



Universidade de Brasília
Departamento de Estatística

RÔMULO COUTINHO ARAÚJO

Modelos Estocásticos em Acidentes de Trânsito

Brasília 2018

RÔMULO COUTINHO ARAÚJO

Modelos Estocásticos em Acidentes de Trânsito

Trabalho de Conclusão de Curso – TCC –
apresentado para obtenção do título de Ba-
charel em Estatística

Departamento de Estatística - Universidade de Brasília

Orientador: CIRA ETHEOWALDA GUEVARA

Brasília

2018

Resumo

Neste trabalho, foram utilizados processos de Poisson e de renovação para modelar dados de acidentes graves de trânsito de algumas rodovias federais do Brasil, ocorridos no período 2016 a 2017. A distribuição dos tempos entre ocorrências dos acidentes foi modelada por uma distribuição Weibull e por uma mistura de distribuições Weibull. Os parâmetros dessas distribuições foram estimados por máxima verossimilhança, com o suporte do algoritmo EM. Quando a distribuição dos dados dos tempos entre ocorrências dos acidentes apresentou uma distribuição Weibull com parâmetro de forma um, o processo de contagem utilizado para analisar o número de acidentes foi de Poisson, porém foi utilizado o processo de renovação quando a distribuição dos tempos entre ocorrências dos acidentes se ajustou melhor por uma mistura de distribuições Weibull.

Palavras-chave: Distribuição Weibull. Processo de Poisson. Processo de Renovação. Mistura de distribuições. Acidentes de trânsito.

Lista de ilustrações

Figura 1 – Curvas da função de densidade de uma variável Weibull com $\alpha = 1.5$ e β variando.	14
Figura 2 – Curvas da função de densidade de uma variável Weibull com $\beta = 1$ e α variando.	14
Figura 3 – Misturas de distribuições <i>Weibull</i> para parâmetros da Tabela 1.	17
Figura 4 – Processo de Contagem Poisson ($\lambda = 2$).	19
Figura 5 – Processo de Renovação.	21
Figura 6 – Gráfico dos tempos entre acidentes (teórico e empírico).	30
Figura 7 – Distribuição acumulada e <i>qqplot</i> dos tempos.	31
Figura 8 – Ajuste tempos entre acidentes com motos à distribuição exponencial.	32
Figura 9 – Distribuição acumulada e <i>qqplot</i> dos tempos de acidentes com motos.	32
Figura 10 – Distribuição acumulada e <i>qqplot</i> dos tempos de acidentes sem motos.	32
Figura 11 – Processo de Contagem Poisson com cada taxa da tabela 5.	34
Figura 12 – Probabilidade dos eventos de acidentes por grupo ocorrerem em t dias.	34
Figura 13 – Modelos para os tempos entre acidentes BR 110/BA.	35
Figura 14 – Distribuições acumuladas para comparativas.	36
Figura 15 – Gráficos <i>qqplot</i> : Weibull e MMW.	37
Figura 16 – Contagem do Processo de Renovação com Mistura.	38

Lista de tabelas

Tabela 1 – Vetor de parâmetros do MMW (2.11).	17
Tabela 2 – Resumo da base de dados de acidentes da PRF.	29
Tabela 3 – Estatísticas dos tempos (em dias) entre acidentes – BR 101/RJ.	30
Tabela 4 – Estatísticas dos tempos (em dias) com e sem motos – BR 101/RJ.	31
Tabela 5 – Estimativas e valor de teste de KS para os tempos de acidentes.	33
Tabela 6 – Probabilidades para $N(t)$ conforme t (dias) e n (acidentes).	33
Tabela 7 – Estatísticas dos Tempos entre Acidentes – BR 110/BA.	35
Tabela 8 – Estimativas dos Parâmetros – BR 110/BA.	36
Tabela 9 – Teste de Kolmogorv-Smirnov para os Dois Casos.	37
Tabela 10 – Probabilidades para $N(t)$ (mistura) conforme t (dias) e n (acidentes).	38

Sumário

1	INTRODUÇÃO	11
2	CONCEITOS PRELIMINARES	13
2.1	Distribuição de Probabilidade Weibull	13
2.1.1	Definição	13
2.2	Mistura de Distribuições	15
2.2.1	Mistura de Distribuições Weibull	16
2.3	Processos de Poisson e Renovação	17
2.3.1	Processos de Poisson	18
2.3.2	Processos de Renovação	21
2.3.3	Distribuição Assintótica de um Processo de Renovação.	22
3	ESTIMAÇÃO	25
3.1	Estimação por Máxima Verossimilhança dos Tempos entre Acidentes	25
3.1.1	O Algoritmo EM	25
3.1.2	Algoritmo EM para Mistura de Distribuições Weibull	26
3.1.3	Procedimentos para Modelagem dos Dados	28
4	APLICAÇÃO	29
4.1	Dados de Acidentes Graves na BR 101/RJ: Modelo Poisson	30
4.2	Dados de Acidentes Graves: Processo de Renovação	35
4.2.1	BR 110/BA	35
5	CONCLUSÃO	41
	REFERÊNCIAS	43
	APÊNDICE A – EQUAÇÃO DE RENOVAÇÃO PARA MISTURA DE WEIBULL	45
	ANEXO A – CÓDIGOS USADOS NO R	47

1 Introdução

Processos estocásticos de contagem são modelos probabilísticos naturais para analisar o número de acidentes de trânsito que ocorrem em uma determinada região para um determinado intervalo de tempo.

O Brasil é um país que escolheu o transporte rodoviário como principal meio de circulação de seus bens e mercadorias, ou seja, a maior parcela de sua riqueza. Mas, junto aos veículos destinados a esse tipo de transporte, também circulam pessoas com interesses diversos, seja para simples deslocamento, viagens de férias, trabalho, etc., e que, aliado às condições desfavoráveis de infraestrutura, ao comportamento inadequado dos condutores e pedestres, um fluxo tão heterogêneo e envolvendo veículos de dimensões e peso muito desiguais, favoreceu o grande número de acidentes graves em todo o país.

A Polícia Rodoviária Federal – PRF – atende cerca de setenta mil quilômetros de rodovias por todo o Brasil e um dos seus principais serviços é o de atendimento aos acidentes de trânsito ocorridos na malha das rodovias federais. Além de realizar o atendimento, a PRF é responsável por fazer estudos e aplicar procedimentos para reduzir a quantidade desses acidentes, principalmente aqueles que deixam vítimas graves ou fatais.

As informações de cada acidente de trânsito atendido pela PRF são coletadas em um sistema próprio chamado BAT (Boletim de Acidente de Trânsito), em um modelo único para todo o Brasil, e então seus dados são disponibilizados para extração ao público em geral.

A justificativa para o uso dos dados da PRF dá-se pela facilidade em ter acesso a essa informação e pela qualidade das informações registradas em seu boletim de acidente; sendo que tal estudo com abrangência nacional seria muito difícil através de dados de outros órgãos de trânsito de nível estadual, tendo em vista que nem todos divulgam com qualidade as informações de acidentes registradas, além da falta de uniformidade dos modelos de boletins de acidentes confeccionados por esses órgãos.

O estudo também se justifica pela necessidade de entender melhor as características dos acidentes graves em rodovias federais, e com isso auxiliar na elaboração de políticas de enfrentamento às ocorrências dessa natureza.

O objetivo principal deste trabalho é aplicar a teoria desses processos de contagem para analisar dados de acidentes de trânsito graves ocorridos na malha rodoviária federal do Brasil.

Um processo de contagem com incrementos independentes e estacionários geral é um "processo de renovação", que por sua vez é caracterizado pela "distribuição entre a

ocorrência dos eventos".

Uma análise exploratória dos dados de acidentes graves em rodovias Brasileiras indicaram que a distribuição entre a ocorrência dos acidentes segue em alguns casos distribuição exponencial ou Weibull e em outros casos, aparentemente, uma mistura de distribuições Weibull. O processo de contagem correspondente à distribuição Exponencial é um processo de Poisson e nos outros casos o processo de contagem é de renovação mais amplo.

O trabalho está organizado de forma que no capítulo 2 é apresentado uma revisão da distribuição Weibull, assim como da mistura de duas distribuições, além da teoria de processos de contagem, processo de Poisson e renovação. No capítulo 3 uma revisão sobre estimação por máxima verossimilhança e algoritmo EM é apresentada. No capítulo 4, o ferramental teórico é utilizado na aplicação de dados de acidentes em algumas rodovias federais.

2 Conceitos Preliminares

2.1 Distribuição de Probabilidade Weibull

Neste capítulo descrevem-se os principais conceitos básicos e necessários para desenvolver o trabalho.

2.1.1 Definição

A distribuição Weibull é uma distribuição de probabilidade contínua com diversas aplicações práticas da distribuição (Weibull, 1951).

Por definição, uma variável aleatória X tem distribuição Weibull se a função de densidade de probabilidade f é dada por:

$$f(x; \alpha; \beta) = \begin{cases} \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^\alpha}, & x \geq 0 \\ 0, & x < 0, \end{cases} \quad (2.1)$$

sendo α o parâmetro de forma, β parâmetro de escala. Quando $\alpha = 1$ o modelo X se reduz a uma variável exponencial.

Abaixo seguem gráficos em que um dos parâmetros da distribuição varia enquanto o outro permanece constante:

Na figura 1, em que se manteve o valor constante para o parâmetro α (forma), variando o parâmetro β , observa-se uma mudança drástica na escala, com a expansão e compressão do gráfico.

Na figura 2, com o valor do parâmetro β constante, a mudança no parâmetro α ocasiona variação acentuada na forma do gráfico. E quando temos um valor para $\alpha < 1$, há a predominância do aspecto exponencial do gráfico.

A função de distribuição acumulada da distribuição Weibull, com parâmetros α e β , denotada por $Weibull(\alpha, \beta)$, é expressa por:

$$F(x; \alpha; \beta) = \begin{cases} 1 - e^{-\left(\frac{x}{\beta}\right)^\alpha} & x \geq 0, \\ 0 & x < 0. \end{cases} \quad (2.2)$$

Os momentos de $X \sim Weibull(\alpha, \beta)$ são obtidos através de:

$$E[e^{t \cdot \log(X)}] = \beta^t \Gamma\left(\frac{t}{\alpha} + 1\right), \quad (2.3)$$

em que Γ representa a função gama.

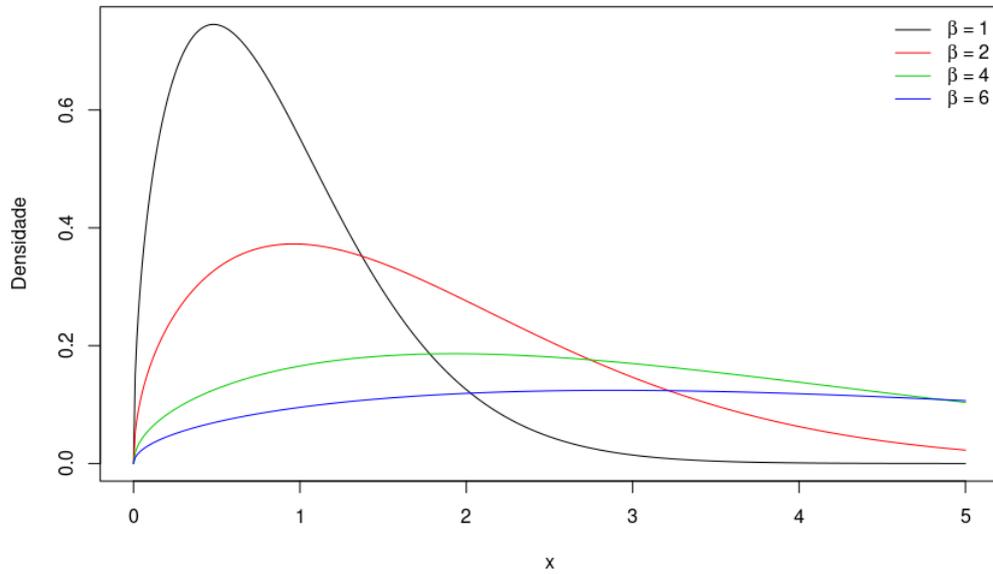


Figura 1 – Curvas da função de densidade de uma variável Weibull com $\alpha = 1.5$ e β variando.

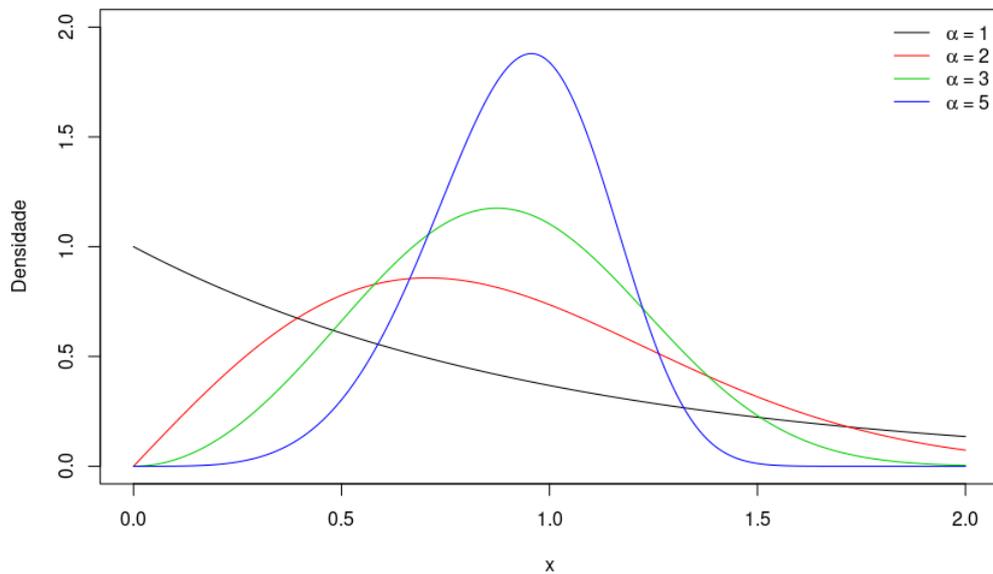


Figura 2 – Curvas da função de densidade de uma variável Weibull com $\beta = 1$ e α variando.

Assim:

$$M_n = \beta^n \Gamma\left(\frac{n}{\alpha} + 1\right) \quad (2.4)$$

Dessa forma, de (2.4), temos, respectivamente, a esperança e variância de X

$\sim Weibull(\alpha, \beta)$:

$$E[X] = \beta \Gamma\left(\frac{1}{\alpha} + 1\right); \quad (2.5)$$

$$Var[X] = \beta^2 \left[\Gamma\left(\frac{2}{\alpha} + 1\right) - \left(\Gamma\left(\frac{1}{\alpha} + 1\right) \right)^2 \right]. \quad (2.6)$$

A distribuição Weibull tem sido utilizada em diversas áreas para modelar tempo de vida, força ou tamanho, entre outras variáveis. Por exemplo, Weibull (1951) a utilizou para modelar forças de tensão necessárias para torcer uma barra de aço ou para rompimento de fibras de algodão indiano. Na última década, destacam-se as aplicações da distribuição Weibull na área estatística de "Análise de Sobrevivência", como pode ser visto em Colosimo e Giolo (2006).

Neste trabalho, utilizamos a distribuição Weibull para modelar o tempo entre ocorrências de acidentes na malha rodoviária Brasileira.

A distribuição Weibull pode ainda apresentar um terceiro parâmetro de *locação*, μ , que é responsável por transladar a curva ao longo do eixo x , e a função de distribuição tem a forma:

$$f(x; \mu; \alpha; \beta) = \frac{\alpha}{\beta} \left(\frac{x - \mu}{\beta} \right)^{\alpha-1} e^{-\left(\frac{x - \mu}{\beta} \right)^\alpha}, \quad (2.7)$$

em que $x \geq \mu$ e $\beta, \alpha > 0$. Desta forma, sua distribuição acumulada é dada por:

$$F(x; \mu; \alpha; \beta) = 1 - e^{-\left(\frac{x - \mu}{\beta} \right)^\alpha}. \quad (2.8)$$

2.2 Mistura de Distribuições

Conforme McLachlan & Peel (2000), misturas finitas de distribuições de qualquer natureza têm sido objeto de atenção crescente principalmente nas últimas décadas, pois possibilitam uma flexibilização bastante útil na modelagem de diversos fenômenos aleatórios. Tais fenômenos estão presentes em diversas áreas do conhecimento, em que se faz uso de misturas de distribuições, como na engenharia, astronomia, física, medicina, etc. As aplicações nessas áreas levam em consideração que a população em questão é composta por duas ou mais subpopulações. Desta forma a densidade da população pode apresentar mais de uma moda e/ou assimétrica que não vem de uma única população; nesse sentido, Pearson (1984) fez uso de uma mistura entre duas distribuições normais para ajustar dados referentes à medida do tamanho de carangueijos da Baía de Nápoles.

Ainda segundo McLachlan & Peel(2000), dadas as funções de densidades f_1, \dots, f_g de uma família F, um modelo de mistura finita dessas g densidades é a combinação convexa

$$f(y_j, \Theta) = \sum_{i=1}^g p_i f_i(y_j, \Theta_i), \quad (2.9)$$

em que $f_i(y_j)$ representa as densidades componentes da mistura e p_i são as proporções de mistura, onde:

$$\sum_{i=1}^g p_i = 1, \quad (2.10)$$

com $0 < p_i < 1$ e $i = 1, \dots, g$.

Neste trabalho, o tempo entre ocorrências em um processo de contagem seguirá uma distribuição homogênea Weibull ou um distribuição heterogênea, mistura de duas distribuições Weibull.

2.2.1 Mistura de Distribuições Weibull

Ao substituir (2.7) em (2.9) com $n=2$, obtemos o modelo de mistura de duas distribuições Weibull, conforme a seguinte equação:

$$f(x|\Theta) = p_1 \frac{\alpha_1}{\beta_1} \left(\frac{x}{\beta_1}\right)^{\alpha_1-1} e^{-\left(\frac{x}{\beta_1}\right)^{\alpha_1}} + p_2 \frac{\alpha_2}{\beta_2} \left(\frac{x}{\beta_2}\right)^{\alpha_2-1} e^{-\left(\frac{x}{\beta_2}\right)^{\alpha_2}}, \quad (2.11)$$

em que $\alpha_i > 0, \beta_i > 0$ são os parâmetros de forma e escala, respectivamente, da subpopulação i , p_i proporções da mistura com $p_2 = 1 - p_1$, e $\Theta = (p_1, \alpha_1, \alpha_2, \beta_1, \beta_2)$ é o vetor de parâmetros do modelo de mistura Weibull (MMW) definido em (2.11).

Os momentos para uma mistura de distribuições Weibull são obtidos da seguinte forma:

$$\begin{aligned} M_F(t) &= \int_0^{\infty} x^t d(p_1 F_1(x) + p_2 F_2(x)) dx = p_1 \int_0^{\infty} x^t df_1(x) dx + p_2 \int_0^{\infty} x^t df_2(x) dx = \\ &= p_1 M_{F_1}(t) + p_2 M_{F_2}(t). \end{aligned}$$

Logo, temos que:

$$M_F(t) = p_1 \beta_1^t \Gamma\left(\frac{t}{\alpha_1} + 1\right) + p_2 \beta_2^t \Gamma\left(\frac{t}{\alpha_2} + 1\right). \quad (2.12)$$

Sendo assim, os dois primeiros momentos são calculados:

$$\mu = E(X) = M_F(1) = p_1 \beta_1 \Gamma\left(\frac{1}{\alpha_1} + 1\right) + p_2 \beta_2 \Gamma\left(\frac{1}{\alpha_2} + 1\right); \quad (2.13)$$

$$E(X^2) = p_1 \beta_1^2 \Gamma\left(\frac{2}{\alpha_1} + 1\right) + p_2 \beta_2^2 \Gamma\left(\frac{2}{\alpha_2} + 1\right).$$

Para ilustrar o modelo (2.11), apresentamos na Figura 6 as densidades do MMW para os quatro vetores de parâmetros dados na Tabela 1:

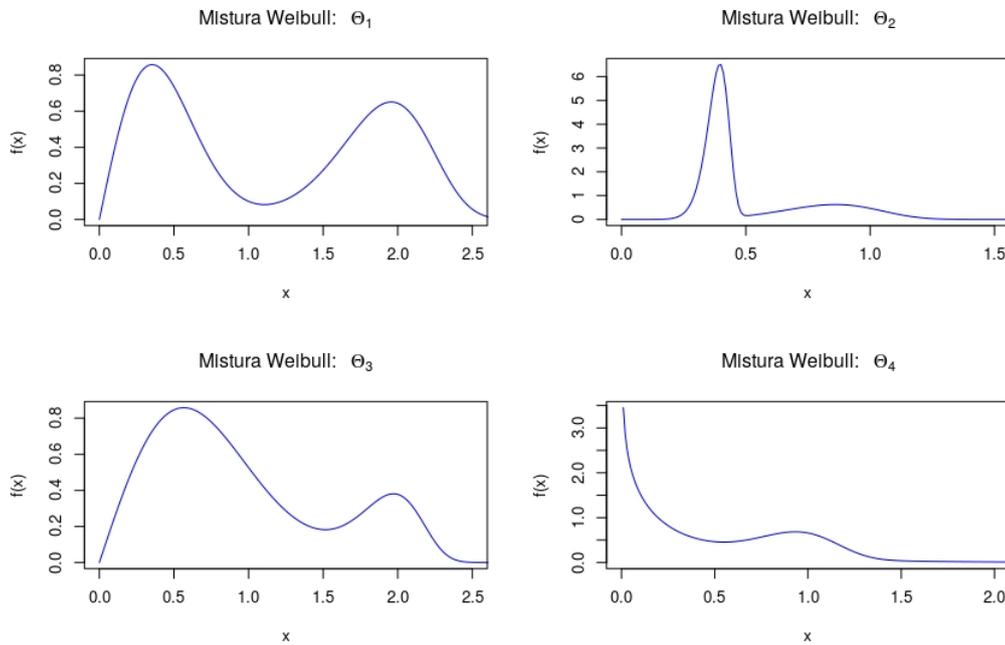


Figura 3 – Misturas de distribuições *Weibull* para parâmetros da Tabela 1.

Tabela 1 – Vetor de parâmetros do MMW (2.11).

Θ_i	α_1	β_1	α_2	β_2	p_1
Θ_1	2	0.5	7	2	0.5
Θ_2	5	0.9	10	0.4	0.3
Θ_3	2	0.8	10	2	0.8
Θ_4	0.8	0.3	5	1	0.7

Observa-se, principalmente a partir da mistura dada pelo vetor de parâmetros Θ_4 , que um valor de $\alpha \leq 1$ dá à mistura uma tendência exponencial visível.

2.3 Processos de Poisson e Renovação

De acordo com ROSS (2010), um processo estocástico $\{X(t), t \in T\}$ é uma coleção de variáveis aleatórias, em que para cada $t \in T$, $X(t)$ é uma variável aleatória; no qual o índice t é frequentemente associado ao tempo, sendo então $X(t)$ o *estado* do processo no tempo t .

O conjunto T é chamado de conjunto de índices do processo. Quando T é um conjunto contável, o processo estocástico é chamado de processo a *tempo-discreto*, caso T seja um intervalo no conjunto de números reais, temos um processo a *tempo-contínuo*. Então, o conjunto $\{X_n, n = 0, 1, 2, \dots\}$ é um processo estocástico a *tempo-discreto* indexado

por valores inteiros não negativos; enquanto que $\{X_t, t \geq 0\}$ é um processo estocástico a *tempo-contínuo* indexado por valores reais não negativos.

O *espaço de estados* de um processo estocástico é definido como o conjunto de todos os valores possíveis que a variável aleatória $X(t)$ pode assumir.

Portanto, um processo estocástico é uma família de variáveis aleatórias que descrevem a evolução no tempo de algum processo (físico).

Definição 2.1: Um processo de contagem, conforme ROSS (2010), é definido como um processo estocástico a tempo contínuo com espaço de estados discreto, $\{N(t), t \geq 0\}$, em que $N(t)$ representa o número de eventos ocorridos no tempo t e que satisfaz as seguintes propriedades:

- (i) $N(t) \geq 0$;
- (ii) $N(t)$ é um valor inteiro;
- (iii) Se $s < t$, então $N(s) \leq N(t)$;
- (iv) Para $s < t$, $N(t) - N(s)$ é o número de eventos no intervalo $(s, t]$.

Se o número de eventos que ocorre em intervalos de tempo disjuntos são independentes, o processo de contagem possui então *incrementos independentes*. Como exemplo, considere $N(t)$ como o número de veículos que chegam em determinado cruzamento da via a um tempo t , então $\{N(t), t \geq 0\}$ é um processo de contagem que corresponde a chegada desses veículos no cruzamento, e nesse caso é razoável entender que esse processo de contagem possui incrementos independentes.

Um processo de contagem é dito ter *incrementos estacionários*, se a distribuição do número de eventos que ocorrem em um intervalo de tempo depende somente do tamanho desse intervalo, ou seja, o processo possui incrementos estacionários se o número de eventos no intervalo $(s, s + t]$ tem mesma distribuição para qualquer valor de $(0, t]$, independente do valor de s . Isso pode ocorrer, por exemplo, no caso anterior de veículos chegando em um determinado cruzamento da via, porém, se houver nesse local horários em que o fluxo aumenta (como num horário de *rush*), não se poderá com razoabilidade considerar que esse processo possui incrementos estacionários.

2.3.1 Processos de Poisson

Definição 2.2: O processo de contagem $\{N(t), t \geq 0\}$ é um processo de Poisson com taxa $\lambda > 0$ se:

- (i) $N(0) = 0$.

- (ii) O processo tem incrementos independentes.
- (iii) O número de eventos em qualquer intervalo de tamanho t tem distribuição Poisson com média λt ; ou seja, para todo $s, t \geq 0$

$$P\{N(t+s) - N(s) = n\} = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \quad n = 0, 1, \dots \quad (2.14)$$

Nota-se que a partir da condição (iii) o processo de Poisson tem incrementos estacionários e também e

$$E[N(t)] = \lambda t, \quad (2.15)$$

em que λ é a ser a taxa do processo.

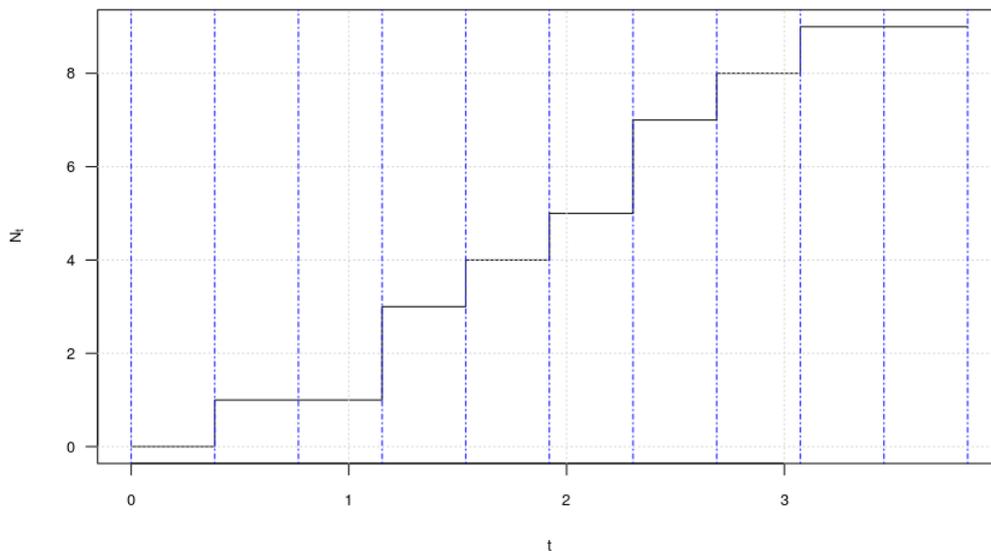


Figura 4 – Processo de Contagem Poisson ($\lambda = 2$).

Na figura 4 observa-se uma possibilidade de processo de contagem Poisson com taxa $\lambda = 2$.

Considere um processo de Poisson e seja o tempo do primeiro evento dado por T_1 . Ainda, para $n > 1$, T_n denota o tempo entre o $(n - 1)$ -ésimo e o n -ésimo evento. A sequência $\{T_n, n = 1, 2, \dots\}$ representa os tempos entre as chegadas dos eventos.

Mostra-se que os tempos entre as chegadas de um Processo de Poisson com taxa λ , $\{T_n, n = 1, 2, \dots\}$, são variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) com distribuição exponencial de parâmetro λ , ver ROSS (2010). Isto é, $F(t) =$

$P(T_n \leq t) = 1 - e^{-\lambda t}$. Isso significa que o processo não possui memória e a cada ocorrência temos um recomeço do processo.

Uma outra medida de interesse é o tempo total até a ocorrência do n -ésimo evento, S_n , ou tempo de espera até o n -ésimo evento, sendo que

$$S_n = \sum_{i=1}^n T_i, \quad n \geq 1, \quad (2.16)$$

conforme a proposição anterior, então S_n tem distribuição gama com parâmetros n e λ :

$$f_{S_n}(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}, \quad t \geq 0. \quad (2.17)$$

Exemplo 1: Suponha que pessoas imigrem para um território de acordo com uma distribuição Poisson com taxa diária de $\lambda = 1$. Qual seria o tempo esperado até que o décimo imigrante chegue, e qual a probabilidade que o tempo entre o décimo e o décimo primeiro imigrante seja maior que dois dias?

Para o primeiro caso, temos que $E[S_{10}] = \frac{10}{\lambda} = 10$ dias.

No segundo caso, temos que $P\{T_{11} > 2\} = e^{-2\lambda} = e^{-2} \approx 0.133$.

Um importante resultado de um processo de Poisson afirma que um processo de contagem $\{N(t), t > 0\}$ é um processo de Poisson homogêneo com taxa λ se e somente se os tempos entre as chegadas $\{T_n, n = 1, 2, \dots\}$ são v.a.'s i.i.d. com distribuição $\exp(\lambda)$.

Diversos são os exemplos de aplicação do processo de Poisson, os quais são caracterizados por possuir eventos que ocorrem de forma repetida e aleatorizada ao longo do tempo, isto é, o número de eventos que ocorre em intervalos de tempos disjuntos, com o mesmo tamanho, são independentes e identicamente distribuídos. Além disso, os tempos entre as ocorrências são variáveis aleatórias independentes e identicamente distribuídas de forma exponencial.

Alguns processos são considerados homogêneos, ou seja, a taxa desses eventos é constante, ou heterogêneos, com mais de um tipo de ocorrência. Como exemplo, podemos modelar dessa forma o número de e-mails que são recebidos, e que, havendo uma classificação etária das pessoas que recebem os e-mails é possível obter um caso de taxas heterogêneas de recebimento. Acidentes de trânsito, considerando o envolvimento de tipos diversos de veículos (automóvel, motocicleta, caminhão, etc.), traz uma situação de provável heterogeneidade das taxas desses acidentes; número de clientes que chegam a uma loja, etc. Para mais detalhes sobre processos de Poisson não-homogêneos ver ROSS (2010).

Nas situações que serão tratadas neste trabalho, nem sempre os tempos entre as ocorrências de eventos (acidentes) estarão ajustados por uma distribuição exponencial.

Uma análise descritiva prévia dos dados em estudo nos mostrou que tal distribuição pode ser uma Weibull ou ainda uma mistura de duas distribuições Weibull, por hipótese. Nesse último caso, precisa-se de um processo de contagem mais geral que Poisson para garantir que a distribuição entre as chegadas pode ser qualquer distribuição F , em particular, uma mistura de Weibull.

2.3.2 Processos de Renovação

Um processo de contagem $N(t), t \geq 0$ é um processo de Renovação com função de renovação $m(t) = E[N(t)]$ somente se os tempos entre as chegadas dos eventos, $\{X_n, n = 1, 2, \dots\}$ são variáveis aleatórias i.i.d. F qualquer. Para um processo de renovação em que X_1, X_2, \dots são os tempos entre as ocorrências, e o valor de $S_0 = 0$, define-se

$$S_n = \sum_{i=1}^n X_i, \quad n \geq 1, \quad (2.18)$$

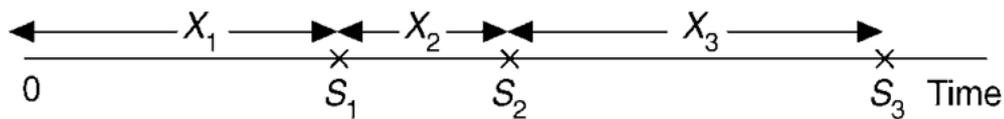


Figura 5 – Processo de Renovação.

em que $S_1 = X_1$ representa o tempo até a primeira renovação do processo, $S_2 = X_1 + X_2$ o tempo da segunda renovação, e assim sucessivamente; ou seja, S_n representa o tempo da n -ésima renovação.

Assim, sendo $X_i, i \geq 1$, variáveis aleatórias independentes identicamente distribuídas com distribuição F ; $S_n = \sum_{i=1}^n X_i$ tem distribuição F_n (convoluções de F), então a distribuição do processo e a função de renovação são dadas, respectivamente, por:

$$\begin{aligned} P\{N(t) = n\} &= P\{N(t) \geq n\} - P\{N(t) \geq n + 1\} = \\ &= P\{S_n \leq t\} - P\{S_{n+1} \leq t\} \Rightarrow \\ &= P\{N(t) = n\} = F_n(t) - F_{n+1}(t), \end{aligned} \quad (2.19)$$

$$m(t) = E[N(t)] = \sum_1^{\infty} P\{N(t) \geq n\} = \sum_1^{\infty} P\{S_n \leq t\} = \sum_1^{\infty} F_n(t). \quad (2.20)$$

A função $m(t)$ é conhecida como o valor médio da função de renovação.

Na teoria de probabilidade, a convolução de distribuições de probabilidade corresponde à operação de adição de variáveis aleatórias independentes, formando combinações lineares dessas variáveis. Para mais detalhes sobre convolução ver Hogg *et al* (2004).

Exemplo 2: Suponha um processo de renovação cuja função de valor médio é dada por:

$$m(t) = 2t, \text{ com } t \geq 0.$$

Qual a distribuição do número de renovações que ocorrem no tempo $t = 10$?

Solução: Da equação (2.13) temos que $E[N(t)] = \lambda t$, e como $m(t) = E[N(t)] = 2t$, temos que $N(t)$ é um processo de Poisson com taxa 2. Assim:

$$P\{N(10) = n\} = e^{-20} \frac{(20)^n}{n!}, \quad n \geq 0.$$

O processo de Poisson é o único caso em que o processo de renovação tem uma função de valor médio linear.

Mostra-se que $m(t)$ unicamente determina o processo de renovação, ou seja, há uma correspondência um para um entre a distribuição F dos tempos entre as ocorrências e $m(t)$. Isso pode ser evidenciado ao assumir que a distribuição dos tempos entre as ocorrências possui função de densidade f e distribuição acumulada F , tem-se que a função de renovação é determinada por F , ao resolver a "equação de renovação":

$$m(t) = \int_0^t [1 + m(t-x)]f(x)dx = F(t) + \int_0^t m(t-x)f(x)dx. \quad (2.21)$$

2.3.3 Distribuição Assintótica de um Processo de Renovação.

De (2.20) tem-se que a distribuição de um processo de renovação $\{N(t), t > 0\}$ é obtido pela convolução da distribuição dos tempos entre as chegadas, F . Porém, essas convoluções nem sempre têm expressões fechadas simples, o que dificulta o cálculo da distribuição exata do processo de renovação. Uma alternativa para o cálculo da distribuição aproximada do processo de renovação é dada considerando-se o limite assintótico do processo. A prova de todos os resultados apresentados a seguir podem ser encontrados em ROSS (2010).

Proposição 2.3: Seja $\{N(t), t > 0\}$ um processo de renovação com função de renovação $m(t)$ e sejam $\{X_n, n > 0\}$ os tempos entre as renovações variáveis aleatórias i.i.d F com $E(X_n) = \mu$. Então, com probabilidade 1, tem-se que

$$\frac{N(t)}{t} \rightarrow \frac{1}{\mu}, \text{ quando } t \rightarrow \infty. \quad (2.22)$$

Algumas considerações: i) A proposição anterior é válida mesmo que a média dos tempos entre as renovações, μ , for infinita; nesse caso, interpreta-se $\frac{1}{\mu}$ como zero; ii) O valor de $\frac{1}{\mu}$ é chamado de taxa de renovação; iii) Como a taxa entre renovações é μ , é

intuitivo que a taxa média em que ocorre as renovações é de 1 para cada μ unidades de tempo.

A proposição 2.3 afirma que com probabilidade 1, para um tempo suficientemente longo a taxa média de renovações do processo é de $\frac{1}{\mu}$.

Proposição 2.4: Seja $\{N(t), t > 0\}$ um processo de renovação com função de renovação $m(t)$ e sejam $\{X_n, n > 0\}$ os tempos entre as renovações v.a.'s i.i.d F com $E(X_n) = \mu$. Então

$$\frac{m(t)}{t} \rightarrow \frac{1}{\mu}, \text{ quando } t \rightarrow \infty. \quad (2.23)$$

O valor $\frac{1}{\mu}$ é chamado de taxa de renovação do processo.

Proposição 2.5: (Teorema do Limite Central para um Processo de Renovação) Seja $\{N(t), t > 0\}$ um processo de renovação e sejam $\{X_n, n > 0\}$ os tempos entre as renovações v.a.'s i.i.d F com $E(X_n) = \mu$ e $Var(X_n) = \sigma^2$. Então

$$\lim_{t \rightarrow \infty} P \left[\frac{N(t) - \frac{t}{\mu}}{\sqrt{\frac{t\sigma^2}{\mu^3}}} < x \right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{x^2}{2}} dx \Rightarrow \quad (2.24)$$

$$N(t) \sim N\left(\frac{t}{\mu}, \frac{t\sigma^2}{\mu^3}\right).$$

Exemplo 3: Uma máquina processa determinado trabalho ininterruptamente. O tempo que a máquina leva para executar um trabalho após o outro tem distribuição gama com parâmetros $n = 4$ e $\lambda = 2$.

a) Qual é o número médio de trabalhos processados em até $t = 100$ horas?

Para um valor de t suficientemente grande, de acordo com (2.22), tem-se que $m(t) = E(N(t)) = \frac{t}{\mu}$.

b) Qual é a probabilidade aproximada de que a máquina processe no máximo 90 trabalhos até o tempo t ?

Se designamos $N(t)$ como o número de trabalhos que a máquina pode processar até o tempo t , então $\{N(t), t \geq 0\}$ é um processos de renovação.

Temos que $P(N(t) \leq 90) = P\left(Z = \frac{90 - \frac{t}{2}}{\sqrt{\frac{t}{8}}}\right)$. Utilizando a proposição 2.5, $\mu = 2$ e $\sigma^2 = 1$. Então, para um tempo suficientemente grande, $N(t) \sim N\left(\frac{t}{2}, \frac{t}{8}\right)$.

3 Estimação

3.1 Estimação por Máxima Verossimilhança dos Tempos entre Acidentes

Como o objetivo deste trabalho é modelar dados de acidentes de trânsito em algumas rodovias federais, baseado na revisão da teoria de processos de contagem dada no capítulo 2, primeiro procurou-se a distribuição F que melhor se ajuste aos dados dos tempos entre os acidentes. Empiricamente notou-se que essa distribuição pode ser de uma única população ou de uma mistura de duas populações. Aqui, utilizamos o MMV (Método da Máxima Verossimilhança) para estimar os parâmetros dessa distribuição. O método de MV para uma população $X \sim F$ consiste em, dada uma amostra $\{X_1, X_2, \dots, X_n\}$, de $X \sim F$, então a função de verossimilhança é

$$L(\theta) = f(x_1; \theta)f(x_2; \theta) \cdots f(x_n; \theta) = \prod_{i=1}^n f(x_i; \theta). \quad (3.1)$$

A função $L(\theta)$ fornece a verossimilhança que as variáveis aleatórias assumem para um valor particular de x_1, x_2, \dots, x_n . A verossimilhança é o valor da função de densidade conjunta da amostra, para variáveis contínuas, e a função de probabilidade conjunta, para variáveis discretas.

O valor de θ que maximiza $L(\theta)$ é o estimador de máxima verossimilhança (EMV).

Isto é, o EMV é a solução de

$$\frac{\partial L(\theta)}{\partial \theta} = 0, \quad (3.2)$$

$$\frac{\partial^2 L(\theta)}{\partial \theta^2} > 0;$$

sendo que, por conveniência, pode-se substituir na derivada $L(\theta)$ pela função log-verossimilhança, $\log[L(\theta)]$, pois o valor que maximiza uma também maximiza a outra função.

Quando o modelo da população vem de uma mistura de duas sub populações, do tipo (2.9), o procedimento de MV tem que utilizar o algoritmo EM.

3.1.1 O Algoritmo EM

Definido por Dempster et al. (1977), o algoritmo EM tem seu uso aplicado ao cálculo do estimador de máxima verossimilhança (EMV), fazendo uso, basicamente, de dois

passos: o primeiro, passo "E" que calcula o valor esperado do logaritmo da verossimilhança e o passo "M", que consiste em calcular o valor máximo da função anterior. Os passos são repetidos em sequência até se chegar numa convergência, a partir da definição de um critério de parada.

De forma geral, as aplicações do algoritmo EM leva em consideração a incompletude dos dados e quando a maximização da função de verossimilhança apresenta dificuldades as quais são resolvidas com a admissão de valores adicionais, resultando em maior simplificação do seu cálculo, pois, conforme os trabalhos de Zhu e Lee (2001), na maioria das vezes, a função log-verossimilhança dos dados completos apresenta maior simplificação em sua forma que a log-verossimilhança dos dados observados.

O algoritmo EM converge para o EMV e tal prova, com correção a partir da prova dada por Dempster et al. (1977), pode ser encontrada em Boyles (1983) e Wu (1983).

Considerando o conjunto de dados \mathbf{y} observados e o conjunto de dados faltantes \mathbf{m} , temos que o conjunto de dados completos é estabelecido por $\mathbf{y}_c = (\mathbf{y}, \mathbf{m})$, com função densidade de probabilidade dada por $f(\mathbf{y}_c|\theta)$, com $\theta \in \Theta \subseteq \mathfrak{R}^p$. Temos que a função log-verossimilhança para os dados completos é dada por $l_c(\theta|\mathbf{y}_c)$, e assim temos que os dois passos do algoritmo EM são dados por:

- E (*Expectation*): calculamos $Q(\theta|\theta^{(k)}) = E[l_c(\theta|\mathbf{y}_c)|\mathbf{y}, \theta^{(k)}]$, em que a esperança é tomada em relação à distribuição condicional $f(\mathbf{y}|\mathbf{m}, \theta^{(k)})$.
- M (*Maximization*): encontrar o valor de $\theta^{(k+1)}$ que maximiza $Q(\theta|\theta^{(k)})$.

Os passos "E" e "M" são repetidos até atingir a convergência sujeita ao critério de parada em que $|\theta^{(k+1)} - \theta^{(k)}| < \epsilon$.

O uso do algoritmo EM para estimar os parâmetros de uma mistura de distribuições Weibull foi amplamente utilizado e, apesar de se encontrar outras abordagens envolvendo esse tipo de estimação, McLachlan & Peel (2000), referência para este trabalho no que se refere à estimação, enfatizou o uso do algoritmo EM em sua obra.

3.1.2 Algoritmo EM para Mistura de Distribuições Weibull

Um dos objetivos deste trabalho é encontrar os parâmetros da mistura de distribuições Weibull relacionados aos dados de tempo entre acidentes graves ocorridos em algumas rodovias federais do Brasil. Os parâmetros são encontrados com a utilização da técnica de algoritmo EM aplicadas em uma mistura de distribuições Weibull conforme equação (2.9), ou seja, encontrar valores para $\Theta = (\pi, \alpha_1, \alpha_2, \beta_1, \beta_2)$.

Seja X uma variável aleatória com valores X_1, X_2, \dots, X_n e distribuição de probabilidade F dada conforme (2.11), e sendo $K = (k_1, k_2, \dots, k_N)$ subconjunto de dados

incompletos da variável X de tamanho N , e $Z = (K, Y)$ os dados completos, em que $Y = (y_1, y_2, \dots, y_n)$. A função de distribuição e a função de verossimilhança dos dados incompletos, respectivamente, é definida por $f(x; \Theta)$ e $L(\Theta|K)$.

O algoritmo EM encontra primeiro a esperança dos valores da função log-verossimilhança dos dados completos, com relação aos dados desconhecidos Y , dado os valores de K e os parâmetros estimados, ou seja,

$$\begin{aligned} Q(\Theta; \Theta^{(k)}) &= E[\ln(h(K, Y|\Theta))|K, \Theta^{(k)}] = \\ &= \sum_{l=1}^2 \sum_{i=1}^N \log(p_l) f(l|x_i; \Theta_l^{(k)}) + \sum_{l=1}^2 \sum_{i=1}^N \log[f_l(x_i; \theta_l^{(k)})] f(l|x_i; \Theta_l^{(k)}), \end{aligned} \quad (3.3)$$

Em que Θ^k são os atuais parâmetros estimados que se usa para calcular o valor da esperança e Θ são os novos parâmetros que otimizados e incrementados em Q , definindo assim o primeiro passo do algoritmo (*expectation*). O segundo passo, *maximization*, é definido como a maximização da esperança computada anteriormente.

A partir de então, faz-se a alternância entre cada um dos passos do algoritmo EM.

No segundo passo, na $(k+1)$ -ésima interação, maximizar $Q(\Theta; \Theta^{(k)})$ com relação a p_1 é equivalente a atualizar p_1^{k+1} por

$$p_1^{(k+1)} = \frac{1}{N} \sum_{i=1}^N f(1|x_i, \Theta_1^{(k)}), \quad (3.4)$$

em que,

$$f(l|x_i; \Theta_l^{(k)}) = \frac{p_l \frac{\alpha_l}{\beta_l} \left(\frac{x}{\beta_l}\right)^{\alpha_l-1} e^{-\left(\frac{x}{\beta_l}\right)^{\alpha_l}}}{\sum_{l=1}^2 p_l \frac{\alpha_l}{\beta_l} \left(\frac{x}{\beta_l}\right)^{\alpha_l-1} e^{-\left(\frac{x}{\beta_l}\right)^{\alpha_l}}}, \quad l = 1, 2; \quad (3.5)$$

e atualiza-se $\Theta_l^{(k+1)}$, $l = 1, 2$ ao se resolver a equação

$$\sum_{i=1}^N f(l|x_i, \Theta_l^{(k)}) \frac{\partial}{\partial \Theta_l^{(k)}} \log(f_l(x_i; \Theta_l^{(k)})) = 0$$

Então, na $(k+1)$ -ésima interação a atualização das estimativas de $\alpha_l^{(k+1)}$ e $\beta_l^{(k+1)}$, para $l = 1, 2$, são obtidos, respectivamente, ao solucionar o seguinte sistema de equações:

$$\sum_{i=1}^N f(l|x_i; \Theta_l^{(k)}) \left\{ \frac{1}{\alpha_l} + \ln(x_i) + \ln(\beta_l) - \frac{\ln(x_i)}{\beta_l} e^{\alpha_l \frac{\ln(x_i)}{\beta_l}} \right\} = 0 \quad (3.6)$$

e de β_l , $l = 1, 2$, ao calcular

$$\beta_l = \left\{ \frac{-\alpha_l \sum_{i=1}^N f(l|x_i; \Theta_l^{(k)})}{\sum_{i=1}^N f(l|x_i; \Theta_l^{(k)}) x_i^{\alpha_l}} \right\}^{-\alpha_l}. \quad (3.7)$$

3.1.3 Procedimentos para Modelagem dos Dados

A seguir, são apresentados os passos necessários para o ajuste dos dados de acidentes:

Passo 1: Utilizar as equações (3.1) ou (3.3) a (3.5) para estimar os parâmetros da distribuição F dos tempos entre acidentes, definida em (2.2) ou por (2.11), respectivamente. O primeiro caso é para uma Weibull e o segundo para a mistura de duas Weibull.

Passo 2: Fazer o estudo do número de acidentes de acordo com as equações (2.21) a (2.23).

4 Aplicação

Foram coletados a partir de diversas rodovias federais do país os dados de acidentes graves (acidente com pelo menos uma vítima morta ou com ferimentos graves), ao longo dos anos de 2016 e 2017, conforme o quadro de variáveis abaixo:

Tabela 2 – Resumo da base de dados de acidentes da PRF.

Variável	Descrição
id	Número identificador da ocorrência (acidente)
pesid	Número identificador do envolvido no acidente
data	Data em que ocorreu o acidente
dia semana	Dia da semana do acidente, formato "Y-m-d"
horario	Horário do acidente, formato "HH:MM:SS"
uf	Unidade da Federação onde ocorreu o acidente
br	Número identificador da rodovia federal do acidente
km	Quilômetro onde ocorreu o acidente
municipio	Município de ocorrência do acidente
causa acidente	Causa principal presumida do acidente, conforme classificação da PRF
tipo acidente	Especificação do tipo de acidente, conforme tabela classificação da PRF
condicao meteorologica	Classificação meteorológica no momento do acidente
tipo pista	Tipo de pista conforme classificação da PRF
id veiculo	Número identificador do veículo envolvido
tipo veiculo	Classificação do veículo quanto ao seu tipo
estado fisico	Classificação do estado físico do envolvido quanto às lesões
idade	Idade do envolvido no acidente
sexo	Sexo do envolvido no acidente

Inicialmente, as variáveis de interesse serão a data e o horário de cada acidente, para a criação de uma nova variável definida como a diferença (em dias) de uma ocorrência para outra, em que, ao considerar esses eventos sucessivos como independentes e identicamente distribuídos com uma distribuição qualquer de probabilidade, é possível associar um processo de contagem desses eventos em determinados períodos e, portanto, obtém-se um processo de renovação.

Na análise descritiva de alguns trechos de rodovias, é possível observar uma aproximação da densidade dos tempos entre os eventos a uma distribuição Weibull. Em vários casos, o "parâmetro de forma" encontrado da distribuição Weibull teórica tem valor próximo a 1, o que denota uma aproximação à distribuição exponencial e, por isso, chega-se a um Processo de Poisson.

Em algumas rodovias federais, no entanto, a distribuição dos tempos entre os acidentes aparenta ser composta por uma mistura de distribuições, aqui considerada hipoteticamente como uma mistura de distribuições Weibull.

4.1 Dados de Acidentes Graves na BR 101/RJ: Modelo Poisson

Um caso particular em que um processo de Poisson pode ser usado para modelagem de $N(t)$, contagem dos acidentes graves, está exemplificado nos dados dos tempos de acidentes na BR 101/RJ, ao longo dos anos de 2016 e 2017, com as estatísticas principais de acordo com a tabela 3.

Tabela 3 – Estatísticas dos tempos (em dias) entre acidentes – BR 101/RJ.

Observações	Mín	Quartil 1	Mediana	Média	Quartil 3	Máx	Desvio Padrão
550	0.0000	0.340	0.9000	1.3230	1.965	7.6200	1.3339

A distribuição dos dados pode ser visualizada abaixo na figura 6, com o histograma associado, a curva (em vermelho) da distribuição empírica e (em azul) a distribuição teórica exponencial com parâmetro $\lambda = 0.7545$ (taxa).

Na figura 7, compara-se as distribuições acumuladas empírica e teórica, além do gráfico *qqplot* (comparação dos quantis), mostrando um bom ajuste.

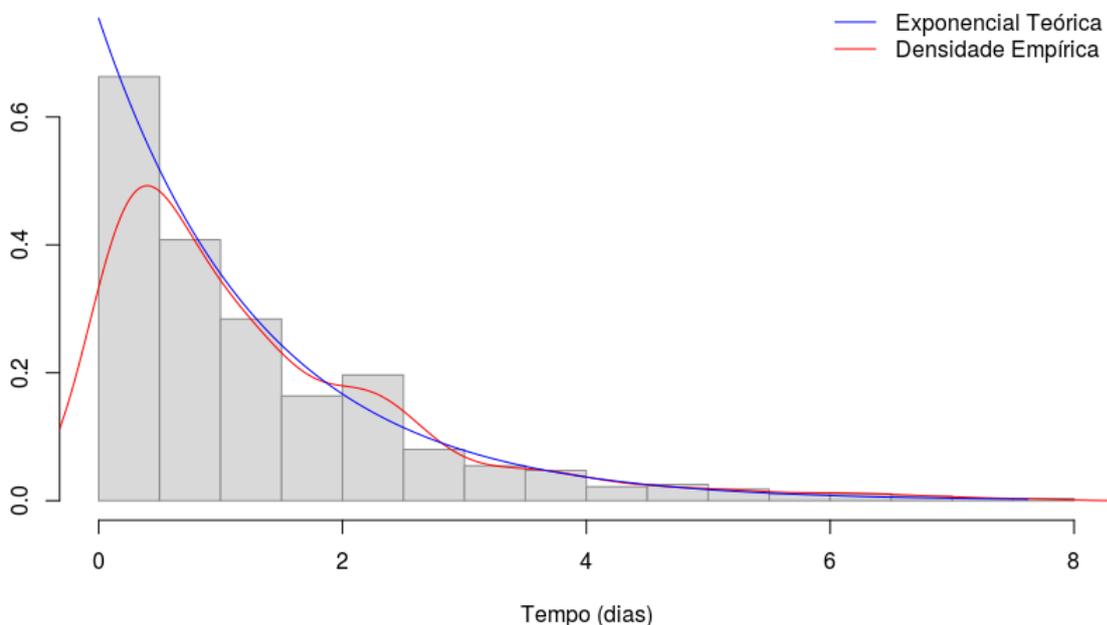


Figura 6 – Gráfico dos tempos entre acidentes (teórico e empírico).

Uma outra situação que se pode propor é, a partir do particionamento entre acidentes graves envolvendo motocicletas ou não, tentar verificar as diferenças entres esses grupos.

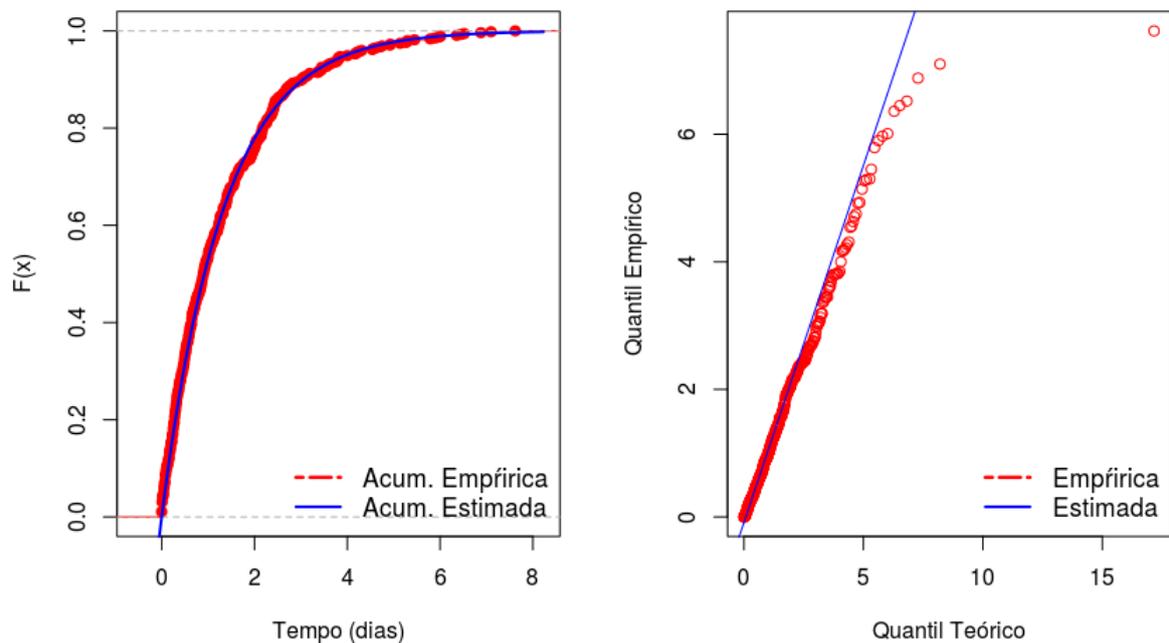


Figura 7 – Distribuição acumulada e *qqplot* dos tempos.

Para o caso dos acidentes em que houve a participação de alguma motocicleta no acidente e para as ocorrências sem nenhuma moto envolvida, as principais estatísticas dos tempos entre as ocorrências estão detalhadas na tabela 4.

Tabela 4 – Estatísticas dos tempos (em dias) com e sem motos – BR 101/RJ.

Tipo	Obs.	Mín	Quartil 1	Mediana	Média	Quartil 3	Máx	DP
Com Motos	173	0.000	0.810	2.710	4.193	6.040	27.060	4.496
Sem Motos	377	0.000	0.560	1.400	1.930	2.690	12.460	1.834

A distribuição dos tempos pode ser visualizada na figura 8, referente ao acidentes (a) em que houve alguma moto envolvida e (b) quando não houve, com o respectivo histograma associado, a curva (em vermelho) da distribuição empírica e (em azul) a distribuição teórica da curva exponencial (via Algoritmo EM).

As figuras 9 e 10 apresentam as distribuições acumuladas empírica e teórica, além do gráfico *qqplot* (comparação dos quantis), para cada um dos dois casos.

Em seguida, é apresentada a tabela 5 apresentando as estimativas do parâmetro λ da distribuição exponencial dos tempos entre as ocorrências, conforme método de máxima verossimilhança, considerando todos os tipos de veículos e para os casos em que houve uma partição entre os acidentes envolvendo motos ou não. Os testes de Kolmogorov-Smirnov apresentados na tabela reforçam a não rejeição da hipótese de que esses tempos têm distribuição exponencial.

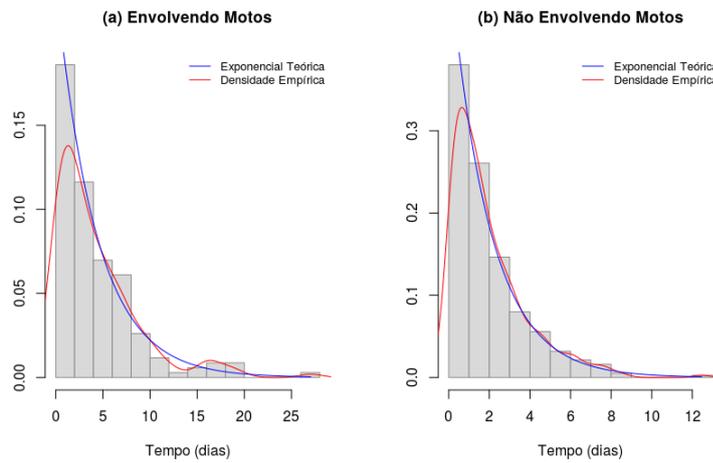


Figura 8 – Ajuste tempos entre acidentes com motos à distribuição exponencial.

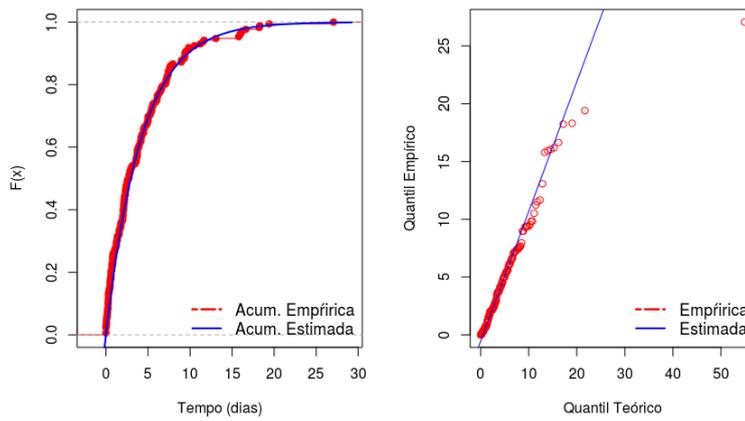


Figura 9 – Distribuição acumulada e *qqplot* dos tempos de acidentes com motos.

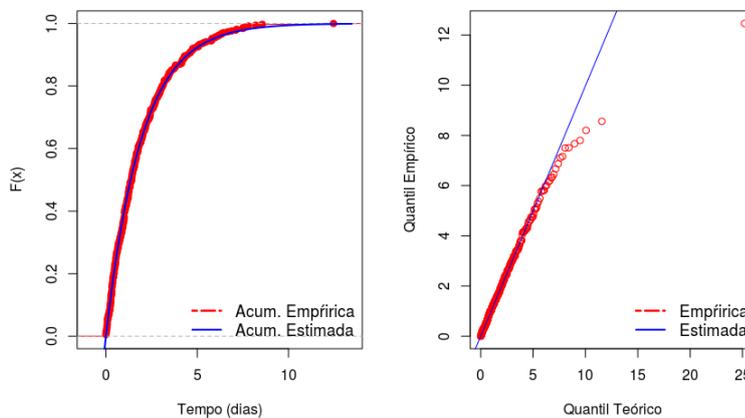


Figura 10 – Distribuição acumulada e *qqplot* dos tempos de acidentes sem motos.

Tabela 5 – Estimativas e valor de teste de KS para os tempos de acidentes.

Tipo	Taxa λ	Erro Padrão	KS	valor-p
Com Motos	0.24	0.0180	0.0762	0.2706
Sem Motos	0.51	0.0266	0.0325	0.8207
Total	0.75	0.0322	0.0275	0.8015

Observa-se que os valores dos parâmetros λ para os dois grupos somados chegam ao valor do parâmetros.

Considerando o total de acidentes ocorridos na BR 101/RJ, como exemplificação das probabilidades de ocorrer nenhum, um ou mais de um acidente grave dentro de determinado número de dias, a tabela 6 mostra algumas dessas probabilidades para a contagem de acidentes de acordo com o número de dias entre 0 e 7.

Tabela 6 – Probabilidades para $N(t)$ conforme t (dias) e n (acidentes).

t/n	0	1	2	3	4	5	6
1	0.47024	0.35480	0.13385	0.03366	0.00635	0.00096	0.00012
2	0.22113	0.33369	0.25177	0.12664	0.04778	0.01442	0.00363
3	0.10398	0.23537	0.26638	0.20099	0.11373	0.05149	0.01942
4	0.04890	0.14758	0.22269	0.22403	0.16903	0.10203	0.05132
5	0.02299	0.08675	0.16362	0.20576	0.19406	0.14642	0.09206
6	0.01081	0.04895	0.11080	0.16720	0.18922	0.17133	0.12927
7	0.00508	0.02685	0.07092	0.12485	0.16485	0.17413	0.15328

Um exemplo de gráfico de $N(t)$ pelo tempo t (em dias) é mostrado na figura 11, em que, para vários intervalos de tempo t , apresentam-se as possibilidades para contagem de acidentes graves $N(t)$, conforme o modelo Poisson especificado abaixo, levando-se em conta os valores das estimativas para λ da tabela 5:

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

para todo $n, t \geq 0$.

O gráfico da figura 10 evidencia o quanto a contagem de acidentes graves com envolvimento de moto é inferior ao caso complementar (com todos os outros tipos excluindo as motos), no mesmo intervalo de dias.

Como exemplo prático, para a ocorrência do 100º acidente grave nessa rodovia (com e sem motos), e, de acordo com o que foi visto na revisão teórica, calcula-se esse tempo da seguinte forma:

$$E[S_{100}] = \frac{n}{\lambda} = \frac{100}{0.7545} \approx 133,$$

ou seja, deve ocorrer o centésimo acidente após 133 dias.

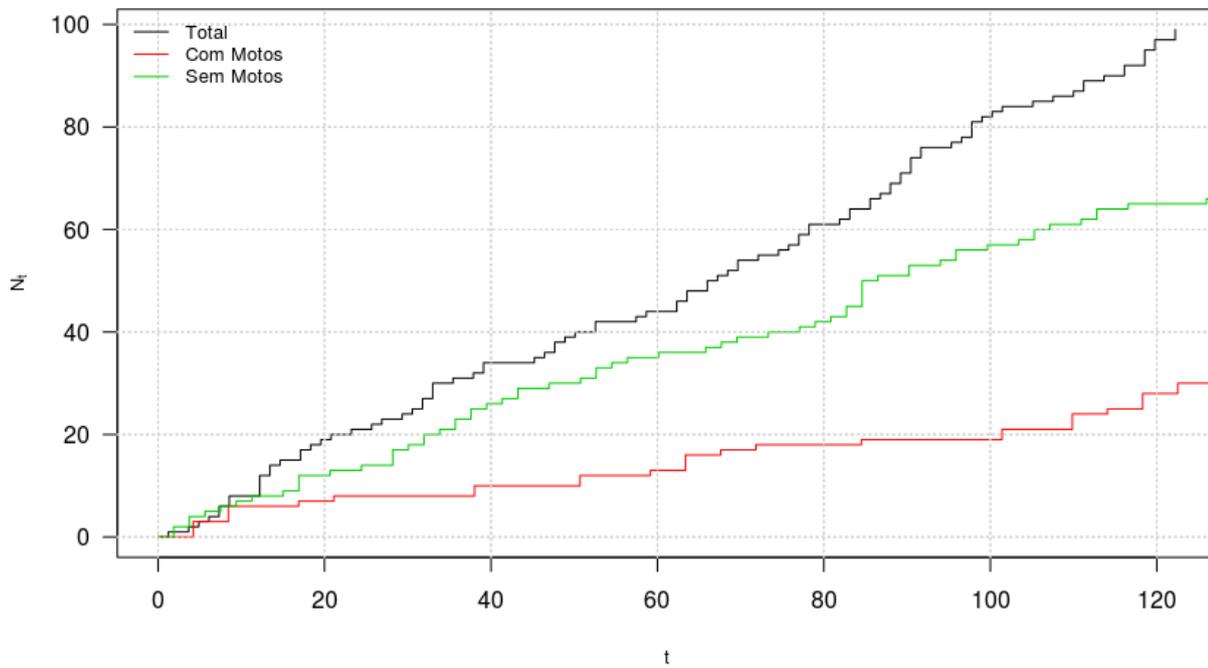


Figura 11 – Processo de Contagem Poisson com cada taxa da tabela 5.

Uma outra forma de se observar o comportamento para cada grupo, é observando as probabilidades de ocorrer um acidente com motos, sem motos ou geral, de acordo com o intervalo de dias, conforme figura 12.

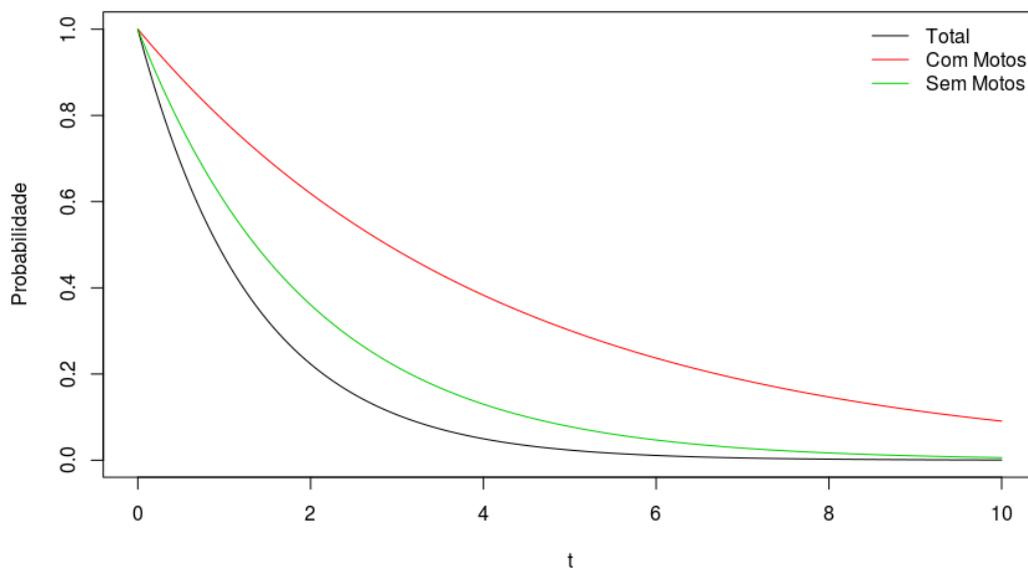


Figura 12 – Probabilidade dos eventos de acidentes por grupo ocorrerem em t dias.

Observa-se que os acidentes graves com motos ocorrem com maior probabilidade num intervalo mais espaçado de dias.

4.2 Dados de Acidentes Graves: Processo de Renovação

4.2.1 BR 110/BA

O próximo caso a ser discutido é referente aos dados dos tempos entre acidentes graves ocorrido na malha da BR 110/BA ao longo dos anos de 2016 e 2017. Diferentemente do caso anterior, BR 101/RJ, em que os tempos entre os acidentes graves estão ajustados a uma distribuição exponencial, os tempos entre os acidentes ocorridos na BR 110/BA serão analisados sob a hipótese de ajustarem-se a uma mistura entre duas distribuições Weibull, o que implica numa análise sob um processo de renovação.

As estatísticas principais dos tempos (em dias) do trecho da BR 110/BA são apresentadas conforme tabela abaixo:

Tabela 7 – Estatísticas dos Tempos entre Acidentes – BR 110/BA.

Observações	Mín	Quartil 1	Mediana	Média	Quartil 3	Máx	Desvio Padrão
254	0.0000	0.7125	1.8850	2.8747	4.4350	19.2000	2.8691

Os tempos entre as ocorrências, de 2016 a 2017, referente à BR 110/BA, considerando uma distribuição Weibull e, comparativamente, uma modelo de mistura Weibull (MMW), estão ilustrados na figura 13:

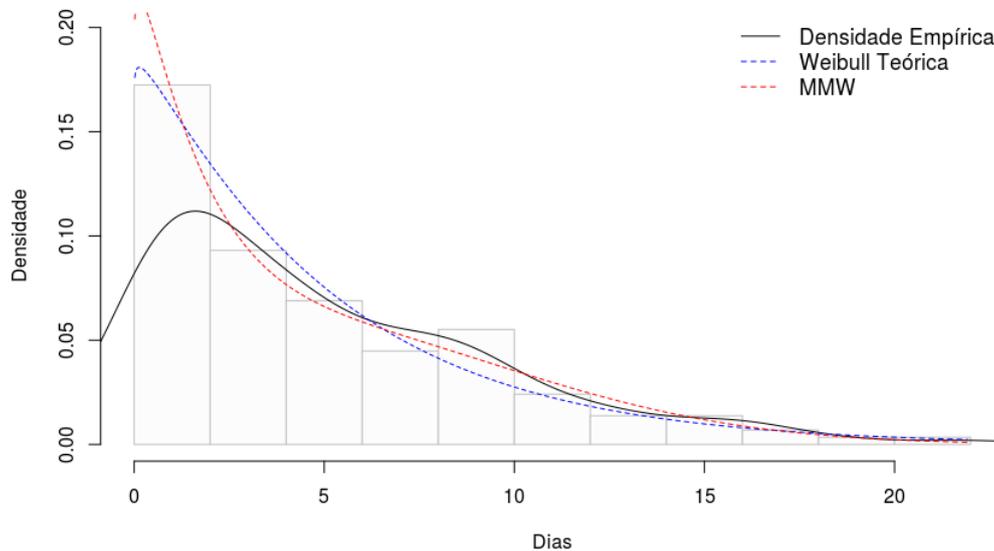


Figura 13 – Modelos para os tempos entre acidentes BR 110/BA.

Visualmente o modelo com mistura adere melhor à densidade empírica.

Os valores dos parâmetros para cada situação acima encontra-se na tabela 8.

Tabela 8 – Estimativas dos Parâmetros – BR 110/BA.

Modelo					
Weibull Simples	$\alpha = 1.025182$	$\beta = 5.049275$			
Weibull Mistura	$\alpha_1 = 1.87274$	$\beta_1 = 8.889720$	$\alpha_2 = 1.04734$	$\beta_2 = 2.02114$	$p = 0.50746$

A figura 14 apresenta as comparações entre as distribuições acumuladas dos tempos com um modelo teórico Weibull (a) e com uma mistura de modelos (b), conforme parâmetros estimados na tabela 8. Apresentando melhor conformidade o modelo de mistura.

A figura 15 ilustra a comparação entre os quantis dos valores empíricos versus uma Weibull teórica (a) e um MMW (b), conforme parâmetros estimados na tabela 8, o que mostra melhor ajuste também ao MMW.

O teste de Kolmogorov-Smirnov foi utilizado para verificação do melhor ajuste aos tempos entre acidentes graves e o resultado é indicado na tabela 9, onde se pode verificar que o modelo MMW apresentou melhor ajuste.

Apesar do ajuste melhor para os dados do modelo com mistura de Weibull, não houve uma diferença tão significativa com relação ao ajuste com um modelo simples.

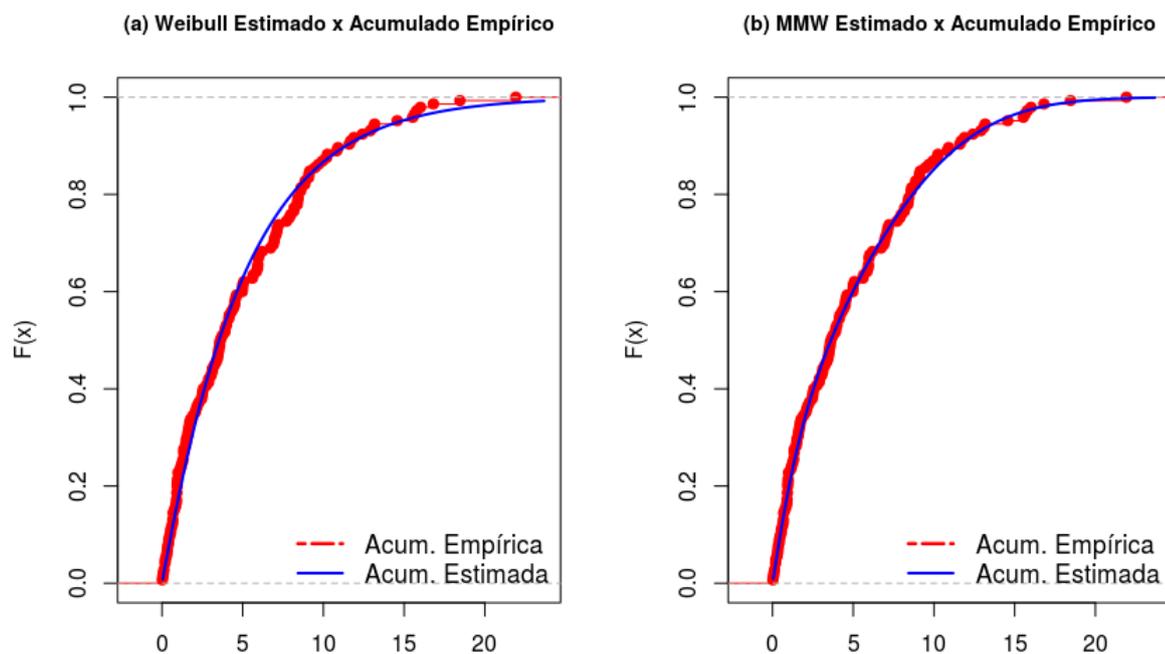


Figura 14 – Distribuições acumuladas para comparativas.

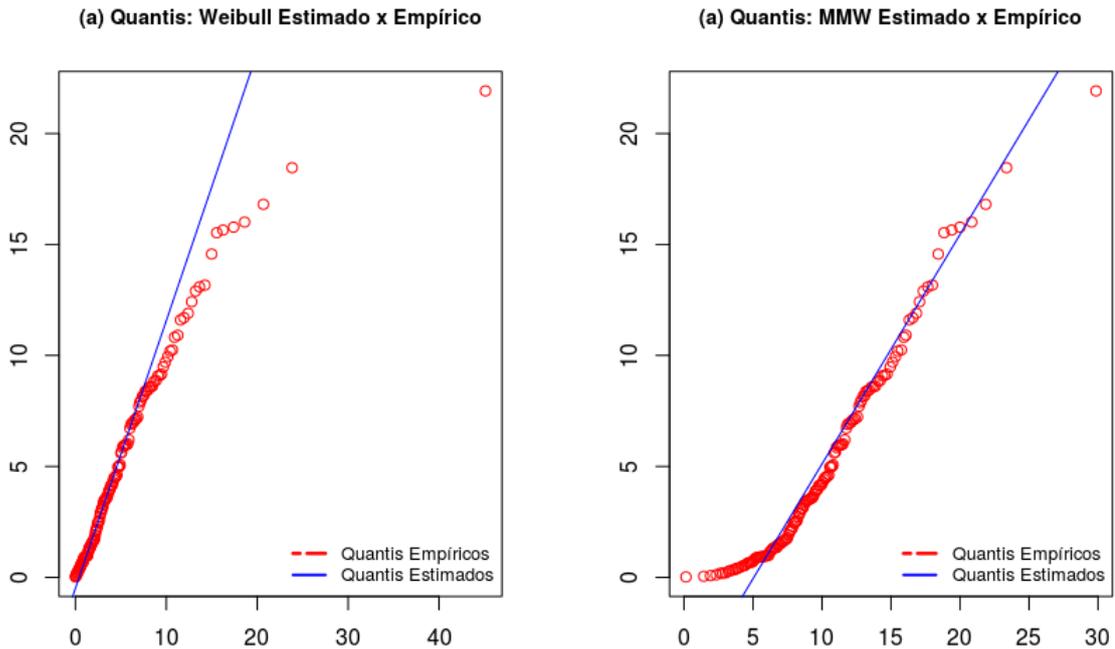
Figura 15 – Gráficos *qqplot*: Weibull e MMW.

Tabela 9 – Teste de Kolmogorv-Smirnov para os Dois Casos.

Modelo	KS	p-Valor
Weibull Simples	0.062069	0.9427
Weibull Mistura	0.041379	0.9997

De acordo com a proposição 2.5, que trata do teorema do limite central para um processo de renovação, sendo t um valor suficientemente grande, a distribuição das contagens de ocorrências, $N(t)$, é dada por $N(t) \sim \mathfrak{N}\left(\frac{t}{\mu}, \frac{t\sigma^2}{\mu^3}\right)$, sendo $\mu = E(X)$ e $\sigma^2 = Var(X)$.

Conforme discutido na sessão 2.2.1, o n -ésimo momento para a mistura é dado por:

$$Var(X) = \sigma^2 = E(X^2) - [E(X)]^2,$$

ou seja, ao considerar esse caso como um processo de renovação, em que os tempos são modelados por uma MMW, a distribuição das contagens de acidentes, $N(t)$, apresenta uma distribuição normal com média $\frac{t}{\mu}$ e variância $\frac{t\sigma^2}{\mu^3}$:

e para a mistura referente aos tempos entre os acidentes na BR 110/BA, tem-se que $\mu = E(X) = 4.9821$ e $\sigma^2 = Var(X) = 20.2183$:

$$N(t) \sim \mathfrak{N}\left(\frac{t}{4.9821}, 0.1635 \cdot t\right)$$

A tabela 10 apresenta diversos valores para a probabilidade de que ocorra um determinado número (n) de contagens de acidentes de acordo com o número de dias (t), além do valor esperado para esse número de dias conforme o valor para a média $\frac{t}{4.9821}$.

Tabela 10 – Probabilidades para $N(t)$ (mistura) conforme t (dias) e n (acidentes).

t/n	01	06	08	10	12	16	20	$E[N(t)]$
30	0.00841	0.17381	0.14348	0.05312	0.00882	0.00002	0.00000	6.02
40	0.00215	0.09566	0.15178	0.13164	0.06242	0.00229	0.00001	8.03
60	0.00015	0.01457	0.04469	0.09147	0.12496	0.06933	0.00763	12.04
80	0.00001	0.00159	0.00681	0.02147	0.04996	0.10867	0.07010	16.06
100	0.00000	0.00015	0.00080	0.00328	0.01052	0.05210	0.09744	20.07
120	0.00000	0.00001	0.00008	0.00041	0.00161	0.01384	0.05270	24.09
150	0.00000	0.00000	0.00000	0.00001	0.00007	0.00105	0.00817	30.11

É possível perceber que o valor médio das contagens, $E[N(t)]$, está próximo do número de acidentes onde a probabilidade é maior. Por exemplo, a probabilidade de termos 10 acidentes graves na BR 110/BA ao longo de 40 dias é de 0.13164, mas ter 8 acidentes graves para esse mesmo período a probabilidade sobre para 0.15178.

Uma exemplificação de um processo de renovação conforme discutido acima é apresentado na figura 16, com os valores de contagem média para os períodos descritos na tabela 10.

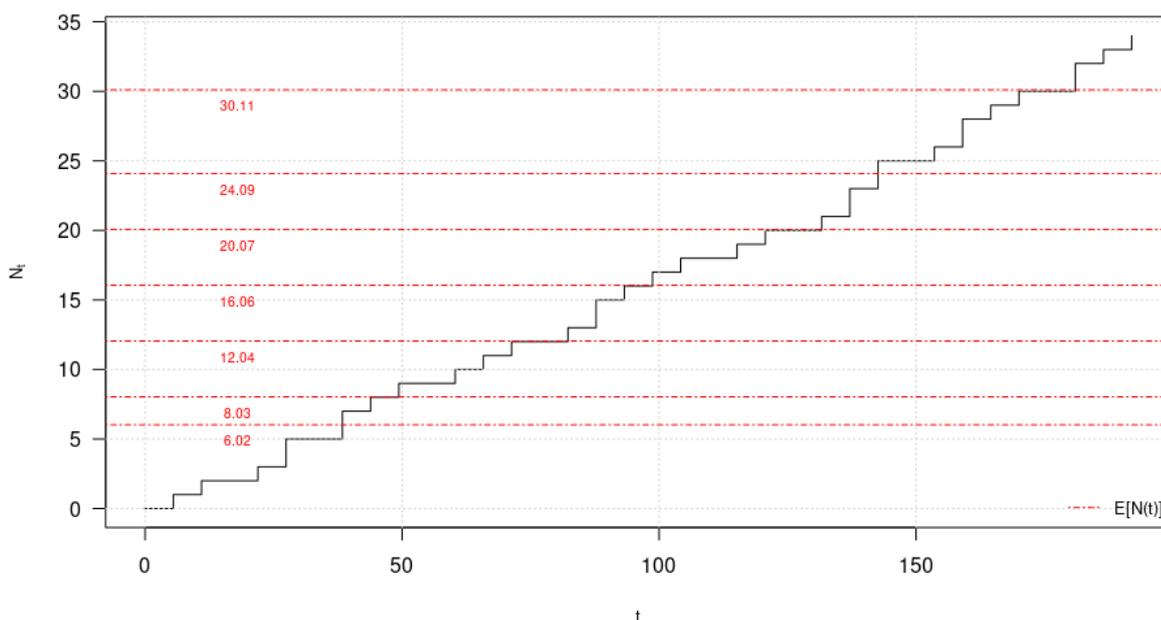


Figura 16 – Contagem do Processo de Renovação com Mistura.

Verifica-se de acordo com a figura 16 que para um período de 100 dias a contagem de acidentes graves na BR 1110/BA encontra-se em torno de 16.

5 Conclusão

Para o caso, muito comum, de modelar os dados dos tempos entre acidentes graves nas rodovias conforme uma simples distribuição Weibull (na maior parte das vezes particularizada pela distribuição exponencial), exemplificada aqui na BR 101/RJ, os resultados mostrados produziram informações bastante úteis sobre o comportamento dos acidentes nesse local, dando margem ainda para diversas outras análises considerando outras variáveis não utilizadas neste trabalho.

O modelo proposto para a mistura de distribuições Weibull, usado nos dados de tempos entre acidentes graves na BR 110/BA, mostrou uma melhoria quanto à descrição desses dados, mesmo que para esse caso em particular o teste de Kolmogorov-Smirnov não evidenciou uma grande diferença em relação ao modelo com apenas uma distribuição Weibull. Como proposta para essa situação, é possível testar uma estimação para a MMW em que haja mais um parâmetro para cada componente do modelo, um parâmetro de locação, e então verificar se a translação de qualquer um dos componentes da mistura resulte num modelo ainda mais bem ajustado.

Como proposta para trabalhos posteriores, é interessante investigar as causas da mistura entre as populações dos tempos entre acidentes, de acordo com as diversas variáveis presentes nos dados fornecidos pela Polícia Rodoviária Federal.

Referências

- [1] CARVALHO, T.M; *Modelagem da LGD Via Mistura de Distribuições Kumaraswamy*. 2015. Dissertação (Mestrado em Estatística) – Departamento de Estatística, Universidade de Brasília, Brasília.
- [2] COLOSIMO, E. A; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. São Paulo: Blucher, 2006. .
- [3] DEMPSTER, A. P. et al. *Maximum likelihood from incomplete data via the EM algorithm*. 1977. Journal of the royal statistical society. Series B (methodological), 1–38.
- [4] HOGG, R. V.; MCKEAN, J. W.; CRAIG, A. T. *Introduction to Mathematical Statistics*. Upper Saddle River, New Jersey: Prentice Hall, 2004.
- [5] MCLACHLAN, G.; PEEL, D. *Finite Mixture Models*. [S.1.]: John Wiley Sons, 2000.
- [6] PRF. *Dados Abertos da Polícia Rodoviária Federal*. Disponível em: <<https://www.prf.gov.br/dadosabertos/>>. Acesso em: 25 jan. 2018.
- [7] ROSS, S. M. *Introduction to Probability Models*. [S.1.]: Elsevier, 2010.
- [8] WEIBULL, W. *A Statistical Distribution Function of Wide Applicability*. J. Appl. Mech. 18, 293–296 (1951).

APÊNDICE A – Equação de Renovação para Mistura de Weibull

Para o caso da mistura de distribuições weibull, a equação de renovação é ainda mais complexa:

$$m(t) = F(t) + \int_0^t m(t-x)f(x)dx = p_1 \left[1 - e^{-\left(\frac{x}{\beta_1}\right)^{\alpha_1}} \right] + p_2 \left[1 - e^{-\left(\frac{x}{\beta_2}\right)^{\alpha_2}} \right] +$$

$$+ \int_0^t m(t-x) \left\{ p_1 \frac{\alpha_1}{\beta_1} \left(\frac{x}{\beta_1}\right)^{\alpha_1-1} e^{-\left(\frac{x}{\beta_1}\right)^{\alpha_1}} + p_2 \frac{\alpha_2}{\beta_2} \left(\frac{x}{\beta_2}\right)^{\alpha_2-1} e^{-\left(\frac{x}{\beta_2}\right)^{\alpha_2}} \right\} dx.$$

A dificuldade de se trabalhar com a equação acima justifica ainda mais o uso da teoria assintótica para estudar casos de processos de renovação mais complexos.

ANEXO A – Códigos Usados no R

Distribuição Weibull Simples – Processo de Poisson

```
##### ANÁLISE DESCRITIVA BR 116/BA - DISTRIBUIÇÃO POISSON #####
```

```
library(dplyr);
library(ggplot2);
library(lubridate);
library(fitdistrplus);
library(tseries)
```

```
load("../TCC/Análise dos Dados/db.RData")
db2 <- db[year(db$data_inversa) == 2016:2017,]
```

```
ba116 <- db2[db2$uf == "BA" & db2$br == 116,]
ba116 <- ba116[complete.cases(ba116),]
```

```
dados <- ba116 %>%
  group_by(data_inversa, horario, uf, br, km) %>%
  summarise(gravas=n_distinct(id))
dados$tempo <- as.POSIXlt(paste(dados$data_inversa,dados$horario),
                          "%Y-%m-%d %H:%M:%S")
dados$tempo <- sort(dados$tempo)
dados$dif <- vector(length = length(dados$tempo))
for(i in 2:length(dados$tempo)){
  dados$dif[i] <- round(difftime(dados$tempo[i], dados$tempo[i-1],
                                units = "days"),2)
}
```

```
summary(dados$dif)
length(dados$dif)
sd(dados$dif)
```

```
dados2 <- dados[-1,-7]
hist(dados2$dif, main=paste0("Tempo entre Ocorrências de Acidentes Graves
- BR 116/BA"),col="grey85",
```

```

    freq=F, xlab="", ylab="", cex.main=1.1, col.main="blue",
    ylim = c(0,0.55), col.main="black", border="grey55")
lines(density(dados2$dif), col="red")
fit <- fitdist(dados2$dif[dados2$dif!=0], "weibull")
ks.test(dados2$dif, "pexp", rate=fit2$estimate[1])
x <- seq(min(dados2$dif),max(dados2$dif),0.001)
lines(x, y=dweibull(x, shape=fit$estimate[1], scale=fit$estimate[2]),
      col="blue")
# lines(x, y=dexp(x, rate=fit2$estimate[1]), col="green")

legend('topright', legend=paste0("Exponencial: ",names(fit2$estimate),
                                " = ",round(fit2$estimate,2)),
      lty=, bty='n', cex=1, text.col="blue")

fit2 <- fitdist(dados2$dif, "exp")
plot(fit2)

# Probabilidade para N(t):

Nt <- function(t, n){
  exp(-0.62*t)*(0.62*t)^n/factorial(n)
}
Nt(1,2)

m <- matrix(0,7,7,byrow = F)
n <- c(0,1,2,3,4,5,6)
for(i in 1:7){
  for(j in 1:7){
    m[i,j] <- round(Nt(i,n[j]),5)
  }
}
m

# Esperança para o n-ésimo acidente

E <- n/0.62
n <- seq(0,100,1)
plot(E ~n)

```

```

#ACIDENTE COM MOTOS
table(db$tipo_veiculo)
db2 <- db[year(db$data_inversa) == 2016:2017,]
db2$tipo_veiculo <- trimws(db2$tipo_veiculo)
motos <- c("Motocicleta","Motocicletas","Motoneta","Ciclomotor")
motos <- db2[db2$tipo_veiculo %in% motos,]
vetor <- motos$id[!duplicated(motos$id)]
motos <- db2[db2$id %in% vetor,]
nmotos <- db2[!(db2$id %in% vetor),]

ba116 <- motos[motos$uf == "BA" & motos$br == 116,]
ba116 <- ba116[complete.cases(ba116),]

dados <- ba116 %>%
  group_by(data_inversa, horario, uf, br, km) %>%
  summarise(grades=n_distinct(id))
dados$tempo <- as.POSIXlt(paste(dados$data_inversa,dados$horario),
  "%Y-%m-%d %H:%M:%S")
dados$tempo <- sort(dados$tempo)
dados$dif <- vector(length = length(dados$tempo))
for(i in 2:length(dados$tempo)){
  dados$dif[i] <- round(difftime(dados$tempo[i], dados$tempo[i-1],
    units = "days"),2)
}

summary(dados$dif)
length(dados$dif)
sd(dados$dif)

dados2 <- dados[-1,-7]
hist(dados2$dif, main=paste0("Tempo entre Ocorrências de Acidentes Graves com Motos
  - BR 116/BA"),col="grey85",
  freq=F, xlab="", ylab="", cex.main=1.1, col.main="blue",
  ylim = c(0,0.14), col.main="black", border="grey55")
lines(density(dados2$dif), col="red")
fit <- fitdist(dados2$dif[dados2$dif!=0], "weibull")
ks.test(dados2$dif, "pexp", rate=fit2$estimate[1])
x <- seq(min(dados2$dif),max(dados2$dif),0.001)

```

```

lines(x, y=dweibull(x, shape=fit$estimate[1], scale=fit$estimate[2]), col="blue")
# lines(x, y=dexp(x, rate=fit2$estimate[1]), col="green")
fit
legend('topright', legend=paste0("Exponencial: ",names(fit2$estimate),
                                " = ",round(fit2$estimate,2)),
      lty=, bty='n', cex=1, text.col="blue")

fit2 <- fitdist(dados2$dif, "exp")

### Acidentes sem Motos

ba116 <- nmotos[nmotos$uf == "BA" & nmotos$br == 116,]
ba116 <- ba116[complete.cases(ba116),]

dados <- ba116 %>%
  group_by(data_inversa, horario, uf, br, km) %>%
  summarise(grades=n_distinct(id))
dados$tempo <- as.POSIXlt(paste(dados$data_inversa,dados$horario),
                        "%Y-%m-%d %H:%M:%S")
dados$tempo <- sort(dados$tempo)
dados$dif <- vector(length = length(dados$tempo))
for(i in 2:length(dados$tempo)){
  dados$dif[i] <- round(difftime(dados$tempo[i], dados$tempo[i-1],
                                units = "days"),2)
}

summary(dados$dif)
length(dados$dif)
sd(dados$dif)

dados2 <- dados[-1,-7]
hist(dados2$dif, main=paste0("Tempo entre Ocorrências de Acidentes
                            Graves sem Motos - BR 116/BA"),col="grey85",
     freq=F, xlab="", ylab="", cex.main=1.1, col.main="blue",
     ylim = c(0,0.40), col.main="black", border="grey55")
lines(density(dados2$dif), col="red")
fit <- fitdist(dados2$dif[dados2$dif!=0], "weibull")
ks.test(dados2$dif, "pexp", rate=fit2$estimate[1])

```

```

x <- seq(min(dados2$dif),max(dados2$dif),0.001)
lines(x, y=dweibull(x, shape=fit$estimate[1], scale=fit$estimate[2]), col="blue")
# lines(x, y=dexp(x, rate=fit2$estimate[1]), col="green")
fit
legend('topright', legend=paste0("Exponencial: ",names(fit2$estimate),
                                " = ",round(fit2$estimate,2)),
      lty=, bty='n', cex=1, text.col="blue")

fit2 <- fitdist(dados2$dif, "exp")

```

Grafico probabilidades:

```

t <- seq(0,10, 0.0001)
f <- exp(-0.62*t)
f1 <- exp(-0.21*t)
f2 <- exp(-0.41*t)
plot(f ~ t, type="l", main="Probabilidade de dois Eventos Ocorrerem
      entre t Dias", ylab="Probabilidade", col="1")
lines(f1 ~ t, col="2")
lines(f2 ~ t, col="3")
legend('topright', legend=c("Total", "Com Motos", "Sem Motos"),
      bty='n', cex=1, col= 1:3, lty = 1)

```

Mistura de Distribuições Weibull – Processo de Renovação

```

## Criação da Base de Dados
library(dplyr);
library(ggplot2);
library(lubridate);
library(fitdistrplus);

load("../TCC/Análise dos Dados/db.RData")
db2 <- db[year(db$data_inversa) == 2016:2017,]
ba324 <- db2[db2$uf == "BA" & db2$br == 110,] # BA/110 e BA/324
ba324 <- ba324[complete.cases(ba324),]
dados <- ba324 %>%
  group_by(data_inversa, horario, uf, br, km) %>%
  summarise(grades=n_distinct(id))
dados$tempo <-as.POSIXlt(paste(dados$data_inversa,dados$horario),

```

```

                                "%Y-%m-%d %H:%M:%S")
dados$tempo <- sort(dados$tempo)
dados$dif <- vector(length = length(dados$tempo))
for(i in 2:length(dados$tempo)){
  dados$dif[i] <- round(difftime(dados$tempo[i], dados$tempo[i-1]),
                        units = "days"),2)
}
dados2 <- dados[-1,-7]

#### ALGORITMO EM PARA MISTURA DE WEIBULL ####

x <- dados2$dif
res <- matrix(rep(NA, 1*7), 1, 7)
set.seed(1182)
## Mistura de duas distribuições Weibull (função de densidade)

mweib = function(x){
  p[1]*((a[1]/b[1])*((x/b[1])^(a[1]-1))*exp(-(x/b[1])^a[1])) +
  p[2]*((a[2]/b[2])*((x/b[2])^(a[2]-1))*exp(-(x/b[2])^a[2]))
}

## Estimação dos Valores iniciais
erro <- Inf
iter <- 0
a <- c(1,1)
b <- c(4,3)
p <- c(0.5,0.5)

while(erro > 0.000001){
## Passo E (Expressão decorrente da Regra de Bayes)
fpost = function(x,j){
  fp1 = p[j]*(a[j]/b[j])*((x/b[j])^(a[j]-1))*exp(-(x/b[j])^a[j])
  fp2 = mweib(x)
  fp = fp1/fp2
  return(fp)
}
## Maximizar Proporção
for(j in 1:2){
  pi[j]=mean(fpost(x,j))
}

```

```
}

veros <- function(param){
  a = c(param[1], param[2])
  b = c(param[3], param[4])
  if((a[1]>0) && (a[2] > 0) && (b[1]>0) && (b[2]>0))
    return((-1)*sum(log((a[1]/b[1])*((x/b[1])^(a[1]-1))
                      *exp(-(x/b[1])^a[1])*pi[1] +
                      (a[2]/b[2])*((x/b[2])^(a[2]-1))
                      *exp(-(x/b[2])^a[2])*pi[2])))
  else return(-Inf)
}

vero_ant <- veros(c(a, b))

## Valor Inicial
vi <- c(a, b)

## Estimação com o Optim [Nelder-Mead]
estima <- optim(vi, veros, NULL, hessian=FALSE)

a <- c(estima$par[1], estima$par[2])
b <- c(estima$par[3], estima$par[4])
vero_atu <- veros(c(a, b))
erro=abs(vero_atu-vero_ant)/vero_ant
erro=abs(erro)
iter = iter + 1
res[1,] <- c(iter, a, b, pi)
}

round(res[1,], 5)

## Gráfico Comparativo com As Estimativas Unimodais
par(mfrow=c(1,1))

# antes
hist(x, main="", col="grey85", freq=F, xlab="Dias", ylab="Densidade",
     cex.main=1.1, col.main="blue",
     col.main="black", border="grey55", cex.main=0.9, ylim=c(0,0.18))
```

```

lines(density(x), col="red")
fit <- fitdist(x[x!=0], "weibull")
k <- seq(min(x),max(x),0.00001)
lines(k, y=dweibull(k, shape=fit$estimate[1], scale=fit$estimate[2]),
      col="blue")

## depois
hist(x, main="",col="grey85",
     freq=F, xlab="Dias", ylab="Densidade", cex.main=1.1, col.main="blue",
     col.main="black", border="grey55", cex.main=0.9, ylim=c(0,0.18))
lines(density(x), col="red")
lines(k,res[1,6]*dweibull(k,shape = res[1,2],scale = res[1,4])+res[1,7]
      *dweibull(k,shape = res[1,3],
                scale = res[1,5]), col="blue")

# Acumulada
f <- function(j){
  res[1,6]*dweibull(j,shape = res[1,2],scale =res[1,4])+res[1,7]
  *dweibull(j,shape = res[1,3],scale = res[1,5])
}

Fx <- function(x){
  res[1,6]*(1-exp(-(sort(x)/res[1,4])^res[1,2]))+res[1,7]
  *(1-exp(-(sort(x)/res[1,5])^res[1,3]))
}

Fx2 <- function(x){
  1-exp(-(sort(x)/fit$estimate[2])^fit$estimate[1])
}

par(mfrow=c(1,2))
acum <- ecdf(x)
plot(acum, col="grey", ylab="F(x)", main="Weibull Simples",
     cex.main=0.9, xlab="")
curve(Fx2, add=TRUE, col="red", lwd=2)
legend("bottomright", legend=c("Acum. Empírica", "Acum. Estimada"),
     lty=c(6,1), bty='n',
     cex=1.1, col=c("grey","red"), lwd = c(2.5,2))

```

```
plot(acum, col="grey", ylab="F(x)", main="Mistura Weibull",
     cex.main=0.9, xlab="")
curve(Fx, add=TRUE, col="blue", lwd=2)
legend("bottomright", legend=c("Acum. Empírica", "Acum. Estimada"),
      lty=c(6,1), bty='n',
      cex=1.1, col=c("grey","blue"), lwd = c(2.5,2))

## qqplot
set.seed(1182)
par(mfrow=c(1,2))
dist = rweibull(10000, shape=fit$estimate[1], scale = fit$estimate[2])
qqplot(dist, x, main="Weibull Simples", xlab="",cex.main=0.9, ylab="")
qqline(x, distribution = function(p) qweibull(p,shape=fit$estimate[1],
                                             scale = fit$estimate[2]),
      col="red")

dist = rweibull(5000, shape=res[1,2], scale = res[1,4])+
  rweibull(5000, shape=res[1,3], scale = res[1,5])
qqplot(dist, x, main="Mistura Weibull", xlab="",cex.main=0.9, ylab="")
qqline(x, distribution = function(p) qweibull(p,shape=res[1,2],
                                             scale = res[1,4])+
      qweibull(p,shape=res[1,4], scale = res[1,5]), col="blue")

## KS Teste
y1 <- sort(acum(x))
y2 <- Fx(x)
y3 <- Fx2(x)
ks.test(y1,y2)
ks.test(y1,y3)
```