UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

CIÊNCIAS ATUARIAIS

Método de Panjer para a obtenção da distribuição do sinistro agregado: uma aplicação em dados reais

FELIPE OLIVEIRA MUNIZ DA SILVA
RAFAEL SOUZA DOS SANTOS

Rio de Janeiro
2014

FELIPE OLIVEIRA MUNIZ DA SILVA
RAFAEL SOUZA DOS SANTOS

Método de Panjer para a obtenção da distribuição do sinistro agregado: uma aplicação em dados reais

Projeto de Graduação apresentado ao curso de Ciências Atuariais do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências Atuariais.

Orientador:

Prof. Paulo Pereira Ferreira

Rio de Janeiro
Fevereiro 2014

FELIPE OLIVEIRA MUNIZ DA SILVA
RAFAEL SOUZA DOS SANTOS

Método de Panjer para a obtenção da distribuição do sinistro agregado: uma aplicação em dados reais

Projeto de Graduação apresentado ao curso de Ciências Atuariais do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências Atuariais.

Examinado por:

____

Prof. Paulo Pereira Ferreira

____

Prof.ª Flávia Maria P. F. Landim

____

Thereza Christina M. de Oliveira

Rio de Janeiro
Fevereiro 2014

AGRADECIMENTOS

Agradecemos primeiramente às nossas famílias, por todo amor, carinho e por nos ensinarem sempre a importância de se dedicar aos estudos. Além, é claro, de todo apoio e confiança que colocaram em nossas escolhas pessoais e profissionais, que nos permitiram chegar até aqui.

Agradecemos a todos os amigos do Fundão, tanto nossos contemporâneos quanto companheiros que ingressaram em outros anos, por todos os momentos de descontração e amizade. Graças a esses amigos todo o percurso árduo da graduação se tornou muito mais leve, além de todo apoio mútuo nos estudos das disciplinas.

Agradecemos também aos docentes do DME – UFRJ (Departamento de Métodos Estatísticos) pelos grandes ensinamentos passados e por toda a paciência que tiveram para conosco ao longo desses anos, nos garantindo toda a base necessária para se atingir o sucesso profissional. Agradecemos em especial ao professor Paulo Ferreira, por toda a boa vontade e paciência ao orientar esse nosso projeto, enriquecendo-o com toda sua experiência.

Por fim, agradecemos ainda à professora Flávia Maria Pinto Ferreira Landim e Thereza Christina Moreno de Oliveira por terem aceitado fazer parte da banca avaliadora desse projeto.

RESUMO

Atualmente, há uma preocupação da maioria dos indivíduos com relação à ocorrência de eventos inesperados que possam trazer consequências negativas significativas em sua vida e como estarem preparados para isso. Para as perdas de caráter financeiro, como um acidente envolvendo seu carro, ou sua residência, há a opção do indivíduo fazer um contrato de seguro, que atenua esse prejuízo. Isso justifica o grande crescimento do mercado de seguros no Brasil e no mundo. Porém, para que as empresas seguradoras possam garantir aos seus clientes essa segurança, é necessário que elas tenham o máximo de conhecimento possível acerca das distribuições de probabilidade envolvidas com o risco de ocorrência de sinistros e com as perdas financeiras decorrentes destes. Então, esse trabalho tem como objetivo apresentar ferramentas para se obter essas distribuições e demonstrá-las através da aplicação em uma base de dados real recente sobre sinistros de automóveis (Casco). Inicialmente falaremos de forma breve sobre esse ramo. Depois serão apresentados os resultados teóricos envolvendo o cálculo das distribuições do número de sinistros, do valor de cada sinistro individualmente e do valor total de sinistros; com destaque para o Método de Panjer. Por fim, essa metodologia será aplicada na base de dados, ilustrando e comprovando sua relevância para solucionar o problema descrito.

Palavras-chave: Seguro de automóvel, Cálculo do sinistro agregado, Método Panjer,

LISTA DE FIGURAS

SUMÁRIO

1. Introdução

A finalidade deste projeto é o estudo da distribuição do número de sinistros, do valor de um sinistro e do sinistro agregado de uma carteira de seguro de automóveis, a partir da aplicação do Método de Panjer em uma base de dados real.

O conhecimento da distribuição do sinistro agregado de uma carteira é de grande importância, pois possibilita utilização das melhores práticas no que tange à precificação e ao gerenciamento do risco. A obtenção dessa distribuição era bastante complexa, até que Panjer desenvolveu uma fórmula recursiva para encontrá-la de modo mais simples. Para aplicação do método, foi escolhida uma carteira de seguros de automóveis, em virtude da grande massificação desse seguro e da consequente competição existente no mercado segurador brasileiro.

No capítulo 2, temos uma breve descrição do ramo escolhido, bem como a apresentação de algumas estatísticas que revelam a importância desse ramo no mercado segurador. Nos capítulos 3 e 4 são apresentados respectivamente os modelos de risco individual e coletivo. No capitulo 5 é apresentada a fórmula recursiva de Panjer. No capítulo 6 temos a descrição da base de dados e da metodologia utilizada. No capítulo 7 encontram-se os resultados obtidos e, por fim, a conclusão é apresentada no capítulo 8.

2. Seguro de Automóveis - Casco

Seguro é o nome dado a todo contrato pelo qual uma das partes, denominada seguradora, se obriga a indenizar a outra, denominada segurado, em caso de ocorrência de determinado sinistro, em troca do recebimento de um prêmio de seguro.

No caso de seguro de automóvel (casco), o segurado paga regularmente uma quantia financeira à seguradora e em caso de algum acidente envolvendo o automóvel objeto do seguro, a seguradora repõe o valor monetário perdido em decorrência desse acidente.

A figura 1 mostra a evolução dos prêmios deste ramo por ano, no Brasil, desde a última década, evidenciando o seu crescimento e sua assimilação em nossa sociedade, até mesmo em consequência do crescimento da economia em nosso país e do maior poder de compra do brasileiro. A maior facilidade de acesso ao crédito e à financiamentos acarretou um aumento no número de automóveis nas ruas, que por conseguinte, aumentou a demanda por seguros de automóveis.

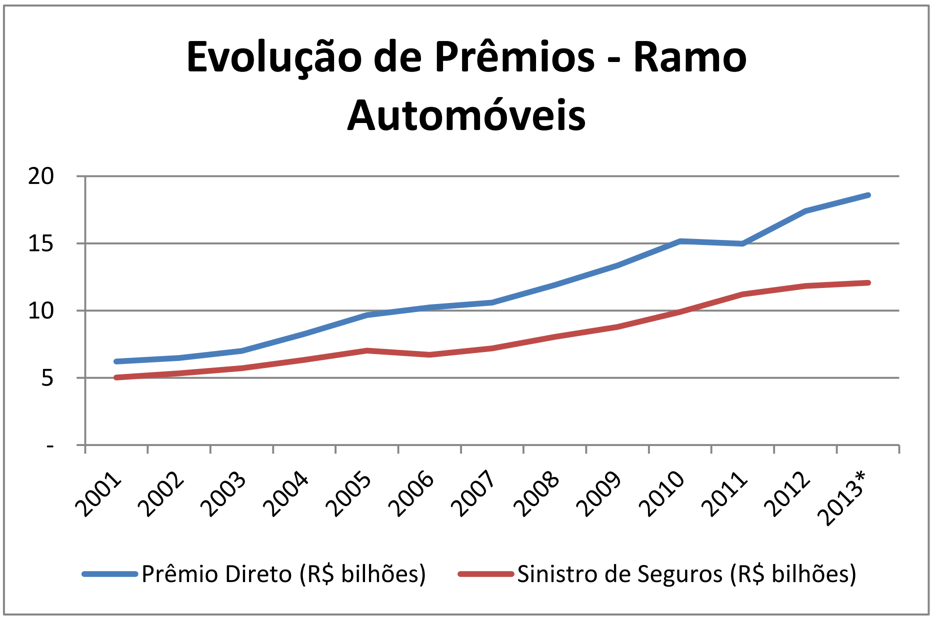

Figura 1 - Evolução de Prêmios - Ramo Automóveis

* Para o ano de 2013, ainda não estão disponibilizados os dados do mês de Dezembro.

Além disso, destacamos no gráfico da figura 1 que o volume de sinistros neste ramo cresceu em menor proporção que o volume de prêmios, o que é explicado pelo aperfeiçoamento dos estudos de sinistralidade nas seguradoras, visto que é cada vez maior a necessidade de se conhecer bem o risco assumido.

O gráfico da figura 2 mostra o aumento da frota de veículos em nosso país desde o ano de 1998 até 2013, com base em dados do DENATRAN - Departamento Nacional de Trânsito. Observamos que a frota de veículos quase triplicou nesse período de 15 anos.

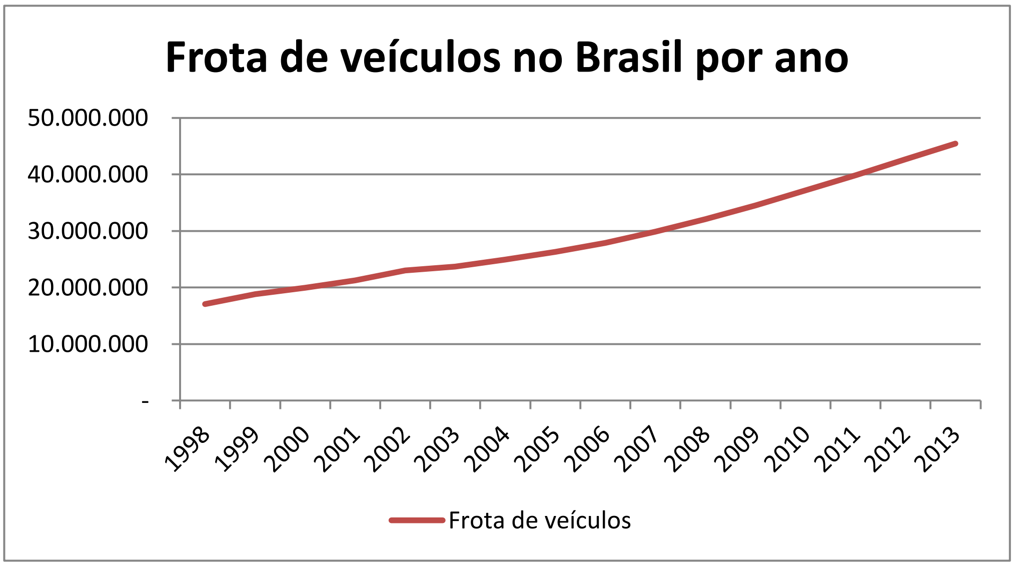

Figura 2 - Frota de veículos no Brasil por ano

A estatística apresentada na figura 2 explica em grande parte o crescimento do ramo de seguros de automóveis, verificado na figura 1.

Para demonstrar a representatividade deste ramo perante os demais, as figuras 3, 4 e 5 apresentam a distribuição do volume de prêmios diretos entre os principais ramos de seguros, indicando a ocorrência de grandes mudanças no padrão de distribuição, no período de 2001 a 2012. O ramo de seguros de Automóveis - Casco apresentava o maior volume de prêmios em 2001, tendo sido ultrapassado em 2006 pelo VGBL, produto de acumulação de recursos, inserido no âmbito dos seguros de pessoas. Essa tendência se acentuou em 2012, com participação ainda maior do VGBL e de outros ramos de seguros de pessoas, tais como os seguros prestamista, de acidentes pessoais e de vida em grupo.

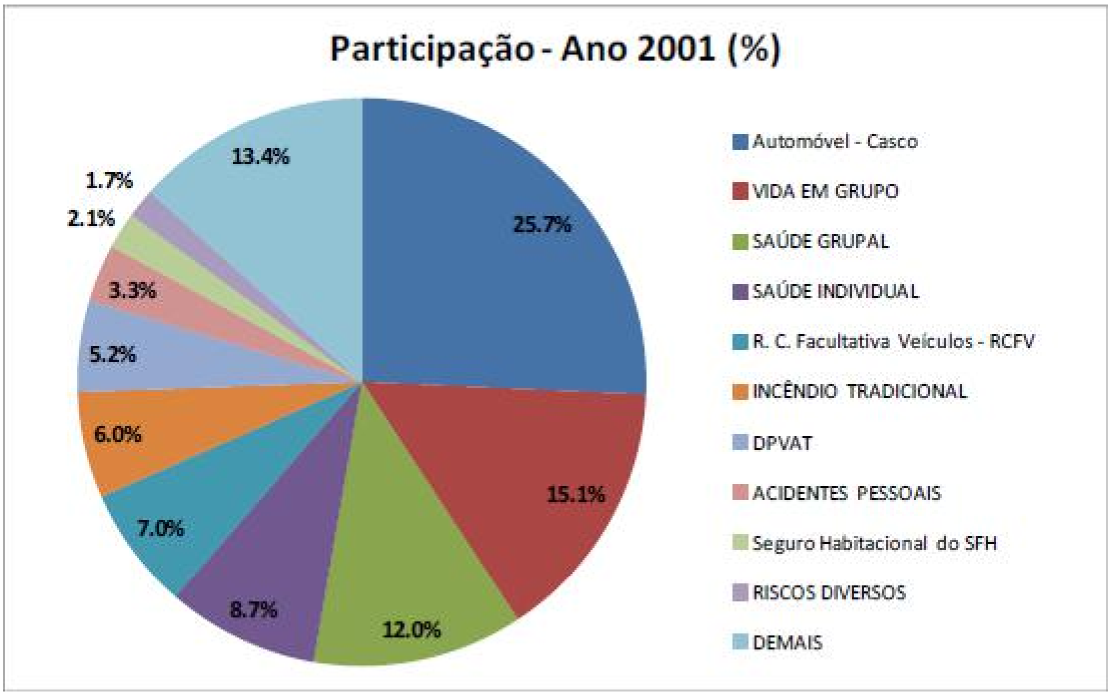

Figura 3 - Distribuição do volume de prêmios entre os principais ramos de seguros - 2001

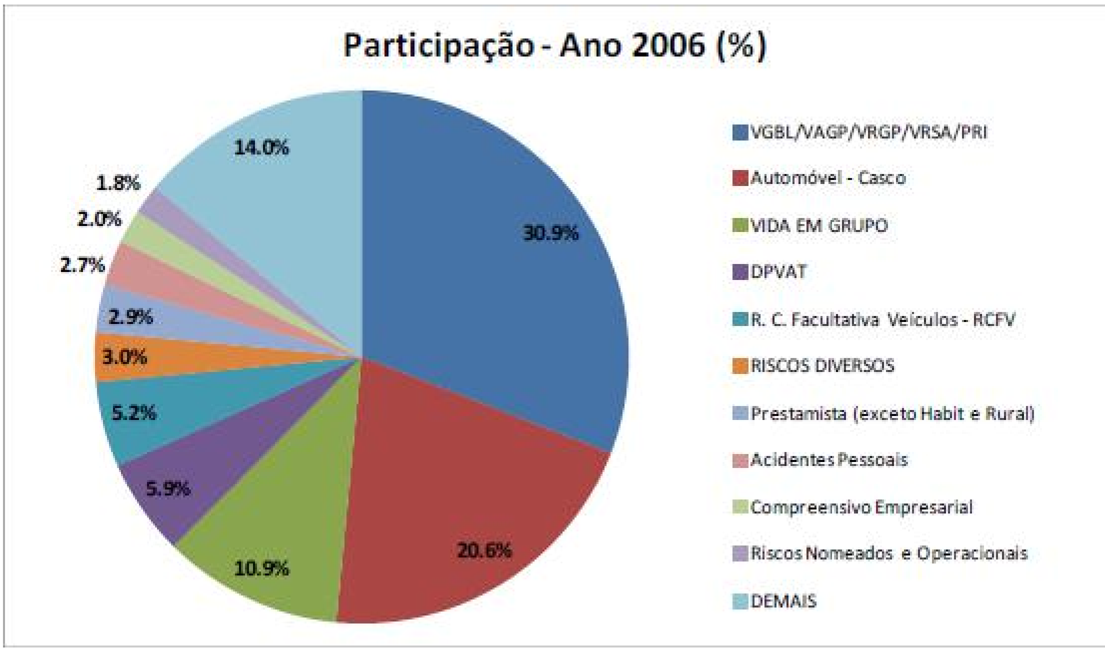

Figura 4 - Distribuição do volume de prêmios entre os principais ramos de seguros - 2006

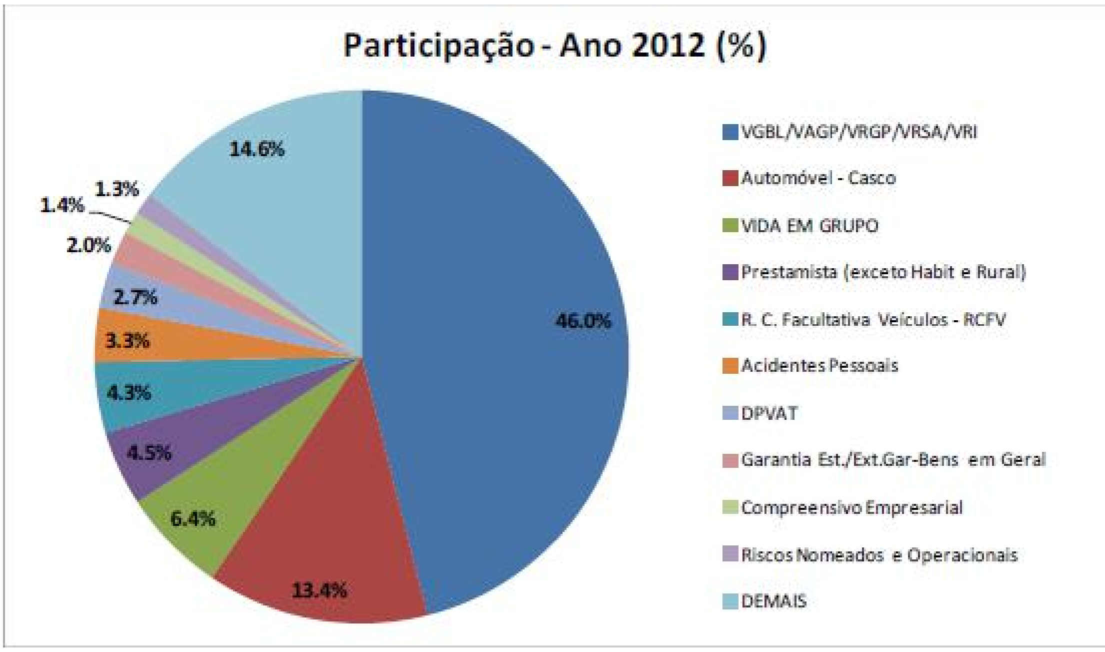

Figura 5 - Distribuição do volume de prêmios entre os principais ramos de seguros - 2010

3. Modelo do Risco Individual

Neste Modelo, todo o enfoque para a obtenção do valor total de sinistros é individual, pois utilizamos as distribuições das variáveis “Valor do sinistro” e “Ocorrência do sinistro” individualmente em cada apólice. A descrição do modelo foi baseada em FERREIRA (2010).

Comecemos pela descrição das hipóteses para elaboração do modelo:

- É conhecida a probabilidade de ocorrência de sinistros no período estudado de cada apólice, que será denotada por q_i;
- É conhecida a distribuição da variável aleatória “Valor do sinistro de cada apólice”, que será denotada por B_i;
- É desconsiderada a probabilidade de ocorrência de mais de 1 sinistro por apólice;
- É conhecido e fixo o número de apólices, não levando em conta entradas e saídas;
- Os riscos assumidos em cada apólice são independentes.

Notação:

Seja $S^{ind} = X_1 + X_2 + ... + X_N$

Em que $X_1 \perp X_2 \perp ... \perp X_N$ (ou seja, são independentes) e $X_i = I_i B_i$

S^{ind} é a variável aleatória que representa o valor total das indenizações na carteira no período;

X_i é a variável aleatória que representa o valor do sinistro da *i* -ésima apólice, caso tenha ocorrido. (X_i assume valor 0 caso não tenha ocorrido sinistro);

I_i é a variável aleatória indicadora que representa se houve ocorrência ou não de sinistro da apólice i no período, ou seja:

$$I_i = \begin{cases} 1 & \text{com probabilidade } q_i \\ 0 & p_i = 1 - q_i \end{cases}$$

B_i é a variável aleatória que representa o valor do sinistro da apólice *i* dado que o sinistro ocorreu no período.

3.1. Cálculo da função de distribuição acumulada de X_i

$$F_{X_i}(x) = P(X_i \leq x) = \sum_{k=0}^{1} P(X_i \leq x, I_i = k) = P(X_i \leq x \mid I_i = k)P(I_i = 1) + P(X_i \leq x \mid I_i = 0)P(I_i = 0)$$
$$= F_{B_i}(x) q_i + I(x) p_i$$

em que $I(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$ e $F_{B_i}(x)$ representa a função de distribuição acumulada de B_i.

3.2. Obtenção da distribuição de S^{ind}

Podemos obter a distribuição de S^{ind} de duas formas:

3.2.1. Por Convolução, a partir das distribuições de X_i

$$F_{S^{ind}}(x) = \sum_{n=0}^{\infty} P(S^{ind} \leq x) = F_{X_1} * F_{X_2} * ... * F_{X_n}$$

Onde o sinal de asterisco ($*$) representa a convolução entre as funções.

O processo de Convolução é um processo recursivo, onde primeiro calculamos a distribuição de X_1 e, a partir da distribuição de X_1, se calcula a distribuição de $X_1 + X_2$ e, assim sucessivamente, até se calcular a distribuição de $S^{ind} = X_1 + X_2 + ... + X_n$.

3.2.2. Pela Função Geratriz de Momentos

Notação: Seja $M_X(t)$ a função geratriz de momentos da variável aleatória X.

$$M_{S^{ind}}(t) = E[e^{tS^{ind}}] = E[e^{t(X_1+X_2+...+X_n)}] = E[e^{tX_1}]E[e^{tX_2}]...E[e^{tX_n}] = M_{X_1}(t)M_{X_2}(t)...M_{X_n}(t)$$

Assim se conhecermos $M_{X_i}(t)$, obteremos $M_{S^{ind}}(t)$.

3.3. Cálculo de $E[S^{ind}]$ e $Var[S^{ind}]$

Para esses cálculos, assumiremos as seguintes hipóteses adicionais: $I_i \perp B_i$ e I_i iid.

Ou seja, o valor do sinistro em cada apólice independe da sua ocorrência e as variáveis aleatórias "Ocorrência de sinistro" em cada apólice são independentes e identicamente distribuídas.

Desta forma, então,

$$E[S^{ind}] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} E[I_i]E[B_i] = \sum_{i=1}^{n} q_i E[B_i]$$

$$Var[S^{ind}] = \sum_{i=1}^{n} Var[X_i] = \sum_{i=1}^{n} E[Var[X_i \mid I_i]] + Var[E[X_i \mid I_i]] =$$

$$= \sum_{i=1}^{n} Var[X_i \mid I_i = 0]P(I_i = 0) + Var[X_i \mid I_i = 1]P(I_i = 1) + Var[E[I_i B_i \mid I_i]] =$$

$$= 0 + \sum_{i=1}^{n} Var(B_i) q_i + Var[I_i E[B_i]] = \sum_{i=1}^{n} Var(B_i) q_i + E[B_i]^2 Var(I_i) =$$

$$= \sum_{i=1}^{n} Var(B_i) q_i + E[B_i]^2 q_i p_i$$

4. Modelo do Risco Coletivo

No Modelo do Risco Coletivo, analisamos a distribuição dos sinistros como um todo, em uma carteira. É dispensável que conheçamos o comportamento dos sinistros individuais, como seria necessário no Modelo do Risco Individual. Esta seção foi baseada em FERREIRA (1998).

Notação:

$$S^{col} = X_1 + X_2 + ... + X_N$$

em que:

S^{col} é a variável aleatória que representa o sinistro agregado da carteira no período;

N é a variável aleatória que representa o número de sinistros no período;

X_i é a variável aleatória que representa o valor do *i*-ésimo sinistro da carteira;

Pode-se notar que S^{col} é uma soma aleatória de N parcelas, onde N é uma v.a., das variáveis aleatórias X_i.

Para tornar o modelo aplicável na prática, usualmente são feitas as seguintes suposições fundamentais:

i. $X_1, X_2, ..., X_N$ são independentes e identicamente distribuídos
ii. $X_1, X_2, ..., X_N$ são independentes de N

4.1. Obtenção da distribuição de S^{col}

Podemos obter a distribuição de S^{col} de três formas:

4.1.1. Por Convolução, a partir das distribuições de X e N

$$F_{S^{col}}(x) = \sum_{n=0}^{\infty} P(S^{col} \le x | N = n) P(N = n)$$

$$= \sum_{n=0}^{\infty} P(X_1 + X_2 + \ldots + X_N \le x) P(N = n)$$

$$= \sum_{n=0}^{\infty} P^{*n}(x) P(N = n)$$

$$f_{S^{col}}(x) = \sum_{n=0}^{\infty} p^{*n}(x) P(N = n)$$

Observações:

i. $P^{*n}(x)$ e $p^{*n}(x)$ são, respectivamente, a função de distribuição acumulada e a função de probabilidade da variável aleatória do valor de n sinistros.

ii. Se X tem distribuição discreta, então S^{col} terá distribuição discreta; Se X tem distribuição contínua, então S^{col} terá distribuição contínua.

Para calcular $p^{*n}(x)$ e $P^{*n}(x)$ utiliza-se o processo de Convolução, conforme desenvolvido seguir:

- X discreto

Seja y um dos possíveis valores que X pode assumir, então,

$$p^{*n}(x) = P(X_1 + X_2 + \ldots + X_N = x)$$

$$= \sum_{y} P(X_1 + X_2 + \ldots + X_{n-1} = x - y) P(X_n = y)$$

$$= \sum_{y} p^{*n-1}(x - y) p(y)$$

onde $p^{*n}(x)$ é chamada de n-ésima convolução de $p(x)$, e pode ser representada por $p^{*n} = p^{*n-1} * p$

Da mesma forma, $P^{*n}(x) = \sum_{y} P^{*n-1}(x-y)p(y)$. Onde $P^{*n} = P^{*n-1} * P$

- X contínuo

$$p^{*n}(x) = \int_{y} p^{*n-1}(x-y)p(y)dy = p^{*n-1} * p$$

O desenvolvimento dessa igualdade é análogo ao caso discreto, substituindo o somatório por integral já que se tratam agora de variáveis contínuas.

Da mesma forma, $P^{*n}(x) = \int_{y} P^{*n-1}(x-y)p(y)dy$

Logo, repetindo esse procedimento n vezes, temos que $P^{*n} = P * P * P * ... * P$

Ou seja, se $M_X(t)$ é a Função Geratriz de Momentos associada a $P(x)$, então, a Função Geratriz de Momentos associada a $P^{*n}(x)$ será: $M_{X^{*n}}(t) = [M_X(t)]^n$

Observação:

A determinação da distribuição de S^{col} por esse método é bastante trabalhosa, tanto quando X possui distribuição paramétrica conhecida, o que implica em cálculos complexos de integrais e somatórios, tanto quando trabalhamos com distribuição empírica para X, o que requer recursos computacionais não triviais.

4.1.2. Pela Função Geratriz de Momentos

Sabemos que:

$$M_X(t) = E[e^{tX}]$$

$$M_N(t) = E[e^{tN}]$$

$$M_{S^{col}}(t) = E[e^{tS^{col}}] = E\left\lfloor E[e^{tS^{col}} \middle| N]\right\rfloor$$

$$E[e^{tS^{col}} | N = n] = E[e^{t(X_1 + X_2 + ... + X_n)}] = E[e^{tX_1} e^{tX_2} ... e^{tX_n}] = M_{X_1}(t) M_{X_2}(t) ... M_{X_n}(t) = M_X(t)^n$$

Pois as variáveis X_i são independentes e o conhecimento de N afeta apenas o número de parcelas na exponencial.

Então,

$$M_{S^{col}}(t) = E[M_X(t)^N] = E[e^{N \log M_X(t)}] = M_N(\log M_X(t))$$

Dessa forma, se conhecermos as distribuições de X e N, então conheceremos $M_N(t)$, $M_X(t)$ e, consequentemente, $M_{S^{col}}(t)$.

4.1.3. Pelo Teorema Central do Limite

O Teorema Central do Limite aplicado a variáveis i.i.d's é um importante resultado da estatística, que é enunciado da seguinte forma:

Seja $X_1, X_2, \cdots X_N$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas com média $E[X]$ finita e $0 < Var(X) < \infty$. Então, vale a seguinte convergência em distribuição:

$$\frac{\sum_{i=1}^{n} X_i - E\left[\sum_{i=1}^{n} X_i\right]}{\sqrt{Var\left[\sum_{i=1}^{n} X_i\right]}} \xrightarrow{D} N(0,1), \text{ quando n} \to \infty$$

Então, precisamos conhecer $E[S^{col}]$ e $Var[S^{col}]$ pois $S^{col} = \sum_{i=1}^{n} X_i$

- Cálculo de $E[S^{col}]$

$$E[S^{col}] = E\left[E\left[S^{col} \middle| N\right]\right] = E\left[E\left[X_1 + X_2 + ... + X_N \middle| N\right]\right] = E\left[NE\left[X\right]\right] = E[N]E[X]$$

pois os X_i são variáveis aleatórias independentes e identicamente distribuídas com distribuição X.

Este resultado é bastante intuitivo pois o valor esperado do sinistro agregado é igual ao número médio de sinistros multiplicado pelo valor médio de 1 sinistro.

Podemos também calcular $E[S^{col}]$ do seguinte modo, utilizando a função geradora de momentos:

$$E[S^{col}] = M'_{S^{col}}(0)$$

$$M'_{S^{col}}(t) = M'_N(\log M_X(t)) \frac{d}{dt} \log M_X(t) = M'_N(\log M_X(t)) \frac{M'_X(t)}{M_X(t)}$$

$$M'_{S^{col}}(0) = M'_N(\log 1) \frac{M'_X(0)}{1} = M'_N(0) M'_X(0) = E[N]E[X]$$

- Cálculo de $Var[S^{col}]$

$$Var[S^{col}]=E\left[Var\left[S^{col}\middle|N\right]\right]+Var\left[E\left[S^{col}\middle|N\right]\right]=E\left[Var\left[X_1+X_2+...+X_N\middle|N\right]\right]+$$
$$+Var\left[E\left[X_1+X_2+...+X_N\middle|N\right]\right]=E\left[NVar\left[X\right]\right]+Var\left[NE\left[X\right]\right]=Var[X]E[N]+E[X]^2Var[N]$$

pois os X_i são variáveis aleatórias independentes e identicamente distribuídas com distribuição X.

Este resultado nos mostra que a variância do sinistro agregado é diretamente proporcional à variância do número de sinistros e à variância do valor de 1 sinistro.

Podemos, também, calcular $Var[S^{col}]$ do seguinte modo, utilizando função geradora de momentos:

$$Var[S^{col}]=M''_{S^{col}}(0)-E[S^{col}]^2=M''_{S^{col}}(0)-E[X]^2E[N]^2$$

$$M''_{S^{col}}(t)=M''_N(\log M_X(t))\frac{M'_X(t)}{M_X(t)}\frac{M'_X(t)}{M_X(t)}+M'_N(\log M_X(t))\frac{M''_X(t)M_X(t)-M'_X(t)M'_X(t)}{M_X(t)^2}$$

$$M''_{S^{col}}(0)=M''_N(0)E[X]E[X]+M'_N(0)(E[X^2]-E[X]E[X])=E[N^2]E[X]^2+E[N](E[X^2]-E[X]^2)=$$
$$=E[N^2]E[X]^2+E[N]Var[X]$$

Logo,

$$Var[S^{col}]=M''_{S^{col}}(0)-E[X]^2E[N]^2=E[N^2]E[X]^2+E[N]Var[X]-E[X]^2E[N]^2=$$
$$=E[X]^2(E[N^2]-E[N]^2)+E[N]Var[X]=E[X]^2(Var[N])+E[N]Var[X]$$

4.2. Distribuição da variável aleatória Valor de 1 sinistro

A distribuição do valor de 1 sinistro é obtida a partir da observação histórica da carteira, levando-se em conta fatores tais como: tendência e sazonalidade.

Tendência e sazonalidade são tratadas estatisticamente com ferramentas de séries temporais. Nesse caso, destacamos o amortecimento exponencial, o qual atribui um peso maior às informações de anos mais recentes.

Para se obter essa distribuição podem ser utilizados métodos de obtenção paramétricos ou não-paramétricos.

Os métodos paramétricos são utilizados quando o número de dados é pequeno. Nesse caso, em função da experiência existente em relação a fenômenos semelhantes, atribuímos uma distribuição conhecida, como Log normal, Pareto ou Gama. A utilização de distribuições paramétricas também é conveniente quando se deseja fazer previsões em um tempo futuro.

Os métodos não-paramétricos devem ser utilizados quando o número de dados é grande. Nesse caso, aplicamos a distribuição empírica. Esses métodos podem ser satisfatórios quando se deseja apenas fazer descrições de dados históricos.

4.2.1. Principais distribuições paramétricas utilizadas nesse contexto

A seguir encontram-se algumas das distribuições paramétricas mais utilizadas na prática.

Log Normal

$$X \sim Log - normal(\mu, \sigma)$$

$$f_X(x) = \frac{1}{\sigma x \sqrt{2\pi}} Exp\left\{ -\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma} \right)^2 \right\}, \ x > 0$$

Os principais momentos da Log Normal são:

$$E\left[X^k\right] = Exp\left(k\mu + \frac{1}{2} k^2 \sigma^2 \right)$$

$$Média = Exp\left(\mu + \frac{1}{2}\sigma^2 \right)$$

$$Variância = Exp\left(2\mu + \sigma^2\right)\left[Exp\left(\sigma^2 - 1\right)\right]$$

Observação:

Se $X \sim Log - normal\left(\mu, \sigma^2\right) \rightarrow Y = LnX \sim Normal\left(\mu, \sigma^2\right)$

Pareto

$$X \sim Pareto(\lambda, \alpha)$$

$$f_X(x) = \frac{\alpha \lambda^\alpha}{(\lambda + x)^{\alpha+1}}, \ x > 0$$

Os principais momentos da Pareto são:

$$E\left[X^k\right] = \frac{\lambda^k k!}{\prod_{i=1}^{k}(\alpha - i)}, \ \alpha > k$$

$$Média = \frac{\lambda}{\alpha - 1}, \ \alpha > 1$$

$$Variância = \frac{\lambda^2 \alpha}{(\alpha - 1)^2 (\alpha - 2)}, \ \alpha > 2$$

Gama

$$X \sim Gama(\alpha, \beta)$$

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha - 1}, \ x \geq 0$$

Os principais momentos da Gama são:

$$E\left[X^k\right] = \frac{\prod_{i=0}^{k-1}(\alpha + i)}{\beta^k}$$

$$Média = \frac{\alpha}{\beta}$$

$$Variância = \frac{\alpha}{\beta^2}$$

Weibull

$$X \sim Weib(\lambda, \delta)$$

$$f_X(x) = \frac{\delta}{\lambda}\left(\frac{x}{\lambda}\right)^{\delta - 1} \exp\left\{-\left(\frac{x}{\lambda}\right)^{\delta}\right\}, \ x \geq 0$$

Os principais momentos da Weibull são:

$$Média = \lambda \Gamma\left(1 + \frac{1}{\delta}\right)$$

$$Variância = \lambda^2 \left[\Gamma\left(1+\frac{2}{\delta}\right) - \left(\Gamma\left(1+\frac{1}{\delta}\right)\right)^2 \right]$$

O ajuste é feito em duas etapas. A primeira é a determinação dos parâmetros, que pode ser feita de acordo com os métodos dos momentos, mínimos quadrados ou máxima verossimilhança. A segunda etapa é o teste de aderência, no qual comparamos a distribuição ajustada com a distribuição analítica através de um teste de aderência, como o Teste Qui-quadrado ou teste de Kolmogorov-Smirnov.

4.3. Distribuições para o Número de Sinistros

Conforme já vimos, os principais resultados necessários para utilizarmos o Modelo Coletivo dependem das distribuições de X e N.

Já tratamos das distribuições para o ajuste de X e agora veremos duas distribuições importantes para o ajuste de N: Poisson e Binomial Negativa.

Quando N tem distribuição de Poisson, dizemos que S^{col} tem distribuição de Poisson Composta e, quando N tem distribuição de Binomial Negativa, dizemos que S^{col} tem distribuição de Binomial Negativa Composta.

4.3.1. Distribuição de Poisson para N

Na maioria dos casos o processo de ocorrência de sinistros satisfaz as condições do Processo de Poisson, que são:

- A distribuição do processo do número de sinistros no intervalo de tempo t (N_t) só depende da duração desse intervalo t, não sendo portanto afetado pelo instante em que o processo iniciou;
- Só é possível 1 sinistro no intervalo dt, ou seja, como tratamos o tempo como sendo contínuo, a probabilidade de mais de 1 sinistro ocorrer exatamente no mesmo instante é nula;
- As variáveis aleatórias "número de sinistros" em intervalos de tempo disjuntos são independentes;
- A probabilidade de 1 sinistro no intervalo dt é λdt;
- $P(N_0 = 0) = 1$, ou seja, não há possibilidade de haverem sinistros anteriores ao instante inicial do processo.

Assim, $N_t \sim Poisson\ (\lambda t)$

Algumas propriedades da Poisson

- Seja T a variável aleatória “intervalo de tempo entre 2 sinistros”, então temos que $T \sim Exp(\lambda)$, pois:

$$P(T > t) = P(N_t = 0) = e^{-\lambda t}$$

Como $1 - e^{-\lambda t}$ é exatamente a função distribuição acumulada da $Exp(1)$ avaliada em t, segue que $T \sim Exp(\lambda)$

Além disso, segue que $E[T] = \dfrac{1}{\lambda}$.

Ou seja, o tempo médio entre a ocorrência de 2 sinistros é igual a $1/\lambda$. Este resultado pode ser útil na definição do número de reguladores de sinistros pela seguradora, por exemplo.

- A média é igual à variância: $E[N] = Var(N) = \lambda$
- Sua função geradora de momentos é dada $M_N(t) = \exp\{\lambda(e^t - 1)\}$

4.3.2. Distribuição Binomial Negativa para N

Quando há indícios que $Var(N) > E[N]$, então a distribuição de Poisson para N não é adequada, pois ela estaria subestimando a variância. Uma boa alternativa é utilizar a distribuição Binomial Negativa.

Notação: $N \sim Binomial\ Negativa\,(r, p) \iff N \sim NB(r, p)$

$$P(N = n) = \binom{r + n - 1}{n} p^r q^n, \quad n = 0,1,2,\ldots, \quad \begin{array}{l} r > 0 \\ 0 < p < 1 \\ q = 1 - p \end{array}$$

Nota: Está sendo utilizada a interpretação de que a Binomial Negativa modela número de fracassos até atingir *r* sucessos.

Algumas propriedades da Binomial Negativa

- Sua função geradora de momentos é dada por $M_N(t) = \left(\dfrac{p}{1 - qe^t}\right)^r$;
- $E[N] = \dfrac{rq}{p}$, $Var(N) = \dfrac{rq}{p^2}$. Logo, $Var(N) > E[N]$ pois $p^2 < p$;
- Se $r = 1$, $P(N = n) = pq^n$, que é a função de probabilidade da distribuição Geométrica (p);

- $NB\left(r, \frac{\lambda}{\lambda + r}\right) \xrightarrow{r \to \infty} Pois(\lambda)$, ou seja, se $r \to \infty$, então a Binomial Negativa converge em distribuição para uma Poisson;
- Se N_i, com $i = 1,...,r$ são variáveis aleatórias i.i.d's com distribuição Geométrica(p), então,

$$N = \sum_{i=1}^{r} N_i \sim NB(r, p)$$

4.4. Distribuições para o Sinistro Agregado

4.4.1. Distribuição de Poisson Composta para S^{col}

Quando N possui distribuição de Poisson(λ), dizemos que S^{col} possui distribuição de Poisson Composta $(\lambda, P(x))$, de modo que:

$$F_{S^{col}}(x) = \sum_{n=0}^{\infty} P^{*n}(x) \frac{e^{-\lambda}(\lambda)^n}{n!}$$

Vamos agora encontrar a Função Geratriz de Momentos da Poisson Composta

Conforme demonstrado na seção 4.1.2, a Função Geratriz de Momentos de S^{col} é expressa por:

$$M_{S^{col}}(t) = M_N(\log M_X(t))$$

Além disso, a Função Geratriz de Momentos da Poisson é expressa por:

$$M_N(t) = \exp\{\lambda(e^t - 1)\}$$

Conjugando esses dois resultados, temos que:

$$M_{S^{col}}(t) = e^{\lambda(M_X(t) - 1)}$$

Temos que a média e variância da Poisson Composta são dadas por:

$$E[S^{col}] = E[N]E[X] = \lambda E[X]$$

$$Var[S^{col}] = E[X]^2 Var[N] + E[N] Var[X] = E[X]^2 \lambda + \lambda Var[X] = \lambda E[X^2]$$

Uma importante propriedade da Poisson Composta é a seguinte:

Sejam S_1^{col}, S_2^{col}, ... , S_m^{col} variáveis aleatórias independentes de modo que:

$$S_i^{col} \sim Poisson\,Composta\,(\lambda_i, P_i(x))$$

Então,

$$S^{col} = \sum_{i=1}^{m} S_i^{col} \sim Poisson\,Composta\,(\lambda, P(x))$$

onde $\lambda = \sum_{i=1}^{m} \lambda_i$ e $P(x) = \sum_{i=1}^{m} \frac{\lambda_i}{\lambda} P_i(x)$

Dessa propriedade decorrem algumas aplicações práticas, apresentadas a seguir:

- Várias partições de carteira com distribuição de Poisson Composta geram carteira com distribuição de Poisson Composta;
- Vários subintervalos de tempo com distribuição de Poisson Composta geram um intervalo de tempo, que é a soma dos subintervalos, com distribuição de Poisson Composta.

4.4.2. Distribuição de Binomial Negativa Composta para S^{col}

Quando N possui distribuição de Binomial Negativa (r, p), dizemos que S^{col} possui distribuição de Binomial Negativa Composta $(r, p, P(x))$, de modo que:

$$F_{S^{col}}(x) = \sum_{n=0}^{\infty} P^{*n}(x) \binom{r+n-1}{n} p^r q^n$$

Vamos agora encontrar a Função Geratriz de Momentos da Binomial Negativa Composta:

Novamente, conforme demonstrado na seção 4.1.2, a Função Geratriz de Momentos de S^{col} é expressa por:

$$M_{S^{col}}(t) = M_N(\log M_X(t))$$

Sabemos também que a Função Geratriz de Momentos da Binomial Negativa é expressa por:

$$M_N(t) = \left(\frac{p}{1-qe^t} \right)^r$$

Conjugando esses dois resultados, temos que:

$$M_{S^{col}}(t) = \left(\frac{p}{1-qM_X(t)} \right)^r$$

Temos que a média e variância da Binomial Negativa Composta são dadas por:

$$E[S^{col}] = E[N]E[X] = \frac{rq}{p}E[X]$$

$$Var[S^{col}] = E[X]^2 Var[N] + E[N]Var[X] = \frac{rq}{p^2}E[X]^2 + \frac{rq}{p}\left[E[X^2] - E[X]^2\right] =$$

$$= \frac{rq}{p}E[X^2] + E[X]^2\left(\frac{rq}{p^2} - \frac{rq}{p}\right) = \frac{rq}{p}E[X^2] + E[X]^2\,\frac{rq^2}{p^2}$$

4.4.3. Aproximações para S^{col}

Conforme visto na seção 4.1, a determinação da distribuição de S^{col} é extremamente trabalhosa, seja por convolução ou a partir da Função Geratriz de Momentos. Vamos apresentar então as aproximações Normal e Gama para S^{col}, que são muito utilizadas na prática.

Aproximação Normal para S^{col}

Temos que S^{col} é a soma de variáveis aleatórias X_i independentes e identicamente distribuídas, então, pelo Teorema Central do Limite, é sabido que quando o número de sinistros for suficientemente grande, S^{col} terá distribuição aproximadamente Normal com média $E[S^{col}]$ e variância $Var(S^{col})$, ou seja:

$$\frac{S^{col} - E[S^{col}]}{\sqrt{Var(S^{col})}} \sim N(0,1)$$

Uma utilização prática desse resultado é, por exemplo, o cálculo do total de prêmio puro (P) tal que:

$P(S^{col} > P) = \alpha$, onde $P = E[S^{col}] + Z_{1-\alpha}\sqrt{Var(S^{col})}$

Quando $S^{col} \sim Poisson\ \ Composta\ \ (\lambda, P(x))$, foi demonstrado no item 4.4.1 que:

$$E[S^{col}] = \lambda E[X]$$

$$Var(S^{col}) = \lambda^2 E[X^2]$$

Então, ao aplicarmos a aproximação Normal, chegamos ao seguinte resultado:

$$\frac{S^{col} - \lambda E[X]}{\sqrt{\lambda E[X^2]}} \xrightarrow[\lambda \to \infty]{} N(0,1)$$

Quando $S^{col} \sim Binomial\ Negativa\ Composta\ (r, p, P(x))$, foi demonstrado no item 4.4.2 que:

$$E[S^{col}] = r\frac{q}{p}E[X]$$

$$Var[S^{col}] = r\frac{q}{p}E[X^2] + r\frac{q^2}{p^2}E[X]^2$$

Então, ao aplicarmos a aproximação Normal, chegamos ao seguinte resultado:

$$\frac{S^{col} - r\frac{q}{p}E[X]}{\sqrt{r\frac{q}{p}E[X^2] + r\frac{q^2}{p^2}E[X]^2}} \xrightarrow[r \to \infty]{} N(0,1)$$

Nota: A aproximação Normal costuma ser muito boa no extremo superior da distribuição de S^{col}, mesmo quando o número esperado de sinistros é pequeno, o que tem uma aplicação prática muito grande pois são exatamente os valores mais elevados de S^{col} que mais interessam nos cálculos atuariais, uma vez que eles podem trazer complicações para a seguradora, em termos de solvência.

Aproximação Gama para S^{col}

Se tivermos indícios de que a assimetria da distribuição de S^{col} é muito forte, então não é adequado utilizar a aproximação Normal. Uma alternativa é aplicar a aproximação Gama, que tem assimetria à direita.

Um inconveniente dessa aproximação é que estão sendo consideradas probabilidades positivas a montante de indenizações desprezíveis (ε). Porém, isto é muito difícil de acontecer com um grande número de apólices na carteira, ou seja:

$$P(0 < S^{col} < \varepsilon) \to 0$$

Para solucionar esse problema, faz-se uma translação de *t* para direita na curva da função de densidade, onde *t* é o montante mínimo de indenizações, ou seja:

$$H(x, \alpha, \beta, t) = G(x - t, \alpha, \beta)$$

H é chamada de distribuição Gama Transladada de t. A forma geral da densidade H pode ser vista na figura 6.

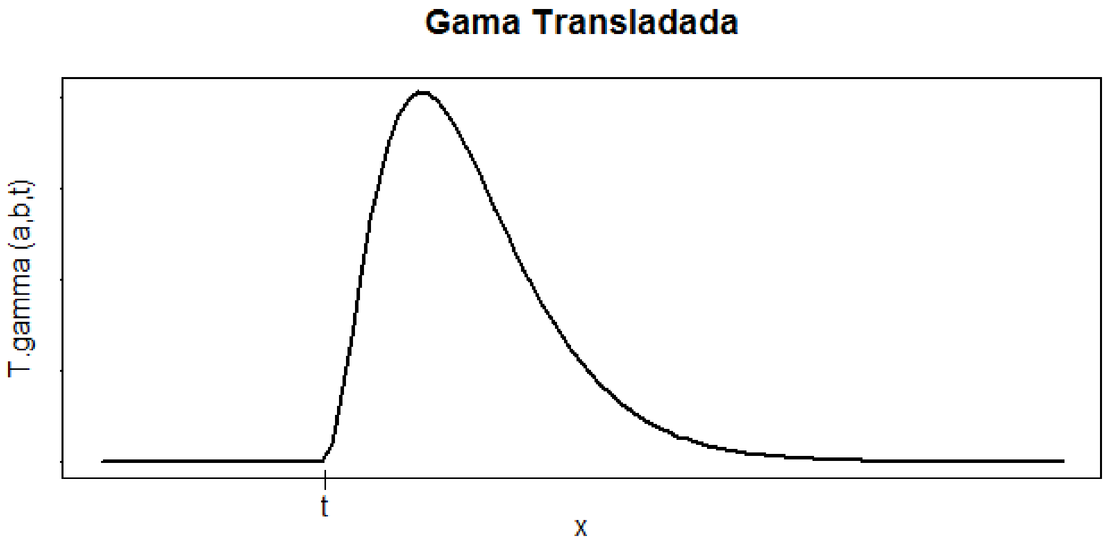

Figura 6 - Densidade da Gama (a,b) transladada de t para a direita

Assim sendo, se S^{col} tem distribuição Gama Transladada, então:

$$F_{S^{col}}(x) = H(x,\alpha,\beta,t) = G(x-t,\alpha,\beta)$$

Para encontrar os parâmetros dessa gama transladada, será necessário utilizar as seguintes relações associadas aos momentos da distribuição gama transladada:

$$E[S^{col}] = t + \frac{\alpha}{\beta}$$

$$Var[S^{col}] = \frac{\alpha}{\beta^2}$$

$$E\left[\left(S^{col} - E[S^{col}]\right)^3\right] = \frac{2\alpha}{\beta^3}$$

Resolvendo esse sistema de 3 incógnitas e 3 equações, temos:

$$\alpha = \frac{4\left(Var(S^{col})\right)^3}{E\left[\left(S^{col} - E[S^{col}]\right)^3\right]^2}$$

$$\beta = \frac{2\,Var(S^{col})}{E\left[\left(S^{col} - E[S^{col}]\right)^3\right]}$$

$$t = E[S^{col}] - \frac{2\left(Var(S^{col})\right)^2}{E\left[\left(S^{col} - E[S^{col}]\right)^3\right]}$$

Então, para calcularmos α, β, t que melhor aproximam uma distribuição Gama à variável S^{col} é preciso conhecer $E[S^{col}], Var[S^{col}], E\left[\left(S^{col} - E[S^{col}]\right)^3\right]$.

Quando $S^{col} \sim Poisson\,Composta\,(\lambda, P(x))$ temos que:

Conforme demonstrado no item 4.4.1, temos:

$$E[S^{col}] = \lambda E[X]$$

$$Var(S^{col}) = \lambda E[X^2]$$

$$E\left[\left(S^{col} - E[S^{col}]\right)^3\right] = \lambda E[X^3]$$

Logo, os parâmetros da aproximação gama transladada são estimados por:

$$\alpha = \frac{4\lambda\left(E[X^2]\right)^3}{E[X^3]^2}$$

$$\beta = 2\frac{E[X^2]}{E[X^3]}$$

$$t = \lambda E[X] - 2\lambda\frac{(E[X^2])^2}{E[X^3]}$$

Quando $S^{col} \sim Binomial\ Negativa\ Composta\ (r, p, P(x))$, temos que:

$$E[S^{col}] = r\frac{q}{p}E[X]$$

$$Var[S^{col}] = r\frac{q}{p}E[X^2] + r\frac{q^2}{p^2}E[X]^2$$

$$M_{S^{col}}(t) = \left(\frac{p}{1 - qM_X(t)}\right)^r$$

Com a função geradora de momentos, pode-se mostrar que:

$$E\left[\left(S^{col} - E[S^{col}]\right)^3\right] = \frac{d^3}{dt}\log M_{S^{col}}(t)\Big|_{t=o} = \frac{rq}{p}E[X^3] + \frac{3rq^2}{p^2}E[X]E[X^2] + \frac{2rq^3}{p^3}E[X]^3$$

Agora, basta substituir esses valores nas igualdades de α, β, t para encontrá-los.

Nota: a aproximação Gama pode ser vista como uma generalização da distribuição Normal. Observe:

$$\begin{array}{l} \alpha \to \infty \\ \beta \to \infty \\ t \to \infty \end{array} \text{ tal que} \begin{cases} t + \alpha/\beta = \mu(cte) \\ \\ \alpha/\beta^2 = \sigma^2(cte) \end{cases} \to H(x, \alpha, \beta, t) \xrightarrow{D} N(\mu, \sigma^2)$$

5. Fórmula Recursiva de Panjer

A obtenção da distribuição exata da variável aleatória "Valor total dos sinistros" é extremamente trabalhosa, conforme abordado anteriormente. Para contornar esse inconveniente, foi desenvolvida a Fórmula Recursiva de Panjer, que permite a obtenção da distribuição exata do sinistro agregado de modo mais simples, a partir de recursos computacionais menos complicados. Os resultados e demonstrações apresentados a seguir foram baseados em FERREIRA (1998).

A base para essa fórmula foi a descoberta por PANJER que diversas distribuições discretas possuem a seguinte característica:

$$P_n = \left(a + \frac{b}{n} \right) P_{n-1}, \ n = 1,2,...$$

Sendo a, b constantes que dependem da distribuição e P_n a função de probabilidade no ponto n.

Abaixo temos a verificação dessa característica em algumas distribuições discretas mais usuais:

- $Poisson \ (\lambda)$

$$P_n = \frac{\lambda^n e^{-\lambda}}{n!} = \frac{\lambda}{n} P_{n-1}$$

Então, temos $a = 0$ e $b = \lambda$.

- *Geométrica* (p)

$$P_n = pq^n = \left(q + \frac{0}{n}\right)P_{n-1}$$

Então, temos $a = q$ e $b = 0$.

- *Binomial Negativa* (r, p)

$$P_n = \frac{(r+n-1)!}{(r-1)!n!}p^r q^n = \frac{r+n-1}{n}qP_{n-1}$$

Então, temos $a = q$ e $b = (r-1)q$.

As distribuições que possuem essa característica pertencem à chamada família (a,b) de Panjer.

A fórmula recursiva de Panjer é dada por:

$$P(S=0) = P(N=0)$$
$$P(S=n) = \frac{1}{1-aP(X=0)}\sum_{i=1}^{n}P(X=i)P(S=n-i)\left(a+b\frac{i}{n}\right), \text{para } n = 1,2,...$$

onde X é a variável aleatória "Valor de 1 sinistro" e S é variável aleatória Sinistro agregado.

Essa fórmula permite calcular a distribuição exata de S de forma recursiva, de modo que, dado que conhecemos a distribuição de X e conhecemos a distribuição de N e, consequentemente, os valores de a e b. Começamos calculando $P(S=0) = P(N=0)$. Em seguida, podemos calcular $P(S=\alpha_1)$, sendo α_1 o primeiro valor que S pode assumir após zero. Calculando $P(S=\alpha_1)$, podemos calcular $P(S=\alpha_2)$, sendo α_2 o primeiro valor que S pode assumir após α_1, e assim sucessivamente.

5.1. Demonstração da Fórmula Recursiva de Panjer

Seja

$$M_N(t) = E[t^N] = P_0 + P_1 t + P_2 t^2 + P_3 t^3 + K$$

Sendo P_k a probabilidade do número de sinistros em 1 ano ser igual a k.

$$M_X(t) = E[t^X] = f(t) = f_0 + f_1 t + f_2 t^2 + f_3 t^3 + K$$

Sendo f_k a probabilidade do valor de 1 sinistro ser igual a k.

$$M_S(t) = E[t^S] = g(t) = g_0 + g_1 t + g_2 t^2 + g_3 t^3 + K$$

Sendo g_k a probabilidade do valor do sinistro agregado ser igual a k.

Mas, note que,

$$g(t) = E[t^S] = E\big[E[t^S|N]\big] = E\big[E[t^{NX}]\big] = E\big[E[t^X]^N\big] = M_N[f(t)]$$

$$g' = M'_N(f)f'$$

Além disso,

$$P_n = \left(a + \frac{b}{n}\right)P_{n-1}$$

$$nP_n = anP_{n-1} + bP_{n-1} = a(n-1)P_{n-1} + (a+b)P_{n-1}$$

$$\sum_{n=1}^{\infty} nP_n t^n = \sum_{n=1}^{\infty} a(n-1)P_{n-1}t^n + (a+b)P_{n-1}t^n$$

$$tM'_N = at^2 M'_N + (a+b)tM'_N$$

$$M'_N = atM'_N + (a+b)M_N \quad \Rightarrow \quad (a+b)M_N - (1-at)M'_N = 0 \quad \Rightarrow \quad M'_N = \frac{a+b}{1-at}M_N$$

Logo,

$$g' = \frac{a+b}{1-af}M_N(f)f' = \frac{a+b}{1-af}gf' \quad \Rightarrow \quad g' - ag'f = (a+b)gf'$$

$$\sum_{n=0}^{\infty} ng_n t^{n-1} - a\left(\sum_{n=0}^{\infty} f_n t^n\right)\left(\sum_{n=0}^{\infty} ng_n t^{n-1}\right) = (a+b)\left(\sum_{n=0}^{\infty} nf_n t^{n-1}\right)\left(\sum_{n=0}^{\infty} g_n t^n\right)$$

$$\sum_{n=0}^{\infty} ng_n t^{n-1} - a\sum_{n=0}^{\infty}\left(\sum_{i=0}^{\infty} f_i(n-i+1)g_{n-i+1}\right)t^n = (a+b)\sum_{n=0}^{\infty}\left(\sum_{i=0}^{\infty} g_i(n-i+1)f_{n-i+1}\right)t^n$$

Para o coeficiente de t^n temos:

$$(n+1)g_{n+1} - a\sum_{i=0}^{\infty} f_i(n-i+1)f_i g_{n-i+1} = (a+b)\sum_{i=0}^{\infty}(n-i+1)g_i f_{n-i+1}$$

Seja $g_0 = f_0$, então,

$$(n+1)g_{n+1} - a(n+1)f_0 g_{n+1} - a\sum_{i=1}^{n}(n-i+1)f_i g_{n-i+1} = (a+b)(n+1)g_0 f_{n+1} - \sum_{i=1}^{n}(n-i+1)g_i f_{n-i+1}(a+b)$$

$$g_{n+1}\left((n+1) - a(n+1)f_0\right) = \sum_{i=1}^{n}(n-i+1)\left[af_i g_{n-i+1} - (a+b)g_i f_{n-i+1}\right]$$

$$g_n\left(n - anf_0\right) = \sum_{i=1}^{n}(n-i)\left[af_i g_{n-i} - (a+b)g_i f_{n-i}\right]$$

$$g_n = \frac{1}{1-af_0}\sum_{i=1}^{n}\left(1-\frac{i}{n}\right)\left[af_i g_{n-i+1} - (a+b)g_i f_{n-i}\right] = \frac{1}{1-af_0}K$$

$$K = \sum_{i=1}^{n} af_i g_{n-i} - (a+b)g_i f_{n-i} - \sum_{i=1}^{n}\frac{i}{n}(af_i g_{n-i} - (a+b)g_i f_{n-i})$$

$$K = \sum_{i=1}^{n}\left(1-\frac{i}{n}\right)af_i g_{n-i} - (a+b)\left(1-\frac{i}{n}\right)g_i f_{n-i}$$

$$K = \sum_{i=1}^{n}\left(1-\frac{i}{n}\right)af_i g_{n-i} - (a+b)\left(1-\frac{n-i}{n}\right)f_i g_{n-i}$$

$$K = \sum_{i=1}^{n} f_i g_{n-i}\left[\left(1-\frac{i}{n}\right)a + (a+b)\frac{i}{n}\right]$$

$$K = \sum_{i=1}^{n} f_i g_{n-i}\left[a + b\frac{i}{n}\right]$$

Logo,

$$g_n = \frac{1}{1-aP(X=0)}\sum_{i=1}^{n}P(X=i)P(S=n-i)\left(a+b\frac{i}{n}\right)$$

Ou seja,

$$P(S=n) = \frac{1}{1-aP(X=0)}\sum_{i=1}^{n}P(X=i)P(S=n-i)\left(a+b\frac{i}{n}\right)$$

6. Base e Metodologia

6.1. Descrição da base

Para aplicação do conhecimento teórico estudado, será utilizado um conjunto de dados com algumas informações sobre 420.390 apólices sinistradas no ramo de seguro de automóveis – Casco. Esse ramo apresenta hoje o segundo maior volume de prêmios entre todos os ramos de seguros, atrás apenas do ramo que engloba o VGBL. A demanda por esse tipo de seguro é crescente junto ao aumento das vendas de carros novos, o que desafia as seguradoras a criarem produtos menos custosos ou com coberturas mais sofisticadas e atraentes aos olhos do consumidor.

As informações de nossa base de dados são:

- Data do início de vigência
- Data do término de vigência
- Data da ocorrência do(s) sinistro(s)
- Valor total dos sinistros da apólice (em reais)
- Número de sinistros ocorridos no período

Esses dados são todos referentes a apólices sinistradas entre Jan/2010 e Maio/2013, pertencentes a quatro grandes seguradoras que foram misturados e não foram identificados a fim de preservar a confidencialidade das informações dessas companhias.

6.2. Mineração e tratamento dos dados

Antes de começar a utilizar os dados, é preciso sempre realizar uma análise e efetuar eventuais alterações cabíveis de modo a garantir que os valores estão coerentes com a finalidade para a qual serão aplicados.

Para tal, primeiro foram excluídos alguns poucos valores de sinistros que certamente foram incluídos incorretamente ou não se encaixavam no objetivo do trabalho, como por exemplo: Valores negativos, valores com mais de 2 casas decimais, entre outros.

Depois disso, foram excluídas também todas as apólices que constavam de mais de 1 sinistro, que não eram muitas mas poderiam viesar a distribuição do sinistro agregado.

Feito isso, ainda restaram 407.682 observações, o que é um tamanho de amostra considerável.

Por fim, como os valores de sinistro que temos são de um período de 4 anos, foi feita uma padronização dos valores, efetuando uma correção monetária de modo a colocá-los no mesmo instante de tempo. Todos os valores foram levados para a data de Julho de 2013, utilizando o Índice Nacional de Preços ao Consumidor Amplo – IPCA, que é utilizado pelo Banco Central como medidor oficial da inflação do país.

6.3. Estatísticas descritivas

Antes de entrar no estudo específico dos dados, é sempre bom observar as principais estatísticas descritivas referentes à base de dados, já que elas são úteis para se ter um melhor entendimento da informação proveniente desses valores.

Seguem abaixo as principais estatísticas descritivas e gráficos, obtidas com auxílio do software estatístico R.

- Tamanho da base: 407.682
- Média: 8.975,37
- Mediana: 3.232,54
- Máximo: 4.554.921
- Variância: 416.752.093
- Desvio padrão: 20.414,51

Nas figuras 7 e 8, temos respectivamente o histograma e boxplot associados a essa base de dados.

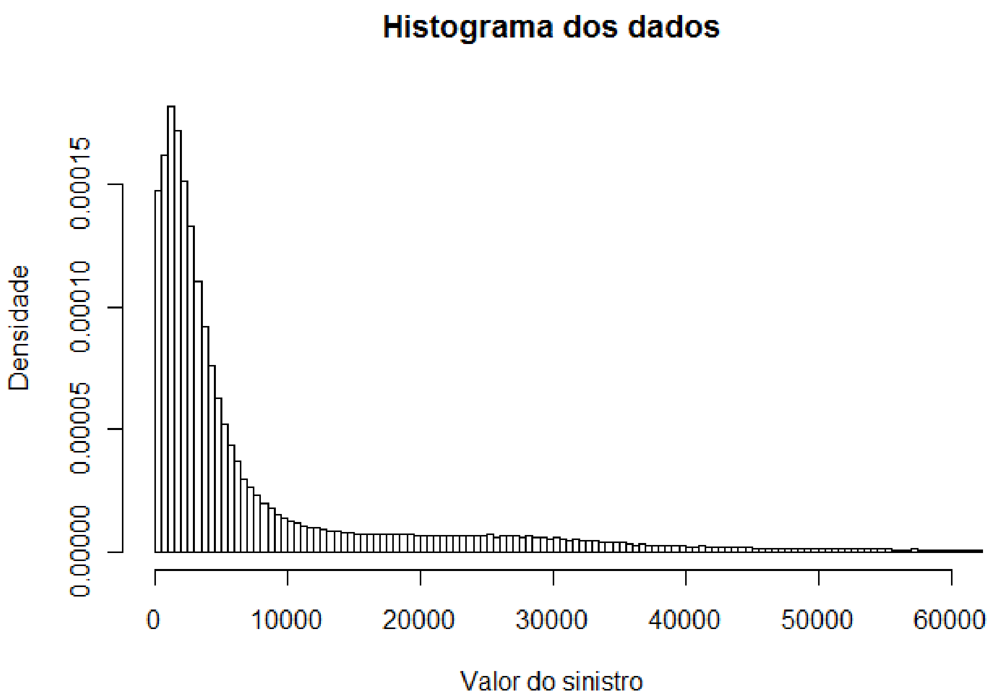

Figura 7 - Histograma dos dados

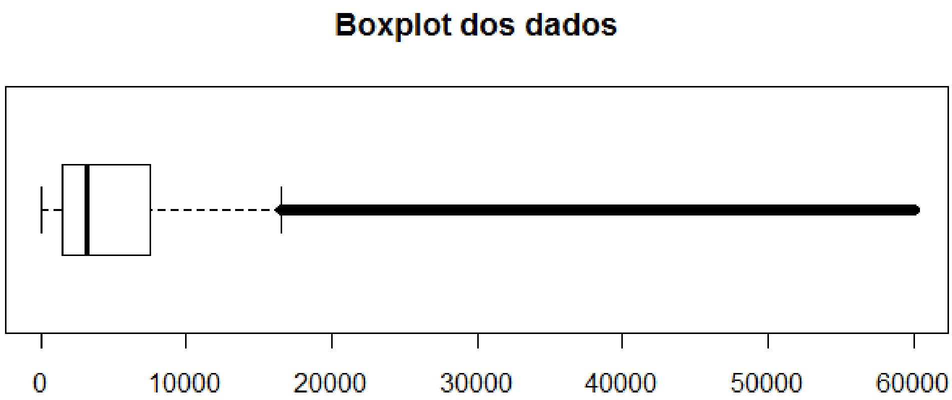

Figura 8 - Box-plot dos dados

Vale observar que tanto o histograma quanto o boxplot estão indo apenas até o valor 60.000, embora tenha alguns valores acima disso. Tal escolha foi feita porque são realmente poucos valores acima desse (menos de 500, em uma base de mais de 400.000 valores) e se construíssemos os gráficos até o valor máximo (4.554.921), eles deixariam de ser apresentáveis visualmente, por questões de escala.

Com relação à interpretação, a informação que mais chama a atenção com essas figuras é que os dados possuem uma assimetria à direita muito forte, ou seja, a maioria das observações em valores menores e algumas poucas observações de valor muito alto. Tal característica é de fato comum e esperada quando se analisa valor de sinistros em automóveis.

6.4. Metodologia

Conforme visto anteriormente, a distribuição do sinistro agregado depende de 2 variáveis aleatórias: N, número total de sinistros, e X, valor de 1 sinistro; e o método de Panjer só pode ser aplicado após se ter ajustado uma distribuição para cada uma dessas v.a's.

Para a distribuição de N, podemos ver que não há informação na base de dados sobre o número de apólices não sinistradas, tornando assim inviável a realização de um ajuste para a variável N. Para contornar isso foi feita arbitrariamente a suposição do número de sinistros ter distribuição Poisson(50), já que essa é uma das principais distribuições utilizadas para modelar o N e 50 é um valor de parâmetro alto, mas não tanto a ponto de tornar o método complicado de ser aplicado com programação computacional simples (quanto maior o parâmetro, maior tempo computacional é gasto na execução do método).

Para a distribuição do X, foram testadas 4 distribuições que usualmente são mais adequadas para descrever valor de sinistro. Foram elas: Exponencial, Gama, Weibull e Lognormal. Os parâmetros dessas distribuições foram estimados pelo método de momentos, que consiste em igualar os momentos amostrais com os teóricos. A única exceção foi a distribuição Weibull, pois não é possível estimar seus parâmetros por método dos momentos de forma analítica, então seus parâmetros foram estimados com auxílio de algumas ferramentas do software R.

Após os 4 ajustes, foi escolhido aquele que melhor se adequou a nossa base de dados. Como critério de escolha foi feita uma análise gráfica, sobrepondo a densidade dessas 4 distribuições sobre o histograma do valor dos sinistros. Tal critério foi suficiente pois foi possível detectar claramente a que ficou melhor e ela de fato se aproxima bastante do histograma.

Um critério formal muitas vezes utilizado para confirmar que uma determinada distribuição realmente se ajusta bem aos dados é a utilização de determinados testes de hipóteses, como bondade de ajustes ou Kolmogorov-Smirnov, que assumem em sua hipótese nula que os dados seguem determinada distribuição. Esses testes também foram feitos e todas as distribuições foram rejeitadas ao nível de significância de 5%. Porém, tal fato ocorreu porque o tamanho da amostra é muito grande (mais de 400 mil observações) e é sabido que uma grande falha dos testes de hipóteses é sempre rejeitar a hipótese nula quando o tamanho da amostra se aproxima do infinito. Então, por usarmos uma amostra muito grande, esses testes acabam não se aplicando.

Definidas as distribuições a serem utilizadas para N e X_i, resta ainda realizar a discretização da densidade de X_i, pois para a aplicação do método Panjer é exigido que a distribuição do valor dos sinistros também esteja apresentada na forma discreta.

Essa discretização foi feita do seguinte modo: Tomaram-se K intervalos de tamanho C partindo do 0 (pois o valor de sinistro só assume valores positivos), ou seja, ficamos com os segmentos [0,C), [C,2C), [2C,3C), ..., [(K-1)C,KC). Seja F a função de distribuição acumulada de X, cuja densidade queremos discretizar. A distribuição de X discreta será dada por:

$$\begin{cases} X = C/2, & \text{com prob. igual a } \mathrm{F}(\mathrm{C}) - F(0) \\ X = 3C/2, & \text{com prob. igual a } \mathrm{F}(2\mathrm{C}) - F(C) \\ \vdots & \\ X = (2j-1)C/2, & \text{com prob. igual a } \mathrm{F}(\mathrm{jC}) - F((j-1)C) \\ \vdots & \\ X = (2k-1)C/2, & \text{com prob. igual a } \mathrm{F}(\mathrm{kC}) - F((k-1)C) \end{cases}$$

Ou seja, a v.a. X discretizada assume valor igual ao ponto médio de um daqueles intervalos com probabilidade igual a da v.a. X contínua pertencer ao respectivo intervalo.

O valor C (amplitude dos intervalos) foi escolhido de modo a garantir que todos os intervalos tenham probabilidade pequena de ocorrer. O intervalo de maior probabilidade continha apenas cerca de 2% da massa total de probabilidade.

O valor K (número de intervalos) foi escolhido de modo a garantir que a soma das probabilidades de todos os intervalos contemplassem mais de 99,998% da massa de probabilidade total da distribuição de X contínua.

Por fim, a pequena massa de probabilidade acima do ponto KC foi alocada nos intervalos de forma ponderada pela probabilidade de cada intervalo.

7. Resultados

Todos os resultados e gráficos apresentados nessa seção foram obtidos com auxílio do software estatístico R.

7.1. Ajuste da distribuição para X

7.1.1. Ajuste exponencial

Usando a estimação pelo método dos momentos para $\exp(\lambda)$, temos que:

$$E\left[\exp(\lambda)\right] = \frac{1}{\lambda} \quad \Leftrightarrow \quad \hat{\lambda} = \frac{1}{\bar{x}} = 0.0001114$$

Então a distribuição exponencial ajustada aos dados pelo método dos momentos será a Exp(0,0001114). Na figura 9, podemos ver a sobreposição da densidade dessa exponencial com o histograma dos dados.

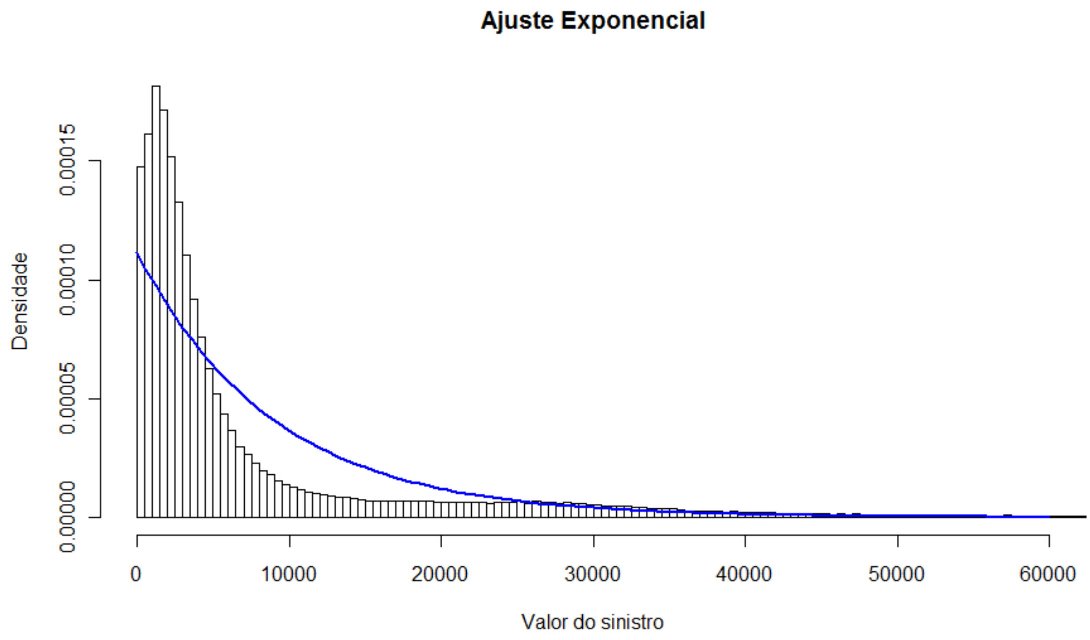

Figura 9 - Ajuste exponencial aos dados

7.1.2. Ajuste Gama

Usando a estimação pelo método dos momentos para $Gama(\alpha,\beta)$, temos que:

$$\begin{cases} E\left[Gama(\alpha,\beta)\right] = \dfrac{\alpha}{\beta} \\ Var\left[Gama(\alpha,\beta)\right] = \dfrac{\alpha}{\beta^2} \end{cases} \Leftrightarrow \begin{cases} \hat{\beta} = \dfrac{\overline{x}}{Var(X)} = 0.0000215 \\ \hat{\alpha} = \hat{\beta}\,\overline{x} = 0.193298 \end{cases}$$

Então a distribuição gama ajustada aos dados através do método dos momentos será a Gama(0,193298 ; 0,0000215). Na figura 10, podemos ver a sobreposição da densidade dessa Gama com o histograma dos dados.

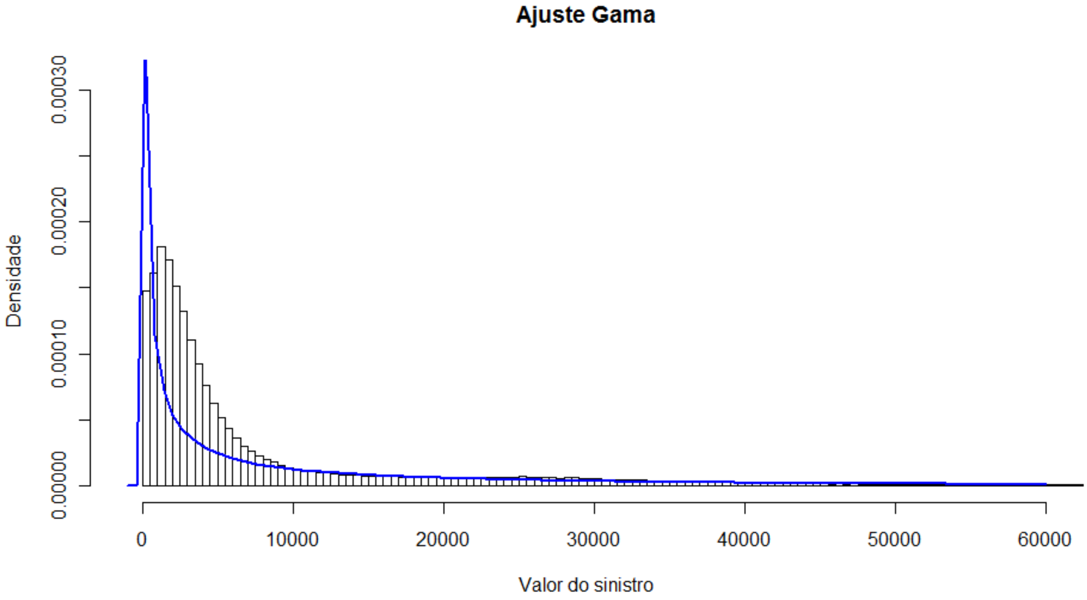

Figura 10 - Ajuste Gama aos dados

Vale observar que a distribuição $\exp(\lambda)$ equivale a $Gama(1,\lambda)$, então não era preciso testar a adequação de uma distribuição Exponencial, já que ela está contida na distribuição Gama. Tal ajuste foi feito mais para efeito didático, colocando ao menos uma distribuição uniparamétrica entre as tentativas.

7.1.3. Ajuste Weibull

A esperança e variância de uma distribuição $Weibull\,(\lambda,\delta)$ são dadas por:

$$\begin{cases} E\left[Weibull\,(\lambda,\delta)\right] = \lambda\,\Gamma\left(1+\dfrac{1}{\delta}\right) \\ Var\left[Weibull\,(\lambda,\delta)\right] = \lambda^2\left[\Gamma\left(1+\dfrac{2}{\delta}\right) - \left(\Gamma\left(1+\dfrac{1}{\delta}\right)\right)^2\right] \end{cases}$$

, onde Γ é a função gama

Podemos ver que não é possível estimar os parâmetros da Weibull por método de momentos de forma analítica. Então, com auxílio de ferramentas do software R, obtivemos os seguintes parâmetros estimados:

$$\hat{\lambda} = 7001.055 \qquad \hat{\delta} = 0.72658$$

Então a distribuição Weibull ajustada aos dados será a Weibull(7.001,055 ; 0,72658). Na figura 11, podemos ver a sobreposição da densidade dessa Weibull com o histograma dos dados.

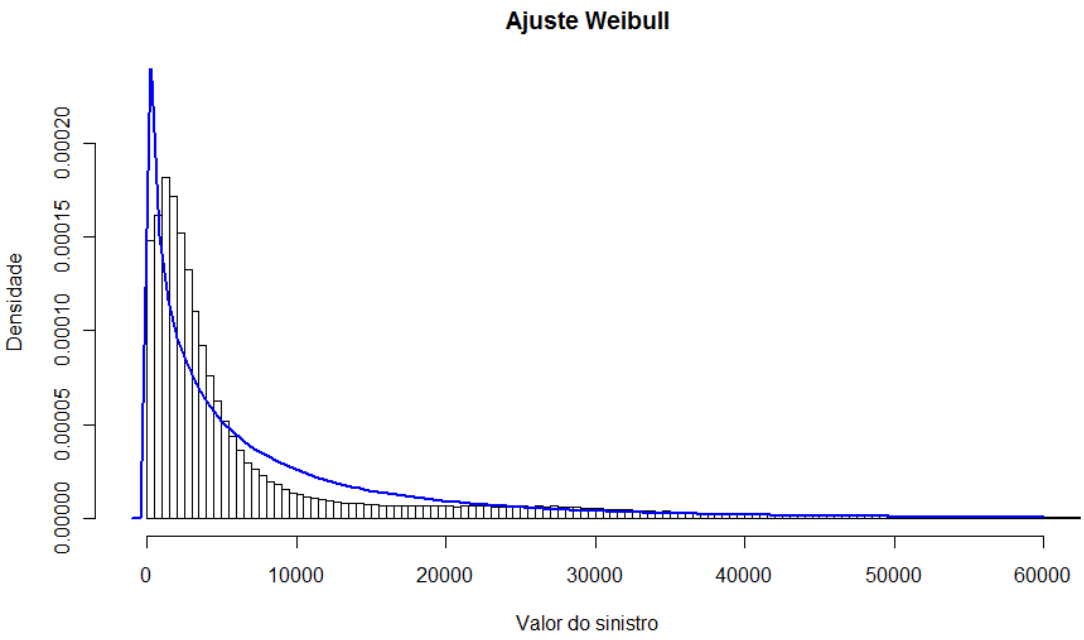

Figura 11 - Ajuste Weibull aos dados

7.1.4. Ajuste Lognormal

Usando a estimação pelo método dos momentos para a $LogNorm(\mu,\sigma)$, temos que:

$$\begin{cases} E\left[\,LogNorm(\mu,\sigma)\,\right] = \exp\left\{\mu + \dfrac{1}{2\sigma^2}\right\} \\ Var\left[\,LogNorm(\mu,\sigma)\,\right] = \left(\exp\left\{2\mu+\sigma^2\right\}\right)\left(\exp\left\{\sigma^2\right\}-1\right) \end{cases} \Leftrightarrow \begin{cases} \hat{\sigma}^2 = \ln\left(1+\dfrac{Var(X)}{\bar{x}^2}\right) = 1.8202 \\ \hat{\mu} = \ln(\bar{x}) - \dfrac{\hat{\sigma}^2}{2} = 8.19212 \end{cases}$$

$$\hat{\sigma}^2 = 1.8202 \quad \Rightarrow \quad \hat{\sigma} = 1.34916$$

Então a distribuição LogNormal ajustada aos dados através do método dos momentos será a LogNorm(8,19212 ; 1,34916). Na figura 12, podemos ver a sobreposição da densidade dessa LogNormal com o histograma dos dados.

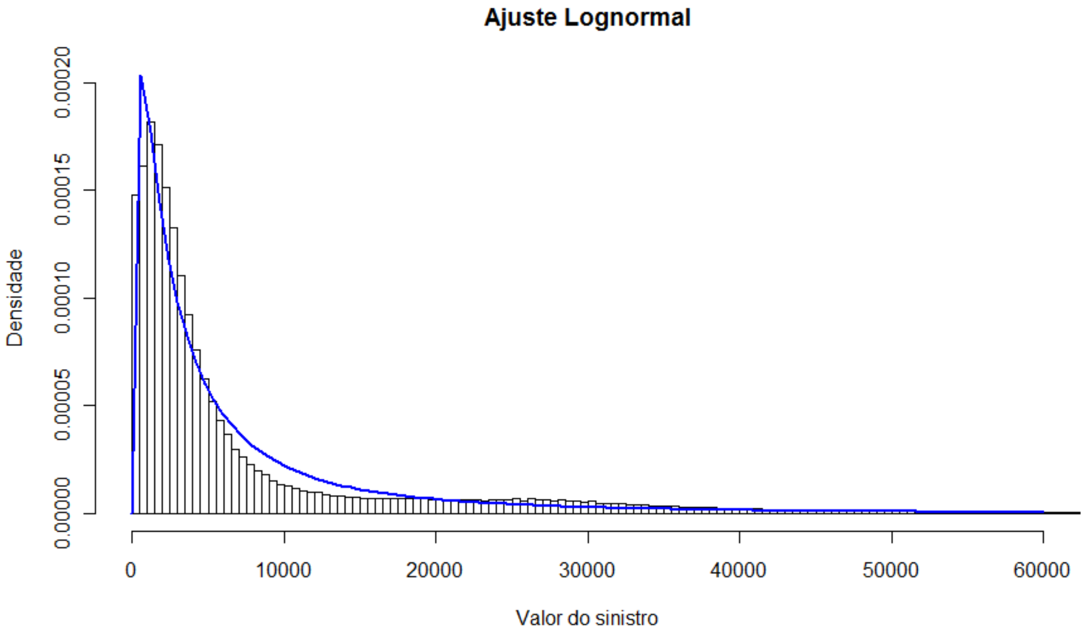

Figura 12 - Ajuste Log Normal aos dados

Por fim, comparando os 4 ajustes, através dos gráficos, fica evidente que o ajuste LogNormal (figura 12) foi o que mais se aproximou dos dados. Logo, a distribuição LogNorm(8,19212 , 1,34916) será a utilizada para aplicação do método Panjer.

7.2. Discretização da distribuição de X

Utilizando o método de discretização descrito na seção 1.4, o valor utilizado para C, amplitude dos intervalos, foi 100. Com isso o intervalo que conteve maior massa de probabilidade foi o [500,600), com 2.03%.

O valor utilizado para K, número de intervalos, foi 10000, pois com esse valor conseguimos que a soma dos intervalos contemple 99.99846% da massa de probabilidade da distribuição LogNormal ajustada ao X_i, que é um valor bastante próximo dos 100% e que ainda é viável computacionalmente.

Com isso, os intervalos são: [0,100), [100,200), [200,300), ... , [999.900,1.000.000).

A distribuição discretizada fica:

$$\begin{cases} X = 50, & \text{com prob. igual a } 0.00392 \\ X = 150, & \text{com prob. igual a } 0.01206 \\ X = 250, & \text{com prob. igual a } 0.01658 \\ \vdots & \end{cases}$$

7.3. Método Panjer e distribuição do sinistro agregado

Para distribuição de N, número de sinistros, foi utilizada a distribuição Poisson(50), definida arbitrariamente por não haver dados referentes a apólices sinistradas, não sendo possível portanto ajustar uma distribuição para N.

Na aplicação do método, é preciso obter os parâmetros a e b da fórmula recursiva de Panjer, que decorrem da distribuição discreta ajustada para N. Como no caso utilizamos a Poisson(50), pode ser visto na seção 5 que teremos $a = 0$ e $b = \lambda$.

Para a distribuição de X, o melhor ajuste obtido foi a Lognormal(8,19212 ; 1,34916), que para aplicação do método de Panjer foi preciso ser transformada em uma distribuição discreta. Tal discretização ocorreu com a utilização de 10.000 intervalos de tamanho 100.

Nas figuras 13 e 14, temos respectivamente os gráficos das funções de probabilidade do sinistro individual e do sinistro agregado, obtido através do método Panjer.

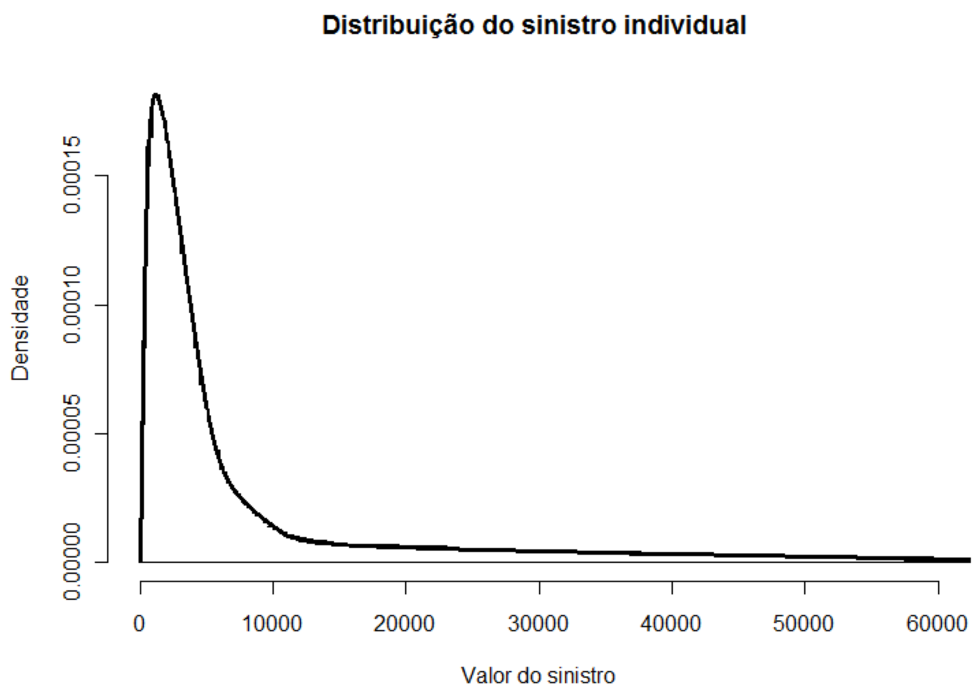

Figura 13 - Função de probabilidade do sinistro individual

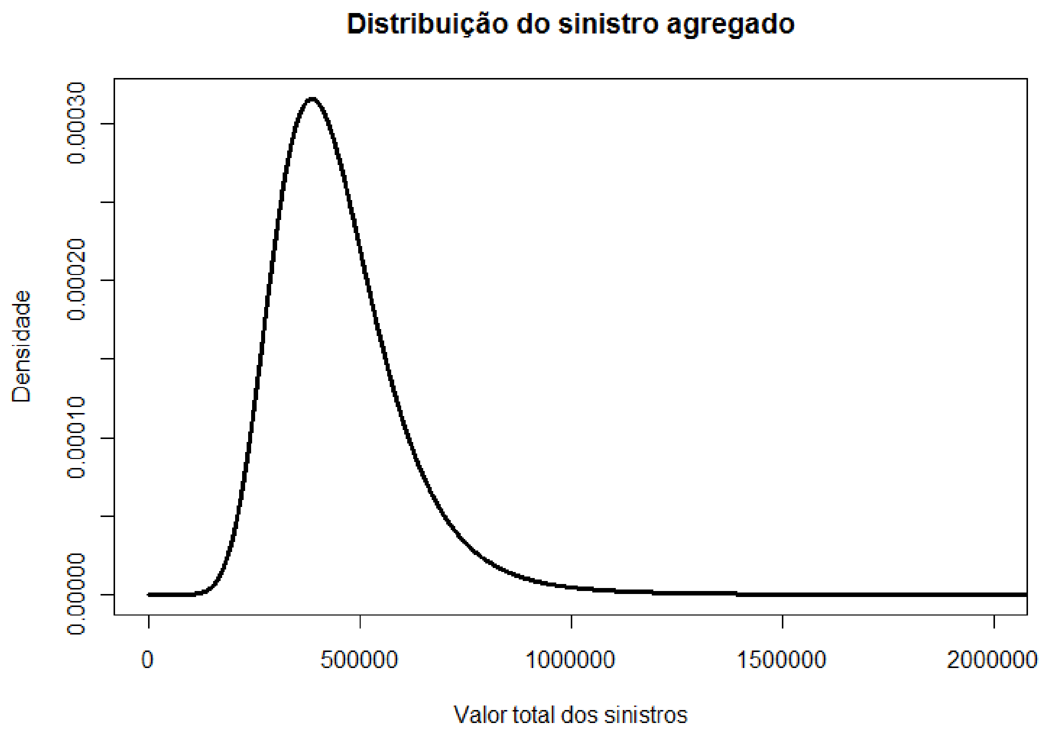

Figura 14 - Função de probabilidade do sinistro agregado

Na figura 15, temos a distribuição acumulada do sinistro agregado.

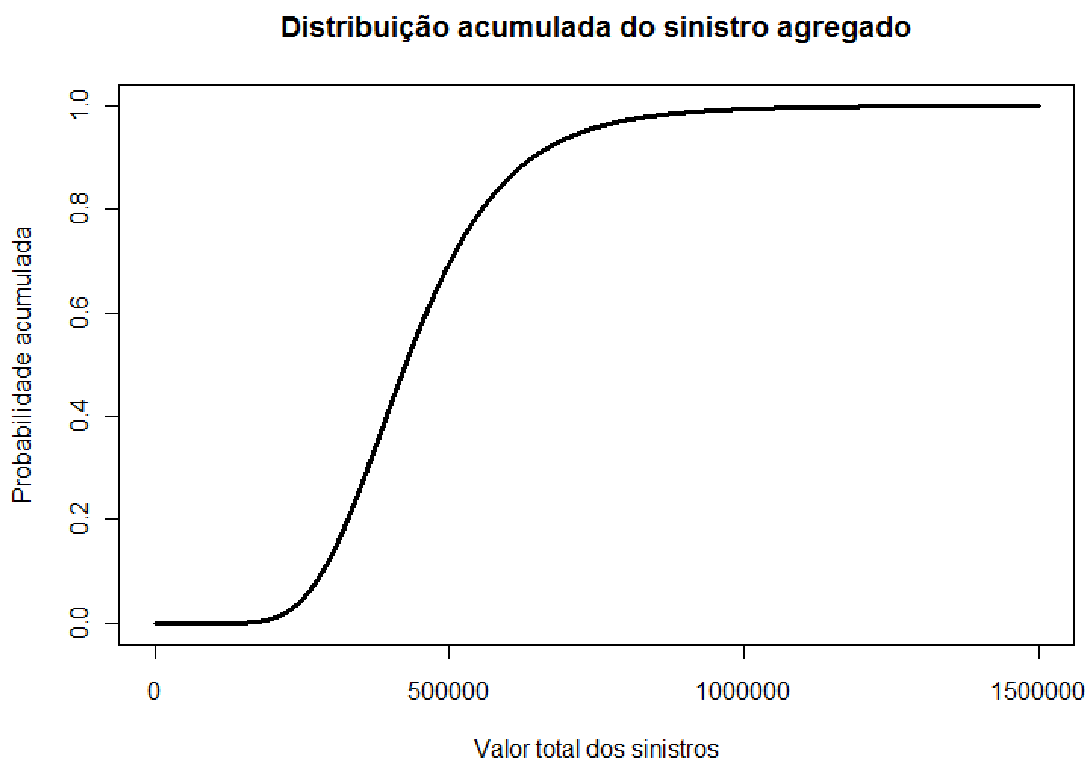

Figura 15 - Distribuição acumulada do sinistro agregado

Algumas informações relevantes da distribuição do sinistro agregado:

- Média: 449.477,9
- Mediana: 426.050
- Moda: 387.150
- Variância: 23.058.301.599
- Desvio padrão: 151.849,6

Pode-se perceber que a mediana é inferior à média, que decorre do fato da distribuição do sinistro agregado ser assimétrica à direita.

Além disso, temos que, de acordo com o item 4.1.3:

$$\begin{cases} E[N]E[X] = 448768.5 \\ Var[X]\mathrm{E}[N] + \mathrm{E}[X]^2 Var[N] = 24865468364 \end{cases}$$

Comparando esses valores com a média e variância da distribuição de S_{col} estimada pelo método Panjer, podemos ver que eles são bastante similares, indicando que a distribuição encontrada está coerente.

Com a distribuição acumulada é possível descobrir valores na cauda com certas probabilidades pequenas de serem ultrapassados, como por exemplo, a probabilidade de S^{col} ultrapassar 955.050 é de apenas 1%.

A realização dos cálculos recursivos para o método Panjer levou um tempo computacional de aproximadamente 3 horas e 40 minutos. Aumentar o número de intervalos utilizados na discretização ou o parâmetro da Poisson ajustada no N acarretaria em um aumento desse tempo computacional.

Por fim, temos na tabela 1 alguns valores da distribuição de S^{col} com suas respectivas probabilidades e probabilidades acumuladas. Não foram incluídos todos os valores porque a tabela ficaria muito extensa. Temos que x é o valor do sinistro agregado, $P(S=x)$ é a probabilidade desse sinistro assumir valor x e $F_S(x)$ é a probabilidade de $s \leq x$.

Tabela 1: Distribuição do sinistro agregado								
x	P(s=x)	$F_s(x)$	x	P(s=x)	$F_s(x)$	x	P(s=x)	$F_s(x)$
0	$1.9x10^{-22}$	$1.9x10^{-22}$	255050	$1.3x10^{-4}$	0.05182	515050	$2.0x10^{-4}$	0.72721
50	$3.8x10^{-23}$	$2.3x10^{-22}$	260050	$1.4x10^{-4}$	0.05850	520050	$2.0x10^{-4}$	0.73721
5050	$4.8x10^{-19}$	$4.2x10^{-18}$	265050	$1.5x10^{-4}$	0.06572	525050	$1.9x10^{-4}$	0.74690
10050	$4.0x10^{-17}$	$5.2x10^{-16}$	270050	$1.6x10^{-4}$	0.07348	530050	$1.9x10^{-4}$	0.75630
15050	$9.6x10^{-16}$	$1.6x10^{-14}$	275050	$1.7x10^{-4}$	0.08178	535050	$1.8x10^{-4}$	0.76541
20050	$1.2x10^{-14}$	$2.5x10^{-13}$	280050	$1.8x10^{-4}$	0.09063	540050	$1.7x10^{-4}$	0.77422
25050	$9.4x10^{-14}$	$2.3x10^{-12}$	285050	$1.9x10^{-4}$	0.10001	545050	$1.7x10^{-4}$	0.78275
30050	$5.5x10^{-13}$	$1.6x10^{-11}$	290050	$2.0x10^{-4}$	0.10994	550050	$1.6x10^{-4}$	0.79100
35050	$3.6x10^{-12}$	$8.4x10^{-11}$	295050	$2.1x10^{-4}$	0.12038	555050	$1.6x10^{-4}$	0.79897
40050	$1.0x10^{-11}$	$3.7x10^{-10}$	300050	$2.2x10^{-4}$	0.13133	560050	$1.5x10^{-4}$	0.80667
45050	$3.4x10^{-11}$	$1.4x10^{-9}$	305050	$2.3x10^{-4}$	0.14278	565050	$1.5x10^{-4}$	0.81410
50050	$1.0x10^{-10}$	$4.6x10^{-9}$	310050	$2.4x10^{-4}$	0.15471	570050	$1.4x10^{-4}$	0.82128
55050	$2.8x10^{-10}$	$1.4x10^{-8}$	315050	$2.5x10^{-4}$	0.16709	575050	$1.4x10^{-4}$	0.82820
60050	$6.8x10^{-10}$	$3.6x10^{-8}$	320050	$2.6x10^{-4}$	0.17990	580050	$1.3x10^{-4}$	0.83487
65050	$1.6x10^{-9}$	$9.1x10^{-8}$	325050	$2.7x10^{-4}$	0.19312	585050	$1.3x10^{-4}$	0.84131
70050	$3.4x10^{-9}$	$2.1x10^{-7}$	330050	$2.8x10^{-4}$	0.20673	590050	$1.2x10^{-4}$	0.84750
75050	$7.0x10^{-9}$	$4.7x10^{-7}$	335050	$2.8x10^{-4}$	0.22068	595050	$1.2x10^{-4}$	0.85347
80050	$1.4x10^{-8}$	$9.7x10^{-7}$	340050	$2.9x10^{-4}$	0.23496	600050	$1.1x10^{-4}$	0.85922
85050	$2.5x10^{-8}$	$1.9x10^{-6}$	345050	$2.9x10^{-4}$	0.24952	605050	$1.1x10^{-4}$	0.86475
90050	$4.4x10^{-8}$	$3.6x10^{-6}$	350050	$3.0x10^{-4}$	0.26436	610050	$1.0x10^{-4}$	0.87008
95050	$7.6x10^{-8}$	$6.6x10^{-6}$	355050	$3.0x10^{-4}$	0.27941	615050	$1.0x10^{-4}$	0.87520
100050	$1.2x10^{-7}$	$1.1x10^{-5}$	360050	$3.1x10^{-4}$	0.29467	620050	$9.7x10^{-5}$	0.88019
105050	$2.0x10^{-7}$	$1.9x10^{-5}$	365050	$3.1x10^{-4}$	0.31009	625050	$9.3x10^{-5}$	0.88485
110050	$3.1x10^{-7}$	$3.2x10^{-5}$	370070	$3.1x10^{-4}$	0.32565	630050	$8.9x10^{-5}$	0.88940
115050	$4.7x10^{-7}$	$5.1x10^{-5}$	375050	$3.1x10^{-4}$	0.34131	635050	$8.6x10^{-5}$	0.89377
120050	$6.9x10^{-7}$	$8.0x10^{-5}$	380050	$3.2x10^{-4}$	0.35704	640050	$8.2x10^{-5}$	0.89797
125050	$9.9x10^{-7}$	$1.2x10^{-4}$	385050	$3.2x10^{-4}$	0.37281	645050	$7.9x10^{-5}$	0.90201
130050	$1.4x10^{-6}$	$1.8x10^{-4}$	390050	$3.2x10^{-4}$	0.38859	650050	$7.6x10^{-5}$	0.90588
135050	$1.9x10^{-6}$	$2.6x10^{-4}$	395050	$3.1x10^{-4}$	0.40437	655050	$7.3x10^{-5}$	0.90960
140050	$2.6x10^{-6}$	$3.8x10^{-4}$	400050	$3.1x10^{-4}$	0.42008	660050	$7.0x10^{-5}$	0.91317
145050	$3.5x10^{-6}$	$5.3x10^{-4}$	405050	$3.1x10^{-4}$	0.43574	665050	$6.7x10^{-5}$	0.91659
150050	$4.6x10^{-6}$	$7.4x10^{-4}$	410050	$3.1x10^{-4}$	0.45130	670050	$6.4x10^{-5}$	0.91988
155050	$6.0x10^{-6}$	$1.0x10^{-3}$	415050	$3.1x10^{-4}$	0.46674	675050	$6.2x10^{-5}$	0.92304
160050	$7.7x10^{-6}$	$1.3x10^{-3}$	420050	$3.0x10^{-4}$	0.48204	680050	$5.9x10^{-5}$	0.92606
165050	$9.7x10^{-6}$	$1.8x10^{-3}$	425050	$3.0x10^{-4}$	0.49718	685050	$5.7x10^{-5}$	0.92897
170050	$1.2x10^{-5}$	$2.3x10^{-3}$	430050	$3.0x10^{-4}$	0.51214	690050	$5.5x10^{-5}$	0.93175
175050	$1.5x10^{-5}$	$3.0x10^{-3}$	435050	$2.9x10^{-4}$	0.52691	695050	$5.2x10^{-5}$	0.93442
180050	$1.8x10^{-5}$	$3.8x10^{-3}$	440050	$2.9x10^{-4}$	0.54146	700050	$5.0x10^{-5}$	0.93699
185050	$2.2x10^{-5}$	$4.8x10^{-3}$	445050	$2.8x10^{-4}$	0.55578	705050	$4.8x10^{-5}$	0.93944
190050	$2.6x10^{-5}$	$6.0x10^{-3}$	450050	$2.8x10^{-4}$	0.56986	710050	$4.6x10^{-5}$	0.94180
195050	$3.1x10^{-5}$	$7.4x10^{-3}$	455050	$2.7x10^{-4}$	0.58369	715050	$4.4x10^{-5}$	0.94406
200050	$3.6x10^{-5}$	$9.1x10^{-3}$	460050	$2.7x10^{-4}$	0.59725	720050	$4.2x10^{-5}$	0.94622
205050	$4.2x10^{-5}$	0.01108	465050	$2.6x10^{-4}$	0.61054	725050	$4.1x10^{-5}$	0.94830
210050	$4.9x10^{-5}$	0.01334	470050	$2.6x10^{-4}$	0.62355	730050	$3.9x10^{-5}$	0.95029
215050	$5.6x10^{-5}$	0.01594	475050	$2.5x10^{-4}$	0.63627	735050	$3.7x10^{-5}$	0.95220
220050	$6.3x10^{-5}$	0.01890	480050	$2.5x10^{-4}$	0.64869	740050	$3.6x10^{-5}$	0.95403
225050	$7.1x10^{-5}$	0.02224	485050	$2.4x10^{-4}$	0.66082	745050	$3.4x10^{-5}$	0.95579
230050	$7.9x10^{-5}$	0.02601	490050	$2.3x10^{-4}$	0.67265	750050	$3.3x10^{-5}$	0.95747
235050	$8.8x10^{-5}$	0.03021	495050	$2.3x10^{-4}$	0.68417	755050	$3.2x10^{-5}$	0.95908
240050	$9.8x10^{-5}$	0.03487	500050	$2.2x10^{-4}$	0.69539	760050	$3.0x10^{-5}$	0.96063
245050	$1.1x10^{-4}$	0.04002	505050	$2.2x10^{-4}$	0.70630	765050	$2.9x10^{-5}$	0.96211
250050	$1.2x10^{-4}$	0.04566	510050	$2.1x10^{-4}$	0.71691	770050	$2.8x10^{-5}$	0.96353

x	P(s=x)	$F_s(x)$	x	P(s=x)	$F_s(x)$	x	P(s=x)	$F_s(x)$
775050	2.7×10^{-5}	0.96489	1040050	3.4×10^{-6}	0.99400	1305050	6.6×10^{-7}	0.99833
780050	2.6×10^{-5}	0.96620	1045050	3.3×10^{-6}	0.99416	1310050	6.4×10^{-7}	0.99836
785050	2.5×10^{-5}	0.96746	1050050	3.2×10^{-6}	0.99432	1315050	6.2×10^{-7}	0.99839
790050	2.4×10^{-5}	0.96866	1055050	3.1×10^{-6}	0.99448	1320050	6.0×10^{-7}	0.99842
795050	2.3×10^{-5}	0.96981	1060050	3.0×10^{-6}	0.99463	1325050	5.8×10^{-7}	0.99845
800050	2.2×10^{-5}	0.97092	1065050	2.9×10^{-6}	0.99477	1330050	5.6×10^{-7}	0.99848
805050	2.1×10^{-5}	0.97198	1070050	2.8×10^{-6}	0.99491	1335050	5.5×10^{-7}	0.99851
810050	2.0×10^{-5}	0.97299	1075050	2.7×10^{-6}	0.99505	1340050	5.3×10^{-7}	0.99853
815050	1.9×10^{-5}	0.97397	1080050	2.6×10^{-6}	0.99518	1345050	5.1×10^{-7}	0.99856
820050	1.8×10^{-5}	0.97491	1085050	2.5×10^{-6}	0.99531	1350050	4.9×10^{-7}	0.99858
825050	1.8×10^{-5}	0.97581	1090050	2.4×10^{-6}	0.99544	1355050	4.8×10^{-7}	0.99861
830050	1.7×10^{-5}	0.97667	1095050	2.4×10^{-6}	0.99556	1360050	4.6×10^{-7}	0.99863
835050	1.6×10^{-5}	0.97750	1100050	2.3×10^{-6}	0.99567	1365050	4.5×10^{-7}	0.99865
840050	1.6×10^{-5}	0.97829	1105050	2.2×10^{-6}	0.99578	1370050	4.3×10^{-7}	0.99868
845050	1.5×10^{-5}	0.97905	1110050	2.1×10^{-6}	0.99589	1375050	4.2×10^{-7}	0.99870
850050	1.4×10^{-5}	0.97979	1115050	2.1×10^{-6}	0.99600	1380050	4.0×10^{-7}	0.99872
855050	1.4×10^{-5}	0.98049	1120050	2.0×10^{-6}	0.99610	1385050	3.9×10^{-7}	0.99874
860050	1.3×10^{-5}	0.98116	1125050	1.9×10^{-6}	0.99620	1390050	3.8×10^{-7}	0.99876
865050	1.3×10^{-5}	0.98181	1130050	1.9×10^{-6}	0.99630	1395050	3.6×10^{-7}	0.99878
870050	1.2×10^{-5}	0.98244	1135050	1.8×10^{-6}	0.99639	1400050	3.5×10^{-7}	0.99879
875050	1.2×10^{-5}	0.98303	1140050	1.8×10^{-6}	0.99648	1405050	3.4×10^{-7}	0.99881
880050	1.1×10^{-5}	0.98361	1145050	1.7×10^{-6}	0.99656	1410050	3.3×10^{-7}	0.99883
885050	1.1×10^{-5}	0.98416	1150050	1.7×10^{-6}	0.99665	1415050	3.1×10^{-7}	0.99884
890050	1.0×10^{-5}	0.98469	1155050	1.6×10^{-6}	0.99673	1420050	3.0×10^{-7}	0.99886
895050	1.0×10^{-5}	0.98520	1160050	1.6×10^{-6}	0.99681	1425050	2.9×10^{-7}	0.99887
900050	9.6×10^{-6}	0.98569	1165050	1.5×10^{-6}	0.99689	1430050	2.8×10^{-7}	0.99889
905050	9.3×10^{-6}	0.98617	1170050	1.5×10^{-6}	0.99696	1435050	2.7×10^{-7}	0.99890
910050	8.9×10^{-6}	0.98662	1175050	1.4×10^{-6}	0.99703	1440050	2.6×10^{-7}	0.99891
915050	8.6×10^{-6}	0.98705	1180050	1.4×10^{-6}	0.99710	1445050	2.5×10^{-7}	0.99893
920050	8.2×10^{-6}	0.98747	1185050	1.3×10^{-6}	0.99717	1450050	2.4×10^{-7}	0.99894
925050	7.9×10^{-6}	0.98788	1190050	1.3×10^{-6}	0.99724	1455050	2.3×10^{-7}	0.99895
930050	7.6×10^{-6}	0.98827	1195050	1.3×10^{-6}	0.99730	1460050	2.2×10^{-7}	0.99896
935050	7.3×10^{-6}	0.98864	1200050	1.2×10^{-6}	0.99736	1465050	2.1×10^{-7}	0.99897
940050	7.1×10^{-6}	0.98900	1205050	1.2×10^{-6}	0.99742	1470050	2.1×10^{-7}	0.99898
945050	6.8×10^{-6}	0.98935	1210050	1.2×10^{-6}	0.99748	1475050	2.0×10^{-7}	0.99899
950050	6.5×10^{-6}	0.98968	1215050	1.1×10^{-6}	0.99754	1480050	1.9×10^{-7}	0.99900
955050	6.3×10^{-6}	0.99000	1220050	1.1×10^{-6}	0.99760	1485050	1.8×10^{-7}	0.99901
960050	6.1×10^{-6}	0.99031	1225050	1.1×10^{-6}	0.99765	1490050	1.8×10^{-7}	0.99902
965050	5.9×10^{-6}	0.99061	1230050	1.0×10^{-6}	0.99770	1495050	1.7×10^{-7}	0.99903
970050	5.6×10^{-6}	0.99090	1235050	1.0×10^{-6}	0.99775	1500050	1.6×10^{-7}	0.99904
975050	5.4×10^{-6}	0.99117	1240050	9.7×10^{-7}	0.99780	1550050	1.1×10^{-7}	0.99911
980050	5.2×10^{-6}	0.99144	1245050	9.4×10^{-7}	0.99785	1600050	7.0×10^{-8}	0.99915
985050	5.0×10^{-6}	0.99169	1250050	9.2×10^{-7}	0.99790	1650050	4.6×10^{-8}	0.99918
990050	4.9×10^{-6}	0.99194	1255050	8.9×10^{-7}	0.99794	1700050	3.0×10^{-8}	0.99920
995050	4.7×10^{-6}	0.99218	1260050	8.6×10^{-7}	0.99799	1750050	2.0×10^{-8}	0.99921
1000050	4.5×10^{-6}	0.99241	1265050	8.4×10^{-7}	0.99803	1800050	1.3×10^{-8}	0.99922
1005050	4.4×10^{-6}	0.99264	1270050	8.1×10^{-7}	0.99807	1850050	8.5×10^{-9}	0.99922
1010050	4.2×10^{-6}	0.99285	1275050	7.9×10^{-7}	0.99811	1900050	5.7×10^{-9}	0.99923
1015050	4.1×10^{-6}	0.99306	1280050	7.7×10^{-7}	0.99815	1950050	3.9×10^{-9}	0.99923
1020050	3.9×10^{-6}	0.99326	1285050	7.4×10^{-7}	0.99819	2000050	2.6×10^{-9}	0.99923
1025050	3.8×10^{-6}	0.99345	1290050	7.2×10^{-7}	0.99822			
1030050	3.7×10^{-6}	0.99364	1295050	7.0×10^{-7}	0.99826			
1035050	3.5×10^{-6}	0.99382	1300050	6.8×10^{-7}	0.99829			

CONCLUSÃO

Este trabalho constituiu em uma aplicação do método Panjer para obtenção da distribuição do sinistro agregado a uma base de dados real de seguros de automóveis (Casco) constituída de mais de 400.000 valores de sinistros referentes aos anos de 2010 a 2013.

Foram apresentadas diversas ferramentas para a obtenção das distribuições do valor de um sinistro, do número de sinistros e do sinistro agregado, bem como propriedades relevantes delas, seus prós e contras. Destacou-se o método Panjer, que consegue obter a distribuição do sinistro agregado apenas utilizando uma fórmula recursiva, tornando desnecessário o cálculo de integrais complexas ou o uso de programação computacional pesada. A obtenção dessas distribuições é essencial para que as seguradoras possam fazer um gerenciamento adequado do risco e garantir as coberturas que oferecem aos seus segurados.

Para os dados utilizados, a distribuição que melhor se ajustou à variável aleatória “valor de um sinistro” foi a Lognormal. Para a variável aleatória “número de sinistros”, infelizmente a base de dados obtida não permitia ajustar uma distribuição (não havia informação sobre o número de apólices não sinistradas), então a alternativa foi a utilização da distribuição Poisson(50), uma vez que a Poisson usualmente se ajusta bem a essa variável. Conjugando essas duas distribuições, foi possível aplicar o método Panjer e obter uma distribuição final para o sinistro agregado.

Os resultados obtidos permitem ratificar que o método Panjer é de fato aplicável na prática, mesmo com um grande número de iterações e a eficácia de seus resultados, haja visto que a distribuição encontrada teve média e variância consistentes com as calculadas utilizando outro método.

Por fim, vale mencionar que uma possível melhoria num trabalho sobre esse tema seria a utilização de dados que permitam ajustar também uma distribuição para o número de sinistros e aplicar o método Panjer utilizando essa distribuição. Porém, as empresas são muito reservadas com relação à confidencialidade de suas informações, o que dificulta a obtenção de tal base de dados.

REFERÊNCIAS

BOWERS, N. et al; GERBER, H.U; HICKMAN, J.C; JONES, D.A; NESBITT, C.J; **Actuarial Mathematics**. 2nd. ed. The Society of Actuaries, 1997.

FERREIRA, Paulo Pereira; **Modelos de Precificação e Ruína para Seguros de Curto Prazo**. Funenseg, 2010.

FERREIRA, Paulo Pereira; **Uma Aplicação do Método de Panjer à Experiência Brasileira de Sinistros do Ramo de Seguros de Automóveis**. Funenseg, 1998.

JAMES, Barry R.; **Probabilidade: Um Curso em Nível Intermediário**. 3rd. ed. IMPA, 2011.

MAGALHÃES, Marcos Nascimento; **Probabilidade e Variáveis Aleatórias**. 2nd ed. Edusp, 2006.

PANJER, H.H., **Recursive Evaluation of a Family of Compound Distributions**. Astin Bulletin, 1981

Site da Susep - Superintendência de Seguros Privados, pode ser acessado em: http://www.susep.gov.br

Site do Denatran - Departamento Nacional de Trânsito, pode ser acessado em: http://www.denatran.gov.br

APÊNDICE

Comandos utilizados no software R para obtenção dos resultados apresentados

```
#Carregando a base de dados no software:
dados=read.table("E:\\UFRJ\\Projeto Final\\PANJER\\dados projeto v3 ipca.txt")
valor.sinistro=as.numeric(as.character(dados[2:length(dados[,5]),5]))
y=valor.sinistro

#Análise descritiva:
length(y)
mean(y)
median(y)
var(y)
sd(y)

#Gráficos iniciais:
hist(y, nclass=10000, xlim=c(0,60000), prob=T, ylab="Densidade", xlab="Valor do sinistro", main="Histograma dos dados")
boxplot(y[y<60000], main="Boxplot dos dados")

#Ajuste exponencial:
lambda=1/mean(y)
hist(y, nclass=10000, xlim=c(0,60000), prob=T, ylab="Densidade", xlab="Valor do sinistro", main="Ajuste Exponencial")
curve(dexp(x,lambda), add=T,col=4, lwd=2)

#Ajuste Gama:
beta=mean(y)/var(y)
alfa=beta*mean(y)
hist(y, nclass=10000, xlim=c(-1000,60000), ylim=c(0,0.00031), ylab="Densidade", xlab="Valor do sinistro", main="Ajuste Gama")
curve(dgamma(x,alfa,beta),add=T,col=4, lwd=2)

#Ajuste Weibull:
library(survival)
weib <- survreg(Surv(y)~1, dist='weibull')
alpha <- exp(weib$coefficients[1])
gama <-1/weib$scale
hist(y, nclass=10000, xlim=c(-1000,60000), ylim=c(0,0.00023), prob=T, ylab="Densidade", xlab="Valor do sinistro", main="Ajuste Weibull")
curve(dweibull(x,gama,alpha), add=T,col=4, lwd=2)
```

```
#Ajuste Lognormal:
sigma2=log(1+(var(y)/(mean(y)^2)))
sigma=sqrt(sigma2)
mi=log(mean(y))-sigma2/2
hist(y, nclass=10000, xlim=c(-10,60000), ylim=c(0,0.0002), prob=T,
ylab="Densidade", xlab="Valor do sinistro", main="Ajuste Lognormal")
curve(dlnorm(x,mi,sigma),add=T,col=4, lwd=2)

#Testes de aderência: (não conclusivos)
ks.test(y,plnorm)
chisq.test(y,y=NULL,plnorm)

#Discretização da distribuição do sinistro individual:
x=0.01
M=0.01
n=10000
c=100
for (i in 1:n) {
  x[i] = plnorm(c*(i),mi,sigma)-plnorm(c*(i-1),mi,sigma)
  z=max(x)
  M[i]=i*c-c/2 }
a=0
b=50

#Programação da formula iterativa de Panjer:
t1=proc.time()
p=NA
acum=NA
it=100000 #15000
p=c(rep(0,it))
p[1]=exp(-L)
acum[1]=p[1]
for (i in 2:it) {
  k=0
  for (j in max(1,i-n):(i-1)) {
    k=k+(a+(b*(abs(j-i))/(i-1)))*p[j]*x[i-j] }
  p[i]=k
  acum[i]=sum(p) }
t2=proc.time()
t2-t1

#Gráficos referentes a distribuição do sinistro agregado:
b=seq(50,it*c-(c+50),by=c)
```

```
b=c(0,b)
plot(b,p, cex=0.4, ylab="Densidade", xlab="Valor total dos sinistros", main="Distribuição do sinistro agregado")
plot(b,acum, cex=0.4, ylab="Probabilidade acumulada", xlab="Valor total dos sinistros", main="Distribuição acumulada do sinistro agregado")

#Cálculo da média, mediana e variância de Scol:
m=0
for (i in 1:15000) {
  m[i]=b[i]*p[i] }
sum(m)
mean(y)*L
acum[4260]
acum[4261]    #Mediana=426050
m2=0
for (i in 1:15000) {
  m2[i]=(b[i]^2)*p[i] }
Var=sum(m2)-(sum(m)^2)
dp=sqrt(Var)
var(y)*50+(mean(y)^2)*50
acum[9551]    #Valor de x tal que P(Scol > x) = 0.01
```