Proceso de reclamaciones en una aseguradora: Estimaciones y simulaciones de Severidad total

Bernardo Mondragón Brozon, Karen Delgado Curiel, Diego Gonzalez Garcia-Santoyo

Diciembre 10, 2018

El modelo colectivo

El negocio de seguros está sujeto a dos tipos esencialmente diferentes de riesgo, riesgos comerciales y riesgos de seguro. Hay dos puntos de vista desde los cuales se puede considerar la teoría del riesgo, la colectiva y la individual o clásica.

En la teoría del riesgo colectivo, se busca investigar directamente la empresa en su conjunto. El interés primario se centra no en las ganancias, pérdidas o reclamaciones de pólizas individuales, sino en el monto de las reclamaciones totales o la ganancia total que surja de todas las pólizas en la cartera considerada.

La teoría del riesgo colectivo considera dos problemas principales: encontrar las funciones de distribución de la ganancia total o el monto total de las reclamaciones en una cartera o empresa de riesgo, y encontrar la probabilidad de que la reserva de riesgo de una empresa se agote.

En este trabajo se ilustran los métodos que sigue una compañía aseguradora para obtener la distribución de la severidad total a la que está expuesta a lo largo de un año. Primero, se supondrá una distribución para la frecuencia con la que llegan las reclamaciones y una distribución para cada monto reclamado. De esta manera se pondrán obtener las probabilidades teóricas con las que la compañía estará egresando cantidades anualmente. Después, mediante técnicas de muestreo, se realizarán simulaciones de estas cantidades para obtener distribuciones empíricas a las cuales se les realizarán pruebas de bondad de ajuste para comprobar que efectivamente los datos observados provienen de la distribución propuesta al principio.

En la práctica, los actuarios no proponen una distribución para la frecuencia y la severidad, sino que ocupan la información de las bases de datos y realizan pruebas para determinar la distribución de los datos que se están observando. Una vez determinada las distribuciones, realizan las simulaciones bajo diferentes escenarios para analizar el impacto que tienen estos eventos en la situación financiera de la compañía. Cuando no se cuentan con bases de datos, entonces se procede suponiendo distribuciones y ajustándolas bayesianamente conforme se observan más datos. Posteriormente, para calcular probabilidades sobre estas distribuciones se realizan las simulaciones. En este trabajo no se realizará ningun analisis bayesiano sobre las distribuciones.

1.1 La severidad total

En el proceso de reclamaciones de una aseguradora, la severidad total, que es la cantidad total de dinero que la aseguradora terminará pagando en el periodo de estudio, esta dada por la siguiente cantidad:

$$S = x_1 + x_2 + \cdots + x_N$$

En donde los riesgos $\{x_i\}_{i=1\ldots N}$ son variables aleatorias independientes e idénticamente distribuidas que indican los montos de las reclamaciones. Se puede suponer que x_i sigue una distribución exponencial con media $1/\lambda_{exp} = 10^4$ para toda i. Como la llegada de una reclamación a la aseguradora por parte de un asegurado es un suceso extraño, podemos suponer que el proceso de llegada de reclamaciones a la aseguradora es un proceso de Poisson, en donde el número N de reclamaciones que llegan a la aseguradora dentro de un año es variable aleatoria que tiene una distribución Poisson con media $\lambda_{Po} = 0.1$ independiente de las x_i's, de manera que a lo largo del año se observará en promedio 0.1 reclamaciones.

Por el teorema de probabilidad total, se tiene que

$$Pr\{S \leq s | N = n\} = \frac{Pr\{S \leq s, N = n\}}{Pr\{N = n\}}$$

$$\Rightarrow \quad Pr\{S \leq s, N = n\} = Pr\{N = n\} Pr\{S \leq s | N = n\}.$$

Por lo tanto se tiene que

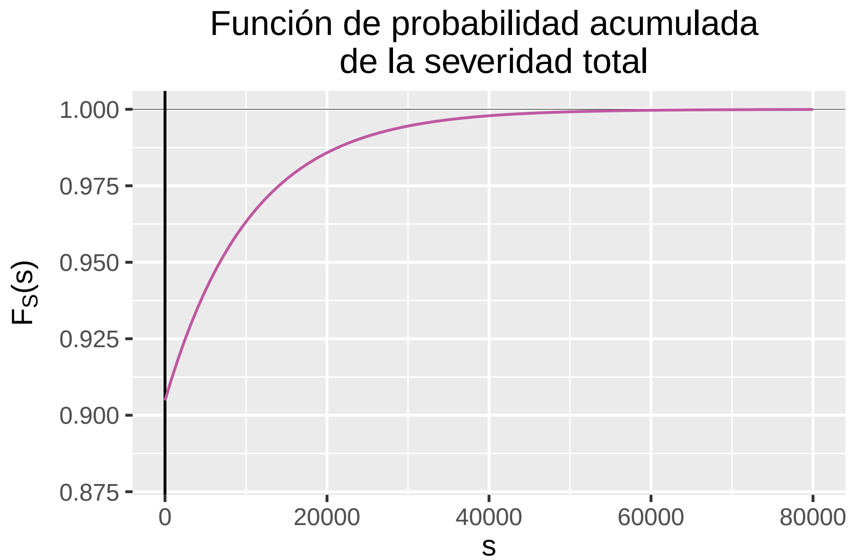
$$\begin{aligned} F_S(s) &= \sum_{n=0}^{\infty} Pr\{S \leq s, N = n\} \\ &= \sum_{n=0}^{\infty} Pr\{N = n\} Pr\{S \leq s | N = n\} \\ &= \sum_{n=0}^{\infty} Pr\{N = n\} Pr\{\sum_{i=n}^{N} x_i \leq s | N = n\} \\ &= \sum_{n=0}^{\infty} Pr\{N = n\} F_X^{*n}(s). \end{aligned}$$

En donde $F_X^{*n}(s)$ es la convolución de las variables aleatorias $\{x_i\}_{i=1\ldots N}$, o bien, la función de distribución de probabilidad acumulada de $\sum_{i=n}^{N} x_i$ con $x_i \sim exp(\lambda_{exp} = 1/10^4)$. Sean $M_{\sum_{i=n}^{N} x_i}(t)$ y $M_{x_1}(t)$ las funciones generadoras de momentos de las variables aleatorias $\sum_{i=n}^{N} x_i$ y x_1 respectivamente, entonces, sin pérdida de generalidad, se tiene que

$$M_{\sum_{i=n}^{N} x_i}(t) = \prod_{i=1}^{n} M_{x_1}(t) = \prod_{i=1}^{n} \frac{\lambda_{exp}}{\lambda_{exp} + t} = \left(\frac{\lambda_{exp}}{\lambda_{exp} + t}\right)^n = \left(1 - \frac{t}{\lambda_{exp}}\right)^{-n}.$$

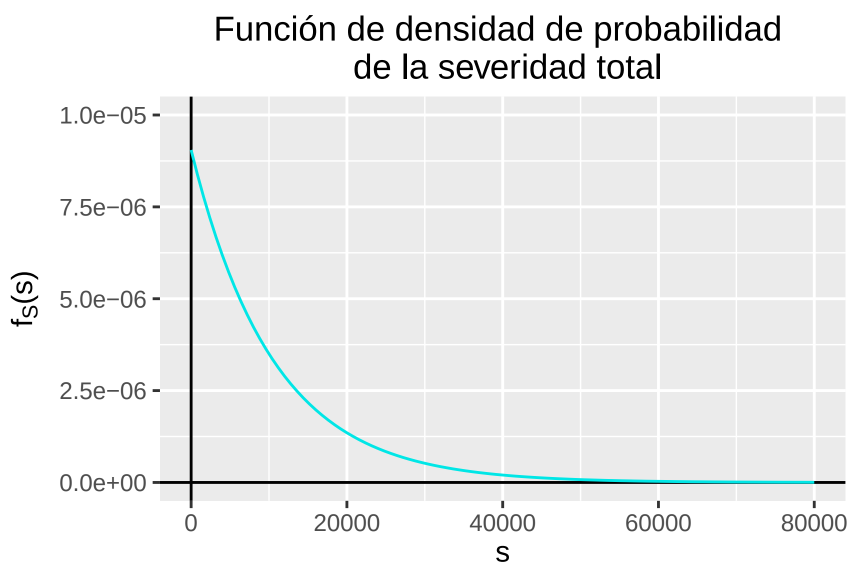
Lo cual corresponde a la funcion generadora de momentos de una variable aleatoria distribuida Gamma con parámetros de $\alpha = n$ y $\beta = \lambda_{exp}$. Ademas, la distribución de la frecuencia es Poisson con parametros $\lambda_{Po} = 0.1$, entonces se sigue que

$$F_S(s) = \sum_{n=0}^{\infty} \frac{\lambda_{Po}^n e^{-\lambda_{Po}}}{n!} \int_0^s \frac{\lambda_{exp}^n}{\Gamma(n)} t^{n-1} e^{-\lambda_{exp} t} \, \mathrm{d}t$$



La funcion de densidad de probabilidad de la severidad total esta dada por

$$\begin{aligned} f_S(s) &= \frac{d}{ds} F_S(s) \\ &= \frac{d}{ds} \sum_{n=0}^{\infty} \frac{\lambda_{Po}^n e^{-\lambda_{Po}}}{n!} \int_0^s \frac{\lambda_{exp}^n}{\Gamma(n)} t^{n-1} e^{-\lambda_{exp} t} \, \mathrm{d}t \\ &= \sum_{n=0}^{\infty} \frac{\lambda_{Po}^n e^{-\lambda_{Po}}}{n!} \frac{d}{ds} \int_0^s \frac{\lambda_{exp}^n}{\Gamma(n)} t^{n-1} e^{-\lambda_{exp} t} \, \mathrm{d}t \\ &= \sum_{n=0}^{\infty} \frac{\lambda_{Po}^n e^{-\lambda_{Po}}}{n!} \frac{\lambda_{exp}^n}{\Gamma(n)} s^{n-1} e^{-\lambda_{exp} s} \end{aligned}$$



A continuación se presenta una tabla que indica la probabilidad que se acumula en ciertos puntos de la distribución de la severidad total:

s	F(s)
0	0.9048374
1	0.9048465
50	0.9052887
100	0.9057377
1000	0.9134693
10000	0.9632416
20000	0.9858116
40000	0.9978908
50000	0.9991875

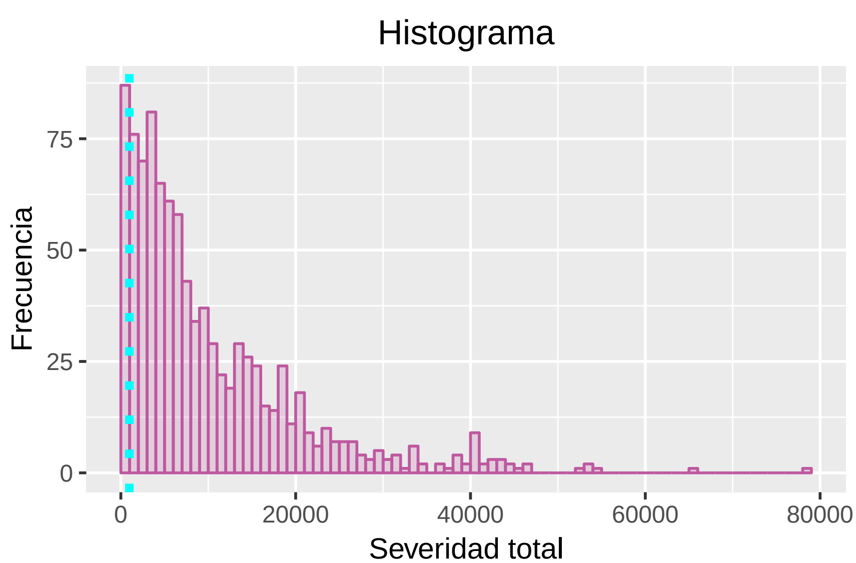
Como se puede observar, casi toda la probabilidad se acumula en valores muy pequeños para la severidad total.

1.2 Simulación

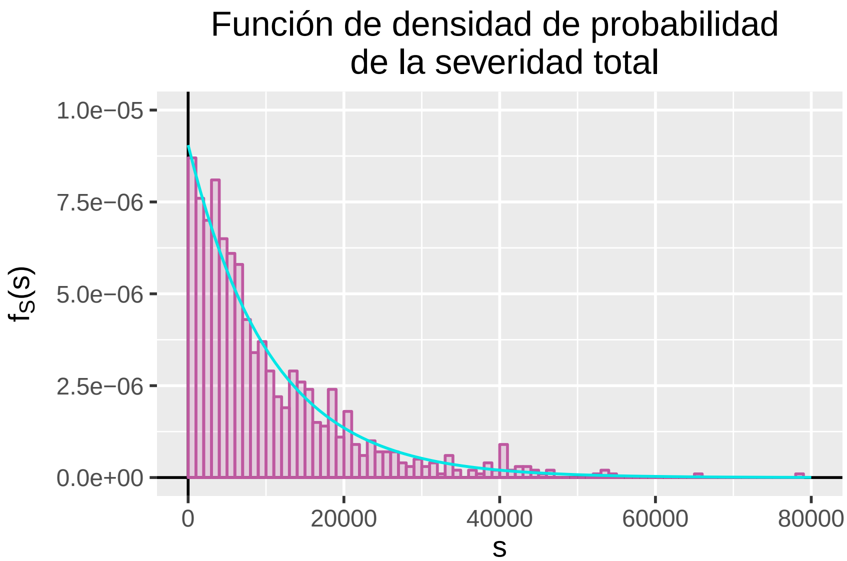
Para hacer simulaciones de la severidad total, primero hay que simular el número N de siniestros ocurridos, el cual proviene de la distribución de Poisson con media $\lambda_{Po} = 0.1$. Una vez simulado el número N de siniestros, se simulan N siniestros, que son variables aleatorias exponenciales con parámetros $\lambda_{exp} = 1/10^4$. Después de obtener los N siniestros, se suman y de esta manera se obtiene un primer valor de la severidad total. Para concluir con la simulación, se repite el proceso anterior 10000 veces. A continuación se muestran los primeros 100 valores obtenidos de la simulacion:

0.000	0	0.000	3970.997	0	0.00	0	0	0	0.000	0	16429.20
0.000	0	0.000	0.000	0	23687.38	0	0	0	0.000	0	0.00
0.000	0	9991.035	0.000	0	0.00	0	0	0	0.000	0	0.00
0.000	0	3805.708	0.000	0	0.00	0	0	0	0.000	0	0.00
0.000	0	0.000	0.000	0	0.00	0	0	0	0.000	0	0.00
0.000	0	0.000	0.000	0	0.00	0	0	0	8190.671	0	0.00
0.000	0	0.000	8896.043	0	0.00	0	0	0	0.000	0	0.00
41315.282	0	0.000	0.000	0	0.00	0	0	0	0.000	0	0.00
1058.154	0	0.000	0.000	0	0.00	0	0	0	0.000	0	41315.28

El histograma debe parecerse a la función de densidad de probabilidad (la linea punteada indica la seveidad promedio):



En la siguiente gráfica se puede apreciar que, en efecto, el histograma se parece a la función de densidad de probabilidad (nótese el cambio de escala en el eje y):



1.3 Comparación de los valores reales con los estimados

Se vio que la severidad total, que sigue uns distribucion de Poisson Compuesta, está dada por

$$S = \sum_{i=1}^{N} x_i.$$

Entonces el valor esperado de la severidad total está dado por

$$\begin{aligned} E(S) &= E\left(\sum_{i=1}^{N} x_i\right) = E(E(S|N=n)) = E\left(\sum_{i=1}^{n} x_i \middle| N=n\right) = E(NE(x_1)) \\ &= E\left(N\left(\frac{1}{\lambda_{exp}}\right)\right) = \frac{1}{\lambda_{exp}E(N)} = \frac{\lambda_{exp}}{\lambda_{Po}} = \frac{10000}{0.1} = 1000 \end{aligned}$$

y su varianza está dada por

$$\begin{aligned} Var(S) &= Var(E(S|N)) + E(Var(S|N)) = Var(NE(x_1)) + E(NVar(x_1)) \\ &= \frac{1}{\lambda_{exp}}Var(N) + \frac{1}{\lambda_{exp}^2}E(N) = \frac{\lambda_{Po}}{\lambda_{exp}} + \frac{\lambda_{Po}}{\lambda_{exp}^2} = (10000)(0.1) + (10000)^2(0.1) = 10001000, \end{aligned}$$

de manera que su desviacion estandar es la siguinete:

$$\sqrt{Var(S)} = \sqrt{10001000} = 3162.43577.$$

La función generadora de momentos de S está dada por

$$M_S(t) = E\left(e^{N\ln\left(e^{x_1 t}\right)}\right) = M_N(\ln(x_1 t)) = e^{\lambda_{Po}\left(\frac{\lambda_{exp}}{\lambda_{exp}+t}-1\right)}$$

se sigue que el sesgo de la distribucion es positivo y está dado por

$$E\left((S-E(S))^3\right) = \lambda_{Po}E(x_1^3) = \lambda_{Po}M_{x_1}^{(3)}(0) = \lambda_{Po}\frac{6}{\lambda_{exp}^3} = (0.1)(6)(10000)^3 = 6\times 10^{11},$$

y el coeficiente de Kurtosis es

$$\frac{E(S^4)}{Var(S)^2} = \frac{M_s^{(4)}(0)}{\left(\frac{\lambda_{Po}}{\lambda_{exp}} + \frac{\lambda_{Po}}{\lambda_{exp}^2}\right)^2} = \frac{\frac{\lambda_{Po}^4+12\lambda_{Po}^3+36\lambda_{Po}^2+24\lambda_{Po}}{\lambda_{exp}^4}}{10001000^2} = \frac{\frac{(0.1)^4+12(0.1)^3+36(0.1)^2+24(0.1)}{(1/10000)^4}}{10001000^2} = 277.1545.$$

Con valores simulados, se tienen las siguientes estimaciones para la severidad total

Estimaciones por Jack Knife

```
## [1] 4416.58
```

```
## [1] 19506181
```

1.4 Aproximación de la distribución de la severidad total

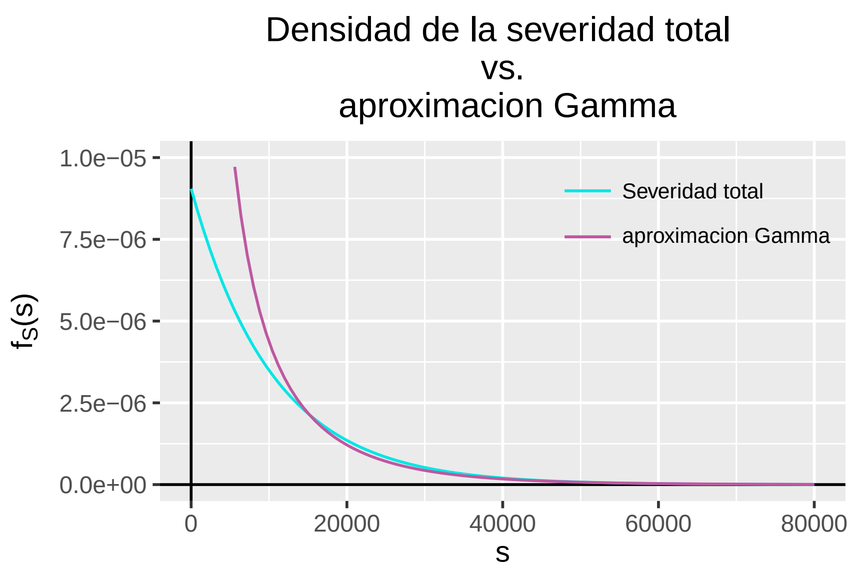
La función de distribución Poisson Compuesta puede ser aproximada mediante una distribución Gamma trasladada con los siguientes parámetros:

$$\alpha = \frac{4\lambda_{Po}E\left(x_1^2\right)^3}{E\left(x_1^3\right)^2} = \frac{4\lambda_{Po}\left(\frac{2}{\lambda_{exp}^2}\right)^3}{\left(\frac{6}{\lambda_{exp}^3}\right)^2} = \frac{4}{45}, \quad \beta = \frac{2E\left(x_1^2\right)}{E\left(x_1^3\right)} = \frac{2\left(\frac{2}{\lambda_{exp}^2}\right)}{\left(\frac{6}{\lambda_{exp}^3}\right)} = \frac{1}{15000}$$

y un desplazamiento

$$x_0 = \lambda_{Po}E\left(x_1\right) - \frac{2\lambda_{Po}E(x_1^2)^2}{E(x_1^3)} = \lambda_{Po}\left(\frac{1}{\lambda_{exp}}\right) - \frac{2\lambda_{Po}\left(\frac{2}{\lambda_{exp}^2}\right)^2}{\left(\frac{6}{\lambda_{exp}^3}\right)} = -\frac{1000}{3}$$

Observe que la aproximación es buena para valores distantes de la media de la distribución Poisson Compuesta:

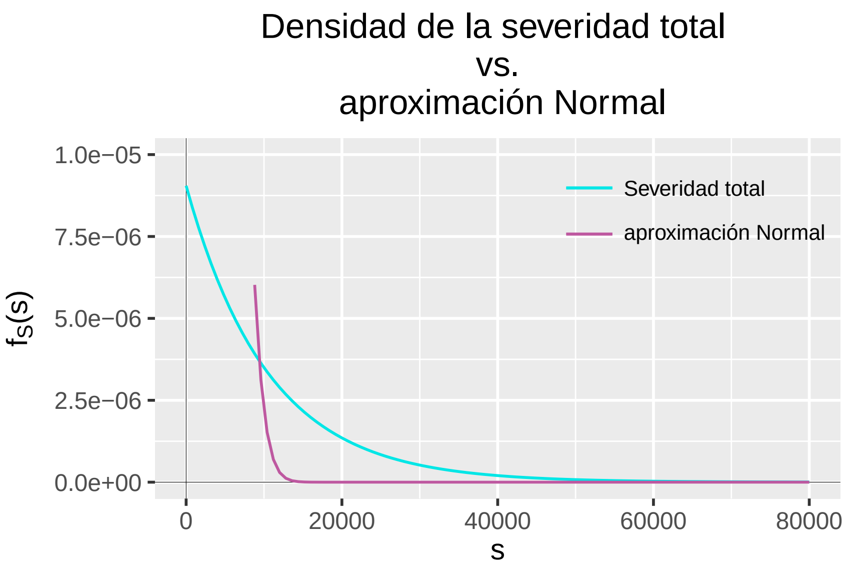


Para valores cercanos a la media es mas conveniente aproximar la distribucion Poisson Compuesta con una distribucion Normal con los siguientes parametros:

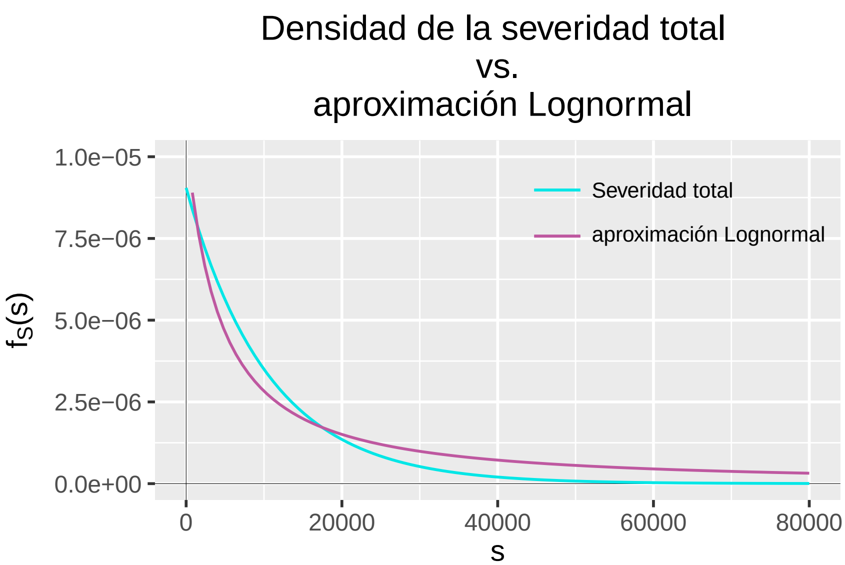
$$\mu = E(N)E(X_1) = \frac{\lambda_{Po}}{\lambda_{exp}} = (0.1)(10000) = 1000,$$

$$\sigma^2 = E(x_1)Var(N) + Var(x_1)E(N) = \frac{\lambda_{Po}}{\lambda_{exp}} + \frac{\lambda_{Po}}{\lambda_{exp}^2} = (10000)(0.1) + (10000)^2(0.1) = 10001000$$

Sin embargo, esto solo funciona para valores grandes de λ_{Po}. En el siguinete gráfico se muestra que la aproximación Normal para valores alrededor de la media es bastante mala, pues como λ_{Po} es pequeña la media de la suma de variables aleatorias no corverge a la media de la aproximación Normal:



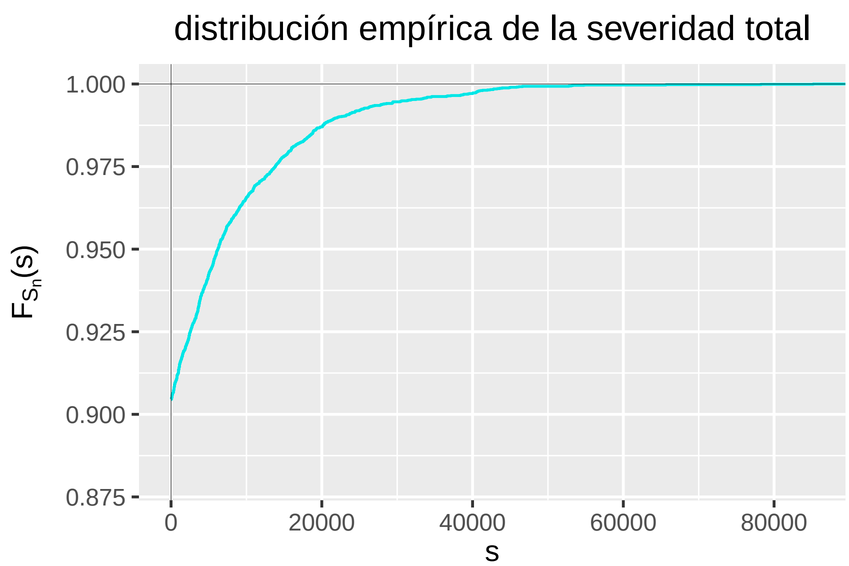
En este caso, la media de la distribución de la frecuencia de los siniestros en un año es un valor muy pequeño dado por $\lambda_{Po} = 0.1$, entonces, lo más recomendable para estimar percentiles alrededor de la media, es aproximar la función de distribución mediante una distribución lognormal:



Esta proximación Lognormal tiene un corrimiento igual 5000, una media de 4 y una desviación estándar de 5.

1.5 Distribucion empirica

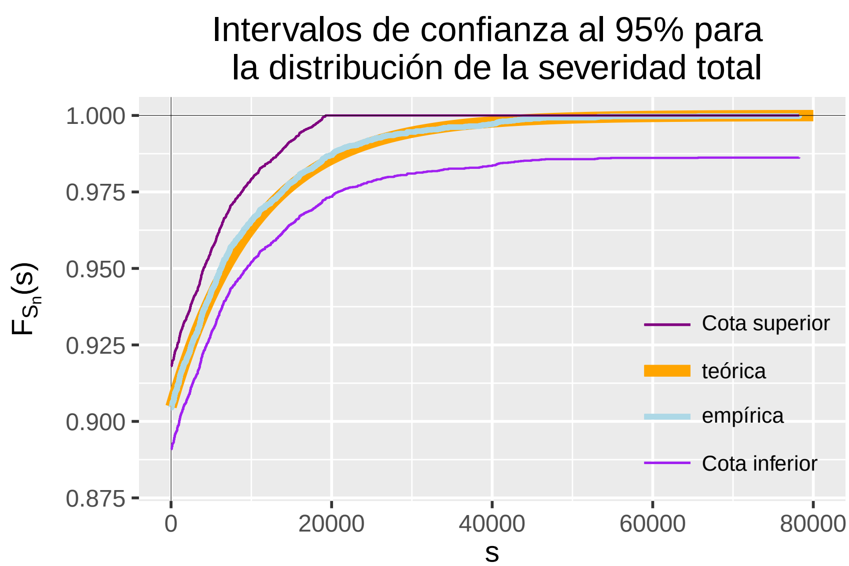
La distribución empírica de la severidad total está dada por



Entonces, un intervalo de confianza a un nivel de $(1-\alpha)100\%$ para distribución de la serveridad total $F_S(s)$ está dado por

$$\left(F_{s_n}(s) - \sqrt{\frac{\ln\frac{2}{\alpha}}{2n}}, F_{s_n}(s) + \sqrt{\frac{\ln\frac{2}{\alpha}}{2n}} \right)$$

Si se construyen los intervalos de confianza al 95% para $F_S(s)$ utilizando la distribución empírica se tiene lo siguiente:



Estimacion de la frecuencia y severidad

2.1 Simulación de los pares (T_i, x_i)

Como la frecuencia de los eventos en un año se distribuye Poisson con media $\lambda_{Po} = 0.1$, entonces las diferencias de tiempo entre las ocurrencias de los eventos siguen una distribución exponencial con media $1/\lambda = \lambda_{Po} = 0.1$, es decir, los tiempos inter-arribos siguen una distribución exponencial con media $1/\lambda = \lambda_{Po} = 0.1$. Simulando los tiempos inter-arribos exponenciales, se puden simular las fechas de ocurrencia de los siniestros a partir de hoy. A continuación se muestran las primeras 5 y las últimas 5 realizaciones de las fechas de los siniestros y los montos correspondientes distribuidos exponencialmente con media $1/\lambda_{exp} = 10^4$:

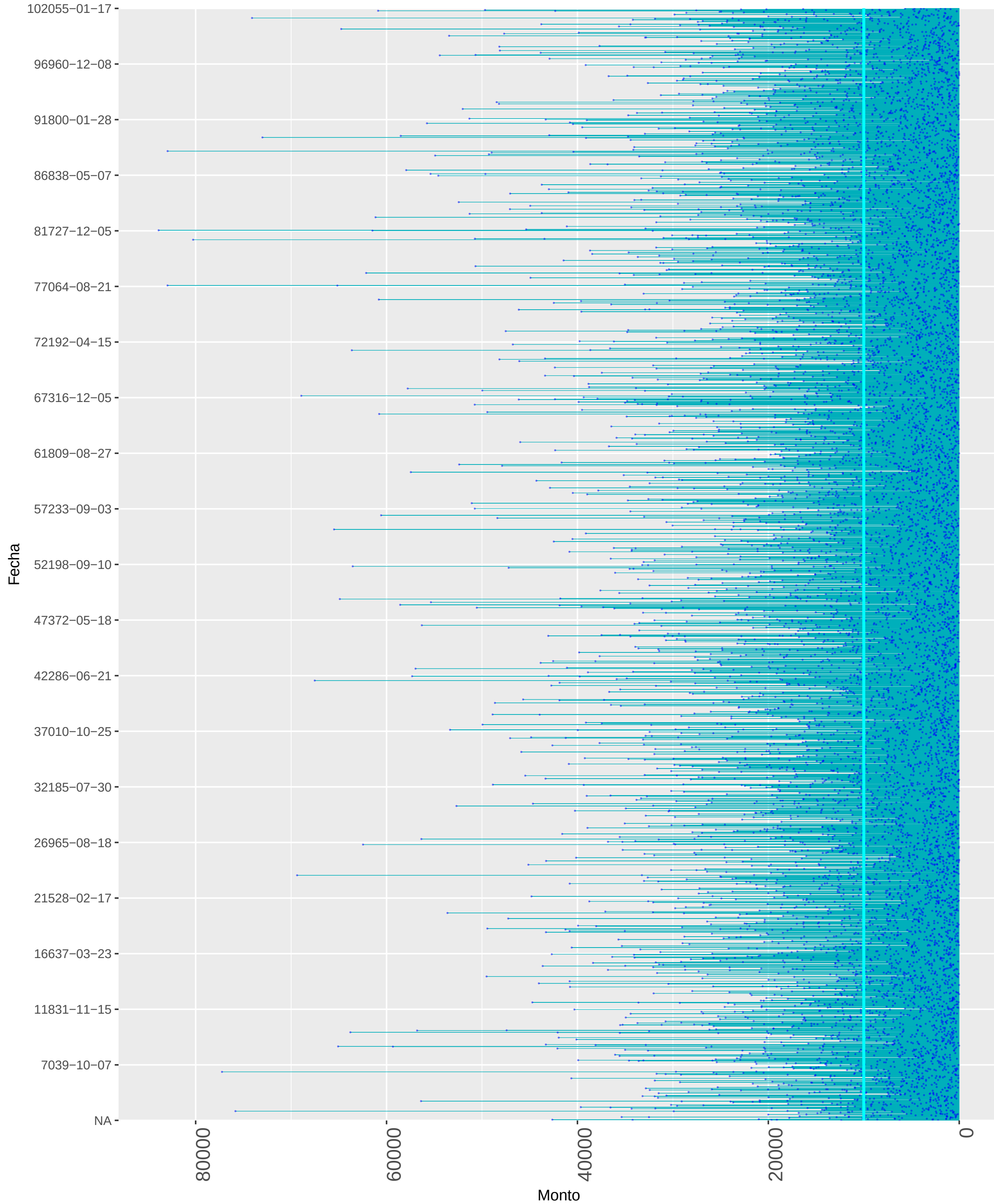
Fechas	Montos
2024-07-13 16:58:14	6245.128
2056-09-16 18:50:47	12363.807
2058-07-09 07:47:12	2213.956
2077-06-14 21:27:15	19672.506
2092-03-29 04:22:01	34219.872

$$\vdots$$

Fechas	Montos
102050-03-14 22:16:24	4676.792
102054-08-23 11:58:27	4012.666
102054-09-27 09:17:37	3706.924
102055-01-16 13:48:17	5680.881
102055-01-17 17:04:02	1956.398

2.2 Serie de tiempo

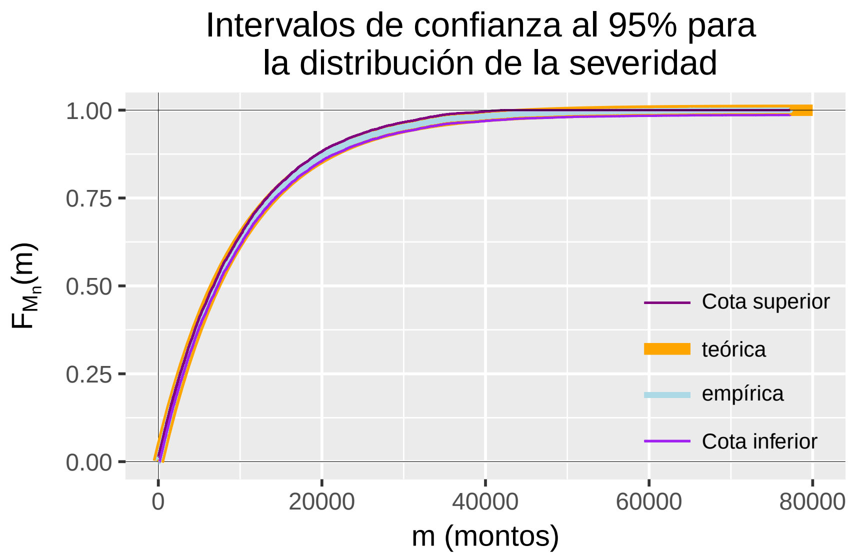
A continuación se muestra la serie de tiempo generada con la informacion de 1.00037×10^5 años:



La linea azul claro indica la media de la distribución teorica a partir de la cual fueron simulados los datos.

2.3 Distribución empírica de la severidad

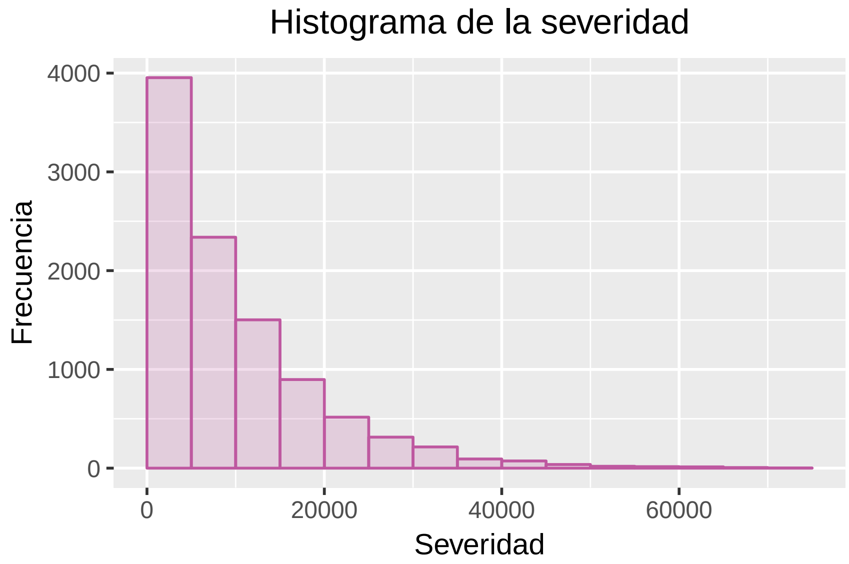
Con los valores simulados de los montos, se puede construir la función de distribución de probabilidad empírica de la severidad:



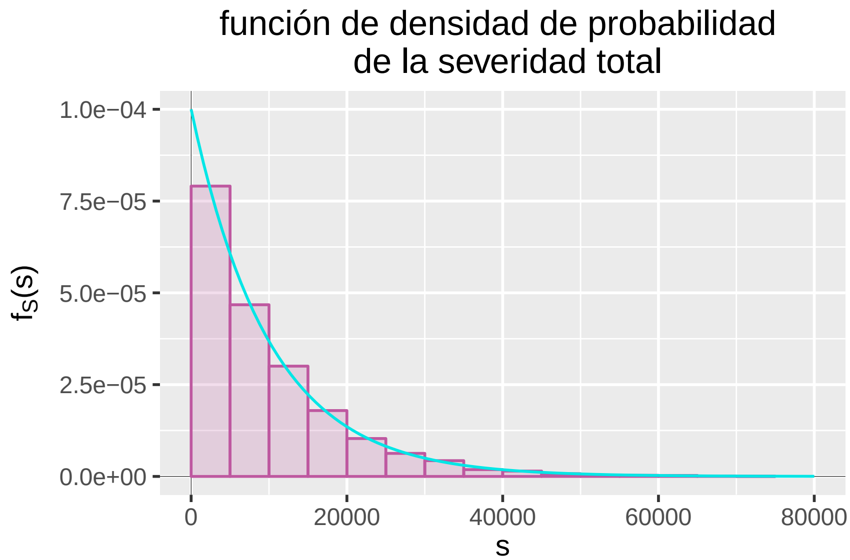
Dado el numero de observaciones (simulaciones) $n = 10000$, se tiene que la función de distribución empírica de la severidad se aproxima bastante bien a la función de distribución teórica y la longitud de los intervalos de confianza no es muy grande. Note que las cotas superiores e inferiores "abrazan estrechamente" a la función de distribución teórica.

2.4 Histograma de la severidad

Agrupando los montos de los siniestros en intervalos de longitud 5000, se obtiene el siguinete histograma:



Con un escalamiento adecuado sobre las frecuencias (note el cambio en la escala en el eje y), se puede observar la enorme similitud entre la densidad y el histograma:

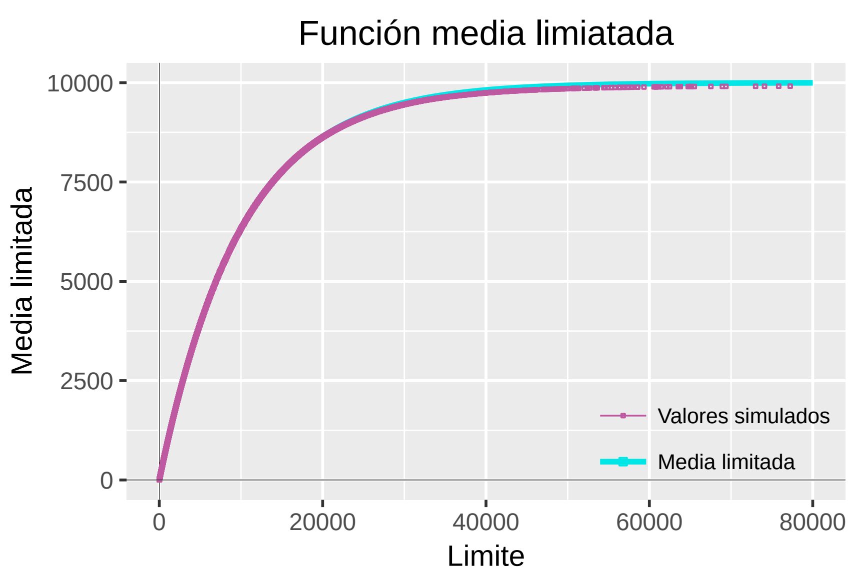


2.5 Características de la distribución empírica

Sea M la variable aleatoria que indica el monto de un siniestro, entonces el valor esperado de M limitado a l está dada por

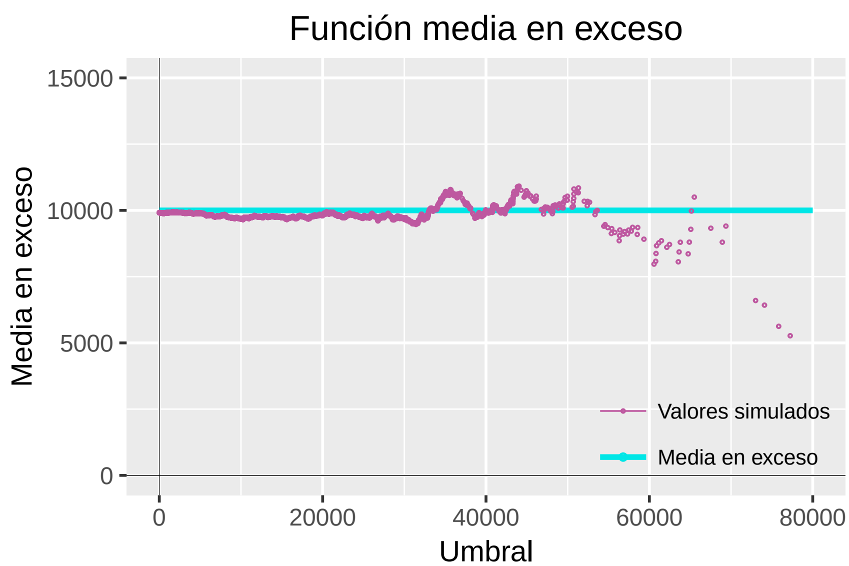
$$E\left(\min M, l\right) = \int_0^l S_{x_1}(t)\,\mathrm{d}t = \int_0^l e^{-\lambda_{exp} t}\,\mathrm{d}t = \frac{1 - e^{-\lambda_{exp} l}}{\lambda_{exp}}$$

A continuacion se presenta a la **media limitada** de los montos como funcion del limite:



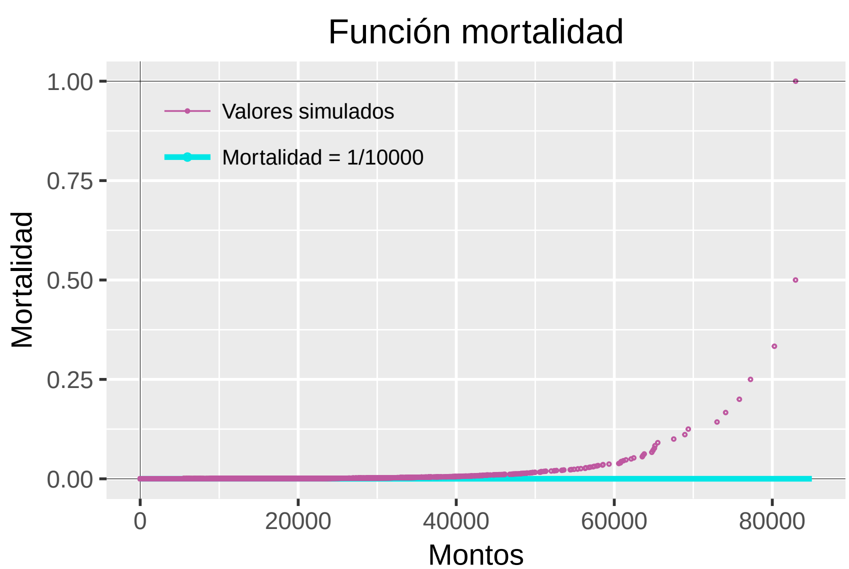
Como los montos son distribuidos exponencialmente con media $E(x_1) = 1/\lambda_{exp}$, no se observan indicios de una distribucion de cola pesada.

La **media en exceso** como funcion del umbral es la siguiente:



Claramente los montos de los siniestros no son un fenomeno de colas pesadas, pues se observa que la media en exceso descrece.

La mortalidad empirica como funcion del monto es la siguiente:



2.6 Prueba Ji-Cuadrada de Pearson

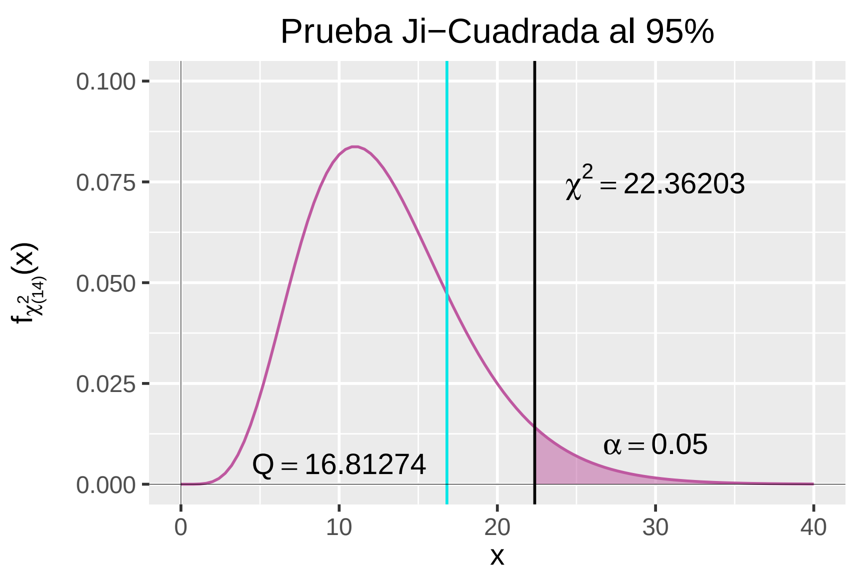
Se agrupan los datos de tal forma que todas las frecuencias de los datos observados sean mayores a 5:

cj_1	cj	V3
0	5000	3952
5000	10000	2340
10000	15000	1502
15000	20000	896
20000	25000	517
25000	30000	314
30000	35000	215
35000	40000	93
40000	45000	73
45000	50000	37
50000	55000	19
55000	60000	15
60000	65000	13
65000	70000	6
70000	85000	8
85000	inf	0

Realizando la prueba Ji-Cuadrada, se tiene que la estadistica de prueba toma el siguite valor:

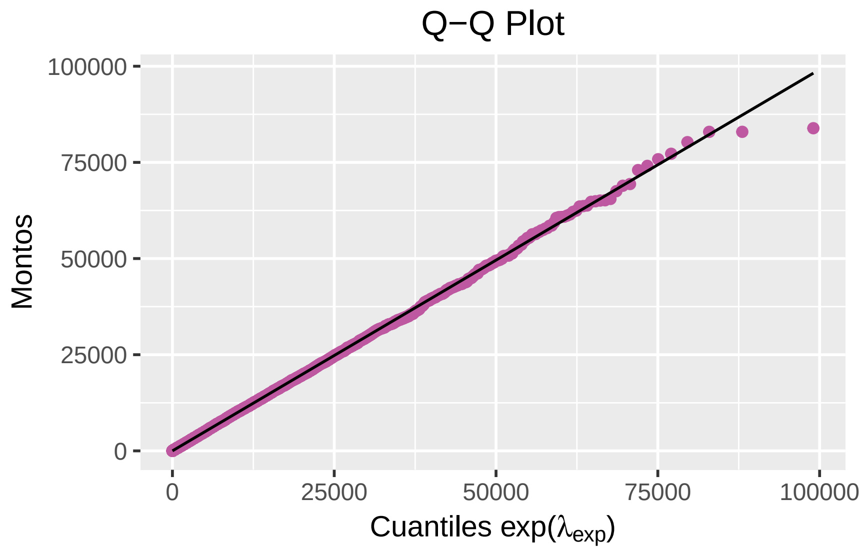
$$Q = 16.8127387 < \chi^2_{(13),0.05} = 22.3620325$$

Por lo tanto, la hipótesis de que la severidad sigue una distribución exponencial con media 10000 no es rechazada.



2.7 Q-Q Plot

Derivado del q-q plot podemos concluir que el modelo subestima las probabilidades de eventos de monto alto:



2.8 Estimacion para la frecuencia

Suponiendo que la frecuencia de los siniestros se distribuye Poisson, entonces, estimando por máxima verocimilitud se obtiene que, en promedio, ocurren

$$\widehat{\lambda}_{MV} = \frac{1}{T}\sum_{j=0}^{T} N_i = \frac{1}{1.00037 \times 10^5} \sum_{j=0}^{1.00037\times 10^5} N_i = 0.099963$$

siniestros al año. En donde T es el total de años sobre los cuales se tiene información y N_j es el número de siniestros que ocurrieron en el j-ésimo año. Además, dados estos valores se tiene que un intervalo al 95% de confianza para el parámetro de esta distribución Poisson es el siguiente

$$\widehat{\lambda} \pm 1.96\sqrt{\frac{\widehat{\lambda}}{n}} = 0.099963 \pm 0.0061969 = (0.0937661, 0.1061599).$$

Estimación con censura y truncamiento

3.1 Efectos de un deducible y un límite

Sobre los 10000 siniestros simulados se considera que aplicarán un deducible de $5,000$ y un límite de $30,000$. El archivo de indemnizaciones tendrá, entonces, el siguiente aspecto:

Fechas	Montos
2024-07-13 16:58:14	6245.128
2056-09-16 18:50:47	12363.807
2077-06-14 21:27:15	19672.506
2092-03-29 04:22:01	30000.000
2097-12-21 19:09:05	19105.984

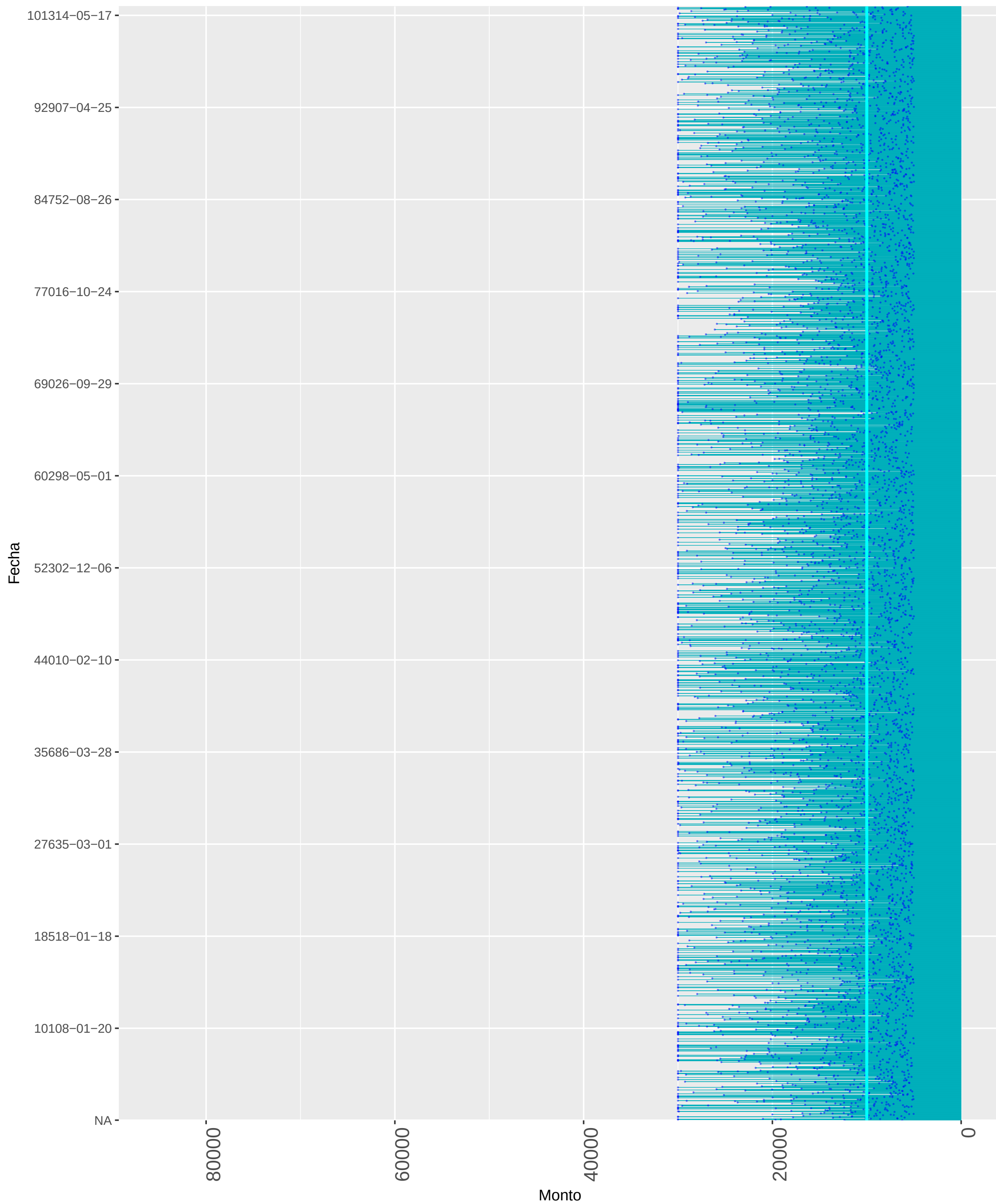
$\vdots$

Fechas	Montos
101947-01-26 23:33:39	8360.947
101991-12-31 03:36:50	22860.955
102000-08-25 01:35:51	20202.464
102026-10-06 13:37:54	16348.527
102055-01-16 13:48:17	5680.881

Como resultado de imponer un deducible y un límite, se ha eliminado un 39.52% de registros en el archivo de indemnizaciones. Se aceptó la hipótesis de que los montos de los siniestros provienen de una distribución exponencial de media 10^4, entonces se puede estimar de manera anaítica la proporción de registros eliminados mediante la probabilidad de que el monto de un siniestro se encuentre por debajo del deducible de 5000. Esta probabilidad es la siguiente:

$$Pr\{M \leq 5000\} = 1 - e^{-\frac{5000}{10^4}} = 0.3934693.$$

La nueva seríe de tiempo generada por el proceso de riesgo es la siguiente:



La línea azul claro indica el valor de la media en exceso con un umbral de 5000.

3.2 Estimación para montos con deducible y límite

Primero vamos a definir la función con el deducible de $5,000$ y un límite de $30,000$

$$y(x)=\begin{cases}0 & x\leq 5,000\\ x-5000 & 5,000\leq x\leq 30,000\\ 25,000 & 30,000\leq x\end{cases}$$

$$F_Y(y)=\begin{cases}0 & y=0\\ \frac{F_x(y+5000)-F_x(5000)}{1-F_x(5000)} & 0\leq y\leq 900\\ 1 & 900\leq y\end{cases}$$

$$f_Y(y)=\begin{cases}\frac{f_x(y+5000)}{S_X(5000)} & 0\leq y\leq 25,000\\ \frac{1-F_x(25000)}{1-F_x(5000)} & y=25,000\end{cases}$$

Un vez teniendo bien definidas funciones de Distribución y Densidad, podemos hacer la función de Máxima Verosimilitud.

$$L(\lambda)=\left(\frac{1-F_x(25000)}{1-F_x(5000)}\right)^a\cdot\prod_{i=1}^{b}\frac{f_x(y_i+5000)}{1-F_{x(5000)}}$$

Ahora, sustituyendo los valores de la función definida, se tiene que:

$$L(\lambda)=\left(\frac{e^{-\lambda(25000)}}{e^{-\lambda(5000)}}\right)^a\cdot\prod_{i=1}^{b}\frac{\lambda e^{-\lambda(y_i+5000)}}{e^{-\lambda(5000)}}$$

sacando Logaritmo a la función:

$$\ln(\lambda)=a(-\lambda(25000)+\lambda(5000))+\sum_{i=1}^{b}(\ln(\lambda)-\lambda(y_i+5000)+\lambda(5000)a)$$

Agrupando los terminos comunes llegamos a la función a maximizar

$$l(\lambda)=a(-\lambda(2000))+b(\ln(\lambda))-\sum_{i=1}^{b}y_i$$

Después de resolver con Solver en Excel, se obtuvo que el parámetro de la distribución exponencial estimado por máxima verosimilitud es .000076511 y su media $13,069.95$. Comparada con la media que se obtuvo de los datos, que fue de $13,734.35$, se puede notar que el parámetro estimado considera los efectos de censura y truncamiento adecuadamente.