Universidad de Costa Rica

Distribuciones de Pérdidas

Pérdidas ligadas a los daños a la propiedad y a las personas en aeropuertos de Estados Unidos

Autores
David Zumbado
Leonardo Blanco
Ignacio Barrantes

3 de diciembre de 2022

Indice de contenidos

Listado de Figuras

Listado de Tablas

1 Bitácora 1

1.1 Parte de planificación

1.1.1 Sección 1.

Para esta bitácora hemos estado usando rmarkdown y por alguna razón no está renderizando bien el signo de pregunta invertido. Estamos concientes del problema y en proceso de solucionarlo.

1.1.1.1 1- Nombres de los integrantes

El grupo de trabajo estará integrado por:

- Ignacio Barrantes Valerio, carné B50939
- Leonardo Blanco Villalobos, carné B71139
- David Zumbado Fernández, carné B88751

1.1.1.2 2- Idea

Se buscará modelar las pérdidas ligadas al extravío de equipajes en aeropuertos a partir de su frecuencia y severidad.

1.1.1.3 3- Reformulación de la idea de investigación

1.£Cómo se puede modelar las pérdidas ligadas al extravío de equipajes en aeropuertos a partir de su frecuencia y severidad?

Justificación:

Las pérdidas se pueden modelar utilizando la base proveniente del departamento de seguridad nacional de Estados Unidos, quienes manejan la seguridad en más de 400 aeropuertos del país a través del TSA (transportation security Administration). De esta manera, los datos serán obtenidos directamente del

aparato gobernamental de los Estados Unidos. Además, esta base ha sido usada en trabajos similares como los de Flores (2022) y Chen (2020) (ética). Según Flores (2022), se puede encontrar las distribuciones marginales de cada tipo de reclamo utilizando MLE, luego se pueden incorporar en un modelo de cópulas multivariadas, y finalmente se pueden evaluar utilizando las medidas de riesgo como el VaR y el TVaR (lógica). La cantidad de personas que se ven afectadas por extravíos o daños al equipaje durante el chequeo en los aeropuertos es significante y representan costos que deben ser previstos para mantener el buen funcionamiento de la empresa y para que se puedan hacer los pagos correspondientes a los afectados (emocional).

2.£Cuáles distribuciones probabilísticas permiten modelar las pérdidas ligadas al extravío de equipajes en aeropuertos a partir de su frecuencia y severidad?

Justificación:

La base utilizada sería la misma por lo que la justificación ética se mantiene. Para la parte lógica , se puede seguir una metodología parecida a la del punto anterior, sin embargo se tendrá que hacer más énfasis en la escogencia de la distribución para cada variable aleatoria de pérdida y frecuencia, y también a la hora de escoger la cópula. En Flores (2022), esto se hace a través de la estimación del MLE para diferentes distribuciones como la Poisson, Geométrica, binomial y luego se comparan usando la prueba ^{2} de bondad de ajuste para la frecuencia. Un proceso similar se lleva a cabo con la severidad, comparando Log-Laplace, Johnson SU, Logística generalizada y Lognormal. Un proceso similar se lleva a cabo con las cópulas al comparar la Frank, Clayton y Gumbel utilizando la métrica empírica UTDC para compararlas. Entonces bajo esta pregunta habrá un énfasis más fuerte en la exploración de estas y otras distribuciones para tratar de hallar la que mejor se ajusta en cada caso. Se puede comparar y contrastar con los hallazgos de otras personas que han hecho un trabajo similar (lógica). La justificación emocional es la misma.

3.£Por qué es importante modelar las pérdidas ligadas al extravío de equipajes en aeropuertos?

Justificación:

En general, el modelado de las pérdidas es de vital importancia ya que permite a las empresas, entidades financieras y aseguradoras a tener reservas para lograr mitigar el impacto de estas.Tal y como establece Ondieki et al. (2018) para las aseguradoras poder liquidar los siniestros que puedan llegar a producirse es fundamental, por lo que es imperativo que se modele adecuadamente los datos históricos y actuales sobre la experiencia de los siniestros, permitiendo de esta forma proyectar las experiencia de los siniestros futuros esperados y establecer reservas suficientes.

Los métodos estadísticos ha implementar serían iguales o similares a los mencionados en las propuestas anteriores de investigación. No obstante, en esta propuesta el enfoque es el de comparar el impacto positivo (importancia) en las finanzas del ente responsable de llevar a cabo los pagos por reclamos al contar con un modelo que le permita poseer reservas para ser frente a dichos pagos, en contraste con la de no hacer un estudio de pérdidas (lógico).

Los datos fueron tomados de una base de datos reales, esta proviene de US Terminal Security Agency (TSA) la cual registra los reclamos efectuados por los usuarios del transporte aéreo en Estados Unidos.

Esta base de datos ha sido ampliamente usada en diversos estudios entre ellos los de Flores (2022) y Chen (2020) confirmando así la validez de los datos. Con el fin de verificar los resultados obtenidos, se aplicaran medidas y pruebas estadísticas (ética).

En particular, poder contar con un modelo que permita modelar las pérdidas para hacer frente a los reclamos ligados a extravíos de equipajes en aeropuertos permitiría al ente responsable hacer frente con dichos montos de reclamo al contar con la reserva suficiente (emocional).

4.£Cuáles métodos no paramétricos pueden emplearse en la modelación de las pérdidas ligadas al extravío o daños de equipajes en aeropuertos?

Justificación:

En la revisión de la literatura se ubican dos fuentes que emplean la misma base de datos que la propuesta en la presente investigación y que persiguen el mismo objetivo, modelar la distribución de los costos de los reclamos, que son los trabajos de Flores (2022) y Chen (2020). De esta manera, se comprueba que la fuente de la base de datos ha sido validada antes en investigaciones de corte académico y estrechamente relacionadas, además de estar adecuadamente referenciada y poder consultarse en Kelly & Wang (2020) (ética). Se encontró que ambos trabajos utilizan métodos paramétricos; por esta razón, resulta de interés explorar también métodos no paramétricos alternativos que puedan llegar a usarse en el contexto de modelación de pérdidas en una aseguradora, por ejemplo (emocional). Un método no paramétrico que puede utilizarse es la estimación de densidades por medio de kernels (lógica). En Pitt et al. (2011), se advierte que este método suele ser inadecuado en presencia de asimetría, por lo cual, si se llegase a comprobar dicha condición, una manera de proceder es aplicar una transformación previa a los datos, concretamente una perteneciente a la *shifted power transformation family* y aplicar la estimación por kernels a los datos transformados, obteniéndose la densidad estimada de los datos originales mediante un proceso de inversión explicado en el mismo artículo.

1.1.1.4 Sección 5.

Fuente de Información:

Los datos se obtuvieron del Departmento de Seguridad Nacional, un organismo del gobierno de Estados Unidos y se puede encontrar en Homeland Security (2015)

Contexto temporal y espacial de los datos:

La base registra la ocurrencia de reclamos entre 2002 y 2015 en 466 aeropuertos alrededor de Estados Unidos.

Facilidad de obtener la información:

La base fue extraída de la página oficial del departamento de seguridad nacional la cual es accesible por cualquier persona por lo que se considera fácil de obtener.

Población de estudio:

Los reclamos realizados a aeropuertos de Estados Unidos.

Muestra observada:

Reclamos registrados por el TSA por daños realizados a los pasasjeros durante los chequeos de seguridad en Aeropuertos estadounidenses.

Unidad estadística o individuos:

La unidad estadística es el registro de una ocurrencia de un reclamo.

Descripción de las variables de la tabla:

Los datos se conforman por 13 variables: claim_number es el identificador del reclamo, date_received es la fecha que se registró el reclamo, incident_date es la fecha que ocurrió el incidente que ameritó el reclamo, airport_code son las 3 letras que identifican el aeropuerto donde ocurrión el incidente, y airport_name es el nombre del aeropuerto. Claim_type es el tipo de daño ocasionado (daño a propiedad, daño a personas, entre otras), claim_site es el lugar dentro del aeropuerto donde sucedió el incidente. Item es el ítem que sufrió el daño, claim_amount es la cantidad en dólares que la persona pide, status es el estado del reclamo (se llegó a un acuerdo, se negó, etc...), y close_amount fue el monto que efectivamente se pagó.

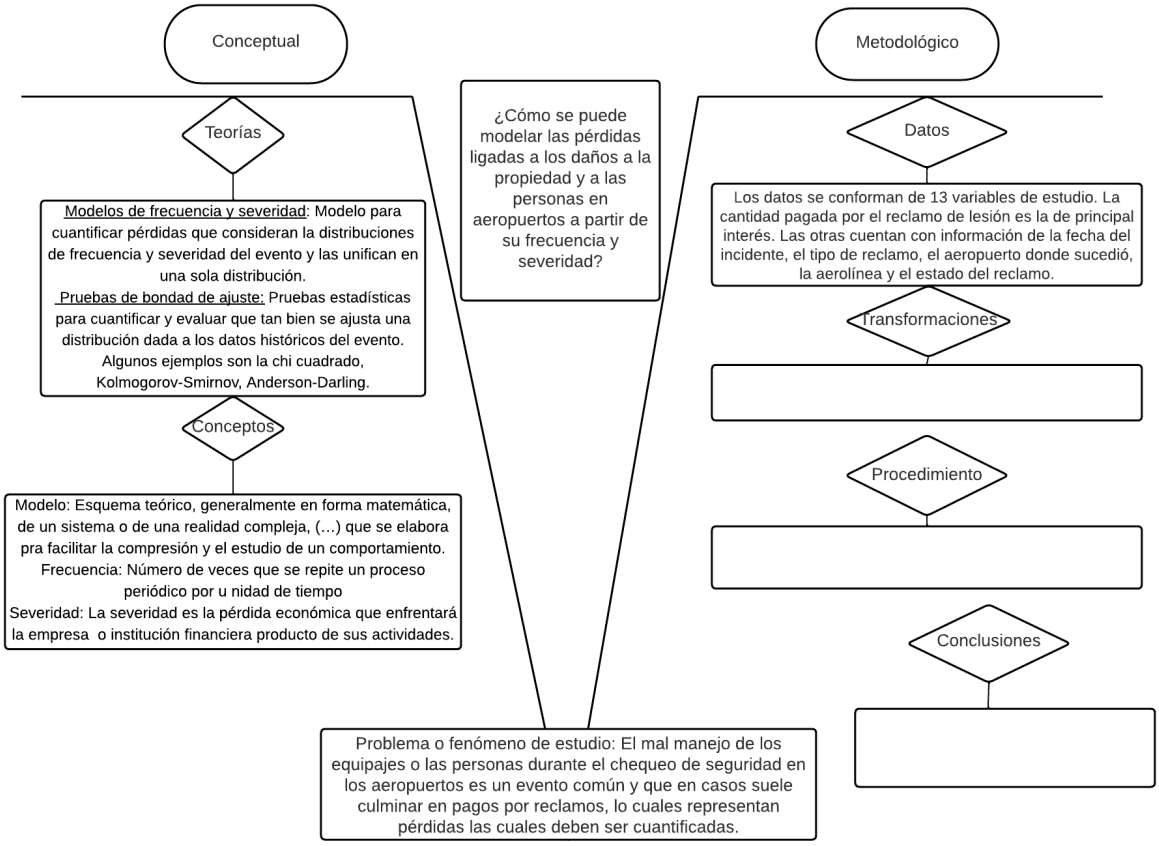
1.1.1.5 6. UVE Heurística

Objeto de estudio: El mal manejo de los equipajes o las personas durante el chequeo de seguridad en los aeropuertos es un evento común y que en casos suele culminar en pagos por reclamos, lo cuales representan pérdidas las cuales deben ser cuantificadas.

conceptos que delimitan la pregunta:

1. Modelo: Según (**RAE?**), un modelo es un esquema teórico, generalmente en forma matemática, de un sistema o de una realidad compleja, (...) que se elabora pra facilitar la compresión y el estudio de un comportamiento.
2. Frecuencia: Según Feldman & Brown (2005) es el número de pérdidas que se producirán en un periodo determinado. Esta variable aleatoria del número de pérdidas se denomina comúnmente frecuencia de pérdidas y su distribución de probabilidad se llama distribución de frecuencia.
3. Severidad: Según Feldman & Brown (2005) esta es la variable aleatoria del monto de la pérdida, dado que una pérdida ha ocurrido. Este monto suele denominarse severidad, y la distribución de probabilidad para el monto de la pérdida se denomina distribución de severidad.

Figura 1.1: Borrador de la UVE Heurística



1.1.1.6 Sección 7. Descripción detallada de la tabla de datos

Claim Number: Es una variable de tipo *string* que indica el identificador del reclamo, cada vez que una persona procede a efectuar un reclamo se le asigna este.

Date Received: Es una variable de tipo caracter, sin embargo para efectos de estudio debe ser transformada a una variable tipo fecha. El objetivo de esta variable es registrar el momento donde se realiza el reclamo en el siguiente formato: día-mes-año.Hay un total de 263 NA , es decir donde se registro reclamo respectivo pero no así la fecha. Esta variable es importante para nuestro estudio,ya que se debe tomar en cuenta el numero de reclamos al modelar las perdidas por frecuencia.

Incident Date: Se observa una diferencia de fechas desde el momento que se lleva acabo el incidente y el momento de reclamo correspondiente. Por esta razón se registra a fecha del incidente.

Esta es una variable de tipo caracter , la cual para efectos de estudio debe ser transformada a una variable tipo fecha. Esta registra el momento en que se lleva a cabo el incidente siguiendo el siguiente formato: día/mes/año.Hay un total de 2183 NA , es decir , se llevo a cabo el registro del reclamo pero no se tiene información de la fecha en que se llevo a cabo el incidente.

Para efectos del modelado de la perdida por frecuencia la variable de mayor interés para nuestro estudio es la ya mencionada *Date Received*.

Airport Name: Esta variable es una variable de tipo *string* categórica , hay un total de 466 Aeropuertos registrados en la base de datos, un total de 8524 NA y 441 reclamos donde no se especifica el nombre del Aeropuerto (se les establece el símbolo : -).

En esta se registra el nombre del aeropuerto donde se lleva acabo el incidente.Esta variable aleatoria es importante , ya es posible saber cuales son aquellos aeropuertos donde se presenta mayor número de reclamos.

Airport Code: Los códigos de aeropuertos están formados por grupos de tres letras, que designan a cada aeropuerto del mundo y asignadas por la Asociación Internacional de Transporte Aéreo.

Esta variable es de tipo caracter, y registra dicho código del aeropuerto donde se lleva a cabo el incidente por lo que aporta la misma información que la variable aleatoria *Airport Name* , para efectos del estudio es importante considerar eliminar alguna de estas dos variables.

Airline Name: Es una variable que registra la aerolínea en la que viajaba la persona que sufrió el incidente. Es una variable tipo *string* categórica. Hay un total de 232 aerolíneas registradas en la base de datos, un total de 34374 NA y 4247 reclamos donde no se especifica el nombre de la aerolínea (se les establece el símbolo : -).

Esta variable es de interés para el desarrollo de nuestro estudio ya que es importante saber cuales son las aerolíneas que presentan mayor número de reclamos, y por ende si tienen una mayor impacto en las pérdidas.

Claim type: Esta variable es de tipo *string* categórica, la cual registra el tipo de reclamo realizado por la persona. Hay un total 10 tipos de reclamos registrados en la base de datos:

- *Bus Termina* : Categoría que registra reclamos relacionados a la terminal de buses, hay un 1 reclamo.
- *Complaint*: Categoría que registra reclamaciones de forma general, hay un total de 49 reclamos.
- *Compliment*: Categoría que registra quejas de forma general, hay un total de 3 reclamos.
- *Employee Loss (MPCECA)*: Categoría que registra reclamos por pérdidas de empleados (MPCECA).Es decir, reclamos de perdida realizados por los mismos empleados. Hay un total de 485.
- *Motor Vehicle*: Categoría que registra reclamos asociados vehiculos automotores , hay un total de 369 reclamos.

- *Passenger Property Loss* : Categoría que registra reclamos por pérdidas de bienes de los pasajeros. Esta es la categoría con mayor número de reclamos con un total de 117868 reclamos.
- *Passenger Theft*: Categoría que registra reclamos asociados robos realizados a bienes de los pasajeros , hay un total de 465 reclamos.
- *Personal Injury* : Categoría que registra reclamos asociados a daños personales (siendo este un término legal para una lesión al cuerpo, la mente o las emociones, en contraposición a una lesión a la propiedad) hay un total de 1465 reclamos.
- *Property Damage*: Categoría que registra reclamos asociados daños a la propiedad. Esta es la segunda categoría con mayor cantidad de reclamos hay un total de 75364 reclamos.
- *Wrongful Death*: Categoría que registra reclamos asociados a muerte por negligencia , es un reclamo contra una persona que puede ser considerada responsable de la muerte de otra persona. Hay un total de 4 reclamos.

Finalmente, para esta variable hay un total de 7913 NA y 282 reclamos donde no se especifica el tipo de reclamo (se les establece el símbolo : -).

Esta variable es de interés , puesto que puede existir una relación directa con el monto de los reclamos , y por ende tener impacto sobre la perdida por severidad.

Claim Site: Esta variable es de tipo string categórica, la cual indica el sitio del reclamo. Hay un total de 5 categorías registras para esta variable: *Bus Station* , *Checked Baggage*, *Checkpoint*, *Motor Vehicle* y otra categoría llamada *others*.

Se observa que las categorías con mayor número de reclamos son: *Checked Baggage* con 159753 reclamos , y *Checkpoint* con 40133 reclamos. Además, hay un total de 740 NA y 276 reclamos donde no se especifica el nombre del Aeropuerto (se les establece el símbolo : -).

Item: Esta variable es de tipo *string* ,la cual se encarga de describir el motivo de reclamo (bien material perdido, daño material sufrido, daño personal).

Claim Amount: Esta variable es de tipo numérico,la cual se encarga de registar el monto de reclamo.Es decir, el monto solicitado por la persona que sufrió el incidente. Hay un total de 4043 NA y 12752 reclamos donde no se especifica el monto de reclamo (se les establece el símbolo : -).

Status: Esta variable indica el estado intermedio del reclamo, es de tipo *string* categórica. Al no ser el status final del reclamo cuenta con una cantidad considerable de categorías (14 registradas en la base de datos, 5 NA y 12752 reclamos donde no se especifica el monto de reclamo donde se les establece el símbolo : -) las cuales se terminan asignando en alguna de las tres categorías de la variable *Disposition* que a continuación se describe.

Disposition: Esta variable a diferencia de *Status* muestra la disposición , es decir el acuerdo final sobre el reclamo. Es una variable de tipo string categórica , con tres categorías: *Approve in Full*, *Deny* y *Settle*. A continuación, se describe cada una de estas:

- *Approve in Full* : Esta categoría registra aquellos reclamos cuyo *Claim Amount* (monto reclamado por la persona perjudicada) fue aprobado de forma total, es decir, el monto acordado a pagar (*Close Amount*) es igual al *Claim Amount*. Hay un total de 35010 reclamos aprobados de forma completa.
- *Deny*: Esta categoría registra los reclamos denegados, es decir aquellos reclamos donde no se paga por el reclamo. Hay un total de 68382 reclamos denegados.
- *Settle*: En esta categoría se registran aquellos reclamos cuyo *Close Amount* (monto a pagar) , es menor al *Claim Amount* (Monto de reclamo). Es decir, son aquellos reclamos en cuyo acuerdo final se estableció un monto inferior de pago por el incidente.

Close Amount: Esta variable es de tipo numérica, y es el monto final acordado por ambas partes.Es decir, es el monto que debe ser pagado a la persona como producto del reclamo realizado. Es importante señalar que este monto es igual o inferior al *Claim Amount* y depende de la variable disposición que se describió anteriormente. Esta variable es relevante para nuestro estudio pues esta relacionado de forma directa con la severidad, y por ende, con la perdida por severidad.

1.1.1.7 Sección 8. Literatura

1. **Título**: Modelling Dependencies in Airport Passenger Claim Data Using Copulas (Flores, 2022)
   Autor: Roberto Carcache Flores

   Nombre del tema: Modelación del riesgo utilizando cópulas

   Forma de organizarlo:

   - Cronológico: Febrero 2022
   - Metodológico: Cópulas bivariadas y multivariadas y simulaciones
   - Temático: Funciones de distribución y dependencia de variables aleatorias
   - Teoría: Probabilidad y estadística

   Resumen en una oración: Se encuentra la mejor distribución para la severidad y frecuencia de cada reclamo y luego estas distribuciones marginales se incorporan en diferentes modelos de cópulas

   Argumento central: En la metodología tradicional del modelamiento del riesgo se asume independencia entre frecuencia y severidad, lo cual no se hace en esta investigación. Además, se utiliza un proceso de eliminación de la tendencia con respecto al tiempo para mejorar los resultados.

Problemas con el argumento o el tema: Las medidas de riesgo utilizando cópulas resultan en medidas de riesgo más altas que en los datos históricos.

Resumen en un párrafo: Se eliminan los reclamos que fueron negados justificando el hecho de que el punto de la investigación es cuantificar los pagos que efectivamente fueron hechos, además del gran volumen de los datos. La agregación de los datos se hace por mes y con suma para la severidad y por frecuencia de los reclamos. El autor nota que hay un tendencia negativa de la frecuencia y severidad con respecto al tiempo por lo que procede a eliminar la tendencia. Luego determina la mejor marginal para cada variable utilizando MLE. Se encuentra que la binomial negativa se ajusta mejor a las frecuencias. Por otro lado la Log-Laplace se ajusta mejor a los reclamos por daños a la propiedad y la lognormal se ajusta mejor a los reclamos por pérdidas de los bienes por lo que se utilizan estas dos para modelar la severidad. Luego se procede a hacer algo similar con los resultados de eliminar la tendencia. Se encuentra que el proceso de eliminación de la tendencia facilita la búsqueda de una distribuciones. Se encuentra que todas las variables pares muestran algún tipo de dependencia en las colas. Finalmente, las cópulas multivariadas se comparan utilizando log verosimilitud y se obtiene que las cópulas elípticas (Gaussiana y t-Student) se ajustan mejor que las arquimedianas (Clayton y Gumbel).

2. **Título**: Aggregate loss model with Poisson - Tweedie loss frequency (Chen, 2020)

   Autor: Si Chen.

   Nombre del tema: Modelado perdidas usando la familia de distribuciones Poisson - Tweedie .

   Forma de organizarlo:

   - Cronológico: Año 2020.
   - Metodológico: Modelado de la frecuencia de perdida a partir de una distribución Poisson-Tweedie , simulaciones y modelado de perdida agregada.
   - Temático: Modelos de perdida agregada.
   - Teoría: Distribuciones de perdidas.

   Resumen en una oración: Uso de la familia de distribuciones Poisson-Tweedie con la finalidad de modelar la frecuencia de las perdidas y ver el impacto que tiene este sobre el modelo de perdidas agregadas.

   Argumento central: Pese a que el impacto de la perdida por severidad en un modelo de perdida agregada ha sido bien estudiado a través de los años, se ha prestado menos atención a la influencia de la perdida por frecuencia en dichos modelos, esto motiva el estudio de un modelo de perdidas por frecuencias no tradicional.

   Problemas con el argumento o el tema: Dado el estudio , no se pudo captar por completo las relaciones entre las perdidas por severidad , perdida por frecuencia y perdida agregada.

Resumen en un párrafo: En este estudio, se modela la perdida por frecuencia usando la familia de distribuciones Poisson-Tweedie, esto bajo el argumento que dichas familias presentan características como: el ajuste de la frecuencia de pérdidas es más flexible , reducen la posibilidad de una especificación errónea del modelo y dichas familias presentan una convolución cerrada. Mediante estudios de simulación , se investiga y encuentra el impacto de una mala especificación de la distribución perdida de la frecuencia al cuantil de perdidas agregadas, así como el sesgo del estimador de máxima verosimilitud del índice de la familia de Poisson-Tweedie.

3. **Título**: *Estimation of Parametric and Nonparametric Models for Univariate Claim Severity Distributions - an approach using R* (Pitt et al., 2011)

   Autores: David Pitt, Montserrat Guillen y Catalina Bolancé

   Nombre del tema: Comparación de métodos paramétricos y no paramétricos apra modelar la severidad de reclamos en una aseguradora

   Forma de organizarlo:

   - Cronológico: mayo de 2011
   - Metodológico: estimación de densidades por Kernels modificados,
   - Temático: Modelación de reclamos métidos y de seguros de automóviles
   - Teoría: Probabilidad y estadística

   Resumen en una oración: Se encuentra que la estimación por kernels modificados es adecuada para modelar la distribución tanto de costos médicos como de reclamos en seguros de automóviles.

   Argumento central: Se pueden usar métodos no paramétricos para estimar distribuciones de reclamos en seguros de vehículo y de costos médicos.

   Problemas con el argumento o el tema: Los métodos clásicos de estimación de densidades por kernels suelen ser inadecaudos en presencia de asimetría, lo cual es común en datos de montos de reclamos en el contexto de seguros.

   Resumen en un párrafo: Se utilizan datos de costos de reclamos hechos a una aseguradora española por accidentes ocurridos en el año 2000 y recopilados en 2002, que incluye tanto los ligados a costos por daños a la propiedad como por costos médicos. El tamaño de muestra es de 518 reclamos. Para estimar la densidad para cada uno de los costos (daños a la propiedad y médicos) por separado, se utilizan métodos paramétricos y no paramétricos. Dentro de los paramétricos, se utilizaron aproximaciones normales y log-normales. Dentro de los no paramétricos, se utilizó una aproximación por kernels modificada, donde la modificación consiste en que primero se aplica una transformación a los datos originales para corregir la asimetría, se hace una aproximación con un kernel gaussiano a los datos modificados, y luego se calcula la aproximación de los datos originales a partir de la calculada para los modificados. La transformación aplicada a los datos se enmarca en la *shifted power transformation*

family. Para evaluar la bondad de ajuste de todas las estimaciones propuestas, se utilizan distintas versiones log-verosimilitud tanto la versión clásica como modificaciones ponderadas, mientras que para evaluar solamente los métodos no paramétricos se usan distintas versiones de una aproximación a errores cuadráticos integrados ponderados. Se concluye que la log-verosimilitud no es una buena medida de bondad de ajuste para comparar los ajustes no paramétricos, debido a su relación inversa con la magnitud del ancho de banda empleado. En general, de las propuestas paramétricas, la log-normal tuvo un mejor desempeño mientras que la estimación por kernel modificada tuvo un desempeño adecuado y se recomienda para modelar distribuciones con colas pesadas.

4. **Título**: *Estimation of Parametric and Nonparametric Models for Univariate Claim Severity Distributions - an approach using R* (Ondieki et al., 2018)

   Autores: Cyprian Ondieki, Shalyne Gathoni y Joan Wairimu

   Nombre del tema: Estimación de distribuciones de frecuencia y severidad en seguros de automóviles

   Forma de organizarlo:

   - Cronológico: febrero de 2018
   - Metodológico: Distribuciones continuas para la modelización de la severidad y discretas para la modelización de la frecuencia, donde los parámetros se estiman por máxima verosimilitud y los ajustes se miden con pruebas chi cuadrad, Kolmogorov-Smirnov, Anderson-Darling y los modelos se seleccionan de acuerdo a sus medidas de AIC y BIC.
   - Temático: Modelización de la frecuencia y severidad en seguros de automóviles
   - Teoría: Probabilidad y estadística, distribuciones probabilísticas, pruebas de bondad de ajuste, estimación por máxima verosimilitud

   Resumen en una oración: Se ponen a prueba varias distribuciones para estimar la distribución tanto de frecuencia como la severidad de reclamos de automóviles de tres bases de datos distintas.

   Argumento central: En el contexto de seguros de automóviles, la distribución lognormal es apropiada para modelizar la severidad y la binomial negativa y geométrica lo son para modelizar la frecuencia.

   Problemas con el argumento o el tema: Se advierte que pronósticos realizados con los modelos seleccionados pueden ser C:tiles solamente en el corto plazo. Además, no se consideran distribuciones de la clase $(a, b, 1)$.

   Resumen en un párrafo: Con tres bases de seguros de automóviles gratuitas en R (*AutoCollision*, *dataCar*, *dataOhlsson*) se proponen distribuciones continuas para la modelización de la severidad (Exponencial, Gamma, Pareto, Lognormal y Weibull) y discretas para la modelización de la frecuencia (Binomial, Geométrica, Binomial Negativa, Poisson), donde los parámetros se estiman por máxima

verosimilitud y los ajustes se miden con pruebas chi cuadrado (para la frecuencia) y Kolmogorov-Smirnov y Anderson-Darling (para la severidad), así como se usa el Criterio de Información de Akaike (AIC) y el Criterio de Información Bayesiano (BIC) para determinar el mejor modelo de los no descartados con las pruebas anteriores. Se concluye que la distribución que constituye el mejor modelo para la severidad es la lognormal, mientras que en cuanto a la frecuencia, las más adecuadas son la binomial negativa y la geométrica.

1.1.2 Teorías, principios o metodologías

Elección de los modelos de frecuencia y severidad:

Con la finalidad de modelar la frecuencia con la que ocurren los reclamos por extravío y la severidad de estos, se desea contar con los modelos que mejor se ajusten a nuestro estudio en cuestión.

No obstante, existe una serie de distribuciones de probabilidad estándar que se podrían utilizar para aproximar las distribuciones de las variables aleatorias de la frecuencia de reclamaciones y la severidad o monto de estas reclamos. Las distribuciones binomial, geométrica, binomial negativa y Poisson se consideran para la modelización de la frecuencia.

Por otro lado, entre las distribuciones estándar para modelar la severidad se tienen las siguientes distribuciones: exponencial, gamma, Weibull, Pareto y lognormal.

Tal como lo establece Ondieki et al. (2018) una forma de abordar la escogencia de la distribución correcta es ajustando los datos a las distribuciones estadísticas seleccionadas y los parámetros se estiman mediante el método de máxima verosimilitud.

Una vez ajustas las distribuciones a los datos y estimados los parámetros es posible hacer pruebas de bondad de ajuste para ambos modelos, y pruebas para elegir entre las distribuciones que compiten entre sí. A continuación se establecen cuales son estas pruebas.

Elección de los modelos de frecuencia y severidad:

Una prueba de bondad de ajuste es “un procedimiento estadístico que describe qué tan bien se ajusta una distribución a un conjunto de observaciones mediante la medición de la compatibilidad cuantificable entre las distribuciones teóricas estimadas y la distribución empírica de los datos muestrales” (Ondieki et al., 2018). Estas pruebas se pueden basar en la función de densidad o masa o en la función de distribución y adoptan la estructura de prueba de hipótesis donde la hipótesis nula consiste en que los datos siguen una distribución particular, mientras que la alternativa en que los datos no siguen dicha distribución particular.

Se presenta ahora una idea general de las tres pruebas de bondad de ajuste que se proponen para este análisis:

- Prueba Chi-Cuadrado de bondad de ajuste: Esta prueba propone un estadístico compuesto de frecuencias observadas y esperadas, calculado a partir de una partición de la muestra, el cual presenta bajo la hipóstesis nula una distribución Chi-Cuadrado con grados de libertad que dependen de la cantidad de datos, la cantidad de intervalos de la partición y la cantidad de parámetros de la distribución propuesta calculados por medio de los datos muestrales.
- Prueba Kolmogorov-Smirnov: Esta prueba se basa en comparar la función de distribución propuesta con la función de distribución empírica de los datos para medir el ajuste, partiendo de que la función de distribución caracteriza a una distribución de probabilidad. Esta comparación se realiza mediante un estadístico que mide la distancia entre ambas distribuciones, del cual se conocen ciertos resultados de convergencia y distribución que fundamentan la efectividad del método.
- Prueba Anderson-Darling: Esta prueba se asemeja a la de Kolmogorov-Smirnov pero mide de una forma distinta la diferencia en tre las funciones de distribución empírica y teórica. Además, de acuerdo a Klugman et al. (2019) el estadístico de prueba de Anderson-Darling suele priorizar un mejor ajuste en las colas de la distribución en comparación con las regiones más centrales.

En las bitácoras posteriores se ampliará en los aspectos técnicos de las pruebas anteriormente mencionadas.

1.2 Parte de escritura

1.2.1 Sección 1. Escogencia de la pregunta de investigación

La pregunta seleccionada es la primera: *£Cómo se puede modelar las pérdidas ligadas a los daños a la propiedad y a las personas en aeropuertos a partir de su frecuencia y severidad?*

Se cambió "extravío de equipajes" por "daños a la propiedad y a las personas" para poder abarcar el resto de eventos que aparecen en la base de datos.

1.2.2 Sección 2. Propuesta de argumentación

En general, el modelado de las pérdidas es de vital importancia ya que permite a las empresas, entidades financieras y aseguradoras tener reservas para lograr mitigar el impacto de dichas pérdidas. Tal y como establece Ondieki et al. (2018), para las aseguradoras poder liquidar los siniestros que puedan llegar a producirse esto es fundamental, por lo que es imperativo que se modelice adecuadamente los datos históricos y actuales sobre la experiencia de los siniestros, permitiendo de esta forma proyectar las experiencia de los siniestros futuros esperados y establecer reservas suficientes.

Una forma de realizar este ejercicio de modelación es utilizando una base proveniente del Departamento de Seguridad Nacional de Estados Unidos, quienes manejan la seguridad en más de 400 aeropuertos de

dicho país a través de TSA (Transportation Security Administration). En la revisión de la literatura se ubican dos fuentes que emplean esta base de datos y que persiguen objetivos similares, que son los trabajos de Flores (2022) y Chen (2020). De esta manera, se comprueba que la fuente de la base de datos ha sido validada antes en investigaciones de corte académico y estrechamente relacionadas con el tema del presente escrito, además de estar adecuadamente referenciada y poder consultarse en Kelly & Wang (2020).

Para contestar la pregunta de investigación, se propone seguir el procedimiento adoptado en Ondieki et al. (2018), en el se parte de un grupo de distribuciones continuas para la modelización de la severidad (Exponencial, Gamma, Pareto, Lognormal y Weibull) y otro de distribuciones discretas para la modelización de la frecuencia (Binomial, Geométrica, Binomial Negativa, Poisson), donde los parámetros se estiman por máxima verosimilitud y los ajustes se miden con pruebas Chi-Cuadrado (para la frecuencia) y Kolmogorov-Smirnov y Anderson-Darling (para la severidad). También, en dicha investigación se usa el Criterio de Información de Akaike (AIC) y el Criterio de Información Bayesiano (BIC) para determinar el mejor modelo de los no descartados con las pruebas anteriores. Adicionalmente, se contempla incorporar distribuciones truncadas y modificadas en búsqueda de ajustes superiores.

1.2.3 Sección 3. Resumen del problema hasta el momento.

El TSA (Transportation Security Administration) es la agencia establecida luego del 2001 que se ocupa del chequeo de los pasajeros y su equipaje en los aeropuertos de Estados Unidos. Como consecuencia de sus labores, es común que se causen daños y extravíos de las pertenencias de los pasajeros lo que resulta en reclamos por parte de los mismos en la forma de compensación monetario por los daños ocassionados. Estos pagos han sido registrados en la tabla que se utilizará para esta investigación, además de otra información pertinente a cada incidente reclamado. El propósito de este trabajo será el de modelar estas pérdidas para lograr cuantificarlas. Esto es importante pues el TSA tiene que tener previsto estos costos para poder continuar sus operaciones, y los pasajeros que sí son víctimas del mal manejo de sus pertenencias puedan conseguir su dinero devuelta.

Al consultar la literatura se ha visto que este es un problema que ha sido tratado por al menos dos trabajos anteriores en los cuales se basará esta investigación con el fin de expandir y poder comparar y contrastar los resultados que se obtienen con los de ellos. En general se han identificados varios pasos que se deberán seguir para poder lograr el cometido. Primero se busca la densidad apropiada para la frecuencia y severidad por separado mediante la estimación del parámetro MLE. Las densidades candidatas para la frecuencia son la binomial negativa, geométrica, Poisson y binomial. Por otro lado, las candidatas para la densidad de la severidad son la log-normal, log-laplace, Johnson SU y la logística generalizada. Sin embargo, a través del trabajo se investigará otras posibles distribuciones que pueden ser de utilidad. La bondad de ajuste para comparar las distribuciones anteriores se lleva a cabo utilizando pruebas como la chi cuadrado, Kolmogorov-Smirnov y Anderson Darling.

Una vez obtenidas las distribuciones se deben unificar en un modelo agregada que represente las pérdidas totales utilizando cópulas bivariadas o multivariadas. Algunos ejemplos de las que sea han utilizado son

la Clayton y la Gumbel. Estas son comparadas mediante la utilización de medida empírica UTDC para escoger la más apropiada para el problema en cuestión. Otra observación importante es que algunos autores noten que existe una tendencia de las pérdidas con respecto al tiempo por lo que el primer paso en realidad es la eliminación de dicha tendencia. Este paso puede ser importante pues notan que al comparar el proceso descrito anteriormente al haber eliminado la tendencia se lograban mejores resultados. Sin embargo, esta es otra capa de complejidad que será evaluada durante el proceso si se incluye o no.

Algunos otros hallazgos importantes aparte del impacto positivo que tiene la elimnación de la tendencia, es que encontró que la binomial negativa se ajusta mejor a las frecuencias, mientras que la Log-Laplace se ajusta mejor a la severidad de los daños a la propiedad, y la Log-lognormal a los extravíos. Luego para las cópulas se encuentra que cópulas elípticas (Gaussiana y t-Student) se ajustan mejor que las arquimedianas (Clayton y Gumbel).

2 Bitácora 2

Para esta bitácora se decidió mudar el análisis a otra base de datos (también de reclamos a TSA), pues al revisar con más detalle la anterior, que comprendía datos del periodo 2002-2015, como parte del análisis descriptivo se notó que los datos de la variable más importante para este trabajo, que es *close_amount* (monto final pagado por cada reclamo), no estaba presente del todo a partir del año 2010. Esto marca una inconsistencia ya que al revisar los archivos de TSA para el periodo 2010-2013 se comprobó que los datos para la mencionada variable sí estaban disponibles. Por esta razón, se decidió trabajar con esta segunda base de datos, es decir la que contempla solamente de 2010 a 2013 y en lo sucesivo los análisis se refieren a este periodo de menor duración.

2.1 Ordenamiento de la literatura

Tabla 2.1: Ordenamiento de la literatura

Tipo de grupo	**Nombre del grupo**	**Nombre del tema**	**Título**	**Año**	**Autor (es)**
Metodológico	Modelo de cópulas	Modelación de distribuciones de pérdidas	Modelling Dependencies in Airport Passenger Claim Data Using Copulas	2022	Roberto Carcache Flores
Metodológico	Modelos de pérdidas agregadas	Modelación paramétrica de distribuciones de pérdidas	Aggregate loss model with Poisson - Tweedie loss frequency	2020	Si Chen
Metodológico	Estimación de densidades por kernels	Modelación no paramétrica de distribuciones de pérdidas	Estimation of Parametric and Nonparametric Models for Univariate Claim	2011	David Pitt, Montserrat Guillen y Catalina Bolancé
Metodológico	Modelos de frecuencia y severidad	Modelación de las distribuciones de frecuencia y severidad	Severity Distributions - an approach using R	2018	Cyprian Ondieki, Shalyne Gathoni y Joan Wairimu

2.2 Enlaces de la literatura

En Flores (2022) se establece el procedimiento base para conseguir la distribución agregada al igual que algunos hallazgos y metodologías que son de alta utilidad. Primero la agregación de los datos se hace mensualmente con suma para la severidad y por frecuencia para los reclamos. El autor nota que hay un tendencia negativa de la frecuencia y severidad con respecto al tiempo por lo que procede a eliminarla. Luego, determina la mejor distribución para cada variable utilizando estimación de máxima verosimilitud (MLE). Se encuentra que la binomial negativa se ajusta mejor a las frecuencias. Por otro lado, la Log-Laplace se ajusta mejor a los reclamos por daños a la propiedad y la lognormal se ajusta mejor a los reclamos por pérdidas de los bienes, por lo que se utilizan estas dos para modelar la severidad. Durante este proceso el autor nota que la eliminación de la tendencia facilita el proceso de ajustar una distribución a la frecuencia y la severidad. Finalmente, las cópulas multivariadas se comparan utilizando log_añoverosimilitud y se obtiene que las cópulas elípticas (Gaussiana y t-Student) se ajustan mejor que las arquimedianas (Clayton y Gumbel).

En un estudio similar, Pitt et al. (2011) utilizan datos de costos de reclamos hechos a una aseguradora española por accidentes ocurridos en el año 2000 y recopilados en 2002, que incluye tanto los ligados a costos por daños a la propiedad como por costos médicos. El tamaño de muestra es de 518 reclamos. Al igual que el estudio anterior, para estimar la densidad para cada uno de los costos (daños a la propiedad y médicos) se utilizan métodos paramétricos como las aproximaciones normales y log-normales. En contraste al estudio pasado también recurren a estimadores no paramétricos como la aproximación por kernels modificada, donde la modificación consiste en que primero se aplica una transformación a los datos originales para corregir la asimetría, se hace una aproximación con un kernel gaussiano a los datos modificados, y luego se calcula la aproximación de los datos originales a partir de la calculada para los modificados. La transformación aplicada a los datos se enmarca en la *shifted power transformation family*.

Adicionalmente, los mismos autores exponen métodos para evaluar la bondad de ajuste de las distribuciones encontradas. Para evaluar todas las estimaciones propuestas se utiliza la log-verosimilitud, tanto la versión clásica como modificaciones ponderadas, mientras que para evaluar solamente los métodos no paramétricos se usan distintas versiones de una aproximación a errores cuadráticos integrados ponderados. Se concluye que la log-verosimilitud no es una buena medida de bondad de ajuste para comparar los ajustes no paramétricos, debido a su relación inversa con la magnitud del ancho de banda empleado. En general, de las propuestas paramétricas, la log-normal tuvo un mejor desempeño, el cual es un hallazgo que concuerda con el de Flores (2022), mientras que la estimación por kernel modificada tuvo un desempeño adecuado y se recomienda para modelar distribuciones con colas pesadas.

El modelado de pérdidas agregadas es una técnica estadística ampliamente utilizada en el ámbito actuarial, cuyo objetivo es la obtención de una función de distribución de perdidas agregadas, a partir de la distribución de frecuencia de reclamos, y de la distribución de la severidad de estos.

Un claro ejemplo de la implementación de esta técnica es el estudio realizado por (Chen, 2020) . La principal motivación de este estudio es la de modelar la frecuencia de las pérdidas mediante el uso de la familia de

distribuciones Poisson-Tweedie con la finalidad de modelar la frecuencia de las perdidas y ver el impacto que tiene este sobre el modelo de pérdidas agregadas.

Esto bajo el argumento que dichas familias presentan características como: el ajuste de la frecuencia de pérdidas es más flexible , reducen la posibilidad de una especificación errónea del modelo y dichas familias presentan una convolución cerrada.

Mediante el uso de la distribución de la familia Poisson-Tweedie y el estudio de simulación basados en: Percentil de la distribución de pérdidas agregadas bajo diferentes distribuciones de frecuencia de pérdidas (diferentes valores del parámetros de la familia) y la investigación de estimadores de parámetros para frecuencia de pérdidas vía simulaciones de Monte Carlo, se investiga y encuentra el impacto de una mala especificación de la distribución perdida de la frecuencia al cuantil de pérdida agregadas, así como el sesgo del estimador de máxima verosimilitud del índice de la familia de Poisson-Tweedie.

Una de las principales diferencias de los métodos implementados en el estudio realizado por (Chen, 2020) es el uso de máxima verosimilitud y la implementación de simulaciones vía Monte Carlo. A diferencia de los métodos empleados por (Pitt et al., 2011) donde su estudio se centra en la comparación entre métodos paramétricos tradicionales, y métodos no paramétricos basados en la estimación de densidades por Kernels modificados.

No obstante, pese a que según (Pitt et al., 2011) se logra estimar de forma adecuada la distribución tanto de costos médicos como de reclamos en seguros de automóviles, los métodos clásicos de estimación de densidades por kernels suelen ser inadecuados en presencia de asimetría, siendo esto habitual en datos de montos de reclamos.

Pese a que el método de estimación de densidades por kernels es técnicamente más sencillo que la implementación de los métodos utilizados por (Chen, 2020) , es importante tomar en consideración los problemas presentes en el estudio realizados por (Pitt et al., 2011) , ya que pueden generar dificultades técnicas importantes.

Además, el uso de máxima verosimitud por parte de (Chen, 2020) para la escogencia de los parámetros es un método de uso más frecuente, para resolver problemas de esta índole.

Sin embargo, debido al enfoque de hace uso de una familia particular para modelar la frecuencia.No obstante, es importante señalar que existen test y pruebas especificas para escoger las distribuciones más adecuadas dada la base de datos de un estudio en particular. Estas técnicas estadísticas para la escogencia de las mejores distribuciones tanto de a frecuencia como la severidad son empleadas por (Ondieki et al., 2018) .

(Ondieki et al., 2018) hace uso de tres bases de seguros de automóviles gratuitas en R (AutoCollision, dataCar, dataOhlsson), en su estudio propone el modelado de la severidad mediante distribuciones continuas (Exponencial, Gamma, Pareto, Lognormal y Weibull) y discretas (Binomial, Geométrica, Binomial Negativa, Poisson) para el caso de la frecuencia, donde los parámetros se estiman vía máxima verosimilitud y los ajustes se miden con pruebas chi cuadrado (para la frecuencia) , Kolmogorov-Smirnov y Anderson-Darling_año(para la severidad).

Tabla 2.2: Primeras cinco filas de la tabla de datos

claim_number	date_received	incident_date	airport_code	airport_name	airline_name	claim_type	claim_site	item_category	close_amount	disposition
2.010011e+12	2010-01-04	2010-01-03 14:30:00	SLC	Salt Lake City International Airport	Delta Air Lines	Property Damage	Checked Baggage	Cosmetics & Grooming	0	Deny
2.010011e+12	2010-01-04	2010-01-02 00:00:00	LAX	Los Angeles International Airport	Southwest Airlines	Passenger Property Loss	Checked Baggage	Other	0	Deny
2.010011e+12	2010-01-04	2010-01-02 05:00:00	SEA	Seattle-Tacoma International	Delta Air Lines	Passenger Property Loss	Checked Baggage	Cameras; Cameras	0	Deny
2.010011e+12	2010-01-04	2010-01-01 00:00:00	DEN	Denver International Airport	Southwest Airlines	Passenger Property Loss	Checked Baggage	Clothing	NA	-
2.010011e+12	2010-01-04	2010-01-02 00:00:00	LAS	McCarran International	American Airlines	Passenger Property Loss	Checked Baggage	Travel Accessories	0	Deny
2.010011e+12	2010-01-04	2010-01-03 00:00:00	DFW	Dallas-Fort Worth International Airport	American Airlines	Passenger Property Loss	Checked Baggage	Travel Accessories	0	Deny

Una vez obtenidos los parámetros y realizadas las pruebas de ajuste, se seleccionan los modelos de acuerdo a sus medidas del Criterio de Información de Akaike (AIC) AIC y el Criterio de Información Bayesiano (BIC).

Se concluye que la distribución que constituye el mejor modelo para la severidad es la lognormal, mientras que en cuanto a la frecuencia, las más adecuadas son la binomial negativa y la geométrica.

A diferencia del estudio realizado por (Ondieki et al., 2018) en el cual usa conjuntamente métodos paramétricos y no paramétricos con el objetivo de compara estos, en este caso (Ondieki et al., 2018) se enfoca en utilizar un método paramétrico ampliamente utilizado como lo es la estimación de parámetros vía máxima verosimilitud.Sin embargo, se enfoca en implementar una gran variedad de pruebas, test y métricas para obtener las mejores distribuciones posibles tanto de la frecuencia como la severidad.

Es importante señalar que en el estudio de (Ondieki et al., 2018) no se realiza ninguna técnica para encontrar una distribución de perdidas agregadas, a diferencia del estudio de (Chen, 2020) donde si construyen esta distribución agregada.

Pese a que las técnicas estadísticas implementadas por Ondieki et al. (2018) son las tradicionales, a diferencia de los otros estudios mencionados en este apartado, sí se tiene como objetivo escoger las mejores distribuciones para la frecuencia y severidad para nuestro estudio, es prudente seguir una línea de investigación similar a las empleadas en este estudio. El implementar métodos no paramétricos como el de densidades por kernels puede traer complejidades técnicas. Además, el uso de una sola familia en particular como la Poisson-Tweedie tal y como lo expuesto por (Chen, 2020) puede limitar la escogencia del mejor modelo que describa de forma apropiada nuestra pregunta de investigación.