Proyecto I Métodos MonteCarlo Aplicados Cuantificación en Riesgo Operacional

Yael Fabián Olivares Cruz

December 4, 2024

Contents

1 Introducción

La simulación Montecarlo es una herramienta para modelar problemas donde tenemos un entorno de incertidumbre. En este proyecto, se aplicó específicamente para cuantificar el riesgo operacional en una organización, centrándose en los eventos que generan pérdidas económicas significativas.

El objetivo principal es estimar las pérdidas potenciales debido a eventos de riesgo operacional, todo esto a través de medidas de riesgo conocidas.

Se analizó una base de datos con eventos simulados. Se describen de manera breve algunos de los riesgos en una empresa tecnológica y algunas de las estrategias de mitigación de los mismos, para posteriormente describir el riesgo residual de manera cuantitativa creando simulaciones a través del método MonteCarlo. Nos centraremos en los eventos que tienen un impacto económico en la organización.

Debido a que los incidendes no ocurren con la misma tasa en diferentes meses se utilizará un proceso poission no homogeneo para describir la frecuencia de los eventos. Para la severidad de los incidentes de intentará ajustar una distribución lognormal ó algún método de simulación no paramétrico en caso de que no se pueda ajustar la distribución. Posteriormente se utilizará un modelo Poisson compuesto no homogeneo para describir las pérdidas en un mes determinado, se ajustará una distribución a los valores finales obtenidos para cada simulación.

A continuación, para describir los valores extremos se ajustará alguna distribución de cola pesada a las simulaciones de pérdidas para poder estudiar los valores extremos, en los cuales se tiene eventos de baja probabilidad con alto impacto.

Finalmente, se ajustará una cópula para encontrar la función de densidad conjunta que describa las pérdidas. Esto debido a que la ocurrencia de un incidente puede afectar a más de un riesgo, por lo que las distribuciones marginales tienen una dependencia inherente.

2 Planteamiento del problema

En una organización tecnológica, los riesgos operacionales son inherentes a los procesos, productos y sistemas. Estos riesgos pueden originarse en fallos técnicos, errores humanos o desastres externos. La naturaleza del entorno digital y tecnológico implica una alta dependencia de sistemas críticos que pueden fallar, generando incertidumbre sobre la frecuencia y severidad de los eventos adversos.

Riesgo Inherente

Este tipo de riesgo existe intrínsecamente en toda actividad, independientemente de las acciones de mitigación. Cuando se ejecuta un proceso, este no puede ser eliminado, por lo que su identificación debe contemplarse en los planes de gestión de la organización. Ejemplos incluyen:

- Fallos de hardware o software en sistemas operativos críticos.
- Errores humanos en procesos de manejo de datos sensibles.

Entorno de Control

Antes de la ejecución de cualquier proceso, se deben reconocer los riesgos inherentes e implementar estrategias de mitigación, estas son conocidas como EMTA, las cuales incluyen:

1. **Evitar:** Consiste en anular el proceso que se está llevando a cabo con la finalidad de eliminar el riesgo al que se está expuesto.
2. **Mitigar:** Como se mencionó anteriormente, consiste en diseñar un entorno de control, con la finalidad de reducir la probabilidad e impacto de situaciones desfavorables.
3. **Transferir:** Es posible delegar las responsabilidades que implica la exposición de riesgo a otro equipo que acepte llevar a cabo otras estrategias para mitigar situaciones adversas durante la ejecución de procesos.
4. **Aceptar:** Cuando las acciones de mitigación de riesgo llegan a tener un mayor costo que las situaciones a las que se está expuesto, lo más recomendable es continuar con el proceso sin acciones mitigantes.

Una vez elegida la estrategia de mitigación, se diseñan herramientas para hacer frente al riesgo tales como:

- **Indicadores Clave de Riesgo (KRIs):** Monitorean señales tempranas de posibles eventos adversos.
- **Recomendaciones de auditoría:** Detectan vulnerabilidades en los sistemas y procesos.
- **Planes de mitigación:** Reducen la probabilidad o impacto de eventos negativos mediante controles específicos.
- **Certificación de controles:** Evalúan la eficacia de las medidas implementadas.

Riesgo Residual

Este es el riesgo que persiste después de implementar el entorno de control. Es el foco principal de la cuantificación en este proyecto. Ejemplos incluyen:

- Vulnerabilidades que no pudieron eliminarse completamente tras una actualización de seguridad.
- Eventos impredecibles, como ataques cibernéticos de tipo *zero-day*.

Desafíos en el Modelado de Riesgos

Los riesgos operacionales presentan características que dificultan su análisis:

1. **Frecuencia Variable:** Los eventos no ocurren con una tasa constante. Por ejemplo, los incidentes de seguridad cibernética pueden aumentar durante períodos de alta actividad, como ventas especiales. Esto demanda un modelo que capture tasas variables en el tiempo, como el proceso Poisson no homogéneo.
2. **Severidad Altamente Dispersa:** La severidad de los eventos varía considerablemente. Algunos eventos pueden generar pérdidas pequeñas, mientras que otros tienen impactos financieros catastróficos. Las distribuciones de cola pesada (e.g., Pareto, Lognormal) son necesarias para describir adecuadamente este comportamiento.

3. **Escasez de Datos de Eventos Extremos:** Los eventos de alta severidad son raros y, por tanto, hay pocos datos históricos para analizarlos. Esto hace necesaria la extrapolación mediante distribuciones específicas y la teoría de valores extremos.

4. **Dependencia entre Riesgos:** Un evento en una categoría de riesgo puede desencadenar impactos en otras. Por ejemplo, una interrupción tecnológica podría generar pérdidas económicas, problemas legales y daños reputacionales. Es esencial modelar esta dependencia utilizando herramientas como las cópulas, que capturan correlaciones no lineales entre riesgos.

Medidas Clave para la Cuantificación del Riesgo

El análisis del riesgo operacional requiere métricas específicas que permitan caracterizar tanto los riesgos promedio como los extremos:

Pérdida Esperada ($E[X]$)

Representa el valor promedio de las pérdidas esperadas en un periodo. Es una medida fundamental para presupuestar reservas contra pérdidas.

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx \tag{1}$$

En caso de datos discretos, se estima con:

$$\hat{E}[X] = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2}$$

Volatilidad (σ)

Mide la variabilidad de las pérdidas. Una alta volatilidad indica eventos con impactos muy diferentes, desde leves hasta catastróficos.

$$\sigma = \sqrt{E[X^2] - (E[X])^2} \tag{3}$$

Valor en Riesgo (VaR)

Mide la pérdida máxima esperada en un nivel de confianza específico. Es clave para escenarios donde se necesita evaluar la solvencia financiera frente a riesgos extremos.

$$P(X > VaR_p) = 1 - p \tag{4}$$

Valor en Cola en Riesgo (TVaR)

Es el promedio de las pérdidas que exceden el VaR, proporcionando una visión más completa de los eventos extremos.

$$TVaR_p = E[X \mid X > VaR_p] \tag{5}$$

Impacto Económico de Eventos Extremos

Los eventos extremos son el foco de este análisis porque tienen el potencial de desestabilizar una organización. Ejemplos:

- **Eventos de baja probabilidad y alto impacto (LFHS):** Un ataque cibernético exitoso puede generar multas regulatorias multimillonarias y daños a la reputación. Estos eventos requieren reservas financieras específicas y planes de respuesta rápidos.
- **Eventos de alta frecuencia y baja severidad (HFLS):** Problemas recurrentes en operaciones menores, como errores en la facturación, pueden acumularse y generar costos significativos a largo plazo.

Objetivo del Análisis

El propósito del análisis es desarrollar un modelo cuantitativo que permita:

1. **Identificar patrones:** Estudiar las características temporales y severidad de los eventos.
2. **Evaluar riesgos extremos:** Identificar los eventos que representan el mayor impacto potencial y estimar sus probabilidades.

3. **Proveer herramientas de decisión:** Ayudar a la organización a priorizar estrategias de mitigación, asignar recursos y establecer políticas de gestión del riesgo más efectivas.

3 Medidas de riesgo

3.1 Pérdida esperada

La pérdida esperada es el valor promedio de las pérdias que una empresa anticipa en un periodo dado, para un tipo de riesgo es específico ó todos los posibles eventos de riesgo operacional. Es una medida clave en la gestión de riesgos, ya que proporciona una estimación de las pérdidas promedio que se esperan debido a la ocurrencia de eventos adversos, como fallos en sistemas, falta de disponibilidad, desastres externos, etc.

La pérdida esperada o esperanza $\mathbf{E}[X]$, es el valor promedio ponderado de una variable aleatoria.

Para distribuciones discretas se define como sigue

$$\mathbf{E}[X] = \sum_{i=1}^{n} X_i \cdot P(X_i)$$

donde X_i son las pérdidas y $P(X_i)$ es la probabilidad de que ocurra una pérdida de X_i

Para distribuciones continuas la esperanza se define de la siguiente manera

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx$$

donde $f_X(x)$ es la función de densidad de las pérdidas.

En caso de no tener una función de probabilidad asociada a las pérdidas, podemos estimar la pérdida esperada utilizando la ley de los grandes números.

$$\frac{1}{n} \sum_{i=1}^{n} X_i \underset{n \to \infty}{\longrightarrow} \mathbf{E}[X]$$

Por lo que

$$\widehat{\mathbf{E}(X)} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

será el estimador que utilizaremos para la pérdida esperada.

3.2 Volatilidad

La volatilidad es una medida de la variabilidad ó dispersión de los retornos de un activo o de los resultados de una operación en un periodo específico. En el contexto de riesgo operacional, la volatilidad es importante para entender la incertidumbre en los procesos internos, sistemas y factores externos que puedan impactar a una organización. Ayuda a a medir la inestabilidad y variabilidad en eventos operativos que pueden resultar en pérdidas para la empresa.

Una alta volatilidad implica que las pérdidas operativas pueden fluctuar significativamente, lo cual representa incertidumbre y una posible amenaza para la estabilidad financiera de la organización. Esra variabilidad puede provenir de eventos operativos frecuentes con pérdidas pequeñas ó eventos poco frecuentes con pérdidas grandes que suelen ser impredecibles y tener un impacto financiaero considerable.

Para medir la volatilidad se suele utilizar la desviación estándar σ de las pérdidas, la cual representa el promedio de las distancias de cada pérdida individual respecto de la pérdida medio. Matemáticamente es la raíz cuadrada de la desviación estándar σ^2, la cual se define como sigue

$$\sigma^2 = \mathbf{E}[(X - \mathbf{E}[X])^2]$$

Desarrollando el lado derecho de la ecuación obtenemos que

$$\sigma^2 = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

Donde $\mathbf{E}[X]$ es la pérdida esperada y $\mathbf{E}[X^2]$ es el segundo momento de la distribución de pérdidas, el cual utilizando la ley del estadístico inconsciente puede calcularse de la siguiente forma para distribuciones discretas

$$\mathbf{E}[X^2] = \sum_{i=1}^{n} X_i^2 \cdot P(X_i)$$

y para distribuciones continuas como sigue

$$\mathbf{E}[X^2] = \int_{-\infty}^{\infty} x^2 \cdot f_X(x) dx$$

Como vimos anteriormente, el cálculo de la esperanza requiere conocer la distribución de las pérdidas. En caso de que no tener información sobre su

distribución podemos utilizar el siguiente estimador para σ

$$\widehat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

donde n es el número total de eventos, X_i es la pérdida de cada evento y $\overline{X}$ es el estimador de la pérdida media.

3.3 Valor en riesgo

El valor en riesgo (VaR) es la medida estandar utilizada para evaluar la exposición al riesgo. En términos generales en VaR es el capital necesario para asegurar, con un nivel de confianza, que la empresa no se volverá técnicamente insolvente. El nivel de confianza es elegido arbitrariamente, en la práctica puede ser un número grande como 99.95% para toda la empresa, o puede ser 95% para un sólo riesgo.

Sea X una variable aleatoria que denota pérdidas. El **Valor en Riesgo (VaR)** de X al nivel de confianza $100\%p$, denotado $VaR_p(X)$ ó x_p es el percentil 100_p de la distribución de X.
Para variables continuas simplemente podemos buscar el valor x_p que satisface

$$\mathbf{P}(X > x_p) = p$$

El VaR es utilizado en riesgos de traiding y gestión de riesgo en un periodo fijo de tiempo. En estas situaciones la distribución normal es utilizada para describir ganancias ó pérdidas según sea el caso. Sin embargo, la distribución normal generalmente no es utilizada para describir pérdidas operacionales ya que la mayoría de las distribuciones tienen una asimetría considerable.

3.4 Valor de cola en riesgo

Sea X una variable aleatoria que denota pérdidas. El **Valor de cola en Riesgo (TVaR)** de X al nivel de confianza $100\%p$, denotado por $TVaR_p(X)$, es la pérdida esperada dado que la pérdida excede el cuantil 100_p de la distribución de X. Consideraremos únicamente distribuciones continuas.

Podemos definir el $TVaR_p(X)$ para la variable aleatoria X como

$$TVaR_p(X) = \mathbf{E}(X|X > x_p) = \frac{\int_{x_p}^{\infty} x dF(x)}{1 - F(x_p)}$$

Más aún, la cantidad anterior es finita, podemos usar integración por partes y sustitución como sigue

$$\begin{aligned} TVaR_p(X) &= \frac{\int_{x_p}^{\infty} x f(x)}{1 - F(x_p)} \\ &= \frac{\int_{x_p}^{\infty} VaR_u(X) du}{1 - p} \end{aligned}$$

De este modo, podemos observar que el $TVaR$ promedia todos los valores del VaR por encima de un nivel de confianza p. Por lo que el $TVaR$ nos dice más sobre la distribución de la cola que sólo el VaR.

Finalmente, el $TVaR$ también puede ser escrito como sigue

$$\begin{aligned} TVaR_p(X) &= \mathbf{E}(X|X > x_p) \\ &= x_p + \frac{\int_{x_p}^{\infty} (x - x_p) dF(x)}{1 - F(x_p)} \\ &= VaR_p(x) + e(x_p) \end{aligned}$$

Donde $e(x_p)$ es la función de pérdida esperada.

De este modo, el $TVaR$ es más grande que el VaR. Se argumenta que el VaR es una medida de riesgo de "todo ó nada", si ocurre un evento de pérdidas grandes que sobrepase el umbral del VaR, no habrá capital para cubrir las pérdidas. Mientras que el $TVaR$ da una definición de "malos tiempos" donde las pérdidas exceden el umbral del VaR.

3.5 Ejemplos

1. Distribución Normal
   Considere a X como una variable aleatoria normal con media μ y desviación estándar σ. Entonces su función de densidad es

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

La pérdida esperada de X es μ
La volatilidad de X es σ
$VaR_p(X) = \mu + \sigma\Phi^{-1}(p)$
$TVaR_(X) = \mu + \sigma\frac{\phi[\Phi^{-1}(p)]}{1-p}$

Donde $\phi(x)$ representa la función de densidad de una normal estándar y $\Phi(x)$ representa la función de distribución acumulada de una normal estándar

2. Distribución t de student
   Considere a X como una variable aleatoria t de student con parámetro de localización μ y parámetro de escala σ, con ν grados de libertad. Su función de densidad es

$$f_X(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\sigma\Gamma(\frac{\nu}{2})}\left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}}$$

   La pérdida esperada de X es μ
   La volatilidad de X es $\sigma\sqrt{\frac{\nu}{\nu-2}}$
   $VaR_p(X) = \mu + \sigma T^{-1}(p)$
   $TVaR_p(X) = \mu + \sigma\frac{t[T^{-1}(p)]}{1-p}\left\{\frac{\nu+[T^{-1}(p)]^2}{\nu-1}\right\}$

3. Distribución exponencial
   Considere a X como una distribución exponencial con media θ. Su función de densidad es

$$f_X(x) = \frac{1}{\theta}exp\left(-\frac{x}{\theta}\right),\ x > 0$$

   La pérdida esperada de X es θ
   La volatilidad de X es θ^2
   $VaR_p(X) = \theta ln(1-p)$
   $TVaR_p(X) = VaR_p(X) + \theta$

   El excedente del $TVaR$ sobre el VaR es la constante θ para todos los valores de p por la propiedad de pérdida de memoria de la distribución exponencial.

4 Severidad de eventos

En esta sección nos centraremos en modelos que pueden usarse para medir el tamaño de las pérdidas. Nos restingiremos a distribuciones con soporte no negativo. En este enfoque nos centraremos en eventos que no consideran ganancias si no pérdidas.

4.1 Distribuciones de un parámetro

4.1.1 Distribución exponencial

$$f(x) = \frac{1}{\theta} e^{-x/\theta}$$

$F(x) = 1 - e^{-x/\theta}$
$\mathbf{E}[X^k] = \theta^k \Gamma(k+1), \quad k > -1$

La exponencial es la única distribución continua con una función de riesgo constante $h(x) = \frac{1}{\theta}$ y una pérdida esperada excedente condicional que también es constante, $e_d(x) = \theta$. Por lo tanto el tamaño esperado del excedente sobre un umbral no depende del umbral

4.1.2 Distribución inversa exponencial

$$f(x) = \frac{\theta e^{-\theta/x}}{x^2}$$

$F(x) = e^{-\theta/x}$
$\mathbf{E}[X^k] = \theta^k \Gamma(1-k), \quad k < 1$

La inversa exponencial está relacionada con la exponencial. Esta distribución tiene esperanza no finita

4.1.3 Pareto de un parámetro

$$f(x) = \alpha \theta^{\alpha-1} x^{-\alpha-1}, \quad x > \theta$$

$F(x) = 1 - \left(\frac{\theta}{x}\right)^{\alpha}, x > \theta$
$\mathbf{E}[X^k] = \frac{\alpha \theta^k}{\alpha - k}, \quad k < \alpha$

La Pareto es descrita como una distribución de cola pesada. El soporte de esta distribución inicia en θ.

4.2 Distribuciones de dos parámetros

4.2.1 Distribución Gamma

$$f(x) = \frac{(x/\theta)^{\alpha} e^{-x/\theta}}{x\Gamma(\alpha)}$$

$F(x) = \Gamma(\alpha; x/\theta)$
$\mathbf{E}[X^k] = \frac{\theta^k \Gamma(\alpha+k)}{\Gamma(\alpha)}, \quad k > -\alpha$

Si α es un entero, la distribución gamma puede verse como la suma de α variables aleatorias independientes identicamente distribuidas con distribución exponencial.

4.2.2 Distribución Lognormal

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{ln(x)-\mu}{\sigma}\right)^2\right] = \frac{\phi\left(\frac{ln(x)-\mu}{\sigma}\right)}{\sigma x}$$

$F(x) = \Phi\left(\frac{ln(x)-\mu}{\sigma}\right)$
$\mathbf{E}[X^k] = exp(k\mu + \frac{1}{2}k^2\sigma^2)$

Esta distribución es obtenida a partir de una variable aleatoria Y normal con media μ y varianza σ^2 a través de la transformación e^Y

4.2.3 Distribución Weibull

$$f(x) = \lambda\alpha(\lambda x)^{\alpha-1} e^{-(\lambda x)^{\alpha}}, \quad x > 0$$

$F(x) = 1 - e^{-(\lambda x)^{\alpha}}$
$\mathbf{E}[X^k] = \frac{1}{\lambda^k}\Gamma\left(1 + \frac{n}{\alpha}\right)$

4.2.4 Distribución Loglogística

$$f(x) = \frac{\gamma(x/\theta)^{\gamma}}{x[1 + (x/\theta)^{\gamma}]^2}$$

$F(x) = \frac{(x/\theta)^\gamma}{1+(x/\theta)^\gamma}$
$\mathbf{E}[X^k] = \theta^k \Gamma(1+k/\gamma)\Gamma(1-k/\gamma), \quad -\gamma < k < \gamma$

Esta es una distribución de cola pesada

4.2.5 Distribución Pareto

$$f(x) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}$$

$F(x) = 1 - \left(\frac{\theta}{x+\theta}\right)^\alpha$
$\mathbf{E}[X^k] = \frac{\theta^k \Gamma(k+1)\Gamma(\alpha-k)}{\Gamma(\alpha)}$

La distribución Pareto tiene una cola muy pesada y es muy utilizada en la modelación cuando hay una probabilidad muy alta de pérdidas grandes.

4.3 Distribución PERT

La distribución PERT (Program Evaluation and Review Technique) es una familia de distribuciones Beta ajustadas que se emplean para estimaciones de tiempo, costos y otros factores inciertos en la planificación de proyectos.
Tiene la siguiente función de densidad

$$f(x) = \frac{(x-a)^{\alpha-1}(c-x)^{\beta-1}}{\mathbf{B}(\alpha,\beta)(c-a)^{\alpha+\beta-1}}$$

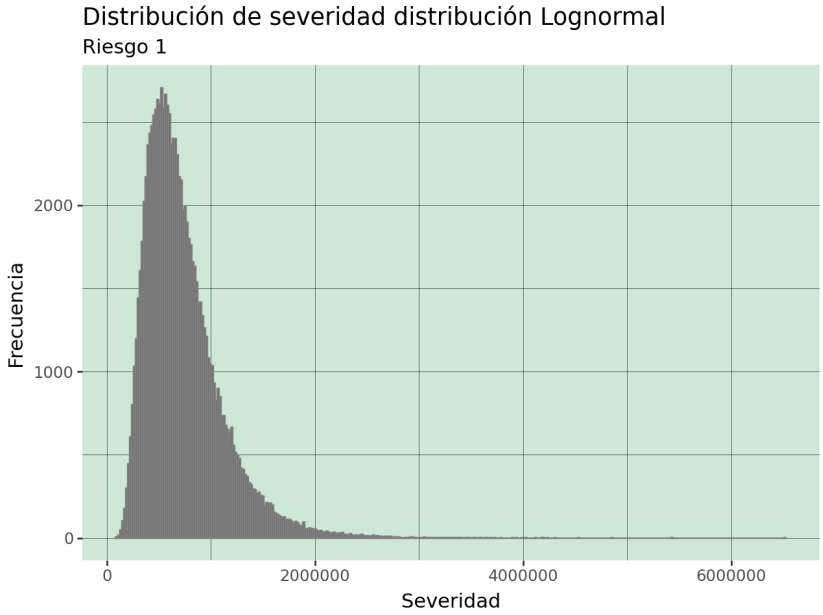
Donde
$\alpha = 1 + \lambda\frac{b-a}{c-a}$
$\beta = 1 + \lambda\frac{c-b}{c-a}$
$a < b < c$; a es el valor mínimo, b es el valor más frecuente y c es el valor máximo
$\lambda \in \mathbf{R}$ usualmente se usa $\lambda = 4$

4.4 Simulación

La simulación MonteCarlo se hizo a través de Python 3, el código con los resultados queda anexo a este documneto.
Se hicieron tres simulaciones de pérdidas con diferente distribución de probabilidad con el código utilizado a continuación.

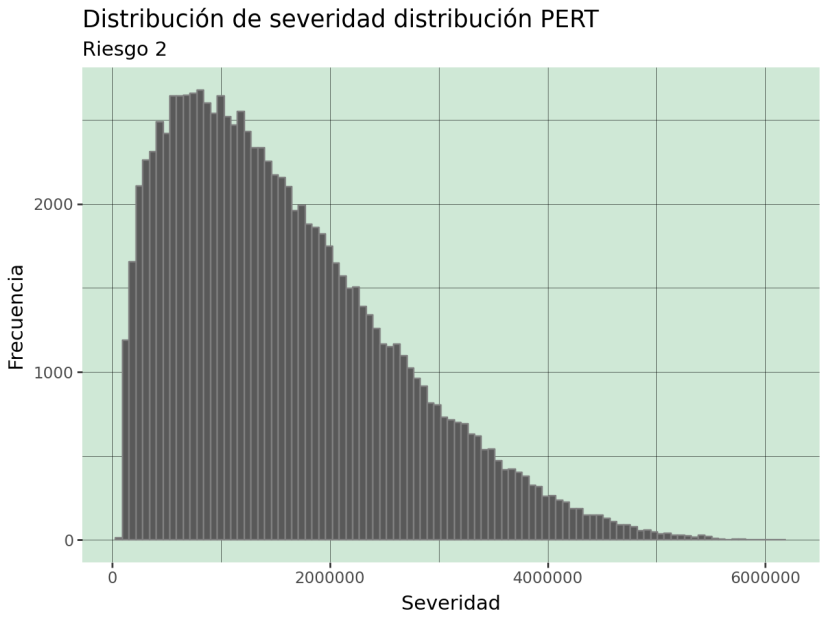
1. **Lognormal**

```
# Parametros de la distribucion Lognormal
n = 100000
perdidaMedia = 650000
sdPerdida = 1.62

# Generacion de la Lognormal
np.random.seed(2024)
lognormal = np.random.lognormal(mean=np.log(perdidaMedia),
sigma=np.log(sdPerdida), size = n)

# Grafico de la distribucion Lognormal
(ggplot() +
 geom_histogram(mapping=aes(x=lognormal), color='grey') +
 labs(title="Distribución de severidad distribución Lognormal",
 subtitle='Riesgo 1', x="Severidad", y="Frecuencia") +
 theme(panel_background=element_rect(fill="#cfe8d6", color=None),
       panel_grid=element_line(color="black", size=0.2)
       )
 )
```



2. **PERT**

```
def rpert(n, min, media, max, lamb=4):
  valorAlpha = 1 + lamb*((media-min)/(max-min))
  valorBeta = 1 + lamb*((max-media)/(max-min))
  np.random.seed(2024)
  rbeta = beta.rvs(valorAlpha, valorBeta, size=n)
  return min + (max - min)*rbeta

pert = rpert(n, min, media, max)

# Grafico de la distribucion PERT
(ggplot() +
 geom_histogram(mapping=aes(x=pert), color='grey') +
 labs(title="Distribución de severidad distribución PERT",
 subtitle='Riesgo 2', x="Severidad", y="Frecuencia") +
 theme(panel_background=element_rect(fill="#cfe8d6", color=None),
       panel_grid=element_line(color="black", size=0.2)
       )
 )
```



3. **Distribución empírica** Cuando no la distribución de pérdidas no tiene una función de densidad conocida, podemos utilizar el método de la transformada inversa con los datos empíricos para hacer simulaciones que aproximen el comortamiento de los eventos.

   Para obtener simulaciones de la distribución utilizaremos el siguiente algoritmo:

   1. Obtenemos el número de eventos (n)
   2. Ordenamos los eventos por severidad
   3. La función de distribución acumulada (CDF) valdrá $\frac{i}{n}$, donde i es el i-ésimo evento
   4. Generar un número aleatorio $U \sim U(0,1)$
   5. Tomar $X = F^{-}U$

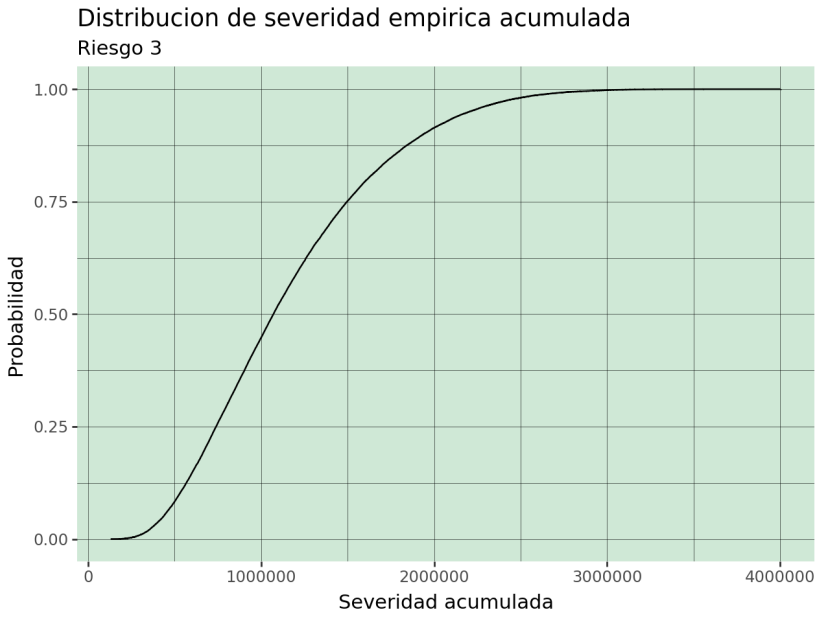
   Se utilizará un DataFrame llamado tickets

```
tickets.head()
```

	Incident_ID	Impact	Category	Open_Time	Severidad
0	IM0000004	4	incident	2012-02-05 13:32:00	2.036414e+06
1	IM0000005	3	incident	2012-03-12 15:44:00	1.087896e+06
3	IM0000011	4	incident	2012-07-17 11:49:00	3.459975e+05
4	IM0000012	4	incident	2012-08-10 11:01:00	1.198250e+06
5	IM0000013	4	incident	2012-08-10 11:27:00	6.721282e+05

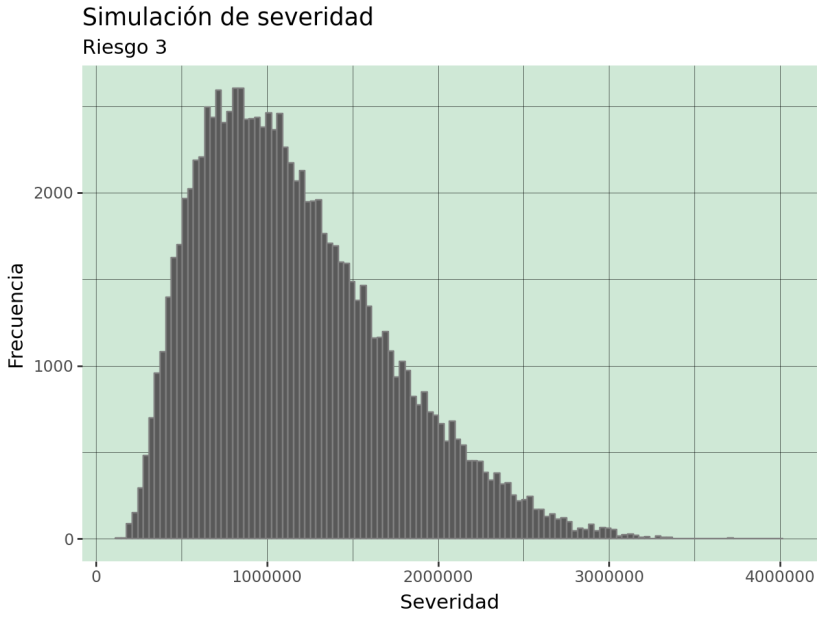
```
# Funcion para calcular la distribucion empirica
def distribucionEmpirica(eventos):
  n = len(eventos)
  eventosOrdenados = np.sort(eventos)
  CDF = np.arange(1, n+1)/n
  return eventosOrdenados, CDF

# Obtener la funcion de distribucion empirica
```

```
eventos = tickets['Severidad']
eventos, cdf = distribucionEmpirica(eventos)

# Graficar la distribucion empirica acumulada
(ggplot() +
 geom_step(mapping=aes(x=eventos, y=cdf)) +
 labs(title="Distribucion de severidad empirica acumulada",
 subtitle='Riesgo 3', x="Severidad acumulada", y="Probabilidad") +
 theme(panel_background=element_rect(fill="#cfe8d6", color=None),
       panel_grid=element_line(color="black", size=0.2)
       )
 )
```




```
# Funcion para simular valores de los incidentes
def simulacionDistEmpirica(n):
  perdidas = [] # Lista donde se almacenaran las perdidas
  for _ in range(n):
    u = np.random.uniform(0,1)
    # Encontrar a que altura se encontro la uniforme
    ubicacionU = np.searchsorted(cdf, u)
```

```
        # Mapear la altura con su perdida
        perdidaSimulada = eventos[ubicacionU]
        # Añadir las perdidas simuladas a la lista
        perdidas.append(perdidaSimulada)
    return np.array(perdidas)

# Generacion de los numeros de que siguen la distribucion empirica
n = 100000
np.random.seed(2024)
distEmpirica = simulacionDistEmpirica(n)

# Graficar la distribucion empirica de perdidas
(ggplot() +
 geom_histogram(mapping=aes(x=distEmpirica), color='grey') +
 labs(title="Simulación de severidad",
 subtitle='Riesgo 3', x="Severidad", y="Frecuencia") +
 theme(panel_background=element_rect(fill="#cfe8d6", color=None),
       panel_grid=element_line(color="black", size=0.2)
       )
 )
```



5 Frecuencia de eventos

En el contexto del riesgo operacional, las distribuciones de conteo describen el número de eventos causantes de pérdidas, la función de probabilidad p_k denota la probabilidad de que exactamente k eventos ocurran. Sea N una variable aleatoria que representa el número de esos eventos. Entonces

$$p_k = \mathbf{P}(N = k), \qquad k = 0, 1, 2, ...$$

5.1 Distribución Poisson

La función de probabilidad Poisson es

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, ...$$

Para esta distribución sabemos que

$$E(N) = \lambda = Var(N)$$

Esta distribución tiene al menos dos propiedades importantes.

1. Sean $N_1, ..., N_n$ variables aleatorias Poisson independientes con parámetros $\lambda_1, ...\lambda_n$. Entonces $N = \sum_{i=1}^{n} N_i$ tiene una distribución poisson con parámetro $\lambda = \lambda_1, ...\lambda_n$
2. La segunda propiedad es útil en el modelado de eventos de riesgo operacional. Supongamos que tenemos un número fijo de eventos en un periodo fijo de tiempo, que sigue una distribución Poisson. Más aún, supongamos que las pérdidas pueden clasificarse en m tipos distintos. Por ejemplo, aquellas que estén por debajo y por encima de cierto umbral. Si queremos estudiar aquellas que están por encima del umbral, esa nueva distribución será también Poisson con otro parámetro

Esta última propiedad puede ser útil cuando consideramos el impacto de remover o añadir tipos de riesgos. Supongamos que el número de eventos de un conjunto de tipos de riesgo sigue una distribución Poisson. Si uno de los tipos de riesgo es eliminado, la distribución del número de pérdidas seguirá siendo Poisson pero con un nuevo parámetro.

Una variable aleatoria discreta X tiene distribución de la familia exponencial si su función de probabilidad puede expresarse como

$$f(x;\theta) = e^{\eta(\theta)T(x)-A(\theta)}h(x)$$

Veamos que la distribución Poisson pertenece a la familia exponencial

$$f(x) = \frac{\theta^x}{x!}e^{-\theta}, \quad \theta = 0, 1, 2, ...$$

puede ser reescrita como

$$f(x) = \frac{1}{x!}exp(xln(\theta) - \theta)$$

si tomamos

$$\begin{aligned} \eta(\theta) &= ln(\theta) \\ T(x) &= x \\ A(\theta) &= \theta \\ h(x) &= \frac{1}{x!} \end{aligned}$$

podemos confirmar que la distribución Poisson pertenece a la familia exponencial

5.2 Simulación

Para obtener la distribución de las ocurrencias de cada evento, consideramos que cada mes tiene su propio parámetro λ asociado a una distribución Poisson. Para obtener cada λ se crearon columnas para el mes y año en que ocurrió cada evento. Posteriormente se obtuvieron grupos jerarquizados por el mes y el año. A continuación se obtuvo el número de eventos en cada mes, para posteriormente promediar los eventos de cada mes

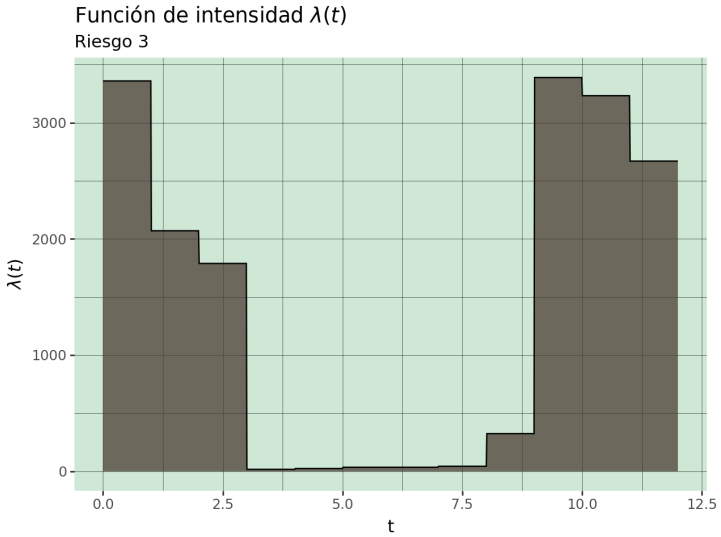
```
# Obtener el promedio de eventos por mes
valoresLambda = (tickets >>
group_by(_.Año, _.Mes) >> summarize(count(_)) >>
group_by(_.Mes) >> summarize(mediaOcurrencias = _.n.mean()))

valoresLambda
```

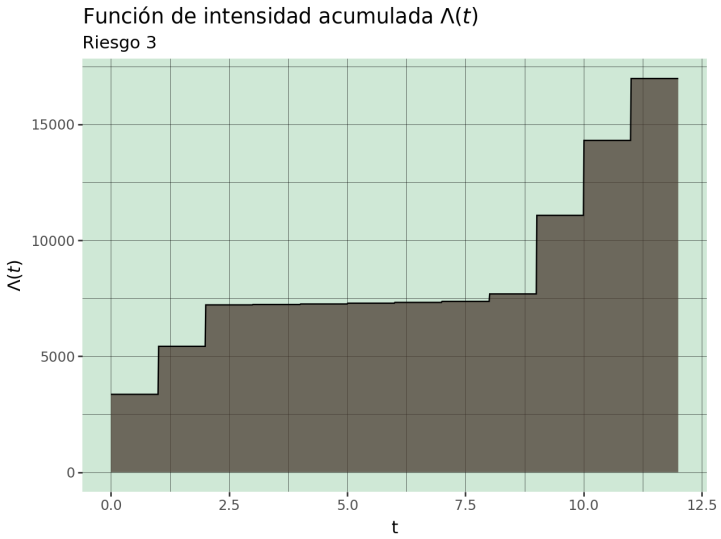
	Mes	mediaOcurrencias
0	1	3359.500000
1	2	2069.666667
2	3	1788.333333
3	4	16.000000
4	5	23.000000
5	6	34.000000
6	7	34.000000
7	8	42.500000
8	9	323.500000
9	10	3389.000000
10	11	3232.500000
11	12	2669.500000

Los resultados anteriores nos permiten construir una función $\lambda(t)$ que devuelve la intensidad de los eventos en un mes determinado. A partir de ella podemos construir la función de distribución acumulada.

```
valoresX = np.linspace(0,12,1000)
valoresY = parametrosPoissonVectorizada(valoresX)

(ggplot()+
 geom_area(mapping=aes(x=valoresX, y=valoresY),
 color='black', fill='#4B3D33', alpha=0.75, size=0.5) +
 labs(title="Función de intensidad $\lambda(t)$", subtitle='Riesgo 3',
```

```
x="t", y="$\lambda(t)$") +
theme(panel_background=element_rect(fill="#cfe8d6", color=None),
      panel_grid=element_line(color="black", size=0.2)
      )
)
```





6 Procesos Poisson

Un proceso de Poisson de parámetro $\lambda > 0$ es un proceso a tiempo continuo $\{X_t : t \geq 0\}$ con espacios de estados $\{0, 1, ...\}$, con trayectorias no decrecientes y que cumple las siguientes propiedades:

1. $X_0 = 0$
2. Tiene incrementos independientes
3. $X_{t+s} - X_s \sim Poisson(\lambda t)$, para cualesquiera $s \geq 0,\ t > 0$

6.1 Proceso Poisson no Homogéneo

Consideraremos ahora que el parámetro λ del proceso Poisson es una función del tiempo t.
Un proceso Poisson no homogéneo es un proceso estocátstico a tiempo continuo $\{X_t : t \geq 0\}$, con espacio de estados $\{0, 1, ...\}$, con parámetro la función positiva y localmente integrable $\lambda(t)$, y que cumple lo siguiente:

a) $X_0 = 0$

b) Los incrementos son independientes

c) Para cualquier $t \geq 0$, y cuando $h \to 0^+$ se tiene que

   i) $\mathbf{P}(X_{t+h} - X_t = 1) = \lambda(t)h + o(h)$
   ii) $\mathbf{P}(X_{t+h} - X_t = 0) = 1 - \lambda(t)h + o(h)$
   iii) $\mathbf{P}(X_{t+h} - X_t \geq 2) = o(h)$

Con esta definición se pueden demostrar las siguientes propiedades

1. La variable X_t en un proceso Poisson no homogéneo de parámetro $\lambda(t)$ tiene distribución $Poisson(\Lambda(t))$, en donde se define

$$\Lambda(t) = \int_0^t \lambda(s)ds$$

es decir, para $n = 0, 1, ...$

$$\mathbf{P}(X_t = n) = e^{-\Lambda(t)} \frac{[\Lambda(t)]^n}{n!}$$

2. Para el proceso Poisson no homogéneo, la variable incremento $X_{t+s} - X_s$ tiene distribución $Poisson(\Lambda(t+s) - \Lambda(s))$

3. Sea $\{X_t : t \geq 0\}$ un porceso Poisson no homogéneo de parámetro $\lambda(t)$, y función de intensidad $\Lambda(t) = \int_0^t \lambda(s)ds$. Definimos la función

$$\Lambda^{-1}(t) = inf\{u \geq 0 : \Lambda(u) = t\}$$

Entonces el proceso $\{X_{\Lambda^{-1}(t)} : t \geq 0\}$ es un proceso Poisson homogéneo de parámetro $\lambda = 1$

6.2 Proceso Poisson Compuesto no Homogéneo

Sea $\{N_t : t \geq 0\}$ un proceso Poisson y sea $Y_1, Y_2, ...$ una sucesión de variables aleatorias independientes, idénticamente distribuidas e independientes del proceso Poisson. Sea $Y_0 = 0$. El proceso Poisson compuesto se define de la siguiente forma:

$$X_t = \sum_{n=0}^{N_t} Y_n$$

Notemos que la variable X_t del proceso Poisson compuesto es la sumade las variables aleatorias Y_i, donde el número de sumandos es aleatorio. En el contexto del riesgo operacional puede interpretarse como el total de pérdidas ocasionadas por una cantidad aleatoria de eventos que a su vez tienen un valor aleatorio de pérdida.

El proceso Poisson compuesto simple las siguientes propiedades:

1. Tiene incrementos independientes y estacionarios
2. $\mathbf{E}(X_t) = \Lambda(t)\mathbf{E}(Y); \quad \Lambda(t) = \int_0^t \lambda(x)dx$
3. $Var(X_t) = \Lambda(t)\mathbf{E}(Y^2)$
4. $Cov(X_t, X_s) = \lambda\mathbf{E}(Y^2)min\{s, t\}$
5. La función generadora de momentos de la variable X_t es

$$M_{X_t}(u) = \mathbf{E}(e^{uX_t}) = exp(\lambda t(M_Y(u) - 1))$$

6.3 Simulación

Describiremos los algoritmos necesarios para la simulación de un proceso Poisson Compuesto no Homogéneo

1. **Trayectorias de un proceso poisson homogéneo**

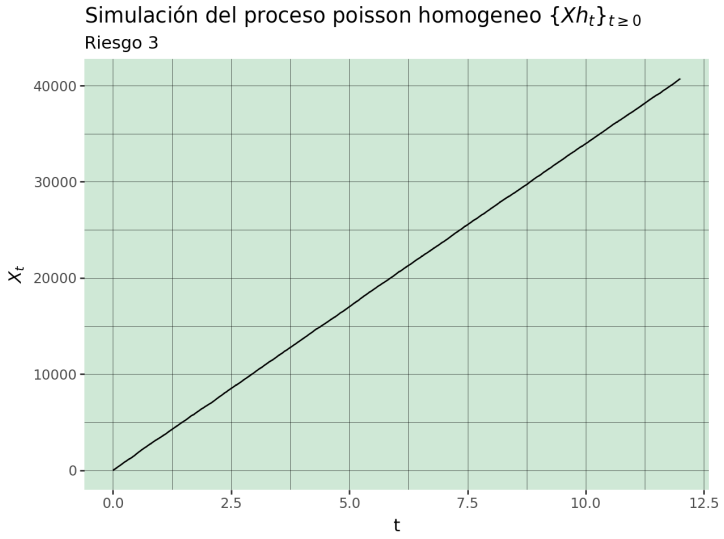
   (a) Generar un número $n \sim Poisson(l * T)$, donde l es la tasa λ de ocurrencias y T es el tiempo

   (b) Hacer $t = 0$ y $X_0 = 0$

   (c) Generar un vector t de n variables $U(0, T)$. Cada uniforme representará una ocurrencia del proceso Poisson Homogéneo

```
# Hacemos la funcion para el proceso poisson homogeneo
def procesoPoissonHom(l,T):
  n = np.random.poisson(l*T)
  if n != 0:
    vectorUniformes = np.random.uniform(low=0, high=T, size=n)
    t = np.sort(vectorUniformes)
    Xt = np.arange(1,len(t)+1)
  # Hacemos que el proceso inicie en el tiempo cero
  con cero conteos
  else:
    t = [0]
    Xt = [0]
  return t, Xt

# Llamamos al proceso poisson homogeneo
# Encontramos el valor de lambda estrella
lambdaEstrella = valoresLambda['mediaOcurrencias'].max()
np.random.seed(2024)
t, Xt = procesoPoissonHom(lambdaEstrella, 12)

# Funcion Lambda(t)
x = np.linspace(0,12)
y = lambdaEstrella*x

(ggplot() +
```

```
#geom_point(mapping=aes(x=t, y=Xt), color="red") +
geom_step(mapping=aes(x=t, y=Xt)) +
#geom_line(mapping=aes(x=x,y=y)) +
labs(title='Simulación del proceso poisson homogeneo
       $\{Xh_t\}_{t \geq 0}$', subtitle='Riesgo 3', x="t",
       y="$X_t$") +
theme(panel_background=element_rect(fill="#cfe8d6", color=None),
      panel_grid=element_line(color="black", size=0.2)
      )
)
```



2. **Trayectorias de un proceso Poisson no homogéneo X_t con intensidad $\lambda(t)$**

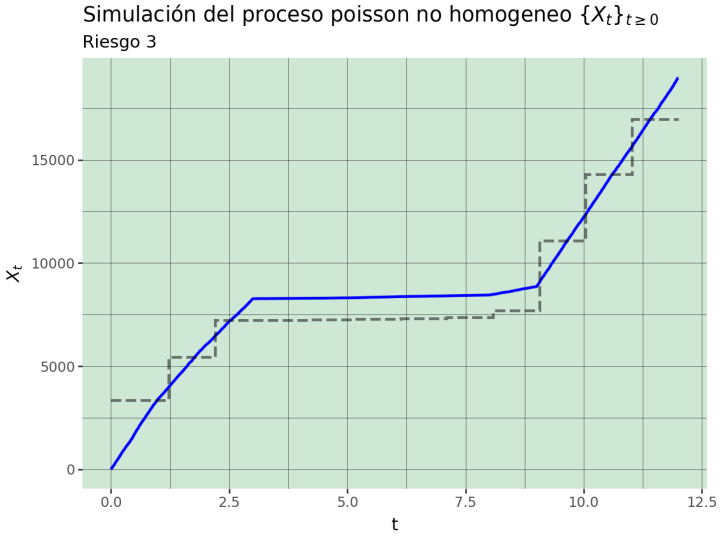
   (a) Genera un proceso poisson Homogénego Xh_t con parámetro λ que acote las intensidades del proceso Poisson no Homogéneo

   (b) Para cada observación x del proceso Xh_t genera una variable aleatoria $u \sim U(0,1)$

   (c) Si $u \leq \frac{\lambda(t)}{\lambda}$ acepta x, si no, rechaza

```
# Funcion que simula el proceso poisson no homogeneo
def procesoPoissonNoHom(T):
```

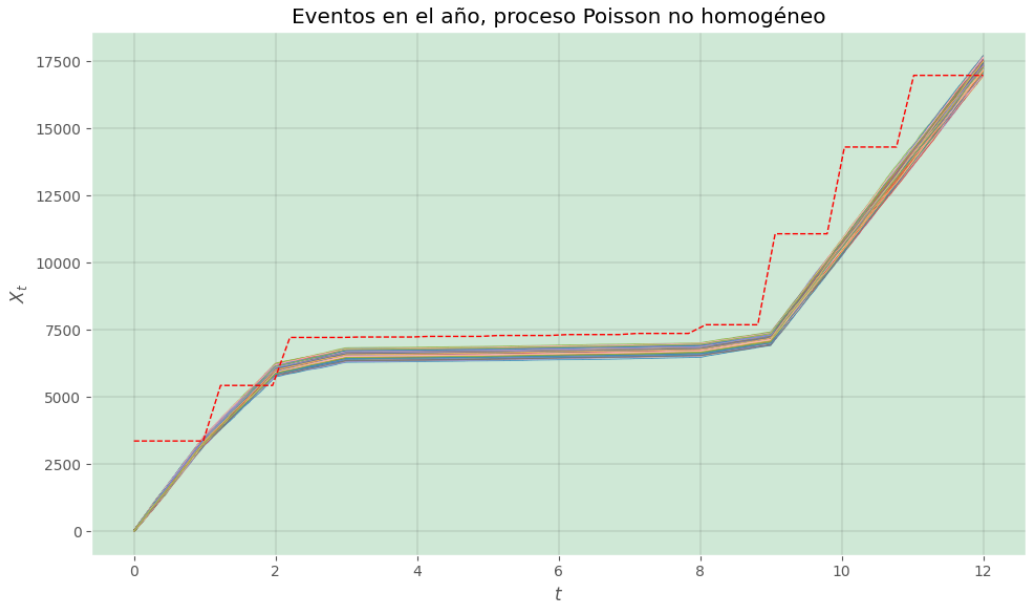
```
  valoresX = np.linspace(0,T)
  poisHom, Xt = procesoPoissonHom(lambdaEstrella, T)
  t = []
  for poisson in poisHom:
    u = np.random.uniform()
    if u <= (parametrosPoissonVectorizada(poisson) /
    parametrosPoissonVectorizada(T)):
      t.append(poisson)
  Xt = np.arange(1, len(t)+1)
  return t, Xt

# Llamar a la funcion que simula el proceso poisson no homogeneo
np.random.seed(2024)
t, Xt = procesoPoissonNoHom(12)

# Trayectoria del proceso poisson no homogeneo
x = np.linspace(0,12)
y = parametrosPoissonAcumuladosVectorizada(x)

(ggplot() +
 geom_line(mapping=aes(x=t, y=Xt), color='blue', size=1) +
 geom_step(mapping=aes(x=x, y=y), linetype='dashed',
 size=1, alpha=0.5) +
 labs(title="Simulación del proceso poisson no homogeneo
 $\{X_t\}_{t \geq 0}$",
      subtitle='Riesgo 3',
      x="t", y="$X_t$") +
 theme(panel_background=element_rect(fill="#cfe8d6", color=None),
       panel_grid=element_line(color="black", size=0.2)
       )
 )
```



3. **Generación de n trayectorias del proceso Poisson no homogéneo**
   Una vez obtenido el algoritmo para simular una trayectoria del proceso poisson no homogéneo, podemos obtener n trayectorias para analizar la distribución de los eventos totales al final del año ó en un intervalo determinado

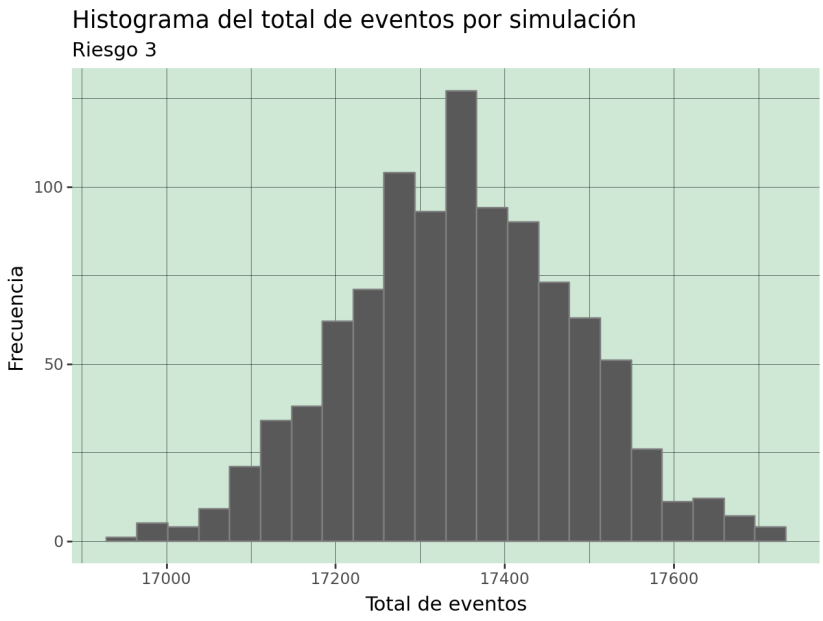
```
n = 1000
trayectorias = pd.DataFrame()
np.random.seed(2024)
for i in range(n):
  t, Xt = procesoPoissonNoHom(12)
  dataFrameTemporal = pd.DataFrame({"Trayectorias":t,
  "Conteos":Xt, "Numero de trayectoria":[i]*len(t)})
  trayectorias = pd.concat([trayectorias,
  dataFrameTemporal], ignore_index=True)

  # Grafico de las trayectorias simuladas
    # Funcion Lambda(x) = integral(lambda(t))
    x = np.linspace(0,12)
    y = parametrosPoissonAcumuladosVectorizada(x)
    plt.style.use('ggplot')
    plt.figure(figsize=(10, 6))
    # Grafico de escalera para cada simulacion
```

```
for num_trayectoria in trayectorias
['Numero de trayectoria'].unique():
    # Datos de la trayectoria actual
    df_trayectoria = trayectorias[trayectorias
    ['Numero de trayectoria'] == num_trayectoria]
    # Grafico de escalera para la trayectoria actual
    plt.step(df_trayectoria['Trayectorias'],
    df_trayectoria['Conteos'], where='post',
    alpha=0.9, linewidth=0.6)
# Graficar funcion Lambda(x) = integral(lambda(t))
plt.plot(x, y, color='red', linestyle='--',
linewidth=1, label='Lambda(x)')
plt.gca().set_facecolor("#cfe8d6") # Fondo
plt.grid(color='black', linewidth=0.1) #Cuadricula
plt.xlabel('$t$')
plt.ylabel('$X_t$')
plt.title('Eventos en el año, proceso Poisson no homogéneo')
plt.tight_layout()
plt.show()
```



Con las simulaciones realizadas podemos analizar el número de eventos obtenidos al final del año

```
# Obtenemos el numero total de eventos en cada simulacion
NoEventos = (trayectorias >> group_by('Numero de trayectoria') >> sum

# Vemos el histograma de eventros por simulacion
display(ggplot() +
        geom_histogram(mapping=aes(x=NoEventos["TotalEventos"]),
        color='grey') +
        labs(title="Histograma del total de eventos por
        simulación",
              subtitle='Riesgo 3',
              x="Total de eventos", y="Frecuencia") +
        theme(panel_background=element_rect(fill="#cfe8d6",
        color=None),
              panel_grid=element_line(color="black", size=0.2)
              )
        )
```



4. **Trayectorias de un proceso Poisson Compuesto no homogeneo**

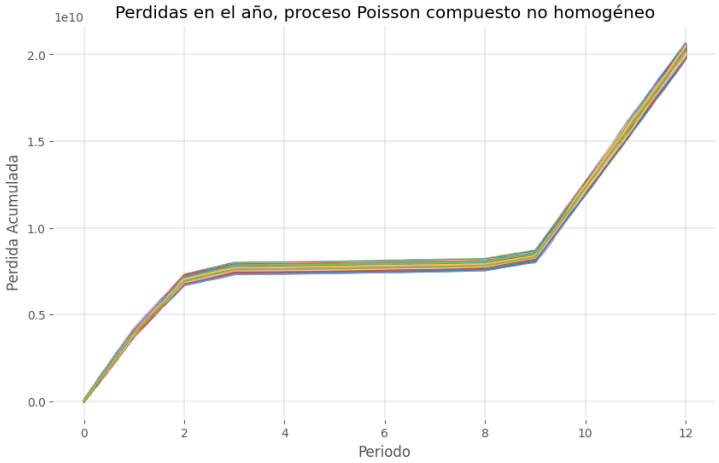
Una vez obtenidas las trayectorias del proceso Poisson compuesto, simulamos pérdidas que siguen la distribución que ya hemos estimado anteriormente y las añadimos al DataFrame para obtener las pérdidas acumuladas X_t en cada tiempo t

```
# Obtenemos el numero de trayectorias
NoTrayectorias = trayectorias['Numero de trayectoria'].max() + 1

# Creamos un dataframe vacio para almacenar las trayectorias
SimulacionesPerdidas = pd.DataFrame()

# A cada trayectoria le agregaremos una columna con la
severidad por evento y la severidad acumulada
for i in range(NoTrayectorias+1):
  # Filtramos por el numero de trayectoria
  Perdida = (trayectorias >>
                filter(_['Numero de trayectoria'] == i))
  # Agregamos una severidad a cada evento y posteriormente
  hacemos la suma acumulada de severidades
  Perdida = (Perdida >>
                mutate(Severidad =
                simulacionDistEmpirica(len(Perdida))) >>
                mutate(SeveridadAcumulada =
                _.Severidad.cumsum()))
  SimulacionesPerdidas = pd.concat([SimulacionesPerdidas, Perdida])

SimulacionesPerdidas
```

	Trayectorias	Conteos	Numero de trayectoria	Severidad	SeveridadAcumulada
0	0.000211	1	0	1.461069e+06	1.461069e+06
1	0.000402	2	0	9.714231e+05	2.432492e+06
2	0.000592	3	0	1.339064e+06	3.771556e+06
3	0.000973	4	0	7.206655e+05	4.492221e+06
4	0.001210	5	0	7.304510e+05	5.222672e+06
...	...	...	...	...	...
17346921	11.998058	17262	999	1.811253e+06	2.018512e+10
17346922	11.998379	17263	999	9.550226e+05	2.018607e+10
17346923	11.998704	17264	999	8.151342e+05	2.018689e+10
17346924	11.999390	17265	999	3.130421e+06	2.019002e+10
17346925	11.999450	17266	999	6.544119e+05	2.019067e+10

Posteriormente graficamos las trayectorias del proceso simulado

```
# Crear la figura y los ejes
plt.figure(figsize=(10, 6))

# Agrupar por 'Numero de trayectoria' y hacer una gráfica de
líneas para cada trayectoria
for trayectoria, datos in SimulacionesPerdidas.
groupby('Numero de trayectoria'):
    plt.plot(datos['Trayectorias'], datos['SeveridadAcumulada'],
    label=f'Trayectoria {trayectoria}')

# Añadir etiquetas y título
plt.gca().set_facecolor("white") # Fondo
plt.grid(color='black', linewidth=0.1) #Cuadricula
plt.xlabel('Periodo')
plt.ylabel('Perdida Acumulada')
plt.title(f'Perdidas en el año, proceso Poisson compuesto
            no homogéneo')
plt.show()
```



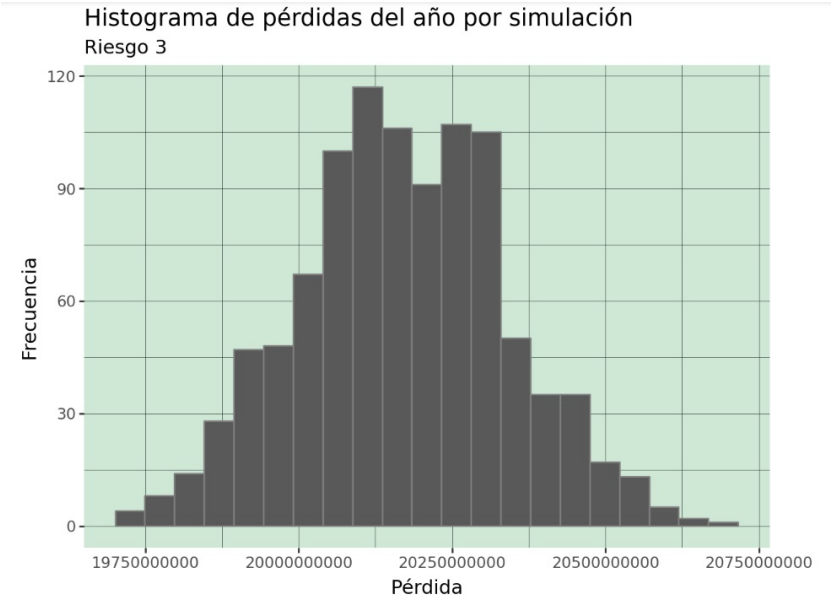
Finalmente podemos obtener las pérdidas totales en el año por simulación

```
perdidaFinal = (SimulacionesPerdidas >> group_by(_['Numero de trayect

perdidaFinal
```

	Numero de trayectoria	perdidaFinal
0	0	2.014987e+10
1	1	2.028924e+10
2	2	2.012437e+10
3	3	2.045330e+10
4	4	2.012979e+10
...	...	...
995	995	2.024784e+10
996	996	2.016655e+10
997	997	2.029237e+10
998	998	2.045088e+10
999	999	2.019067e+10

1000 rows × 2 columns

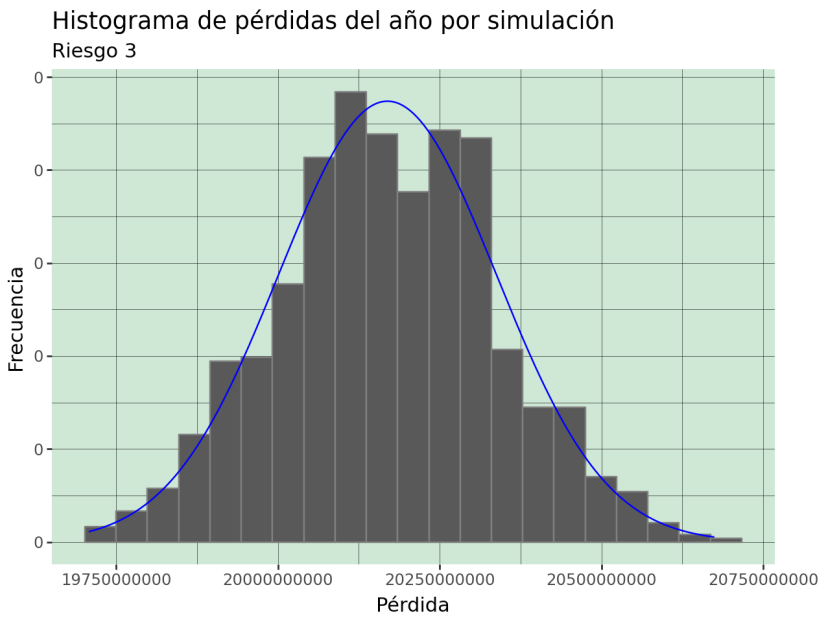
```
mediaPerdidas = np.mean(perdidaFinal['perdidaFinal'])
sdPerdidas = np.std(perdidaFinal['perdidaFinal'])
print(f'La media de las perdidas es: {mediaPerdidas} ')
print(f'La desviacion estandar de las perdidas es:
{sdPerdidas} ')

(ggplot(data = perdidaFinal) +
 geom_histogram(mapping=aes(x='perdidaFinal'),
                color='grey') +
 labs(title="Histograma de pérdidas del año por
        simulación",
      subtitle='Riesgo 3',
      x="Pérdida", y="Frecuencia") +
 theme(panel_background=element_rect(fill="#cfe8d6", color=None),
       panel_grid=element_line(color="black", size=0.2)
       )
 )
```



Podemos observar que tiene una distribución aproximadamente normal, por lo que podemos hacer pruebas estadístics para confirmar esta hipótesis

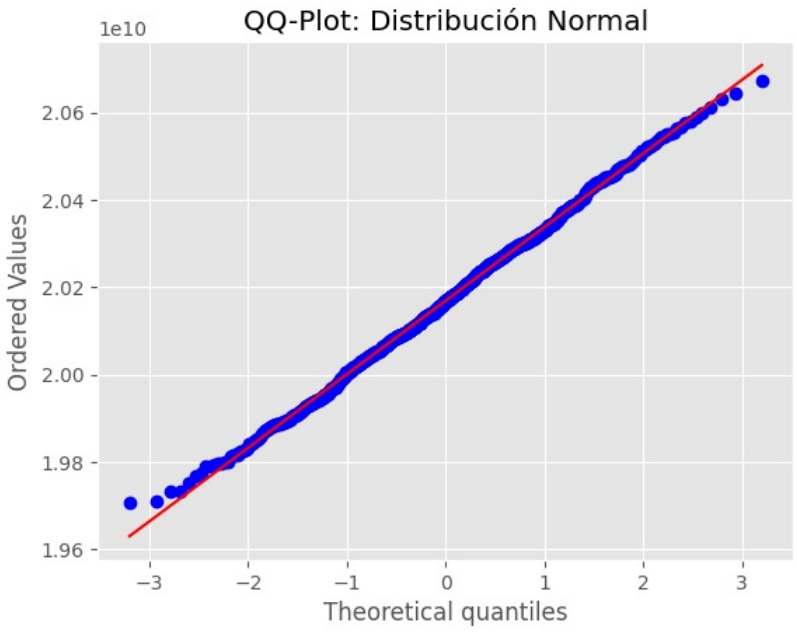
```
stat, p_value = stats.shapiro(perdidaFinal['perdidaFinal'])

if p_value > 0.05:
    print(f"p-valor = {p_value},
    los datos siguen una distribución normal (no se rechaza H0)")
else:
    print(f"p-valor = {p_value},
    los datos no siguen una distribución normal (se rechaza H0)")

p-valor = 0.34962107166858747,
los datos siguen una distribución normal (no se rechaza H0)

min = np.min(perdidaFinal['perdidaFinal'])
max = np.max(perdidaFinal['perdidaFinal'])
mediaPerdidas = np.mean(perdidaFinal['perdidaFinal'])
sdPerdidas = np.std(perdidaFinal['perdidaFinal'])
```

```
ejeX = np.linspace(min, max, 10000)
ejeY = norm.pdf(ejeX, loc=mediaPerdidas, scale=sdPerdidas)


(ggplot() +
 geom_histogram(data = perdidaFinal, mapping=aes(x='perdidaFinal',
 y='..density..'), color='grey') +
 geom_line(mapping=aes(x=ejeX, y=ejeY), color='blue') +
 labs(title="Histograma de pérdidas del año por simulación",
      subtitle='Riesgo 3',
      x="Pérdida", y="Frecuencia") +
 theme(panel_background=element_rect(fill="#cfe8d6", color=None),
       panel_grid=element_line(color="black", size=0.2)
       )
 )
```




```
normales = np.random.normal(loc=0, scale=1, size=1000)

# Generar el QQ-plot
fig = plt.figure()
ax = fig.add_subplot(111)
```

```
# Realizar el QQ-plot comparando los cuantiles de la normal
stats.probplot(perdidaFinal['perdidaFinal'], dist="norm", plot=ax)

# Mostrar el gráfico
plt.title('QQ-Plot: Distribución Normal')
plt.show()
```



7 Valores extremos

Los tipos de riesgos pueden variar desde los de alta frecuencia - baja severidad (HFLS), hasta los de baja frecuencia - alta severidad (LFHS). Las pérdidas de tipo HFLS son descritas por el modelo de un proceso Poisson Compuesto no homogéneo. Si bien podría decrise lo mismo de las pérdidas de tipo LFHS, se puede hacer un enfoque en este tipo de riesgos que tienen pérdidas muy grandes cuando ocurren.

Llamaremos pérdidas Jumbo a aquellas que con una sola ocurrencia pueden tener un gran impacto en una organización. Se tienen algunos problemas gestionando este tipo de riesgos. En primer lugar, hay varios tipos de riesgo cuya pérdida nunca ha ocurrido, por lo que los datos no aportan información para hacer un análisis estadístico. En segundo lugar, al estudiar valores extremos sólo se tiene una observación por año de la pérdida más grande.

Existen alternativas para estos casos. Se puede estudiar la pérdida más grande por mes en vez de cada año, esto incrementa a 12 la cantidad de valores extremos, posteriormente se pueden trasladar los resultados mensuales a anuales. Otra alternativa es estudiar varias de lás pérdidas "más grandes" cada año en vez de la más grande, tomando algún umbral para decidir cuáles son las "más grandes".

La teoría de valores extremos (EVT) estudia la forma asintótica de la distribución de las observaciones más grandes. En el entorno del riesgo operacional nos concentramos no solo en el evento más severo, porque estamos interesados en el impacto de todas las pérdidas operacionales. El estudio de valores extremos **no** nos otorga una visión completa de todo el impacto, a menos que las pérdidas extremas sean mucho mayores a las menores. Si este es el caso, entender el impacto potencial de las pérdidas extremas nos da la oportunidad de desarrollar acciones de mitigación que favorezcan el entorno de control para estos valores extremos.

Un resultado clave en EVT es que la distribución límite de la observación más grande debe ser una de un número muy pequeño de distribuciones. Similarmente, en un resultado muy cercano, la distribución límite de una muestra por encima de un umbral debe ser una de un número muy pequeño de distribuciones. Esta teoría nos permite extrapolar las pérdidas cantidades de pérdidas que son muy superiores a cualquier pérdida histórica y, por lo tanto, darnos una idea de la probabilidad de pérdidas enormes, incluso cuando estas pérdidas nunca antes han ocurrido.

Hay tres distribuciones relacionadas a valores extremos.

1. Distribución Gumbel

$$F_X(x) = exp\left(-exp\left(-\frac{x-\mu}{\theta}\right)\right),\ x > 0,\ \theta > 0$$

2. Distribución de Fréchet

$$F_X(x) = exp\left(-\left(\frac{x-\mu}{\theta}\right)^{-\alpha}\right),\ x \geq \mu;\ \alpha, \theta > 0$$

3. Distribución Weibull

$$F_X(x) = exp\left(-\left(-\frac{x-\mu}{\theta}\right)^{-\alpha}\right),\ x \leq \mu,\ \alpha < 0$$

La función de valores extremos generalizada es la familia de distribuciones que incorpora en una sola expresión las tres mencionadas anteriormente. La expresión general para la función de densidad estandarizada para valores extremos es

$$F_X(x) = G_\gamma(x) = exp\left(-(1+\gamma x)^{-\frac{1}{\gamma}}\right)$$

Cuando $\gamma > 0$, $G_\gamma(x)$ toma la forma de una distribución de Fréchet.
Cuando $\gamma < 0$, $G_\gamma(x)$ toma la forma de una distribución Weibull.
Además, $\lim_{\gamma \to 0} G_\gamma(x) = e^{-e^x}$, que es una Gumbel estandarizada.

7.1 Distribuición del máximo

7.1.1 Número fijo de péridas

Sea n fijo y $\{X_i\}_{i=1}^n$ una muestra aleatoria no negativa con distribución $F_X(x)$, donde n es un número fijo. Sea $M_n = max\{X_i\}$ y denotemos su función de distribución acumulada por $F_n(x)$
Entonces tenemos que

$$F_n(x) = \mathbf{P}(M_n \leq x) = [F_X(x)]^n$$

Por lo que la distribución del máximo es una función de la distribución de la varibale aleatoria original. Cuando $n \to \infty$ F_n se aproxima a 0 ó 1 dependiendo

si $F_X(x) < 1$ ó $F_X(x) = 1$. De este modo la distribución límite es degenerada, para evitar esto para valores grandes de n se requiere normalizar.
Para variables aleatorias no negativas, la media del máximo (si existe) puede calcularse como

$$\mathbf{E}(M_n) = \int_0^\infty x f_n(x) dx = \int_0^\infty [1 - F_X(x)^n] dx$$

El segundo momento (si existe) puede ser calculado como

$$\mathbf{E}(M_n^2) = \int_0^\infty x^2 f_n(x) dx = 2 \int_0^\infty x[1 - F_X(x)^n] dx$$

Ejemplo. Máximo mensual a anual
Supongamos que hemos hecho un estudio de las mayores pérdidas en cada mes y determinamos la distribución del máximo mensual dada por $F(x)$. La distribución del máximo anual estará dada por $[F(x)]^{12}$.

Si el máximo mensual sigue una distribución Gumbel, entonces el máximo anual tendra la siguiente distribución:

$$\begin{aligned}
[F(x)]^{12} &= \left(exp\left(-exp\left(-\frac{x-\mu}{\theta} \right) \right) \right)^{12} \\
&= exp\left(-12exp\left(-\frac{x-\mu}{\theta} \right) \right) \\
&= exp\left(-exp\left(-\frac{x-\mu^*}{\theta} \right) \right)
\end{aligned}$$

Donde $\mu^* = \mu + \theta ln 12$
Si ahora el máximo mensual sigue una distribución de Fréchet, entonces el máximo anual tendra la siguiente distribución:

$$\begin{aligned}
[F(x)]^{12} &= exp\left[-12\left(\frac{x-\mu}{\theta} \right)^{-\alpha} \right] \\
&= exp\left[-12\left(\frac{x-\mu}{\theta^*} \right)^{-\alpha} \right]
\end{aligned}$$

Donde $\theta^* = 12^{-\frac{1}{\alpha}}\theta$

7.1.2 Número aleatorio de pérdidas

Recordemos que si F_n es la distribución del máximo de una muestra aleatoria, entonces

$$F_n(x) = \prod_{i=1}^{n} [F_X(x)]^n$$

Esto asume que el tamaño de la muestra en cada periodo es fijo. Sin embargo, en modelos de riesgo operacional el número de pérdidas es desconocido, por lo que nos interesa estudiar el comportamiento de un número aleatorio de pérdidas. Para este caso podemos obtener una expresión para el número (aleatorio) de pérdidas en un periodo de tiempo fijo.

Sea N una variable aleatoria que denota el número de pérdidas en un tiempo determinado, su función de densidad será denotada $P_N(z)$, además conservaremos las hipótesis de de independencia. Ahora consideraremos la distribución del máximo de las pérdidas M_N donde N es un valor aleatorio.

$$\begin{aligned} F_{M_N} &= \mathbf{P}(M_N \leq x) \\ &= \sum_{n=0}^{\infty} \mathbf{P}(M_N \leq x | N = n)\mathbf{P}(N = n) \\ &= \sum_{n=0}^{\infty} \mathbf{P}(N = n)[F_X(x)]^n \\ &= \mathbf{P_N}(F_X(x)) \end{aligned}$$

Entonces, si podemos obtener la distribución de la frecuencia y la severidad de las pérdidas, fácilmente podemos obtenr la distribución de la pérdida máxima, esta última tiene soporte para valores no negativos de x por lo que tiene un salto en el valor $x = 0$ con valor $\mathbf{P_N}(F_X(0))$, que es la probabilidad de que no haya costo en ningún incidente. Más aún, si $F_X(0) = 0$ todos los incidentes tendrán un costo positivo, como es común en la práctica.

Ejemplo Sea un proceso Poisson que genera pérdidas Poisson a tasa λ por año. Entonces tenemos que

$$P(N = n) = \frac{e^{\lambda}\lambda^n}{n!}$$

De la ecuación obtenida anteriormente tenemos que la función de probabilidad acumulada de la pérdida máxima está dada por

$$\begin{aligned} F_{M_N}(x) &= \mathbf{P_N}(F_X(x) \\ &= \sum_{n=0}^{\infty} \mathbf{P}(N=n)[F_X(x)]^n \\ &= \sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!}[F_X(x)]^n \\ &= \sum_{n=0}^{\infty} \frac{e^{-\lambda}(\lambda F_X(x))^n}{n!} \\ &= e^{-\lambda}\sum_{n=0}^{\infty} \frac{(\lambda F_X(x))^n}{n!} \\ &= e^{-\lambda}e^{\lambda F_X(x)} \\ &= e^{-\lambda(1-F_X(x))} \end{aligned}$$

y para una periodo de k años será

$$F_{M_N}(x) = exp[-\lambda k(1 - F_X(x))]$$

Ejemplo Supongamos además que las pérdidas independientes tienen una distribución exponencial con la siguiente función de distribución acumulada

$$F_X(x) = 1 - exp\left(-\frac{x}{\theta}\right),\ x > 0$$

Entonces la distribución de la pérdida máxima en un periodo de k años será

$$F_{M_N}(x) = exp\left[-k\lambda exp\left(-\frac{x}{\theta}\right)\right]$$

La cual puede ser reescrita como

$$F_{M_N}(x) = exp\left[-exp\left(-\frac{x-\mu}{\theta}\right)\right],\ x > 0$$

donde $\mu = \theta ln(k\lambda)$.
Esta es la distribución de una $Gumbel(x; \theta, \mu)$

Ejemplo Supongamos que las pérdidas individuales siguen una distribución de Pareto con con la siguiente distribución acumulada

$$F(x) = 1 - \left(\frac{x+\beta}{\beta}\right)^{\alpha}, \quad x \geq 0; \quad \alpha, \beta > 0$$

Entonces la distribución de la pérdida máxima en un periodo de k años tiene la siguiente distribución

$$F_{M_N}(x) = exp\left[-k\lambda\left(\frac{x+\beta}{\beta}\right)^{-\alpha}\right]$$

La cual puede ser reescrita como

$$F_{M_N}(x) = exp\left[-\left(\frac{x-\mu}{\theta}\right)^{-\alpha}\right]$$

Donde $\theta = \frac{\beta}{(k\lambda)^{\frac{1}{\alpha}}}$ y $\mu = -\beta$
Esta es es la distribución de una $Frechet(x; \alpha, \mu, \theta)$

7.2 Estabilidad de la distribución del máximo

Las distribuciones Gumbel, Fréchet y Weibull tienen otra propiedad llamada estabilidad del máximo, la cual es muy útil en teoría de valores extremos.
Podemos notar que

$$\begin{aligned}[G_0(x + ln(n))]^n &= exp[-nexp(-x - ln(n)] \\ &= exp[-exp(-x)] \\ &= Gumbel(x)\end{aligned}$$

Equivalentemente,

$$[Gumbel(x)]^n = Gumbel(x - ln(n))$$

Esto muestra que la distribución del máximo de n observaciones de una Gumbel estandarizada tiene dentro una distribución Gumbel después de una traslación de

$ln(n)$, de modo que

$$\begin{aligned}
[Gumbel(x;\mu,\theta)]^n &= \left[Gumbel\left(\frac{x-\mu}{\theta}\right)\right]^n \\
&= Gumbel\left(\frac{x-\mu}{\theta} - ln(n)\right) \\
&= Gumbel\left(\frac{x-\mu-\theta ln(n)}{\theta}\right) \\
&= Gumbel\left(\frac{x-\mu^*}{\theta}\right) \\
&= Gumbel(x;\mu^*,\theta)
\end{aligned}$$

Donde $\mu^* = \mu + \theta ln(n)$

Similarmente para la distribución Fréchet tenemos que

$$\begin{aligned}
[Frechet(n^{\frac{1}{\alpha}}x;\alpha)]^n &= exp\left(-n\left(n^{1/\alpha}x\right)^{-\alpha}\right) \\
&= exp\left(-x^{-\alpha}\right) \\
&= Frechet(x;\alpha)
\end{aligned}$$

De manera equivalente

$$[Frechet(x;\alpha)]^n = Frechet\left(\frac{x}{n^{1/\alpha}};\alpha\right)$$

Esto muestra que la distribución del máximos de n observaciones de una Fréchet estandarizada, despúes de un cambio de escala, tiene una distribución de Fréchet.

$$\begin{aligned}
[Frechet(x;\alpha,\mu,\theta)]^n &= Frechet\left(\frac{x-\mu}{\theta n^{1/\alpha}};\alpha\right) \\
&= Frechet(x;\alpha,\mu,\theta^*)
\end{aligned}$$

Donde $\theta^* = \theta n^{1/\alpha}$

La idea clave es que la distribución del máximo después de una normalización de localización ó escala para cada distribución de valores extremos (VE) tiene la misma distribución VE

7.3 Teorema de Fisher-Tippett

Ahora estudiaremos la distribución del valor máximo de una muestra de tamaño n fijo cuando la muestra viene de cualquier distribución. Cuando $n \to \infty$, la distribución del máximo es degenerada, por lo que para entender la forma de la distribución para valores grandes de n, será necesario normalizar la variable aleatoria que representa el máximo. Requeriremos transformaciones lineales tales que

$$\lim_{n\to\infty} F_n\left(\frac{x-b_n}{a_n}\right) = G(x)$$

Para todos los valores de x, donde $G(x)$ es una distribución no degenerada.
Si dicha transformación lineal existe, el teorema siguiente da un resultado muy útil que constituye un elemento fundamental en la teoría de valores extremos.

Teorema de Fisher-Tippet
Si $\left[F\left(\frac{x-b_n}{a_n}\right)\right]^n$ tiene una distribución límite no degenerada cuando $n \to \infty$, para algunas constantes a_n y b_n que dependen de n, entonces

$$\left[F\left(\frac{x-b_n}{a_n}\right)\right]^n \longrightarrow G(x)$$

cuando $n \to \infty$, para todos los valores de x, para alguna distribución de valores extremos G, la cual es una entre Gumbel, Frechet, Weibull con algún parámetro de localización y escala.

Este es un resultado muy importante si estamos interesados en entender cómo se comportan las pérdidas grandes. Sólo tenemos que mirar en tres elecciones para nuestro modelo de cola derecha (incluso sólo dos, ya que la Weibull tiene un límite superior).

El teorema de Fisher Tippett requiere una normalización usando las constantes apropiadas que dependen de n. Para algunas distribuciones especiíficas, estas constantes pueden ser identificadas. en la sección "Numero aleatorio de pérdidas" ya hemos visto algunos ejemplos.

Ejemplo. Máximo de exponenciales.
Sin pérdida de generalidad, por conveniencia de notación, se utilizará la versión

estandarizada de la distribución exponencial. Utilizando las constantes de normalización $a_n = 1$ y $b_n = -ln(n)$ la distribución del máximo está dada por

$$\begin{aligned}
\mathbf{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) &= \mathbf{P}(M_n \leq a_n x + b_n) \\
&= [\mathbf{P}(X \leq a_n x + b_n)]^n \\
&= [\mathbf{P}(X \leq x + ln(n))]^n \\
&= [1 - exp(-x - ln(n))]^n \\
&= \left[1 - \frac{e^{-x}}{n}\right]^n \underset{n\to\infty}{\longrightarrow} e^{e^{-x}}
\end{aligned}$$

Habiendo elegido las constantes de normalización adecuadas, podemos ver que la distribución límite del máximo de una variable aleatoria exponencial es una distribución Gumbel.

Ejemplo. Máximo de una Pareto.
Utilizando la densidad de una Pareto tenemos que

$$\begin{aligned}
S(x) &= \left(\frac{x+\theta}{\theta}\right)^{-\alpha} \\
&= \left(1 + \frac{x}{\theta}\right)^{-\alpha}, \quad x \geq 0, \quad \alpha, \theta > 0
\end{aligned}$$

y las constantes de normalización $a_n = \theta n^{1/\alpha}/\alpha$ y $b_n = \theta n^{1/\alpha} - \theta$

$$\begin{aligned}
\mathbf{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) &= \mathbf{P}(M_n \leq a_n x + b_n) \\
&= [\mathbf{P}(X \leq a_n x + b_n)]^n \\
&= \left[\mathbf{P}\left(X \leq \frac{\theta n^{1/\alpha}}{\alpha} x + \theta n^{1/\alpha} - \theta\right)\right]^n \\
&= \left[1 - \left(1 + \frac{\frac{\theta n^{1/\alpha}}{\alpha} x + \theta n^{1/\alpha} - \theta}{\theta}\right)^{-\alpha}\right]^n \\
&= \left[1 - \frac{1}{n}\left(1 + \frac{x}{\alpha}\right)^{-\alpha}\right]^n \underset{n\to\infty}{\longrightarrow} exp\left(-\left(1 + \frac{x}{\alpha}\right)^{-\alpha}\right)
\end{aligned}$$

Esto muestra que el máximos de una variable aleatoria Pareto tiene una distribución de Fréchet con $\mu = -\alpha$ y $\theta = \alpha$

7.4 Dominio de atracción máximo

El **dominio de atracción máximo** (MDA) para cualquier distribución G, es el conjunto de de todas las distribuciones que tienen a G como distribución límite cuando $n \to \infty$ para la normalización del máximo $\frac{M_n - b_n}{a_n}$ con algunas constantes de normalización a_n y b_n.

Las distribuciones con límites no degenerados se pueden dividir en tres clases: Gumbel ($G_{0,\mu,\theta}$), Fréchet ($G_{1,\alpha,\mu,\theta}$) y Weibull ($G_{2,\alpha,\mu,\theta}$). Si podemos encontrar la distribución límite y estamos interesados en modelar los valores extremos podemos tratar la distribución límite como una representación aproximada de la distribución del valor extremo.

Teorema Caracterización del dominio de atracción máximo (MDA) por colas
Una distribuciín F pertenece al dominio de atracción máximo de una distribución de valor extremo Gi con constantes de normalización a_n y b_n sí y sólo sí

$$\lim_{n\to\infty} nS(a_n x + b_n) = -ln(G_i)$$

Ejemplo Máximo de exponenciales
Se utilizará la versión estandarizada de la distribución exponencial. Usado las constantes de normalización $a_n = 1$ y $b_n = -ln(n)$, la distribución del máximo está dada por

$$\begin{aligned} nS(x + b_n) &= n\mathbf{P}(X > x + ln(n)) \\ &= ne^{-x-ln(n)} \\ &= n\frac{e^{-x}}{n} \\ &= -ln(G_0(x)) \end{aligned}$$

La distribución límite del máximo de una variable exponencial es una distribución Gumbel.

Es conveniente identificar funciones que tienen una cola con un forma asintóticamente igual. El siguiente ejemplo muestra que si alguna distribución tiene una cola asintóticamente igual a la de una exponencial, entonces la distribución límite

del máximo debería ser una Gumbel. Entonces dos distribuciones F_X y F_Y son equivalentes en colas sí

$$\lim_{x\to\infty} \frac{S_X(x)}{S_Y(x)} = c$$

donde c es una constante.
Si dos distribuciones son equivalentes en colas, entonces tendrán el mismo dominio de atracción máximo, ya que la constante c será absorbida por las constantes de normalización.

Ejemplo Máximo de Pareto

$$S(x) = \left(\frac{x+\theta}{\theta}\right)^{-\alpha},\ x \geq 0,\ \alpha, \theta > 0$$

Utilizaremos las constantes de normalización $a_n = \theta n^{-1/\alpha}$ y $b_n = 0$ y la equivalencia de cola

$$S(x) \sim \left(\frac{x}{\theta}\right)^{\alpha}$$

Para valores grandes de x tenemos que

$$\begin{aligned}\lim_{n\to\infty} nS(a_n x + b_n) &\sim \lim_{n\to\infty} n\left(\frac{\theta x}{\theta n^{1/\alpha}}\right)^{-\alpha} \\ &= x^{-\alpha} \\ &= -ln(G_1(x))\end{aligned}$$

Esto muestra que el máximo de una distribución Pareto tiene una distribución Fréchet.

Por la equivalencia de colas, todas las distribuciones de la forma $cx^{-\alpha}$ están en el dominio de atracción máximo de una Fréchet.
Y todas las distribuciones con una cola asintóticamente de la forma $ke^{-x/\theta}$ están en el dominio de atracción máximo de una Gumbel.

Teorema Si una distribución tiene una cola derecha caracterizada por $S(x) \sim x^{-\alpha}C(x)$, donde $C(x)$ es una función que varía lentamente, entonces está en el dominio de atracción máximo de una Fréchet.
Una función $C(x)$ varía lentamente sí

$$\lim_{x\to\infty} \frac{C(tx)}{C(x)} = 1,\ \text{para todo } t > 0$$

Las distribuciones que están en el dominio de atracción máximo de una Gumbel no son fáciles de caracterizar. Las colas de las distribuciones en el dominio de atracción máximo son muy diferentes entre sí, desde colas ligeras como es el caso de la distribución normal, hasta colas muy pesadas como la distribución Gaussiana inversa.

7.5 Estimación de parámetros

Según lo visto anteriormente, dada una muestra aleatoria $\{X_i\}_{i=1}^n$, podemos encontrar la distribución de la cola derecha. Ahora el siguiente paso es encontrar los parámetros de la distribución

7.5.1 Regresión lineal

La regresión lineal simple es un modelo estadístico utilizado para describir la relación entre una variable dependiente y y una variable independiente x. El objetivo principal es ajustar una linea recta que minimice la diferencia entre los valores observados de y y los valores predichos por el modelo.

El modelo de regresión lineal simple se puede expresar de la siguiente manera:

$$y = \beta_0 + \beta_1 + \varepsilon$$

Donde:

- y es la variable dependiente
- x es la variable independiente
- β_0 es el intercepto
- β_1 es la pendiente
- ε es el error

Los parámetros β_0 y β_1 se estiman mediante el método de mínimos cuadrados ordinarios, que minimiza la suma de los cuadrados de las diferencias entre los

valores observados y los predichos de y, esto es, minimizar la función

$$\begin{aligned} f(\beta_0, \beta_1; \underline{x}) &= \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \\ &= \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \end{aligned}$$

Los valores de los parámetros son los siguientes

$$\begin{aligned} \hat{\beta}_0 &= \frac{1}{n} \left(\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i \right) \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \end{aligned}$$

7.5.2 Distribución Gumbel

La distribución Gumbel tiene la siguiente función de distribución acumulada

$$F(t; \mu\lambda) = exp\left(-exp\left(-\frac{t-\mu}{\lambda} \right) \right)$$

La cual podemos obtener de manera empírica a paritr de la muestra

Sea

$$\begin{aligned} z &= ln(-ln(F(t))) \\ &= -\frac{t-\mu}{\lambda} \\ &= \frac{\mu}{\lambda} - \frac{1}{\lambda} t \end{aligned}$$

Tomamos el modelo $y = \beta_0 + \beta_1 x$
Donde $y = ln(-ln(F(t)))$
Entonces tenemos que

$$\begin{aligned} \beta_0 &= \frac{\mu}{\lambda} \\ \beta_1 &= -\frac{1}{\lambda} \\ x &= t \end{aligned}$$

Por lo que los parámetros serán

$$\lambda = -\frac{1}{\beta_1}$$
$$\mu = -\frac{\beta_0}{\beta_1}$$

7.5.3 Distribución Fréchet

La distribución Fréchet tiene la siguiente función de distribución acumulada

$$F(x;\mu,\lambda,\alpha) = exp\left(-\frac{x-\mu}{\theta}\right)^{-\alpha},\ x \geq \mu,\ \alpha,\theta > 0$$

La cual podemos obtener de manera empírica a paritr de la muestra
El parámetro μ representa la mínima pérdida que se puede presentar, consideraremos este valor conocido.

Sea

$$\begin{aligned} z &= ln(-ln((F_X(x))) \\ &= -\alpha ln\left(\frac{x-\mu}{\theta}\right) \\ &= -\alpha ln(\theta) - \alpha ln(x-\mu) \end{aligned}$$

Tomamos el modelo $y = \beta_0 + \beta_1 x^*$
Donde $y = ln(-ln(F_X(x)))$ y $x^* = ln(x-\mu)$
Entonces tenemos que

$$\beta_0 = \alpha ln(\theta)$$
$$\beta_1 = -\alpha$$

Por lo que los parámetros serán

$$\alpha = -\beta_1$$
$$\theta = exp\left(-\frac{\beta_0}{\beta_1}\right)$$

7.5.4 Distribución Weibull

La distribución Weibull tiene la siguiente función de distribución acumulada

$$F(x;\mu,\lambda,\alpha) = exp\left(-\left(-\frac{x-\mu}{\theta}\right)^{-\alpha}\right),\ x \leq \mu,\ \alpha,\theta > 0$$

La cual podemos obtener de manera empírica a paritr de la muestra
El parámetro μ representa la mínima pérdida que se puede presentar, consideraremos este valor conocido.

Sea

$$\begin{aligned} z &= ln(-ln((F_X(x))) \\ &= -\alpha ln\left(-\frac{x-\mu}{\theta}\right) \\ &= -\alpha ln(\theta) - \alpha ln(\mu - x) \end{aligned}$$

Tomamos el modelo $y = \beta_0 + \beta_1 x^*$
Donde $y = ln(-ln(F_X(x)))$ y $x^* = ln(\mu - x)$
Entonces tenemos que

$$\begin{aligned} \beta_0 &= \alpha ln(\theta) \\ \beta_1 &= -\alpha \end{aligned}$$

Por lo que los parámetros serán

$$\begin{aligned} \alpha &= -\beta_1 \\ \theta &= exp\left(-\frac{\beta_0}{\beta_1}\right) \end{aligned}$$

7.6 Simulación de pérdidas totales

Para hacer un estudio de los valores extremos primero necesitamos definir un umbral a partir del cual los valores serán considerados como tal. Podemos utilizar un valor fijo o un cuantil, en las organizaciones comúnmente se utiliza un valor fijo para determinar si un evento es de alta severidad. Para efectos de esta simulación utilizaremos el valor 20500000000 como umbral, valores posteriores

serán considerados como pérdidas acumuladas de alto impacto en el año.

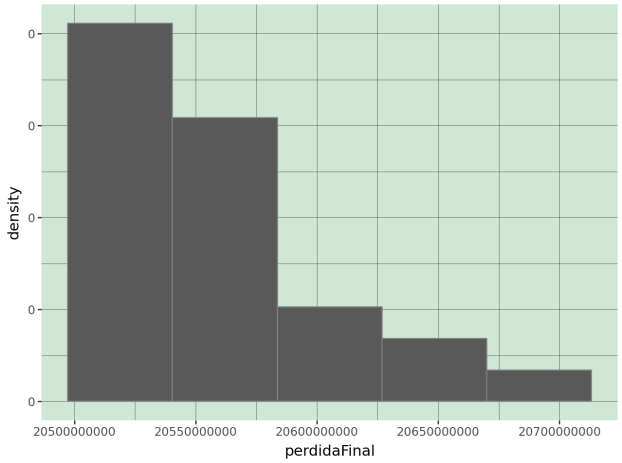
Una vez obtenidos los valores de alto impacto podemos estimar la función de distribución empírica con el objetivo de determinar si existe la convergencia a alguna de las funciones de valores extremos. El primer candidato es la distribución Gumbel, ya que no es una distribución acotada como las otras dos, por lo que procederemos a estimar sus parámetros a través de una regresión lineal.

```
n=perdidaFinal.shape[0]
regresionLineal = (perdidaFinal >>
                   arrange(-_.perdidaFinal) >>
                   mutate(i=range(1, n+1),
                          St=_.i/(n+0.0000000000001),
                          # sumamos 1*10^{-13} porque
                          en la perdida mas grande Ft da infinito
                          Ft = 1-_.St,
                          y = np.log(-np.log(_.Ft)),
                          const = 1,)
                   )

regresionLinealSeveros = (regresionLineal >>
                 filter(_.perdidaFinal >= 20500000000))
regresionLinealSeveros.tail()
```

	Numero de trayectoria	perdidaFinal	i	St	Ft	y	const
485	485	2.051274e+10	23	0.023	0.977	-3.760649	1
425	425	2.051256e+10	24	0.024	0.976	-3.717580	1
732	732	2.050448e+10	25	0.025	0.975	-3.676247	1
57	57	2.050331e+10	26	0.026	0.974	-3.636516	1
582	582	2.050170e+10	27	0.027	0.973	-3.598264	1

```
(ggplot() +
 geom_histogram(data = regresionLinealSeveros,
 mapping=aes(x='perdidaFinal', y='..density..'),
 color='grey') +
 theme(panel_background=element_rect(fill="#cfe8d6", color=None),
       panel_grid=element_line(color="black", size=0.2)
```

```
        )
)
```




```
predictors = ["const", "perdidaFinal"]
target = "y"
fit = sm.OLS(regresionLinealSeveros[target],
        regresionLinealSeveros[predictors]).fit()
display(fit.summary())
```

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.993
Model:	OLS	**Adj. R-squared:**	0.992
Method:	Least Squares	**F-statistic:**	3411.
Date:	Sun, 01 Dec 2024	**Prob (F-statistic):**	2.98e-28
Time:	03:43:29	**Log-Likelihood:**	33.209
No. Observations:	27	**AIC:**	-62.42
Df Residuals:	25	**BIC:**	-59.83
Df Model:	1		
Covariance Type:	nonrobust		

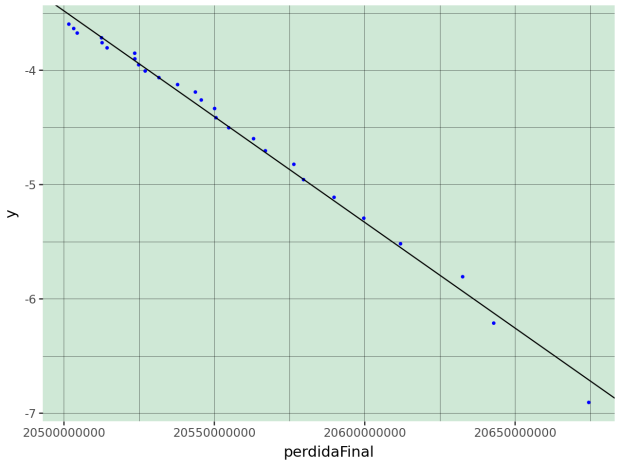
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	375.9153	6.514	57.710	0.000	362.500	389.331
perdidaFinal	-1.851e-08	3.17e-10	-58.402	0.000	-1.92e-08	-1.79e-08

Omnibus:	3.967	**Durbin-Watson:**	0.929
Prob(Omnibus):	0.138	**Jarque-Bera (JB):**	2.499
Skew:	-0.711	**Prob(JB):**	0.287
Kurtosis:	3.449	**Cond. No.**	9.47e+12

Al hacer la regresión lineal obtenemos muy buenos resultados de ajuste, una R cuadrada de 0.993 y p-valores asociados a las variables muy significativos al ser

numéricamente cero. Por lo que podemos tener una buena aproximación a la distribución real con nuestro modelo, lo cual podemos confirmar confirmar al graficar la regresión lineal

```
(ggplot(regresionLinealSeveros) +
 geom_point(mapping=aes(x='perdidaFinal', y='y'), color='blue', size=0.5) +
 geom_abline(intercept=betas[0], slope=betas[1]) +
 theme(panel_background=element_rect(fill="#cfe8d6", color=None),
       panel_grid=element_line(color="black", size=0.2))
 )
```



De nuestro modelo de regresión lineal obtenemos los siguientes parámetros para la distribución Gumbel

$$\lambda = 54033185.48415679, \quad \mu = 20311902419.13288$$

Recordando que la función de distribución es la siguiente

$$F(t;\mu,\lambda) = exp\left(-exp\left(-\frac{t-\mu}{\lambda}\right)\right)$$

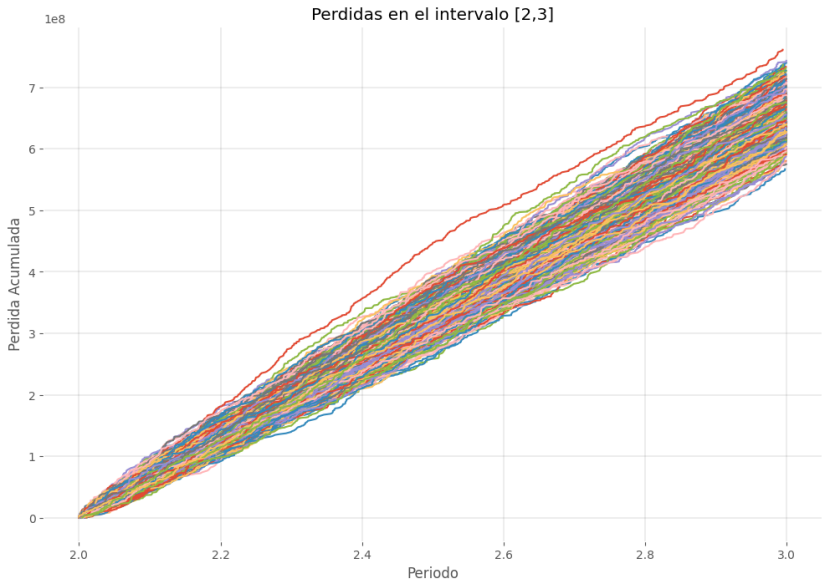
7.7 Simulación de pérdidas por mes

Ahora podemos analizar los valores extremos para las pérdidas en un mes, para este ejemplo tomaremos las pérdidas simuladas en el mes de marzo. El umbral que utilizaremos será de 725000000.

Primero tomaremos las trayectorias simuladas en el proceso poisson no Homogeneo y filtraremos las que se encuentran entre los valores de Trayectorias 2 y 3 que corresponden al mes de marzo, posteriormente simularemos las pérdidas con la distribución que previamente ya hemos simulado.

```
a = 2
b = 3

SimulacionesAcotadas = (SimulacionesPerdidas >>
                        # Quitamos las severidades acumuladas
                        select(-_.SeveridadAcumulada) >>
                        filter((_.Trayectorias > a) & (_.Trayectorias < b)))
                        # Tomamos solo los eventos que ocurrieron en el
                        periodo [a,b]

### Añadimos la nueva severidad acumulada ###

# Obtenemos el numero de trayectorias
NoTrayectorias = SimulacionesAcotadas['Numero de trayectoria'].max() + 1

# Creamos un dataframe vacio para almacenar las trayectorias
SimulacionesPerdidasAcotadas = pd.DataFrame()

# A cada trayectoria le agregaremos una columna con la severidad por
evento y la severidad acumulada
for i in range(NoTrayectorias+1): # Sumamos 1 por el slicing
  # Filtramos por el numero de trayectoria
  Perdida = (SimulacionesAcotadas >>
            filter(_['Numero de trayectoria'] == i))
  # Agregamos una severidad a cada evento y posteriormente hacemos
  la suma acumulada de severidades
  Perdida = (Perdida >> mutate(Severidad =
                simulacionDistEmpirica(len(Perdida))) >>
                mutate(SeveridadAcumulada = _.Severidad.cumsum()))
  SimulacionesPerdidasAcotadas = pd.concat([SimulacionesPerdidasAcotadas,
                                    Perdida])

# Volvemos a graficar
```

```
plt.style.use('ggplot')
plt.figure(figsize=(12, 8))

# Agrupar por 'Numero de trayectoria' y hacer un gráfico de líneas
para cada trayectoria
for trayectoria, datos in SimulacionesPerdidasAcotadas.
groupby('Numero de trayectoria'):
    plt.plot(datos['Trayectorias'], datos['SeveridadAcumulada'],
    label=f'Trayectoria {trayectoria}')

# Añadir etiquetas y título
plt.gca().set_facecolor("white") # Fondo
plt.grid(color='black', linewidth=0.1) #Cuadricula
plt.xlabel('Periodo')
plt.ylabel('Perdida Acumulada')
plt.title(f'Perdidas en el intervalo [{a},{b}]')
plt.show()
```




```
# Construimos la funcion de distribucion acumulada
n=perdidaFinal.shape[0]
regresionLineal = (perdidaFinalAcotada >>
                   arrange(-_.perdidaFinal) >>
```

```
                    mutate(i=range(1, n+1),
                           St=_.i/(n+0.0000000000001),
                           Ft = 1-_.St,
                           y = np.log(-np.log(_.Ft)),
                           const = 1,)
                    )

regresionLinealSeveros = (regresionLineal >>
                          filter(_.perdidaFinal >= 725000000))
regresionLinealSeveros.tail()

predictors = ["const", "perdidaFinal"]
target = "y"
fit = sm.OLS(regresionLinealSeveros[target],
        regresionLinealSeveros[predictors]).fit()
display(fit.summary())
```

OLS Regression Results

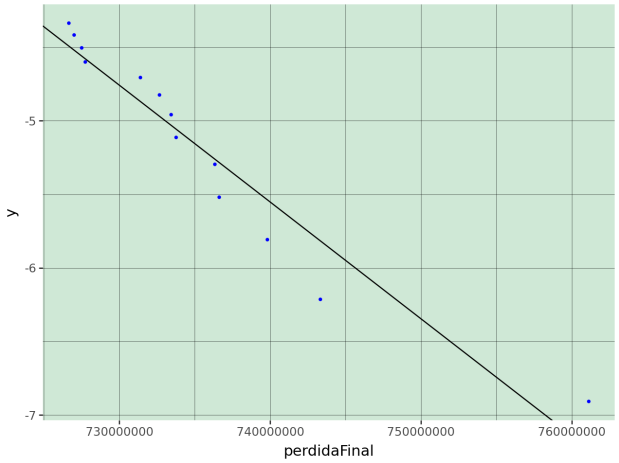
Dep. Variable:	y	**R-squared:**	0.931
Model:	OLS	**Adj. R-squared:**	0.925
Method:	Least Squares	**F-statistic:**	148.9
Date:	Sun, 01 Dec 2024	**Prob (F-statistic):**	9.78e-08
Time:	04:55:44	**Log-Likelihood:**	2.9171
No. Observations:	13	**AIC:**	-1.834
Df Residuals:	11	**BIC:**	-0.7043
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	53.1977	4.783	11.122	0.000	42.670	63.725
perdidaFinal	-7.939e-08	6.51e-09	-12.204	0.000	-9.37e-08	-6.51e-08

Omnibus:	1.053	**Durbin-Watson:**	1.322
Prob(Omnibus):	0.591	**Jarque-Bera (JB):**	0.729
Skew:	-0.531	**Prob(JB):**	0.694
Kurtosis:	2.533	**Cond. No.**	6.03e+10

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.03e+10. This might indicate that there are strong multicollinearity or other numerical problems.

Al hacer la regresión lineal obtenemos muy buenos resultados de ajuste, una R cuadrada de 0.931 y p-valores asociados a las variables muy significativos al ser numéricamente cero. Por lo que podemos tener una buena aproximación a la distribución real con nuestro modelo, aunque gráficamnte podemos ver que el ajuste no es tan bueno como el anterior.



De nuestro modelo de regresión lineal obtenemos los siguientes parámetros para la distribución Gumbel

$$\lambda = 12595799.735945726, \quad \mu = 670067257.6197054$$

8 Cálculo de Medidas

Una vez obtenidas las distribuciones asociadas a nuestro modelo, podemos obtener medidas de riesgo que nos permitan obtener información relevante sobre el comportamiento del negocio con el fin de poder tomar decisiones basadas en datos.

Recordemos que

1. $\mathbf{E}(X_t) = \Lambda(t)\mathbf{E}(Y); \quad \Lambda(t) = \int_0^t \lambda(x)dx$
2. $Var(X_t) = \Lambda(t)\mathbf{E}(Y^2)$

Notemos que $\Lambda(12) = \int_0^{12} \lambda(x)dx = 16981.5$

Usaremos como estimadores

$$\widehat{\mathbf{E}(Y)} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{Var(Y)} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Por lo que

$$\widehat{\mathbf{E}(X_t)} = \Lambda(t) * 1081694.923262261$$

$$\widehat{\mathbf{E}(X_{12})} = 16981.5 * 1081694.923262261 = 18368802339.3780$$

Y además

$$\widehat{Var(X_t)} = 233962639949.6137 * \Lambda(t) + 1081694.9232^2 * \Lambda(t)$$

$$\widehat{Var(X_{12})} = 233962639949.6137 * 16981.5 + 1081694.9232^2 * 16981.5 = 1.986945860571556e{+}16$$

De donde

$$\widehat{Std(X_t)} = 140959067.1284$$

Ahora, para hacer el cálculo del $VaR_p(X)$ y del $TVaR_p(X)$ tomaremos $p = 0.95$ y serán calculados de la siguiente manera

```
# VaR
var = np.quantile(perdidaFinal['perdidaFinal'], 0.95)
Out : 20450878141.960884

# TVaR
np.mean((perdidaFinal >> filter(_.perdidaFinal > 20450878141.960884))
        ['perdidaFinal'])
Out: 20516671910.37621
```

Por lo que para los eventos ocurridos en el año tenemos los siguientes resultados.

1. La pérdida esperada es 18368802339.3780
2. La volatilidad es 40959067.1284
3. El $VaR_{0.95}(X)$ es 20450878141.960884
4. El $TVaR_{0.95}(X)$ es 20516671910.37621

Si tomamos en cuenta que el umbral es 20500000000, podemos calcular el TVaR considerendo los valores mayores a dicho umbral. Dado que la cola derecha se distribuye como una gumbel de parámetros

$$\lambda = 54033185.48415679 \text{ y } \mu = 20311902419.13288$$

entonces el $TVaR_p(X) = \mathbf{E}(X|X > x_p) = \mu + \lambda\gamma$

```
# Calculo del TVaR
lambGumbel = 54033185.48415679
muGumbel = 20311902419.13288
gamma = 0.57721566

muGumbel + lambGumbel*gamma
Out: 20343091219.95402
```

Por lo que en este caso el TVaR es 20343091219.95402

9 Dependencia entre riesgos

Toda la teoría descrita en capítulos anteriores se enfocó en el desarrollo de modelos univariados, sin embargo en las instituciones existen varios tipos de riesgos que pueden estar relacionados entre sí, es decir, un evento puede vivir en más de un riesgo ó desencadenar otro evento que viva en otro riesgo.

Hay varias formas de describir la dependencia entre variables aleatorias, por ejemplo el coeficiente de correlación lineal ρ

$$\rho = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Para dos variables aleatorias X y Y, el coeficiente de correlación lineal es cercano a 1 ó -1 sí existe una gran relación lineal. Es importante resaltar mencionar que si $rho \approx 0$ esto no implica que no extista algún tipo de relación entre ambas variables, sólo que no hay relación **lineal**. Sólo se puede asegurar la inexistencia de relación cuando ambas variables aleatorias tienen una distribución normal

En el desarrollo de modelos para riesgo operacional estamos interesados principalmente en describir el comportamiento en la cola derecha de las distribuciones. Nos gustaría responder preguntas como "Si un riesgo tiene un evento de pérdida grande, ¿Es más probable que otro riesgo tenga un evento de pérdida similar?

9.1 Cópulas

A partir de la función de distribución conjunta $F_{(X_1,...,X_d)}$ podemos encontrar las marginales $F_{X_1}, ..., F_{X_d}$ integrando sobre las demás variables aleatorias, es decir,

$$F_{X_i} = \int_{\mathbf{R}^{d-1}} F_{(X_1,...,X_d)} dx_1, ...dx_{i-1}, dx_{i+1}, ...dx_d$$

Para el caso de variables aleatorias independientes, a partir de las distribuciones marginales F_{X_i} podemos encontrar la función de distribución conjunta $F_{(X_1,...,X_d)}$ multiplicando las marginales. Sin embargo esto no es válido cuando las marginales tienen algún tipo de dependencia entre ellas, como es el caso de la ocurrencia de eventos en diferentes tipos de riesgo. Para estos casos utilizaremos

funciones de distribución conjunta llamadas cópulas.

Definimos $U_i := F_{X_i}(x_i)$

$$\begin{aligned}\mathbf{P}[u_i \leq t] &= \mathbf{P}[F_{X_i}(x_i) \leq t] \\ &= \mathbf{P}[x_i \leq F_{X_i}^{-1}(t)] \\ &= F_{X_i}(F_{X_i}^{-1}(t)) \\ &= t, \quad t \in [0,1]\end{aligned}$$

$(U_1, ..., U_n),\ U_i \sim U(0,1)$ no necesariamente independientes.

Definimos una cópula C como la función de distribución conjunta de d variables aleatorias uniformes en el intervalo $(0,1)$. Esto es

$$\begin{aligned}C : [0,1]^d &\longrightarrow [0,1] \\ C(U_1, ..., U_d) &:= \mathbf{P}[U_1 \leq u_1, ..., U_n \leq u_n]\end{aligned}$$

Podemos observar que

1. C es una función de distribución conjunto en el cubo unitario
2. F_{X_i} tienen la información de cómo se comportan las X_i marginalmente
3. C tiene la información de la dependencia de las X_i sin tener ninguna información de cómo se comportan marginalmente

Teorema de Sklar
Si F es una función de distribución conjunta en $\mathbf{R}^d$ con distribuciones marginales F_i, existe una función cópula $C : [0,1]^d \longrightarrow [0,1]$ que satisface las propiedades de distribución conjunta tal que

$$F(x_1, ...x_d) = C(F_{X_1}(x_1), ..., F_{X_d}(x_d))$$

9.2 Medidas de dependencia

La correlación lineal mencionada anteriormente es una función de las distribuciones marginales. Si cambiamos la forma de las marginales, esto afectará el valor de la correlación lineal. Cuando describimos la dependencia utilizando cópulas,

esta no depende de la forma de las marginales, ya que la dependencia está capturada exclusivamente en la cópula.

Existen otras medidas de dependencia, las dos más populares son la rho de Spearman y la Tau de Kendall, desarrolladas en el campo de la estadística no paramétrica.

9.2.1 Rho de Spearman

Considere una función continua bivariada (X_1, X_2) con distribuciones marginales F_{X_1} y F_{X_2}. La medida de asociación **rho de spearman** $\rho_s(X_1, X_2)$ está dada por

$$\rho_s(X_1, X_2) = \rho(F_1(X_1), F_2(X_2))$$

donde ρ dentora correlación lineal.

Por lo que la rho de Spearman representa la relación lineal entre las variables F_{X_1} y F_{X_2}. Como ambas son variables aleatorias uniformes$(0, 1)$ con media $\frac{1}{2}$ y varianza $\frac{1}{12}$, podemos reescribir la rho de Spearman como

$$\begin{aligned}\rho_s(X_1, X_2) &= \frac{\mathbf{E}[F_{X_1}(x_1)F_{X_2}(x_2)] - \mathbf{E}[F_{X_1}(x_1)]\mathbf{E}[F_{X_2}(x_2)]}{\sqrt{Var(F_{X_1}(x_1))Var(F_{X_2}(x_2))}} \\ &= 12\mathbf{E}[F_{X_1}(x_1)F_{X_2}(x_2)] - 3\end{aligned}$$

En términos de cópulas, la rho se Spearman es

$$\begin{aligned}\rho_s(X_1, X_2) &= 12\mathbf{E}[F_{X_1}(x_1)F_{X_2}(x_2)] - 3 \\ &= 12\mathbf{E}[UV] - 3 \\ &= 12\int_0^1\int_0^1 uv\ dC(u, v) - 3 \\ &= 12\int_0^1\int_0^1 C(u, v)\ du\ dv - 3\end{aligned}$$

De este modo, la rho de Spearman es el coeficiente de correlación lineal entre las funciones de distribución acumulada de las variables aleatorias.

9.2.2 Tau de Kendall

Considere dos variables aleatorias bivariadas independientes e idénticamente distribuidas (X_1, X_2) y (X_1^*, X_2^*) con distribución marginal $F_{X_1}(x_1)$ para X_1 y X_1^* y distribución marginal $F_{X_2}(x_2)$ para X_2 y X_2^*. La medida de asociación **Tau de Kendall** $\tau_k(X_1, X_2)$ está dada por

$$\tau_k(X_1, X_2) = \mathbf{P}[(X_1 - X_1^*)(X_2 - X_2^*) > 0] - \mathbf{P}[(X_1 - X_1^*)(X_2 - X_2^*) < 0]$$

El primer término mide la concordancia, en el sentido de que para cada una de las dos dimensiones, las diferencias entre las variables aleatorias tienen el mismo signo. El segundo término mide la discordancia. De la definición podemos ver que la tau de Kendall puede reescribirse como

$$\tau_k(X_1, X_2) = \mathbf{E}[sign(X_1 - X_1^*)(X_2 - X_2^*)]$$

Ahora obtendremos una expresión para la tau de Kendall en términos de la cópula

$$\begin{aligned}
\tau_k(X_1, X_2) &= \mathbf{P}[(X_1 - X_1^*)(X_2 - X_2^*) > 0] - \mathbf{P}[(X_1 - X_1^*)(X_2 - X_2^*) < 0] \\
&= \mathbf{P}[(X_1 - X_1^*)(X_2 - X_2^*) > 0] - \{1 - \mathbf{P}[(X_1 - X_1^*)(X_2 - X_2^*) > 0]\} \\
&= 2\mathbf{P}[(X_1 - X_1^*)(X_2 - X_2^*) > 0] - 1 \\
&= 4\mathbf{P}[X_1 < X_1^*, X_2 < X_2^*] - 1 \\
&= 4\mathbf{E}\{\mathbf{P}[X_1 < X_1^*, X_2 < X_2^* | X_1^*, X_2^*)]\} - 1 \\
&= 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{P}[X_1 < x_1, X_2 < x_2] dF(x_1, x_2) - 1 \\
&= 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x_1, x_2) dF(x_1, x_2) - 1 \\
&= 4\int_0^1\int_0^1 C(F_{X_1}(x_1), F_{X_2}(x_2))\ dC(F_{X_1}(x_1), F_{X_2}(x_2)) - 1 \\
&= 4\int_0^1\int_0^1 C(u, v)\ dC(u, v) - 1 \\
&= 4\mathbf{E}[C(F_{X_1}(x_1), F_{X_2}(x_2))] - 1
\end{aligned}$$

Si la cópula es absolutamente continua puede reescribirse como

$$\tau_k(X_1, X_2) = 4\int_0^1\int_0^1 C(u, v)\frac{\partial^2 C(u, v)}{\partial u \partial v}\ du\ dv - 1$$

9.2.3 Dependencia de colas

El estudio de las variables aleatorias no sólo debe ocuparse de analizar el comportamiento general, si no también en responder preguntas acerca de la distribución de las colas. Cuando existe dependencia entre variables aleatorias también es necesario entender el comportamiento de manera conjunta cuando existen valores atípicos, específicamente aquellos donde se tienen eventos de baja frecuencia con alto impacto.

Se ha observado que si un evento atípico ocurre en un riesgo, esto aumenta la probabilidad de que exista otro evento extremo en otro riesgo. Las medidas de dependencia en colas han sido desarrolladas para analizar qué tan fuerte es la correlación en la cola derecha de las distribuciones.

Considere dos variables aleatorias X y Y con distribuciones marginales $F(x)$ y $G(x)$. El índice de alta dependencia de cola λ_U se define como

$$\lambda_U = \lim_{u \to 1} \mathbf{P}[X > F^{-1}(u) | Y > G^{-1}(u)]$$

De manera general, el índice de alta dependencia de cola mide la probabilidad de que X tome valores grandes dado que Y tomó valores grandes, donde los "valores grandes" son medidos en términos de cuantiles. Este índice puede ser reescrito como sigue:

$$\begin{aligned} \lambda_U &= \lim_{u \to 1} \mathbf{P}[F(X) > u | G(Y) > u] \\ &= \lim_{u \to 1} \mathbf{P}[U > u | V > v] \end{aligned}$$

Donde $U, V \sim U(0,1)$

Esto puede reescribirse como

$$\begin{aligned} \lambda_U &= \lim_{u \to 1} \frac{1 - \mathbf{P}[U \leq u] - \mathbf{P}[V \leq u] + \mathbf{P}[U \leq u, V \leq u]}{1 - \mathbf{P}[V \leq u]} \\ &= \lim_{u \to 1} \frac{1 - 2u + C(u,u)}{1 - u} \end{aligned}$$

Esto muestra que la dependencia de las colas de X y Y como se definió anteriormente puede ser medida utilizando la cópula en vez de la distribución original.

9.3 Tipos de cópulas

9.3.1 Cópula arquimediana

Las cópulas arquimedianas son aquellas funciones de distribución acumulada de dimensión d que tienen la forma

$$C(u_1, ..., u_d) = \phi^{-1}(\phi(u_1) + ... + \phi(u_d))$$

Donde $\phi(u)$ es una función estríctamente decreciente, convexa y continua llamada **generador** tal que

$$\phi : [0, 1] \longrightarrow [0, \infty]$$

9.3.2 Cópula independiente

Para n variables aleatorias independientes con función de distribución acumulada $F(x_j), j = 1, ...d$, la función de distribución conjunta está dada por $\prod_{j=1}^{d} F(x_j)$. La cópula correspondiente es llamada cópula independiente y está dada por

$$C(u_1, ..., u_d) = \prod_{j=1}^{d} u_j$$

Esta es una cópula Arquimediana con generador $\phi(u) = -lnu$

9.3.3 Cópula de Cook-Johnson

La cópula de Cook-Johnson tiene generador $\phi(u) = u^{-\theta} - 1, \theta > 0$, por lo que tiene la forma

$$C(u_1, ...u_d) = (u_1^{-\theta} + u_d^{-\theta} - d + 1)^{1/\theta}$$

Tiene un parámetro θ que puede ser estimado por máxima verosimilitud. En el contexto bivariado es conocida como cópula de Clayton.

9.3.4 Cópula de Gumbel-Hougaard

La cópula de Gumbel-Houggard tiene generador $\phi(u) = (-lnu)^{\theta},\ \theta \geq 1$, por lo que tiene la forma

$$C(u_1, ..., u_d) = exp\{-[(-lnu_1)^{\theta} + ... + (-lnu_d)^{\theta}]\}^{1/\theta}$$

En el caso bivariado se conoce como cópula de Gumbel

9.3.5 Cópula de Frank

La cópoula de Frank tiene generador

$$\phi(u) = -ln\frac{e^{-\theta u-1}}{e^{-\theta}-1}, \quad -\infty < \theta < \infty,\ \theta \neq 0$$

Por lo que la cópula de Frank tiene la forma

$$C(u_1, ...u_d) = -ln\left\{1 + \frac{(e^{-\theta u_1}-1)...(e^{-\theta u_d}-1)}{e^{-\theta}-1}\right\}^{1/\theta}$$

9.3.6 Cópula de Ali-Mikhail-Haq

La cópula de Ali-Mikhail-Haq tiene generador

$$\phi(u) = ln\frac{1-\theta(1-u)}{u}, \quad -1 \leq \theta < 1$$

Por lo que tiene la forma

$$C(u_1, ..., u_d) = \frac{\prod_{j=1}^{d} u_j}{1-\theta\prod_{j=1}^{d}(1-u_j)}$$

9.3.7 Cópula de Joe

La cópula de Joe tiene generador

$$\phi(u) = -ln(1-(1-u)^{\theta}), \quad \theta \geq 1$$

Por lo que tiene la forma

$$C(u_1, ..., u_d) = 1 - \left[\sum_{j=1}^{d}(1-u_j)^{\theta} - \prod_{j=1}^{d}(1-u_j)^{\theta}\right]^{1/\theta}$$

9.3.8 Cópulas elípticas

Las cópulas elípticas son aquellas que son asociadas con distribuciones elípticas. Los dos modelos principales son la cópula Gaussiana, asociada con la distribución normal multivariada y la cópula t (de Student), asociada con la distribución multivariada t de student

9.3.9 Cópula Gaussiana

La cópula Gaussiana está dada por

$$C(u_1, ...u_d) = \Phi_P(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_d))$$

Donde $\Phi(x)$ es la función de distribución acumulada de la distribución normal univariada y $\Phi_P(x_1, ..., x_d)$ es la función de distribución acumulada normal multivariada (con media 0 y variazna 1 para cada componente) y matriz de correlación P. Como la matriz de correlación tiene $\frac{d(d-1)}{2}$ pares de coeficientes de correlación, este es el número de parámetros de la cópula. En el caso bivariado, la cópula Gaussiana puede escribirse como

$$C(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left\{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right\} dydx$$

donde ρ es la correlación entre las variables aleatorias

9.3.10 Cópula t

La cópula t está dada por

$$C(u_1, ..., u_d) = t_{\nu,P}^{-1}(t_\nu^{-1}(u_1), ..., t_\nu^{-1}(u_d))$$

Donde $t_\nu(x)$ es la función de distribución conjunta de la distribución t univariada con ν grados de libertad y $t_{\nu,P}(x_1, ..., x_d)$ es la función de distribución multivariada t con ν grados de libertad para cada componente y donde P es la matriz de correlación. Para el caso bivariado, la cópula t puede ser escrita como

$$C(u_1, u_2) = \int_{-\infty}^{t_\nu^{-1}(u_1)} \int_{-\infty}^{t_\nu^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \left[1 + \frac{x^2 - 2\rho xy + y^2}{\nu(1-\rho^2)}\right]^{-1-\frac{\nu}{2}} dydx$$

9.3.11 Cópulas de valores extremos

Otra clase importante de cópulas es la de valores extremos. Esta clase de cópulas es definida en terminos de la propiedad de escalabilidad de las distribuciones de valores extremos. Una cópula es de valores extremos (EV) si satisface que

$$C(u_1^n, ..., u_d^n) = C^n(u_1, ..., u_d)$$

9.4 Simulación

Para poder simular la dependencia entre dos riesgos, se han simulado otro proceso de eventos con diferentes parámetros de intensidad por mes en la distribución poisson y diferente densidad de severidad. Debido a que ambos riesgos simulados tienen una cantidad de eventos muy grande, se tomará una muestra de 100 elementos para simular cópulas.

```
riesgo1 = perdidaFinal2['perdidaFinal'].tolist()
riesgo3 = perdidaFinal['perdidaFinal'].sample(n=100).tolist()

eventos = pd.DataFrame({'Riesgo1':riesgo1, 'Riesgo3':riesgo3})
eventos
```

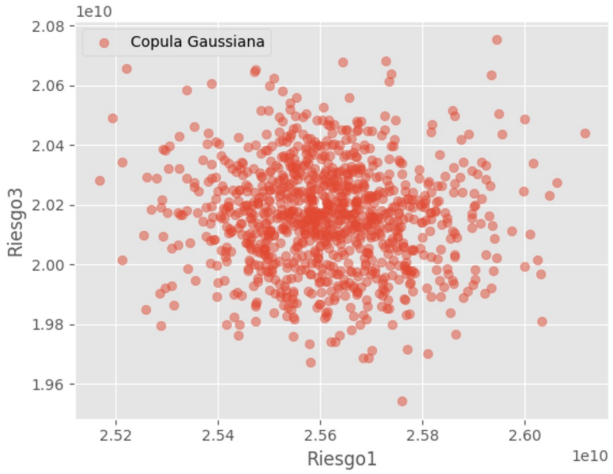
	Riesgo1	Riesgo3
0	2.551122e+10	2.028611e+10
1	2.550625e+10	2.027440e+10
2	2.548260e+10	2.035134e+10
3	2.592228e+10	2.028732e+10
4	2.563061e+10	2.028170e+10
...	...	...
95	2.563335e+10	1.995748e+10
96	2.574654e+10	2.007013e+10
97	2.546119e+10	1.997042e+10
98	2.555244e+10	2.053164e+10
99	2.577414e+10	2.013950e+10

100 rows × 2 columns

Como las pérdidas finales de las simulaciones de ambos riesgos tienen una distribución normal, utilizaremos la cópula Gaussiana para simular la dependencia. Utilizaremos el siguiente código para realizar la simulación

```
from copulas.multivariate import GaussianMultivariate
from copulas.univariate import GaussianKDE
from copulas.bivariate import Clayton, Gumbel, Frank
from sklearn.preprocessing import QuantileTransformer

# Ajustamos la copula gaussiana
```

```
copula_gauss = GaussianMultivariate()
copula_gauss.fit(eventos)

# Obtenemos muestras de la copula gaussiana
muestrasConjunta = copula_gauss.sample(1000)

# Visualización de las muestras generadas
plt.scatter(muestrasConjunta['Riesgo1'], muestrasConjunta['Riesgo3'],
alpha=0.5, label='Copula Gaussiana')
plt.legend()
plt.xlabel('Riesgo1')
plt.ylabel('Riesgo3')
plt.show()
```



Como medida de dependencia utilizaremos la tau de Kendal, para la distribución normal se calcula de la siguiente forma

$$\tau = \frac{2}{\pi} arcsen(\rho)$$

donde ρ es el coeficiente de correlación de Spearman

```
# Obtener el coefifience de correlación de Spearman
correlacion = muestrasConjunta['Riesgo1'].corr(muestrasConjunta['Riesgo3'])
tauGauss = (2/np.pi)*np.arcsin(correlacion)
tauGauss
```

```
Out: -0.034810178767449115
```

Podemos observar que la tau de Kendal es un valor cercano a cero, por lo que la dependencia entre los eventos de ambos riesgos es casi nula.

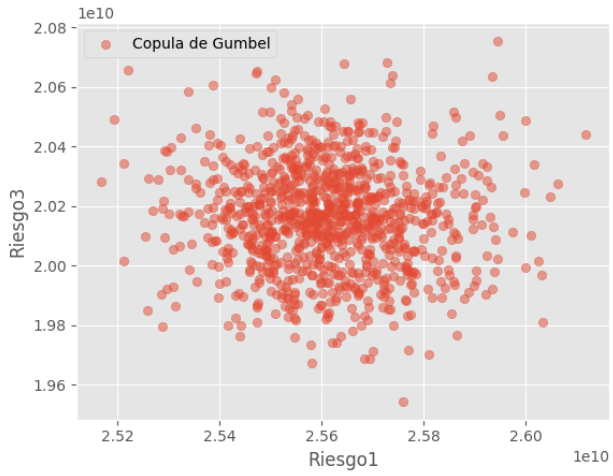
Realizaremos un estudio adicional del comportamiento de los valores extremos. Como vimos, las colas tienen un buen ajuste a una distribución Gumbel, por lo que podemos utilizar la cópula de Gumbel para simular la dependencia entre estos eventos. Utilizaremos el siguiente código para realizar la simulación.

```
# Dataframe de las colas
riesgo1 = regresionLinealSeveros2['perdidaFinal'].sample(n=
            len(regresionLinealSeveros)).tolist()
riesgo3 = regresionLinealSeveros['perdidaFinal'].tolist()

colas = pd.DataFrame({'Riesgo1':riesgo1, 'Riesgo3':riesgo3})
colas
```

	Riesgo1	Riesgo3
0	2.548260e+10	7.398551e+08
1	2.564338e+10	7.382467e+08
2	2.562725e+10	7.366971e+08
3	2.550016e+10	7.320921e+08
4	2.568851e+10	7.314424e+08
5	2.562074e+10	7.295398e+08
6	2.536371e+10	7.293771e+08
7	2.555333e+10	7.291668e+08
8	2.555244e+10	7.291646e+08
9	2.554063e+10	7.287386e+08
10	2.549851e+10	7.277952e+08
11	2.546752e+10	7.276607e+08
12	2.554898e+10	7.273050e+08
13	2.535872e+10	7.256028e+08

```
# Ajustamos la copula de Gumbel
copula_Gumbel = Gumbel()
```

```
# Transformamos los datos a una escala uniforme
scaler = QuantileTransformer(output_distribution='uniform')
colasReescaladas = scaler.fit_transform(colas)

copula_Gumbel.fit(colasReescaladas)

# Obtenemos muestras de la copula gaussiana
muestrasConjuntaGumbel = copula_Gumbel.sample(1000)

# Convertimos el resultado en un DataFrame
muestrasConjuntaGumbel = pd.DataFrame(muestrasConjuntaGumbel,
columns=['Riesgo1', 'Riesgo1'])
muestrasConjuntaGumbel

# Visualización de las muestras generadas
plt.scatter(muestrasConjunta['Riesgo1'], muestrasConjunta['Riesgo3'],
alpha=0.5, label='Copula de Gumbel')
plt.legend()
plt.xlabel('Riesgo1')
plt.ylabel('Riesgo3')
plt.show()
```




```
tauGumbel = 1 - (1/copula_Gumbel.theta)
tauGumbel
```

```
Out: 0.4065934065934066
```

La tau de Kendal nos da una dependencia positiva moderada entre ambos riesgos

10 Conclusiones

Las simulaciones realizadas utilizando procesos de Poisson compuestos no homogéneos lograron modelar tanto la severidad como la frecuencia de eventos de riesgo. Esto permitió estimar métricas de riesgo como la pérdida esperada, la volatilidad el VaR, el TVaR y la tau de Kendal.

Las pruebas de normalidad y el análisis de valores extremos confirmaron que las distribuciones utilizadas son adecuadas para modelar las pérdidas. En particular, las distribuciones ajustadas para eventos extremos son esenciales para planificar escenarios de alto impacto financiero.

El uso de cópulas permitió estudiar la dependencia entre diferentes tipos de riesgos. Aunque algunas dependencias fueron bajas, los modelos lograron capturar las interacciones relevantes, ofreciendo un enfoque más completo para la gestión de riesgos.

11 Common Shock Poisson Models

2. El modelo

2.1 Frecuencia de pérdidas

Supongamos que hay m diferente tipos de eventos y, para $e = 1, ...m$ sea

$$\{N^{(e)}(t)\}_{t\geq 0}$$

un proceso Poisson con intensidad $\lambda^{(e)}$ el cual cuenta el número de eventos que ocurren en $(0, t]$.

Supongamos además que esos procesos de conteo son independientes. Considera pérdidas de n diferentes tipos, para $j = 1, ...n$, sea

$$\{N_j(t)\}_{t\geq 0}$$

un proceso de conteo de la frecuencia de pérdidas del tipo "j" que han ocurrido en $(0, t]$.

Para la r-ésima ocurrencia de un evento de tipo "e", la variable Bernoulli $I^{(e)}_{j,r}$ indica cuándo ocurrió una pérdida del tipo "j".

Los vectores

$$\mathbf{I}^{(e)}_r = (I^{(e)}_{1,r}, ..., I^{(e)}_{n,r})$$

para $r = 1, ...N^{(e)}(t)$ son cosiderados independientes e idénticamente distribuidos con distribución Multinoulli (Bernoulli Multivarida). En otras palabras, cada nuevo evento representa una nueva oportunidad independiente de tener una pérdida, pero, para un evento fijo, la **variable de activación*** podría ser dependiente. El tipo de dependencia depende de la **especificación*** de la distribución Multinoulli y su independencia es un caso especial. Usaremos la siguiente notación para p **densidades¿?** marginales de esta distribución (donde se omite el subíndice r por simplicidad)

$$\mathbf{P}(I^{(e)}_{j_1} = i_{j_1}, ..., \mathbf{I}_{j_p} = i_{j_p}) = p_{j_1,...j_p}(i_{j_1}, ..., i_{j_p}), i_{j_1}, ..., i_{j_p} \in \{0, 1\}$$

También escribimos $p^{(e)}_j(1) = p^{(e)}_j$ para densidades marginales unidimiensionales, entonces en este caso especial de independencia condicional tenemos

$$p^{(e)}_{j_1,...,j_p}(1, , , 1) = \prod_{k=1}^{p} p^{(e)}_{jk}$$

Los proceso de conteo para eventos y pérdidas son vinculados por

$$N_j(t) = \sum_{e=1}^{m} \sum_{r=1}^{N^{(e)}(t)} \mathbf{I}_{j,r}^{(e)}$$

Bajo el supuesto Poisson para los procesos de eventos y el supuesto Bernoulli para la indicadora de pérdidas, los procesos $\{N_j(t)\}_{t\geq 0}$ son Poisson, como se obtuvieron superponiendo m procesos Poisson (posiblemente **diluidos¿?**) generados por los m procesos de eventos subyacentes. $(N_1(t), ..., N_n(t))'$ puede considerarse como si tuviera una distribución Poisson multivariada **subyacente¿?**

Sin embargo, el número total de pérdidas $N(t) = \sum_{j=1}^{n} N_j(t)$ en general no es Poisson si no Poisson Compuesto. Es la suma de m variables aleatorias Poisson compuestas independientes como se escribirá a continuación

$$N(t) = \sum_{e=1}^{m} \sum_{r=1}^{N^{(e)}} \sum_{j=1}^{n} \mathbf{I}_{j,r}^{(e)}$$

La distribución conjunta del e-ésimo proceso Poisson compuesto es la distribución de $\sum_{j=1}^{n} \mathbf{I}_j^{(e)}$ que en general es una suma de variables Bernoulli independientes

12 Bibliografía

References

[1] Panjer, H. H. (2006). *Operational Risk: Modeling Analytics*. John Wiley & Sons, Inc.

[2] Rodríguez Bermúdez, M. A. (2016). *Modelación Cuantitativa de Riesgo Operativo* (Tesis de maestría, Universidad Nacional de Colombia, Facultad de Ingeniería, Departamento de Ingeniería Industrial). Universidad Nacional de Colombia.

[3] Baltazar-Larios, F., & López Ortega, S. I. (2024). *Simulación estocástica*. Facultad de Ciencias, Universidad Nacional Autónoma de México.

[4] Erdely, A. (2009). Cópulas y dependencia de variables aleatorias: Una introducción. *Miscelánea Matemática, 48*, 7-28.

[5] Lindskog, F., & McNeil, A. J. (2003). Common Poisson shock models: Applications to insurance and credit risk modelling. *ASTIN Bulletin, Vol. 33 No. 2, pp. 209-238.*

[6] Rincón, L. (2012). Introducción a los procesos estocásticos (1a ed.). UNAM Facultad de Ciencias, Las Prensas de Ciencias.