10.0313	0.0313	0.0060	0.6787	-0.0212	0.0784
	σ=2	10.0972	0.0972	0.0506	1.9241	-0.0758	0.6238
	σ=10	10.4801	0.4801	1.2640	9.6707	-0.3292	15.0970

Tabla 4: Estimación Clásica (Newton-Raphson, n=500)

		$\mathbf{E}(\hat{\mu})$	**Sesgo**	**ECM**	$\mathbf{E}(\hat{\sigma})$	**Sesgo**	**ECM**
μ=0.5	σ=0.7	0.5017	0.0017	0.0001	0.6989	-0.0010	0.0037
	σ=2	0.5076	0.0076	0.0014	1.9881	-0.0118	0.0298
	σ=10	0.5250	0.0250	0.0355	10.0004	0.0004	0.7344
μ=2	σ=0.7	2.0026	0.0026	0.0001	0.6990	-0.0009	0.0033
	σ=2	2.0052	0.0052	0.0015	1.9942	-0.0057	0.0282
	σ=10	2.0309	0.0309	0.0371	9.9893	-0.0106	0.721
μ=10	σ=0.7	10.0023	0.0023	0.0001	0.6967	-0.0032	0.0035
	σ=2	10.0060	0.0060	0.0015	1.9928	-0.0071	0.0307
	σ=10	10.0352	0.0352	0.0358	9.9151	-0.0848	0.6817

Tabla 5: Estimación Clásica (Nelder-Mead, n=500)

		$\mathbf{E}(\hat{\mu})$	**Sesgo**	**ECM**	$\mathbf{E}(\hat{\sigma})$	**Sesgo**	**ECM**
μ=0.5	σ=0.7	0.5023	0.0023	0.0001	0.6975	-0.0024	0.0031
	σ=2	0.5077	0.0077	0.0015	1.9815	-0.0184	0.0296
	σ=10	0.5230	0.0230	0.0563	9.9858	-0.0141	0.6901
μ=2	σ=0.7	2.0015	0.0015	0.00018	0.7006	0.0006	0.0034
	σ=2	2.0050	0.0050	0.0014	1.9910	-0.0089	0.0285
	σ=10	2.0248	0.0248	0.0355	9.9904	-0.0095	0.6773
μ=10	σ=0.7	10.0018	0.0018	0.0001	0.7000	-0.00007	0.0034
	σ=2	10.0089	0.0089	0.0015	1.9858	-0.0142	0.0292
	σ=10	10.0290	0.0290	0.0380	9.9687	-0.0312	0.7162

4.2.. Aproximación Bayesiana

Al realizar el estudio de simulación bajo diferentes escenarios, empleando la estimación Bayesiana y utilizando las distribuciones a priori no informativas para los parámetros μ y σ establecidas en la sección 3.2., se encontró que, en el escenario de simulación donde el tamaño de la muestra es pequeño ($n = 30$) las estimaciones estuvieron más lejanas al parámetro real (ver tabla 6), mientras que cuando el tamaño de muestra es grande ($n = 500$) (ver tabla 7) las estimaciones estuvieron más cercanas al valor real del parámetro. Además se encontró que en las regiones de credibilidad se reduce el rango del intervalo a medida que aumenta el tamaño de muestra.

Se debe resaltar que contrario a la estimación Clásica donde el parámetro de escala σ se subestima, en este caso se sobreestima, mientras que el parámetro de localización μ se subestima en la mayoría de los casos.

Tabla 6: Estimación Bayesiana (n=30)

		Media posterior (μ)	**Regiones de Credibilidad (μ)**	**Media posterior (σ)**	**Regiones de Credibilidad (σ)**
μ=0.5	σ=0.7	0.3137	(0.0509 ; 0.4928)	1.1701	(0.4915 ; 2.2217)
	σ=2	0.6707	(0.2596 ; 0.8930)	1.6577	(0.7683 ; 3.0437)
	σ=10	0.6883	(0.0389 ; 1.4785)	8.2089	(4.1710 ; 13.5360)
μ=2	σ=0.7	1.9991	(1.7895 ; 2.1033)	0.7257	(0.3238 ; 1.3558)
	σ=2	1.5733	(0.4131 ; 2.2307)	5.0576	(2.4149 ; 9.0549)
	σ=10	0.7825	(0.0338 ; 1.9206)	17.4177	(9.1186 ; 28.0846)
μ=10	σ=0.7	9.9374	(9.6677 ; 10.0590)	0.7426	(0.2986 ; 1.5395)
	σ=2	10.1370	(9.5604 ; 10.4477)	1.7979	(0.7083 ; 3.6149)
	σ=10	10.0000	(7.1714 ; 11.5021)	8.7200	(3.6811 ; 16.594)

Tabla 7: Estimación Bayesiana (n=500)

		Media posterior (μ)	**Regiones de Credibilidad (μ)**	**Media posterior (σ)**	**Regiones de Credibilidad (σ)**
μ=0.5	σ=0.7	0.5084	(0.4828 ; 0.5283)	0.6468	(0.5482 ; 0.7579)
	σ=2	0.5442	(0.4593 ; 0.6129)	1.9688	(1.6581 ; 2,3248)
	σ=10	0.4085	(0.0485 ; 0.7898)	11.9144	(10.0963 ; 13.8591)
μ=2	σ=0.7	1.9933	(1.9660 ; 2.0149)	0.6741	(0.5698 ; 0.7895)
	σ=2	2.0492	(1.9697 ; 2.1138)	2.0572	(1.7444 ; 2.4046)
	σ=10	1.7575	(1.3056 ; 2.1259)	10.4897	(8.7735 ; 12.3836)
μ=10	σ=0.7	9.9716	(9.9397 ; 9.9980)	0.7995	(0.6765 ; 0.9369)
	σ=2	10.0226	(9.9369 ; 10.0937)	2.1115	(1.7824 ; 2.4776)
	σ=10	10.3949	(10.0700 ; 10.6515)	8.6931	(7.3663 ; 10.1277)

5.. APLICACIÓN A DATOS REALES

Datos de concentración de anticuerpos contra el sarampión

Con el propósito de ajustar la distribución Lévy a dos conjuntos de datos reales, se utiliza en primer lugar los datos empleados en el artículo de Moulton y Halsey [15], los cuales hacen referencia a un estudio de seguridad e inmunogenicidad de la vacuna contra el sarampión en niños haitianos entre los años 1987 y 1990. Este estudio se realizó a 330 niños a los 12 meses de edad, en quienes se observó la concentración de anticuerpos contra el sarampión por medio de instrumentos de medición los cuales tenían un límite inferior de detección para la prueba de 0.1 Unidades Internacionales (UI), donde 86 casos presentaron mediciones iguales o inferiores a este límite, representando un 26.1 % de las observaciones las cuales se consideraron como censuradas. Los autores asumen que el comportamiento de la variable analizada sigue una distribución Log-Normal por su comportamiento asimétrico y la presencia de valores extremos. Debido a este comportamiento en la variable de interés en esta investigación se excluyen las observaciones censuradas obteniendo una muestra final de 244 niños. La concentración promedio de anticuerpos es de 1.59 UI con una desviación estándar de 2.32 UI, fluctuando en un rango entre 0.2 UI y 15.475 UI, donde el 50 % de los niños presentan niveles de concentración de anticuerpos inferiores a 0.7 UI.

Para identificar la distribución de probabilidad que se ajusta mejor a este conjunto de datos bajo el método Clásico, es necesario realizar las estimaciones de los parámetros de las distribuciones propuestas y de la distribución de referencia por medio del método de máxima verosimilitud considerando el algoritmo de Nelder-Mead. En la tabla 8 se presentan los resultados de las estimaciones para cada una de las distribuciones consideradas con su respectivo error estándar de estimación y el criterio de selección AIC.

Para realizar la estimación de parámetros bajo el enfoque Bayesiano se emplea el algoritmo Metropolis-Hasting [3], el cual es un método de Cadenas de Markov Monte Carlo, donde se generan muestras de la distribución posterior para estimar algunos indicadores de interés. Además, se debe tener en cuenta que los hiperparámetros de las distribuciones a priori se realizan empleando el método empírico de Bayes descrito en la sección 3.2. debido a que estos valores son desconocidos.

En primer lugar se considera la distribución Lévy, donde se asume una distribución a priori no informativa Uniforme en el intervalo $[0, 0.2]$ para el parámetro de localización, donde 0.2 es el valor mínimo observado en los datos y para el parámetro de escala se considera una distribución a priori no informativa Gamma $[5.86, 0.05]$. Ahora, en la distribución Lévy Estándar se asume para el parámetro de escala una distribución a priori no informativa Gamma $[32.12, 0.01]$. Finalmente, en el caso de la distribución Log-Normal y de acuerdo con Zellner [24], para el parámetro σ^2 se asume una distribución a priori no informativa de Jeffreys $1/1.156$ y para $\mu \mid \sigma^2$ se asume una distribución a priori no informativa Normal $[0.151, 1.156]$. En la tabla 8 se presentan los diferentes resultados obtenidos de la media posterior y las regiones de credibilidad del 95 % para cada una de las distribuciones de probabilidad consideradas.

Tabla 8: Estimación Clásica y Bayesiana (datos de anticuerpos)

Distribución	**Parámetro**	**EMV (EE)**	**AIC**	**Media Posterior**	**Región de Credibilidad**	**DIC**
Lévy	μ	0.176 (0.005)	593.042	0.176	(0.163 ; 0.186)	1239.169
	σ	0.226 (0.026)		0.226	(0.179 ; 0.281)	
Lévy Estándar	σ	0.556 (0.050)	708.060	0.630	(0.460 ; 0.823)	2086.349
Log-Normal	μ	-0.176 (0.066)	631.953	-0.214	(-0.349 ; -0.078)	1338.926
	σ	1.045 (0.047)		1.029	(0.937 ; 1.133)	

En la figura 1, se observa el histograma de los datos de la concentración de anticuerpos contra el sarampión frente a las funciones de densidad de las distribuciones de probabilidad propuestas, donde se destacan que las curvas de las distribuciones Lévy y Log-Normal presentan un mejor ajuste a los datos.

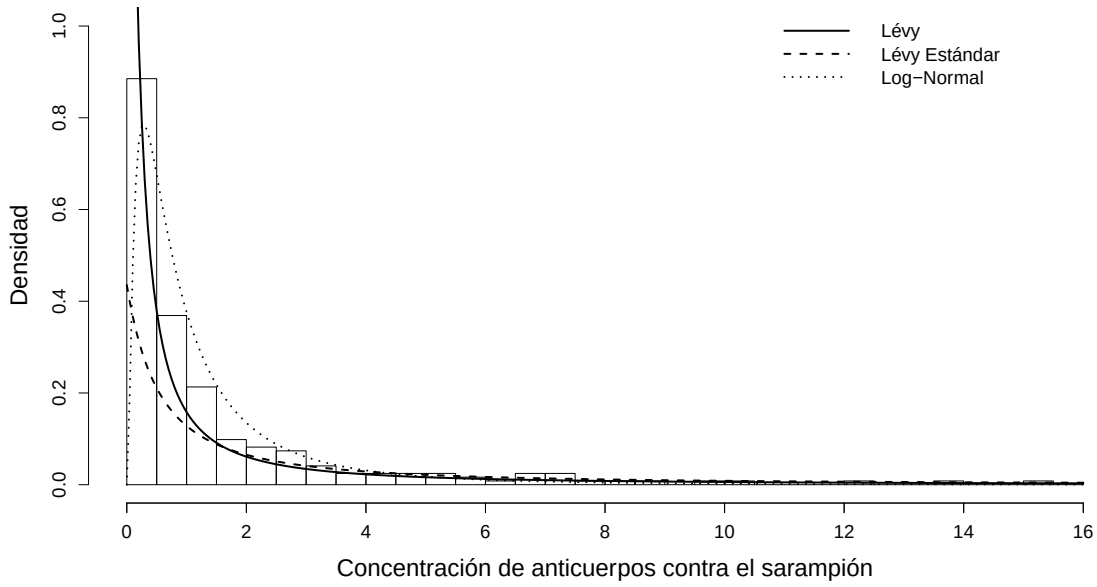

Figura 1: Funciones de densidad ajustadas usando EMV para los datos de anticuerpos.

Según lo observado en la figura 1 y el criterio del AIC en la tabla 8, las distribuciones que mejor se ajustan a los datos son la distribución Lévy (AIC=593.042) y la distribución Log-Normal (AIC=631.953). De forma analoga, para elegir la distribución que se ajusta mejor al conjunto de datos propuesto mediante la estimación Bayesiana, se utiliza el Criterio de Información de la Devianza (DIC). De acuerdo a los resultados de la tabla 8, la distribución Lévy es aquella con mejor ajuste, puesto que presenta el menor valor del DIC.

Datos de niveles de tiroglobulina post ablación

Ahora se empleará un segundo conjunto de datos empleado por Alvear y Tovar [4], los cuales hacen referencia a las mediciones de tiroglobulina post ablación en 91 pacientes con diagnóstico de cáncer diferenciado de tiroides que fueron llevados a terapia ablativa post-quirúrgica con yodo radiactivo, entre enero de 2006 y enero de 2010, en una clínica de nivel IV de atención ubicada en Bogotá Colombia. Los valores observados de esta variable están comprendidos entre 0 y 586 ng/ml, resaltando la presencia de tres valores relativamente grandes como son 142, 220 y 586 ng/ml. Los valores inferiores al límite de detección (0.1 ng/ml) son reportados como 0 ng/ml, lo cual se presenta en 34 casos que representan el 37.3 % de las observaciones, las cuales se consideran como censuradas. Dado que este conjunto de datos también considera valores censurados, estos se excluyen del análisis conformado un total de 57 observaciones, sobre las cuales se ajustarán las tres distribuciones de probabilidad y se realizaron las estimaciones de los parámetros empleando métodos Clásicos y Bayesianos.

Para emplear la metodología Bayesiana se considero en la distribución Lévy, una distribución a priori no informativa Uniforme en el intervalo $[0; 0.1]$ para el parámetro de localización y para el parámetro de escala se considera una distribución a priori no informativa Gamma $[139.55; 0.0007]$. En la distribución Lévy Estándar se asume para el parámetro de escala una distribución a priori no informativa Gamma $[219.38; 0.0012]$. Finalmente, en la distribución Log-Normal, para el parámetro σ^2 se asume una distribución a priori no informativa de Jeffreys $1/2.2298$ y para $\mu \mid \sigma^2$ se asume una distribución a priori no informativa Normal $[-0.0250; 2.2298]$.

En la tabla 9 se presentan los resultados de las estimaciones para cada una de las distribuciones consideradas con su respectivo error estándar de estimación y el criterio de selección AIC, bajo el método Clásico y bajo el método Bayesiano se presentan los resultados de la media posterior y las regiones de credibilidad del 95 %.

Tabla 9: Estimación Clásica y Bayesiana (datos de tiroglobulina)

Distribución	**Parámetro**	**EMV (EE)**	**AIC**	**Media Posterior**	**Región de Credibilidad**	**DIC**
Lévy	μ	0.072 (0.010)	223.224	0.070	(0.067 ; 0.073)	2536.798
	σ	0.100 (0.037)		0.093	(0.078 ; 0.109)	
Lévy Estándar	σ	0.274 (0.051)	233.320	0.311	(0.273 ; 0.351)	2810.583
Log-Normal	μ	-0.023 (0.295)	254.536	0.072	(-0.112 ; 0.259)	5659.165
	σ	2.231 (0.208)		2.172	(2.041 ; 2.313)	

En la figura 2, se observa que el histograma de los datos de los niveles de tiroglobulina post presenta una gran asimetría y una alta concentración de valores hacia los niveles de tiroglobulina post más bajos. De acuerdo a los resultados obtenidos en la tabla 9, se aprecia que la distribución Lévy presenta un mejor ajuste para estos datos (AIC=223.224 y DIC=2536.798) y que la distribución Log-Normal es la que menos se ajusta a las características de esta variable (AIC=254.536 y DIC=5659.165).

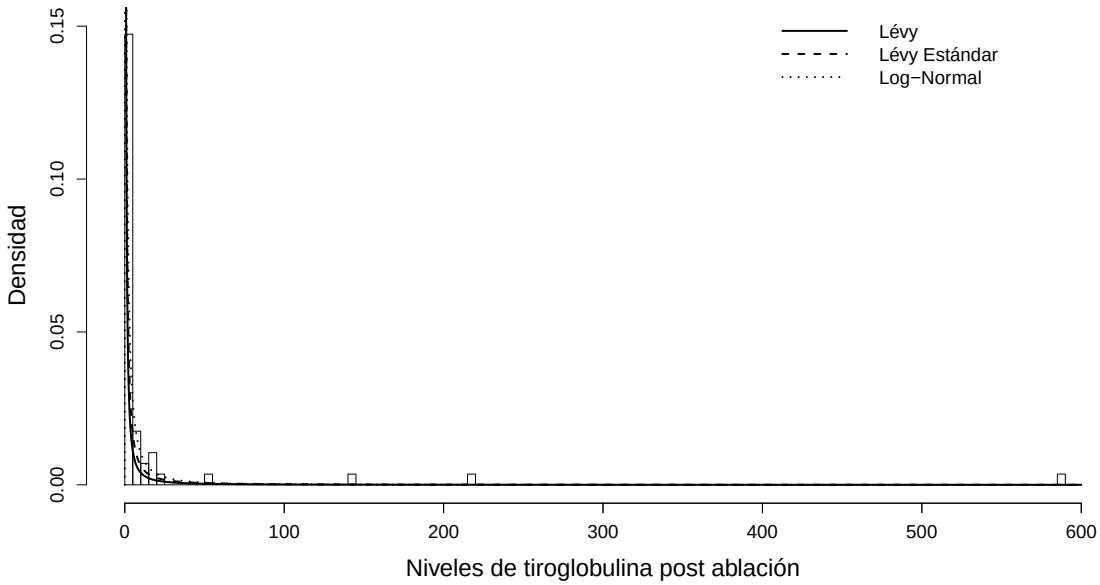

Figura 2: Funciones de densidad ajustadas usando EMV para los datos de tiroglobulina.

6.. CONCLUSIONES

En este trabajo se ha estudiado el comportamiento de la distribución Lévy mediante un análisis de simulación y el ajuste a dos conjuntos de datos con distribución asimétrica y presencia de valores extremos.

En un primer estudio de simulación se evaluaron las estimaciones de los dos parámetros de la distribución Lévy, contemplando diferentes escenarios, donde se utilizaron nueve configuraciones de los parámetros de localización y escala, dos tamaños de muestra y además se emplearon dos métodos numéricos para la estimación Clásica, se encontró que, cuando se cuenta con un tamaño de muestra pequeño las estimaciones con menor sesgo se obtuvieron usando el método de Nelder-Mead, también, se encontró que el parámetro de localización se sobreestima en todos los escenarios de simulación, mientras que el parámetro de escala se subestima. Por otra parte, en el estudio de simulación donde se evaluaron las estimaciones Bayesianas empleando distribuciones a priori no informativas para los parámetros μ y σ se encontró que el parámetro de localización se subestima y el parámetro de escala se sobreestima en la mayoría de los casos.

Se ajustaron las distribuciones a dos conjuntos de datos, el primero contenía registros sobre la concentración de anticuerpos del sarampión en niños de un año de edad y el segundo contenía información sobre los niveles de tiroglobulina post ablación en pacientes con cáncer diferenciado de tiroides. En ambos casos, se presenta asimetría en los datos siendo mucho más marcada para los datos

de tiroglobulina post ablación, para los cuales la cola de la distribución es bastante pesada. Se observó entonces la flexibilidad que tiene la distribución para ajustarse a datos asimétricos, aun cuando la distribución muestra una moderada asimetría a la derecha.

Para futuros trabajos se puede considerar el uso de otro tipo de distribuciones asimétricas con el fin de comparar los resultados tomando como distribución de referencia la Lévy, ya que de acuerdo a los resultados obtenidos esta resulta ser una buena candidata a la hora de ajustar datos con presencia de valores extremos, bajo cualquiera de los dos enfoques de la estadística (Clásica o Bayesiana).

RECEIVED: MARCH, 2019.
REVISED: SEPTEMBER, 2019.

REFERENCIAS

[1] ACHCAR, J. A., COELHO-BARROS, E. A., TOVAR, J. R., and MAZUCHELI, J. (2018): Use of lévy distribution to analyze longitudinal data with asymmetric distribution and presence of left censored data **Communications for Statistical Applications and Methods**, 25(1):43–60.

[2] AHSANULLAH, M. and NEVZOROV, V. B. (2014): Some inferences on the lévy distribution **Journal of Statistical Theory and Applications**, 13(3):205–211.

[3] ALBERT, J. (2009): **Bayesian computation with R** Springer Science & Business Media, New York.

[4] ALVEAR, C. A. and TOVAR, J. R. (2019): Regression models with asymmetric data for estimating thyroglobulin levels one year after the ablation of thyroid cancer **Statistical Methods in Medical Research**, 28(8):2258–2275.

[5] BUCKLE, D. (1995): Bayesian inference for stable distributions **Journal of the American Statistical Association**, 90(430):605–613.

[6] CHIVERS, C. (2012): **MHadaptive: General Markov Chain Monte Carlo for Bayesian Inference using adaptive Metropolis-Hastings sampling** R package version 1.1-8.

[7] EFRON, B. and TIBSHIRANI, R. J. (1994): **An Introduction to the Bootstrap** Chapman and Hall/CRC, New York.

[8] FELLER, W. (1971): **An introduction to probability theory and its applications** John Wiley & Sons, New York.

[9] FRAIN, J. C. (2009): **Studies on the Application of the α-stable Distribution in Economics** PhD thesis, University of Dublin, Ireland.

[10] GONZÁLEZ, D. S. (2008): **Modelos de mezcla de distribuciones α-estables** Master's thesis, Universidad de Granada, España.

[11] HAMEDANI, G., AHSANULLAH, M., and NAJIBI, S. (2015): Characterizations of lévy distribution via sub-independence of the random variables and truncated moments **Pak. J. Statist**, 31(4):417–425.

[12] HENNINGSEN, A. and TOOMET, O. (2011): maxlik: A package for maximum likelihood estimation in r **Computational Statistics**, 26(3):443–458.

[13] IPIÑA, S. L. and DURAND, A. I. (2008): **Inferencia estadística y análisis de datos** Pearson Educación, Madrid.

[14] LEE, P. M. (2012): **Bayesian statistics: an introduction** John Wiley & Sons, New York.

[15] MOULTON, L. H. and HALSEY, N. A. (1995): A mixture model with detection limits for regression analyses of antibody response to vaccine **Biometrics**, 51(4):1570–1578.

[16] NELDER, J. A. and MEAD, R. (1965): A simplex method for function minimization **The computer journal**, 7(4):308–313.

[17] PODOBNIK, B., VALENTINÄŒIÄŒ, A., HORVATIÄ†, D., and STANLEY, H. E. (2011): Asymmetric lÃ©vy flight in financial ratios **Proceedings of the National Academy of Sciences of the United States of America**, 108(44):17883–17888.

[18] PRIETO, V. H., QUINTERO, C., and RODRÍGUEZ, I. (1995): Análisis de bayes empírico mediante un ejemplo **Revista Colombiana de Estadística**, 16(31):81–88.

[19] R CORE TEAM (2017): **R: A Language and Environment for Statistical Computing** R Foundation for Statistical Computing, Vienna, Austria.

[20] SCALAS, E. and KIM, K. (2007): The art of fitting financial time series with levy stable distributions **Journal of the Korean Physical Society**, 50(1):105–111.

[21] SILVA, S. D., MATSUSHITA, R., GLERIA, I., FIGUEIREDO, A., and RATHIE, P. (2005): International finance, lÃ©vy distributions, and the econophysics of exchange rates **Communications in Nonlinear Science and Numerical Simulation**, 10(4):365–393.

[22] SWIHART, B. and LINDSEY, J. (2017): **rmutil: Utilities for Nonlinear Regression and Repeated Measurements Models** R package version 1.1.0.

[23] YPMA, T. J. (1995): Historical development of the newton–raphson method **SIAM review**, 37(4):531–551.

[24] ZELLNER, A. (1971): Bayesian and non-bayesian analysis of the log-normal distribution and log-normal regression **Journal of the American Statistical Association**, 66(334):327–330.