# DOCUMENTO TÉCNICO
# IDENTIFICACIÓN DE SINIESTROS EXTREMOS

**Resumen**— Desde una perspectiva prudencial, los recursos de capital que deben acreditar las aseguradoras deben responder a los riesgos que enfrentan. Bajo la normatividad vigente, el componente más significativo para las entidades de seguros generales es el riesgo de suscripción. En el cálculo de este componente, el Decreto 1349 de 2019 reconoce que, en caso de siniestros que puedan catalogarse como extremos, hay una mayor participación del reaseguro y, por tanto, se da un tratamiento diferencial a estos siniestros siempre que hayan sido pagados. Este documento presenta una aplicación de las metodologías utilizadas internacionalmente en la identificación de valores extremos, tomando como referencia la información de los siniestros avisados entre 2008 y 2018 para las compañías de seguros generales y cooperativas de seguros en Colombia.

**Palabras clave**— Teoría de valores extremos (EVT). Exceso de media. Hill plot. Value at risk (VaR). Expected shortfall (ES). L-momentos. Test Kolmogorov-Smirnov. Gráfico cuantil-cuantil.

**Clasificación JEL:**— C02, C13, C16, G22, G28.

# ÍNDICE

**Introducción** **3**

**Marco teórico y conceptual** **4**
Value at risk (VaR) y Expected Shortfall o Conditional VaR (ES - CVaR) . . . 4
Test de normalidad . . . 5
Teoría de Valores Extremos . . . 5
Distribución generalizada del valor extremo $(GEV)$ (Fisher-Tippett (1928), Gnedenko (1943)) . . . 5
Exceso sobre un umbral (Picklands (1975), Balkema-in Haan (1974)) . . . 6
Elección de umbrales . . . 6
Gráfico de exceso de media (ME-PLOT) . . . 6
Gráfico de Hill (Hill - Plot) . . . 7
L-momentos . . . 7
Umbral óptimo . . . 8
Selección del umbral óptimo metodología L-momentos . . . 8
Selección del umbral óptimo métodos gráficos . . . 8
Prueba de Kolmogorov-Smirnov (Chakravarty et al., 1967, pp.392-394). . . . 9

**Aplicación de la teoría de valores extremos al mercado de seguros de Colombia** **10**
Datos . . . 10
Análisis exploratorio . . . 10
Resumen estadístico del total de siniestros incurridos . . . 10
Resumen estadístico de siniestros incurridos mayores a \$1,000 .($\geq$ 1, en la base de datos) . . . 11
Histograma de frecuencia del total de siniestros incurridos . . . 11
Histograma de frecuencia del 99 % de siniestros incurridos . . . 12
Histograma de frecuencia del 1 % de siniestros incurridos más grandes . . . 12
Gráfica de los siniestros incurridos por año de aviso . . . 13
Gráfica de los siniestros incurridos por año de aviso (excluyendo los tres siniestros incurridos de mayor valor). 13
Aplicación de la teoría de los L-momentos . . . 14
Análisis de las colas . . . 14
Gráfico de exceso de media (ME-PLOT) . . . 15
Hill Plot . . . 16

**Selección de umbral y modelo de exceso** **18**
Prueba Kolmogov-Smirnov . . . 18
Gráfica de cuantil-cuantil Pareto . . . 18

**Conclusiones** **19**

## INTRODUCCIÓN

El Ministerio de Hacienda y Crédito Público mediante el artículo 5 del Decreto 1349 del 26 de julio de 2019, que modificó el artículo 2.31.1.2.6 del Decreto 2555 de 2010, otorgó facultades a la Superintendencia Financiera de Colombia (SFC) para determinar los criterios estadísticos y actuariales para la identificación de siniestros extremos, los cuales son insumo para el cálculo del componente del riesgo de suscripción en las entidades de seguros generales.

El presente documento contiene la metodología empleada para la identificación de siniestros extremos de las entidades de seguros generales, fundamentada en la Teoría de valores extremos (EVT), la cual mediante diversas técnicas estadísticas y probabilísticas permite modelar los valores de baja frecuencia y alta severidad de la distribución de los siniestros. Esta teoría permite establecer un umbral apropiado a partir del cual la distribución de la cola se aproxime a una deseada. Los métodos usuales para la selección de dicho umbral consisten principalmente en: i) métodos gráficos, entre los que se encuentran el Exceso de medias y el Hill plot, ii) medidas de riesgo como el Value at Risk (VaR) y el Expected Shortfall (ES) y iii) metodología de L-momentos. Dichos métodos se emplearon en la identificación de un rango de umbrales para los siniestros avisados entre 2008 y 2018 por las compañías de seguros generales y cooperativas de seguros en Colombia, excluyendo el ramo de terremoto.

En la primera sección se revisarán los conceptos y definiciones asociados a las medidas de riesgo, con especial énfasis en la EVT y en la segunda sección se aplicará la metodología expuesta para encontrar el umbral a partir del cual se considera un siniestro extremo en el mercado.

## MARCO TEÓRICO Y CONCEPTUAL

La adecuada gestión del riesgo de una entidad financiera requiere de la implementación de diferentes medidas que buscan establecer la máxima pérdida esperada que estarían dispuestos a asumir sus inversionistas o administradores, y a partir de la cual se demandarán recursos de capital. En otras palabras, definir el nivel de riesgo que asumiría la compañía ante la ocurrencia de situaciones extremas y para las cuales las reservas técnicas o provisiones serían insuficientes.

La literatura alrededor del cálculo de las medidas de riesgo financiero ha evolucionado, desde la introducción de la media y la varianza en los años 50 con el modelo de Markowitz, donde el comportamiento de los riesgos (en este caso de los retornos financieros) era sin lugar a duda el principal factor de interés. Sin embargo, el supuesto de normalidad o distribución t-student, para predecir dicho comportamiento, ha conllevado a subestimaciones o sobreestimaciones de eventos extremos causantes de las mayores pérdidas en las compañías. Desde la publicación de Artzner et al (1997, 1999) de los primeros resultados sobre medidas que desafían el paradigma de normalidad de los datos, diversos autores han marcado un nuevo status quo en la medición y modelamiento del comportamiento del riesgo ante situaciones de skewness, leptocurtosis y/o colas anchas, por medio de métodos paramétricos o no paramétricos.

En consecuencia, el desarrollo de la literatura de los años 90 consagró al Value At Risk (VaR), como la medida de riesgo por excelencia. No obstante, existen otras metodologías para estimar eventos extremos tales como el Expected Shortfall (ES) o la Teoría de Valores Extremos (EVT).

Con el objetivo de establecer un umbral a partir del cual un siniestro se pueda considerar como extremo y así realizar una adecuada estimación y gestión del riesgo de suscripción, a continuación, se explican las metodologías que constituyen la base de estudio y análisis del comportamiento histórico de los siniestros incurridos para el caso colombiano:

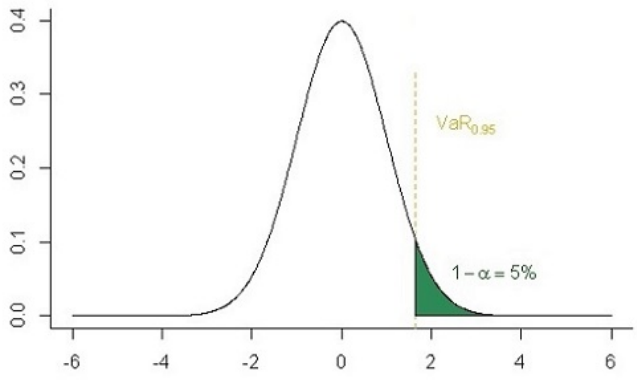
### *Value at risk (VaR) y Expected Shortfall o Conditional VaR (ES - CVaR)*

El VaR se define como "la máxima pérdida esperada en un periodo de tiempo y con un nivel de confianza dado, en condiciones normales de mercado" (Jorion, 2000). En otros términos, dada una cartera $P$, un periodo temporal $T$ y un nivel de probabilidad $Q$, se estima un nivel de pérdidas $L^*$, tal que existe una probabilidad $Q$ de que las pérdidas efectivas $L$, sean iguales o menores que $L^*$ durante el periodo $T$.

A este nivel de pérdidas se le denomina el Valor en Riesgo de una cartera y se expresa de la siguiente manera:

$$VaR_Q = Prob[L^* \geq L] = Q \tag{1}$$

El VaR es la herramienta fundamental en la administración de riesgos financieros (tales como en las tasas de interés o en la tasa representativa del mercado) asociados a la operación de una empresa.




Fuente: Parmentier (2016)

Existen tres tipos de metodologías para la estimación del VaR:

1. **VaR paramétrico:** El cual asume una distribución normal de los datos y puede generar subestimaciones o sobreestimaciones al no capturar el fenómeno de la cola (colas pesadas).
2. **VaR histórico:** Estimado a partir de los datos históricos, utilizando como supuesto fundamental que los eventos futuros tienen la misma dinámica que los eventos pasados.
3. **VaR por simulación de Monte Carlo:** Estimación, paramétrica o no paramétrica, realizada a través de la simulación de los datos y sus posibles trayectorias.

Por otro lado, es aconsejable realizar pruebas de normalidad a los datos para establecer el tipo de metodología a utilizar, así como técnicas estadísticas que permitan disminuir el riesgo de estimación, para lo cual puede recurrirse a la identificación del tipo de cola que presentan los datos. Lo anterior debido a que tanto el VaR, como el ES (que se describe a continuación), se derivan del comportamiento de las colas.

Finalmente, el ES corresponde al promedio de las pérdidas que exceden el valor del VaR, esto es:

$$ES_Q = E(X|X > VaR_Q = Prob[L^* \geq L]) \tag{2}$$

En otras palabras, el ES corresponde al promedio de las pérdidas que exceden la máxima pérdida esperada en un periodo y con un nivel de confianza dados, siendo esta metodología una medida de riesgo mayor al VaR.

### *Test de normalidad*

Para los test de normalidad que involucran las metodologías descritas se van a utilizar los estadísticos de: i) Anderson-Darling, ii) Cramer von Mises, iii) Lilliefors-Kolmogorov-Smirnov y iv) Pearson chi square. Con estos se busca establecer si el comportamiento de los datos corresponde a una distribución normal.

Los estadísticos miden qué tan bien los datos se ajustan a una distribución específica, dado un conjunto de datos y una distribución en particular. Por ejemplo, es posible utilizar el estadístico de Anderson-Darling para determinar si los datos cumplen el supuesto de normalidad para una prueba t.

La hipótesis nula $(H_0)$ y alternativa $(H_1)$ para las pruebas son las siguientes:

$H_0$ : Los datos siguen una distribución especificada (i.e. $X \sim N(\mu, \sigma^2)$).

$H_1$ : Los datos no siguen una distribución especificada (i.e. $X \sim N(\mu, \sigma^2)$).

## *Teoría de Valores Extremos*

La teoría de valores extremos $(EVT)$ tiene dos resultados significativos:

1. Para distribuciones de pérdidas, la distribución de los valores máximos convergerá a una distribución Gumbel, Frechét o Weibull, dependiendo de la cola de dicha distribución, la cual se denomina **Distribución Generalizada de Valor Extremo** $(GEV)$.
2. Para distribuciones de pérdidas, la distribución de excedentes sobre un umbral dado, es una **Distribución Generalizada de Pareto** $(GPD)$.

Dados formalmente por:

### *Distribución generalizada del valor extremo $(GEV)$ (Fisher-Tippett (1928), Gnedenko (1943))*

Sea $X_1, \ldots, X_n$ un conjunto de $n$ variables aleatorias independientes e idénticamente distribuidas (i.i.d). El máximo $X_n = max\{X_1, \ldots, X_n\}$ converge a la siguiente distribución:

$$H_\xi(x) = \begin{cases} exp\{-(1+\xi x)^{-\frac{1}{\xi}}\} & si \quad \xi \neq 0 \\ exp\{-exp^{-x}\} & si \quad \xi = 0 \end{cases} \tag{3}$$

Donde $1 + \xi x > 0$. El parámetro $\xi$ define la forma de la distribución si:

- $\xi > 0$, implica que $H_\xi(x)$ se distribuye Frechét, es decir dicha distribución tiene una cola pesada.
- $\xi = 0$, implica que $H_\xi(x)$ se distribuye Gumbel, es decir dicha distribución tiene una cola media.
- $\xi < 0$, implica que $H_\xi(x)$ se distribuye Weibull, es decir dicha distribución tiene una cola corta.

Este resultado es de gran relevancia, pues sin importar cuál sea la distribución original de las pérdidas, la distribución de los máximos siempre pertenece a alguna de estas tres distribuciones. No obstante, esto no considera el riesgo residual más allá de dichos máximos, por lo cual otros autores introdujeron el segundo resultado que implica estimar la distribución de los valores excedentes luego de un umbral alto.

*Exceso sobre un umbral (Picklands (1975), Balkema-in Haan (1974))*

Sea $X$ una variable aleatoria con una función de distribución $F$, $x_F$ el punto final derecho y un umbral dado $u < x_F$ , $F_u$ es la función de distribución de excesos de $X$ sobre $u$. Donde:

$$F_u(x) = P(X - u \leq x | X > u), \quad x \geq 0 \tag{4}$$

Luego de que $u$ es estimado, la distribución $F_u$ se aproxima a una distribución Generalizada de Pareto $(GPD)$.
Es decir:

$$F_u(x) \approx G_\xi(x), \quad u \to \infty, x \geq 0 \tag{5}$$

Donde:

$$G_\xi(x) = \begin{cases} 1 - (1 + \xi x)^{-\frac{1}{\xi}} & si \quad \xi \neq 0 \\ 1 - exp^{-x} & si \quad \xi = 0 \end{cases} \tag{6}$$

- Cuando $\xi > 0$, se obtiene la distribución Pareto ordinaria.
- Cuando $\xi = 0$, se obtiene la distribución exponencial.
- Cuando $\xi < 0$, se obtiene la distribución Pareto de cola corta.

Este segundo resultado implica que toda distribución tiene un umbral, a partir del cual sus excesos se distribuyen *GPD*. Por ende, la definición de umbral óptimo dependerá de qué tanto se ajuste la distribución de excedentes sobre dicho umbral a la *GPD*.

Para poder estimar el umbral óptimo existen diferentes métodos, algunos de ellos son gráficos y otros computacionales; en este ejercicio se usarán métodos gráficos los cuales permitirán establecer un rango de posibles umbrales, para luego identificar aquel que mejor se ajuste al modelo *GPD*, siendo este último el óptimo para la definición de la pérdida extrema.

### *Elección de umbrales*

Para la selección del umbral, al que se hizo referencia en el apartado anterior, se emplearán el gráfico de exceso de media $(ME - PLOT)$y el gráfico de Hill $(Hill - PLOT)$; estos dos métodos permitirán determinar un rango de umbrales entre los cuales se encontrará el óptimo.

Estos análisis gráficos pueden definirse, en general, como:

- Gráfico de exceso de media $(ME - PLOT)$: donde el umbral se escoge utilizando un gráfico que representa la media de los excedentes versus el umbral,
- Gráfico de Hill $(Hill - PLOT)$: en el cual el umbral se determina graficando el estimador Hill. Este método solo será aplicado si la cola sigue una distribución Frechét (cola gruesa).

*Gráfico de exceso de media (ME-PLOT)*

Este gráfico representa la esperanza de los valores por encima de un umbral u una vez que han superado ese valor, es decir, la media esperada de una función de distribución condicionada $E_{k,n}$.

Dada una variable aleatoria $X = (X_1, \ldots, X_n)$, donde los datos están organizados de mayor a menor $X_{1,n} \geq X_{2,n} \geq, \ldots, \geq X_{n,n}$, si $u = X_{k+1,n}$, entonces $E_{k,n}$ es la media aritmética de los $k$ mayores valores muestrales:

$$E_{k,n} = e_n(X_{k,n}) = \frac{\sum_{i=1}^{k} X_{i,n}}{k} - X_{k+1,n} \quad k = 1, \ldots, n-1 \tag{7}$$

Los puntos del gráfico de exceso de media son $\{(X_{k,n}, e_n(X_{k,n})) : k = 1, \ldots, n\}$, esto es $E_{k,n}$ será la variable dependiente y la variable independiente serán los valores de $u = X_{k+1,n}$.

De acuerdo con el resultado de Picklands y Balkema-in Haan previamente descrito, para un umbral alto dado, la serie de los excedentes converge a una *GPD*, con lo cual es posible elegir el umbral óptimo pues a partir de ese valor se detecta que el gráfico es lineal (Embrechts et al., 2013, Teorema 3.4.13 ).

*Gráfico de Hill (Hill - Plot)*

Este método se basa en el estimador de Hill, el cual se deriva de la estimación de máxima verosimilitud del coeficiente de potencia en la distribución de Pareto. Para realizar la gráfica se usa el hecho de que la función de distribución $F(x)$ de una variable aleatoria $X$ se aproxima a la función de distribución de Pareto dado que excede un umbral $u$.

Dada una variable aleatoria $X = (X_1, \ldots, X_n)$, donde los datos están organizados de mayor a menor, $X_{1,n} \geq X_{2,n} \geq, \ldots, \geq X_{n,n}$ independientes e idénticamente distribuidas (i.i.d). El estimador de Hill para la cola con parámetro $\xi$, usando los $k+1$ estadísticos ordenados, es definido por:

$$\hat{H}_{k,n} = \frac{1}{k}\sum_{i=1}^{k} ln(x_i) - ln(x_{k+1}) = \hat{\xi} \tag{8}$$

Así el gráfico de Hill es definido por el conjunto de puntos: $\{(k, \hat{H}_{k,n}^{-1}) : k = 1, \ldots, n-1\}$. Para este método, el umbral óptimo $u$ se selecciona escogiendo las áreas estables en el gráfico; sin embargo, esta elección no siempre es clara. Por esta razón, se estima en qué rango de umbrales el inverso del estimador se estabiliza, para posteriormente verificar la convergencia a partir de dichos umbrales a una *GPD*.

*L-momentos*

Los L-momentos son medidas de la ubicación, la escala y la forma de las distribuciones de probabilidad o muestras de datos. Se basan en combinaciones lineales de estadísticos de orden. Hosking (1990) y Hosking y Wallis (1997) presentaron exposiciones de la teoría y las relaciones de los L-momentos. Adicionalmente muestran expresiones y algoritmos para estimar los parámetros de las distribuciones más usadas, equiparando los L -momentos de la muestra y la población (el "método de los L -momentos ").

Dada una variable aleatoria $X$ con una función de probabilidad $F$, tal que $E[X] < \infty$, se considera: $\alpha_r = E[X(1-F(X))^r]$, donde los primeros cuatro L-momentos son:

$$\begin{array}{ll} \lambda_1 = \alpha_0 & \text{L-locación = valor esperado} \\ \lambda_2 = \alpha_0 - 2\alpha_1 & \text{L-escala} \\ \lambda_3 = \alpha_0 - 6\alpha_1 + 6\alpha_2 & \text{Tercer L-momento} \\ \lambda_4 = \alpha_0 - 12\alpha_1 + 30\alpha_2 - 20\alpha_3 & \text{Cuarto L-momento} \end{array}$$

Y las razones de los L-momentos, como $\tau_r = \dfrac{\lambda_r}{\lambda_2}, r = 3,4,\ldots$. Estas razones satisfacen que $|\tau_r| < 1$, y dan medidas de la forma de la distribución independientemente de la escala. Por lo tanto, se tiene que:

$$\tau_3 = \frac{\lambda_3}{\lambda_2} \quad \text{L-asimetría}$$
$$\tau_4 = \frac{\lambda_4}{\lambda_2} \quad \text{L-curtosis}$$

Las anteriores razones han sido calculadas para distintas distribuciones de probabilidad. Sin embargo, este documento profundiza particularmente en la distribución generalizada de Pareto $(GPD)$, donde existe una relación particular entre $\tau_3$ y $\tau_4$, dada por:

$$\tau_4 = \tau_3 \frac{1+5\tau_3}{5+\tau_3} \tag{9}$$

Para poder estimar los L-momentos de una muestra ordenada de menor a mayor $x_{1,n} \leq x_{2,n} \leq \cdots \leq x_{n,n}$, se consideran los $\alpha_r$ como combinaciones lineales de los elementos de la siguiente manera:

$$a_r = \frac{1}{n}\sum_{i=1}^{n} \binom{n-i}{r} x_{i,n} \binom{n-1}{r}^{-1} \quad r = 0,1,\ldots,n-1 \tag{10}$$

De donde se construyen los primeros cuatro momentos así:

$$\begin{aligned} l_1 &= a_0 && \text{Media muestral} \\ l_2 &= a_0 - 2a_1 && \text{L-escala muestral} \\ l_3 &= a_0 - 6a_1 + 6a_2 && \text{Tercer L-momento muestral} \\ l_4 &= a_0 - 12a_1 + 30a_2 - 20a_3 && \text{Cuarto L-momento muestral} \end{aligned}$$

Adicionalmente, las razones para la L-asimetría y la L-curtosis vienen dadas por:

$$\begin{aligned} t_3 &= \frac{l_3}{l_2} && \text{L-asimetría muestral} \\ t_4 &= \frac{l_4}{l_2} && \text{L-curtosis muestral} \end{aligned}$$

## *Umbral óptimo*

### *Selección del umbral óptimo metodología L-momentos*

Para la selección del umbral óptimo, se deben definir una serie de posibles candidatos que posteriormente serán evaluados de acuerdo con la distancia euclidiana. El procedimiento es el siguiente:

1. Se define un conjunto de posibles candidatos a umbrales $\{u_i\}_{i=1}^{I}$ como una de las alternativas razonables sería.
   - $I = 10$ cuantiles de la muestra, empezando en el 25 % y dando pasos de 7,5 %.
   - $I = 20$ cuantiles de la muestra, empezando en el 25 % y dando pasos de 3,7 %.
2. Calcular la L-asimetría y la L-curtosis de los excedentes sobre cada candidato a umbral $(t_{3,u_i}, t_{4,u_i})$ y determinar la mínima distancia euclidiana $(d_{u_i})$ entre cada punto y la curva teórica de la distribución generalizada de Pareto $(GPD)$.

$$d_{u_i} = min_{\tau_3} \sqrt{(t_{3,u_i} - \tau_3)^2 + (t_{4,u_i} - g(\tau_3))^2}, \quad i = 1, \ldots, I \tag{11}$$

Donde:

$$g(\tau_3) = \tau_3 \frac{1 + 5\tau_3}{5 + \tau_3} \tag{12}$$

3. El umbral a partir del cual se considera que los datos son aproximadamente una distribución Pareto generalizada, es aquel que cumple con la siguiente condición:

$$u^* = \arg\min_{u_i} \{d_{u_i}\}_{1<i<I} \tag{13}$$

### *Selección del umbral óptimo métodos gráficos*

Con el objetivo de seleccionar el umbral óptimo, se realizará una prueba de bondad de ajuste para determinar a partir de los umbrales escogidos la distribución que se ajusta mejor a una *GPD*. Para esto se realiza la Prueba de Kolmogorov-Smirnov a los datos que exceden el umbral.

*Prueba de Kolmogorov-Smirnov (Chakravarty et al., 1967, pp.392-394).*

Esta prueba se usa para decidir si una muestra proviene de una población con una distribución específica. El test de Kolmogorov-Smirnov es definido por:

$H_0$: Los datos siguen una distribución específica.

$H_1$: Los datos no siguen una distribución específica.

Para aceptar o rechazar la hipótesis nula, el estadístico de la prueba está definido por:

$$D = max_{1\leq i\leq N}(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i)) \tag{14}$$

Donde $F$ es la función de distribución acumulativa teórica que se está testeando.

La hipótesis nula $(H_0)$ con respecto a la forma de distribución se rechaza si el estadístico de prueba $D$ es mayor que el valor crítico obtenido en la tabla de distribución $KS$. En otras palabras, entre más alto sea el $p-valor$ mejor se ajustan los datos a la distribución.

Dado que, para la fijación del umbral a partir del cual se debe considerar un siniestro como extremo, se utiliza la metodología $EVT$, es de esperar que se presenten pocos siniestros incurridos que superen dicho umbral, por esta razón es pertinente aplicar *bootstrapping* a la prueba de bondad de ajuste (Savapandit y Gogoi, 2015, pp.6-7).

# APLICACIÓN DE LA TEORÍA DE VALORES EXTREMOS AL MERCADO DE SEGUROS DE COLOMBIA

## *Datos*

Los datos empleados para el estudio tienen las siguientes características:

1. **Conjunto de datos:** Corresponde a los siniestros incurridos (pagados y/o reservados avisados) avisados en el horizonte de tiempo expresados en pesos constantes de 2018, excluyendo el ramo de terremoto.
2. **Fuente de los datos:** Entidades de seguros generales autorizadas por la SFC.
3. **Horizonte de tiempo:** 11 años comprendidos entre el 1 de enero de 2008 y el 31 de diciembre de 2018.
4. **Actualización por inflación:** Índice de precios al consumidor certificado por el DANE para el periodo de referencia (2008 – 2018).
5. **Unidades:** El valor de los siniestros incurridos se encuentra expresado en miles de pesos.

## *Análisis exploratorio*

Para determinar el comportamiento de los siniestros durante el periodo de estudio y observar a qué distribución se asemejan, se estiman los principales estadísticos de los datos reportados por las aseguradoras.

### *Resumen estadístico del total de siniestros incurridos*

**TABLA 1:** RESUMEN ESTADÍSTICO DEL TOTAL DE SINIESTROS INCURRIDOS.

| Estadístico | Siniestros incurridos |
|---|---|
| N | 12,880,340 |
| Mínimo* | 0.001 |
| Cuartil* | 86.2 |
| Mediana* | 318 |
| Moda* | 0.14 |
| Media* | 4,921 |
| Desv. Estándar* | 1,082,435 |
| Asimetría | 3,418 |
| Curtosis | 12,030,867 |
| Cuartil* | 1,912 |
| Máximo* | 3,819,000,000 |
| $x_{0,99}$* | 46,495 |

∗Cifras en miles

*Resumen estadístico de siniestros incurridos mayores a* $1,000 .*(*$\geq 1$*, en la base de datos)*

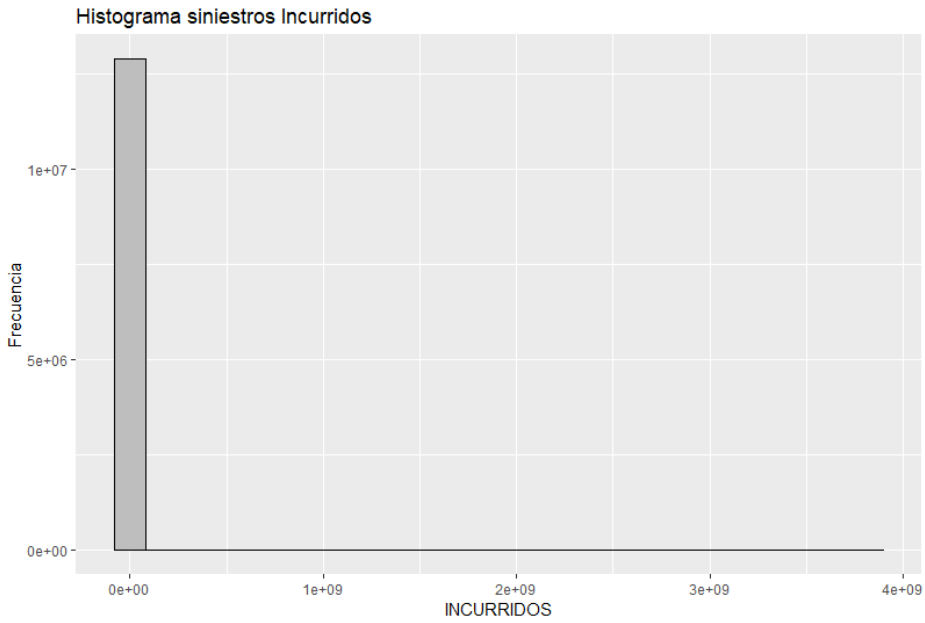
[1].

**TABLA 2:** RESUMEN ESTADÍSTICO DEL TOTAL DE SINIESTROS INCURRIDOS $\geq 1$.

| Estadístico | Siniestros incurridos |
|---|---|
| N | 12,383,355 |
| Mínimo* | 1 |
| Cuartil* | 103 |
| Mediana* | 357 |
| Moda* | 40 |
| Media* | 5,119 |
| Desv. Estándar* | 1,103,941 |
| Asimetría | 3,351 |
| Curtosis | 11,566,675 |
| Cuartil* | 2,071 |
| Máximo* | 3,819,000,000 |
| $x_{0,99}$* | 48,106 |

∗Cifras en miles

*Histograma de frecuencia del total de siniestros incurridos*



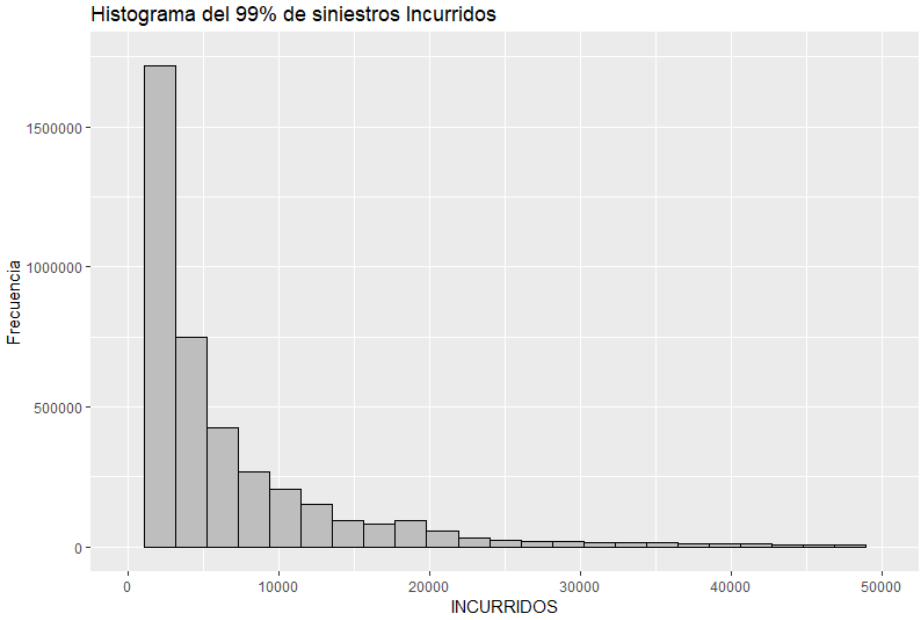

Fuente: Información entidades aseguradoras y cálculos propios.

El gráfico anterior evidencia que existen siniestros de alta cuantía que distorsionan la visualización del histograma. Por tal motivo, se particionan los siniestros en: i) alta cuantía[2] $(1\,\%)$ y ii) media y baja cuantía $(99\,\%)$, con el objetivo de visualizar el comportamiento de los siniestros sin distorsiones, tal y como se muestra a continuación:

[1]Se estiman los estadísticos descriptivos para los siniestros incurridos mayores a $1,000. Lo anterior con el objetivo de evidenciar la influencia de valores pagados de baja cuantía sobre los datos. Sin embargo, se evidencia que los estadísticos no se modifican significativamente y por ende se tomará toda la información disponible.

[2]Este procedimiento sólo se realiza para visualizar el comportamiento de los siniestros incurridos dentro del histograma. Por ende, no se eliminan los siniestros de alta cuantía dentro del análisis y aplicación de las metodologías.

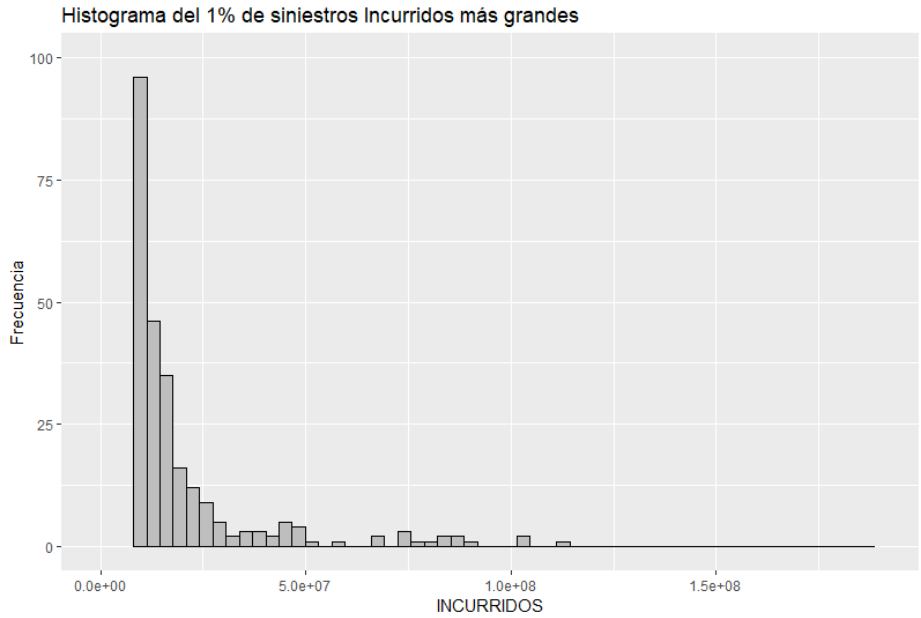
*Histograma de frecuencia del 99% de siniestros incurridos*




Fuente: Información entidades aseguradoras y cálculos propios.

---

*Histograma de frecuencia del 1% de siniestros incurridos más grandes*

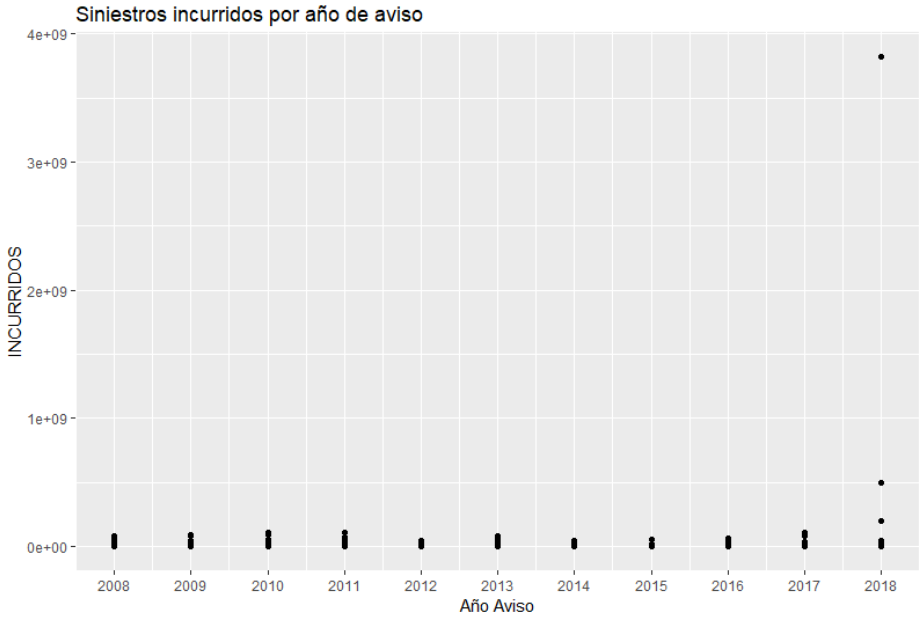



Fuente: Información entidades aseguradoras y cálculos propios.

---

Con relación a los estadísticos observados, en donde la asimetría es mayor que 0, se puede determinar que la función de distribución de los siniestros incurridos tiene una cola a la derecha. En adición el valor de la curtosis, también mayor que cero, muestra una concentración de los datos alrededor de la media, haciendo que la curva de la distribución sea sesgada y considerando que el valor máximo de los siniestros incurridos es de $4 billones, se puede observar una función de distribución con una çola pesada a la derecha ".

*Gráfica de los siniestros incurridos por año de aviso*

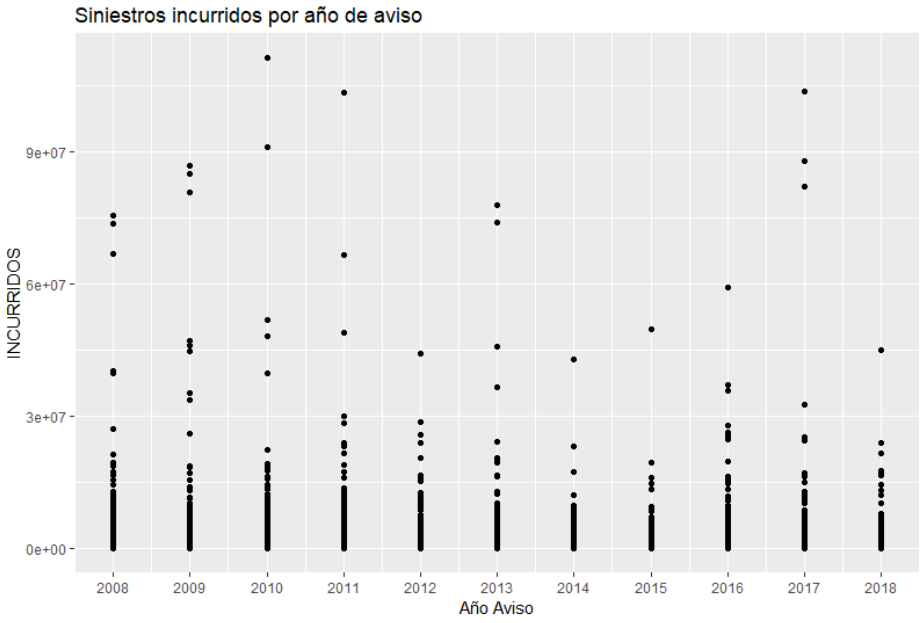



Fuente: Información entidades aseguradoras y cálculos propios.

---

Al graficar los siniestros de acuerdo con su fecha de aviso se observa que para el periodo objeto de análisis, los tres siniestros de mayor valor fueron avisados en 2018, seguidos por un siniestro de $111 mil millones en 2010.

*Gráfica de los siniestros incurridos por año de aviso (excluyendo los tres siniestros incurridos de mayor valor).*




Fuente: Información entidades aseguradoras y cálculos propios.

---

### *Aplicación de la teoría de los L-momentos*

A continuación, se presentan los valores de la L-asimetría y la L-curtosis tanto para la muestra como para la población (curva teórica), junto con la distancia euclidiana entre estos dos.

**TABLA 3:** I=10 CUANTILES DE LA MUESTRA. FUENTE: INFORMACIÓN ENTIDADES ASEGURADORAS Y CÁLCULOS PROPIOS.

| i | Cuantil | Valor | $t_{3,u_i}$ | $t_{4,u_i}$ | $g(\tau_3)$ | $\tau_3$ | $\tau_4)$ | Distancia $d_{u_i}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.250 | 86 | 0.8329 | 0.7135 | 0.7375 | 0.8329 | 0.7375 | 0.0240 |
| 2 | 0.325 | 131 | 0.8257 | 0.7073 | 0.7269 | 0.8257 | 0.7269 | 0.0196 |
| 3 | 0.400 | 189 | 0.8184 | 0.7022 | 0.7162 | 0.8184 | 0.7162 | 0.0140 |
| 4 | 0.475 | 278 | 0.8114 | 0.6988 | 0.7060 | 0.8114 | 0.7060 | 0.0072 |
| 5 | 0.550 | 428 | 0.8057 | 0.6981 | 0.6978 | 0.8057 | 0.6978 | 0.0003 |
| 6 | 0.625 | 700 | 0.8028 | 0.7010 | 0.6936 | 0.8028 | 0.6936 | 0.0074 |
| 7 | 0.700 | 1,262 | 0.8043 | 0.7079 | 0.6959 | 0.8043 | 0.6959 | 0.0120 |
| 8 | 0.775 | 2,353 | 0.8107 | 0.7192 | 0.7051 | 0.8107 | 0.7051 | 0.0142 |
| 9 | 0.850 | 4,411 | 0.8224 | 0.7365 | 0.7221 | 0.8224 | 0.7221 | 0.0143 |
| 10 | 0.925 | 9,784 | 0.8426 | 0.7543 | 0.7517 | 0.8426 | 0.7517 | 0.0025 |

**TABLA 4:** I=20 CUANTILES DE LA MUESTRA. FUENTE: INFORMACIÓN ENTIDADES ASEGURADORAS Y CÁLCULOS PROPIOS.

| i | Cuantil | Valor | $t_{3,u_i}$ | $t_{4,u_i}$ | $g(\tau_3)$ | $\tau_3$ | $\tau_4)$ | Distancia $d_{u_i}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.250 | 86 | 0.8329 | 0.7135 | 0.7375 | 0.8329 | 0.7375 | 0.0240 |
| 2 | 0.288 | 108 | 0.8294 | 0.7103 | 0.7323 | 0.8294 | 0.7323 | 0.0219 |
| 3 | 0.325 | 131 | 0.8257 | 0.7073 | 0.7269 | 0.8257 | 0.7269 | 0.0196 |
| 4 | 0.363 | 157 | 0.8220 | 0.7046 | 0.7215 | 0.8220 | 0.7215 | 0.0169 |
| 5 | 0.400 | 189 | 0.8184 | 0.7022 | 0.7162 | 0.8184 | 0.7162 | 0.0140 |
| 6 | 0.438 | 229 | 0.8148 | 0.7002 | 0.7110 | 0.8148 | 0.7110 | 0.0107 |
| 7 | 0.475 | 278 | 0.8114 | 0.6988 | 0.7060 | 0.8114 | 0.7060 | 0.0072 |
| 8 | 0.513 | 342 | 0.8083 | 0.6981 | 0.7016 | 0.8083 | 0.7016 | 0.0035 |
| 9 | 0.550 | 428 | 0.8057 | 0.6981 | 0.6978 | 0.8057 | 0.6978 | 0.0003 |
| 10 | 0.588 | 539 | 0.8037 | 0.6991 | 0.6950 | 0.8037 | 0.6950 | 0.0041 |
| 11 | 0.625 | 700 | 0.8028 | 0.7010 | 0.6936 | 0.8028 | 0.6936 | 0.0074 |
| 12 | 0.663 | 933 | 0.8030 | 0.7039 | 0.6939 | 0.8030 | 0.6939 | 0.0100 |
| 13 | 0.700 | 1,262 | 0.8043 | 0.7079 | 0.6959 | 0.8043 | 0.6959 | 0.0120 |
| 14 | 0.738 | 1,721 | 0.8070 | 0.7130 | 0.6997 | 0.8070 | 0.6997 | 0.0133 |
| 15 | 0.775 | 2,353 | 0.8107 | 0.7192 | 0.7051 | 0.8107 | 0.7051 | 0.0142 |
| 16 | 0.813 | 3,193 | 0.8156 | 0.7270 | 0.7122 | 0.8156 | 0.7122 | 0.0148 |
| 17 | 0.850 | 4,411 | 0.8224 | 0.7365 | 0.7221 | 0.8224 | 0.7221 | 0.0143 |
| 18 | 0.888 | 6,318 | 0.8319 | 0.7470 | 0.7360 | 0.8319 | 0.7360 | 0.0110 |
| 19 | 0.925 | 9,784 | 0.8426 | 0.7543 | 0.7517 | 0.8426 | 0.7517 | 0.0025 |
| 20 | 0.963 | 17,424 | 0.8444 | 0.7479 | 0.7545 | 0.8444 | 0.7545 | 0.0066 |

Según las tablas anteriores se encontró lo siguiente:

- Para I=10, el umbral óptimo es $u^* = 428$.
- Para I=20, el umbral óptimo es $u^* = 428$.

Para las dos particiones el umbral óptimo sería $428,000. Sin embargo, este último no puede considerarse como siniestro extremo debido a que: i) el elevado número de valores excedentes, hace que estos dejen de ser siniestros de baja frecuencia, en este caso cercano a los 5.8 millones de siniestros (45 % de la población) y ii) es un valor de baja severidad ya que se encuentra en el percentil 55. En otras palabras, este valor contradice la definición de extremo.

**TABLA 5:** FUENTE: INFORMACIÓN ENTIDADES ASEGURADORAS Y CÁLCULOS PROPIOS.

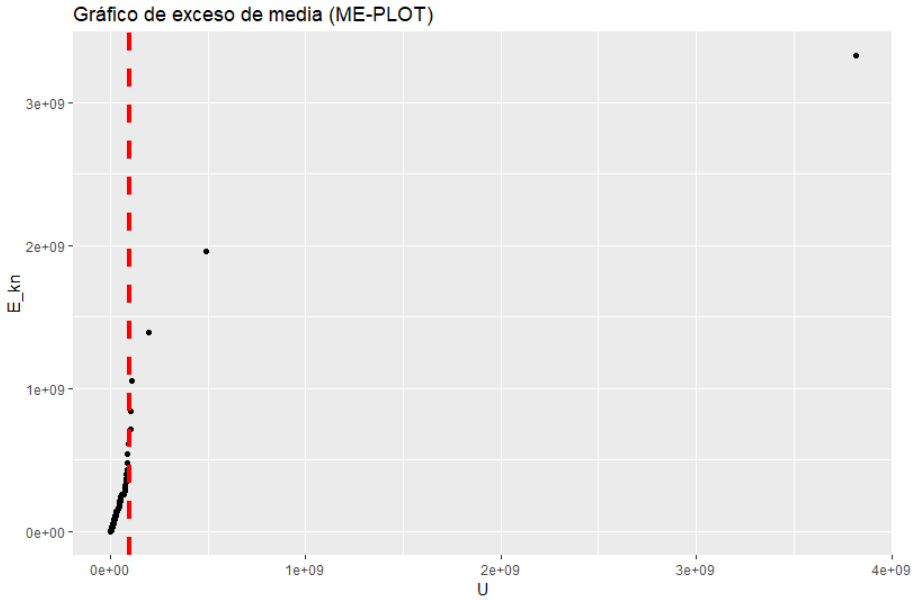
| Límite $(u)$ | Percentil | Excedentes |
|---|---|---|
| 428 | 55 | 5,796,153 |

### *Análisis de las colas*

De acuerdo con el marco conceptual, para el análisis de colas pesadas se emplean los siguientes métodos gráficos.

### *Gráfico de exceso de media (ME-PLOT)*

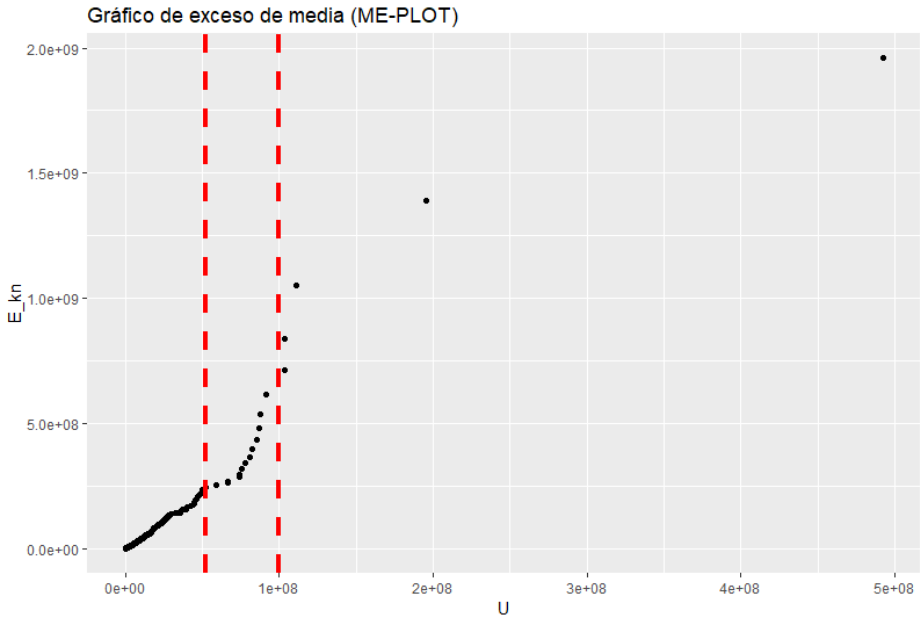
La representación de la función de exceso de media para el periodo de estudio es la siguiente:




Fuente: Información entidades aseguradoras y cálculos propios.

---

En la gráfica anterior se observa que a partir de $u = 1e+08$, la razón de cambio de la pendiente es mayor que en los valores anteriores. A continuación, se incluye la gráfica ME-PLOT excluyendo el dato de mayor valor, para comprobar el cambio de la pendiente de la gráfica.




Fuente: Información entidades aseguradoras y cálculos propios.

---

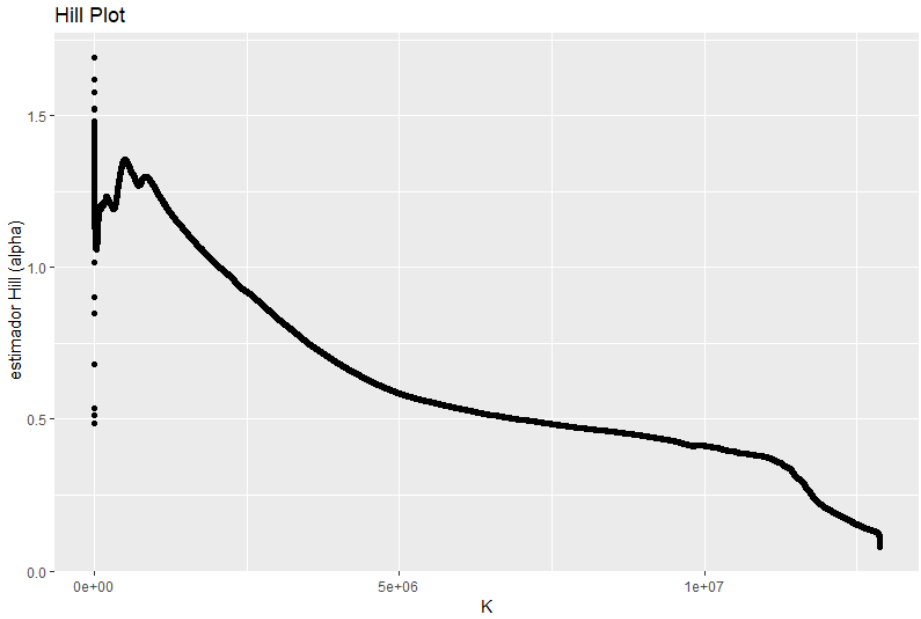
El gráfico de ME-PLOT sin incluir el mayor valor permite observar dos posibles límites $u$ en los cuales la pendiente del gráfico cambia ampliamente en comparación a los demás valores. Donde:

**TABLA 6:** FUENTE: INFORMACIÓN ENTIDADES ASEGURADORAS Y CÁLCULOS PROPIOS.

| Límite ($u$) | Excedentes |
|---|---|
| 103,851,053 | 4 |
| 59,261,306 | 18 |

### *Hill Plot*

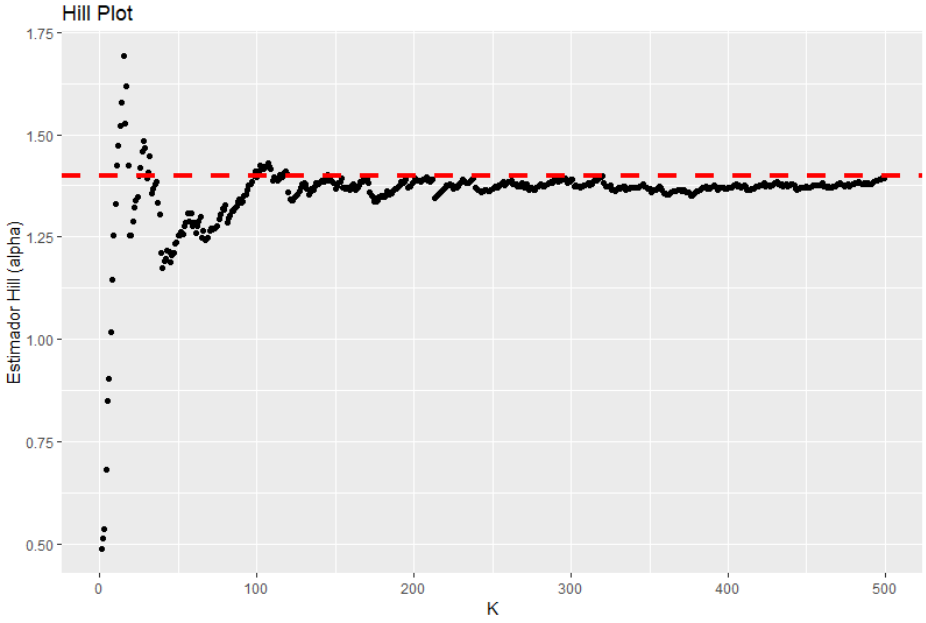
Empleando la base de datos del estudio, la representación del gráfico de Hill es la siguiente, donde los puntos son $\{(k, \hat{H}_{k,n}^{-1}) : k = 1, \ldots, n-1\}$, donde k representa los datos ordenados de mayor a menor y , $\hat{H}_{k,n}^{-1}$ es el inverso del estimador de Hill.




Fuente: Información entidades aseguradoras y cálculos propios.

---

El estimador de Hill es altamente inestable para los valores más altos de $u$, siguiendo por una región amplia de valores decrecientes. Para determinar el inicio del umbral $u$ se selecciona la región donde los estimadores son aproximadamente constantes en un rango. Para ello se realiza un acercamiento en el gráfico para poder determinar dicho umbral.




Fuente: Información entidades aseguradoras y cálculos propios.

Los posibles umbrales son:

**TABLA 7:** FUENTE: INFORMACIÓN ENTIDADES ASEGURADORAS Y CÁLCULOS PROPIOS.

| Límite ($u$) | Excedentes | Alpha Pareto |
|---|---|---|
| 59,261,306 | 18 | 1.43 |
| 45,027,732 | 26 | 1.42 |
| 16,074,811 | 100 | 1.40 |

## SELECCIÓN DE UMBRAL Y MODELO DE EXCESO

- Se asume que la distribución subyacente se comporta como una distribución de Pareto.
- Se estiman los parámetros por el método de máxima verosimilitud (Estimador de Hill).
- Se comprueba la bondad de ajuste de la distribución de Pareto por dos métodos:

### *Prueba Kolmogov-Smirnov*

El umbral óptimo depende del $p-valor$ más alto, que corresponde al umbral más robusto. La siguiente tabla resume los resultados de la prueba, que indica que el umbral óptimo es 103.8 mil millones cuyo p-valor es el más alto.

**TABLA 8:** FUENTE: INFORMACIÓN ENTIDADES ASEGURADORAS Y CÁLCULOS PROPIOS.

| Límite $(u)$ | Excedentes | Alpha Pareto | $P-valor$ | Tamaño cola |
|---|---|---|---|---|
| **103,851,053** | **4** | **0.68** | **0.79** | **0.00003 %** |
| 59,261,306 | 18 | 1.43 | 0.04 | 0,00014 % |
| 49,046,251 | 21 | 1.29 | 0.12 | 0,00016 % |
| 45,027,732 | 26 | 1.42 | 0.25 | 0,00020 % |
| 32,507,732 | 38 | 1.31 | 0.44 | 0,00030 % |
| 25,764,856 | 47 | 1.21 | 0.43 | 0,00036 % |
| 16,074,811 | 100 | 1.40 | 0.25 | 0,00078 % |

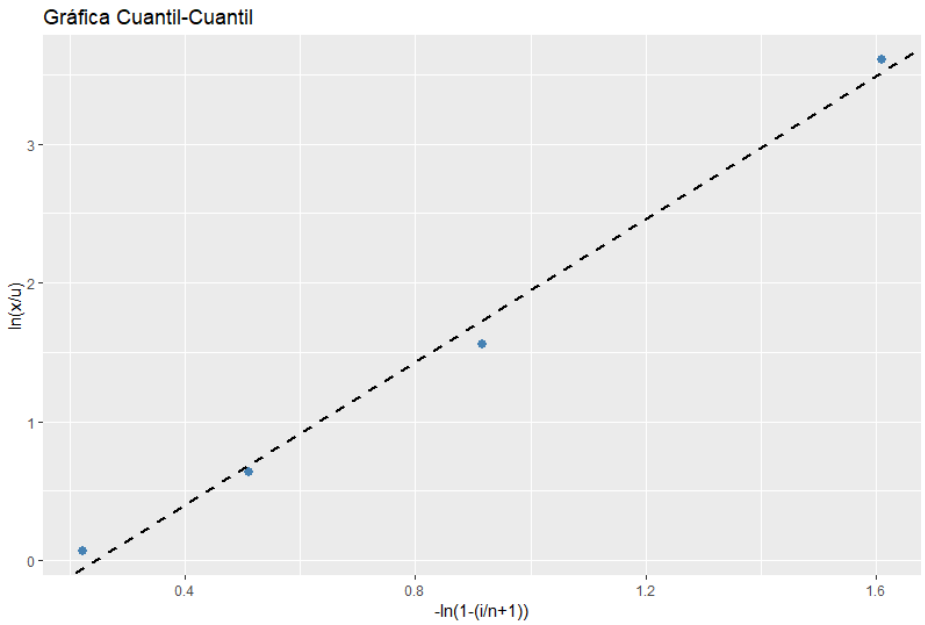
### *Gráfica de cuantil-cuantil Pareto*

Tomando el umbral escogido, se realiza el gráfico de cuantil-cuantil, para determinar si los datos se ajustan a la distribución de Pareto. La prueba de hipótesis es la siguiente:

$H_0$: La distribución de Pareto modela adecuadamente los datos de siniestros.

$H_1$: La distribución de Pareto no modela adecuadamente los datos de siniestros.

Los puntos de la gráfica son: $\{(-ln(1-\frac{i}{n+1}), ln(\frac{X_i}{u})) : i = 1,\ldots,n\}$




Fuente: Información entidades aseguradoras y cálculos propios.

De acuerdo con el gráfico y al valor del estimador $R^2 = 99\,\%$, se determina que la Distribución Generalizada de Pareto modela adecuadamente los siniestros de la cola, es decir, la hipótesis nula no es rechazada.

## CONCLUSIONES

El análisis exploratorio de los datos y la estimación de una medida de riesgo para el mercado de seguros generales de Colombia, evidenció que la distribución de los siniestros incurridos entre 2008 y 2018 presenta colas pesadas. Por tal motivo, las estimaciones realizadas a través del VaR y el ES no tienen en cuenta el comportamiento de los siniestros en las colas, siendo necesario recurrir a la EVT para encontrar umbrales óptimos.

De otro lado, aplicando la metodología de los L-momentos se encuentra que el umbral óptimo para la escogencia del valor extremo no refleja la naturaleza de dicho valor, esto debido a que los cuantiles a partir de los cuales se escoge el conjunto $\{u_i\}_{i=1}^{I}$ de posibles umbrales abarca valores pequeños, que en este caso da como resultado un valor con gran número de excedentes. En consecuencia, métodos gráficos tales como el exceso de medias y Hill plot arrojan una elección consistente.

Para la selección del umbral óptimo, se realizaron pruebas de bondad de ajuste de la distribución generalizada de Pareto a los umbrales determinados a partir de los gráficos, los cuales oscilan entre $16 mil millones y $104 mil millones. La prueba determinó que un siniestro extremo para el caso colombiano sería aquel cuyo monto exceda $104 mil millones; siendo 4 el número de siniestros que superan este umbral en el horizonte de tiempo analizado. En adición, la prueba de hipótesis realizada para determinar la distribución de la cola indicó que ésta se distribuye Pareto.

Por último, se sugiere que el valor determinado en el presente documento sea actualizado anualmente con el Índice de Precios al Consumidor (IPC) certificado por el DANE, con el objetivo de reconocer cambios en los riesgos que asumen las entidades aseguradoras.

## REFERENCIAS

[1] Bensalah, Y. et al. (2000). *Steps in applying extreme value theory to finance: a review*. Citeseer.

[2] Chakravarty, I. M., Roy, J., y Laha, R. G. (1967). "Handbook of methods of applied statistics".

[3] Embrechts, P., Klüppelberg, C., y Mikosch, T. (2013). *Modelling extremal events: for insurance and finance*, volumen 33. Springer Science & Business Media.

[4] Emil Julius, G. (1958). *Statistics of Extremes*. Dover Publications.

[5] Ghosh, S. y Resnick, S. (2010). "A discussion on mean excess plots". *Stochastic Processes and their Applications*, 120(8):1492–1517.

[6] Jorion, P. (2000). "Value at risk: the new benchmark for managing".

[7] Lomba, J. S. y Alves, M. I. F. (2020). "L-moments for automatic threshold selection in extreme value analysis". *Stochastic Environmental Research and Risk Assessment*, pp. 1–27.

[8] Longin, F. (2016). *Extreme events in finance: A handbook of extreme value theory and its applications*. John Wiley & Sons.

[9] Parmentier, V. (2016). *Método de valores extremos aplicado al riesgo operacional*.

[10] Savapandit, M. R. R. y Gogoi, B. (2015). "Bootstrap and other tests for goodness of fit." *Scientiae Mathematicae Japonicae*, 78(4-Special):99–110.

[11] Serra Mochales, I. (2014). *Modelos estadísticos para valores extremos y aplicaciones Statistical models for tails and applications*. Universitat Autònoma de Barcelona,.

[12] Vandewalle, B., Beirlant, J., Christmann, A., y Hubert, M. (2007). "A robust estimator for the tail index of pareto-type distributions". *Computational Statistics & Data Analysis*, 51(12):6252–6268.

[13] Yang, X., Zhang, J., y Ren, W.-X. (2018). "Threshold selection for extreme value estimation of vehicle load effect on bridges". *International journal of distributed sensor networks*, 14(2):1550147718757698.