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

4 de noviembre 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.

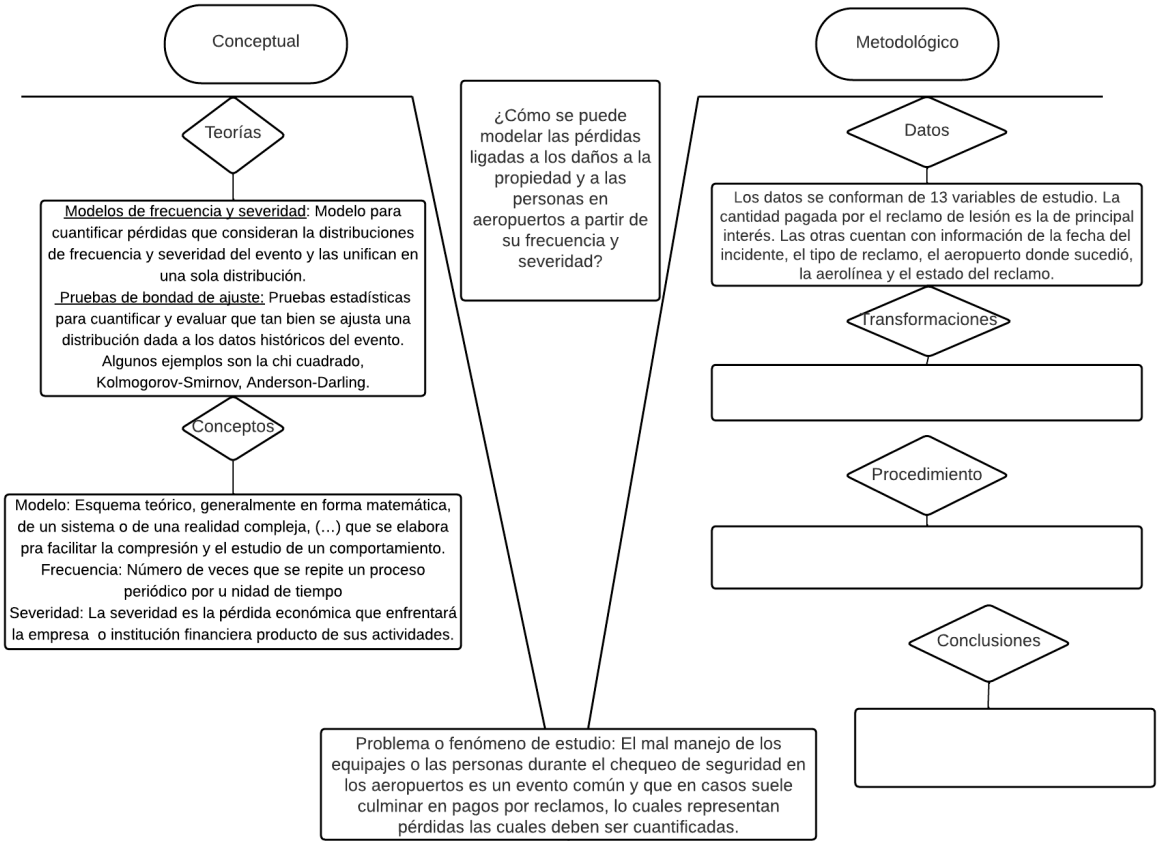
1.1.1.5 6. UVE Heurística

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.

Tabla 2.3: Conteo de reclamos por estado de resolución

Desconocido	Aprobado	Denegado	Acordado
6949	8738	21905	4005

2.3 Análisis Estadístico

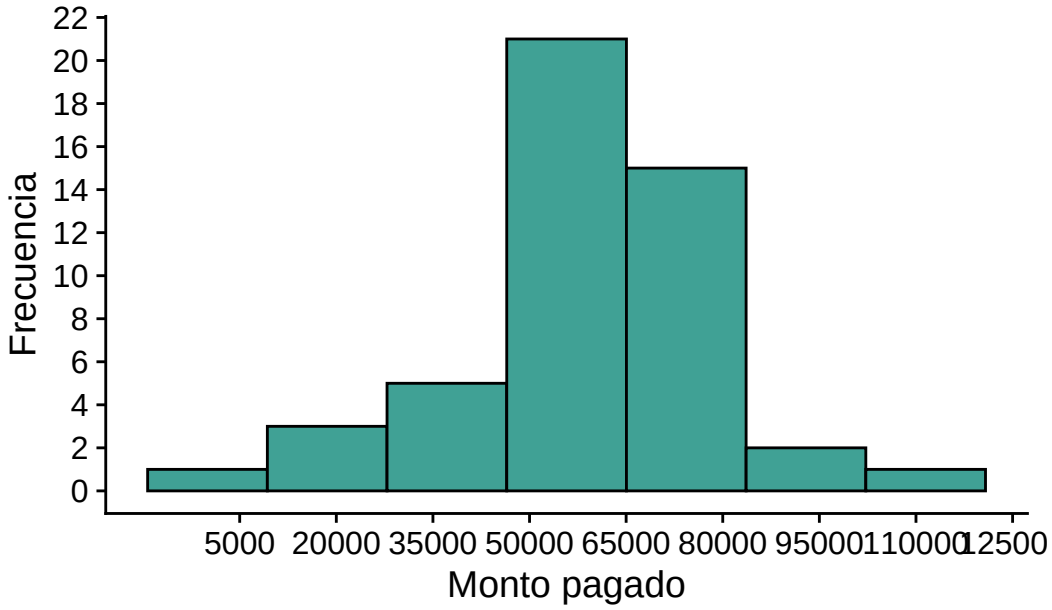
De la Tabla 2.2 se observa que está en formato tidy ya que cada variable tiene su propia columna (11 variables). Cada fila representa una instancia de un reclamo por lo que cada observación tiene su propia fila. Cada Celda tiene un solo valor.

En este punto se observa que la variable de mayor interés es *close_amount*, pues corresponde al desembolso efectivo al atender reclamos. Sin embargo, esta variable no es en sí misma útil para implementar los modelos sugeridos, sino que se tienen que construir los datos de frecuencia y severidad de los reclamos a TSA. Siguiendo a Flores (2022) y Chen (2020) se realizan dos cambios relevantes a este respecto. El primero consiste en filtrar la base de datos para conservar solamente aquellas observaciones en que efectivamente hubo un desembolso al atender el reclamo. Para esto se utiliza la variable *disposition*, que corresponde al estado de resolución del reclamo e indica si el reclamo fue denegado, si se pagó por completo el monto solicitado (aprobado) o si se llegó a un acuerdo (acordado) y se pagó solamente una fracción del monto del reclamo. En la Tabla 2.3 se muestra la frecuencia de cada estado de resolución. Consecuentemente, al filtrar las observaciones se pasa de 41 598 observaciones a 12 743

El segundo paso se refiere propiamente a la construcción de los datos de frecuencia y severidad. Esto se realiza agregando los datos ya filtrados de forma mensual. Para el caso de la frecuencia, esto se traduce en el conteo de reclamos en cada mes. Como el período de estudio comprende cuatro años (de 2010 a 2013), entonces se extraen 48 conteos (4 Œ 12). Ahora bien, para el caso de la severidad, esto se hace de forma similar solo que sumando los montos finales (*close_amount*) de los reclamos en cada mes, obteniéndose 48 valores para la severidad; por ejemplo, el primer valor de la severidad corresponde al monto total pagado por concepto de reclamos a TSA durante enero de 2010.

En la Figura 2.1 se observa la distribución empírica de la severidad. Se observa que la cola izquierda aparenta acumular un mayor peso que la derecha y que la mayor concentración ocurre aproximadamente para los montos pagados entre 50 000 y 60 000 dólares.

Figura 2.1: Histograma de montos pagados agregados por mes



Fuente: Elaboración propia con datos de TSA

En la Tabla 2.4 se muestran algunas estadísticas de los datos de frecuencia y severidad. Sorprende principalmente la asimetría obtenida para la severidad, que marca una discrepancia con los resultados obtenidos tanto por Flores (2022) como por Chen (2020), dado que ambos autores presentan coeficientes de asimetría positivos, sin embargo, debe tenerse en cuenta que el primero utiliza datos del período 2003-2015 (desagregados además por sitio y tipo) y el segundo del período 2008-2012. De la Figura 2.1 ya se observaba que no hay una asimetría positiva marcada en la severidad.

En la tabla Tabla 2.5 se observa la severidad promedio por mes codificado como 1 para enero y 12 para diciembre. Se nota que la severidad la relaciona cercanamente con las temporadas altas: severidades más altas en verano del hemisferio norte y Enero. Aparte de eso, la severidad es aproximadamente uniforme en el resto de los meses. 6

En la Tabla 2.6 se observa que los dos reclamos desproporcionalmente más frecuentes son los de pérdidas y daños a la propiedad de los pasajeros.

En Tabla 2.7 Se observa el número de ocurrencias por aerolínea, esto es, el cantidad de reclamos según la aerolínea en la que viajaba la persona que realizó el reclamo. Se tiene que la *Delta Air Lines* es la presenta mayor cantidad de reclamos reportados, seguidas de *Southwest Airlines* y *American Airlines*, esto es un comportamiento esperable ya que son aerolíneas líderes de mercado y por ende presentan mayor cantidad de viajeros y esto influye en la cantidad de reclamos.

Tabla 2.4: Medidas estadísticas de resumen para la severidad y la frecuencia

	Frecuencia	**Severidad**
Mínimo	44.00	9007.07
Primer cuartil	237.75	50883.81
Mediana	286.00	59722.03
Media	265.48	58393.64
Tercer cuartil	310.00	70566.03
Máximo	396.00	120512.60
Desviación estándar	75.62	20787.52
Rango intercuartil	72.25	19682.22
Asimetría	-1.35	-0.21
Curtosis	1.73	1.21

Tabla 2.5: Severidad promedio mensual

Mes	Severidad promedio
1	72304.02
2	50842.60
3	66050.08
4	58276.72
5	59759.59
6	62022.79
7	62882.68
8	65855.69
9	52256.92
10	55727.93
11	49131.62
12	45612.98

Tabla 2.6: Frecuencia por tipo de reclamo

Tipo de reclamo	Ocurrencias
-	212
Bus Terminal	1
Complaint	11
Employee Loss (MPCECA)	18
Motor Vehicle	143
Passenger Property Loss	25898
Personal Injury	383
Property Damage	14927
Wrongful Death	4
NA	1

Tabla 2.7: Frecuencia de reclamos por aerolínea

Aerolínea	Ocurrencias
Delta Air Lines	7029
Southwest Airlines	5953
American Airlines	5228
UAL	4636
USAir	3055
Jet Blue	1970
Continental Airlines	1713
Alaska Airlines	1505
AirTran Airlines	1020
British Airways	596

La Figura 2.2 muestra el conteo de incidencias menusales entre 2010 y finales del 2013. Se Muestra una tendencia fuerte de incrementeo hasta el mes 40. Esto se puede explicar a partir de que el TSA fue creado en el 2002 y durante su período inicial de funcionamiento se implementaron nuevas prácticas de seguridad en el aeropuerto por lo que los pasajeros y las autoridades tuvieron un período de aprendizaje. Luego del mes 40 se observa una fuerte tendencia de decremento posiblemente porque la población a este punto ya se acostumbró a las nuevas medidas implementadas. Esta tendencia es importante notarla pues Flores (2022) comenta que puede dificultar el proceso de ajustarle una distribución.

Figura 2.2: Número de reclamos mensuales del 2010 al 2013

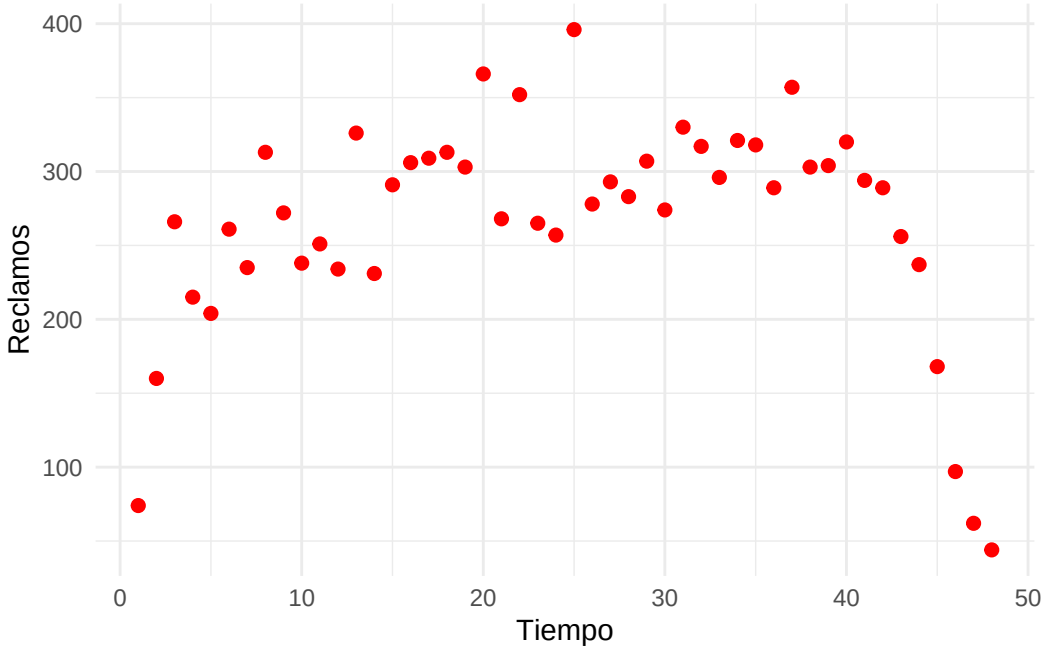


Figura 2.3: Aerolíneas con mayor monto promedio pagado

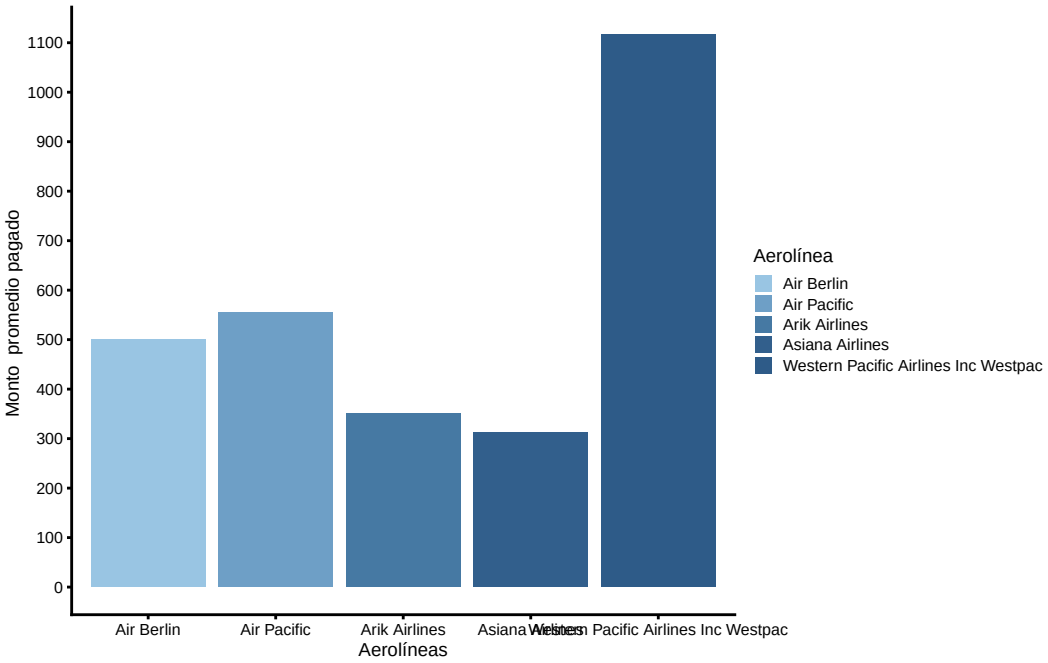
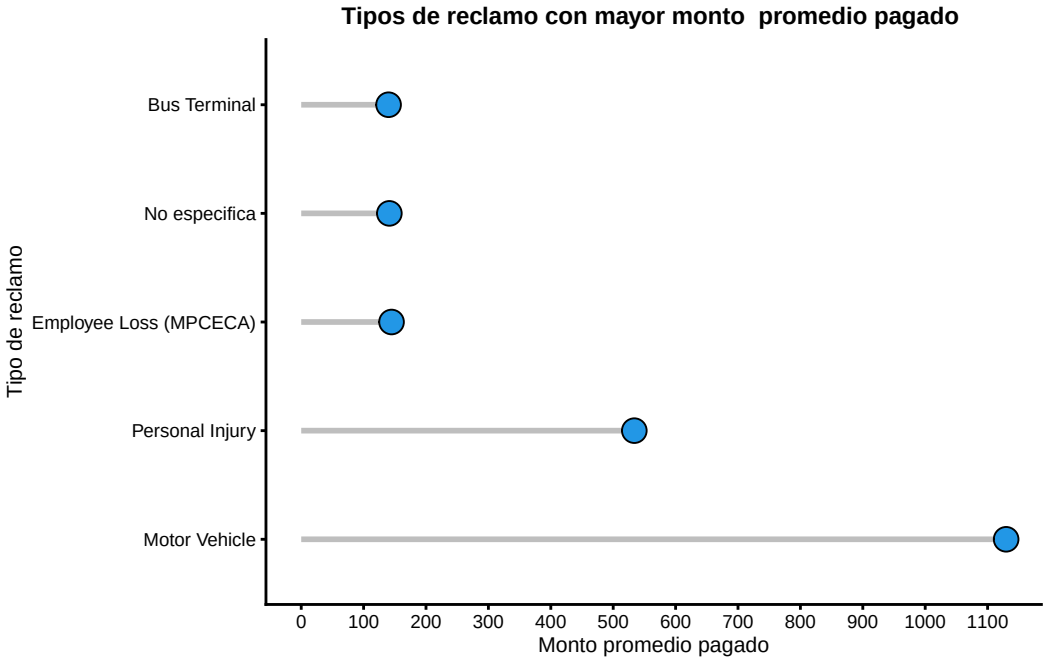


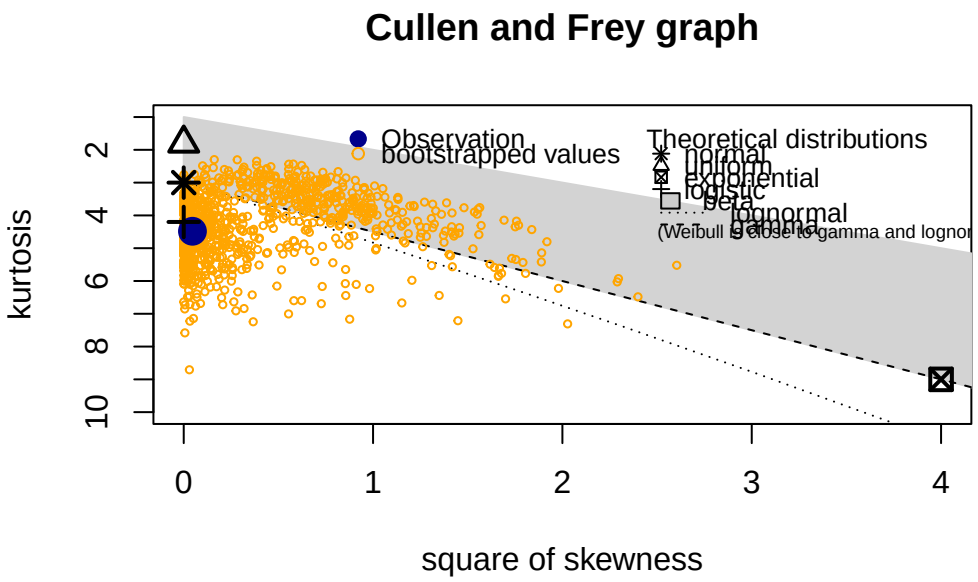
Figura 2.4: Tipos de reclamos con mayor monto promedio pagado



La Figura 2.5 muestra la relación entre curtosis y asimetría de severidad agregadas mensualemente. Se observa que preliminarmente se aproxima a una logística.

```
summary statistics
------
min:  9007.07   max:  120512.6
median:  59722.03
mean:  58393.64
estimated sd:  20787.52
estimated skewness:  -0.2145234
estimated kurtosis:  4.483542
```

Figura 2.5: Comparación de la curtosis y asimetría de la severidad



2.4 Fichas Bibliográficas:

1).

- Nombre de su hallazgo/resultado: Tendencia en la ocurrencia de reclamos
- Resumen en una oración: El número de reclamos mensuales incrementa con respecto al tiempo en los primeros 40 meses hasta alcanzar un máximo y desde entonces se ha mostrado un comportamiento a la baja en la cantidad de reclamos hechos.
- Principal característica: Tendencia aparente
- Problemas o posibles desafíos: En Flores (2022), se comenta que la existencia de una tendencia en los reclamos puede causar problemas al momento de buscar las distribuciones que se ajusten a los datos.
- Resumen en un párrafo: El número de reclamos mensuales parece incrementar rápidamente en los primeros 40 meses del período estudiado. Esto se podría explicar por la poca experiencia en materia de chequeos y procedimientos por parte de los pasajeros y las autoridades. Luego del mes 40 se observa una tendencia a la baja en la cantidad de reclamos, posiblemente porque a este punto ya se habían interiorizado las nuevas prácticas de seguridad. Esta tendencia puede ser un problema porque

en la literatura se expresó que puede complicar el proceso de ajustar una distribución a los reclamos, notando que al eliminar esta tendencia se facilitaba esta búsqueda.

2).

- Nombre de su hallazgo/resultado: Asimetría negativa de la severidad
- Resumen en una oración: La distribución de la severidad está ligeramente sesgada hacia la derecha, según lo indica un valor negativo del coeficiente de asimetría.
- Principal característica: El coeficiente de asimetría es negativo.
- Problemas o posibles desafíos: Esta característica es contrastante respecto de los resultados obtenidos por autores que han utilizado datos de reclamos a TSA, donde la asimetría positiva era muy marcada tanto numérica como visualmente y probalemente signifique que las distribuciones a emplear en el presente trabajo para ajustar la severidad sean muy distintas de las ya estudiadas.
- Resumen en un párrafo: La distribución de la severidad está ligeramente sesgada hacia la derecha, según lo indica un valor bajo pero negativo del coeficiente de asimetría en la Tabla 2.4. Del histograma en la Figura 2.1 ya se observaba que la distribución de la severidad agregada por mes no es claramente sesgada hacia la izquierda. Esta característica sorprende y marca una diferencia notable respecto de los resultados obtenidos por autores que han utilizado datos de reclamos a TSA, como en los trabajos de Flores (2022) y Chen (2020), donde la asimetría positiva era muy marcada tanto numérica como visualmente y probalemente signifique que las distribuciones a emplear en el presente trabajo para ajustar la severidad sean muy distintas de las ya estudiadas.

3).

- Nombre de su hallazgo/resultado: Variables de estudio secundarias con mayor monto promedio pagado.
- Resumen en una oración: Existen variables secundarias de nuestros estudio que tienen alto impacto implícito en la severidad de los montos promedio pagados, estas son:tipo de reclamo y aerolíneas. Donde el tipo de reclamo con mayor monto promedio es *motor vehicule* y para el caso de la aerolínea es *Western Pacific Airlines Inc Westpac*.
- Principal característica:Pese a que las variables principales para nuestro estudio son la frecuencia de reclamos y la severidad de estos. Existen variables secundarias que implícitamente tienen alto impacto en las mencionadas variables principales.
- Problemas o posibles desafíos:Existen observaciones donde no se reporta tanto la aerolínea como el tipo de reclamo por lo que solo se contemplan para el monto promedio de pago aquellas observaciones donde si se registran la información de las variables tipo de reclamo y aerolínea.
- Resumen en un párrafo:

Hay en particular dos variables secundarias altamente ligadas a la severidad para nuestro estudio, estas variables son: Tipo de reclamo y la aerolínea en la que viajaba la persona.

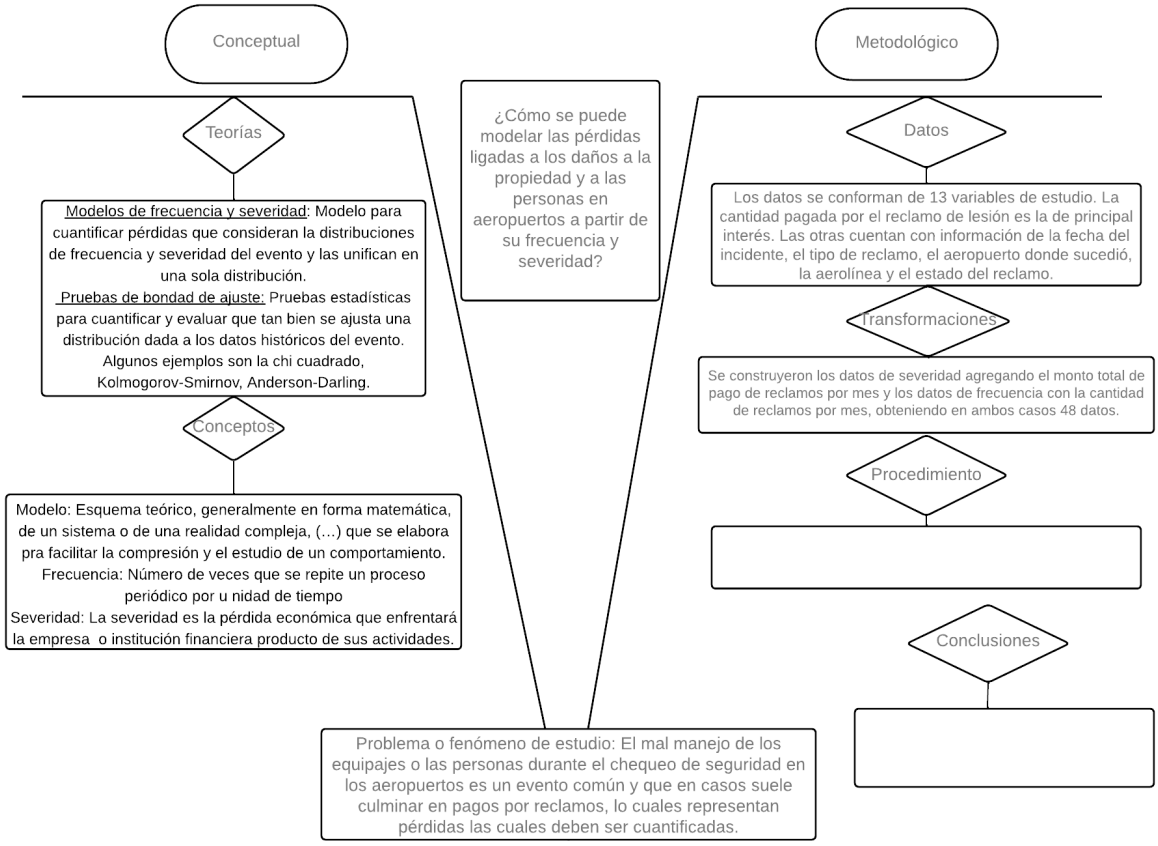
Es importante saber cuales son aquellos tipos de reclamo con mayor monto promedio pagado, como se observa en el Figura 2.4, el reclamo por *motor vehicule* y *personal injury* son aquellos que presentan mayor severidad, seguido de un subgrupo donde el monto de reclamo es menor como lo son: *employee loss* y *bus terminal*.

Análogamente, para el caso de las aerolíneas a las que pertenecen las personas que reportan mayor monto promedio pagado, se observa en el Figura 2.3 que *Western Pacific Airlines Inc Westpac* es la que presenta mayor monto promedio, seguido de *Air Pacific*.

2.5 Parte de reflexión

En la Figura 3.10 se muestra la UVE heurística actualizada, donde se incluyen las transformaciones sobre los datos para obtener los datos mensuales de frecuencia y severidad.

Figura 2.6: Actualización de de la UVE Heurística



En cuanto a las preguntas surgidas, sin duda el punto más sorpresivo consiste en la discordancia de la asimetría hallada para la distribución empírica de la severidad en contraste con los resultados expuestos por Flores (2022) y Chen (2020), quienes realizaron la agregación de la severidad de forma mensual y trabajaron con datos de reclamos a TSA pero para periodos distintos, hallando una marcada asimetría positiva. De esta manera, surge la duda de a qué puede obedecer esta diferencia, aunque debe tenerse en cuenta que contestar esta pregunta no es el objetivo de la presente investigación.

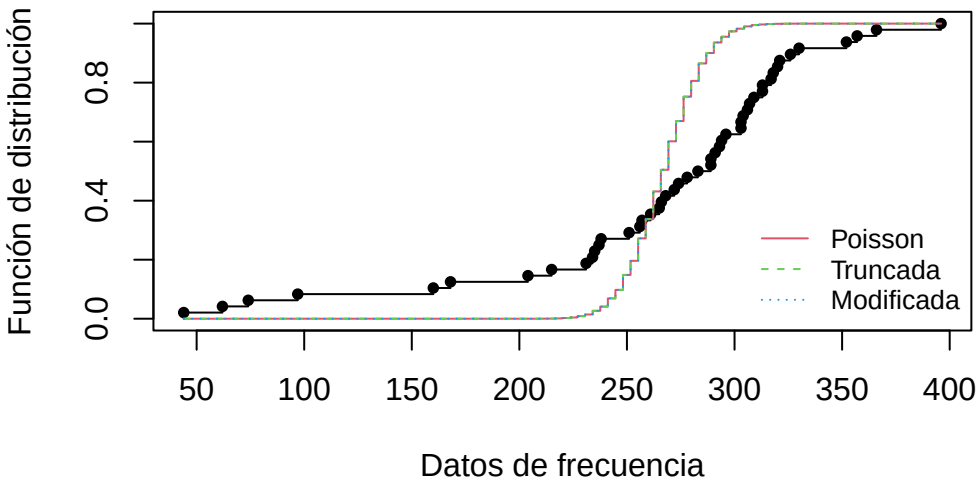
3 Bitácora 3

3.1 Parte de Planificación

3.1.1 Ajuste de la frecuencia

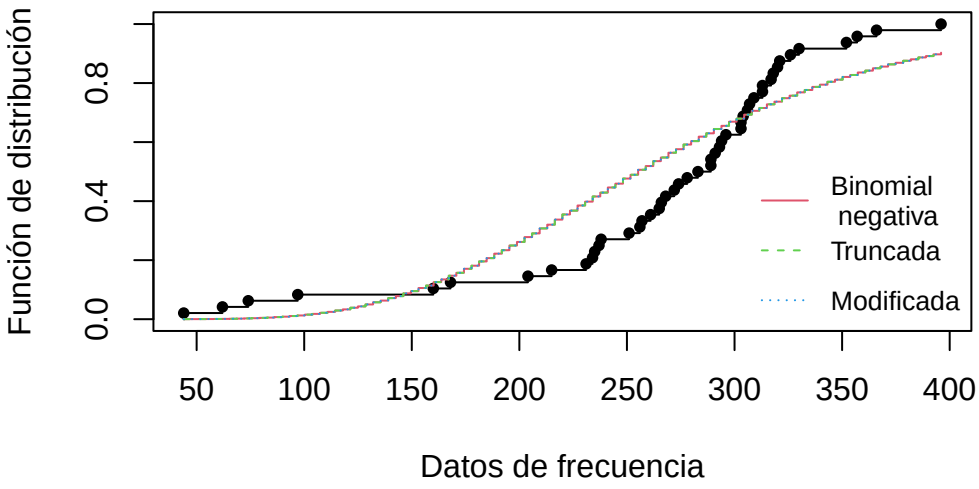
En la Figura 3.1 se muestra el ajuste de la distribución Poisson, y su versión truncada y modificada en cero, donde se comprueba que estas no difieren mucho entre sí, aparte de que se observa que el ajuste no es bueno.

Figura 3.1: Ajustes distribución Poisson



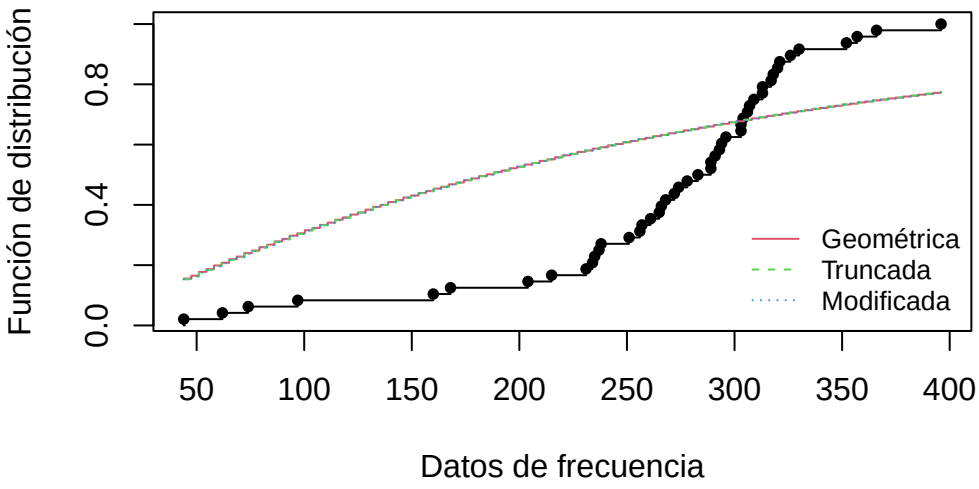
En la Figura 3.2 se muestra el ajuste de la distribución binomial negativa, y su versión truncada y modificada en cero, donde se comprueba que estas no difieren mucho entre sí, aparte de que se observa que el ajuste es un poco mejor que el de la Poisson, sin llegar a ser satisfactorio.

Figura 3.2: Ajustes distribución binomial negativa



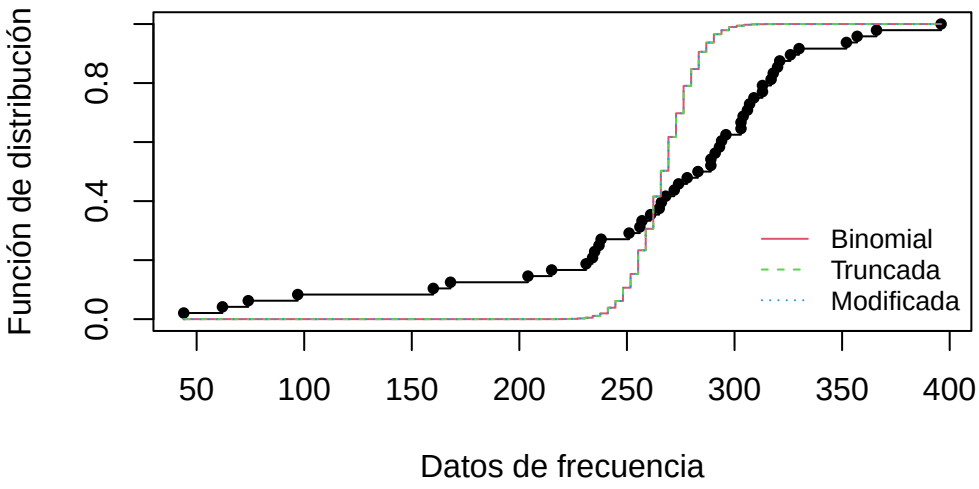
En la Figura 3.3 se muestra el ajuste de la distribución geométrica, y su versión truncada y modificada en cero, donde se comprueba que estas no se diferencian mucho. Siendo esta distribución un caso particular de la binomial negativa, se ve que el ajuste es muy inferior al del caso general. No es claro si el ajuste de la geométrica es mejor al de la Poisson.

Figura 3.3: Ajustes distribución geométrica



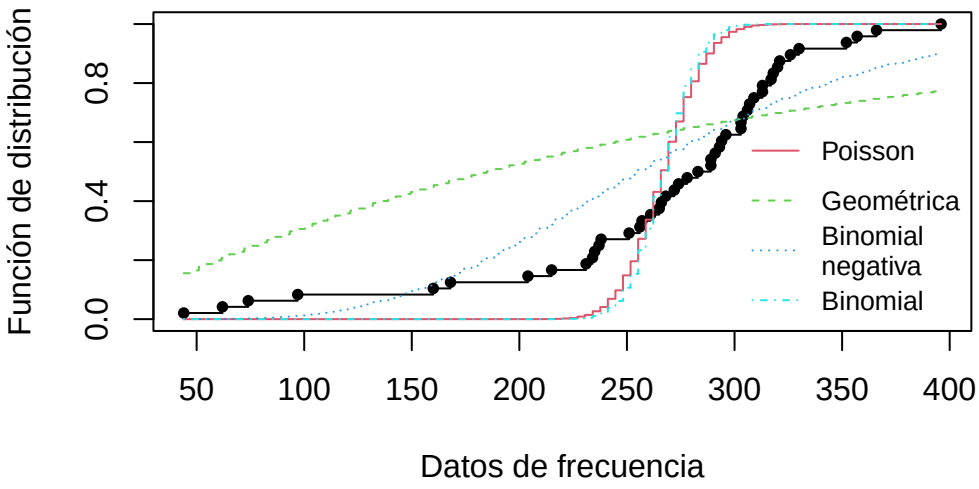
A continuación, en la Figura 3.4, se presenta el ajuste de la distribución binomial, el cual luce similar al de la Poisson y se ve que no es muy bueno.

Figura 3.4: Ajustes distribución binomial



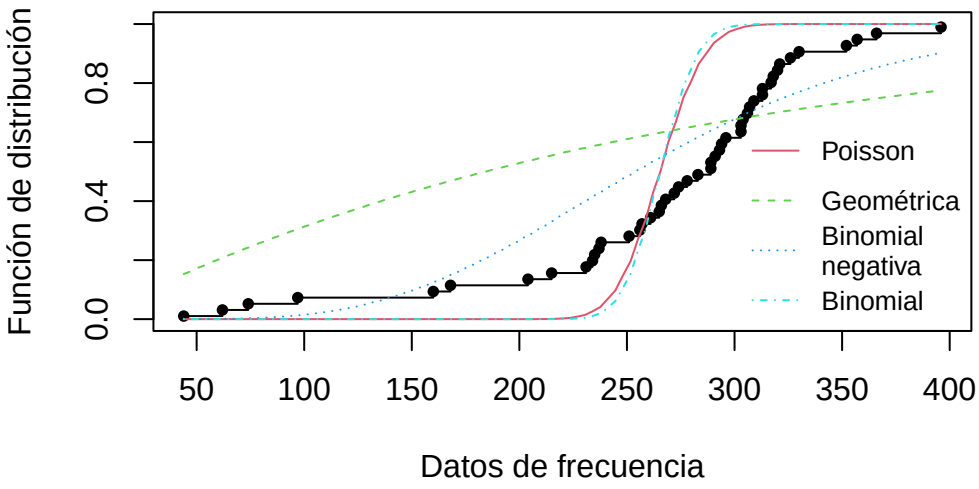
En la Figura 3.5 se muestran todos los ajustes de las distribuciones de la clase $(a, b, 0)$, lo que permite comparar y ver que la binomial negativa es la que mejor emula el comportamiento de la dsitribución empírica de los datos, estado, sin embargo, lejos de ser adecuado.

Figura 3.5: Ajustes con distribuciones de la clase (a,b,0)



A su vez, en la Figura 3.6 están los miembros de la clase $(a, b, 1)$ truncadas en cero, los cuáles no muestran ser mejores que sus distribuciones originales en la clase $(a, b, 0)$.

Figura 3.6: Ajustes con distribuciones de la clase (a,b,1) truncadas en cero



A su vez, en la Figura 3.7 están los miembros de la clase $(a, b, 1)$ modificadas en cero, los cuáles tampoco son visiblemente mejores que sus distribuciones originales en la clase $(a, b, 0)$.

Tabla 3.1: Parámetros de los modelos ajustados para la frecuencia

Distribución	**Parámetros en `R`**
Poisson	`lambda` = 265.479167
ZT-Poisson	`lambda` = 265.479326
ZM-Poisson	`lambda` = 265.473341 , `p0` = 0
Binomial negativa	`size` = 7.764187 , `mu` = 265.491586
ZT-Binomial negativa	`size` = 7.762927 , `prob` = 0.028412
ZM-Binomial negativa	`size` = 7.759127 , `prob` = 0.028396 , `p0` = 0
Geométrica	`prob` = 0.003753
ZT-Geométrica	`prob` = 0.003768
ZM-Geométrica	`prob` = 0.003766 , `p0` = 0
Binomial	`prob` = 0.29629 , `size` = 896.000008
ZT-Binomial	`prob` = 0.29629 , `size` = 896.000008
ZM-Binomial	`prob` = 0.296296 , `size` = 895.999931 , `p0` = 0

Figura 3.7: Ajustes con distribuciones de la clase (a,b,1) modificadas en cero

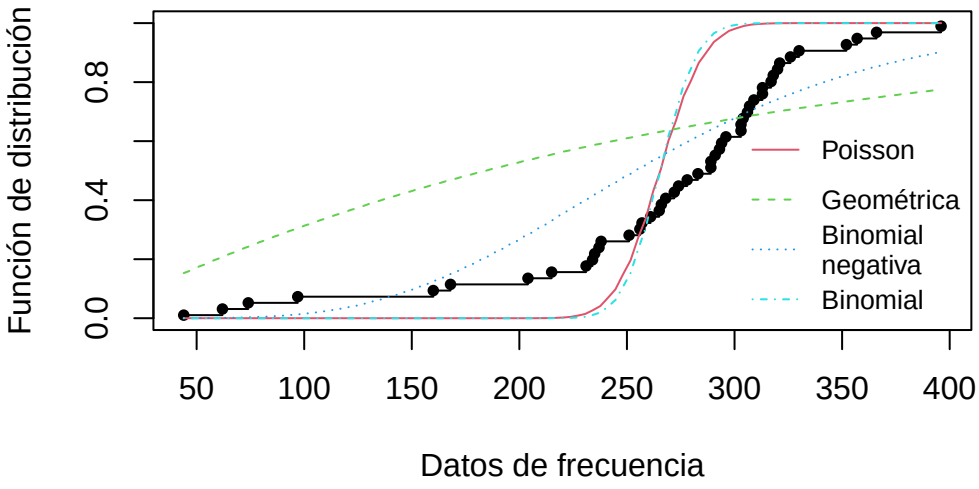


Tabla 3.2: Métricas de los modelos ajustados para la frecuencia

Distribución	**AIC**	**BIC**	**Valor** p
Poisson	1617.17955289	1619.0507539	0
ZT-Poisson	1617.1795529	1619.05075391	0
ZM-Poisson	1619.17955922	1622.92196124	0
Binomial negativa	575.35774711	579.10014913	3.5478e-06
ZT-Binomial negativa	575.35774432	579.10014634	3.5344e-06
ZM-Binomial negativa	577.35774424	582.97134727	1.1145e-06
Geométrica	634.00807001	635.87927102	0
ZT-Geométrica	633.64646075	635.51766176	0
ZM-Geométrica	635.64646629	639.38886831	0
Binomial	2009.5128356	2013.25523762	0
ZT-Binomial	2009.5128356	2013.25523762	0
ZM-Binomial	2011.51283388	2017.12643691	0

3.1.2 Ajuste de la severidad

Figura 3.8: Densidad de las distribuciones ajustadas con MLE

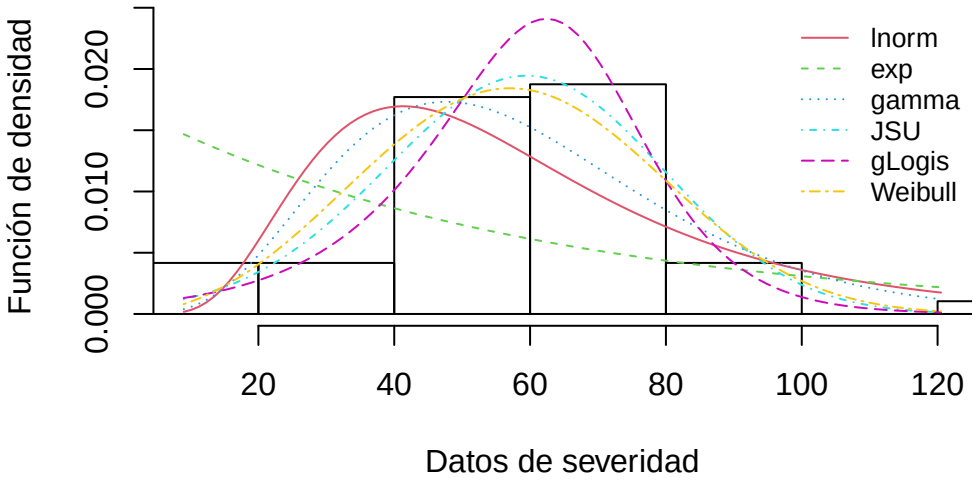
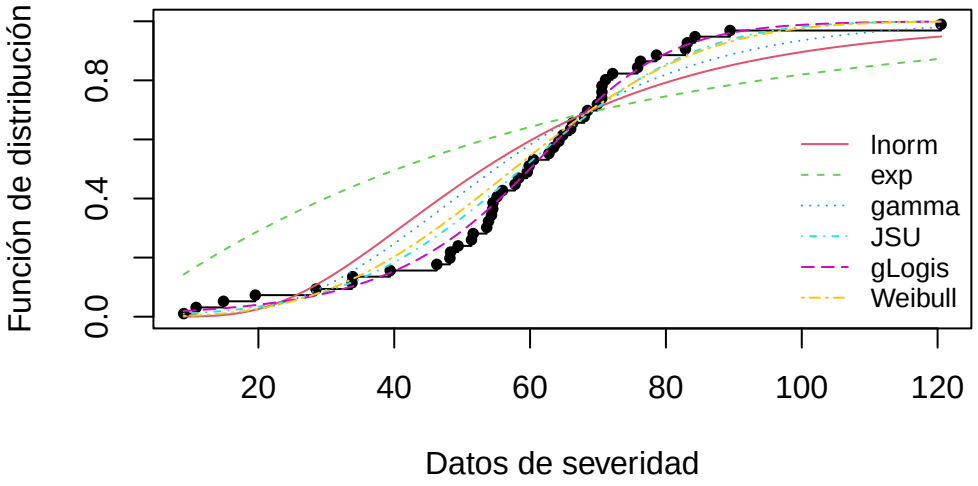


Tabla 3.3: Parámetros de los modelos ajustados para la severidad

Distribución	**Parámetros en `R`**
Lognormal	`meanlog` = 3.972659 , `sdlog` = 0.502333
Exponencial	`rate` = 0.017125
Gamma	`shape` = 5.448032 , `rate` = 0.093297
SU de Johnson	`mu` = 58.400803 , `sigma` = 20.541279 , `nu` = -31.417414 , `tau` = 24.031833
Logística generalizada	`location` = 67.038277 , `scale` = 8.5276 , `shape` = 0.581247
Weibull	`shape` = 3.061525 , `scale` = 64.909845

La Figura 3.8 muestra que la logística generalizada, weibull y Johnson SU muestran la forma que acerca más al histograma. Las otras distribuciones no se aproximan tan bien a los datos.

Figura 3.9: Función cumulativa de las distribuciones ajustadas con MLE



La Figura 3.9 se muestra la distribución cumulativa, la cual muestra más claramente que las distribuciones mencionadas anteriormentes se ajustan mejor. En la Tabla 3.3 se muestran los parámetros ajustados para cada modelo. Se observa que los modelos más adecuados, al menos visualmente, son más complejos al tener tres y cuatro parámetros

Tabla 3.4: Métricas de bondad de ajuste de la severidad

Distribución	AIC	BIC	LogLik	Estad. AD	Estad. KS
log-normal	455.50	459.24	-225.75	3.57	0.24
exponencial	488.45	490.32	-243.23	9.61	0.38
gamma	443.14	446.88	-219.57	2.44	0.20
Johnson SU	434.30	441.79	-213.15	0.78	0.12
glogis	427.78	433.39	-210.89	0.28	0.07
Weibull	432.85	436.59	-214.42	1.21	0.14

3.1.3 Fichas de resultados

1. **Nombre de Resultado**: Ajuste visual de la Severidad

   Resumen en una oración: En la Figura 3.8 y Figura 3.9 se observa uqe logística generalizada, Johnson SU y Weibull ajustan mejor a la severidad.

   Principal característica: Ovservando la Figura 3.9, estas tres distribuciones se acercan más a la función de distribución empírica.

   Problemas o posibles desafíos: La logística generalizada y Johnson SU son distribuciones de 3 y 4 parámtetros respectivamente lo que puede afectar el principio de parsimonía.

   Resumen en un párrafo: En la Figura 3.8 se observa que la exponencial no presenta ningún tipo de simetría, la cual sí se aprecia en el histograma de la severidad. Por otro lado, la log-normal y la gamma no parecen coincidir en la centralidad, aparte de que presetan más peso en la cola derecha. La distirbuciones de logística generalizada, weibull y Johnson SU parecen ajustarse mejor al histograma y las tres presentan una forma similar. Al observar la Figura 3.9 se aprecia un comportamiento similar: la exponencial parece estar muy lejos mientras que las tres ya mencionados se acercan bastante a la distribución empírica, y la logística generalizada parece ser la que sigue más cercanamente.

2. **Nombre de Resultado**: Ajuste bajo la prueba de kolmogorov-Smirnov para la severidad.

   Resumen en una oración: Se confirma que la Johnson SU, logística generalizada y weibull proporcionan buen ajuste para la severidad.

   Principal característica: Se da un no rechazo de la hipótesis nula para las distribuciones mencionadas con un nivel de significancia de 0.05.

   Problemas o posibles desafíos: Ninguna adicional al ya discutido

   Resumen en un párrafo: Para una significancia de 0.05 se tiene un valor crítico de 0.1962991, por lo que se rechaza la hipótesis de bondad de ajuste para la gamma, la exponencial y la log-normal, y se no se rechaza para la Johnson SU, logística generalizada y weibull. Además, note que el estadístico

más bajo obtenido es para la logística generalizada. Bajo una significancia de 0.01 se obtiene un valor crítico de 0.2352702, por lo que se rechaza rotundamente la log-normal y la exponencial como distribuciones candidatas para la severidad bajo esta prueba.

3. **Nombre de Resultado**: Ajuste bajo la prueba de Anderson-Darling para la severidad

   Resumen en una oración: Se confirma que la Johnson SU, logística generalizada y weibull proporcionan buen ajuste para la severidad.

   Principal característica: Se da un no rechazo de la hipótesis nula para las distribuciones mencionadas con un nivel de significancia de 0.05.

   Problemas o posibles desafíos: Ninguna adicional al ya discutido

   Resumen en un párrafo: Para una significancia de 0.05 se tiene un valor crítico de 2.492, por lo que se rechaza la hipótesis de bondad de ajuste para la log-normal y exponencial, y se no se rechaza para la Johnson SU, logística generalizada, weibull y gamma. Además, note que el estadístico más bajo obtenido es para la logística generalizada. Bajo una significancia de 0.01 se obtiene un valor crítico de 3.857, por lo que se rechaza rotundamente la exponencial como distribuciones candidatas para la severidad bajo esta prueba.

4. **Nombre de Resultado**: AIC, BIC y versosimilitud para la severidad

   Resumen en una oración: La logística generalizada obtiene la menor verosimilitud, AIC y BIC lo que confirma que esta es la mejor candidata para la severidad.

   Principal característica: Bajo los criterios de AIC, BIC y verosimilitud se confirma que las mejores candidatas son la logística generalizada, Johnson SU y Weibull

   Problemas o posibles desafíos: Ninguna adicional al ya discutido

   Resumen en un párrafo: Entre las distribuciones ajustadas la logística generalizada presenta el menor valor para la verosimilitud, seguido por la Johnson SU y luego la Weibull. Este mismo comportamiento se repite con el caso del BIC. También se obtiene un menor valor de AIC para la logística, seguido por la Weibull y la Johnson SU . lo cual confirma lo que ya se había observado. Esto proporciona aún más evidencia de que la logística generalizada es la mejor candidata.

5. **Nombre de Resultado**: Ajuste visual de la distribuciones de clase $(a, b, 0)$ para la frecuencia

   Resumen en una oración: Se grafican las CDF empíricas contra las teóricas y ningún ajuste es bueno.

   Principal característica: La distribución binomial negativa parece el mejor ajuste.

   Problemas o posibles desafíos: Ninguna de las distribuciones básicas de frecuencia logra emular la forma de la distribución empírica, por lo que deben buscarse otras distribuciones discretas que posean una mayor flexibilidad. Además, se presentaron muchas dificultades de índole numérica

Resumen en un párrafo: En la Figura 3.5 se presentan los ajustes de las distribuciones de la clase $(a, b, 0)$ para modelar la frecuencia de los reclamos. Ninguna parece emular suficientemente bien la forma de la distribución empírica, siendo la binomial negativa la que parece hacerlo mejor. Para esta distribución, se obtuvieron los parámetros $n = 8$ y $p = 265.49$ según la parametrización de `R`. Los resultado para las demás distribuciones, se resumen en la Tabla 3.1.

6. **Nombre de Resultado**: Ajuste visual de la distribuciones de clase $(a, b, 1)$ para la frecuencia

   Resumen en una oración: Se grafican las CDF empíricas contra las teóricas y ningún ajuste es bueno.

   Principal característica: La distribución binomial negativa parece el mejor ajuste, tanto en el caso truncado como en el modificado.

   Problemas o posibles desafíos: Ninguna de las distribuciones básicas de la familia $(a, b, 1)$ logra emular la forma de la distribución empírica, de forma que debe explorarse alguna alternativa que permita una mayor flexibilidad.

   Resumen en un párrafo: En las figuras 3.6 y 3.7 se presentan los ajustes de las distribuciones de la clase $(a, b, 1)$ para modelar la frecuencia de los reclamos. Ninguna parece emular suficientemente bien la forma de la distribución empírica, siendo la binomial negativa la que parece hacerlo mejor en ambos casos. Además, de las figuras 3.1, 3.2, 3.3, 3.4 se ve aprecia que las distribuciones de la clase $(a, b, 1)$ son prácticamente indiscernibles respecto de las distribuciones correspondientes de la clase $(a, b, 0)$. Además, en la Tabla 3.1 se ve que no hay mucha diferencia en los parámetros y en el caso de las distribuciones modificadas, la probabilidad en cero estimada por máxima verosimilitud se redondea precisamente a cero.

7. **Nombre de Resultado**: Prueba de bondad de ajuste de las distribuciones de frecuencia

   Resumen en una oración: Se conduce una prueba Chi-Cuadrado de bondad de ajuste sobre todas las doce distribuciones ajustadas y ninguna presenta resultados adecuados.

   Principal característica: Se obtienen valores p muy bajos, con lo que se rechaza la hipótesis de bondad de ajuste bajo los niveles de significancia usuales del 10%, 5% y 1%.

   Problemas o posibles desafíos: La eficacia del modelo agraegado va a estar fuertemente comprometido si el ajuste de de la frecuencia es inadecaudo. Se valora probar distribuciones, como la Poisson-Gaussiana inversa, o la Poisson-Geométrica en procura de una mayor flexibilidad.

   Resumen en un párrafo: En la Tabla 3.2 se presenta el valor p resultante de la prueba Chi-Cuadrado de bondad de ajuste. Se obtuvieron valores p muy bajos, con lo que se rechaza la hipótesis de bondad de ajuste bajo los niveles de significancia usuales del 10%, 5% y 1%, de modo que hay evidencia suficiente para rechazar la hipótesis de que la distribución de la frecuencia proviene de cualquiera de las propuestas. Sin embargo, para continuar con las instrucciones de la bitácora 3 y en ausencia de un modelo alternativo mejor, se toman las distribuciones con mayores valores p, que son las binomiales negativas.

8. **Nombre de Resultado**: Medidas AIC y BIC para la frecuencia

   Resumen en una oración: Los modelos binomial negativos tienen los valores más bajos de AIC y BIC, de entre los cuáles el mejor bajo estas medidas es el binomial negativo truncado en cero.

   Principal característica: El modelo de menor AIC y BIC es el binomial negativo truncado en cero, pero por una difencia ínfima respecto al binomial negativo.

   Problemas o posibles desafíos: No hay mucha diferencia entre las medidas para decantarse por el binomial negativo o el binomial negativo modificado en cero, al punto de que hay que recurrir a la sexta cifra decimal para decidir.

   Resumen en un párrafo: En la Tabla 3.2 se presentan las medidas de AIC y BIC para cada modelo. Se ve que en general, los mejores modelos son los binomiales negativos, seguidos de los geométricos (que son casos especiales de los anteriores), Poisson y, por último, los binomiales. El modelo de menor AIC y BIC es el binomial negativo truncado en cero. Sin embargo, se observa que no hay mucha diferencia entre las medidas para decantarse por el binomial negativo o el binomial negativo modificado en cero, al punto de que hay que recurrir a la sexta cifra decimal para decidir.

3.1.4 Tablas

Tabla 3.5: Elementos de reporte

Primarios	Secundario
• Teoría A: Estimación paramétrica por máxima verosimilitud • Teoría B: Selección de modelos distribucionales con pruebas de bondad de ajuste • Teoría C: Selección de modelos con el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC) • Resultado A: Ninguna de las distribuciones de las clases (a,b,0) o (a,b,1) son adecuadas para modelarla frecuencia de los reclamos. • Hipótesis A: La distribución logística generalizada es la más adecuada para modelar la severidad de los reclamos.	• Teoría D: Uso de distribuciones compuestas para el ajuste de la frecuencia • Hipótesis B: Distribuciones compuestas podrían dar un mejor ajuste de la frecuencia al ser más flexibles (se valoran por ejmplo las distribuciones compuestas Poisson-Poisson o Neyman tipo A; Poisson-Geométrica o Polya-Aeppli; y Poisson-Gaussiana inversa) • Resultado B: De las distribuciones de la clase (a,b,0) y (a,b,1), la binomial negativa es la que proporciona el mejor ajuste a los datos de frecuencia.

Tabla 3.6: Distribución de contenidos por sección.

Sección	Temas a tratar
Introducción	1. Introducción al modelado de pérdidas.(primario) 2. Contextualización de la problemática surgida por los daños a la propiedad y a las personas en aeropuertos de Estados Unidos.(primario) 3. Teoría de estimación paramétrica,pruebas de bondad de ajuste y pérdidas agregadas. (primario) 4. Teoría de distribuciones compuestas para el ajuste de frecuencia.(secundario) 5. Resultados de estudios afines. (secundario)
Metodología	1. Introducción de la base de datos y análisis descriptivo. (primario) 2. Método A: Estimación paramétrica vía máxima verosimilitud. (primario) 3. Método B: Uso de distribuciones compuestas para ajustar la frecuencia de reclamos.(secundarios) 4. Selección de modelos mediante pruebas de bondad de ajuste: Chi-Cuadrado, Kolmogorov-Smirnov y Anderson-Darling para modelos obtenidos por método A. (primario) 5. Selección de modelos mediante AIC y BIC. (primario). 6. Método de recursión (Fórmula de Panjer) para los modelos seleccionados según método B. (secundario)
Resultados	1. Resultado A: Ninguna de las distribuciones de las clases (a,b,0) o (a,b,1) son adecuadas para modelar la frecuencia de reclamos.(primario) 2. Resultado B: Entre las distribuciones de clases (a,b,0) y (a,b,1) la binomial negativa es la que proporciona el mejor ajuste a los datos de frecuencia.(secundario) 3. Resultado C: Entre las distribuciones de severidad utilizadas para el ajuste de severidad la logística generalizadas es la más adecuada para modelar la severidad de reclamos. (primario)

3.2 Parte de escritura

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, el autor 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. No obstante, es importante señalar que este autor obtiene muy malos ajustes para las frecuencias de reclamos, ya que al hacer los ajustes respectivos obtiene valores p de 0. Por lo que decide tomar la binomial negativa, debido a que es la que posee menor valor en el estadístico Chi-Cuadrado. Es por esta razón que el autor propone para el modelado de la frecuencia mixturas discretas, dado el mal ajuste obtenido.

En la Figura 3.5 se presentan los ajustes de las distribuciones de la clase $(a, b, 0)$ para modelar la frecuencia de los reclamos de nuestro estudio. Ninguna parece emular suficientemente bien la forma de la distribución empírica sucediendo algo similar a lo observado en el estudio de Flores (2022), siendo la binomial negativa la que parece hacerlo mejor. En las figuras 3.6 y 3.7 se presentan los ajustes de las distribuciones de la clase $(a, b, 1)$ para modelar la frecuencia de los reclamos. Ninguna parece emular suficientemente bien la forma de la distribución empírica, siendo la binomial negativa la que parece hacerlo mejor en ambos casos. Además, de las figuras 3.1, 3.2, 3.3, 3.4 se aprecia que las distribuciones de la clase $(a, b, 1)$ son prácticamente indiscernibles respecto de las distribuciones correspondientes de la clase $(a, b, 0)$. Aunado a esto, se ve en la Tabla 3.1 que no hay mucha diferencia en los parámetros y en el caso de las distribuciones modificadas, la probabilidad en cero estimada por máxima verosimilitud se redondea precisamente a cero.

Ondieki et al. (2018), 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 (para la severidad).

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) y el Criterio de Información Bayesiano (BIC).

En nuestro estudio, se decide seguir esta línea de razonamiento y la implementación de las pruebas de modelos realizadas por Ondieki et al. (2018).

Respecto a la severidad, de la Figura 3.8 se observa que la distribución exponencial no presenta ningún tipo de simetría, la cual sí se aprecia en el histograma de la severidad. Por otro lado, la log-normal y la gamma no parecen coincidir en la centralidad, aparte de que presentan más peso en la cola derecha. La distribuciones de logística generalizada, Weibull y Johnson SU parecen ajustarse mejor al histograma y las tres presentan una forma similar. Al observar la Figura 3.9 se aprecia un comportamiento similar: la exponencial parece

estar muy lejos mientras que las tres ya mencionados se acercan bastante a la distribución empírica, y la logística generalizada parece ser la que sigue más cercanamente.

En la Tabla 3.2 se presenta el valor p resultante de la prueba Chi-Cuadrado de bondad de ajuste en el caso de la frecuencia. Se obtuvieron valores p muy bajos, con lo que se rechaza la hipótesis de bondad de ajuste bajo los niveles de significancia usuales del 10%, 5% y 1%, de modo que hay evidencia suficiente para rechazar la hipótesis de que la distribución de la frecuencia proviene de cualquiera de las propuestas. Establecido esto, en ausencia de un modelo alternativo mejor, se toman las distribuciones con mayores valores p, que son las binomiales negativas. Obsérvese que este inconveniente coincide con el documentado por Flores (2022).

A su vez, en la Tabla 3.2 se presentan las medidas de AIC y BIC para cada modelo de frecuencia. Se ve que, en general, los mejores modelos son los binomiales negativos, seguidos de los geométricos (que son casos especiales de los anteriores), Poisson y, por último, los binomiales. El modelo de menor AIC y BIC es el binomial negativo truncado en cero. Sin embargo, se observa que no hay mucha diferencia entre las medidas para decantarse por el binomial negativo o el binomial negativo modificado en cero, al punto de que hay que recurrir a la sexta cifra decimal para decidir por el primero.

Como se mencionó, Flores (2022) sugiere usar distribuciones compuestas para la frecuencia, enfoque que se tratará de adoptar para conseguir un mejor ajuste, bajo la forma específica de tres modelos: Poisson-Poisson o Neyman tipo A; Poisson-Geométrica o Polya-Aeppli; y Poisson-Gaussiana inversa. Además, como una idea tentativa, se encontró el uso de versiones discretas de distribuciones continuas descritas, por ejemplo, en Chakraborty (2015) y Vila et al. (2019), que además se han aplicado en seguros, tal como se retrata en Lyu & Nadarajah (2022).

Por otro lado, para la severidad, según lo establece Flores (2022), 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.

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. Al igual que el estudio de Flores (2022), 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 general, de las propuestas paramétricas, la log-normal tuvo un mejor desempeño en el estudio de Pitt et al. (2011), que es algo que concuerda con el de Flores (2022).

Para una significancia de 0.05, en el presente estudio se tiene un valor crítico de 0.1962991 según la prueba Kolmogorov-Smirnov, por lo que se rechaza la hipótesis de bondad de ajuste para la gamma, la exponencial y la log-normal, y no se rechaza para la Johnson SU, logística generalizada y Weibull. Además, note que el estadístico más bajo obtenido es para la logística generalizada. Bajo una significancia de 0.01 se obtiene un valor crítico de 0.2352702, por lo que se rechaza rotundamente la log-normal y la exponencial como distribución candidata para la severidad bajo esta prueba.

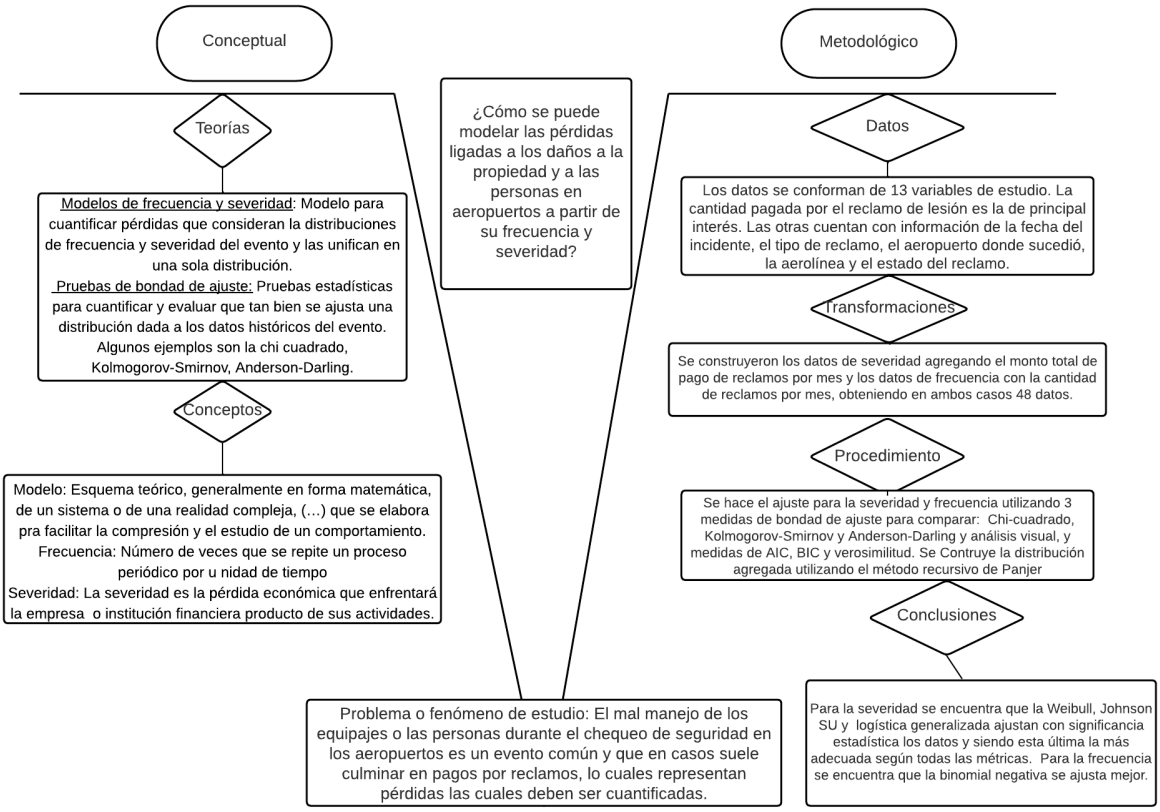
En cuanto a la prueba de Anderson-Darling, con una significancia de 0.05 se tiene un valor crítico de 2.492, por lo que se rechaza la hipótesis de bondad de ajuste para la log-normal y exponencial, y no se rechaza para la Johnson SU, logística generalizada, Weibull y gamma. Además, note que el estadístico más bajo obtenido es para la logística generalizada. Bajo una significancia de 0.01 se obtiene un valor crítico de 3.857, por lo que se rechaza rotundamente la exponencial como distribuciones candidatas para la severidad bajo esta prueba.

Entre las distribuciones ajustadas, la logística generalizada presenta el menor valor para la verosimilitud, seguido por la Johnson SU y luego la Weibull. Este mismo comportamiento se repite con el caso del BIC. También se obtiene un menor valor de AIC para la logística, seguido por la Weibull y la Johnson SU, lo cual confirma lo que ya se había observado. Esto proporciona aún más evidencia de que la logística generalizada es la mejor candidata.

3.3 Parte de reflexión

Luego de hacer la implementación del modelo escogido queda delimitado completamente el alcance del proyecto: Las distribuciones candidatas escogidas, las pruebas de bondad de ajuste: AIC, BIC, Kolmogorov-Smirnov, Anderson-Darling y Chi-cuadrado. Se encuentran resultados muy alentadores para la severidad y se logra ajustar parcialmente la frecuencia. Con esto se logra responder parcialmente la pregunta de investigación. A la UVE se le agrega las conclusiones obtenidas y se adjunta a continuación

Figura 3.10: Actualización de de la UVE Heurística 3



Referencias

Chakraborty, S. (2015). Generating discrete analogues of continuous probability distributions-A survey of methods and constructions. *Journal of Statistical Distributions and Applications*, *2*(1), 1-30.

Chen, S. (2020). *Agrregate Loss Model with Poisson-Tweedie Loss Frequency.*

Feldman, J., & Brown, R. (2005). *Risk and insurance.*

Flores, R. C. (2022). *Modelling dependencies in airport passenger: claim data using copulas* [Tesis doctoral]. Instituto Superior de Economia e Gestão.

Homeland Security, D. of. (2015). *TSA claims data.* https://www.dhs.gov/tsa-claims-data

Kelly, M., & Wang, Z. (2020). A data set for modeling claims processes—tsa claims data. *Risk Management and Insurance Review*, *23*(3), 269-276.

Klugman, S., Panjer, H., & Willmot, G. (2019). *Loss Models From Data ro Decisions, 5 th., vol. 6, no. 1. Hoboken.* New Jersey: John Wiley & Sons, Inc.

Lyu, J., & Nadarajah, S. (2022). Discrete lognormal distributions with application to insurance data. *International Journal of System Assurance Engineering and Management*, *13*(3), 1268-1282.

Ondieki, C., Gathoni, S., & Wairimu, J. (2018). *Modeling the Frequency and Severity of Auto Insurance Claims Using Statistical Distributions.*

Pitt, D., Guillen, M., & Bolancé, C. (2011). *Estimation of parametric and nonparametric models for univariate claim severity distributions: an approach using R.*

Vila, R., Nakano, E. Y., & Saulo, H. (2019). Theoretical results on the discrete Weibull distribution of Nakagawa and Osaki. *Statistics*, *53*(2), 339-363.