Análise de sobrevivência e confiabilidade

Modelos paramétrica

Prof. Paulo Cerqueira Jr
Programa de Pós-Graduação em Matemática e Estatística - PPGME
Instituto de Ciências Exatas e Naturais - ICEN

https://github.com/paulocerqueirajr

Introdução

Introdução

- Agora iremos estudar modelo de probabilidade para dados de sobrevivência.
- Dessa forma, faremos suposições de distribuições de probabilidade para os tempos de falha ou evento.
- Estas distribuições são bastante utilizadas, principalmente para produtos industriais, por se mostrarem adequadas para descrever estes tempos de vida.
- Os modelos paramétricos vêm sendo utilizados com mais frequência na área industrial do que na médica.
- A principal razão deste fato é que os estudos envolvendo componentes e equipamentos industriais podem ser planejados e consequentemente as fontes de perturbação (heterogeneidade) podem ser controladas.
- Nestas condições a busca por um modelo paramétrico adequado fica facilitada e a análise estatística dos dados fica mais precisa.

Distribuições para o tempo de sobrevivência

Distribuições para o tempo de sobrevivência

Distribuições para o tempo de sobrevivência

> As distribuições de probabilidade:

- Exponencial
- Weibull
- Lognormal
- Gamma
- Algumas distribuições mais sofisticadas:
  - Gama Generalizada
  - Exponencial por partes;
  - Distribuições gama-g;
  - Estáveis positivas.

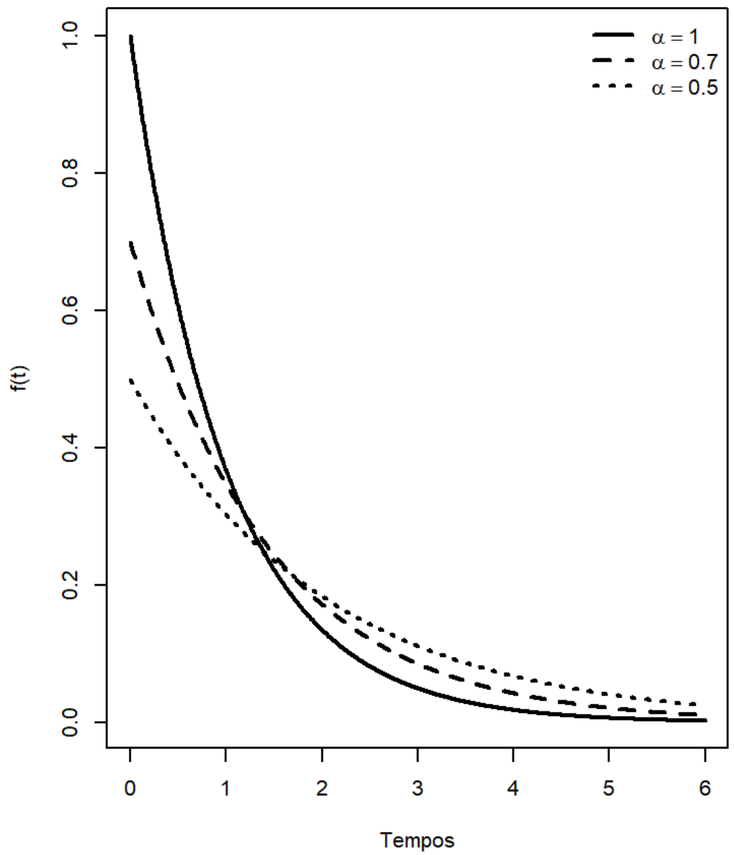
Distribuição exponencial

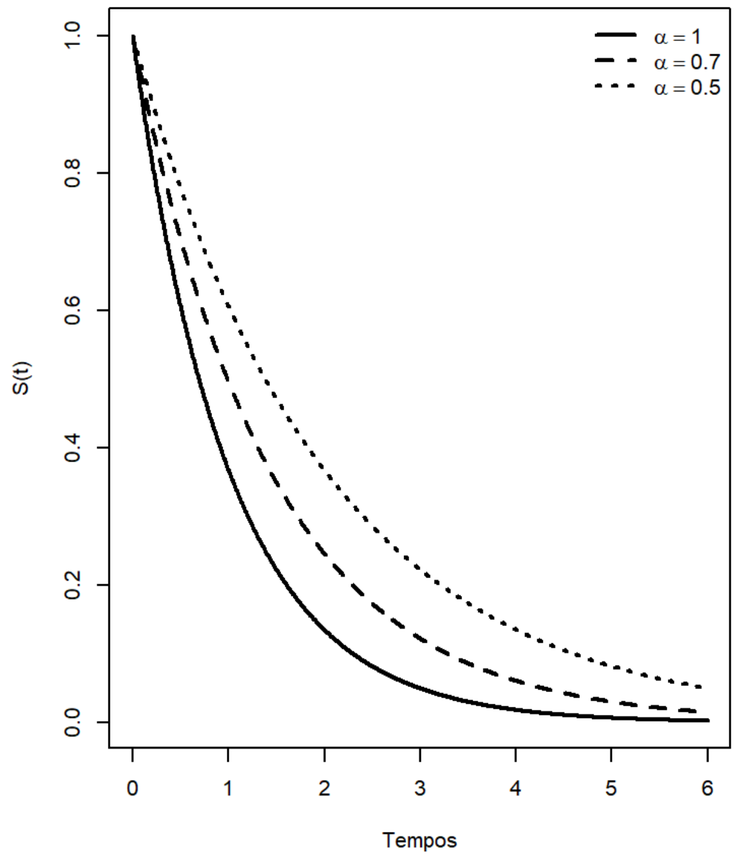
Distribuições para o tempo de sobrevivência

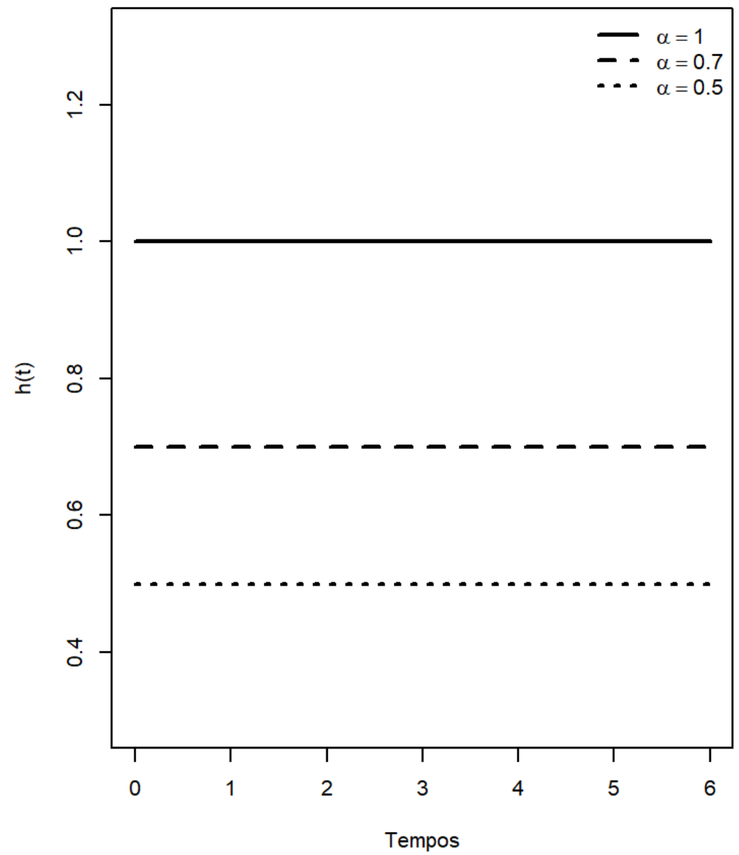
Exponencial

Distribuições para o tempo de sobrevivência

Exponencial





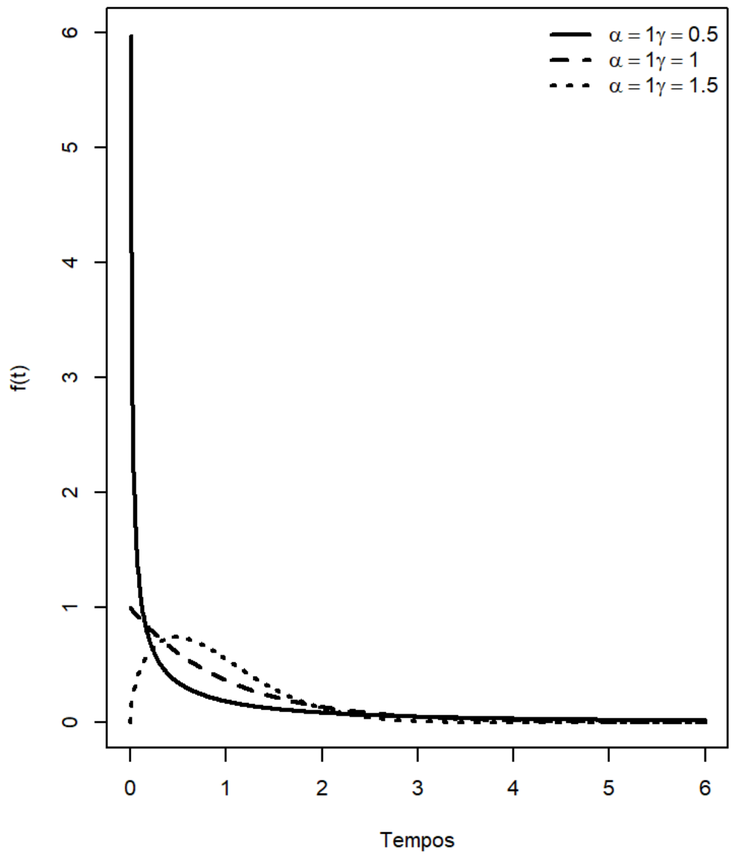


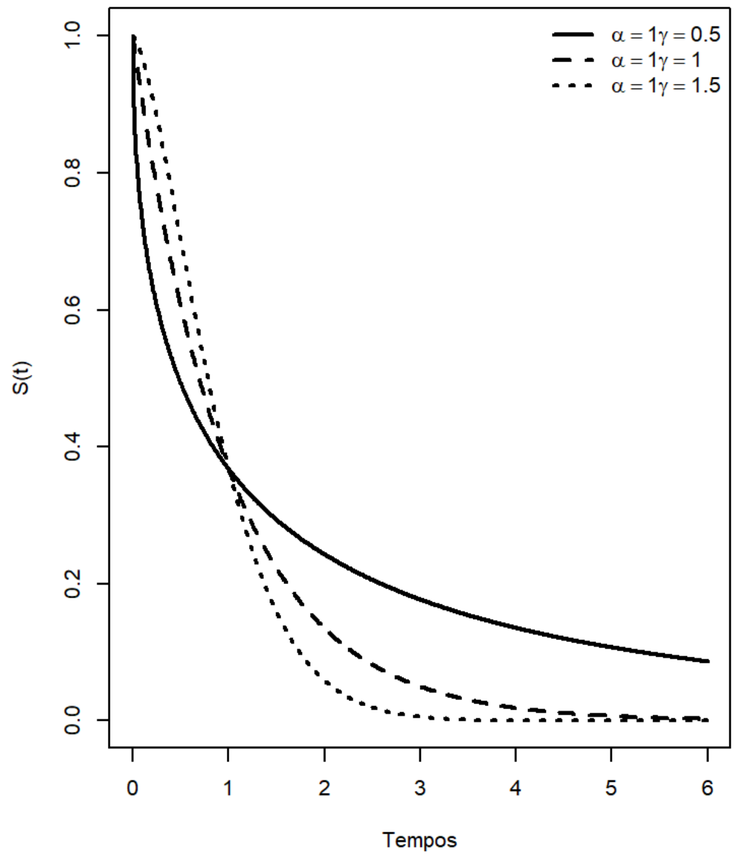
Distribuições para o tempo de sobrevivência

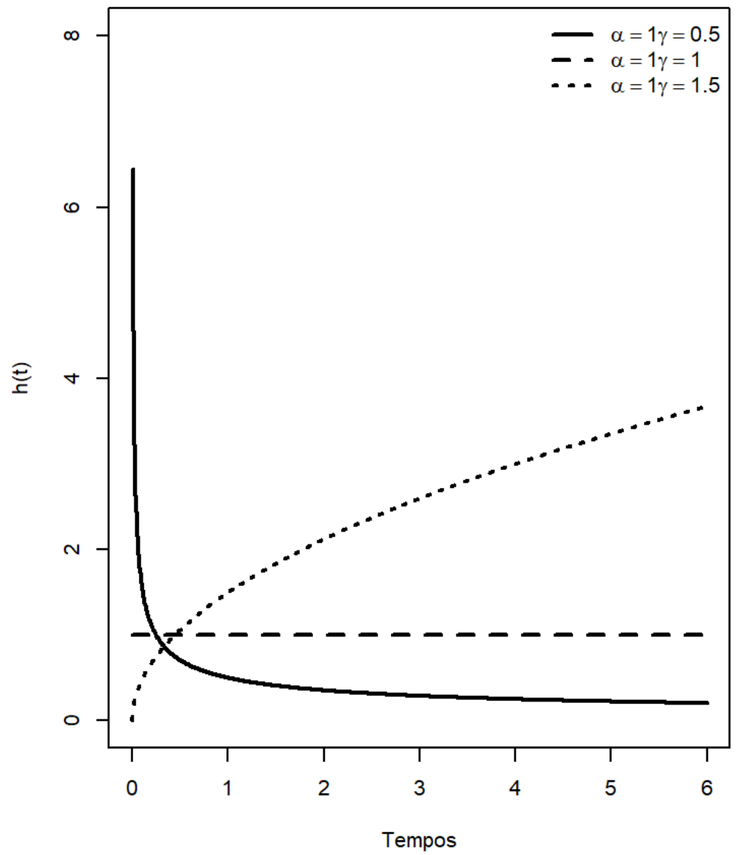
Weibull

Distribuições para o tempo de sobrevivência

Weibull







Distribuições para o tempo de sobrevivência

Lognormal

- O modelo log-normal é, juntamente com o Weibull, um modelo importante em análise de sobrevivência.
- Este modelo apresenta taxas de falhas não monótonas.

Função densidade:

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left(\log(t)-\mu\right)^2\right\}, \quad t \geq 0.$$

Função de sobrevivência:

$$S(t) = \Phi\left\{\frac{-\log(t)+\mu}{\sigma}\right\}, \quad t \geq 0.$$

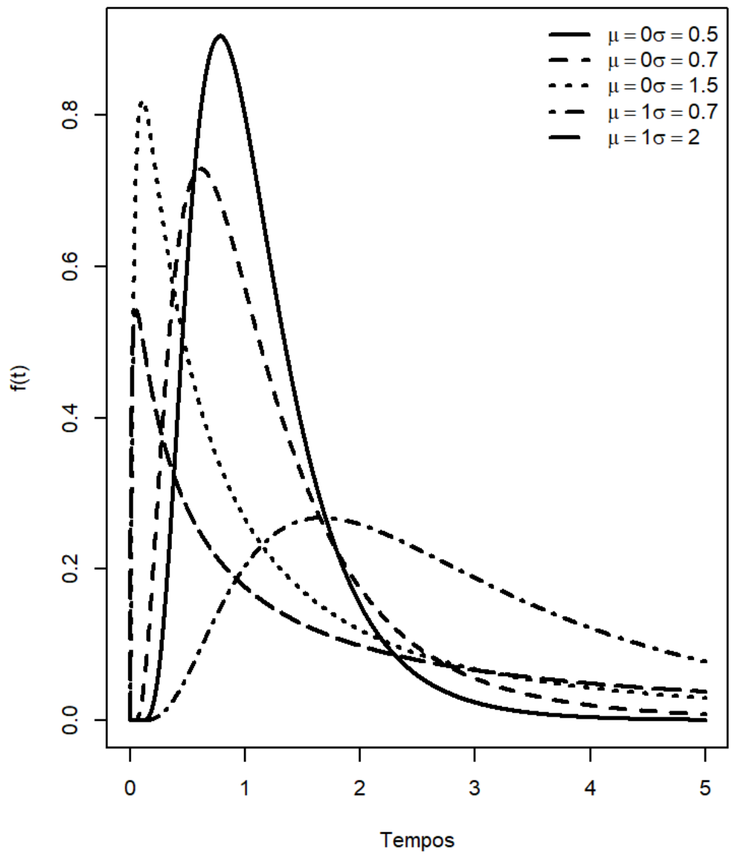
Função taxa de falha:

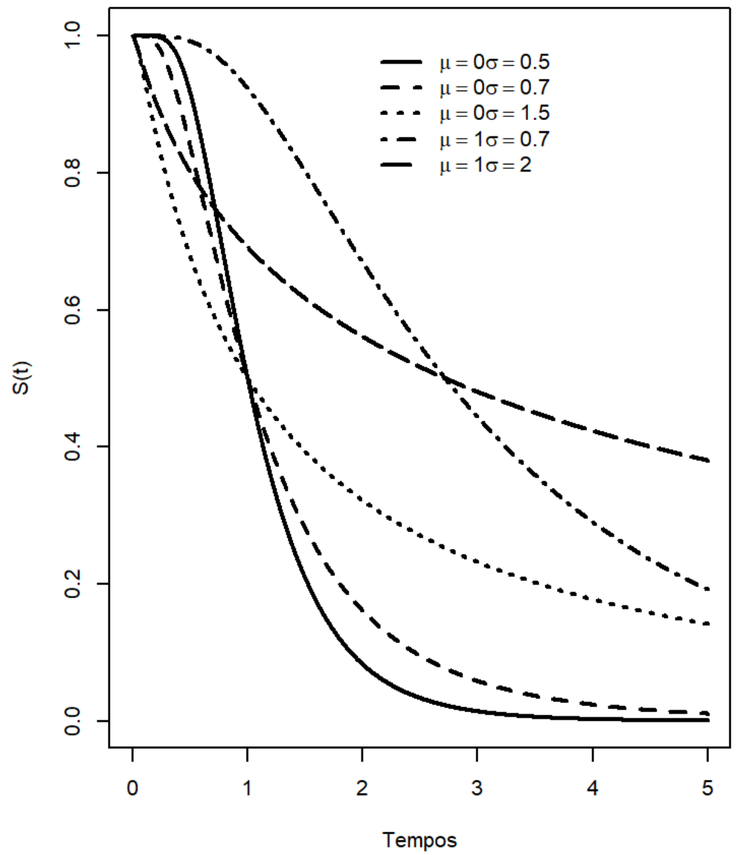
$$h(t) = \frac{f(t)}{S(t)}$$

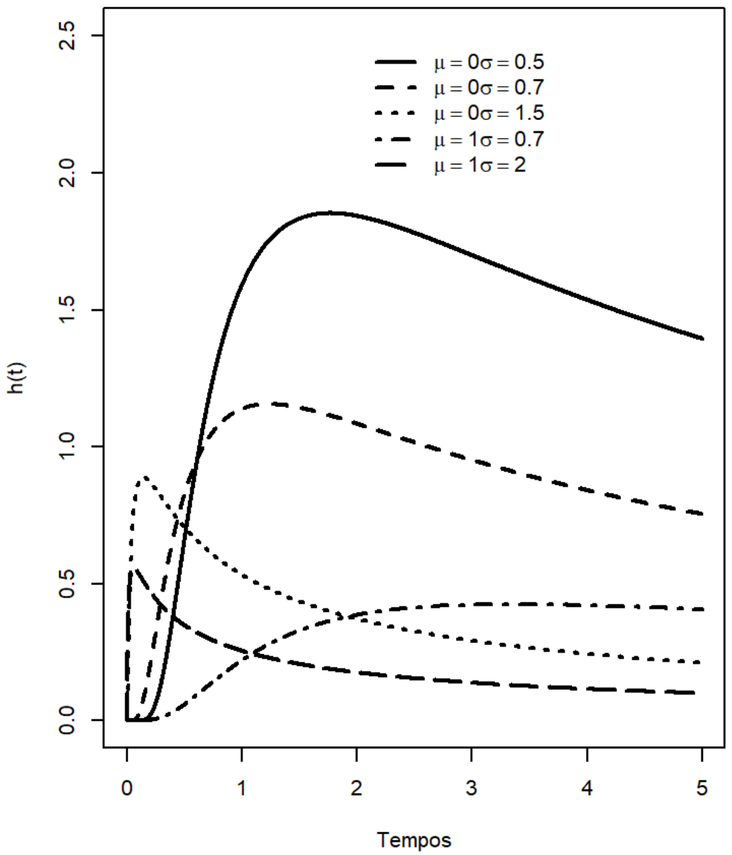
.

Distribuições para o tempo de sobrevivência

Lognormal







Distribuições para o tempo de sobrevivência

Lognormal

- Se T tem distribuição log-normal, então $Y = \log T$ tem distribuição normal ou Gaussiana.
- Observe que o modelo log-normal é definido em termos dos parâmetros da distribuição geradora normal. Ou seja,
  - μ (parâmetro de escala da log-normal) é o parâmetro de locação da distribuição normal.
  - σ (parâmetro de forma da log-normal) é o parâmetro de escala da distribuição normal.

Distribuições para o tempo de sobrevivência

Gama

- O gama é outro modelo importante em análise de sobrevivência.
- Mostramos a seguir as formas de $f(t)$, $S(t)$ e $h(t)$ para o modelo gama.

Função densidade:

$$f(t) = \frac{1}{\Gamma(k)\alpha^k} t^{k-1} \exp\left\{-\frac{t}{\alpha}\right\}, \quad t > 0.$$

Função de sobrevivência:

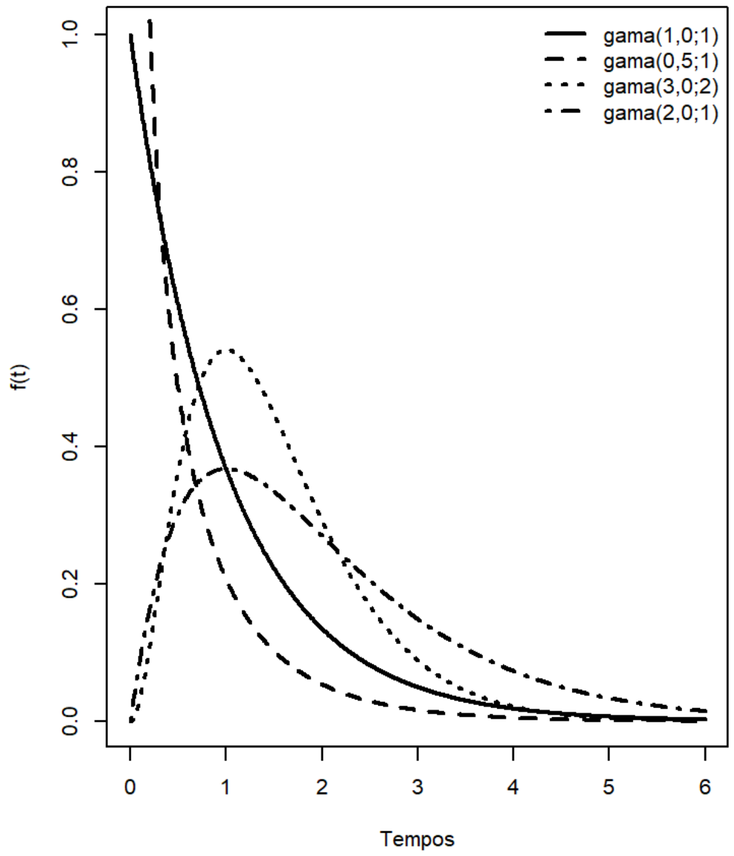
$$S(t) = \int_t^\infty \frac{1}{\Gamma(k)\alpha^k} t^{k-1} \exp\left\{-\frac{t}{\alpha}\right\}, \quad t > 0.$$

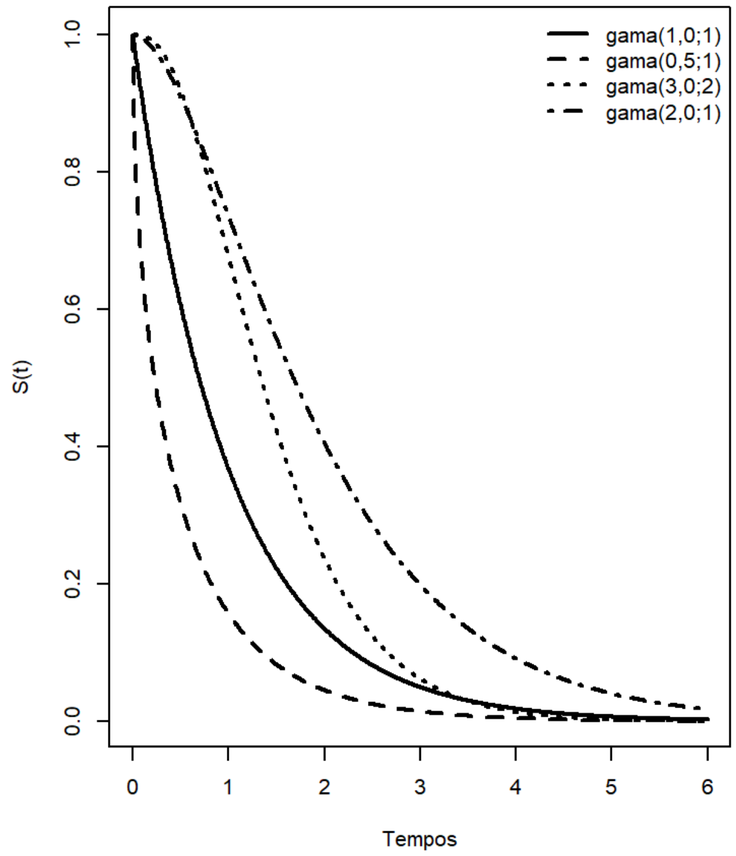
Função taxa de falha:

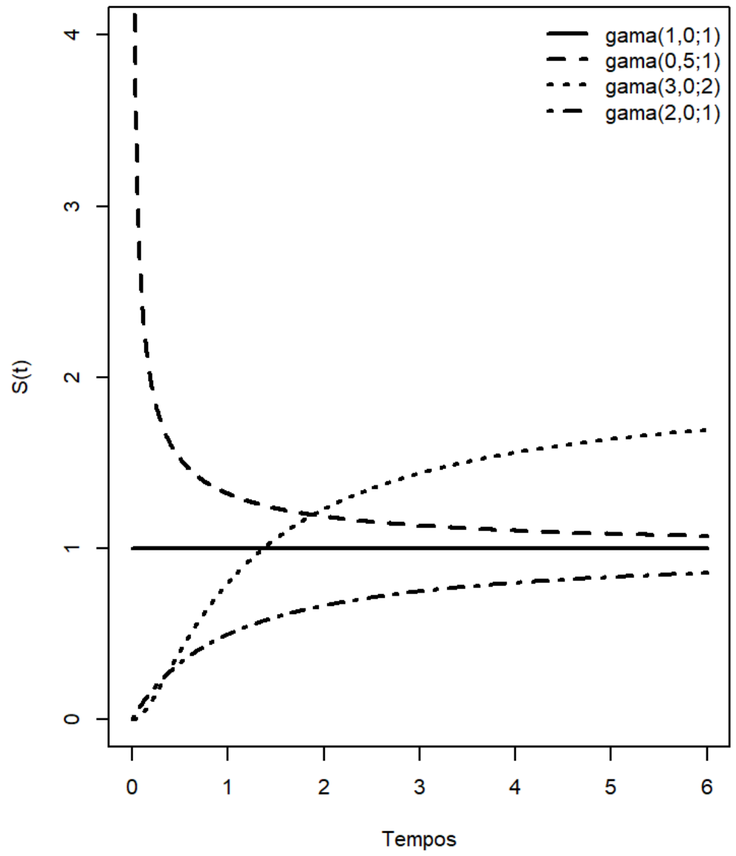
$$h(t) = \frac{f(t)}{S(t)}, \quad t > 0.$$

Distribuições para o tempo de sobrevivência

Gama







Distribuições para o tempo de sobrevivência

Gama generalizado

- O modelo gama generalizado tem um parâmetro de escala e dois de forma.

Função densidade:

$$f(t) = \frac{\gamma}{\Gamma(k)\alpha^{\gamma k}} t^{\gamma k - 1} \exp\left\{ -\left(\frac{t}{\alpha}\right)^{2} \right\}, \quad t > 0.$$

- Os principais modelos em análise de sobrevivência são casos particulares da gama generalizada:
  - para $k = 1$ e $\gamma = 1$ tem-se $T \sim Exp(\alpha)$.
  - para $k = 1$ tem-se $T \sim Weibull(\gamma, \alpha)$.
  - para $\gamma = 1$ tem-se $T \sim Gama(k, \alpha)$.
  - para $k \to \infty$ tem-se $T \sim log - normal$.

Distribuições para o tempo de sobrevivência

Gama generalizado

- O modelo gama generalizado é complexo mas útil na seleção de modelos.
- Os parâmetros, ou uma função deles, não tem interpretação.
- Difícil de ajustar computacionalmente. É comum obtermos falta de convergência.
- O `R` ajusta o modelo gama generalizado no pacote `flexsurv`.

Distribuições para o tempo de sobrevivência

Loglogística

- A distribuição log-logística fornece mais um ajuste paramétrico em análise de sobrevivência.
- A mesma tem várias forma de acordo com o parâmetro de forma beta β.
- Permitindo várias formas da função de taxa de falha.

A função de sobrevivência é dada por

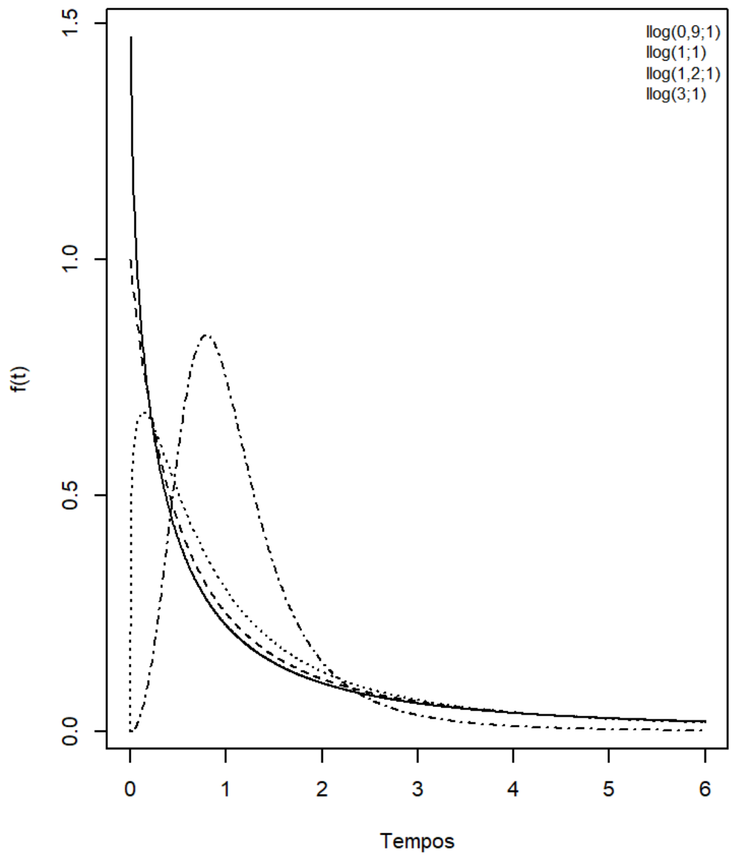
$$S(t) = 1 - F(t) = [1 + (t/\alpha)^{\beta}]^{-1},$$

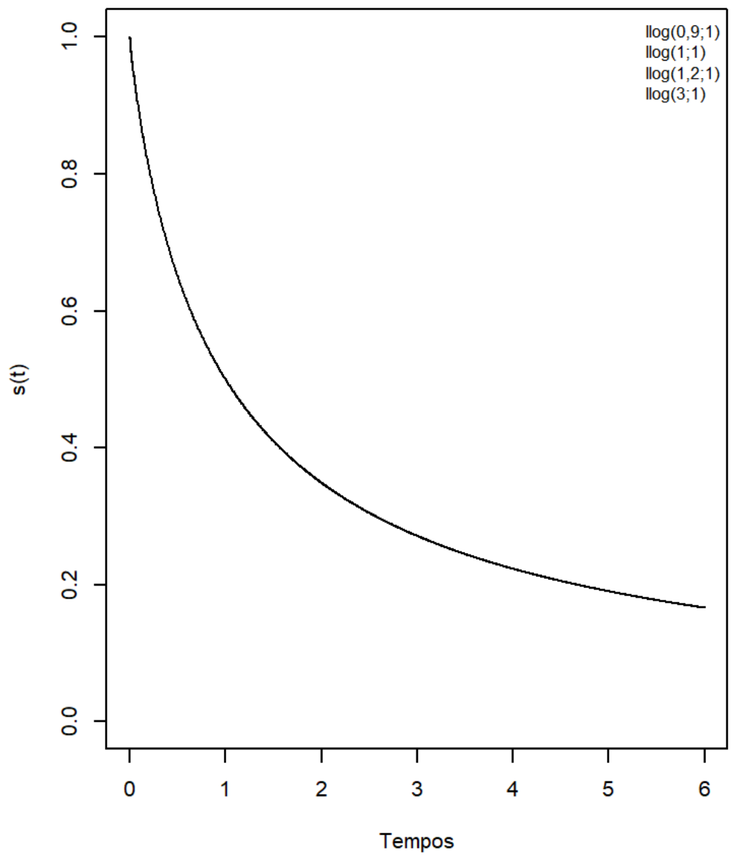
A função taxa de falha

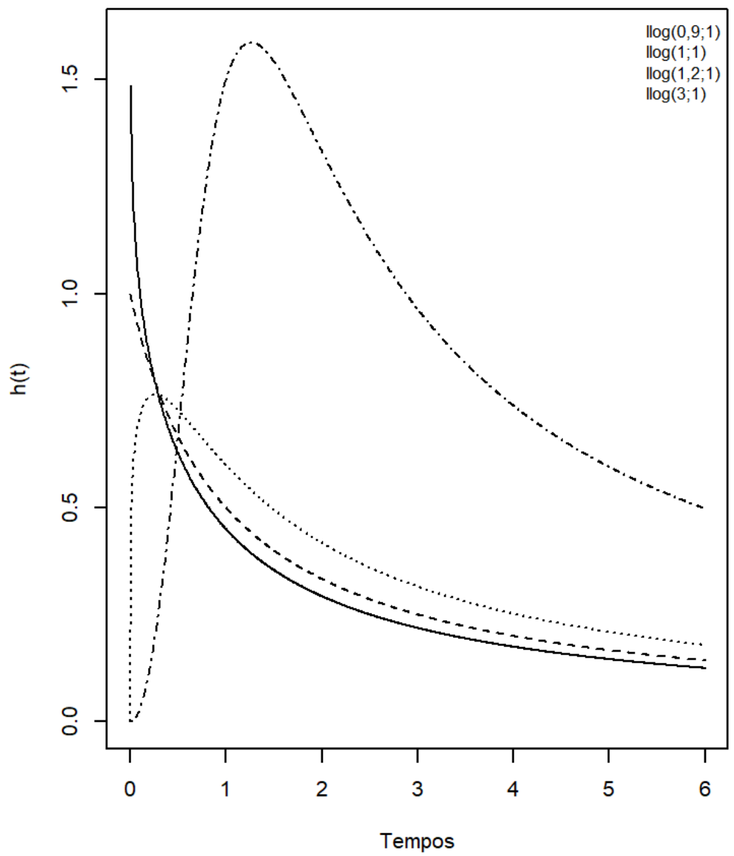
$$h(t) = \frac{f(t)}{S(t)} = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{1 + (t/\alpha)^{\beta}}.$$

Distribuições para o tempo de sobrevivência

Loglogística







Distribuições para o tempo de sobrevivência

Exponencial por partes (EP)

- Segundo Ibrahim (2001), o EP é um dos modelos mais populares na modelagem semiparamétrica de dados de sobrevivência.
- Apesar de ser paramétrico em um senso estrito, não impõe restrições quanto a forma da função taxa de falha, diferentemente de outros modelos paramétricos como: exponencial, Weibull e lognormal, entre outros.
- EP é construído com base em uma aproximação da função taxa de falha por segmentos de retas, cujos comprimentos são determinados por uma grade de tempos τ que divide o eixo dos tempos em um número finito de intervalos.
- Matematicamente, a grade de tempos é definida como $\tau = \{s_0, s_1, \ldots, s_b\}$, em que $0 = s_0 < s_1 < \ldots < s_b = \infty$, que induz intervalos da forma $I_j = (s_{j-1}, s_j]$, para $j = 1, 2, \ldots, b$.

Distribuições para o tempo de sobrevivência

Exponencial por partes (EP)

A função risco é definida da seguinte forma:

$$h(t|\lambda,\tau) = \lambda_j, \ \text{ se } \ t \ \in I_j, \lambda_j > 0 \quad j = 1, \ldots, b,$$

em que $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_b)^\top$.

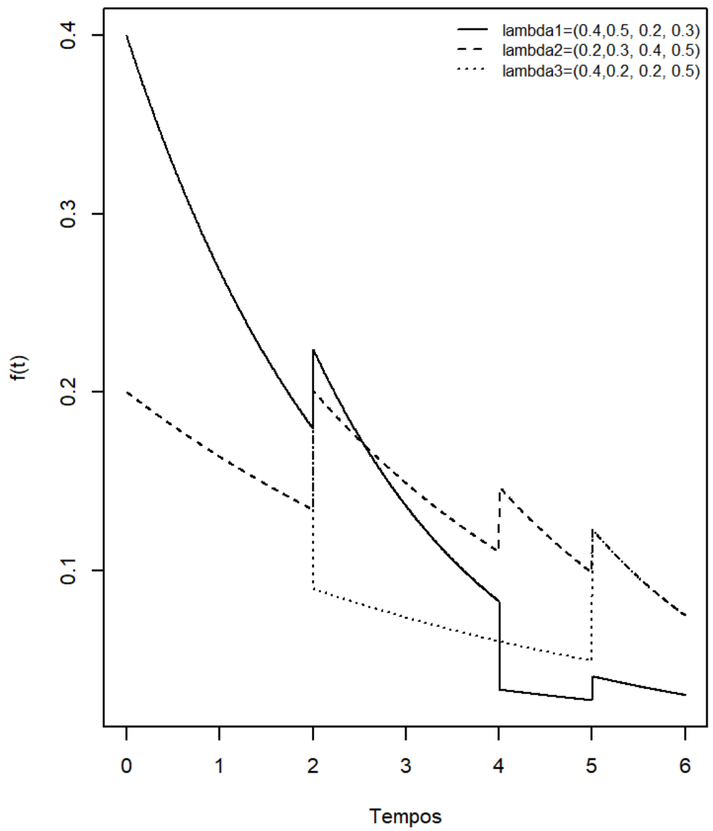
- A função taxa de falha acumulada e a função de sobrevivência são dadas por

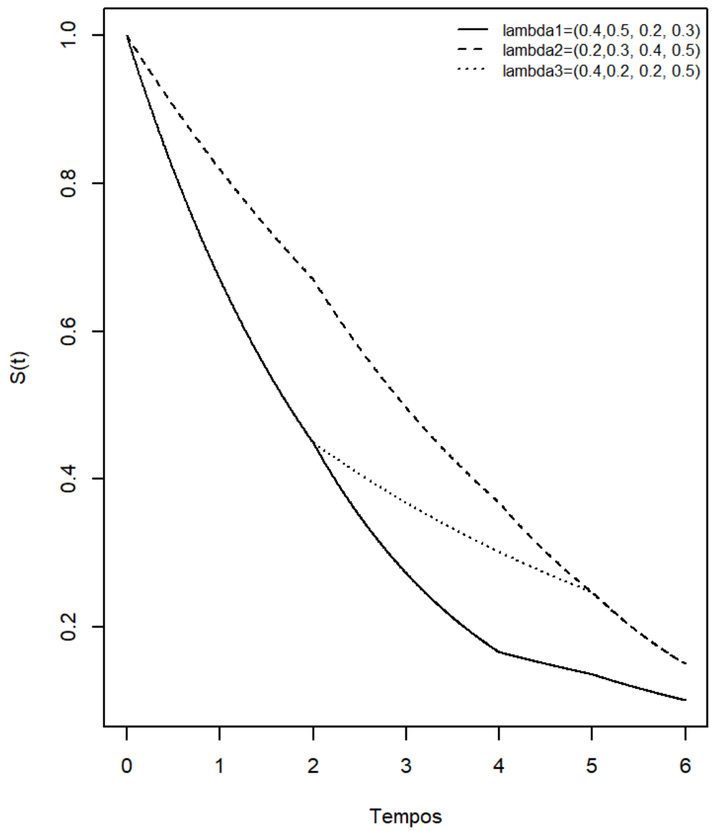
$$H(t|\lambda,\tau) = \lambda_j(t - s_{j-1}) + \sum_{g=1}^{j-1} \lambda_g (s_g - s_{g-1}) \quad \text{e} \quad S(t|\lambda,\tau) = \exp\left\{ -\lambda_j(t - s_{j-1}) + \sum_{g=1}^{j-1} \lambda_g (s_g - s_{g-1}) \right\}$$

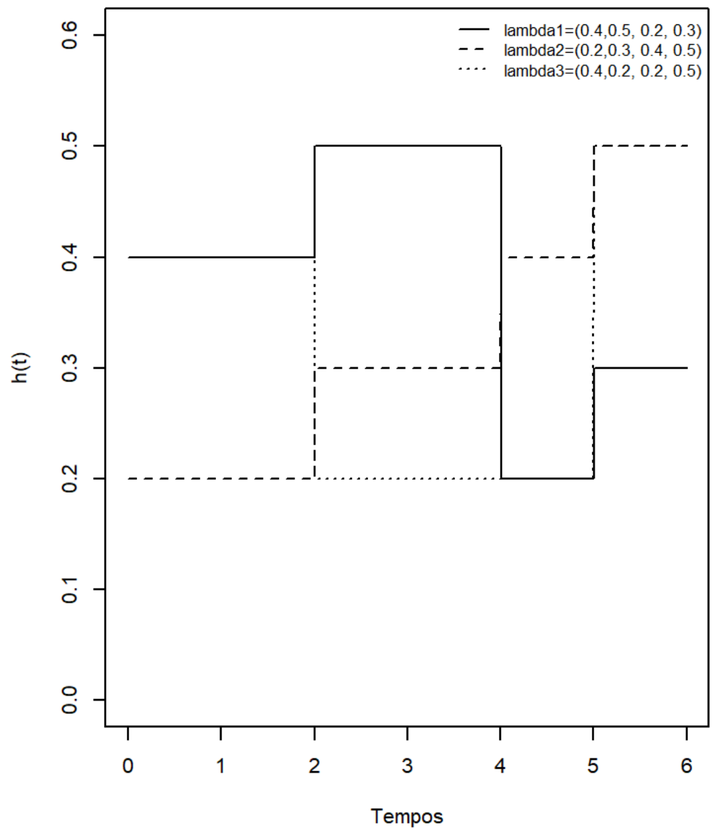
Distribuições para o tempo de sobrevivência

Exponencial por partes (EP)

$$\tau = \{0, 2, 4, 5, \text{inf}\}$$







Estimação dos parâmetros

Estimação dos parâmetros

Introdução

- Se corretamente especificado, os modelos paramétricos são bastante eficientes.
- Inferência para as quantidades desconhecidas dos modelos é baseada na função de verossimilhança e suas propriedades assintóticas.
- Cuidado na incorporação de censuras na função de verossimilhança.
- Má especificação de um modelo paramétrico acarreta em vício na estimação das quantidades de interesse.
- Técnicas de adequação, via resíduos, são fundamentais para verificar a adequação dos modelos paramétricos.

Estimação dos parâmetros

- Sejam,
  - T_i: tempo de falha do $i-$ésimo indivíduo com $f(\cdot)$ e $S(\cdot)$.
  - C_i: tempo de censura do $i-$ésimo indivíduo com $g(\cdot)$ e $G(\cdot)$.
- O tempo observado e a variável indicadora de falha,

$$y_i = \min(T_i, C_i), \delta_i = \begin{cases} 1 & , T_i < C_i \quad \text{tempo de falha,} \\ 0 & , T_i > C_i \quad \text{tempo de censura.} \end{cases}$$

- Supondo que o mecanismo de censura é não informativo, ou seja, T e C são independentes.

Estimação dos parâmetros

A construção da função de verossimilhança segue os seguintes passos:

- O i-ésimo indivíduo é uma censura:

$$P(y_i = t, \delta_i = 0) = P(C_i = t, T_i > C_i) = P(C_i = t, T_i > t) \overset{\text{ind}}{=} g(t)S(t)$$

- O i-ésimo indivíduo é um evento:

$$P(y_i = t, \delta_i = 1) = P(T_i = t, T_i < C_i) = P(T_i = t, C_i > t) \overset{\text{ind}}{=} f(t)G(t)$$

Dessa forma a função de verossimilhança é dada por

$$L(\theta, \nu) = \prod_{i=1}^{n} \left[f(y_i \mid \theta)G(y_i \mid \nu)\right]^{\delta_i} \times \left[g(y_i \mid \nu)S(y_i \mid \theta)\right]^{1-\delta_i},$$

θ e ν são os vetores de parâmetros da distribuição dos tempos de evento e censura, respectivamente.

Estimação dos parâmetros

Usando a suposição de censura não-informativa,

$$\begin{aligned} L(\theta, \nu) &= \prod_{i=1}^{n} \left[f(y_i \mid \theta) G(y_i \mid \nu) \right]^{\delta_i} \times \left[g(y_i \mid \nu) S(y_i \mid \theta) \right]^{1-\delta_i}, \\ &= \prod_{i=1}^{n} f(y_i \mid \theta)^{\delta_i} S(y_i \mid \theta)^{1-\delta_i} \times \prod_{i=1}^{n} g(y_i \mid \nu)^{1-\delta_i} G(y_i \mid \nu)^{\delta_i}, \\ &= L(\theta) L(\nu). \end{aligned}$$

Como o interesse principal consiste em estimar a distribuição dos tempos de evento, temos

$$\begin{aligned} L(\theta, \nu) &\propto \prod_{i=1}^{n} f(y_i \mid \theta)^{\delta_i} S(y_i \mid \theta)^{1-\delta_i} \\ &= \prod_{i=1}^{n} h(y_i \mid \theta)^{\delta_i} S(y_i \mid \theta). \quad \text{usando as relações!!} \end{aligned}$$

Estimação dos parâmetros

Método da máxima verossimilhança

O vetor Escore é dado por,

$$U(\theta) \quad = \quad \frac{\partial \log L(\theta)}{\partial \theta}.$$

O Estimador de Máxima Verossimilhança (EMV) é a solução do seguinte sistema de equações:

$$\hat{U}(\theta) = \mathbf{0}.$$

Métodos numéricos serão utilizados quando não houver solução analítica para este sistema de equações. (Ex: Newton Raphson, etc...)

Estimação dos parâmetros

Método de Newton Raphson

O método segue o seguinte passo de iteração:

$$\hat{\theta}^{k+1} = \hat{\theta}^{k} + \mathcal{I}(\hat{\theta}^{k})^{-1} U(\hat{\theta}^{k})$$

em que U é o vetor escore de primeiras derivadas e $\mathcal{I}$ é a matriz de informação observada.
Usualmente o sistema é inicializado com $\theta^0 = 1$.

Estimação dos parâmetros

Propriedades do EMV

- O EMV tem, assintoticamente, distribuição normal;
- O EMV é consistente;
- A estatística da Razão de Verossimilhança (RV)

$$-2\log(L(\theta) = L(\hat{\theta})),$$

tem, assintoticamente, uma distribuição qui-quadrado com gl igual a dimensão de θ.

- `Invariância`: Se $\hat{\theta}$ é o EMV de θ, então $g(\hat{\theta})$ é o EMV de $g(\theta)$.

Estimação dos parâmetros

Quantidades importantes

- Vetor Escore:

$$U(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = \frac{\partial \ell(\theta)}{\partial \theta}.$$

- Matriz de informação de Fisher:

$$\mathcal{I}(\theta) = E\left[-\frac{\partial^2 \log L(\theta)}{\partial \theta^2}\right] = E\left[-\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right].$$

- Matriz de informação observada:

$$\mathcal{I}(\theta)\big|_{\theta=\hat{\theta}} = -\frac{\partial^2 \log L(\theta)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}} = -\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}}.$$

Estimação dos parâmetros

Resultados importantes

Propriedade Importante:

$$E(U(\theta)) = 0$$

$$Var(\hat{\theta}) \approx \mathcal{I}(\theta)^{-1}$$

e

$$Var(U(\theta) \approx \mathcal{I}(\theta)$$

- Para $\mathcal{I}(\theta)$ ser obtida é necessário especificar uma distribuição para os tempos de censura.
- Isso não é razoável em análise de sobrevivência. No entanto, $\mathcal{I}(\theta)$ é bem estimada por usando o EMV.

$$\hat{Var}(\hat{\theta}) \approx \mathcal{I}(\hat{\theta})^{-1}$$

Estimação dos parâmetros

Estatísticas relacionadas ao EMV

- Estatística de Wald na forma quadrática:

$$W = (\hat{\theta} - \theta)^t \mathcal{I}(\theta)(\hat{\theta} - \theta)$$

tem, para amostra grande, uma distribuição qui-quadrado com gl igual a dimensão de θ.

- Razão de verossimilhança (RV):

$$-2\log(L(\theta)/L(\hat{\theta})) = 2(\ell(\hat{\theta}) - \ell(\theta)).$$

- Estatística Escore S (Estatística de Rao):

$$U(\theta)^t \mathcal{I}(\theta)^{-1} U(\theta),$$

para amostra grande, ambas estatísticas tem uma distribuição qui-quadrado com gl igual a dimensão de θ.

Estimação dos parâmetros

Modelo Exponencial

No caso do modelo exponencial, temos que a função de verossimilhança é dada por

$$\begin{aligned} L(\alpha) &= \prod_{i=1}^{n} \left[\alpha \exp\{-\alpha y_i\}\right]^{\delta_i} \left[\exp\{-\alpha y_i\}\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \alpha^{\delta_i} \left[\exp\{-\alpha y_i\}\right] \end{aligned}$$

A função de log-verossimilhança e Escore:

$$\ell(\alpha) = \log(\alpha) \sum_{i=1}^{n} \delta_i - \alpha \sum_{i=1}^{n} y_i \quad \rightarrow \quad \frac{\partial \log L(\alpha)}{\partial \alpha} = \frac{1}{\alpha} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} y_i .$$

Estimação dos parâmetros

Modelo Exponencial

Dessa forma, igualando a zero

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} y_i} = \frac{\text{Total de eventos}}{\sum_{i=1}^{n} y_i} = \frac{1}{\dfrac{\sum_{i=1}^{n} y_i}{\text{Total de eventos}}}.$$

O termo $\sum_{i=1}^{n} y_i$ é denominado tempo total sob teste.

Observe que se todas as observações fossem não-censuradas, $\hat{\alpha} = \dfrac{1}{\bar{y}}$, em que $\bar{y}$ é a média amostral.

Estimação dos parâmetros

Modelo Weibull

No caso da distribuição de Weibull, temos que a função de verossimilhança é dada por

$$\begin{aligned} L(\alpha,\gamma) &= \prod_{i=1}^{n}\left[\alpha\gamma y_i^{\gamma-1}\exp\{-(\alpha y_i)^\gamma\}\right]^{\delta_i}[\exp\{-(\alpha y_i)^\gamma\}]^{1-\delta_i} \\ &= \prod_{i=1}^{n}\left[\alpha\gamma y_i^{\gamma-1}\right]^{\delta_i}\exp\{-(\alpha y_i)^\gamma\}. \end{aligned}$$

A função de log-verossimilhança:

$$\ell(\alpha) = \log(\alpha)\sum_{i=1}^{n}\delta_i + \log(\gamma)\sum_{i=1}^{n}\delta_i + (\gamma-1)\sum_{i=1}^{n}\log(y_i) - (\alpha y_i)^\gamma.$$

Estimação dos parâmetros

Modelo Weibull

O vetor escore é dado por

$$U(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = \frac{\partial \ell(\theta)}{\partial \theta}.$$

em que $\theta = \{\alpha, \gamma\}$.

Para a obtenção do EMV, usa-se métodos numéricos pois não há solução analítica para este sistema de equações.

Estimação dos parâmetros

Exemplo

- Dados provenientes da UFPR.
- 20 pacientes com câncer de bexiga submetidos a um procedimento cirúrgico a laser.
- `Resposta:` tempo da cirurgia até a reincidência da doença (meses).
- `Objetivo:` mediano de vida destes pacientes.
- `Dados (em meses):` 17 falhas e 3 censuras.

Estimação dos parâmetros

Exemplo

```r
require(survival)
#require(survminer)

tempos<-c(3,5,6,7,8,9,10,10,12,15,15,18,19,
cens<-c(1,1,1,1,1,1,1,0,1,1,0,1,1,1,1,1,1,1
dados <- data.frame(tempos, cens)
ekm <- survfit(Surv(tempos, cens)~1)
ekm
```

```
Call: survfit(formula = Surv(tempos, cens) ~ 1)

      n events median 0.95LCL 0.95UCL
[1,] 20     17     18      10      28
```

```r
summary(ekm)
```

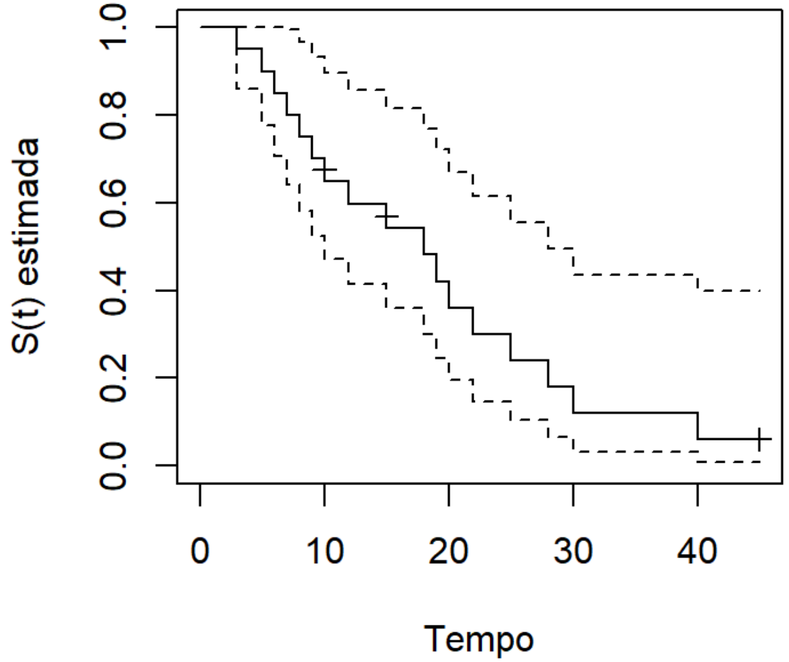
```
Call: survfit(formula = Surv(tempos, cens) ~ 1)

 time n.risk n.event survival std.err lower 95% CI
upper 95% CI
    3     20       1   0.9500  0.0487       0.85913
1.000
    5     19       1   0.9000  0.0671       0.77767
1.000
    6     18       1   0.8500  0.0798       0.70707
1.000
    7     17       1   0.8000  0.0894       0.64257
0.996
```

Estimação dos parâmetros

Exemplo

Estimador de Kaplan-Meier

```
plot(ekm, xlab="Tempo",
  ylab="S(t) estimada", mark.time=TRUE)
```



Estimação dos parâmetros

Exemplo

Ajuste pelo modelo Exponencial usando a função `survreg`.

```
ajust1<-survreg(Surv(tempos,cens)~1,dist='exponential')
summary(ajust1)
```

```
Call:
survreg(formula = Surv(tempos, cens) ~ 1, dist = "exponential")
             Value Std. Error    z      p
(Intercept) 3.016      0.243 12.4 <2e-16

Scale fixed at 1

Exponential distribution
Loglik(model)= -68.3   Loglik(intercept only)= -68.3
Number of Newton-Raphson Iterations: 4
n= 20
```

```
(alpha<-exp(ajust1$coefficients[1]))
```

```
(Intercept)
   20.41176
```

Estimação dos parâmetros

Exemplo

Ajuste pelo modelo Weibull usando a função `\survreg`.

```
ajust2<-survreg(Surv(tempos,cens)~1,
                dist='weibull')
summary(ajust2)
```

```
Call:
survreg(formula = Surv(tempos, cens) ~ 1, dist =
"weibull")
             Value Std. Error     z      p
(Intercept)  3.061      0.160 19.1 <2e-16
Log(scale)  -0.434      0.189 -2.3  0.022

Scale= 0.648

Weibull distribution
Loglik(model)= -66.1   Loglik(intercept only)= -66.1
```

```
alpha<-exp(ajust2$coefficients[1])
gama<-ajust2$scale
cbind(gama, alpha)
```

```
                 gama    alpha
(Intercept) 0.647922 21.33885
```

Estimação dos parâmetros

Exemplo

Ajuste pelo modelo Weibull usando a função `survreg`.

```
ajust3<-survreg(Surv(tempos,cens)~1,
                dist='lognormal')
summary(ajust3)
```

```
Call:
survreg(formula = Surv(tempos, cens) ~ 1, dist =
"lognormal")
             Value Std. Error     z      p
(Intercept)  2.717      0.176 15.42 <2e-16
Log(scale)  -0.268      0.174 -1.54   0.12

Scale= 0.765

Log Normal distribution
Loglik(model)= -65.7   Loglik(intercept only)= -65.7
```

```
mu<-ajust3$coefficients[1]
sigma<-ajust3$scale
cbind(mu, sigma)
```

```
                  mu     sigma
(Intercept) 2.717176 0.7648167
```

Estimação dos parâmetros

Exemplo

Função de sobrevivência.

$$S_E(t) = \exp\{-0.05t\}$$

$$S_W(t) = \exp\left\{-0.05t^{0.65}\right\}$$

$$S_{LN}(t) = \Phi\left[-(\log(t) - 2.72)/0.76\right]$$