UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

FACULTAD DE ECONOMÍA

“MODELACIÓN DE PÉRDIDA AGREGADA APLICADA A SINIESTROS DE CÁNCER DE MAMA: CASO DE UNA EMPRESA ASEGURADORA”

TESINA

QUE PARA OBTENER EL TÍTULO DE:

LICENCIADA EN ACTUARÍA

PRESENTA:
MARÍA STRAULINO GARCÍA RENDÓN

ASESOR:
DR. EN C. MIGUEL ÁNGEL DÍAZ CARREÑO

REVISORES:
M. EN E. JUVENAL ROJAS MERCED
M. EN E. ELIAS EDUARDO GUTIÉRREZ ALVA

TOLUCA, ESTADO DE MÉXICO **FEBRERO 2014**

AGRADECIMIENTOS

Agradezco principalmente a Dios, por este ciclo que me permitió terminar, durante el cual estuvo presente todos los días.

A mi hermano, por apoyarme y acompañarme toda mi vida, especialmente durante la Universidad.

A Erick, por su tiempo, consejos, paciencia y ayuda durante la elaboración de este trabajo.

A mis abuelos, por ser un soporte en mi vida.

A mi papá, por creer en mí y darme ánimos.

A mis tíos, Fernando y Lucía, por su gran apoyo y sus enseñanzas.

A Margarita, por su cariño, tiempo y compañía.

A mis amigos, porque sin ustedes la Universidad no hubiera sido igual para mí.

A mis profesores, por compartir su tiempo, conocimientos y experiencia.

ÍNDICE

INTRODUCCIÓN

En la vida cotidiana, constantemente se deben tomar decisiones bajo condiciones de incertidumbre. La incertidumbre es el origen del riesgo, el cual siempre es asociado con una posible pérdida o ganancia, por tal motivo, medirlo es fundamental (Peng, 2010).

Las medidas de riesgo son representaciones que asocian algún evento de interés a números reales por medio de variables aleatorias; este es el principio de los modelos de pérdidas.

El modelado de la pérdida agregada es uno de los objetivos principales de la teoría y práctica actuarial, especialmente en el sector asegurador cuando se tienen que tomar decisiones importantes del negocio respecto a algunas características de los contratos de seguros (Mohamed, *et al.,* 2010).

Gran parte de la ciencia actuarial está enfocada en el desarrollo y análisis de modelos matemáticos que expliquen el proceso por el cual los fondos entran y salen del sistema asegurador, los modelos de pérdidas explican una de las principales fuentes de egresos, el pago de beneficios (Klugman, *et al.,* 2004).

Una de las variables más importantes para estimar la utilidad, establecer tarifas, proyectar la situación de la compañía en el futuro, crear estrategias de negocio, entre otros, es el monto que la institución tendrá que pagar derivado de las pólizas vendidas que resulten en reclamo, a este monto se le conoce como pérdida agregada (Papush, et al., 2001).

Tener un estimado de la pérdida agregada de una compañía de seguros es básico y necesario para que la empresa sea financieramente sólida y pueda seguir operando con un rumbo definido, por lo cual se han desarrollado varios

modelos para obtener un aproximado de dicha cantidad, estos modelos son la descripción del comportamiento de un conjunto de riesgos formado por un determinado portafolio de pólizas (Burnecki, *et al*., 2010).

Una opción útil para estimar la pérdida agregada es el modelo de riesgo individual, mediante el cual se pueden resolver algunas incógnitas que ayudan en la toma de decisiones y en la creación de estrategias para las compañías de seguros.

El objetivo principal de este trabajo es proponer un modelo de riesgo para estimar la distribución de la pérdida agregada por siniestros de cáncer de mama de una aseguradora, a partir de datos proporcionados por la misma aseguradora de la siniestralidad de gastos médicos en un periodo determinado con el fin de conocer información de interés como:

- El promedio del número de reclamos $E[N]$.
- El monto promedio de pago por póliza $E[X]$.
- El monto esperado de siniestralidad $E[S]$.
- La varianza del monto esperado de siniestralidad $Var(S)$.
- La distribución de la pérdida agregada $P(S \leq s)$.
- El valor en riesgo $VaR_{\alpha} = (F^{-1}(\alpha))$
- El valor esperado de la cola de la distribución $TVaR_{\alpha} = (E[S| S \geq VaR_{\alpha}])$

La razón por la que se pretende crear un modelo de riesgo para pronosticar las pérdidas de una aseguradora en lugar de utilizar solamente datos empíricos, es decir, no utilizar únicamente la experiencia histórica para obtener un promedio y estimar la pérdida esperada de la compañía, es que se estaría asumiendo que el futuro se comportará de la misma manera. Lo anterior implica que si se usan únicamente datos empíricos, se esperaría que el número o proporción de reclamos ocurridos sea muy similar a los años anteriores y, asimismo, su monto. Éste supuesto sería poco útil a la hora de hacer una estimación, dado que el

hecho de que un reclamo catastrófico no haya ocurrido aún, no significa que no pueda ocurrir en un futuro, o que la frecuencia no pueda aumentar o disminuir drásticamente de un año a otro.

En los ramos de daños y gastos médicos es poco probable que la siniestralidad de un año sea igual a la del siguiente. La ventaja de utilizar solamente datos empíricos de severidad y frecuencia es su disponibilidad, pero algunos detalles importantes son omitidos si sólo se analizan estos datos (Heckman y Meyers, 1983). Modelar las pérdidas con distribuciones probabilísticas permitirá obtener más información acerca del comportamiento de las mismas.

Estos modelos están elaborados en términos de distribuciones, que explican el comportamiento del monto y número de reclamos. El describir las pérdidas agregadas en términos de frecuencia y severidad por separado, permite hacer modificaciones sobre los contratos de seguro, como lo son el deducible, suma asegurada máxima y coaseguro.

Este trabajo constará de cuatro secciones. La primera expone los orígenes y el desarrollo de los seguros de gastos médicos, así como la evolución de las técnicas y metodologías para cuantificar el riesgo derivado de éstos hasta lo que hoy son los modelos de pérdidas agregadas. También se mencionan varias razones por las cuales es importante para un asegurador medir este riesgo y algunas aplicaciones de los modelos de pérdidas agregadas.

En la siguiente sección se encuentra el marco conceptual y teórico para el desarrollo de la modelación de la pérdida agregada, se definirán los modelos de riesgo, tanto individual como colectivo y, se mencionarán algunas características importantes de las distribuciones más utilizadas en la modelación de la pérdida agregada. Asimismo, se describirá los métodos *bootstrap* y estimación bayesiana, como herramientas que proveerán una medida de aproximación del error de estimación e intervalos de confianza o de probabilidad.

La tercera sección consiste en la aplicación de la modelación de la pérdida agregada al caso de una empresa seguradora. Se realizarán pruebas de bondad de ajuste, mediante el apoyo de programas estadísticos, para seleccionar una función de distribución que se ajuste a los datos de severidad que serán analizados en el presente trabajo.

De acuerdo a los posibles modelos resultantes, se elegirá el mejor método para aproximar la distribución de la pérdida agregada; una vez obtenida una aproximación de esta, se calcularán algunas variables de interés como el monto de reclamo promedio, el número de reclamos esperados, el monto promedio de la siniestralidad, el valor para el que un determinado porcentaje de los reclamos se encuentra por debajo de éste y la esperanza de la pérdida agregada dado que cierto valor determinado se ha superado.

Una vez calculados los datos de interés, se obtendrán la función de distribución empírica de la pérdida agregada así como las variables anteriormente mencionadas con el objeto de comparar el modelo empírico y el teórico. Finalmente, se comentarán los resultados obtenidos del modelo de pérdida agregada desarrollado en la sección anterior.

I. ORIGEN DEL SEGURO DE GASTOS MÉDICOS

La idea del seguro es parte de nuestra civilización y está basada en la confianza entre el asegurador y el asegurado. Esta confianza mutua debe de estar sustentada en la ciencia y no en creencias o en la especulación. Durante el siglo XX, se desarrollaron las herramientas necesarias para trabajar en algunas cuestiones de seguros, como la teoría de la probabilidad, estadística y procesos estocásticos (Mikosch, 2004).

La función principal de los seguros es distribuir las pérdidas derivadas de la ocurrencia de algún evento aleatorio entre los miembros de un grupo determinado y que ninguna persona en particular se vea fuertemente afectada al ocurrir el evento en cuestión (King, 1915).

Blumenfeld (1961) define al seguro como la distribución del riesgo sobre un grupo de personas, cada una expuesta a un evento en común. Así, mediante la ley de los grandes números, la posibilidad de una gran pérdida se sustituye por un pequeño cargo, cuyo monto es conocido de antemano y, una vez que se unen todas las aportaciones del grupo, este monto sirve para compensar a los miembros que incurrieron en el evento asegurado.

La primera forma del seguro de gastos médicos apareció alrededor de 1840, cuando a los pasajeros de trenes se les pagaba una suma de dinero, en caso de muerte o heridas graves por un accidente ocurrido mientras el pasajero viajaba.

Mientras, el seguro de reembolso de gastos por enfermedad apareció a fines de la década de 1840, pero fue hasta 1890 cuando los aseguradores de pérdidas por accidentes incurrieron en el campo de pérdidas por enfermedad y así, fue que surgió un sistema de seguros de salud. Gradualmente, los seguros de

accidentes y enfermedades se fueron uniendo en una única póliza (Blumenfeld, 1961) hasta lo que hoy conocemos como seguro de gastos médicos mayores.

Algunas aseguradoras han incursionado en la prevención de riesgos como un intento para reducir las pérdidas sufridas por el grupo asegurado, pero aun así, la función más importante del seguro radica en distribuir apropiadamente las pérdidas entre los miembros de un grupo (King, 1915). Es decir, si se establece un costo apropiado del seguro, entonces el riesgo estará distribuido de manera equitativa, ya que sin importar cuál de los individuos del grupo sufra la pérdida, ésta será subsanada por las aportaciones de todo el grupo, por lo tanto, es primordial encontrar una forma de determinar un costo apropiado del seguro.

La matemática actuarial empezó a finales del siglo XVII con la tabla de mortalidad de Edmund Halley, la cual permitió por primera vez un tratamiento matemático y cálculos de valores de anualidades. Actualmente, este modelo matemático no es sólo el más común, sino que es el único en el sector de seguros de vida y, es la aplicación clásica y característica de la ciencia actuarial.

Hoy en día, debido al desarrollo que han tenido los seguros de gastos médicos y daños, desde hace unas cuantas décadas han surgido nuevos desarrollos en el campo actuarial. Estos desarrollos han sido posibles debido al avance en la teoría de probabilidad y estadística matemática desde la década de los 30s y el énfasis que se le ha dado a los métodos matemáticos en la teoría económica. (Bühlmann,1970).

Los matemáticos suecos Filip Lundberg y Harald Cramér fueron pioneros en estas áreas, al percatarse de que la teoría de procesos estocásticos provee el marco más apropiado para modelar la ocurrencia de los montos de reclamo en el sector de seguros. El modelo de Cramér-Lundberg es uno de los fundamentos de las matemáticas de seguros de no vida, el cual ha sido modificado y extendido a varias direcciones y, además, ha motivado la investigación en otros

campos de la teoría de probabilidad aplicada como la teoría de colas, procesos de ramificación, la teoría de renovación, confiabilidad, valores extremos, entre otros.

En 1903, Lundberg estableció las bases de la teoría del riesgo moderna. Mikosch (2004) define la teoría del riesgo como un sinónimo de matemáticas de seguros de no vida, la cual trata sobre los modelos que existen para representar los reclamos que llegan a una aseguradora con el objeto de calcular la prima que debe ser cobrada y, así evitar la banca rota de la institución (teoría de la ruina).

Una de las tareas más comunes de la profesión actuarial en el sector asegurador, es analizar y construir modelos de pérdidas, también llamados modelos de riesgo actuariales, para predecir el monto de las obligaciones que tendrá un asegurador en el futuro.

Una compañía aseguradora se encarga de emitir contratos de seguro, por medio de los cuales se obliga, mediante una prima, a resarcir el daño o a pagar una suma de dinero al verificarse la eventualidad prevista en el contrato.

El objeto de principal interés desde el punto de vista de una aseguradora es el monto total de reclamos o monto de reclamos agregado. Para calcular este monto, se necesitan encontrar modelos probabilísticos suficientemente realistas, pero que sean a la vez simples para modelar el monto agregado de los reclamos y la ocurrencia de estos. La discrepancia entre modelos realistas y simples se relaciona con la pregunta de hasta qué punto un modelo matemático puede describir el complejo dinamismo de un portafolio de seguros, sin ser matemáticamente intratable.

Para una institución de seguros, es primordial determinar una prima que sea suficiente para hacer frente a sus obligaciones y, asimismo, que sea competitiva

dentro del mercado. Hoy en día, con el nuevo marco regulatorio de Solvencia II[1], es muy importante que el sector asegurador implemente los métodos y técnicas, necesarios para obtener los mejores estimados de sus activos y pasivos que le permitan preservar su solvencia y estabilidad cumpliendo con los requerimientos de capital establecidos por la Comisión Nacional de Seguros y Fianzas (CNSF).

Generalmente, es necesario estimar distribuciones de probabilidad para describir el proceso de las pérdidas cubiertas por los contratos de seguros, por ejemplo, para que el costo de un contrato de seguro sea correcto de acuerdo a principios de cálculo razonables, éste debe estar basado en el proceso de pérdidas del contrato. En la práctica es imposible conocer el verdadero proceso de pérdida, pero un estimado preciso de este proceso, dentro de lo razonable, permite tener una base para calcular el costo del contrato de la manera más exacta posible (Patrik, 1980).

Para determinar el monto total de los reclamos, es necesario determinar las propiedades teóricas de los procesos estocásticos de la suma total de los reclamos y, el número total de reclamos ocurridos, entre otras cosas, es de interés conocer sus distribuciones junto con algunas características como sus momentos, varianza y estructura de dependencia (Mikosch, 2004).

Una herramienta útil para calcular el monto de reclamos agregado son los métodos de simulación, los cuales se han hecho cada vez más populares, y muchas veces han remplazado a los métodos probabilísticos y/o estadísticos rigurosos. Esto no significa que la teoría no sea necesaria, más bien quiere decir que la simulación debe estar basada en modelos probabilísticos, el procedimiento de simulación en sí mismo, debe explotar las propiedades teóricas del modelo a simular.

[1] Solvencia II es un proyecto de revisión del régimen de suficiencia de capital para el sector asegurador y tiene como propósito establecer un conjunto de requerimientos de capital, reservas técnicas, estándares de administración de riesgos y mecanismos de revisión para reducir la probabilidad de insolvencia de las aseguradoras y reaseguradoras (Aguilera, 2012).

Cotidianamente, es frecuente tomar decisiones bajo condiciones de incertidumbre. El riesgo se deriva de la incertidumbre y siempre es asociado con una posible pérdida o ganancia, por lo cual medirlo es fundamental (Peng, 2010). Las medidas de riesgo son una representación que asocia los eventos de interés a números reales por medio de variables aleatorias, y éste es el principio de los modelos de pérdidas.

El modelado de la pérdida agregada es uno de los objetivos principales de la teoría y práctica actuarial, especialmente en el sector asegurador cuando se tienen que hacer importantes decisiones de negocios respecto a algunas características de los contratos de seguros (Mohamed, et al., 2010).

El término distribución de pérdida pretende ser muy general, ya que puede representar una distribución de pérdida por reclamo, una distribución de pérdida por ocurrencia, una distribución de razón de pérdidas anuales, etcétera. Esta generalidad y amplia aplicación de las funciones de distribución de probabilidad básicas, es resultado de varios tipos de distribuciones de pérdida específicas y modelos probabilísticos, que pueden ser utilizados en varias áreas de generación de tarifas.

La mayor parte de la profesión actuarial se concentra en el desarrollo y análisis de modelos matemáticos, que expliquen el proceso por el cual los fondos entran y salen del sistema asegurador. Los modelos de pérdidas explican una gran parte de las salidas de los fondos debido al pago de beneficios (Klugman, et al., 2004).

En el sector de seguros de no vida (gastos médicos y daños), una de las principales variables para estimar la utilidad, tarificar productos, proyectar la situación de la compañía en el futuro, crear estrategias de negocio, entre otros, es el monto que la institución tendrá que pagar derivado de las pólizas vendidas

que resulten en reclamo, conocido como pérdida agregada (Papush, et al., 2001). Dicho monto es una variable aleatoria que se desconoce hasta el momento en el que ocurre, es decir, el asegurador no sabe con certeza qué cantidad tendrá que pagar a sus asegurados durante un periodo determinado, hasta que dicho periodo ha transcurrido.

Estimar una distribución de pérdidas agregadas basada en la frecuencia y severidad, permite tarificar el reaseguro de pérdida en exceso incorporando las características del contrato, como lo son deducibles, líneas de negocio y capas de reaseguro. En un contrato de reaseguro de pérdida en exceso que incluye sensibilidades, las primas y gastos se vuelven variables aleatorias que dependen de la pérdida agregada. Así, con una función de distribución apropiada de la pérdida agregada, es posible estimar las primas y los gastos esperados para evaluar la rentabilidad del contrato de reaseguro (Mata et al., 2002).

Tener un estimado de la pérdida agregada de una compañía de seguros es básico y necesario para que la empresa sea financieramente sólida y pueda seguir operando con un rumbo definido, por lo cual, se han desarrollado varios modelos para obtener un aproximado de dicha cantidad, estos modelos son la descripción del comportamiento de un conjunto de riesgos formado por un determinado portafolio de pólizas (Burnecki, et al., 2010).

Una opción útil para estimar la pérdida agregada es el modelo de riesgo individual, mediante el cual se podrán resolver algunas incógnitas, que ayudarán en la toma de decisiones y creación de estrategias para la compañía de seguros. Los modelos probabilísticos, proporcionan una descripción de la exposición al riesgo, no obstante, necesitamos describir esta exposición con algún valor numérico o con unos pocos valores numéricos. Estos valores son funciones del modelo y se conocen como indicadores clave de riesgo. Entonces, estos indicadores son resúmenes numéricos sobre el comportamiento de nuestro

riesgo, parecidos a los que utilizamos para describir una variable aleatoria. Estos indicadores informan a los actuarios y a los administradores de riesgo, sobre el grado en que la compañía está sujeta a un particular aspecto del riesgo.

Los modelos de pérdidas agregadas conforman un papel que se está fortaleciendo y, tomando mayor importancia en la tarificación de las coberturas de seguros. Con la competencia que existe en el sector asegurador, es muy importante obtener estimados lo más exactos posible de las pérdidas que podrían derivarse de tales contratos de seguro y de las variaciones que pudieran presentarse, para así poder ser una compañía sólida financieramente y competitiva en el mercado (Heckman y Meyers, 1983).

La razón por la que se quiere crear un modelo para pronosticar las pérdidas de una aseguradora en lugar de utilizar solamente los datos empíricos, es que si se toman decisiones basadas únicamente en la experiencia pasada, se estaría asumiendo que el futuro se comportará de la misma manera, además, se desconocería la distribución de pérdida para valores no observados en las muestras. En los ramos de daños y gastos médicos es muy poco probable que la siniestralidad de un año sea igual a la del siguiente. Los datos empíricos de severidad y frecuencia tienen la ventaja de estar disponibles, pero algunos detalles importantes están ocultos cuando sólo se observan los datos de pérdidas agregadas (Heckman y Meyers, 1983). Modelar las pérdidas con distribuciones probabilísticas permitirá obtener más información acerca del comportamiento de las mismas.

Estos modelos están calculados en términos de distribuciones que explican el comportamiento del monto y número de reclamos. El describir las pérdidas agregadas en términos de frecuencia y severidad por separado, permite hacer modificaciones sobre los contratos de seguro, como lo son el deducible, suma asegurada máxima y coaseguro.

Una aplicación del modelo de riesgo colectivo es incluir la variable tiempo en el proceso de ocurrencia de los reclamos. Aquí, el proceso de conteo de ocurrencia de las pérdidas es representado por $\{N_t\}_{t\geq 0}$ y el monto de pérdida es definido por $\{X_k\}_{k=1}^{\infty}$ que es una secuencia de variables aleatorias independientes e idénticamente distribuidas. Entonces, el proceso de riesgo estará dado por

$$R_t = u + c(t) - \sum_{i=1}^{N_t} X_i$$

Donde

u = Capital inicial del asegurador.

$c(t)$ = Prima de pólizas vendidas.

$\sum_{i=1}^{N_t} X_i$ = Proceso de pérdida agregada del número de reclamos en el intervalo (o,t].

El modelo de riesgo colectivo es también usado en seguros de salud y seguros en general, donde los principales componentes del riesgo son el número de reclamos y su monto. Este modelo también puede ser usado para modelar el riesgo de productos fuera del sector asegurador, como el riesgo de crédito y el riesgo operacional.

Feria, et al., (2007) aplican el modelo de pérdidas agregadas, junto con el OpVaR, para calcular el capital en riesgo de una institución bancaria dividiendo la información por línea de negocio y por tipo de riesgo, obteniendo la distribución de pérdidas agregadas para cada línea de negocio y tipo de riesgo. Para el cálculo del valor regulatorio, se obtiene el VaR, el cual será definido más adelante, que aplicándolo al contexto de riesgo operacional le llaman OpVaR, el cual representa un percentil de la distribución de pérdidas.

II. MARCO TEÓRICO Y CONCEPTUAL

La pérdida agregada es el monto total que tendrá que pagar un asegurador por los reclamos ocurridos en un periodo de tiempo determinado en un cierto portafolio de pólizas (Klugman, *et al.,* 2004).

Un modelo es una representación simplificada de algún fenómeno real. Un modelo matemático describe el comportamiento de un sistema real mediante el uso de símbolos matemáticos, funciones y ecuaciones (Patrik, 1980) y el construir modelos ayuda a tener una visión estructurada de la realidad y aclarar las alternativas que se tienen para tomar una decisión y anticipar sus efectos (Wagner, 1969).

En un contexto actuarial específico, proponer un modelo para describir una situación, se basa en la experiencia y conocimiento que el Actuario tenga del fenómeno bajo estudio, así como en la información histórica que posea sobre él. El modelo debe proveer un balance entre simplicidad (parsimonia) y conformidad (ajuste) con la información disponible para elaborarlo.

Patrik (1980) define algunas de las características que definen un modelo probabilístico útil para describir las pérdidas del asegurador:

a) Que sea fácil de entender. Sus características principales deben ser claramente descritas y medidas.
b) Puede trabajarse con dicho modelo fácilmente.
c) Debe ser fácil extenderlo a casos generales o análogos de una manera consistente.
d) Puede ser restringido fácilmente a casos particulares, por ejemplo, la distribución de pérdidas de una póliza con un límite determinado puede ser una restricción de una distribución general de pérdidas ilimitadas.
e) Puede ser probado usando métodos estadísticos.

f) Puede usarse para comparar o combinar varios contratos o portafolios.

Schipper (2010) define el riesgo como cualquier pérdida futura incierta, acentuando que lo que separa un riesgo de cualquier gasto ordinario es la incertidumbre sobre su ocurrencia y su magnitud, éste el el fundamento de la industria de riesgos (aseguradoras, fondos de pensiones, bancos y "traders"). Cuando muchos riesgos pequeños se agrupan son más fáciles de predecir y administrar.

Conocer la distribución de la pérdida agregada permitirá estimar de mejor manera la prima de riesgo, la cual es el costo que paga el asegurado al asegurador por transferir su riesgo. Esta prima es el costo esperado de la siniestralidad, y corresponde a la porción de la prima de tarifa que debe destinarse para el pago de las reclamaciones por concepto de siniestros; la prima de tarifa es la cantidad necesaria para cubrir los costos futuros, es decir, es una estimación del valor actual de los costos futuros esperados (Muñoz, 2006) y está compuesta por la prima de riesgo, costos de adquisición, costos de administración y el margen de utilidad. Bowers, et al. (1997) definen a la prima de riesgo como el precio establecido por el asegurador de la cobertura total del seguro, es decir, la pérdida esperada por siniestro, *E[X].*

Larose (1982) define cuatro funciones elementales y fundamentales para la distribución de pérdidas:

a) Función de distribución acumulada

   Esta función representa la probabilidad de que una pérdida dada sea menor o igual que un monto específico x.

$$F(x) = \int_0^x f(t)dt$$

b) Función de pérdidas básica

Esta función representa el porcentaje de las pérdidas totales generado por todos los reclamos que son menores que algún valor específico x.

$$X1(x) = \frac{1}{\alpha}\int_0^x tdF(t),$$

Donde $\alpha = \int_0^\infty t\, dF(t)$

c) Función de pérdidas primaria

Esta función representa el porcentaje de pérdidas totales generado por los montos agregados de los primeros x pesos de cada reclamo (el monto total de reclamo si este es menor o igual que x)

$$X2(x) = \frac{1}{\alpha}\int_0^x tdF(t) + \frac{x}{\alpha}\int_x^\infty dF(t) = X1(x) + \frac{x}{\alpha}[1 - F(x)]$$

d) Función de pérdida en exceso

Esta función representa el porcentaje de pérdidas totales generado por los montos agregados de las pérdidas que excedieron x monto por reclamo.

$$X3(x) = \frac{1}{\alpha}\int_x^\infty (t-x)dF(t) = 1 - X1(x) + \frac{x}{\alpha}[1 - F(x)] = 1 - X2(x)$$

Para trabajar con las funciones anteriores, se requiere contar con una función de distribución que ajuste apropiadamente a la experiencia de la aseguradora y que dicho modelo permita obtener valores de interés como los obtenidos con las funciones anteriores.

El valor esperado se define de la siguiente manera. Considérese una variable aleatoria S como el monto de pérdida relativo a un reclamo ocurrido. Sea $g(S)$ el

pago del reclamo que debe hacerse por un monto de pérdida S y $F_S(x)$ una distribución donde *S* toma sólo valores positivos. Entonces,

$$E[g(S)] = \sum_{k=1}^{\infty} g(k)p_k = \sum_{k=1}^{\infty} g(k) * Prob[S = k]$$

Las probabilidades pueden interpretarse como frecuencias relativas idealizadas, es decir

$$prob[S = k] \sim n_k/n$$

donde n_k representa el número de casos de todos los casos similares para los que el monto de pérdida $S = k$. Así,

$$E[g(S)] \sim \frac{\{\sum g(k)n_k\}}{n} = \frac{pago\ del\ total\ de\ reclamos}{numero\ total\ de\ reclamos}$$

$$= pago\ promedio\ por\ reclamo$$

$E[g(S)]$ representa el valor de la media teórica de $g(S)$; esta media teórica es el valor esperado.

Otra medida de interés para un asegurador es el VaR. Éste es una medida del riesgo de tipo estadístico y probablemente la más utilizada en el sector financiero. Esta medida de riesgo, informa a los Actuarios y administradores de riesgo, sobre el grado en que la compañía está sujeta a un aspecto particular del riesgo.

Una medida de riesgo está en correspondencia con la pérdida asociada a este riesgo. Y proporciona un solo número que intenta cuantificar la exposición a este riesgo.

Wang, et al (1997) introducen ciertos axiomas, que representan propiedades deseables de una medida de riesgo. Artzner et al. (1997) introduce el concepto de coherencia y se considera el parteaguas en medición de riesgo.

Una medida de riesgo coherente, es una medida de riesgo $\rho(x)$ con las cuatro propiedades siguientes. Para cualesquiera dos variables aleatorias de pérdida *X* y *Y*:

a) Subaditividad

$$\rho(X+Y) \leq \rho(X) + \rho(Y)$$

b) Monotonía: Si $X \leq Y$ para todas las posibles consecuencias, entonces

$$\rho(X) \leq \rho(Y)$$

c) Homogeneidad positiva: Para cualquier constante positiva c,

$$\rho(cX) = c\rho(X)$$

d) Invarianza a translaciones: Para cualquier constante positiva c,

$$\rho(X+c) = \rho(X) + c$$

Subaditividad significa que la medida de riesgo (y, por lo tanto, el capital requerido para darle soporte) de dos riesgos combinados, no será mayor que los riesgos considerados por separado. La subaditividad refleja el hecho de que diversificar el riesgo puede ser benéfico para una empresa.

El VaR se ha vuelto la medida estándar para medir exposición al riesgo. En términos generales, el VaR es el capital requerido para asegurar, con alto grado

de certeza, que la empresa no será técnicamente insolvente. El grado de certeza se elige de manera arbitraria. En la práctica se pueden elegir valores grandes como 99.95% para toda la empresa, o 95% para una sola clase de riesgo.

Entonces, el VaR mide la pérdida que se podría sufrir, en condiciones normales del mercado, en un intervalo de tiempo y con un determinado nivel de probabilidad o confianza.

Se define como la máxima pérdida esperada para un nivel de confianza y un periodo de tiempo (Mascareñas, 1998). Se determina con base al 100(1-α)% percentil (q1-α). Su definición formal es, dado un nivel de confianza $1 - \alpha$ con $\alpha \in (0,1)$, el VaR a un nivel de confianza α está dado por el número más pequeño l, tal que la probabilidad de que las pérdidas excedan l es menor o igual a α, i. e.,

$$VaR_{1-\alpha} = \inf\{l \in R : P(S > l) \leq \alpha\} = \inf\{l \in R : F_S(l) \geq 1 - \alpha\}$$

Pese a la popularidad de esta medida, es conveniente aclarar que el VaR no es una medida subaditiva, lo que la convierte en una medida de riesgo incoherente y no refleja el efecto por la diversificación de los riesgos.

Otra medida de riesgo importante es el TVaR (*Tail Value at Risk*) definida de la siguiente manera:

Sea X una variable aleatoria con función de distribución $F(X)$, que denota pérdida. El TVaR de X al 100%p nivel de seguridad, denotado por $TVaR_p(X)$, es la pérdida esperada dado que ésta excede el p-ésimo percentil de la distribución de X.

$$TVaR_p(X) = E\left(X \middle| X > \pi_p\right) = \frac{\int_{\pi_p}^{\infty} xf(x)dx}{1 - F(\pi_p)}$$

Una forma alternativa más interesante de escribir esta cantidad, es:

$$TVaR_p(X) = E\left(X \middle| X > \pi_p\right) = \frac{\int_p^1 VaR_u(X)dx}{1-p}$$

Esta expresión del TVaR implica que puede verse como un promedio de todos los valores VaR por encima del valor de seguridad *p*. Lo que significa que proporciona mucho más información sobre la cola de la distribución que la que da el VaR.

El TVaR recibe otros nombres en el campo de los seguros como Conditional Tail Expectation (CTE), Tail Conditional Expectation (TCE) y Expected Shortfall (ES).

Finalmente, se puede decir que el TVaR es el valor esperado de las perdidas en aquellos casos en que se excede el nivel de seguridad previamente fijado. El TVaR refleja con mayor fidelidad los eventos extremos que pueden amenazar la posición financiera de la entidad.

Contrario al VaR, TVaR es una medida de riesgo coherente, por lo cual refleja el efecto de la diversificación de riesgos.

Dado un umbral de seguridad o confianza, en datos reales, el TVaR es una medida más difícil de calcular que el VaR, ya que ambas se calculan con los datos (generalmente escasos), acumulados en la cola de la distribución (mayor error de estimación).

2.1 Modelos de riesgo

Daykin, *et al.* (1987), mencionan tres fuentes de incertidumbre en el sector asegurador, la primera debido a la ocurrencia de las obligaciones, la segunda

acerca de la suficiencia de activos para hacer frente a las obligaciones en el tiempo en el que ocurren y, por último, la utilidad o pérdida de primas futuras y parte de las primas no devengadas de eventos que aún no ocurren.

Este trabajo está enfocado en la primera fuente de incertidumbre. Como ya se mencionó anteriormente, hay dos enfoques principales para estimar las pérdidas de una institución aseguradora, el modelo de riesgo individual (IRM) y el modelo de riesgo colectivo (CRM).

La teoría del riesgo colectivo, inició con Filip Lundberg y, posteriormente, fue desarrollada por Cramér, Arfwedson, Segerdahl, Saxén, entre otros. Esta teoría considera dos principales problemas, encontrar la función de distribución del monto de reclamos total en un portafolio y, encontrar la probabilidad de que la reserva de riesgo se extinga hasta llegar a la ruina (Markham, 1962).

En el modelo de riesgo individual, la variable aleatoria de interés es el reclamo total de un portafolio de contratos de seguro (o pólizas). El total de reclamos es modelado como la suma de reclamos en las pólizas, que se asume son independientes.

Aunque este modelo es realista, no siempre es conveniente, ya que los datos se utilizan de manera integral. Un modelo utilizado para aproximar el modelo de riesgo individual es el modelo de riesgo colectivo. En este, el portafolio de seguros es visto como un proceso que produce reclamos en el tiempo y, en el cual los montos de reclamo son vistos como variables aleatorias independientes, idénticamente distribuidas e independientes del número de reclamos

En la práctica actuarial, los riesgos generalmente no pueden modelarse como variables aleatorias puramente discretas, ni tampoco puramente continuas, aunque existen dos excepciones, la primera es la probabilidad de no tener reclamos, que es bastante grande y, segunda, la probabilidad de que el monto

de reclamo exceda la suma asegurada máxima. Para calcular el valor esperado de esta mezcla de variables aleatorias se utiliza la integral de Riemann-Stieltjes. Un modelo simple y flexible que produce variables aleatorias de este tipo es un modelo mixto. Dependiendo del resultado del evento: no ocurre reclamo, máximo monto de reclamo u otro reclamo (Kaas et. al, 2008).

Asumiendo que los riesgos de un portafolio son variables aleatorias independientes, la distribución de la suma de éstos puede ser calculada mediante el uso de convoluciones, esta técnica será definida más adelante, aunque es algo laboriosa, por lo cual se usan otras alternativas como, la función generadora de momentos o algunas transformaciones relacionadas como funciones características y funciones generadoras de probabilidad.

Otro enfoque totalmente diferente es aproximar la distribución de la suma de los reclamos totales como la suma de un “gran” número de variables aleatorias, que por virtud del teorema del límite central, aproxima su distribución mediante la distribución normal con la misma media y varianza de esta suma. Esta no es una buena aproximación en la práctica, ya que en las colas de la distribución, se necesitan aproximaciones que reconozcan explícitamente la probabilidad de grandes reclamos. El tercer momento central de la distribución de los reclamos totales es generalmente mayor a cero, mientras que en la distribución normal es igual a 0. Como alternativas existen la aproximación basada en la variable aleatoria gamma trasladada y la *normal power approximation*.

En estadística, casi sin excepción, las variables aleatorias son discretas o continuas, pero este no es el caso en seguros. Muchas de las funciones de distribución que se emplean para modelar pagos de seguros, tienen partes de incremento continuo, pero también saltos positivos, por ejemplo para el pago de un contrato de seguros hay tres posibilidades:

a) El contrato está libre de reclamo (el pago es igual a 0).

b) El contrato genera un reclamo que es mayor a la suma asegurada máxima.
c) El contrato genera un reclamo que es mayor a 0 y menor a la suma asegurada máxima.

Para el último punto se podría usar una función de distribución discreta, ya que el pago será un múltiplo de la unidad monetaria, pero esto produciría un conjunto muy grande de valores posibles, cada uno con una probabilidad muy pequeña, por lo cual es más factible utilizar una función de distribución continua, formando así una función de distribución acumulada que no es puramente continua ni puramente discreta.

En el modelo de riesgo colectivo, así como en el individual, se calcula la distribución del monto total de reclamos en un cierto periodo de tiempo, aunque en algunos casos se observa el portafolio como una población que produce un reclamo en puntos de tiempo aleatorios.

En dicho modelo, se considera el número de reclamos que genera el portafolio y no el número de contratos o pólizas de seguro. Si un portafolio contiene sólo una póliza que puede generar un monto de reclamo alto, este término aparecerá a lo más una vez en el modelo de riesgo individual, mientras que en el modelo de riesgo colectivo pueden ocurrir varios reclamos. Además, en este modelo se requiere que el número de reclamos y el monto de éstos sean independientes.

La ventaja principal del modelo de riesgo colectivo es que es computacionalmente eficiente y, bastante cercano a la realidad. La función de distribución de los reclamos totales bajo este modelo puede calcularse por medio de convoluciones o la recursión de Panjer, entre otros.

Kaas et. al (2008) definen formalmente el IRM de la siguiente manera:

Sea Z el monto de reclamo sobre un contrato de seguro para el cual existen tres posibilidades:

1. *Z=0*
2. *Z=M*
3. *0<Z<M*

Donde *M* es la suma asegurada máxima y sea I una variable aleatoria indicadora con valores $I = 1$ o $I = 0$, donde $I = 1$ significa que el reclamo ha ocurrido. Suponga que la probabilidad del evento es $q = Pr[I = 1]\,, 0 \leq q \leq 1$. Si $I = 1$, el reclamo *Z* puede ser obtenido de la distribución de *X y si* $I = 0$ *Z* puede ser obtenido de la distribución de *Y.* Esto significa:

$$Z = IX + (1 - I)Y$$

Si $I = 1$, *Z* puede reemplazarse por *X*, y si $I = 0$, *Z* puede reemplazarse por *Y*. Entonces la función de distribución acumulada puede escribirse como:

$$\begin{aligned} F(z) &= Pr[Z \leq z] \\ &= Pr[Z \leq z \,\&\, I = 1] + Pr[Z \leq z \,\&\, I = 0] \\ &= Pr[X \leq z \,\&\, I = 1] + Pr[Y \leq z \,\&\, I = 0] \\ &= qPr[X \leq z\,] + (1 - q)Pr[Y \leq z] \end{aligned}$$

En el IRM la variable de interés es:

$$S = S^{ind} = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

En el cual *n* es fijo y representa el número de pólizas en la cartera de estudio. En éste modelo no es necesario que las variables aleatorias X_i sean iguales en distribución, si lo son, a este modelo se le llama modelo de riesgo individual homogéneo.

Varios autores afirman que la diferencia entre el IRM y CRM es la forma de tratar las pólizas, Verral (1989) muestra la relación entre ambos modelos mediante el siguiente desarrollo.

En el IRM, *n* representa el número de pólizas y no el número de reclamos, entonces, algunos X_i tendrán el valor de cero cuando no haya ocurrido reclamo en la póliza. Así, X_i tiene dos componentes aleatorios, el evento de que ocurra reclamo y el tamaño de éste, i. e.,

$$X_i = I_i B_i$$

Donde:

$$I_i = \begin{cases} 1 & si\ existe\ un\ reclamo\ en\ la\ i-ésima\ póliza \\ 0 & cualquier\ otro\ caso \end{cases}$$

y

$B_i = monto\ de\ reclamo\ de\ la\ i-ésima\ póliza$

Así se puede transformar S^{ind} en:

$$S^{ind} = B_1 + B_2 + \cdots + B_r + O_{r+1} + O_{r+2} + \cdots + O_n$$

$$S^{ind} = B_1 + B_2 + \cdots + B_r$$

Escalante y Arango (2004) definen al CRM como la suma de un número aleatorio de variables aleatorias, es decir,

$$S = S^{col} = X_1 + X_2 + \cdots + X_N = \sum_{i=1}^{N} X_i$$

donde *X* y *N* son variables aleatorias independientes e idénticamente distribuidas (v.a.i. i.d.), X_i representa la severidad del i-ésimo reclamo individual y *N* el número de reclamos en un periodo de observación. Kaas et al (2008)

hacen notar que en el caso del modelo de riesgo colectivo, $S=0$ si $N=0$, por lo tanto, los términos de S corresponden a los reclamos ocurridos.

De esta forma, S es una suma aleatoria de variables aleatorias. Nótese que el IRM es una distribución Binomial compuesta, ya que se puede tener a lo más un reclamo por póliza, mientras que el CRM es una distribución Poisson compuesta, eliminando dicha restricción. Así, en el IRM, se ignora parte de la información, ya que si un portafolio contiene una sola póliza, que podría generar un monto de reclamo alto, éste aparecerá a lo más una vez, mientras que en el modelo de riesgo colectivo es posible que ocurra varias veces (Kaas, et al.., 2008). Así, el modelo de riesgo colectivo es un caso especial del modelo de riesgo individual.

La naturaleza aleatoria de la frecuencia y la severidad son supuestos básicos de un modelo de riesgo realista. El supuesto de independencia entre frecuencia y severidad permite modelar el número de reclamos y el monto de pérdida de manera separada.

La frecuencia de las reclamaciones es una medida importante para calcular el monto de reclamo total, que es el número de reclamos en un bloque de pólizas de seguros en un periodo de tiempo. Aunque la frecuencia no muestra directamente las pérdidas monetarias de los reclamos, es una variable importante en la modelación de las pérdidas. El monto de cada reclamo es llamado severidad. La estimación general para modelar la pérdida agregada es considerar la frecuencia y la severidad de manera separada y luego combinarlas (Tse, 2009).

El primer y principal problema que se presenta al querer pronosticar la pérdida agregada es, que dicho monto es una cantidad aleatoria, que depende del número de siniestros que resultan en reclamación a lo largo del periodo de vigencia de las pólizas, así como del monto reclamado en cada uno de éstos, es

decir, se tienen dos componentes estocásticos: frecuencia y severidad. Lo anterior representa una suma con una cantidad aleatoria de variables aleatorias.

Los avances en el desarrollo de técnicas que permitan resumir, analizar, describir, ajustar, simular y estimar dichos fenómenos aleatorios, desde el suavizamiento de gráficas hasta los métodos de momentos, máxima verosimilitud y algunos otros mucho más avanzados, complejos y sofisticados que requieren el uso de programas computacionales, así como el número de modelos y herramientas de diagnóstico ha incrementado en gran manera con el desarrollo tecnológico de la época (Klugman y Rioux, 2006). Es objeto de este trabajo presentar algunas alternativas en la modelación de la pérdida agregada y elegir la que mejor describa la información observada.

Los dos enfoques para modelar la pérdida agregada son el modelo de riesgo individual y el modelo de riesgo colectivo, los cuales han resultado de considerar un portafolio de pólizas en diferentes maneras. El modelo de riesgo colectivo (CRM) se deriva de ver un portafolio como un todo mientras que el modelo de riesgo individual permite tratar cada póliza por separado.

El modelo de riesgo colectivo se define como:

$S = X_1 + X_2 + \ldots + X_N$, con $N = 0, 1, 2, \ldots$

Donde $S=0$ cuando $N=0$ y X_ks son variables aleatorias independientes e idénticamente distribuidas a menos que se especifique de otro modo (Klugman, et al., 2004).

El modelo de riesgo individual (IRM) representa la pérdida agregada como una suma, $S = X1 + \cdots + Xn$ de un número fijo, n, de contratos de seguro. Los montos de pérdida de cada contrato son $(X1, X2, \ldots, Xn)$, donde las Xj se asumen independientes pero no idénticamente distribuidas. Este modelo es usado para

unificar las pérdidas o pagos de un determinado número de contratos de seguro o bloques de contratos. Se usa para modelar pérdidas en seguros colectivos, por ejemplo, el seguro de gastos médicos de los *n* empleados de una compañía, donde cada empleado puede tener diferentes coberturas y diferente probabilidad de reclamo. En el caso en que los montos de reclamo ($X_j s$) sean idénticamente distribuidos, el modelo de riesgo individual se transforma en un caso especial del modelo de riesgo colectivo, en donde la v. a. *N* asigna toda su probabilidad al valor de *n*, es decir, $P\ (N = n) = 1$.

De hecho, el IRM puede ser obtenido de la misma manera que el CRM y puede considerarse como una distribución binomial compuesta. Esto permite que exista un estudio unificado de los modelos de riesgo, simplifica el cálculo de la media y la varianza del IRM y facilita el cálculo de momentos más grandes.

Schipper (2010) menciona claramente la principal desventaja del modelo de riesgo individual, y es que éste sólo permite un reclamo por póliza para lo cual hay dos posibilidades: la primera es que el contrato sólo permita a lo más un reclamo, lo cual no es común y, la segunda es que si ocurren varios reclamos en una póliza éstos se agrupen y se vean como una sola pérdida. Al tener varios reclamos agrupados se pierde información, ya que no se puede distinguir si las pérdidas fueron derivadas de un reclamo o si ocurrieron muchas pérdidas derivadas de muchos reclamos. Por ejemplo, si a un asegurado de gastos médicos mayores se le detecta una enfermedad que requiere un tratamiento largo con varias consultas o periodos de internación en el hospital, ese reclamo (también llamado siniestro) requerirá que el asegurador haga varios pagos en diferentes ocasiones. Sin embargo, este modelo es válido si la probabilidad de que ocurra más de un reclamo durante un periodo de tiempo determinado es pequeña, cuando los datos sólo están disponibles en agregado o cuando el contrato permita sólo un reclamo por póliza.

Una ventaja del IRM contra el CRM es que no es necesario que las pólizas tengan la misma distribución de severidad, al analizar póliza a póliza se permite asignar una distribución de probabilidad distinta a cada una.

Utilizar un modelo de riesgo, como los descritos anteriormente, permite modificar el número de reclamos esperados al cambiar el número de pólizas, lo cual es indispensable en los pronósticos de años futuros basados en datos históricos; facilita el estudio de los impactos provocados por cambiar deducibles y límites en sumas aseguradas; se entiende mejor el impacto que tiene el número de reclamos al cambiar el deducible; los modelos desarrollados para pérdidas no cubiertas por el asegurador, costos de reclamo para el asegurador y el costo de reaseguro son consistentes, esto ayuda cuando el asegurador quiere transferir sus riesgos al reasegurador (Klugman, et al., 2004).

Además, el modelar por separado la severidad y frecuencia, y luego combinar sus distribuciones de probabilidad, permite separar los impactos que afectan a una variable en específico, por ejemplo, el crecimiento de la cartera de pólizas de un asegurador puede afectar el número de reclamos pero no la severidad (Tse, 2009).

El cálculo de la distribución de pérdidas agregadas ha sido complejo debido a sus dos componentes aleatorios, X y N. Por lo anterior, esta distribución no siempre tiene una forma analítica con la cual sea sencillo trabajar. Por dicha razón, se han desarrollado varias alternativas para estimar la distribución de S, como los siguientes:

a) Convoluciones.
b) Transformaciones (vía función generadora de momentos).
c) Aproximaciones (Gamma trasladada y *normal power approximation*).
d) Recursiones (Panjer).
e) Inversión (transformada rápida de Fourier).

f) Simulación.

El cálculo de la pérdida agregada se obtiene con base en la distribución de frecuencias de las pérdidas $p_n = P(N=n)$ y con la función de probabilidad de la severidad de las mismas $f_X(x)$ o la función de distribución acumulada $F_X(x)$.

La convolución es la distribución de probabilidad de la suma de variables aleatorias independientes. Para encontrar la suma de dos variables aleatorias, X_1 y X_2, se define como

$$F_S(s) = \sum_{x_2 \leq s} \Pr\left(X_1 + X_2 \leq s | X_2 = x_2\right) \Pr\left(X_2 = x_2\right)$$

para el caso discreto, y como

$$F_S(s) = \int_0^s F_{X_1}(s - x_2) f_{X_2}(x_2) dx_2$$

para el caso continuo.

Para determinar la suma de más de dos variables aleatorias se usa iterativamente el proceso de convolución. Para $S=X_1+X_2+\ldots+X_n$, donde las X_is son variables aleatorias independientes, F_i es la función de distribución de X_i y $F^{(k)}$ es la función de distribución de $X_1+X_2+\ldots+X_k$, se tiene

$F^{(2)}=F_2*F^{(1)}= F_2*F_1$

$F^{(3)}=F_3*F^{(2)}$

Y así sucesivamente hasta

$F^{(n)}=F_n * F^{(n-1)}$

La aproximación Gamma trasladada es muy utilizada, ya que la mayoría de las distribuciones del monto de reclamos total tiene la misma forma que la distribución Gamma, sesgada a la derecha. A parte de los parámetros α y β, se agrega un tercer grado de libertad, permitiendo un traslado sobre la distancia x_0. Así, se aproxima la función de distribución acumulada *S* mediante la función de distribución acumulada $Z + x_0$, donde $Z \sim Gamma(\alpha, \beta)$. Se eligen α, β y x_0 de manera que la variable aleatoria aproximada tenga los mismos primeros tres momentos que *S*.

La aproximación Gamma trasladada puede ser formulada de la siguiente manera:

$$F_S(s) \approx G(s - x_0; \alpha, \beta),$$

Donde

$$G(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \int_0^x y^{\alpha-1} \beta^\alpha e^{\beta y} dy, \quad x \geq 0$$

Para asegurar que α, β y x_0 se elijan de manera que los 3 primeros momentos concuerden, es decir $\mu = x_0 + \frac{\alpha}{\beta}$, $\sigma^2 = \frac{\alpha}{\beta^2}$ y $\gamma = \frac{2}{\sqrt{\alpha}}$, deben satisfacer:

$\alpha = \frac{4}{\gamma^2}$, $\beta = \frac{2}{\gamma\sigma}$ y $x_0 = \mu - \frac{2\sigma}{\gamma}$

Otra aproximación, muy general, es calcular los momentos de la distribución de pérdidas en términos de los momentos de la distribución de la severidad y el número de reclamos. Se igualan los momentos de la distribución de pérdidas agregadas con la distribución asumida.

Un método muy popular para calcular la distribución de la pérdida agregada es la simulación de Monte Carlo.

Un tercer método para calcular la distribución de pérdidas agregadas se trata de invertir su función característica (transformada rápida de Fourier). Este método requiere que se tenga una representación explícita de la función característica de la distribución del monto de reclamo.

Un cuarto método es el método recursivo, el cual asume una distribución discreta del monto de reclamos. Escogiendo un número suficientemente grande de puntos de la distribución del monto de reclamos, se puede obtener el grado deseado de exactitud invirtiendo la transformada de Laplace de la distribución de pérdidas agregadas (Tilley, 1980).

La recursión de Panjer es un método que ha demostrado ser eficiente para calcular la distribución exacta de un proceso compuesto que satisface las siguientes condiciones:
a) Que la distribución primaria (distribución de *N*) pertenezca a la familia (*a, b, 0*).
b) Que la distribución secundaria (Distribución de *X*) sea discreta y valuada en los enteros no negativos.

El presente trabajo estará enfocado en encontrar un modelo que ajuste el comportamiento de los datos de una compañía aseguradora en un lapso de tiempo determinado, para estimar la prima de riesgo y algunos otros datos de interés para la institución. Lo anterior se hará siguiendo los pasos mencionados por Klugman, et al., (2004):

a) Pueden seleccionarse uno más modelos de acuerdo al conocimiento inicial y experiencia que posea el analista, además de la naturaleza de la información disponible.

b) Ajustar el modelo con la información disponible.

c) Realizar pruebas de diagnóstico del modelo, para determinar si su ajuste es adecuado para los datos utilizados.

d) Considerar, a partir del paso anterior, la posibilidad de utilizar otros modelos.

e) Si existen varios modelos que pueden ser adecuados, entonces, es necesario compararlos con la finalidad de decidir por alguno de ellos.

f) Finalmente, el modelo seleccionado puede adaptarse para aplicarlo en el futuro. Esto puede involucrar algún ajuste de los parámetros, previendo cambios por alguna característica exógena, como inflación, cambios del mercado asegurado o cualquiera otra.

2.2 Métodos de aproximación y distribuciones usadas en la modelación de la pérdida agregada

Burnecki, et al., (2010), proponen el proceso Poisson, el proceso Poisson no homogéneo, el proceso Poisson mixto y el modelo de renovación para modelar el proceso de ocurrencia de reclamos. Bühlmann (1970) menciona las distribuciones más utilizadas para modelar la frecuencia, entre ellas están las distribuciones Binomial, que modela la probabilidad de tener *k* éxitos en *n* intentos independientes e idénticamente distribuidos; la distribución Poisson que surge como un caso de la binomial cuando el número de intentos es muy grande

y la media de la Biniomial es igual a λ, otros ejemplos son las distribuciones Binomial Negativa y la Logarítmica.

Dado que el monto del siniestro es una cantidad no negativa, los modelos probabilísticos asociados deben contemplar esta y otras características.

Concretamente:

La variable asociada al monto de reclamación es mayor o igual que cero (no negativa).

$$X : \Omega \rightarrow [0, \infty)$$

La distribución de esta variable es generalmente sesgada a la derecha.

La distribución puede ser de colas pesadas, lo que podría implicar el uso de distribuciones para valores extremos en su modelación.

Para modelar la severidad, Burnecki, et al., (2005) mencionan tres métodos para aproximar la distribución de la pérdida. La primera es el método empírico, el cual puede ser utilizado sólo cuando el conjunto de datos con los que se cuenta es muy grande, lo cual, es subjetivo y depende, asimismo, del comportamiento de las observaciones. El siguiente enfoque es la aproximación analítica, que es el más común y se reduce a encontrar una expresión analítica adecuada, que se ajuste bien a los datos observados y sea fácil de manejar. Y por último, para los casos en que la forma exacta de la distribución no es necesaria, se puede usar la aproximación por el método de momentos, que consiste en estimar las características más bajas de la distribución (momentos).

Raramente se da el caso de tener un gran volumen de datos para modelar la cola de una distribución, por lo cual, el enfoque empírico es inútil si se requiere modelar la ocurrencia de siniestros extremadamente grandes. En estos casos es

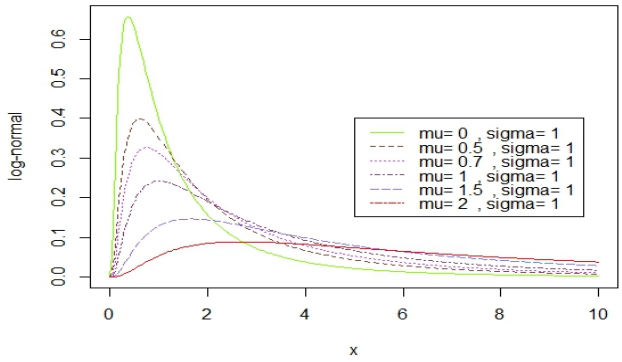
recomendable dividir el rango de valores en dos partes, trabajar con el tamaño de los siniestros hasta cierto límite y reemplazar la cola por una función de distribución acumulada de forma analítica (Burnecki, et al., 2005).

Si los datos con los que se cuenta son muy dispersos para usar el enfoque empírico, se desea encontrar una expresión analítica explícita para la distribución de las pérdidas. Burnecki, et al., (2005), plantean las características más importantes de las distribuciones más usadas para modelar la severidad (Lognormal, Pareto, Burr, Weibull y Gamma).

La distribución más conocida es la distribución normal, la cual no es una curva apropiada para la distribución del monto de reclamos, excepto para algunos casos muy especiales, pero tiene un papel importante como una curva de aproximación para portafolios muy grandes.

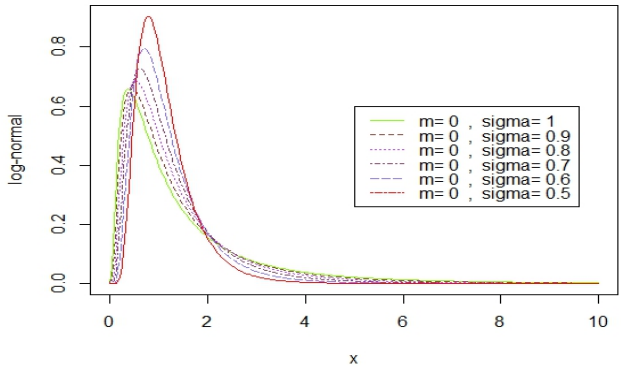
Las distribuciones Lognormal y la Gamma resultan muy útiles para representar la distribución de los montos de reclamo. Como puede observarse en las gráficas 1 y 2, la distribución Lognormal es una buena opción para modelar la severidad; es sesgada por la derecha, y tiene una cola pesada que hace que ajuste bien a varias situaciones. La distribución Gamma es cerrada bajo convolución, (la suma de variables gamma independientes se distribuye también Gamma), es asimétrica hacia la derecha y se aproxima a una normal en el límite cuando α tiende a infinito (véanse gráficas 3 y 4). Es una de las más usadas en el modelado debido a la facilidad matemática con la que se trabajan sus propiedades y, sirve para crear otras distribuciones, pero por sí misma en general no es un modelo útil para modelar la severidad.

Gráfica1. Diversas formas de la distribución Log-Normal (mu variable).



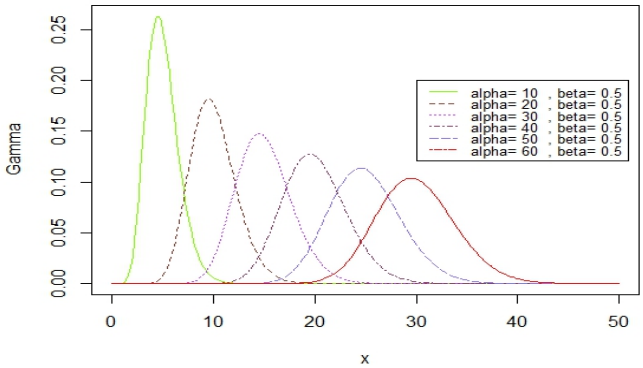
Fuente: elaboración propia

Grafica 2. Diversas formas de la distribución Log-Normal (sigma variable).



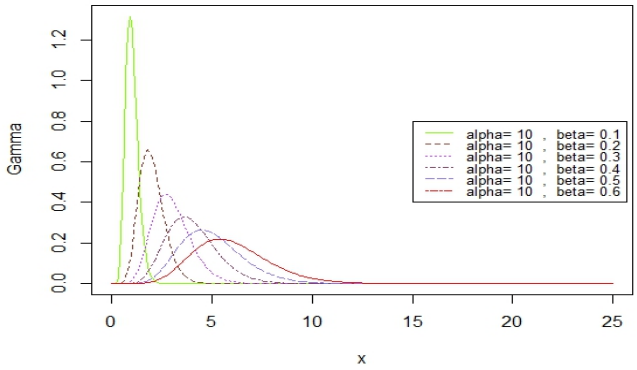
Fuente: elaboración propia

Grafica 3. Diversas formas de la distribución Gamma (alpha variable).



Fuente: elaboración propia

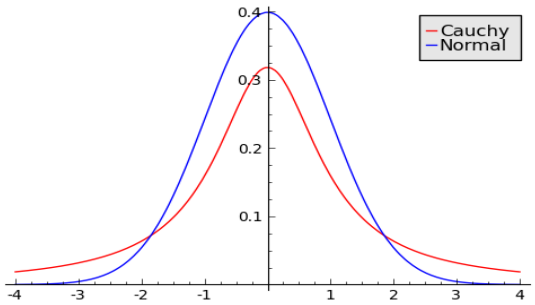
Grafica 4. Diversas formas de la distribución Gamma (beta variable).



Fuente: elaboración propia

La distribución Beta se utiliza para describir distribuciones de grado de daño en seguros de incendio. La distribución de Cauchy que es similar a la curva normal pero converge mucho más despacio a cero en las colas.

Gráfica 5. Comparativo distribución Cauchy vs. Distribución Normal

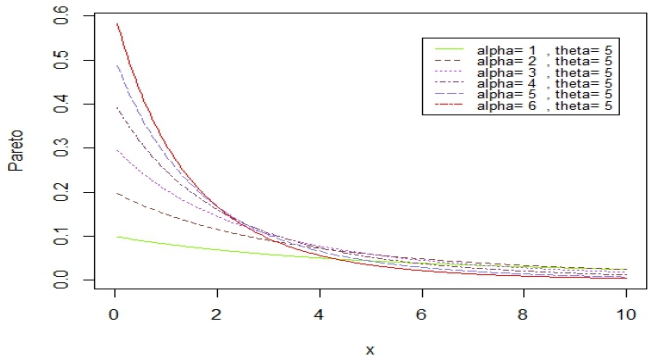


Fuente: Foss et. al (2011)

Como se muestra en las gráficas 6 y 7, la distribución Pareto, también converge lentamente a cero para valores extremos, por lo cual puede ser usada como una distribución del monto de reclamo si se quiere dar énfasis a los valores extremos. Esta distribución tiene una cola pesada, que la hace una buena candidata para modelar el monto de la pérdida, su principal desventaja es que no se puede tratar matemáticamente en algunos casos, al igual que la Lognormal, la transformada de Laplace no tiene una forma cerrada y su función

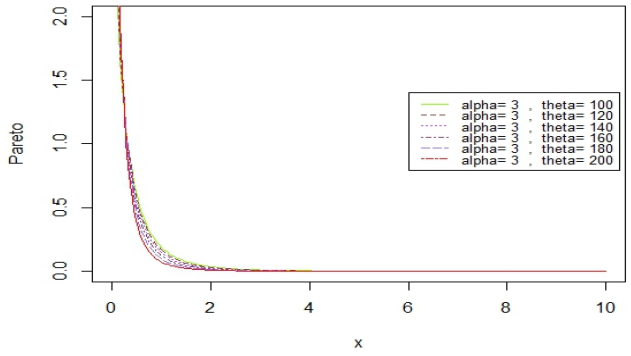
generadora de momentos no existe, además, así como la distribución exponencial, la Pareto es monótona decreciente.

Gráfica 6. Diversas formas de la distribución Pareto (alpha variable).



Fuente: elaboración propia.

Gráfica 7. Diversas formas de la distribución Pareto (theta variable).

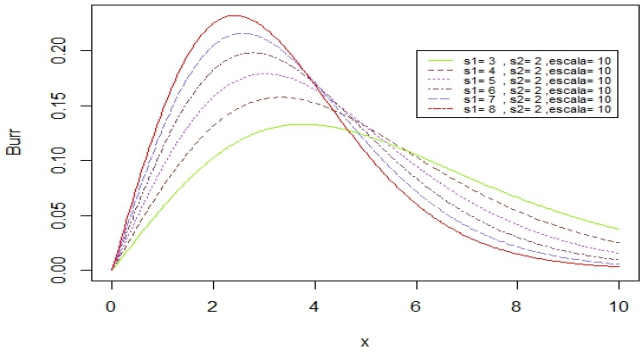


Fuente: elaboración propia.

La distribución exponencial es útil debido a sus propiedades matemáticas, sin embargo, una desventaja es que su densidad es monótona decreciente.

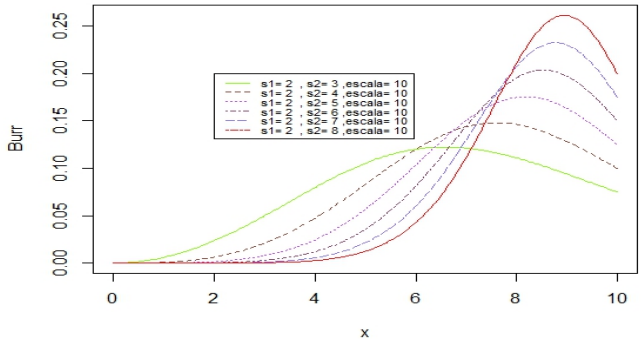
La distribución Burr es de colas pesadas y es más flexible que la Pareto, además de ser una distribución no monótona. Al igual que en la Weibull, el método de momentos y de máxima verosimilitud para encontrar sus parámetros sólo puede evaluarse numéricamente (véanse gráficas 8-11).

Gráfica 8. Diversas formas de la distribución Burr (s1 variable).



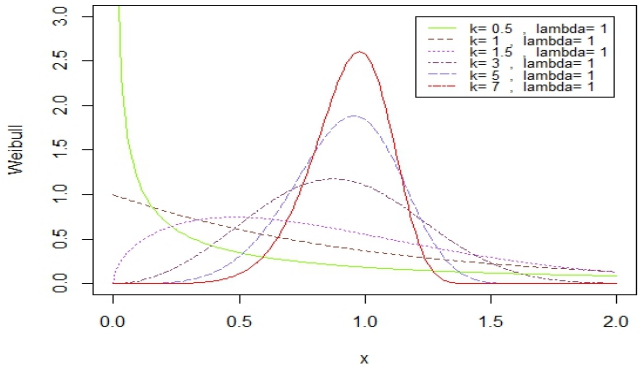
Fuente: elaboración propia.

Gráfica 9. Diversas formas de la distribución Burr (s2 variable).



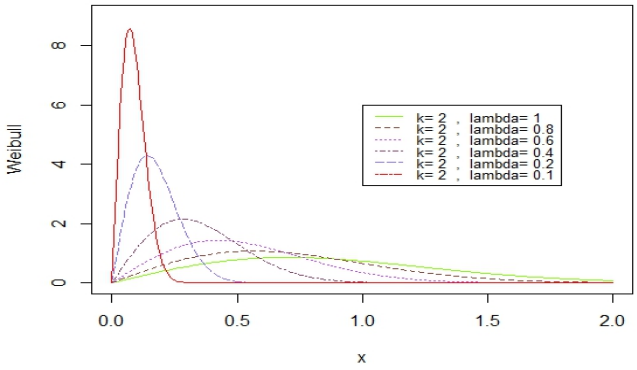
Fuente: elaboración propia.

Gráfica 10. Diversas formas de la distribución Weibull (k variable).



Fuente: elaboración propia.

Gráfica 11. Diversas formas de la distribución Weibull (lambda variable).



Fuente: elaboración propia.

Una vez que se ha seleccionado un modelo y se han calculado sus parámetros, se deben hacer pruebas de bondad de ajuste. Una opción es medir la distancia entre la distribución empírica y la función de distribución ajustada; algunas pruebas basadas en la función de distribución empírica, son la Kolmogorov-Smirnov, la familia Cramer-von Mises y simulaciones de Monte Carlo.

En el caso de tener varias líneas de negocio, cada una con un comportamiento o tipo de riesgo diferente, los datos se dividen por línea de negocio (*i*) y tipo de riesgo (*j*). Así, la distribución de la severidad es denotada por $F_{i,j}$. Asimismo, se asume aleatoriedad en el número de eventos ocurridos durante un periodo de tiempo. Entonces, la variable aleatoria $N(i,j)$ tiene función de probabilidad $p_{i,j}$. La distribución de frecuencia de pérdidas es definida por

$$P_{i,j}(n) = \sum_{k=0}^{n} p_{i,j}(k)$$

De forma que la distribución de pérdida de la línea de negocio *i* y el tipo de riesgo *j* durante un periodo de tiempo determinado es

$$\vartheta(i,j) = \sum_{n=0}^{N(j,i)} \zeta_n(i,j)$$

Donde $G_{i,j}$ es la distribución de $\zeta_n(i,j)$. Entonces $G_{i,j}$ es la distribución compuesta

$$G_{i,j}(x) = \begin{cases} \sum_{n=1}^{\infty} p_{i,j}(n) F_{i,j}^{n*}(x) & x > 0 \\ p_{i,j}(0) & x = 0 \end{cases}$$

Donde F^{n*} es la n-ésima convolución de *F*.

Una vez más, se llega al problema de encontrar una forma analítica para la función de distribución compuesta, lo cual puede solucionarse mediante el método de Monte Carlo, la aproximación por recursión de Panjer y la inversa de la función característica (Frachot, et al., 2001).

En el método de Monte Carlo, se aproxima la distribución $G_{i,j}$ por medio del conjunto $S\langle\vartheta(i,j)\rangle = \{\vartheta_s(i,j), s = 1, \dots, S\}$ de valores simulados de la variable aleatoria $\vartheta(i,j)$, obteniendo un estimado de $G_{i,j}$ mediante la distribución empírica de $S\langle\vartheta(i,j)\rangle$ (Fishman, 1996).

En 1981, Panjer introduce aproximaciones recursivas para calcular convoluciones de orden alto. Si existen las constantes c_1 y c_2 tal que

$$p_{i,j}(n) = \left(c_1 \frac{c_2}{n}\right) p_{i,j}(n-1) \qquad (1)$$

Entonces siguiendo la recursión se sostiene que

$$g_{i,j}(x) = p_{i,j}(1) f_{i,j}(x) + \int_0^x \left(c_1 + c_2 \frac{y}{x}\right) f_{i,j}(y) g_{i,j}(x-y) dy$$

De acuerdo a Sundt y Jewell (1981), las distribuciones de probabilidad que satisfacen (1), son la Poisson, Binomial, Binomial Negativa y las familias geométricas.

Heckman y Meyers (1983) proponen calcular la distribución de la pérdida agregada mediante las propiedades de su función característica de la siguiente forma. Sea *X*, la variable aleatoria con distribución *H*. La función característica se define como

$$\emptyset_H(t) = E\left[e^{itX}\right] = \int_0^\infty e^{itx} dH(x)$$

(La función característica de *m* variables aleatorias independientes es el producto de sus funciones características).

La función característica de $G_{i,j}$ esta dada por

$$\emptyset_{G_{i,j}}(t) = \sum_{n=0}^{\infty} p_{i,j}(n) \left[\emptyset_{F_{i,j}}(t)\right]^n$$

Y finalmente, mediante la transformada de Laplace se tiene que

$$G_{i,j}(x) = \frac{1}{2} - \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{1}{t} e^{-itx} \emptyset_{G_{i,j}}(t) dt$$

Roger (1990) plantea los siguientes pasos para aproximar la distribución de la pérdida agregada mediante el método de Monte Carlo:

1) Seleccionar aleatoriamente el número de reclamos *N* para la frecuencia.

2) Seleccionar aleatoriamente *N* reclamos $X_1, X_2, \ldots, X_n$ de la distribución de la severidad.

3) Calcular una observación de la distribución de *S* mediante la suma $X_1 + X_2 + \cdots + X_n$.

4) Repetir los pasos 1 al 3 “muchas” veces y estimar la distribución de *S* usando los puntos generados.

En los pasos anteriores existe cierta ambigüedad sobre cuántas simulaciones realizar, cómo simular las selecciones aleatorias para lo cual se pueden comparar los resultados de dos conjuntos de simulaciones y si los resultados de ambas son cercanos, entonces la distribución combinada puede ser usada como una aproximación, si los resultados son muy distintos deberán hacerse más simulaciones. Para no perder información, se sugiere asignar una función de distribución para estimar las colas, ya que si no existen límites en sumas aseguradas con los datos empíricos, se perderá información.

2.3 El método bootstrap

Dos de los problemas más importantes dentro de la estadística aplicada, en particular de la estimación puntual, son 1) el determinar un estimador para un parámetro de interés o una función del mismo y 2) la evaluación de la precisión de dicha estadística a través del cálculo del error estándar del estimador. También, otro camino común de la inferencia estadística son los intervalos de confianza.

La teoría de la probabilidad ha demostrado que bajo ciertos supuestos, la media muestral, $\overline{X}$, converge a la media poblacional μ. Uno de los teoremas que respaldan este hecho, es el teorema del límite central, el cual, no solamente

garantiza que $\bar{X}$ se aproxima a μ conforme el tamaño de la muestra crece, sino que también dice de qué forma lo hace; este teorema indica que los errores cometidos al estimar μ por medio del estadístico $\bar{X}$, tienen una distribución normal, con error de estimación igual a $\frac{s}{\sqrt{n}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Es importante calcular dicho error, ya que permite conocer la precisión del estimador; si el error es grande, el valor estimado será poco confiable y, viceversa si el error es pequeño.

La estadística matemática ha logrado estimar el error de estimación de manera analítica, para el caso en el que el parámetro de interés sea la media poblacional; sin embargo, si se desea estimar el error estándar de un estimador $\hat{\theta}$ de algún otro parámetro poblacional distinto a μ, no se tendrá, por lo general, una fórmula analítica que permita conocer el error cometido.

Para utilizar un modelo estadístico que se ajuste a las características de los datos del fenómeno en estudio, es necesario crear restricciones sobre la(s) variable(s) aleatoria(s) analizada, como lo son, por ejemplo, los supuestos de normalidad e independencia entre los datos, de tal forma las técnicas paramétricas incluyen un conjunto de supuestos más restrictivos, provocando que se reduzca el número de casos en el que pueden ser aplicadas; por otro lado, las técnicas no paramétricas tienen menor potencia que las paramétricas, pero se muestran muy útiles cuando no se tiene conocimiento sobre el comportamiento de la población, ya que no es necesario incluir supuestos distribucionales ni de comportamiento a la variable aleatoria (Ledesma, 2008).

A finales de los años 60, se empezaron a desarrollar los métodos de remuestreo (*resampling*) para solucionar problemas en el marco de la teoría de probabilidades y la inferencia estadística y, por otra, la complejidad de los métodos analíticos. El remuestreo está basado en el empleo de simulaciones a través de recursos computacionales (Miranda, 2003).

Dentro de las técnicas de remuestreo el *bootstrap,* es una buena alternativa cuando no se conoce la función de distribución y no puede suponerse una distribución paramétrica, ya que no requiere una hipótesis sobre la distribución de los datos; el mismo método genera una distribución empírica (F_n) del estimador.

El bootstrap es una de las formas, dentro de un gran número de métodos, que vuelven a tomar muestras del conjunto de datos original, por ello el nombre de técnicas de remuestreo. Algunas técnicas de remuestreo datan de mucho tiempo atrás, como el *jackknife* desarrollado por Quenouille y Tukey y los métodos de permutación de Fisher y Pitman alrededor de 1930. Sin embargo, el uso de computadoras para generar simulaciones es más reciente.

Quenoiulle (1949), desarrolló un método para estimar el sesgo de un estimador borrando un dato del conjunto de datos original y, recalculando el estimador basado en el resto de los datos. Sea $T_n=T_n(X_1,\ldots,X_n)$ un estimador del parámetro desconocido θ. Se define el sesgo de T_n como:

$sesgo(T_n) = E(T_n) - \theta$.

Sea $T_{n-1,i}=T_{n-1}(X_1,\ldots,X_{i-1}, X_{i+1},\ldots,X_n)$ el estadístico dado pero ahora basado en n-1 observaciones, y eliminando la *i*-ésima, para *i*=1,…,n. Luego, el estimador jackknife del sesgo de Quenoille es

$b_{jack} = (n-1)(\overline{T_n} - T_n)$, donde

$$\overline{T_n} = \frac{1}{n}\sum_{i=1}^{n} T_{n-1,i}$$

Basado en la técnica de Quenouille para la estimación del sesgo, en 1958, John Tukey descubrió que el jackknife también podía ser usado para construir estimadores de varianza y revolucionó la estimación del error estándar con su método llamado jackknife. Una breve definición del método se expone a continuación.

El error estándar de $\hat{\theta}$, la raíz cuadrada de su varianza,

$$se\{\hat{\theta}; F\} = [var_F\{t(x)\}]^{1/2}, \quad (1)$$

es la medida más común de precisión para los estimadores insesgados $\hat{\theta}$.

Sea $\sigma^2(F)$ la varianza de F,
$\sigma^2(F) = \int_{-\infty}^{\infty}(x - \mu(F))^2 dF$

Una fórmula que relaciona el error estándar $se\{\bar{X}; F\}$ con $\sigma^2(F)$ es,

$$se\{\bar{X}; F\} = [\sigma^2(F)/n]^{1/2} \quad (2)$$

Esto puede parecer no tener sentido ya que $\sigma^2(F)$ es una función de la función de distribución desconocida *F*, sin embargo, existe un estimador insesgado de $\sigma^2(F)$,

$$\sigma^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (3)$$

Y sustituyendo 3 en 2 se tiene un estimado del error estándar para x barra,

$$se\{\bar{X}; F\} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (4)$$

El método de Tukey se basó en una generalización de la fórmula 4 de la siguiente manera. Suponga que un conjunto de datos *x* consta de *n* observaciones x_i independientes e idénticamente distribuidas de una distribución desconocida *F*. Sea $x_{(i)}$ el conjunto de datos con el *i*-ésimo dato eliminado de dicho conjunto. Y sea $t(x_{(i)})$, el estadístico re-evaluado para el conjunto de datos con el punto eliminado. El estimador jackknife del error estándar es

$$se\{\hat{\theta}\} = \left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \bar{\theta}_{(.)}\right)^2\right]^{1/2}$$

Donde $\bar{\theta}_{(.)} = \sum_{i=1}^{n}\hat{\theta}_{(i)}$.

El jackknife de Tukey produce automáticamente un estimado de su error estándar, sin importar la complejidad del mismo. Todo lo que se requiere es recalcular $\hat{\theta}$ *n* veces, una para cada punto eliminado del vector x_i. Este método marcó un cambio decisivo hacia las herramientas computacionales alejándose de la estadística teórica tradicional. El jackknife resultó poco funcional en estimadores como la mediana muestral, pero bastante confiable en cualquier otro resultado.

En 1979, Bradley Efron desarrolla el análisis formal del bootstrap para profundizar en el método jackknife. Esto implicó reexaminar la fórmula (2), la cual había sido evitada al usar directamente la (4). Si en (2) la distribución empírica $\hat{F}$ de los datos observados se sustituye por la función de distribución real desconocida F. Ya que $\sigma^2(\hat{F})=\overline{(x-\bar{x})^2}$, esto da el estimador del error estándar

$se\{\bar{x}, F\}=[\overline{(x-\bar{x})^2}/n]^{1/2}$,

casi igual al estimador tradicional (4).

Los puntos principales que menciona Efron (1979) en su trabajo sobre el bootstrap son los siguientes:

1. Sustituir *F* por $\hat{F}$ en la fórmula (1) da un estimador razonable del error estándar de cualquier estimador, es decir,

$se_{boot}\{\hat{\theta}\} \equiv se\{\hat{\theta}; \hat{F}\}=[var_{\hat{F}}\{t(x*)\}]^{1/2}$.

Con $x*$ indicando un vector hipotético de datos generados por una muestra independiente e idénticamente distribuida de la distribución de $\hat{F}$, distinta del vector *x* de datos observados.

2. Hay un algoritmo simple en computadora para estimar $se_{boot}\{\hat{\theta}\}$.

3. se_{boot} concuerda asintóticamente con se_{jack} y de hecho, el jackknife es una aproximación lineal a un proceso bootstrap más intenso computacionalmente.

4. El bootstrap es un estimador razonable del error, más fácil de entender que el jackknife y más fácil de extenderlo a otras estructuras.

5. El bootstrap puede ser aplicado a problemas de valuación del error estadístico más allá del sesgo y el error estándar, en particular, el establecer intervalos de confianza.

El nombre bootstrap se deriva de la expresión inglesa *to pull oneself up by one's bootstrap*, basada en la obra de Rudolph Erich Raspe, *The Surprising Adventures of Baron Munchausen*, donde el Barón se calló al fondo de un lago y, justo cuando pensó que todo estaba perdido, se le ocurrió jalarse de las agujetas de sus botas para salir de las profundidades del lago; esto puede interpretarse como salir adelante con sus propios recursos.

La teoría estadística se propone resolver tres preguntas básicas:

1. ¿Cómo se deben recolectar los datos?
2. ¿Cómo deben ser analizados?
3. ¿Qué tan precisos son los resultados?

La última pregunta es parte de la inferencia estadística. El bootstrap es una técnica para implementar cierto tipo de inferencias estadísticas. Las ideas básicas de la estadística no cambian, pero sí lo hace la forma de implementarlas gracias al desarrollo tecnológico, donde los recursos computacionales son cada vez más rápidos y más baratos.

Para darse una idea de cómo funciona el bootstrap sin definiciones formales (las cuales se verán más adelante) una descripción simple es, se tiene una muestra de tamaño *n* y se desea estimar un parámetro, determinar el error estándar, un intervalo de confianza para el parámetro o probar una hipótesis sobre dicho parámetro. Se toma la distribución empírica de la muestra que es la distribución de probabilidad, donde a cada elemento se le asigna una probabilidad de 1/*n*. La idea de este método es reemplazar la distribución poblacional desconocida por la distribución empírica conocida. Las propiedades del estimador, como el error estándar son determinadas con base en la distribución empírica. Algunas veces, estas propiedades pueden ser determinadas de forma analítica, pero generalmente son calculadas por métodos de aproximación como Monte Carlo (Chernick, 1999).

Un ejemplo de cómo funciona el bootstrap es el siguiente: se crean dos poblaciones, la primera consta de 319 unos y 11,337 - 319 = 10,018 ceros (muestra bootstrap #1), y la otra población con 698 unos y 11,334 - 698=10,636 ceros (muestra bootstrap #2). Tomamos una muestra con reemplazo de 11,337 objetos de la primera población y otra muestra con reemplazo de 11,334 objetos

de la segunda población. Cada una de éstas se llama muestra bootstrap. Así derivamos una réplica bootstrap de θ:

$\hat{\theta}*$= proporción de unos en la muestra bootstrap # de unos/proporción de unos en la muestra bootstrap #2 (Efron y Tibshirani, 1993).

De esta forma, repetimos el proceso arriba descrito un gran número de veces y obtendremos, digamos, 1,000 réplicas de $\hat{\theta}*$, a partir de éstas, podemos obtener características numéricas de interés sobre el estimador.

Chernick (1999) proporciona una definición más formal del bootstrap de la siguiente manera:

Dada una muestra de *n* vectores aleatorios $X_1, X_2, \ldots X_n$, independientes e idénticamente distribuidos y un estimador $\theta(X1, X2, \ldots, Xn)$ denotado como $\hat{\theta}$ del parámetro θ de la distribución, un procedimiento (bootstrap) para evaluar la precisión de $\hat{\theta}$ es definido en términos de la función de distribución empírica F_n. Esta función de distribución empírica asigna una probabilidad de 1/*n* a cada valor observado de los vectores aleatorios X_i para $i = 1, 2, \ldots, n$.

La distribución empírica es el estimador de máxima verosimilitud de la distribución de las observaciones, cuando no se hacen supuestos paramétricos. La distribución bootstrap para $\hat{\theta} - \theta$ es la distribución obtenida al generar valores de $\hat{\theta}$ mediante el remuestreo independientemente y con reemplazo de la distribución empírica F_n. El estimador bootstrap del error estándar de $\hat{\theta}$ es entonces la desviación estándar de la distribución bootstrap de $\hat{\theta} - \theta$.

La aplicación parcial de dicha técnica requiere de la generación de muestras o remuestras (muestras obtenidas de muestrear independientemente con remplazo de la distribución empírica). Del muestreo bootstrap se obtiene una aproximación del estimador mediante la simulación de Monte Carlo como sigue:

1. Generar una muestra de tamaño *n* (donde *n* es el tamaño de la muestra original) con reemplazo de la distribución empírica (una muestra bootstrap).

2. Calcular θ^*, el valor de $\hat{\theta}$ se obtiene usando la muestra bootstrap en lugar de la muestra original.

3. Repetir los pasos 1 y 2 *k* veces.

Para estimar el error estándar se recomienda que *k* sea al menos 100. Al ejecutar el paso 3, se obtiene una aproximación Monte Carlo de la distribución de θ^*. La desviación estándar de la aproximación de Monte Carlo de la distribución de θ^* es la aproximación Monte Carlo del estimador bootstrap del error estándar de $\hat{\theta}$. Cuando *k* es muy grande, la diferencia entre el estimador bootstrap y la aproximación Monte Carlo es muy pequeña.

El algoritmo del bootstrap comienza generando un gran número de muestras bootstrap independientes, $X^{*1}, X^{*2},\ldots, X^{*k}$, cada una de tamaño *n*. A cada muestra bootstrap le corresponde una réplica bootstrap de θ^{*i}, que es el valor del estadístico θ evaluado para $\hat{\theta}^{*i}$.

El estimador bootstrap del error estándar es la desviación estándar de las réplicas bootstrap.

El error estándar es una de las medidas estadísticas de precisión más simples, pero los métodos bootstrap también sirven para evaluar medidas de precisión más complejas como el sesgo, errores de predicción e intervalos de confianza.

Lo que realmente se quiere saber es la distribución de $\hat{\theta} - \theta$ y lo que se tiene con el método arriba descrito es la aproximación Monte Carlo de la distribución de $\theta * - \hat{\theta}$. La base del bootstrap es que para un *n* suficientemente grande se

espera que ambas distribuciones sean casi iguales. A la idea de que la distribución de que $\theta * - \hat{\theta}$ se comporte casi igual que la distribución de $\hat{\theta} - \theta$ se le llama principio bootstrap.

El bootstrap es un método que usa intensivamente recursos computacionales, ya que en la mayoría de los problemas prácticos donde es considerado útil, la estimación es compleja y se requieren muestras bootstrap (remuestreo).

Un punto importante a enfatizar es que una muestra bootstrap típica difiere de la original debido a que algunas observaciones son repetidas una, dos o más veces, así como algunas observaciones de la muestra original pueden no aparecer.

La idea principal del bootstrap es que la variabilidad de $\theta *$ (basada en F_n) alrededor de $\hat{\theta}$ será similar o idéntica a la variabilidad de $\hat{\theta}$ (basada en la distribución poblacional real *F*) alrededor de σ (principio bootstrap). Esta idea esta sustentada en que, para muestras grandes, a medida que *n* incrementa F_n se acerca más a *F*, entonces si se toman muestras con reemplazo de F_n es casi como tomar muestras aleatorias de *F*.

La ley fuerte de los grandes números para variables aleatorias independientes e idénticamente distribuidas implica que, con probabilidad 1, F_n converge a *F* punto a punto.

En resumen, el bootstrap consiste en extraer un determinado número, *k*, de muestras mediante muestreo aleatorio con reemplazo de la muestra inicial. Sobre cada una de estas muestras se aplica el estimador de interés ($\hat{\theta}$), así, se obtienen *k* estimaciones del parámetro θ y de ésta manera, se puede estimar el error estándar del estimador de interés.

2.4 Estimación Bayesiana

El enfoque Bayesiano, está basado en fundamentos axiomáticos que proveen una estructura lógica y, garantizan consistencia en los métodos propuestos para la inferencia estadística y la toma de decisiones bajo condiciones de incertidumbre. Estos métodos se derivan de un sistema de axiomas, por lo cual proveen una metodología coherente y general. Los métodos bayesianos hacen posible incorporar hipótesis científicas en el análisis (por medio de una distribución a priori) y pueden ser aplicados a problemas cuya estructura es muy compleja de manejar con los métodos convencionales (Bernardo, 2003).

El paradigma Bayesiano consiste en interpretar el concepto de probabilidad como una medida condicional y racional de la incertidumbre, lo cual, se asemeja al sentido que se le da a la palabra en el lenguaje común.

Los métodos bayesianos cubren la necesidad matemática de describir, por medios de distribuciones de probabilidad, todas las incertidumbres en un problema. En particular, los parámetros desconocidos de los modelos probabilísticos, deben tener una distribución de probabilidad conjunta, que describe la información disponible sobre sus valores. Es importante notar que dentro del paradigma Bayesiano, los parámetros son tratados como variables aleatorias. Esto no significa que se describe su variabilidad sino la incertidumbre sobre sus verdaderos valores (Bernardo, 2003).

En ocasiones, la probabilidad de un evento *A* aleatorio puede verse afectada por la ocurrencia de otro evento *B*. Cuando el evento *B* está dado, el espacio muestral se restringe a los elementos que conforman tal evento. Partiendo de esto, se llega al concepto de probabilidad condicional definido de la siguiente manera:

Sean *A, B* dos eventos tales que *P(B)>0*. Se define la probabilidad condicional del evento *A* dado el evento *B*, como

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Otro concepto importante es el de probabilidad total. Supongamos que la colección de eventos $\{B_i\}_{i=1}^{n} \in F$ forma una partición del espacio muestral Ω, es decir, $B_j \cap B_j = \emptyset \; para \; j \neq k \; y \; \cup_{j=1}^{n} B_j = \Omega$. Sea $A \in F$ un evento con probabilidad positiva. Si se conocen $P(A|B_j)$ y $P(B_j)$, para toda $j = 1,2,\dots,n$, entonces,

$$P(A) = \sum_{j=1}^{n} P(A|B_j)P(B_j)$$

Así, utilizando los mismos supuestos del teorema de la probabilidad total, surge el teorema de Bayes:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}$$

Lo anterior puede ser generalizado a variables y vectores aleatorios de la siguiente manera. Si (X,Y) es un vector aleatorio continuo, la función de densidad de probabilidad conjunta del vector aleatorio es una función $f_{X,Y}(x,y)$ que satisface:

$f_{X,Y}(x,y) \geq 0$, para todo $(x,y) \in R^2$

y

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$$

Las funciones de densidad marginales de las variables aleatorias X y Y, se pueden obtener como:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Y la función de densidad condicional de X dado $Y = y$ se define como:

$f_{X|Y=y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$, siempre que $f_Y(y) > 0$

Así, la función de probabilidad conjunta

$$f_{X,Y}(x,y) = f_{X|Y=y}(x|y)f_Y(y) \qquad o \qquad f_{X,Y}(x,y) = f_{Y|X=xy}(y|x)f_X(x)$$

Y por lo tanto, la función de densidad marginal de Y se calcula como:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X=x}(y|x)f_X(x)dx$$

Por último, el teorema de Bayes para variables aleatorias continuas está representado por:

$$f_{X|Y=y}(x|y) = \frac{f_{Y|X=xy}(y|x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X=x}(y|x)f_X(x)dx}$$

Esta versión es muy utilizada en los métodos Bayesianos. Bajo este enfoque, a la función de densidad $f_X(x)$ se le conoce como distribución *a priori*, y a

$f_{X|Y=y}(x|y)$ calculada como se describió anteriormente, se le conoce como distribución *a posteriori*.

Es de interés estimar la proporción θ de individuos o elementos que tienen una característica en común. Al analizar una muestra de *n* elementos, de los cuales *r* poseen determinada propiedad, se pretende usar los resultados de la muestra para establecer regiones de [0,1] donde es aceptable esperar que el valor desconocido de θ que arroja la muestra, no sea el verdadero valor de θ. Esta información está dada por probabilidades de la forma $\Pr(a < \theta < b|r, n, A, K)$, que son una medida condicional de la incertidumbre sobre el evento θ, que pertenece a (a, b) dada la información proporcionada por los datos (r, n), el supuesto A sobre el comportamiento del mecanismo que ha generado los datos y cualquier conocimiento relevante K sobre los posibles valores de θ (Bernardo, 2003).

Se estudiará el caso donde se observa un vector $X = (X_1, X_2, \dots, X_n)$ con densidad discreta o continua en la familia $f(x, \theta)$, con $\theta \in \Theta \subset \mathrm{R}$. El enfoque bayesiano, supone que se tiene algún conocimiento previo sobre θ. Dicha información está expresada por medio de una distribución inicial sobre θ denominada distribución *a priori*, por lo tanto, supondremos que la distribución *a priori* tiene una densidad $f_\Theta(\theta)$.

Una vez observada la muestra aleatoria $X = (X_1, X_2, \dots, X_n)$ es de interés conocer la distribución condicional de Θ dado $X = (X_1, X_2, \dots, X_n)$, $f(x, \theta)$. Esta distribución *a posteriori* está dada por:

$$f(x, \theta) = \frac{f(x|\theta) f_\Theta(\theta)}{\int f(x|\theta) f_\Theta(\theta) d\theta} \dots\dots\dots I$$

Para modelar la frecuencia de los reclamos ocurridos, es de interés hacer inferencias sobre el parámetro de la distribución Bernoulli. Sea

$X = (X_1, X_2, \dots, X_n)$ una muestra independiente de una distribución $Blli(\theta)$ y supongamos que la distribución a priori de θ es una distribución $Beta(a, b)$, es decir con una densidad

$$f_\Theta(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\mathbb{I}_{[0,1]}(\theta)$$

Se demuestra que la esperanza y la varianza de esta distribución son

$$E(X) = \frac{a}{a+b} \qquad y \quad Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

La varianza se puede reescribir como

$$Var(X) = \frac{E(X)(1-E(X))}{(a+b+1)}$$

Luego, si se conocen la media y la varianza de la distribución *a priori* de Θ, se pueden determinar a y b. Además, la segunda forma de escribir la varianza, muestra que, para un valor dado de la esperanza, la varianza depende de $a+b$, tendiendo a 0 cuando $a+b \to \infty$.

Ahora, la distribución de la muestra $X_1, X_2, \dots, X_n$ dado el valor de θ esta dada por

$$\begin{aligned} f_{X_1,X_2,\dots,X_n|\theta}(x_1, x_2, \dots, x_n|\theta) &= f(x, \theta) \\ &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \end{aligned}$$

Usando (I) se tiene

$$f(x,\theta) = \frac{f(x|\theta) f_{\Theta}(\theta)}{\int f(x|\theta) f_{\Theta}(\theta) d\theta}$$

$$= \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{\sum_{i=1}^{n} x_i + a - 1} (1-\theta)^{n - \sum_{i=1}^{n} x_i + b - 1}}{\int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{\sum_{i=1}^{n} x_i + a - 1} (1-\theta)^{n - \sum_{i=1}^{n} x_i + b - 1} d\theta}$$

$$= \frac{\theta^{\sum_{i=1}^{n} x_i + a - 1} (1-\theta)^{n - \sum_{i=1}^{n} x_i + b - 1}}{\int_0^1 \theta^{\sum_{i=1}^{n} x_i + a - 1} (1-\theta)^{n - \sum_{i=1}^{n} x_i + b - 1} d\theta}$$

El denominador de la función de arriba es una constante, por lo que la distribución a posteriori puede expresarse como:

$$f(x,\theta) = k\theta^{\sum_{i=1}^{n} x_i + a - 1} (1-\theta)^{n - \sum_{i=1}^{n} x_i + b - 1}$$

Que corresponde a la forma de una densidad $Beta(\sum_{i=1}^{n} x_i + a, n - \sum_{i=1}^{n} x_i + b)$. La constante k será el factor que depende de las $x_1, x_2, \ldots, x_n$ y que logra que la función a posteriori sea nuevamente una función de densidad de probabilidades.

Con esto se concluye que

$$\Theta|\mathrm{X} = (X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \sim Beta\left(\sum_{i=1}^{n} x_i + a, n - \sum_{i=1}^{n} x_i + b\right)$$

Lo anterior nos servirá para estimar la proporción de reclamaciones por cáncer de mama.

III. APLICACIÓN DEL IRM, CASO DE UNA EMPRESA ASEGURADORA.

Una institución aseguradora proporcionó su base de datos de pagos de siniestros del ramo de Gastos Médicos Individual de los años 2008, 2009, 2010, 2011 y 2012. Cada registro contiene:

1. Monto de reclamo: es el monto que la aseguradora pagó en determinado momento derivado de alguna reclamación.
2. Clave de siniestro: varias reclamaciones pueden provenir de un mismo siniestro, a cada siniestro se le asigna una clave única para identificar las reclamaciones que se deriven de éste.
3. Fecha de pago: Es la fecha en que la aseguradora paga al asegurado el monto de reclamo.
4. Padecimiento: es la enfermedad por la cual la aseguradora paga al asegurado el monto de reclamo.

Se eligió estimar la pérdida agregada por cáncer de mama ya que es un padecimiento representativo de la institución aseguradora por el monto reclamado, como se muestra en la siguiente tabla.

Tabla 1. Los 10 padecimientos más importantes de 2012.

Padecimiento		Monto	# Reclamaciones	# Siniestros
Otros trastornos de los discos intervertebrales	$	52,013,013	2,423	548
Trastorno interno de la rodilla	$	42,290,867	3,080	786
Tumor maligno de la mama	$	41,448,216	2,231	276
Examen y prueba del embarazo	$	35,884,542	4,260	1,359
Otros traumatismos que afectan múltiples regiones del cuerpo, n	$	27,020,923	1,819	421
Enfermedad isquémica crónica del corazón	$	26,379,802	1,813	363
Trastornos relacionados con duración corta de la gestación y con	$	22,020,252	672	119
Cálculo del riñón y del uréter	$	21,161,288	1,681	396
Enfermedad diverticular del intestino	$	20,633,306	1,078	250
Esclerosis múltiple	$	20,144,351	795	92

Fuente: Elaboración propia con datos reales.

Las bases de datos proporcionadas por la aseguradora vienen de los registros que se ingresan al sistema de la institución, por lo cual hay algunos registros que

no deben ser considerados en la base de datos. Las bases fueron depuradas con ayuda de los programas computacionales Excel y Visual Basic de la siguiente manera:

1. Se seleccionaron los registros únicamente de cáncer de mama para formar la base de datos con la que se trabajará.
2. Se eliminaron los registros con monto de reclamo igual a 0.
3. Eliminar los registros negativos, así como su inverso positivo (los registros negativos son cancelaciones de otro registro que se ingresó al sistema por error).
4. Se separaron los registros por fecha de pago.
5. Se agruparon los montos de reclamo por clave de siniestro en cada año. Si un siniestro tuvo varios montos de reclamo efectuados en diferentes años, aparecerá una vez en cada año, con el monto total pagado del año al que corresponda.

Una vez que se hizo lo anterior, esa base de datos se usó para ajustar un modelo a la severidad de cada año (2008 al 2012).

Asimismo, la compañía compartió el número promedio de personas aseguradas en el ramo de Gastos Médicos Individual separados por género para los años 2008, 2009, 2010, 2011 y 2012.

Con el número promedio de asegurados y el número de siniestros pagados por año, se obtuvo el parámetro *p* de una binomial.

3.1 Modelación de la severidad

Dentro de la gama de opciones posibles para modelar la severidad mediante una función de probabilidad, las distribuciones Log-Logistic de tres parámetros, Pearson, Log-Logistic, Lognormal y Burr, entre otras, mostraron buen ajuste a los datos reales proporcionados por la institución aseguradora para todos los años de observación y acorde con las pruebas de bondad de ajuste Anderson Darling y Kolmogorov Smirnov. Los resultados de dichas pruebas se encuentran en el Anexo I.

De las opciones mencionadas se eligió la distribución Lognormal, ya que muestra un buen ajuste en todos los años de estudio y, es uno de los modelos más conocidos y fácil de manejar. Para respaldar la validez del modelo se revisó que la hipótesis nula H_0: la severidad sigue una distribución Lognormal, no fuera rechazada para ningún año. Los resultados de las pruebas de hipótesis Anderson-Darling, Kolmogorov-Smirnov y Chi cuadrada mostraron que con el 99% de confianza, H_0 no se rechaza en ninguno de los casos. Los resultados de las pruebas de hipótesis para la distribución Lognormal se encuentra en el Anexo II.

La Lognormal con parámetros $(\mu, \sigma^2), -\infty < \mu < \infty, 0 < \sigma^2 < \infty$, resulta de una transformación de la distribución Normal donde la variable aleatoria $X = e^Y$ cuando $Y \sim N(\mu, \sigma^2)$ y X tiene una función de densidad de probabilidad

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

para x>0.

Sus principales propiedades son:

Función de distribución:

$$F_X(x) = \Phi\left(\frac{\ln(x) - \mu}{\sigma}\right)$$

Esperanza:

$$E[X] = \exp\,(\mu + \sigma^2/2)$$

Varianza:

$$Var(X) = \exp\,(2\mu + \sigma^2)[exp(\sigma^2) - 1]$$

K-esimo momento:

$$E[X^k] = \exp\,(\mu k + \frac{\sigma^2 k^2}{2})$$

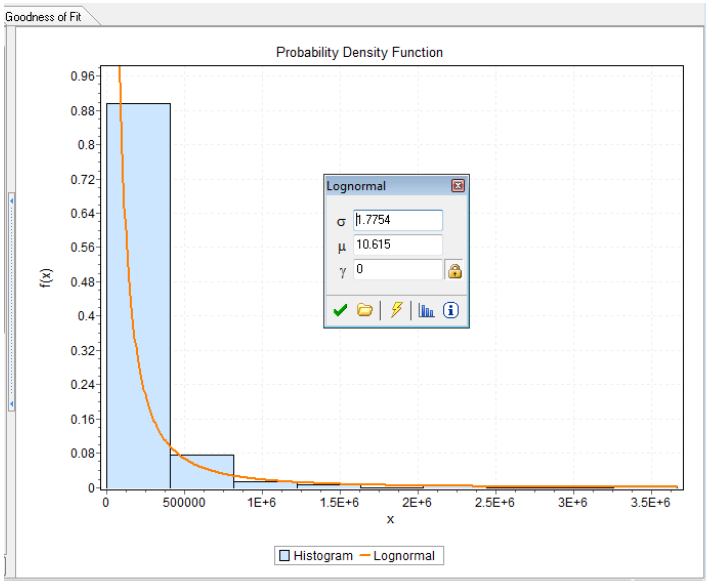
Esperanza condicional

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

No tiene función generadora de momentos.

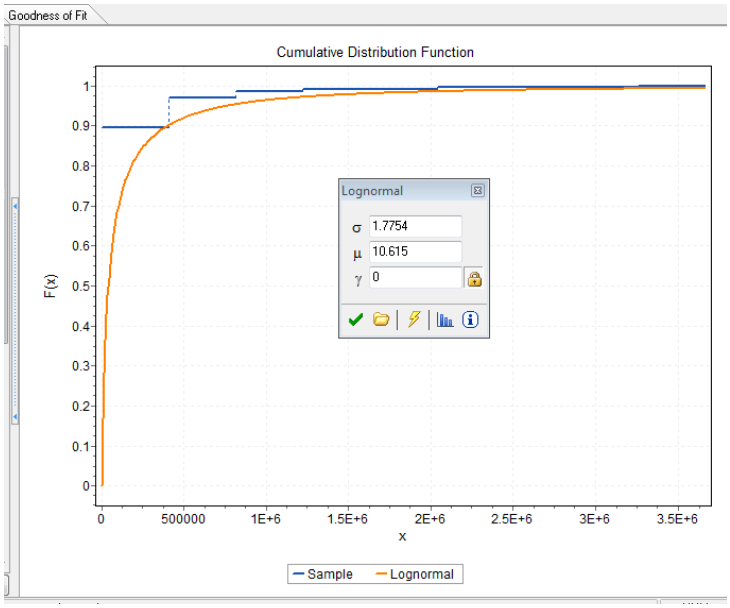
A continuación se muestran las gráficas de las pruebas de bondad de ajuste del modelo Lognormal obtenidos con el programa computacional “Easy Fit” para el año 2012. Se muestra la función de densidad empírica vs teórica, función de distribución empírica vs teórica y su P-P Plot con los estimadores de los parámetros. Los resultados de los años 2008-2011 se encuentran en el Anexo III.

Gráfica 12. Función de densidad Lognormal vs función de densidad empírica para la severidad de 2012.



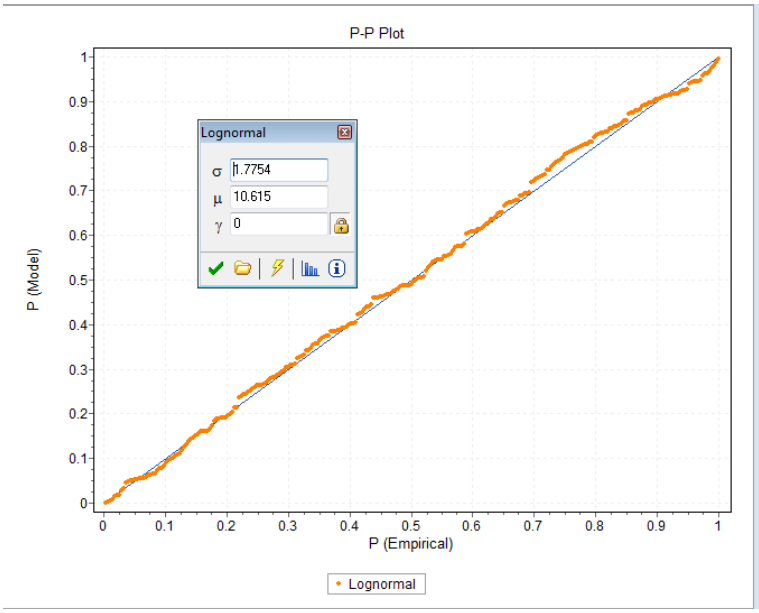
Fuente: elaboración propia.

Gráfica 13. Función de distribución Lognormal vs función de distribución empírica para la severidad de 2012.



Fuente: elaboración propia.

Gráfica 14. PP Plot para la severidad de 2012.



Fuente: elaboración propia.

En las gráficas 12-14, se observa que la función de distribución Log-Normal tiene un buen ajuste a los datos empíricos presentados en el histograma de frecuencias, en el histograma de frecuencias acumuladas y en los gráficos que comparan la probabilidad teórica contra la probabilidad empírica.

Se hicieron pruebas calculando la media y la varianza de la distribución Lognormal para cada año con los estimadores de máxima verosimilitud que arroja el programa “Easy Fit” y los estimadores por momentos calculados de la siguiente forma:

$$\mu = ln(\bar{X}) - \sigma^2/2 \qquad \text{y} \qquad \sigma = \sqrt{\left(\frac{\sum_{i=1}^{n} x^2}{n}\right) - 2 * ln(\bar{X})}$$

Donde:

$\bar{X}$ = Media

$\frac{\sum_{i=1}^{n} x^2}{n}$ = Promedio de cuadrados

n = 123, 150, 163, 177, 276.

Tabla 2. Fn. De distribución empírica, montos de reclamo por cáncer de mama.

Año	n	Media	Promedio de cuadrados
2008	123	90,418	34,893,778,996
2009	150	179,599	130,781,850,562
2010	163	223,405	679,026,555,322
2011	177	173,335	141,195,590,579
2012	276	151,180	130,546,780,323

Fuente: elaboración propia.

De acuerdo a los resultados obtenidos mostrados en las tabla 3 y 4, se decidió usar los estimadores por momentos μ y σ, ya que las varianzas teóricas obtenidas con los estimadores de momentos están más cercan del estimador empírico, que el valor del estimador obtenido mediante el método de máxima verosimilitud.

Tabla 3. Comparativo estimadores empíricos vs teóricos (media).

Año	$Media_{Emp}$	$Media_{MV}$	$Media_{Mom}$	Emp - MV	Emp - Mom
2008	90,418	101,636	90,418	- 11,218	-
2009	179,599	236,069	179,599	- 56,471	-
2010	223,405	268,537	223,405	- 45,132	-
2011	173,335	215,378	173,335	- 42,043	-
2012	151,180	197,013	151,180	- 45,832	-

Emp: Empírica
MV: Máxima Verosimilitud
Mom: Momentos

Fuente: elaboración propia.

Tabla 4. Comparativo estimadores empíricos vs teóricos (varianza).

Año	$Varianza_{Emp}$	$Varianza_{MV}$	$Varianza_{Mom}$	Emp - MV	Emp - Mom
2008	26,937,392,352	15,999,644,052	26,718,389,163	10,937,748,301	219,003,190
2009	99,187,477,606	73,754,336,108	98,526,227,756	25,433,141,498	661,249,851
2010	633,000,271,074	84,216,019,037	629,116,833,828	548,784,252,038	3,883,437,246
2011	111,781,963,440	59,453,513,766	111,150,426,924	52,328,449,674	631,536,517
2012	108,082,874,313	48,033,964,268	107,691,269,695	60,048,910,045	391,604,617

Emp: Empírica
MV: Máxima Verosimilitud
Mom: Momentos

Fuente: elaboración propia.

3.2 Modelación de la frecuencia

Para modelar la frecuencia se decidió usar una función Binomial, utilizando como parámetro *p,* el número de siniestros de cáncer de mama entre el número promedio total de pólizas en vigor, que cubren dicho padecimiento del año correspondiente (al elegir esta función se está suponiendo que sólo puede haber un siniestro al año de este padecimiento por póliza).

La distribución Binomial sirve para modelar el número de éxitos, *I*, ocurridos en *n* ensayos. Su función de masa de probabilidad está dada por

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i} \qquad con\ \ i = 0,1,\dots,n \quad y \quad 0 \leq p \leq 1$$

Sus principales propiedades son:

Función de distribución:

$$P_I(I \leq i) = \sum_{k=0}^{i} \binom{n}{k} p^k (1-p)^{n-k}$$

Esperanza:

$$E[I] = np$$

Varianza:

$$Var(I) = np(1-p)$$

Para modelar la frecuencia, *n* será el número de pólizas en vigor del año correspondiente y *p* será la probabilidad de que una póliza tenga al menos un reclamo por cáncer de mama. Así, *I* es el número de pólizas que tuvieron al menos un reclamo durante el año.

3.3 Estimación de la proporción de pólizas con reclamos por cáncer de mama.

Como se cuenta con poca información sobre la proporción de reclamos por cáncer de mama, se utilizó el enfoque Bayesiano para estimar dicha proporción, θ, utilizada para modelar la frecuencia y, así poder hacer inferencias sobre el valor del parámetro *p,* para estimar la pérdida esperada del año 2013 por cáncer de mama. Así, utilizando el modelo Beta-Binomial

$$\Theta|\mathrm{X} = (X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \sim Beta\left(\sum_{i=1}^{n} x_i + a, n - \sum_{i=1}^{n} x_i + b\right),$$

Se obtuvieron los siguientes estimadores de *p:*

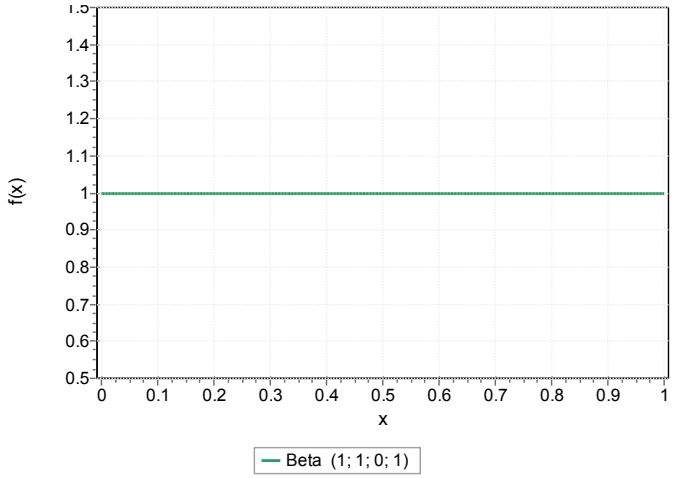
Tabla 5. Estimador Bayesiano p (modelo Beta-Binomial).

Estimador	A priori	2008	2009	2010	2011	2012
α	1	124	274	437	614	890
β	1	34884	72352	112861	158258	206938
Percentil 97.5	0.975	0.00419	0.00423	0.00424	0.00418	0.00457
Percentil 2.5	0.025	0.00295	0.00334	0.00351	0.00357	0.00401
E[p|x]	0.5	0.003542	0.0037728	0.0038571	0.0038647	0.0042824

Fuente: elaboración propia.

La gráfica 15, que se muestra a continuación, representa la función de distribución de Θ, a inicio, α=1 y β=1 suponiendo completo desconocimiento sobre el parámetro.

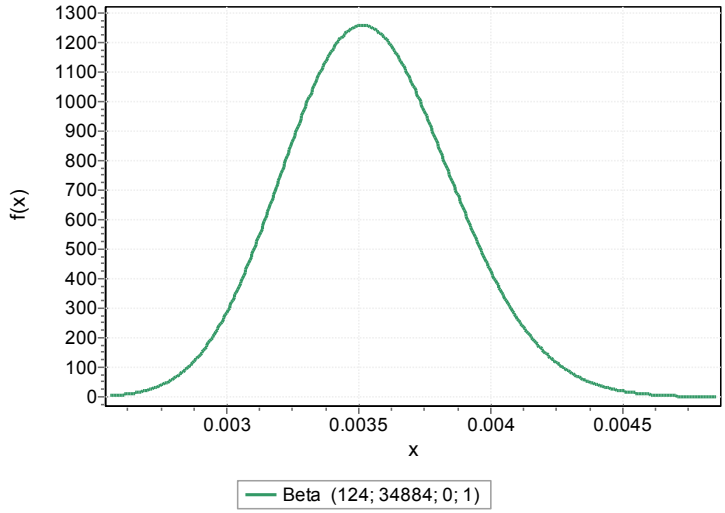
Gráfica 15. Función de densidad de probabilidad a priori.



Fuente: elaboración propia.

Así, cuando se incorpora la experiencia de 2008, es decir, α=124 y β=34,884, se tiene que la función de distribución de Θ se transforma haciéndose puntiaguda alrededor del valor 0.0035.

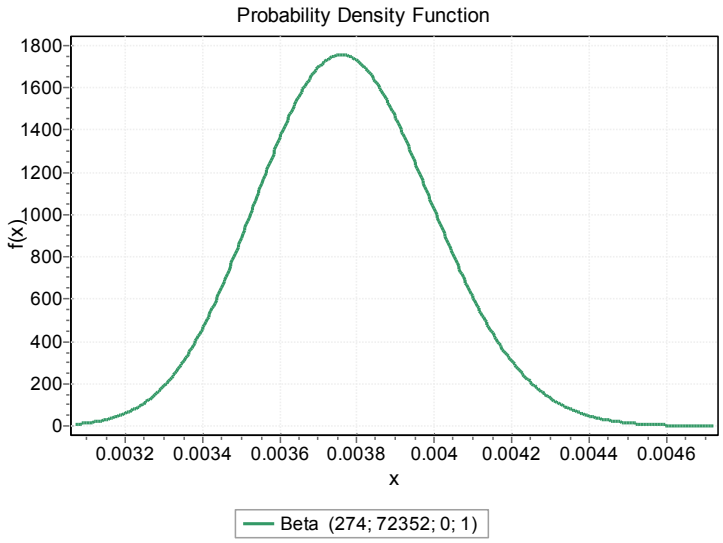
Gráfica 16. Función de densidad de probabilidad de Θ incorporando la experiencia de 2008.



Fuente: elaboración propia.

Incorporando la experiencia de 2009 a la función anterior, la gráfica 17 es más puntiaguda que la anterior con una media de 0.0038 aproximadamente.

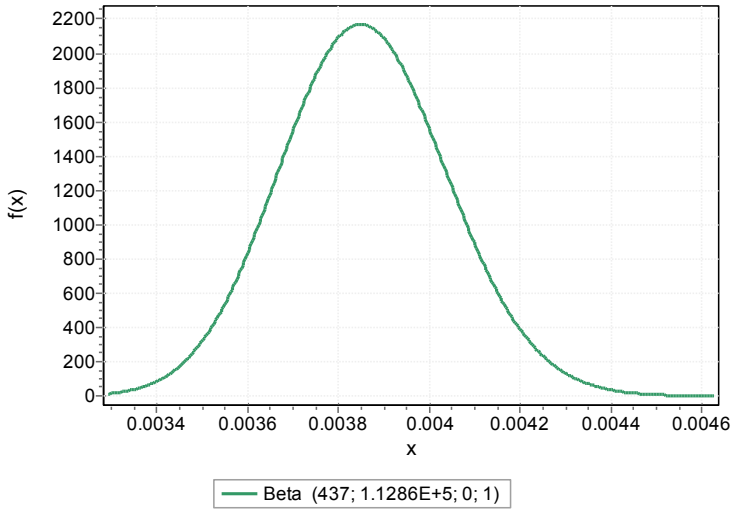
Gráfica 17. Función de densidad de probabilidad de Θ incorporando la experiencia de 2009.



Fuente: elaboración propia.

Incorporando la experiencia de 2010, la gráfica muestra una concentración de valores cerca del 0.0039.

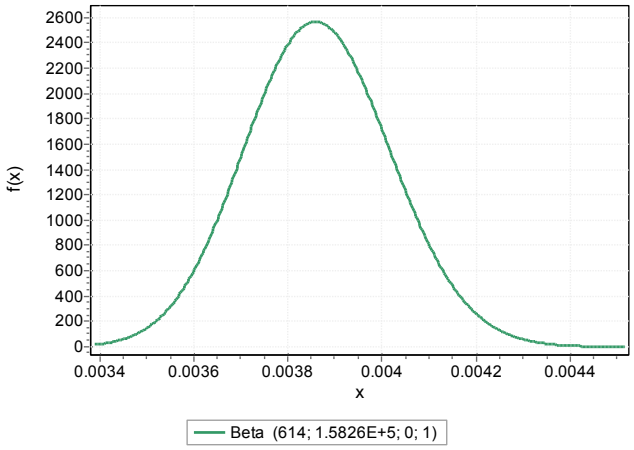
Gráfica 18. Función de densidad de probabilidad de Θ incorporando la experiencia de 2010.



Fuente: elaboración propia.

Incorporando la experiencia de 2011 a la función anterior, se puede observar en la gráfica 19, cómo cada vez hay más valores concentrados cerca de la media, que es aproximadamente 0.0039.

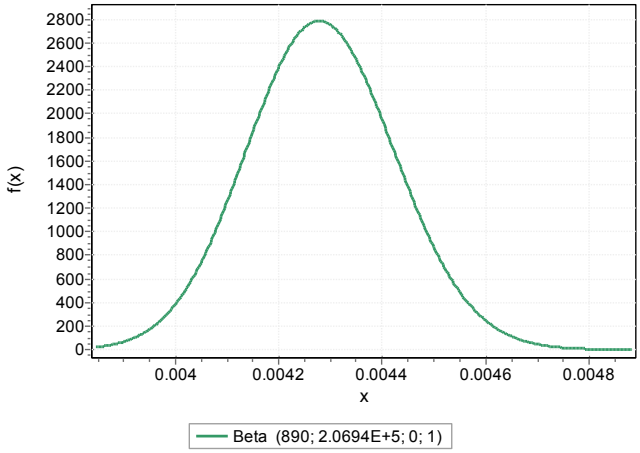
Gráfica 19. Función de densidad de probabilidad de Θ incorporando la experiencia de 2011.



Fuente: elaboración propia.

Y por último, incorporando la experiencia de 2012 a la función anterior se tiene un fuerte cambio en la media de la distribución del estimador Θ, cambiando de 0.0039 a 0.0043

Gráfica 20. Función de densidad de probabilidad de Θ incorporando la experiencia de 2012.



Fuente: elaboración propia.

De las gráficas anteriores se puede observar que conforme se incorpora más información a la muestra, la varianza tiende a cero, ya que la distribución se va a haciendo más puntiaguda, a excepción de 2012 el cual presentó un incremento fuerte respecto a los años 2008-2011. Este incremento es un reflejo de un año de mala experiencia en la compañía debido a cambios en algunos procesos operativos.

3.4 Modelo de pérdidas agregadas

Mediante un código de R mostrado en el Apendice IV se creó el modelo para estimar la pérdida agregada obteniendo su valor esperado, varianza y $TVaR_{95\%}$

Descripción del código

Parte 1:
Se crea una función para calcular el TVaR de un vector de datos. La función ordena el vector (dat) en forma decreciente, selecciona el 5% de los datos más grandes y los promedia.

Parte 2:
Llama la librería MASS para ayudar con los gráficos.

Parte 3:
Se crea la función para estimar la pérdida agregada de un año. Los inputs son:
N = número de pólizas en vigor del año correspondiente.
m = número de simulaciones
p = parámetro *p* de la distribución Binomial del año correspondiente.
mu = parámetro μ de la distribución Lognormal del año correspondiente.
sigma = parámetro σ de la distribución Lognormal del año correspondiente.

La simulación consiste en generar N ensayos Bernoulli, con probabilidad *p* y sumar el número de éxitos, para representar el número de pólizas que tendrán reclamo, llamémosle "*a"* a la suma anterior. Luego se simula el monto de los *a* reclamos y se suma cada uno de ellos. Esta suma representa la perdida agregada de la simulación 1, éste proceso se repite *m* veces y los resultados son guardados en un vector llamado sim.

Parte 4:

Muestra el histograma de frecuencias de las m estimaciones de la pérdida agregada y despliega en una línea la media, varianza y el $TVaR_{95\%}$ y en la siguiente línea despliega el $VaR_{95\%}$, $VaR_{98\%}$ y $VaR_{99.5\%}$

Para modelar la pérdida agregada de cada año se usaron los siguientes estimadores:

Tabla 6. Estimadores utilizados para modelar la pérdida agregada por año.

Año(k)	*Pols en vigor $_k$ (N)*	*p_k*	*μ_k*	*σ_k*
2008	35006	0.003513683	10.68660704	1.204649393
2009	37618	0.003987453	11.39855996	1.183147621
2010	40672	0.004007671	11.0115186	1.615686985
2011	45574	0.003883793	11.28926158	1.243963179
2012	48956	0.005637715	11.05495936	1.320052794

Fuente: elaboración propia.

3.5 Resultados

Los siguientes resultados de la pérdida agregada estimada/teórica fueron obtenidos mediante el modelo Binomial-Lognormal para cada año usando el código de R descrito anteriormente, así como su varianza, $TVaR_{95\%}$, $VaR_{95\%}$, $VaR_{98\%}$ y $VaR_{99.5\%}$. En el Anexo V se pueden encontrar los resultados que arroja directamente R junto con la gráfica del histograma de la pérdida agregada modelada para cada año.

En 2008, la pérdida agregada estimada fue de 11'147,562 pesos, con una varianza de 4'130,783'000,000 pesos, la esperanza de la pérdida agregada dado que ésta es mayor a 14'658,805 pesos ($VaR_{95\%}$) es 16'236,337 pesos ($TVaR_{95\%}$). Adicionalmente, la probabilidad de que la pérdida agregada exceda los 16'128,258 es de 2% y la probabilidad de que exceda 18'442,627 pesos es de 0.5%.

Para 2009, se estimó una pérdida agregada de 27'069,862 pesos, con una varianza de 20'150,010'000,000 pesos. La esperanza de la pérdida agregada dado que ésta es mayor al $VaR_{95\%}$ (34'705,621 pesos) es 37'447,852 pesos ($TVaR_{95\%}$). Adicionalmente, la probabilidad de que la pérdida agregada exceda los 37'870,076 es de 2% y la probabilidad de que exceda 40'658,323 pesos es de 0.5%.

En 2010, el valor esperado de la pérdida agregada fue de 36'489,393 pesos, con varianza de 107'189,000'000,000 pesos. La probabilidad de que la pérdida agregada exceda los 54'380,111 pesos es de 5% ya la media de la pérdida agregada dado que ésta es mayor que el $VaR_{95\%}$ es 64'581,278 pesos ($TVaR_{95\%}$). La probabilidad de que la pérdida agregada sea menor o igual que 62'424,338 pesos es 98% y la probabilidad de que sea menor a 82'359,800 pesos es de 0.5%.

Para 2011, el valor esperado de la pérdida agregada fue de 30’738,932 pesos, con varianza de 26’018,460’000,000 pesos. La probabilidad de que la pérdida agregada exceda los 39’657,647.00 pesos es de 5% y la esperanza de la pérdida agregada dado que ésta es mayor al $VaR_{95\%}$ es 42’982,686 pesos. La probabilidad de que la pérdida agregada sea menor o igual que 42’803,949 pesos es 98% y la probabilidad de que sea menor a 47’873,188 pesos es de 0.5%.

En 2012, la pérdida agregada estimada fue de 41’880,179.00 pesos, con una varianza de 3’879,653’000,000 pesos, la esperanza de la pérdida agregada dado que ésta es mayor a $52’346,892 pesos ($VaR_{95\%}$) es 57’533,624 pesos ($TVaR_{95\%}$). Adicionalmente, la probabilidad de que la pérdida agregada exceda los 56’237,120 es de 2% y la probabilidad de que exceda 64’705,259 pesos es de 0.5%.

La siguiente tabla muestra un resumen de los resultados mencionados anteriormente:

Tabla 7. Resultados del modelo de pérdida agregada 2008-2012.

Año (k)	S_k estimada	Varianza$_k$	$TVaR_{95\%}$	$VaR_{95\%}$	$VaR_{98\%}$	$VaR_{99.5\%}$
2008	11,147,562	4.13078E+12	16,236,337	14,658,805	16,148,258	18,442,627
2009	27,069,862	2.015E+13	37,447,852	34,705,621	37,870,076	40,658,323
2010	36,489,393	1.07189E+14	64,581,278	54,380,111	62,424,338	82,359,800
2011	30,738,932	2.60185E+13	42,982,686	39,657,647	42,803,949	47,873,188
2012	41,880,179	3.87965E+13	57,533,624	52,346,892	56,237,120	64,705,259

Fuente: elaboración propia.

A continuación se muestra un comparativo de los resultados obtenidos de la pérdida agregada mediante el modelo teórico (Binomial-Lognormal) y los datos empíricos de cada año:

Tabla 8. Comparativo entre *S* ocurrida y *S* estimada.

Año (k)	S_k	S_k estimada	Diferencia
2008	$ 11,121,397	$ 11,147,562	-$ 26,165
2009	$ 26,939,776	$ 27,069,862	-$ 130,086
2010	$ 36,414,989	$ 36,489,393	-$ 74,404
2011	$ 30,680,367	$ 30,738,932	-$ 58,565
2012	$ 41,725,788	$ 41,880,179	-$ 154,391

Fuente: elaboración propia.

Una vez que se ha observado que el modelo Binomial-Lognormal proporciona resultados coherentes y cercanos a los ocurridos en el pasado, se incluirá el estimador Bayesiano de la proporción de pólizas con reclamo mostrado en la tabla 5. El estimador que incorpora toda la información disponible se utilizará en el modelo Binomial-Lognormal para estimar la pérdida agregada de 2013.

Incorporando el estimador bayesiano los resultados son los siguientes:

En 2008, se tiene que con el 95% de probabilidad la pérdida agregada se encuentra entre 9’362,692 y 13’246,424 pesos, el $VaR_{95\%}$ se encuentra entre 12’881,204 y 16’969,510, el $VaR_{98\%}$ está entre 13’954,140 y 18’269,530 y el $VaR_{99.5\%}$ se encuentra contenido dentro del intervalo de 16’291,025 a 20’765,013 pesos. El valor esperado de la pérdida agregada dado que ésta superó el $VaR_{95\%}$ está entre 14’261,736.00 y 18’523,302 pesos.

En 2009, se tiene que con el 95% de probabilidad la pérdida agregada se encuentra entre 22’469,522 y 28’691,235 pesos, el $VaR_{95\%}$ se encuentra entre 29’559,186 y 37’142,568, el $VaR_{98\%}$ está entre 31’144,198.00 y 39’374,916 y el $VaR_{99.5\%}$ se encuentra contenido dentro del intervalo de 33’917,253 a 41’886,371 pesos. El valor esperado de la pérdida agregada dado que ésta superó el $VaR_{95\%}$ está entre 31’602,777 y 39’499,198 pesos.

En 2010, se tiene que con el 95% de probabilidad la pérdida agregada se encuentra entre 31'930,427 y 39'279,554 pesos, el $VaR_{95\%}$ se encuentra entre 48'641,892 y 59'055,304, el $VaR_{98\%}$ está entre 55'006,845 y 65'652,500 y el $VaR_{99.5\%}$ se encuentra contenido dentro del intervalo de 64'411,197 a 83'087,643 pesos. El valor esperado de la pérdida agregada dado que ésta superó el $VaR_{95\%}$ está entre 56'033,023 y 70'813,667 pesos.

En 2011, se tiene que con el 95% de probabilidad la pérdida agregada se encuentra entre 27'898,997 y 32'764,917 pesos, el $VaR_{95\%}$ se encuentra entre 36'309,549 y 41'009,617, el $VaR_{98\%}$ está entre 39'176,734 y 44'143,425 y el $VaR_{99.5\%}$ se encuentra contenido dentro del intervalo de 43'606,686 a 54'343,580 pesos. El valor esperado de la pérdida agregada dado que ésta superó el $VaR_{95\%}$ está entre 39'417,334 y 45'263,086 pesos.

En 2012, se tiene que con el 95% de probabilidad la pérdida agregada se encuentra entre 29'745,335 y 33'957,237 pesos, el $VaR_{95\%}$ se encuentra entre 39'422,580 y 43'964,181, el $VaR_{98\%}$ está entre 42'725,502 y 47'306,194 y el $VaR_{99.5\%}$ se encuentra contenido dentro del intervalo de 48'368,123 a 52'644,422 pesos. El valor esperado de la pérdida agregada dado que ésta superó el $VaR_{95\%}$ está entre 43'324,642 y 48'478,127pesos.

La siguiente tabla resume la información obtenida incluyendo en el modelo el estimador bayesiano (los datos obtenidos del modelo se encuentran en el Anexo VI).

Tabla 9. Intervalos de probabilidad a un nivel de confianza del 95% para *p*.

Año	P=Percentil 2.5%	P=Percentil 97.5%	Real	Observaciones
2008	$ 9,362,692	$ 13,246,424	$11,121,397	Dentro del intervalo
2009	$ 22,469,522	$ 28,691,235	$26,939,776	Dentro del intervalo
2010	$ 31,930,427	$ 39,279,554	$36,414,989	Dentro del intervalo
2011	$ 27,898,997	$ 32,764,917	$30,680,367	Dentro del intervalo
2012	$ 29,745,335	$ 33,957,237	$41,725,788	Fuera del intervalo

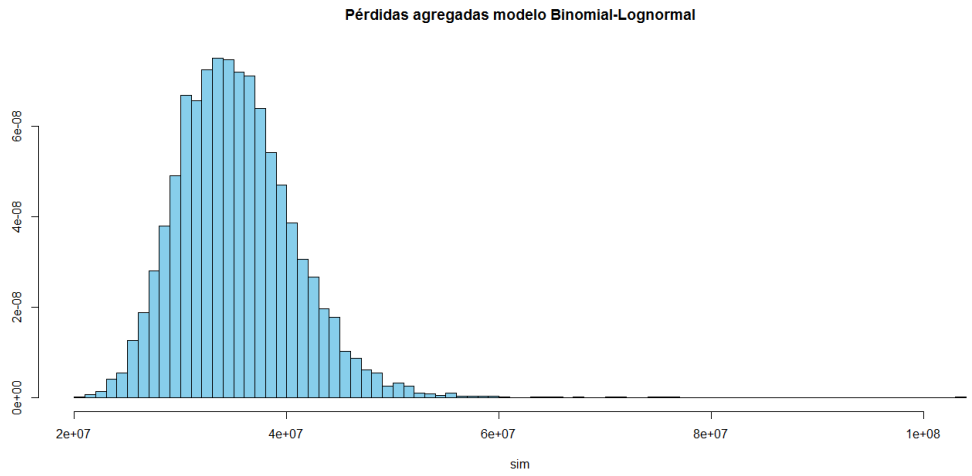
Fuente: elaboración propia.

De la tabla anterior, se puede observar que la pérdida agregada ocurrida en los años 2008 al 2011 se encuentra dentro el intervalo de probabilidad para la pérdida agregada. La pérdida agregada de 2012 se encuentra fuera de su intervalo, pero está contenida dentro del intervalo del $VaR_{95\%;}$ esto indica que la pérdida agregada ocurrida estuvo contenida dentro del 5% de los peores posibles valores de ese año. Este resultado es consistente con la realidad, ya que hubo cambios en la operación de la compañía durante este año.

Para estimar la pérdida agregada de 2013 se utilizó la media de los parámetros μ y σ de la distribución Lognormal de los años 2008-2012, una tasa de crecimiento promedio de pólizas en vigor del 8.8%, la media así como los percentiles 2.5% y 97.5%, del estimador bayesiano *p* para obtener intervalos de probabilidad (véase Anexo VI).

La gráfica 21 muestra el histograma de frecuencias del modelo Beta-Biniomial que se utilizará para hacer inferencias sobre el comportamiento de los reclamos provenientes de cáncer de mama durante 2013.

Gráfica 21. IRM Binomial-Lognormal de la pérdida agregada por cáncer de mama en 2013.



Fuente: elaboracion propia.

CONCLUSIONES

Los resultados que se obtuvieron mediante el modelo de riesgo individual Binomial-Lognormal permiten inferir que la pérdida agregada (*S*) de 2013 a causa de siniestros de cáncer de mama, será de 35’267,891 pesos y con probabilidad del 95% está pérdida se encontrará entre 33’095,293 y 37’689,990 pesos.

Con un 5% de probabilidad, la pérdida agregada por siniestros de cáncer de mama durante 2013 rebasará los 45’176,073 pesos y con un intervalo de probabilidad del 95% el $VaR_{95\%}$ estará contenido en el rango de 42’256,659 a 47’645,540 pesos. Es decir, la máxima pérdida agregada esperada S con un nivel de confianza del 95% estará entre 42’256,659 y 47’645,540 pesos. Asimismo, la esperanza de la pérdida agregada derivada de siniestros de cáncer de mama durante 2013 dado que ésta supera el $VaR_{95\%}$ es de 48’925,085 pesos y con un 95% de probabilidad el $TVaR_{95\%}$ estará entre 46’306,422 y 51’724,984 pesos.

Con probabilidad del 98%, la pérdida agregada por siniestros de cáncer de mama durante 2013 será menor a los 48’237,215 pesos y con un intervalo de probabilidad del 95% el $VaR_{98\%}$ estará contenido en el rango de 46’051,960 a 51’482,890 pesos.

Con probabilidad del 99.5%, la pérdida agregada por siniestros de cáncer de mama durante 2013 será menor a los 54’160,837 pesos y con un intervalo de probabilidad del 95% el $VaR_{99.5\%}$ estará contenido en el rango de 51’976,546 a 56’799,437 pesos.

Resumiendo lo anterior, si la compañía aseguradora quisiera un grado de certeza del 95% sobre el estimado de la pérdida agregada por cáncer de mama, y tomara una postura conservadora, es decir, se preparase para el peor

escenario en 2013, esperará que el monto a pagar en 2013 por siniestros provenientes de cáncer de mama ascienda a lo más a 51’724,984 pesos. Y conforme se incremente el grado de certeza que la compañía requiera en sus estimados, incrementará el monto esperado a pagar debido a esos siniestros que tienen muy poca probabilidad de ocurrir dado que si monto es muy grande.

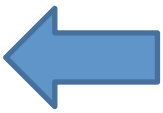
ANEXOS

Anexo I. Pruebas de bondad de ajuste

Imagen 1. Las 10 distribuciones con mejor ajuste de acuerdo a la prueba Anderson Darling para la severidad de 2008.

Goodness of Fit - Summary

#	Distribution	Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
		Statistic	Rank	Statistic	Rank	Statistic	Rank
35	Log-Logistic (3P)	0.04482	4	0.31548	1	3.0337	2
45	Pearson 6	0.04472	3	0.32584	2	2.7601	1
34	Log-Logistic	0.04678	6	0.33519	3	4.2886	7
38	Lognormal	0.05224	9	0.34223	4	5.3025	12
2	Burr	0.04648	5	0.34287	5	3.0349	3
39	Lognormal (3P)	0.05097	8	0.34898	6	5.2594	11
36	Log-Pearson 3	0.0572	11	0.35566	7	4.57	8
27	Inv. Gaussian (3P)	0.04815	7	0.37557	8	4.8417	9
42	Pareto 2	0.05366	10	0.38408	9	3.0885	4
16	Frechet (3P)	0.04369	2	0.44168	10	3.1087	5

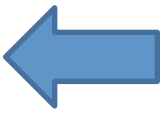


Fuente: elaboración propia.

Imagen 2. Las 10 distribuciones con mejor ajuste de acuerdo a la prueba Anderson Darling para la severidad de 2009.

Goodness of Fit - Summary

#	Distribution	Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
		Statistic	Rank	Statistic	Rank	Statistic	Rank
14	Fatigue Life (3P)	0.05293	2	0.29661	1	6.229	4
36	Log-Pearson 3	0.04964	1	0.35999	2	4.285	1
38	Lognormal	0.05465	5	0.46054	3	6.4161	5
39	Lognormal (3P)	0.05426	4	0.46335	4	5.3448	2
2	Burr	0.05851	6	0.58262	5	6.1097	3
34	Log-Logistic	0.0708	12	0.77155	6	8.2352	6
33	Log-Gamma	0.062	7	0.86386	7	10.302	9
42	Pareto 2	0.06813	9	0.86483	8	9.4391	7
16	Frechet (3P)	0.06943	11	1.2302	9	14.392	14
27	Inv. Gaussian (3P)	0.07973	13	1.2371	10	13.485	13



Fuente: elaboración propia.

Imagen 3. Las 10 distribuciones con mejor ajuste de acuerdo a la prueba Anderson Darling para la severidad de 2010.

Goodness of Fit - Summary

#	Distribution	Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
		Statistic	Rank	Statistic	Rank	Statistic	Rank
35	Log-Pearson 3	0.05355	1	0.71537	1	5.4824	3
38	Lognormal (3P)	0.05718	2	0.76725	2	5.6668	4
37	Lognormal	0.05773	4	0.78774	3	6.9047	6
14	Fatigue Life (3P)	0.05728	3	0.80163	4	4.047	1
32	Log-Gamma	0.07044	7	1.1857	5	5.1927	2
13	Fatigue Life	0.08773	15	1.3009	6	7.6731	7
33	Log-Logistic	0.07795	10	1.3099	7	8.8968	9
2	Burr	0.08534	14	1.4738	8	9.5655	10
41	Pareto 2	0.07283	8	1.4953	9	7.7506	8
16	Frechet (3P)	0.0789	13	1.6174	10	10.943	14

Fuente: elaboración propia.

Imagen 4. Las 10 distribuciones con mejor ajuste de acuerdo a la prueba Anderson Darling para la severidad de 2011.

Goodness of Fit - Summary

#	Distribution	Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
		Statistic	Rank	Statistic	Rank	Statistic	Rank
38	Lognormal	0.04101	2	0.25033	1	1.8822	1
39	Lognormal (3P)	0.04388	5	0.27357	2	2.2356	2
36	Log-Pearson 3	0.043	4	0.32844	3	3.3029	3
33	Log-Gamma	0.04106	3	0.43334	4	3.9691	5
34	Log-Logistic	0.04958	7	0.44802	5	4.3999	6
2	Burr	0.05037	8	0.48033	6	4.8038	7
45	Pearson 6	0.05078	10	0.48858	7	4.8998	9
42	Pareto 2	0.05073	9	0.49193	8	4.8734	8
35	Log-Logistic (3P)	0.0398	1	0.55355	9	3.8133	4
21	Gen. Gamma (4P)	0.04597	6	0.59269	10	5.448	10

Fuente: elaboración propia.

Imagen 5. Las 10 distribuciones con mejor ajuste de acuerdo a la prueba Anderson Darling para la severidad de 2012.

Goodness of Fit - Summary

#	Distribution	Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
		Statistic	Rank	Statistic	Rank	Statistic	Rank
35	Log-Pearson 3	0.03421	1	0.2979	1	5.3635	5
38	Lognormal (3P)	0.03649	4	0.37702	2	5.0166	3
37	Lognormal	0.03466	2	0.39197	3	3.083	1
21	Gen. Gamma (4P)	0.04313	5	0.40512	4	5.1809	4
2	Burr	0.04407	6	0.67148	5	6.7415	9
34	Log-Logistic (3P)	0.03622	3	0.67637	6	5.6157	8
7	Dagum	0.04838	9	0.7053	7	6.8307	11
14	Fatigue Life (3P)	0.05022	10	0.75708	8	3.8896	2
44	Pearson 6	0.04448	7	0.79114	9	6.8237	10
33	Log-Logistic	0.05319	13	0.84135	10	5.5656	6

Fuente: elaboración propia.

Anexo II. Pruebas de hipótesis distribución Lognormal

Imagen 6. Pruebas de bondad de ajuste para el modelo Lognormal 2008.

Lognormal [#38]					
Kolmogorov-Smirnov					
Sample Size Statistic P-Value Rank	123 0.05224 0.87269 9				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	0.09675	0.11027	0.12245	0.13687	0.14688
Reject?	No	No	No	No	No
Anderson-Darling					
Sample Size Statistic Rank	123 0.34223 4				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	1.3749	1.9286	2.5018	3.2892	3.9074
Reject?	No	No	No	No	No
Chi-Squared					
Deg. of freedom Statistic P-Value Rank	6 5.3025 0.50564 12				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	8.5581	10.645	12.592	15.033	16.812
Reject?	No	No	No	No	No

Fuente: elaboración propia.

Imagen 7. Pruebas de bondad de ajuste para el modelo Lognormal 2009.

Lognormal [#38]					
Kolmogorov-Smirnov					
Sample Size Statistic P-Value Rank	150 0.05465 0.7403 5				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	0.08761	0.09986	0.11088	0.12394	0.13301
Reject?	No	No	No	No	No
Anderson-Darling					
Sample Size Statistic Rank	150 0.46054 3				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	1.3749	1.9286	2.5018	3.2892	3.9074
Reject?	No	No	No	No	No
Chi-Squared					
Deg. of freedom Statistic P-Value Rank	7 6.4161 0.49209 5				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	9.8032	12.017	14.067	16.622	18.475
Reject?	No	No	No	No	No

Fuente: elaboración propia.

Imagen 8. Pruebas de bondad de ajuste para el modelo Lognormal 2010.

Lognormal [#37]					
Kolmogorov-Smirnov					
Sample Size Statistic P-Value Rank	163 0.05773 0.62775 4				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	0.08404	0.09579	0.10637	0.1189	0.12759
Reject?	No	No	No	No	No
Anderson-Darling					
Sample Size Statistic Rank	163 0.78774 3				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	1.3749	1.9286	2.5018	3.2892	3.9074
Reject?	No	No	No	No	No
Chi-Squared					
Deg. of freedom Statistic P-Value Rank	7 6.9047 0.43887 6				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	9.8032	12.017	14.067	16.622	18.475
Reject?	No	No	No	No	No

Fuente: elaboración propia.

Imagen 9. Pruebas de bondad de ajuste para el modelo Lognormal 2011.

Lognormal [#38]					
Kolmogorov-Smirnov					
Sample Size Statistic P-Value Rank	177 0.04101 0.91512 2				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	0.08065	0.09193	0.10207	0.1141	0.12244
Reject?	No	No	No	No	No
Anderson-Darling					
Sample Size Statistic Rank	177 0.25033 1				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	1.3749	1.9286	2.5018	3.2892	3.9074
Reject?	No	No	No	No	No
Chi-Squared					
Deg. of freedom Statistic P-Value Rank	7 1.8822 0.96607 1				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	9.8032	12.017	14.067	16.622	18.475
Reject?	No	No	No	No	No

Fuente: elaboración propia.

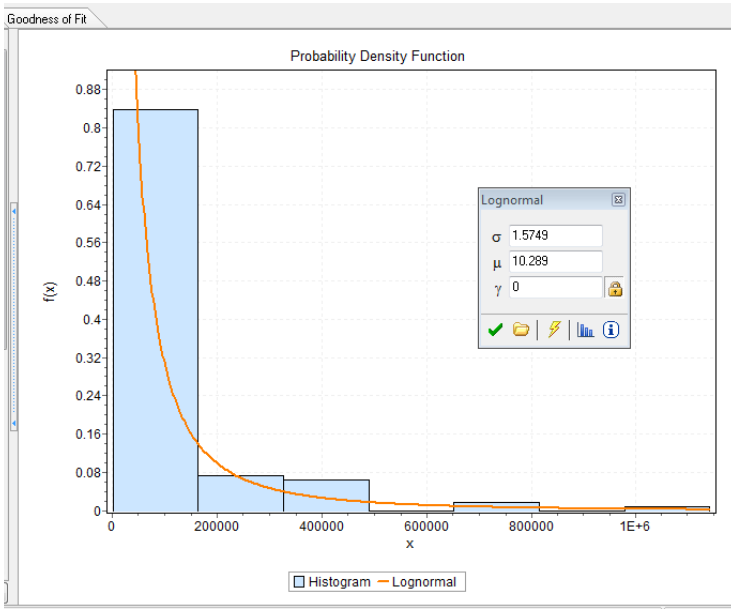
Imagen 10. Pruebas de bondad de ajuste para el modelo Lognormal 2012.

Lognormal [#37]					
Kolmogorov-Smirnov					
Sample Size Statistic P-Value Rank	276 0.03466 0.88303 2				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	0.06459	0.07362	0.08174	0.09137	0.09805
Reject?	No	No	No	No	No
Anderson-Darling					
Sample Size Statistic Rank	276 0.39197 3				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	1.3749	1.9286	2.5018	3.2892	3.9074
Reject?	No	No	No	No	No
Chi-Squared					
Deg. of freedom Statistic P-Value Rank	8 3.083 0.92904 1				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	11.03	13.362	15.507	18.168	20.09
Reject?	No	No	No	No	No

Fuente: elaboración propia.

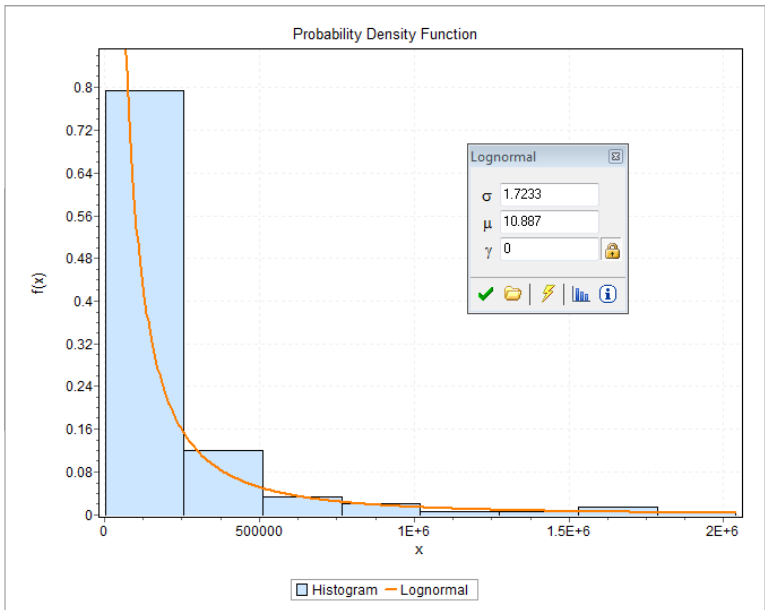
Anexo III. Modelo empírico vs. Teórico (2008-2011)

Gráfica 22. Función de densidad Lognormal vs función de densidad empírica para la severidad de 2008.



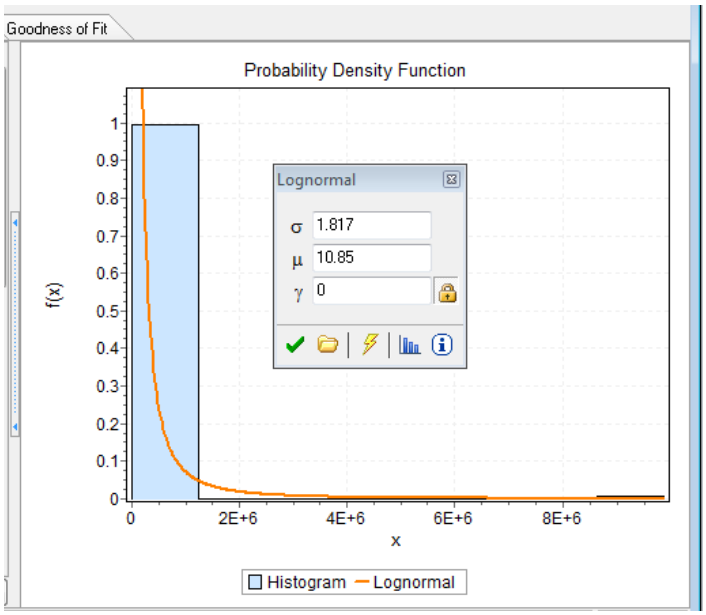
Fuente: elaboración propia.

Gráfica 23. Función de densidad Lognormal vs función de densidad empírica para la severidad de 2009.



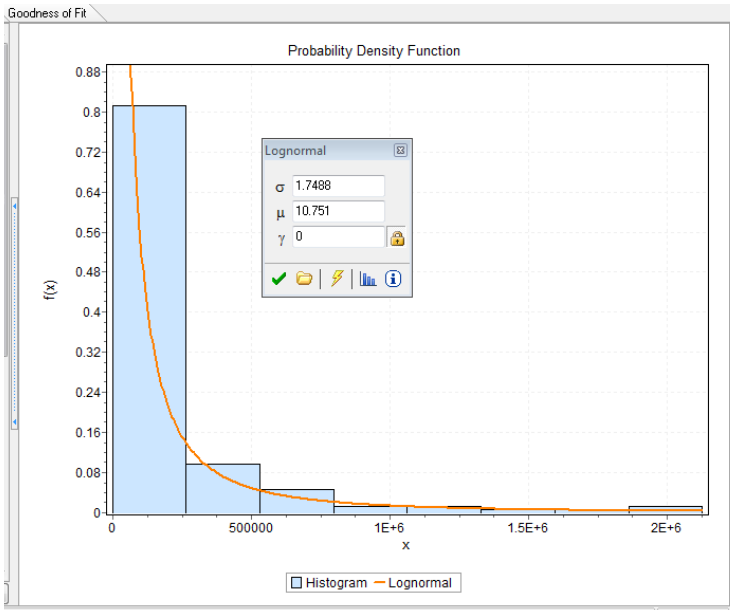
Fuente: elaboración propia.

Gráfica 24. Función de densidad Lognormal vs función de densidad empírica para la severidad de 2010.



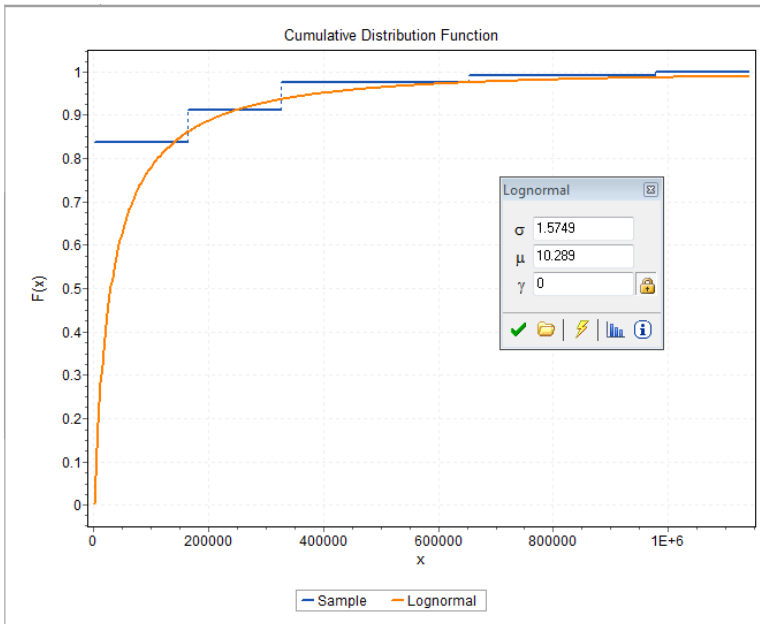
Fuente: elaboración propia.

Gráfica 25. Función de densidad Lognormal vs función de densidad empírica para la severidad de 2011.



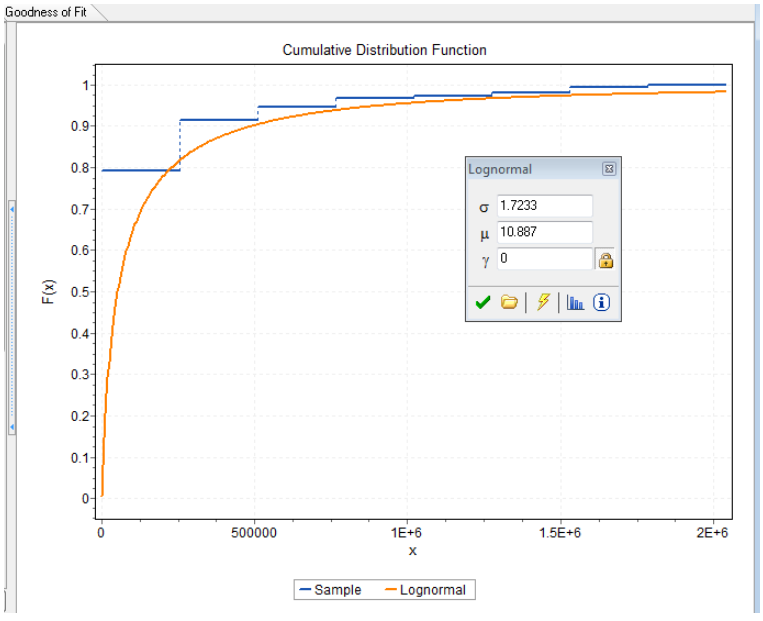
Fuente: elaboración propia.

Gráfica 26. Función de distribución Lognormal vs función de distribución empírica para la severidad de 2008.



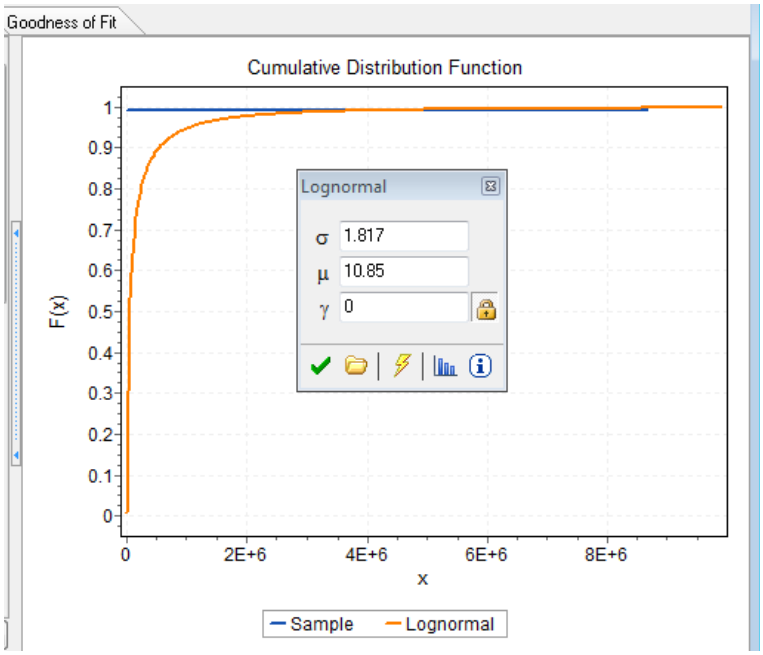
Fuente: elaboración propia.

Gráfica 27. Función de distribución Lognormal vs función de distribución empírica para la severidad de 2009.



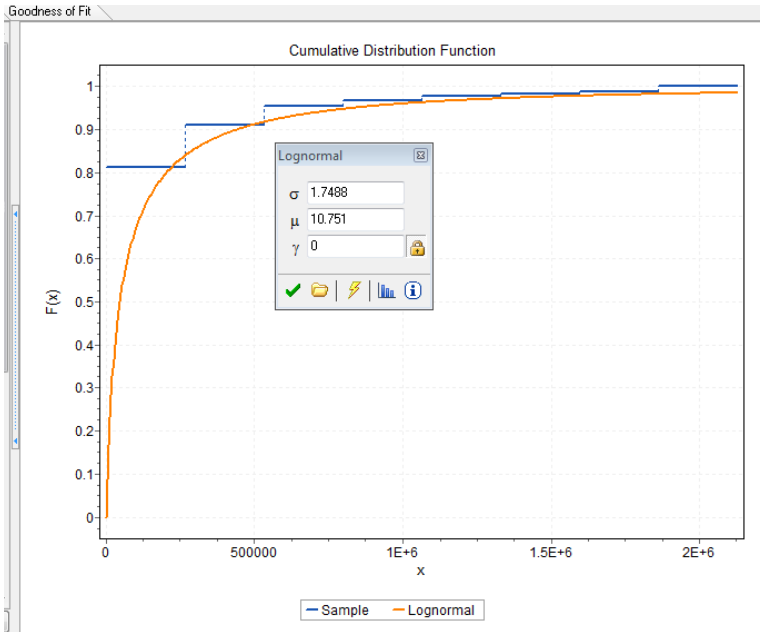
Fuente: elaboración propia.

Gráfica 28. Función de distribución Lognormal vs función de distribución empírica para la severidad de 2010.



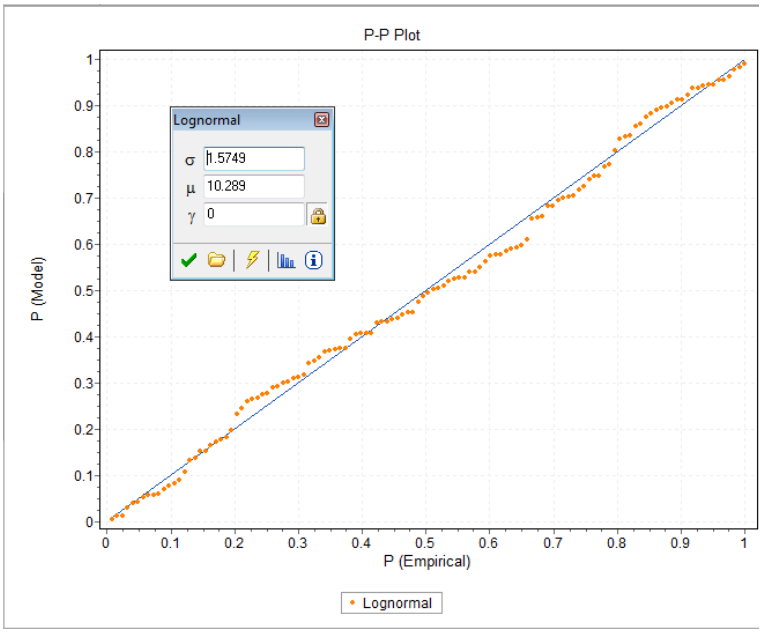
Fuente: elaboración propia.

Gráfica 29. Función de distribución Lognormal vs función de distribución empírica para la severidad de 2011.



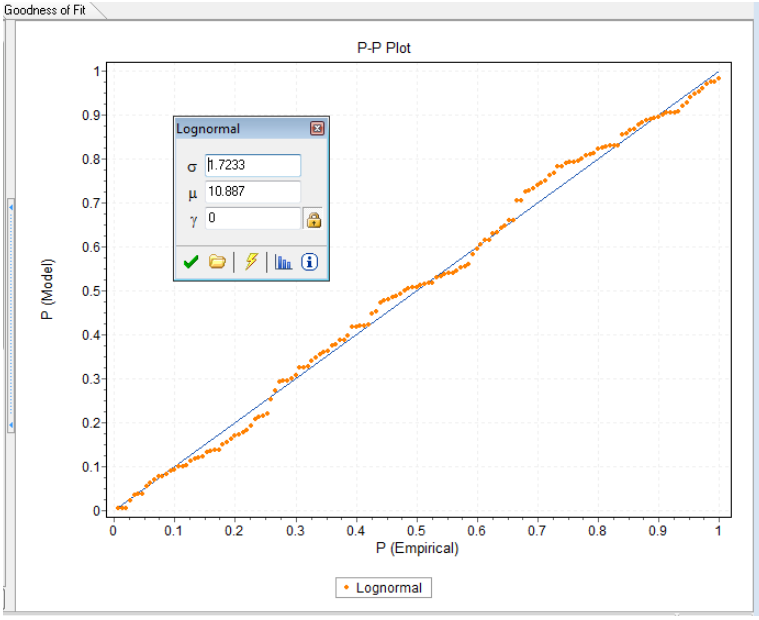
Fuente: elaboración propia.

Gráfica 30. PP Plot para la severidad de 2008.



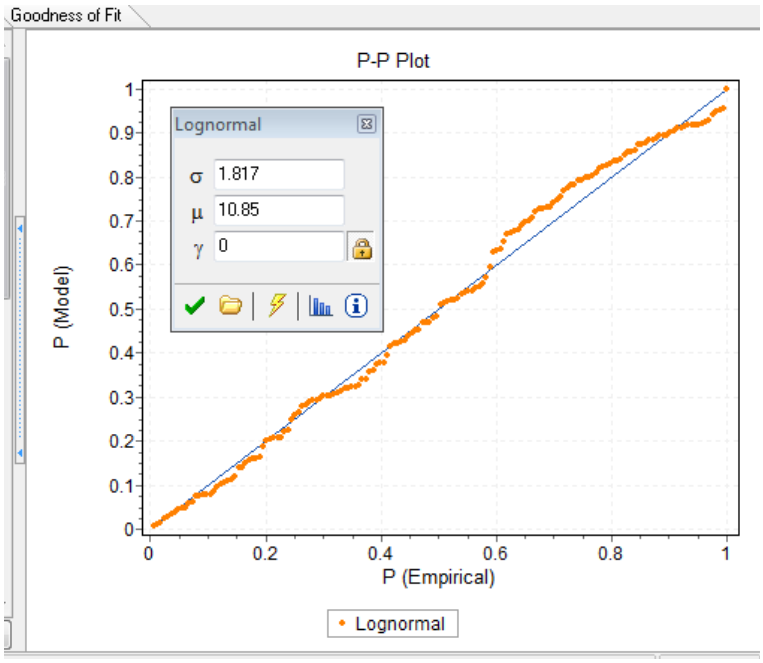
Fuente: elaboración propia.

Gráfica 31. PP Plot para la severidad de 2009.



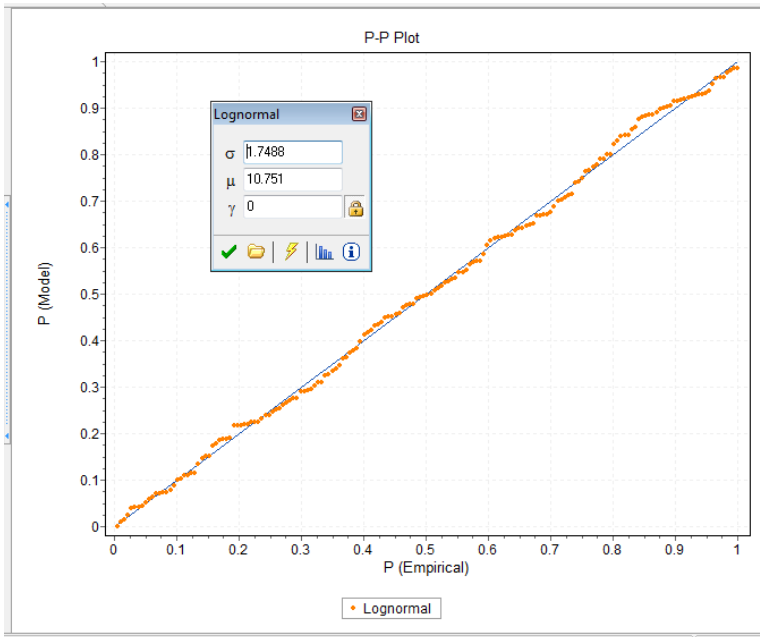
Fuente: elaboración propia.

Gráfica 32. PP Plot para la severidad de 2010.



Fuente: elaboración propia.

Gráfica 33. PP Plot para la severidad de 2011.



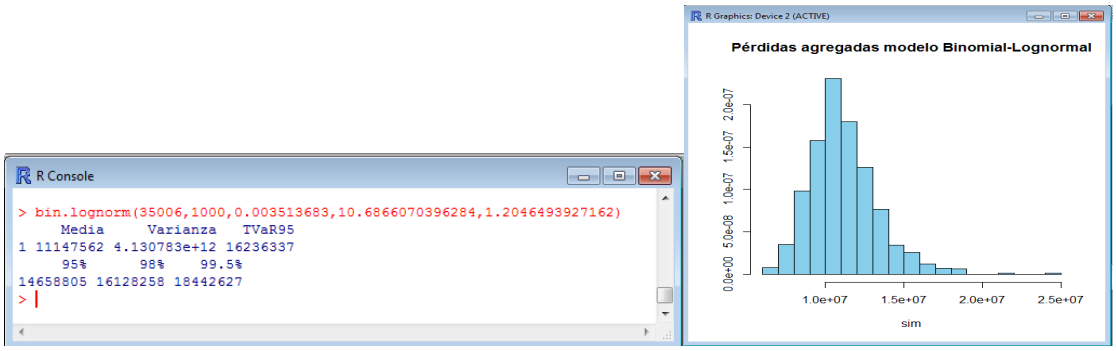
Fuente: elaboración propia.

Anexo IV. Código de R

```
# Modelo de pérdidas agregadas
# Parte 1
es<-function(dat,a)
        {
        n<-trunc((length(dat))*(1-a))
        ord<-sort(dat,decreasing=TRUE)
        return(sum(ord[1:n])/n)
        }
# Parte 2
library(MASS)
# Parte 3
bin.lnorm<-function(N,m,p,mu,sigma)
        {
        sim<-c()
        for(i in 1:m)
                {
                sev<-rlnorm(sum(rbinom(N,1,p)),mu,sigma)
                sim[i]<-sum(sev)
                }
# Parte 4
        truehist(sim,main="Pérdidas agregadas modelo Binomial-Lognormal",col="skyblue")
         vec1<-
        data.frame(cbind(Media=mean(sim),Varianza=var(sim),TVaR95=es(sim,.95)))
        vec<-c(quantile(sim,0.95),quantile(sim,0.98),quantile(sim,0.995))
        print(vec1)
        print(vec)
        }
```

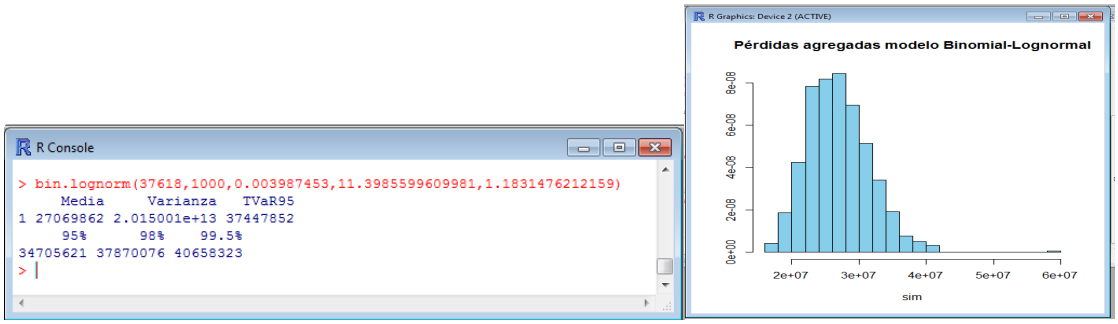
Anexo V. Resultados del modelo Binomial-Lognormal

Imagen 11. Estimación de la pérdida agregada por cáncer de mama en 2008.



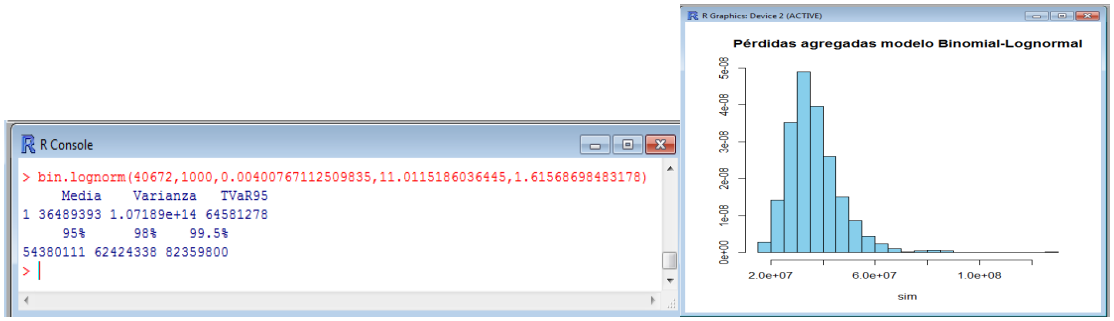
Fuente: elaboración propia.

Imagen 12. Estimación de la pérdida agregada por cáncer de mama en 2009.



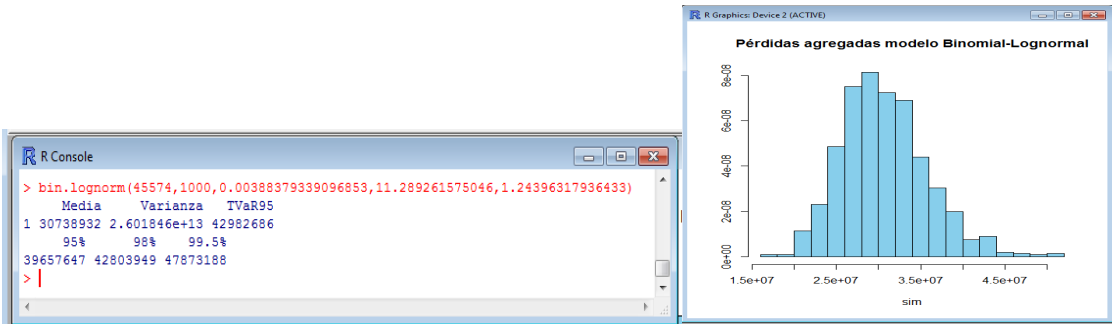
Fuente: elaboración propia.

Imagen 13. Estimación de la pérdida agregada por cáncer de mama en 2010.



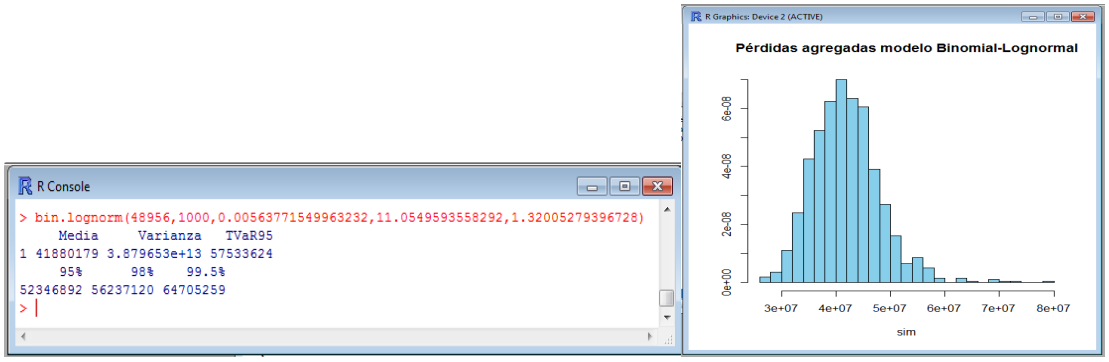
Fuente: elaboración propia.

Imagen 14. Estimación de la pérdida agregada por cáncer de mama en 2011.



Fuente: elaboración propia.

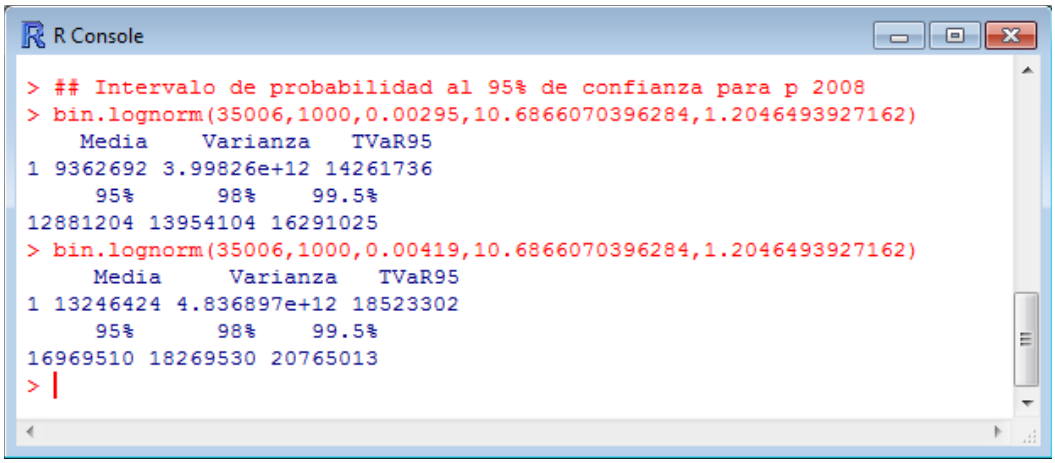
Imagen 15. Estimación de la pérdida agregada por cáncer de mama en 2012.



Fuente: elaboración propia.

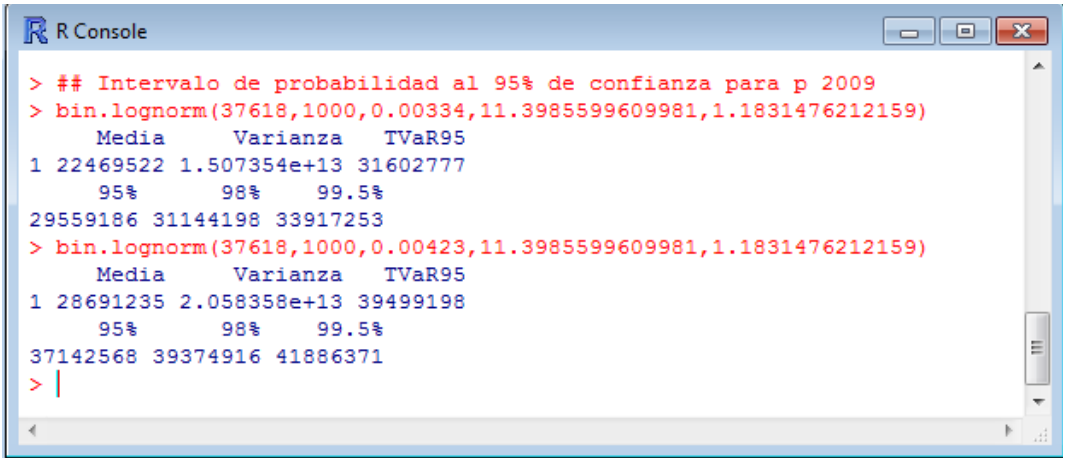
Anexo VI. Resultados incluyendo estimador bayesiano

Imagen 16. Estimación de intervalos de la pérdida agregada para 2008.



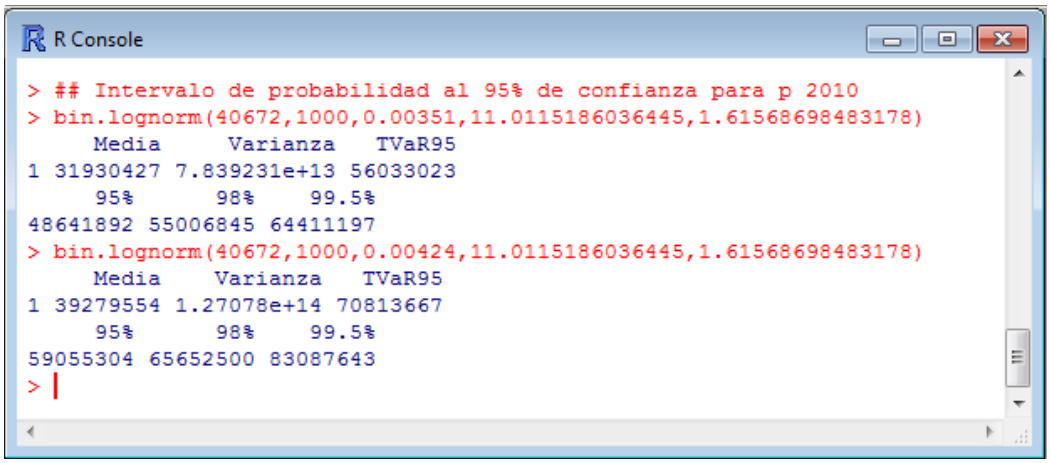
Fuente: elaboración propia.

Imagen 17. Estimación de intervalos de la pérdida agregada para 2009.



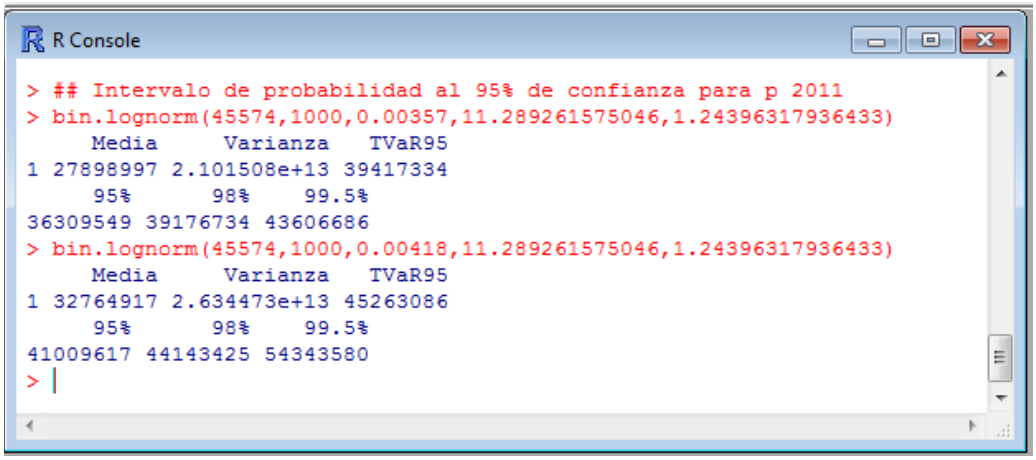
Fuente: elaboración propia.

Imagen 18. Estimación de intervalos de la pérdida agregada para 2010.



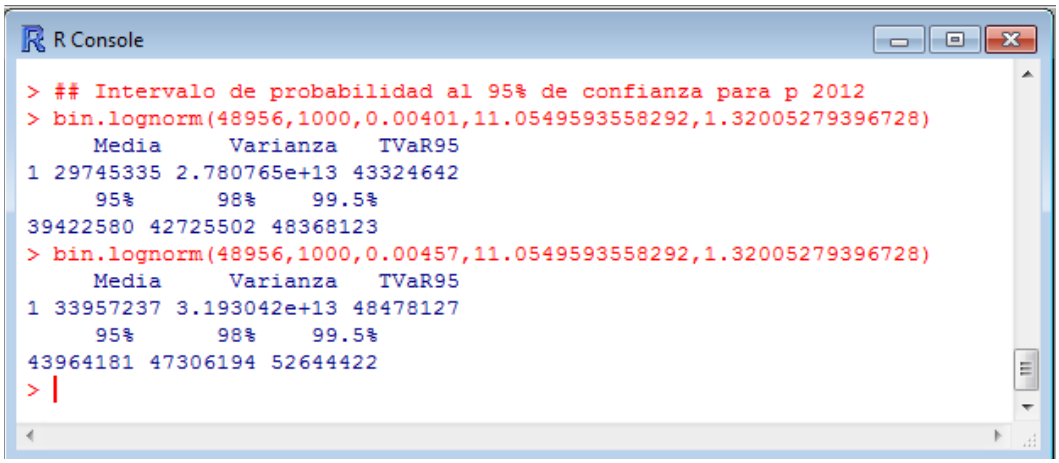
Fuente: elaboración propia.

Imagen 19. Estimación de intervalos de la pérdida agregada para 2011.



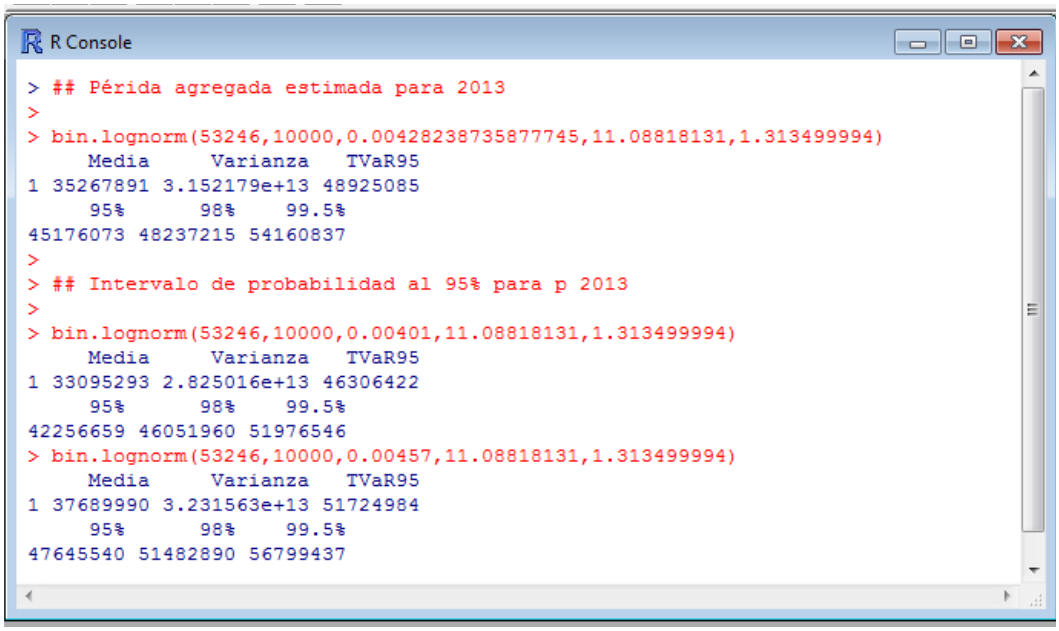
Fuente: elaboración propia.

Imagen 20. Estimación de intervalos de la pérdida agregada para 2012.



Fuente: elaboración propia.

Estimación de intervalos de probabilidad para la pérdida agregada de 2013.



Fuente: elaboración propia.

BIBLIOGRAFÍA

Artzner, P., Delbaen, F., Eber, J. M. and D. Heath (1997). Thinking coherently. Risk 10(11). Pp. 68-71.

Bernardo, J. M. (2003). Bayesian Statistics. Probability and statistics R. Viertl, ed. Oxford Science Publications. Pp. 1-8. Oxford.

Blumenfeld, M. (1961). Recent Trends and Innovations in Individual Hospital Insurance. Proceedings of the Casualty Actuarial Society. Volume XLVIII. Pp. 83-95.

Bowers, N. L. Gerber, H. U. Hickman, J. C. Jones D.A. Nesbitt, C.J. Actuarial Mathematics. (1997). Society of Actuaries. Illinios.

Bühlmann, H. (1970). Mathematical Methods in Risk Theory. Springer-Varlag. Berlin. Pp. 3-62. Instituto Matemático de la Universidad de Zurich.

Burnecki, K. Janczura, J. y Weron, R. (2010). Building Loss Models. MPRA. Wroclaw University of Technology. Polonia.

Burnecki, Misiorek y Weron (2005). Loss Distributions. Wroclaw University of Technology. Santander Consumer Bank S. A. Wroclaw, Polonia.

Chernick, M. R. (1999). Bootstrap Methods. Pp. XI-22. John Wiley & Sons, Inc. Nueva York, NY.

Aguilera, M. (2012). Proyecto Solvencia II. XXIV Congreso Nacional de Actuarios. Comisión Nacional de Seguros y Fianzas. México.

Daykin, C. D., Bernstein, G. D., Coutts, S. M., Devitt, E. R. F., Hey, G. B., Reynolds, D.I.W. Y Smith, P.D. (1987). Assessing the Solvency and Financial Strength of a General Insurance Company. Journal of the Institute of Actuaries. P.1.

Davidson, A. C. y Hinkley, D. V. (1997). Bootstrap methods and their application. Cambridge Uiversity Press. Cambridge, Reino Unido.

Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrual and Applied Mathematics. Estados Unidos.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Stanford University. The Anals of Statistics. Vol. 7. No.1, 1-26.

Efron, B. y Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Pp. 14-38. Editorial Chapman&Hall. Nueva York, NY.

Escalante, C. y Arango, G. O. (2004). Aspectos Básicos del Modelo de Riesgo Colectivo. Pp. 1-8. Matemáticas: Enseñanza Universitaria Universidad del Valle. Vol. XII. Escuela Regional de Matemáticas Universidad del Valle. Cali, Colombia.

Feria, J. M., Jiménez, E. J. y Martín, J. L. (2007). El Modelo de Distribución de Pérdidas Agregadas (LDA): Una Aplicación al Riesgo Operacional. Consejería de Innovación, ciencia y empresa de la Junta de Andalucía. Andalucía, España.

Fishman G. (1996). Monte Carlo: Concepts, Algorithms, and Applications. Springer Verlag. Estados Unidos.

Foss, S. Korshunov D. Zachary S. (2011). An Introduction to Heavy Tailed and Subexponential Distributions. Springer. Alemania.

Frachot, A. Georges, P. y Roncalli, T. (2001). Loss Distribution Approach for Operational Risk.

Heckman P. y Meyers, G. (1983). Loss Distributions form Claim Severity and Claim Count Distributions. PCAS. P. 2. Virginia.

Kaas, R., Goovaerts, M., Dhaene, J. y Denuit, M. (2008) Modern Actuarial Risk Theory Using R. Segunda Edición. Springer. Heidelberg, Alemania.

King, W. (1915), Accident and Health Insurance form an Actuarial ViewPoint. Proceedings of the Casualty actuarial Society. Volume II. Pp. 49-60.

Klugman, S. A. and Rioux, J. (2006). Towards a Unified Approach to Fitting Loss Models. North American Actuarial Journal, Volume 10, Number 1. Illinios.

Klugman, S. A., Panjer, H. H. and Willmot, G. (2004). Loss Models From Data to Decisions. Wiley-Interscience. 2nd. Edition. Canadá. Pp. XVII, 136-137.

Larose G. (1982). A Note on Loss Distributions. Proceediings of the Casualty Actuarial Society. Pp. 15-29

Ledesma, R. (2008). Introducción al Bootstrap. Desarrollo de un ejemplo acompañado de software de apliación. Pp. 51-60. Tutorials in Quantitative Methods for Psychology, Col. 4 (2). Universidad Nacional de Mar de Plata, Argentina.

Markham, P. (1962). An Introduction to Collective Risk Theory and its Application to Stop-Loss Reinsurance. Society of Actuaries. Estados Unidos.

Mascareñas, J. (1998). Introducción al VaR. Universidad Computense de Madrid. P. 2. Madrid, España.

Mata, A. J. Fannin, B. Verheyen, M. A. (2002). Pricing excess of loss treaty with loss sensitive features. An exposure rating approach. General Insurance Convention 2002. Pp. 1,5. London, Reino Unido.

Meyers, G. y Schenker N. (1983). Parameter Uncertainty in the Collectieve Risk Model. Casualty Actuarial Society. Volume LXX, No. 133-134. Virginia.

Mikosch, T. (2006). Non-Life Insurance Mathematics. And introduction with Stochastic Processes. Copenhagen: University of Copenhagen. Pp. 87-118.

Miranda, A. (2003). Tesis de Bioestadística: El Método de Remuestreo y su Aplicación en la Investigación Biomédica. P. 1.Escuela Nacional de Salud.

Mohamed, A. Razali, A. M. and Noriszura, I. (2010). Approximation of Aggregate Losses Using Simulation. Journal of Mathematics and Statistics 6 (3). Bangi, Selangor.

Muñoz, J. (2006). Determinantes en los siniestros de los seguros de gastos médicos mayores grupo y colectivo. Comisión Nacional de Seguros y Fianzas.

Panjer, H. H. y Willmot, G. E. (1992), Insurance Risk Models. Society of Actuaries. Estados Unidos.

Papush, D. E. Patrik, G. and Podgaits, F. (2001). Approximations of the Aggregate Loss Distribution. Casualty Actuarial Society Forum. Pp. 175. Arlington, Virginia.

Patrik, G. (1980). Estimating Casualty Insurance Loss Amount Distributions. Proceedings of the Casualty Actuarial Society, LXVII, Pp. 57-90.

Peng, Jin. (2010). Value at Risk and Tail Value at Risk in Uncertain Enviroment. Huanggang Normal University. P.787. Hubei, China.

Quenouille, M. (1949). Approximation Test of Correlation in Time Series. J. R. Statist. Soc. Pp.18-84.

Schipper N. (2010). Non-life unsurance risk models under inflation. Copenhagen Business School.

Shao, Jun y Tu, Dongsheng (1995). The Jackknife and Bootstrap. Pp. 1-9. Springer. Nueva York.

Sundt B., Jewell W. S. (1981). Further results of recursive evaluation of compound distributions. ASTIN Bulletin. Pp. 27-39.

Tilley, James. (1980). Discussion of Harry Panjer,The Aggregate Claim Distribution and Stop-Loss Reinsurance. p. 538. TSA XxX11.

Tse, Yiu-Kuen. (2009) Nonlife Actuarial Models Theory, Methods and Evaluation. Pp. 4-5, 87-89. Cambrige University Press. New York.

Verrall, R. J. (1989). The individual risk Model: A Compound Distribution. Journal of the Institute of Actuaries. Londres, Inglaterra.

Wagner, H. (1969). Principles of Operations Research with Aplications to Managerial Decisions. Prentice-Hall, Inc.

Wang, S., Young, V.R. and H.H. Panjer (1997). Axiomatic characterization of insurance prices. Insurance : Mathematics and Economics 21, 173-183.