UM CURSO DE ANÁLISE DE SOBREVIVÊNCIA

Prof. Eduardo Yoshio Nakano

1. INTRODUÇÃO

A Análise de Sobrevivência (ou Confiabilidade) consiste em uma coleção de procedimentos estatísticos para a análise de dados relacionados ao tempo até a ocorrência de um determinado **evento de interesse**. Geralmente o termo "Análise de Sobrevivência" refere-se a dados biomédicos, enquanto "Confiabilidade" refere-se a área de engenharia (pesquisas industriais).

Em estudos médicos, em geral, o evento de interesse é o tempo até o óbito do paciente, o tempo de duração da doença, o tempo até a ocorrência de uma doença ou à complicação da mesma, tempo até a cura, etc.

Em pesquisas industriais, o evento em questão está relacionado ao tempo até a falha de um determinado produto manufaturado, o tempo até a ocorrência de um reparo, ou o tempo até a utilização da garantia de um determinado equipamento, etc.

Nota: O termo "evento de interesse" tem sido estendido de forma aplica-lo aos mais variados tipos de eventos. Por exemplo:

- tempo até a ocorrência de um sinistro (área de seguros);
- tempo que um indivíduo permanece desempregado (ou empregado);
- tempo até que um cliente atrase a primeira prestação de um empréstimo;
- número de sessões de um tratamento até a cura, etc...

PRESENÇA DE CENSURAS

Um problema em análise de sobrevivência relaciona-se ao fato de a variável de interesse, tempo de sobrevida, ser temporal e, consequentemente, não ser medida instantaneamente e independentemente do tamanho da resposta. Valores grandes da variável tempo necessitam de mais tempo e persistência para serem observados. Em situações extremas, este fato pode comprometer a observação do valor da variável para alguns indivíduos, uma vez que o evento de interesse pode não ocorrer até o tempo final do estudo. Também, o paciente pode abandonar o estudo antes da observação do evento de interesse por causa de uma possível mudança de localidade ou devido a possíveis efeitos colaterais relacionados à terapia, ou falecer devido a outras causas de morte, mas não a em estudo. Isso pode, inclusive, acontecer antes mesmo do tempo final de recrutamento. Da mesma maneira, este tipo de problema pode estar presente em dados industriais, quando somente uma proporção de componentes em estudo falha até o final do estudo.

Entretanto, apesar de não termos o tempo até a ocorrência do evento de interesse para um determinado indivíduo, temos o tempo até, por exemplo, o abandono do estudo. Essa informação, apesar de incompleta, é útil e importante para a análise.

Dessa forma existe a necessidade da introdução de uma variável dicotômica na análise que indique se o valor do tempo de sobrevida para um determinado indivíduo foi ou não observado. Essa variável é conhecida como variável indicadora de censura, ou simplesmente, censura, e é definida como sendo igual a um, se o tempo de sobrevida é observado e igual à zero, caso o tempo de sobrevida seja censurado.

$$\delta_i = \begin{cases} 0, & \text{se o } i-\text{ésimo tempo foi censurado} \\ 1, & \text{se o } i-\text{ésimo tempo foi observado} \end{cases} \qquad i = 1,2,...,n.$$

Causas de Censuras

As censuras podem ocorrer devido a várias causas, como por exemplo:

Perda de acompanhamento: O paciente após entrar no estudo decide não ir até o fim, seja porque ele mudou de local de residência, de hospital ou simplesmente porque perdeu o interesse no estudo;

Término do estudo: para certos pacientes o estudo termina antes do término do estudo;

Falha devida a outra causa: A falha ou a morte de um paciente pode ocorrer devido a um fato totalmente independente do tratamento. Assim num estudo de tratamentos contra o câncer uma morte causada por um acidente automobilístico pode ser considerada como sendo uma censura;

Número de falhas suficientes: O pesquisador acredita que o número de equipamentos que falharam já é suficiente para obter a informação desejada.

Tipos de censuras

Existem três principais tipos de censura:

Censura do Tipo I: acontece nos casos em que o experimento tem um tempo de duração pré-determinado e ao final deste tempo um ou mais elementos em estudo não falharam e a informação sobre o tempo de falha destes elementos será incompleta. O percentual de censuras neste caso é uma variável aleatória. Na censura do Tipo I, censura-se o tempo de uma unidade em observação antes da ocorrência de falha ao se atingir um tempo pré-estabelecido e definido pelo pesquisador como tempo razoável para o estudo da variável.

Censura do Tipo II: ocorre nos casos em que o estudo é conduzido até que um número ($k \leq n$) de elementos falhe. O número de elementos a falharem deve ser determinado antes do inicio do experimento, desta forma o percentual de censuras será uma constante de escolha do experimentador. O pesquisador determina um número desejável de falhas, por exemplo, k falhas, e observa as unidades em estudo até que as k falhas aconteçam, lembrando-se que $k \leq n$ onde n é o tamanho da amostra a ser estudada.

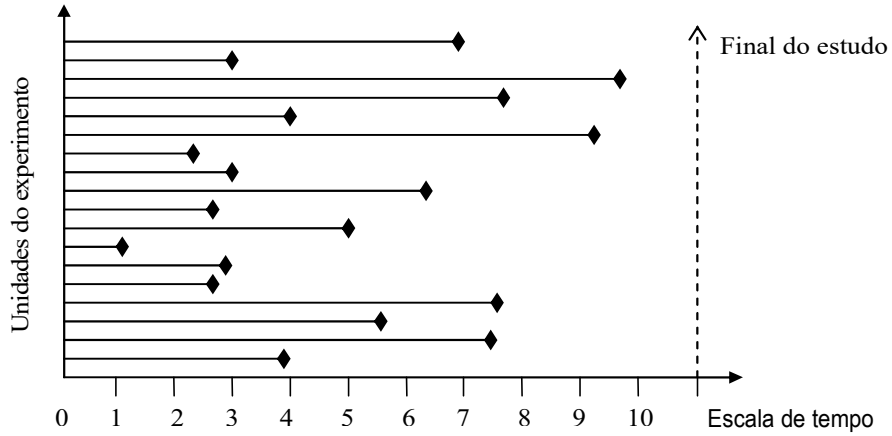
Censura aleatória: é a mais geral e engloba as demais, pode acontecer quando um ou mais componentes não puderem ser acompanhados até o final do experimento ou ainda quando estes falharem por motivos distintos do interesse do estudo. A censura aleatória ocorre de forma natural, ou seja, sem a manipulação do pesquisador, os exemplos anteriores de censuras por perda de acompanhamento, término do estudo e falha devida a outra causa são exemplos de censuras aleatórias.

As censuras definidas acima (Tipo I, II e aleatória) são denominadas censuras à direita. Existem também outras classes de censuras denominadas como censura à esquerda e censuras intervalares. Maiores detalhes sobre essas classes de censuras podem ser encontrados em Lawless (1982).

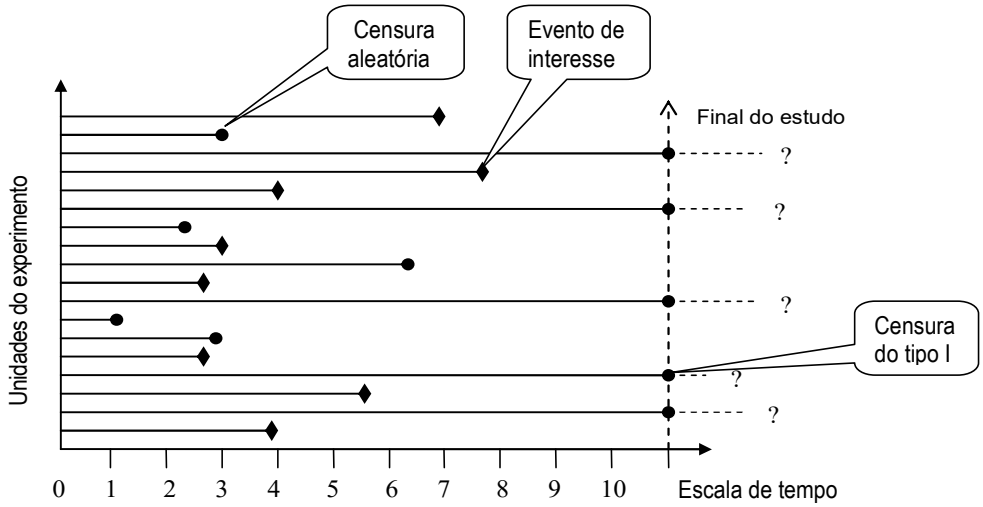
O esquema de censura do Tipo II é predominante em experimentos industriais e da engenharia, enquanto que, na prática médica costuma-se usar esquema de censura do Tipo I.

Mesmo sendo incompletos os dados censurados fornecem informações sobre o tempo de vida dos elementos em teste. A omissão dos dados censurados torna as estimativas viezadas. Nos casos onde ocorrem censuras à direita, os valores populacionais são subestimados no caso de omissão das censuras, pois os dados censurados têm tempos de falhas sabidamente maiores que os não censurados.

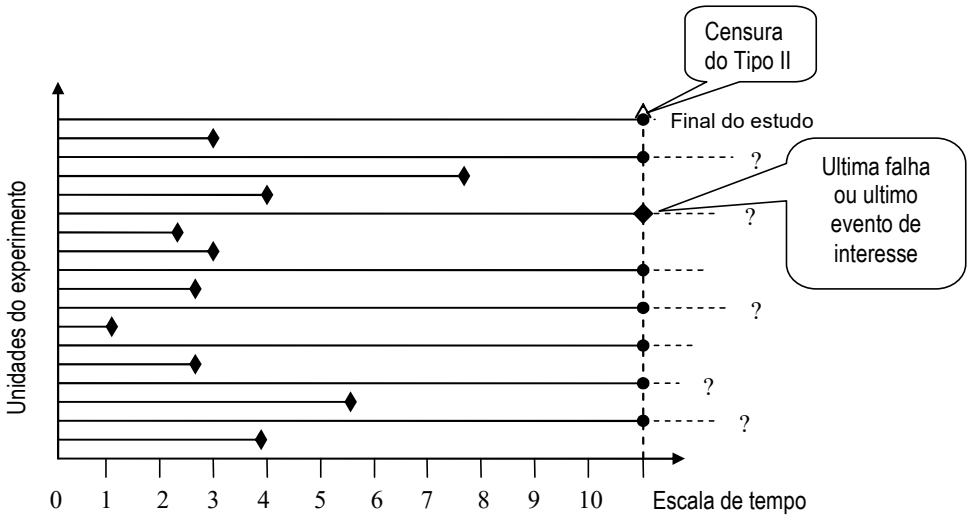
Podemos ver nas figuras a seguir a ilustração das seguintes censuras:



Dados com ausência de censura



Censura Tipo I
e
Aleatórias



Censura Tipo II

PRESENÇA DE VARIÁVEIS EXPLICATIVAS

Além do tempo de sobrevivência e da variável indicadora de censura, também podemos observar nos dados, variáveis que representam tanto a heterogeneidade existente na população, tais como, idade, sexo, entre outras; como também possíveis tratamentos aos quais os indivíduos são submetidos. Estas variáveis são conhecidas como **variáveis explicativas** ou **covariáveis**.

Muitas vezes, o objetivo da análise está centrado na relação entre o tempo de sobrevivência e algumas variáveis explicativas de interesse. A figura abaixo ilustra como o tempo de sobrevivência pode ser influenciado por três variáveis explicativas. A questão é saber se existe o efeito do tratamento, do sexo e da idade do paciente no tempo de sobrevivência, bem como se as interações entre estas variáveis explicativas são importantes.

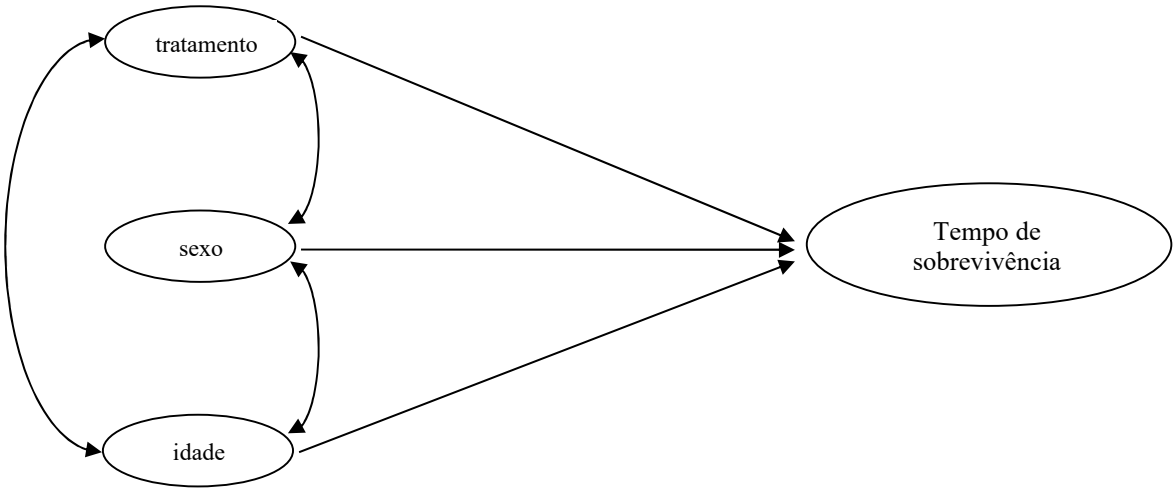


Figura: Relacionamento entre o tempo de sobrevivência e algumas variáveis explicativas.

Nota: Existem outras peculiaridades em Análise de Sobrevivência que não serão trabalhadas neste curso, como a quantidade de causas de falha e o número de eventos recorrentes. O primeiro é tratado como análise de dados de **Riscos Competitivos** (várias causas competem entre si para que o evento de interesse ocorra) e o segundo é tratado como análise de **Eventos Recorrentes** (quando o evento de interesse, este sendo não fatal, é observado várias vezes para um mesmo indivíduo).

2. DESCRIÇÃO DO COMPORTAMENTO DO TEMPO DE SOBREVIVÊNCIA

Na análise de sobrevivência, o comportamento da variável aleatória não-negativa que descreve o tempo até a falha é geralmente especificada pela sua função de sobrevivência ou pela função de risco (ou taxa de falha). Essas duas funções são matematicamente equivalentes, de forma que se uma delas é especificada, a outra pode ser derivada. Este capítulo apresenta a definição da função de sobrevivência, da função de risco e de outras funções relacionadas que descrevem o tempo até a falha. Também serão descritas as relações entre essas funções e outras medidas de interesse como os quantis do tempo de sobrevivência, o tempo de vida médio e a vida média residual.

2.1. Variáveis aleatórias contínuas

O comportamento de uma variável aleatória contínua não-negativa, $T \geq 0$, pode ser caracterizada por meio de várias funções matematicamente equivalentes. Entre elas tem-se a função densidade de probabilidades, a função de sobrevivência e as funções de risco e risco acumulado. Essas funções serão descritas em detalhes a seguir.

2.1.1. Função Densidade de Probabilidades

Seja T uma variável aleatória não-negativa contínua. A Função Densidade de Probabilidades de T, $f(t)$, é uma função que satisfaz as seguintes condições (Meyer, 1983):

i) $f(t) \geq 0$ para todo $t \geq 0$;

ii) $\int_0^\infty f(t)dt = 1$; e

iii) $P(a \leq T \leq b) = \int_a^b f(t)dt$, $\forall\ 0 \leq a < b$.

Essa função pode ser vista como o limite da probabilidade de um indivíduo experimentar o evento de interesse no intervalo de tempo $[t, t+\Delta t)$, dividida pelo comprimento do intervalo e pode ser expressa por

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t},\ t \geq 0. \qquad (2.1)$$

2.1.2. Função de Sobrevivência

Denotada por $S(t)$, a função de sobrevivência é definida como a probabilidade de um indivíduo não falhar até um determinado tempo t, ou seja, a probabilidade desse indivíduo sobreviver além de t. Esta função é uma das principais funções probabilísticas para representar o tempo de sobrevivência e é definida por

$$S(t) = P[T > t] = \int_t^\infty f(t)dt\ ,\ t \geq 0. \qquad (2.2)$$

A função de sobrevivência (2.2) é uma função não-crescente, absolutamente contínua, tal que $\lim_{t \to 0} S(t) = 1$ e $\lim_{t \to \infty} S(t) = 0$.

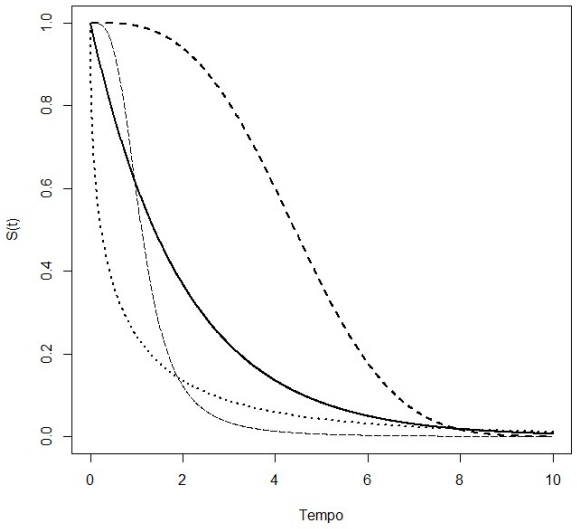


Figura 2.1 - Funções de sobrevivência de variáveis aleatórias contínuas.

A função de sobrevivência também pode ser utilizada para se determinar os quantis do tempo de sobrevivência. Seja t_p o p-ésimo quantil da variável aleatória contínua T, isto é, $P[T \leq t_p] = p$. Assim,

$$t_p = S^{-1}(1-p), \ \forall\ 0 < p < 1. \tag{2.3}$$

2.1.3. Função de Risco (ou Taxa de Falha)

A Função de Risco (ou função taxa de falha), denotada por h(t), representa o risco instantâneo que o indivíduo tem de experimentar o evento de interesse em um terminado tempo t. No caso de uma variável aleatória contínua, esta função é definida como a razão do limite da probabilidade condicional de um indivíduo experimentar o evento de interesse no intervalo de tempo [t, t+Δt) dado que o mesmo não tenha experimentado o evento de interesse antes de t, sobre o intervalo de tempo Δt . A função $h(t)$ é expressa por

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}, \ t \geq 0. \tag{2.4}$$

É importante notar que para variáveis aleatórias contínuas, a função de risco $h(t)$ é uma função que assume valores reais positivos e essa função não é limitada superiormente.

A função $h(t)$ descreve como o risco (taxa de falhar) se modifica com o passar do tempo. Por esse motivo, essa função é muito utilizada para descrever o comportamento do tempo de sobrevivência. Alguns autores consideram que a taxa de falha é mais informativa que a função de sobrevivência, pois diferentes funções de sobrevivência podem ter formas semelhantes, enquanto as respectivas funções de risco podem diferir drasticamente (Colosimo, 2006). A Figura 2.2, apresenta funções de risco com comportamento constante, monótona (crescente e decrescente) e não monótonas (unimodais e em forma de banheira).

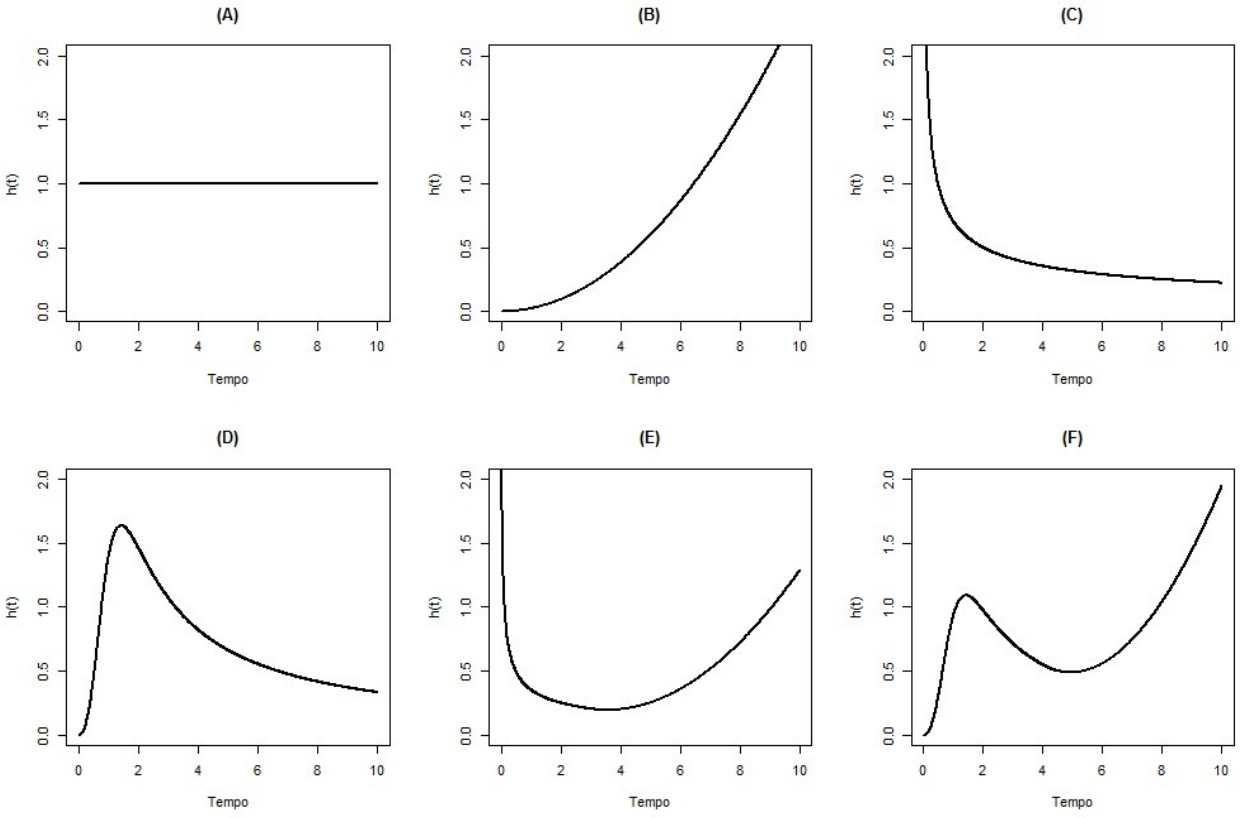


Figura 2.2 - Funções de risco: constante (A), crescente (B), decrescente (C), unimodal (D), banheira (E) e multimodal (F).

A função de risco crescente (B) indica que a taxa de falha de um indivíduo aumenta com o passar do tempo, similar a um efeito de envelhecimento. Em contrapartida, uma função decrescente (C) indica que a taxa de falha diminui com o passar do tempo (rejuvenescimento). A função constante (A) indica que a taxa de falhar não se altera com o passar do tempo (o indivíduo não envelhece/rejuvenesce). O risco unimodal (D) pode descrever o comportamento de um indivíduo que contraiu uma determinada doença que pode apresentar um risco inicialmente crescente, mas que decresce com a introdução de um tratamento. A função em forma de banheira, ou em forma de "U" (E), é uma função que inicialmente é decrescente, podendo se estabilizar por um determinado período de tempo e, finalmente, se torna crescente. Esse tipo de função pode ser adequado para representar o comportamento humano: que apresenta um período inicial com alta taxa de mortalidade que decai rapidamente nos primeiros anos de vida até se estabilizar na idade adulta, crescendo novamente com o aumento da idade. Ainda, podem existir situações em que indivíduos podem apresentar funções de risco mais complexas, com mais de uma moda (F).

2.1.4. Função de Risco Acumulado (ou Taxa de Falha Acumulada)

Uma outra função importante que pode ser obtida a partir da função $h(t)$ é a Função de Risco Acumulada ou Taxa de Falha Acumulada, denotada aqui por $H(t)$. A função $H(t)$ não tem uma interpretação direta, mas ela é útil em procedimentos de estimação não-paramétricos e na seleção de um modelo mais apropriado para ajustar um determinado conjunto de dados. A função $H(t)$ fornece o risco acumulado do indivíduo no tempo t, e, no caso de uma variável aleatória contínua, esta função é definida por

$$H(t) = \int_0^t h(u)du, \;\; t \geq 0. \qquad (2.5)$$

2.1.5. Relações importantes entre *f*(*t*), *S*(*t*), *h*(*t*) e *H*(*t*)

Como citado anteriormente, as funções $f(t)$, $S(t)$, $h(t)$ e $H(t)$, definidas anteriormente, são matematicamente equivalentes. Serão descritas aqui algumas relações matemáticas importantes entre essas funções. Essas relações podem ser utilizadas na obtenção de uma dessas funções quando uma delas é especificada.

Dada as definições (2.1) e (2.4), é fácil ver que

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \cap T \ge t)}{\Delta t P(T \ge t)} = \frac{\lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t}}{P(T > t)} = \frac{f(t)}{S(t)}. \tag{2.6}$$

A função densidade de probabilidades, f(t), é definida como a derivada da Função de Distribuição Acumulada, F(t), isto é,

$$f(t) = \frac{d}{dt} F(t).$$

Visto que $F(t) = 1 - S(t)$, tem-se que

$$f(t) = \frac{d}{dt}\left[1 - S(t)\right] = -\frac{d}{dt} S(t) = -S'(t). \tag{2.7}$$

Substituindo (2.7) em (2.6) obtém-se que

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log S(t). \tag{2.8}$$

Desta forma, integrando ambos os termos (2.8) resulta em

$$\log S(t) = -\int_0^t h(u)du = -H(t),$$

o que implica em

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\} = \exp\{-H(t)\}. \tag{2.9}$$

De (2.6) tem-se que

$$f(t) = h(t)S(t). \tag{2.10}$$

Desta forma, substituir (2.9) em (2.10) resulta em

$$f(t) = h(t)\exp\left\{-\int_0^t h(u)du\right\}. \tag{2.11}$$

A expressão (2.11) pode ser útil para desenvolver procedimentos de estimação baseados somente na função de risco.

2.1.6. Momentos e Vida Média Residual

Duas outras quantidades de interesse na análise de dados de sobrevivência são os momentos da variável aleatória *T* e a Vida Média Residual. Como será visto a seguir, ambas quantidades podem ser obtidas por meio da função de sobrevivência.

Se *T* é uma variável aleatória contínua não-negativa, então o seu *r*-ésimo momento, se existir, pode ser obtido por (James, 2015):

$$E(T^r) = r\int_0^\infty t^{r-1}S(t)dt \text{ , para todo } r \geq 1. \quad (2.12)$$

Esse resultado implica que a média e a variância do tempo de vida são dadas, respectivamente, por:

$$E(T) = \int_0^\infty S(t)dt \quad (2.13)$$

e

$$Var(T) = 2\int_0^\infty tS(t)dt - \left(\int_0^\infty S(t)dt\right)^2 . \quad (2.14)$$

A Equação (2.13) sugere que o tempo de vida médio de um indivíduo é obtido pela área sob a função de sobrevivência.

A Vida Média Residual é a expectativa média de vida de um indivíduo que não falhou até o tempo *t*. Essa quantidade é obtida por meio do tempo de vida residual ($T–t|T\geq t$), que representa o tempo restante de vida para indivíduos com idade *t*. A função Vida Média Residual é definida por $\nu(t) = E(T - t \mid T \geq t)$, para todo $t \geq 0$.

A função $\nu(t)$ pode ser descrita em termos da função de sobrevivência por meio da expressão

$$\nu(t) = \frac{1}{S(t)}\int_t^\infty S(u)du \text{ , para todo } t \geq 0. \quad (2.15)$$

A vida média residual no tempo *t* é área sob a curva de sobrevivência à direita do ponto *t*, dividido pelo valor da função de sobrevivência neste ponto. Note que $S(0)=1$ e, portanto, $\nu(0) = E(T)$.

2.2. Variáveis aleatórias discretas

Uma variável aleatória é dita ser discreta se assume um número finito ou infinito enumerável de valores. Em análise de sobrevivência, a variável de interesse é, em geral, uma variável de contagem representando, por exemplo, número de ciclos até a ocorrência de um evento de interesse ou o número de sessões de um tratamento até a cura. Desta forma, essa variável aleatória assume, na maioria dos casos, números inteiros não negativos, ou seja, $t = 0, 1, 2, ...$

De forma equivalente a uma variável aleatória contínua, o comportamento de uma variável aleatória discreta não-negativa pode ser caracterizado por meio de várias funções matematicamente equivalentes. Entre elas tem-se a função de probabilidades (ou distribuição de probabilidades), a função de sobrevivência e as funções de risco e risco acumulado. Essas funções serão descritas em detalhes a seguir.

2.2.1. Função de Probabilidades ou Distribuição de Probabilidades

Seja T uma variável aleatória discreta que assume valores inteiros não negativos, ou seja, t = 0, 1, 2, ... A função de probabilidades ou distribuição de probabilidades de T é uma função que atribui a cada possível valor da variável aleatória a sua probabilidade. Essa função, denotada aqui por $p(t) = P(T=t)$, deve satisfazer as seguintes condições:

i) $p(t) \geq 0$, para $t = 0, 1, 2,...$; e

ii) $\sum_{t=0}^{\infty} p(t) = 1$.

2.2.2. Função de Sobrevivência

Definida como a probabilidade de um indivíduo não falhar até um determinado tempo t, a função de sobrevivência definida, no caso em que T é uma variável aleatória discreta, por

$$S(t) = P[T > t] = \sum_{k=t+1}^{\infty} p(k) = \sum_{k=t+1}^{\infty} P(T = k), \quad t = 0, 1, 2, ... \tag{2.16}$$

A função de sobrevivência é uma função definida em todos os reais não negativos. De fato, note que, para todo $a \geq 0$ e $t = 0, 1, 2, ...$, é correto dizer que

$$S(a) = P[T > a] = P[T > t] = S(t), \text{ se } t \leq a < t+1. \tag{2.17}$$

Isto é, a função $S(t)$ de uma variável aleatória discreta é uma função que decresce nos pontos em que t tem probabilidade positiva e é constante dos demais pontos de t (função escada). Ainda, no caso de uma variável aleatória discreta, tem-se que $\lim_{t\to 0} S(t) = 1 - P(T = 0)$ e $\lim_{t\to\infty} S(t) = 0$.

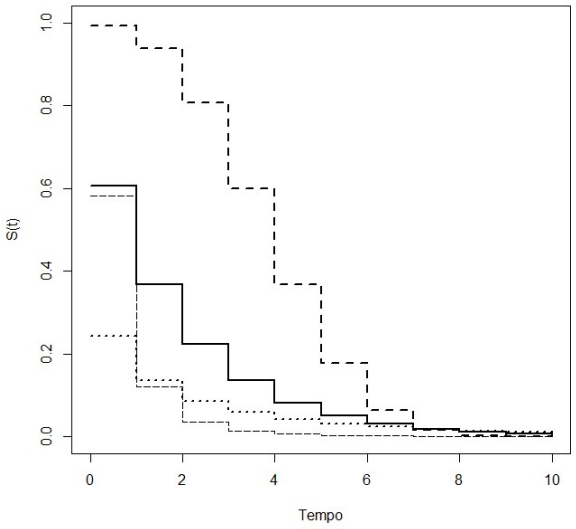


Figura 2.3 - Funções de sobrevivência de variáveis aleatórias discretas.

2.2.3. Função de Risco (ou Taxa de Falha)

No caso de uma variável aleatória discreta, a Função de Risco (ou função taxa de falha), é definida como a probabilidade condicional do indivíduo experimentar o evento de interesse no instante *t*, dado que o mesmo não tenha experimentado o evento de interesse antes de *t*, isto é:

$$h(t) = P(T = t \mid T \geq t), \ t = 0, 1, 2, ... \tag{2.18}$$

Note que para os demais valores de *t* (negativos ou não inteiros), a função de risco (2.18) é igual a zero. Além disso, por ser uma probabilidade condicional, a função de risco para variáveis aleatórias discretas é uma função limitada no intervalo [0, 1]. A Figura 2.4, apresenta funções de risco com comportamento constante, monótona (crescente e decrescente) e não monótonas (unimodais e em forma de banheira).

2.2.4. Função de Risco Acumulado (ou Taxa de Falha Acumulada)

A função de risco acumulado, $H(t)$, fornece o risco acumulado do indivíduo no tempo *t*, e, no caso de uma variável aleatória discreta, esta função é definida por

$$H(t) = \sum_{k=0}^{t} h(k), \ t = 0, 1, 2, ... \tag{2.19}$$

A função de risco acumulado (2.19) não tem uma interpretação direta e é uma função que assume valores reais positivos, não sendo limitada superiormente.

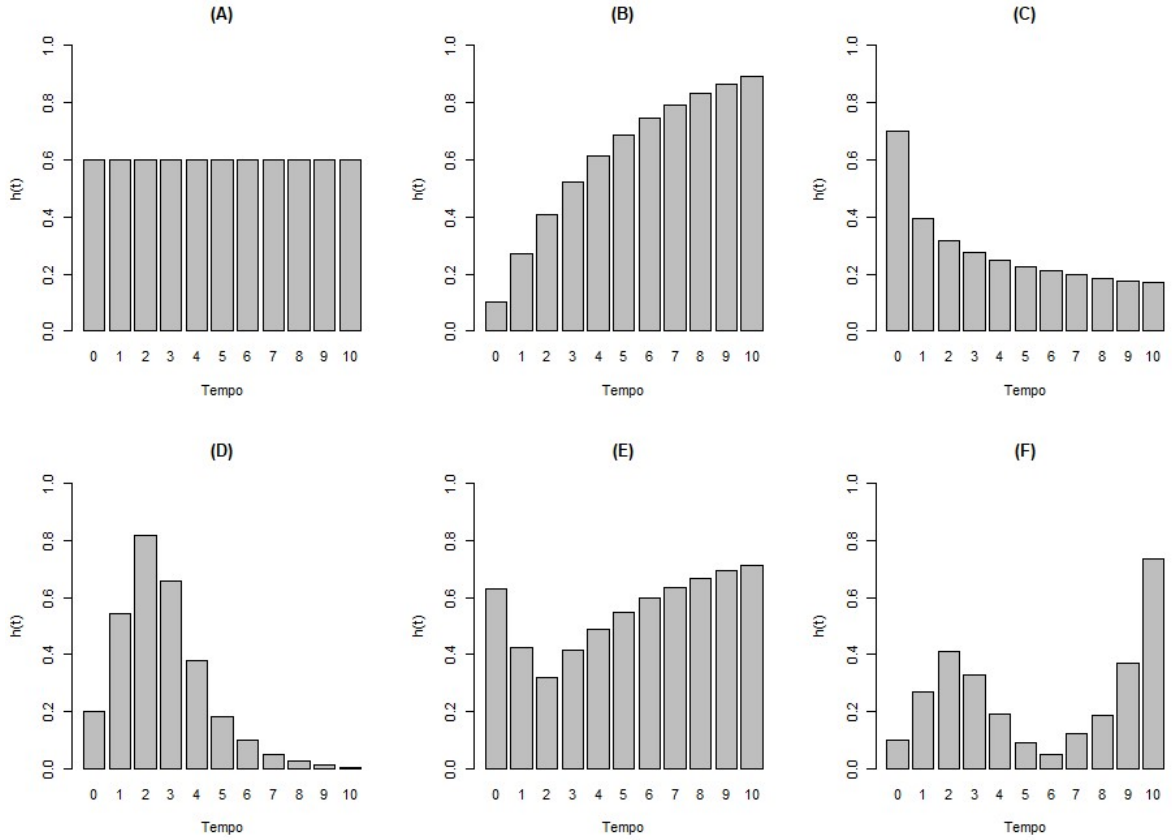


Figura 2.4 - Funções de risco: constante (A), crescente (B), decrescente (C), unimodal (D), banheira (E) e multimodal (F).

2.2.5. Relações importantes entre *p*(*t*), *S*(*t*), *h*(*t*) e *H*(*t*)

Será descrito aqui algumas relações matemáticas importantes entre as funções *f*(*t*), *S*(*t*), *h*(*t*) e *H*(*t*) de uma variável aleatória discreta.

Dada as definições (2.16) e (2.18), é fácil ver que

$$h(t) = P(T = t \mid T \geq t) = \frac{P(T = t \cap T \geq t)}{P(T \geq t)} = \frac{P(T = t)}{P(T = t) + P(T > t)} = \frac{p(t)}{p(t) + S(t)}, \; t = 0,1,2,\ldots, \quad (2.20)$$

que resulta facilmente em

$$p(t) = \frac{h(t)}{1 - h(t)} S(t), \; t = 0, 1, 2, \ldots \quad (2.21)$$

Ainda, a distribuição de probabilidades pode ser escrita em termos da função de sobrevivência por meio da expressão

$$p(t) = \begin{cases} 1 - S(0), & se \;\; t = 0 \\ S(t-1) - S(t), & se \;\; t = 1,2,\ldots \end{cases} \quad (2.22)$$

Veja ainda que, para $t = 1, 2,\ldots$, tem-se que

$$S(t) = \frac{S(0)}{1} \frac{S(1)}{S(0)} \frac{S(2)}{S(1)} \ldots \frac{S(t-1)}{S(t-2)} \frac{S(t)}{S(t-1)} = S(0) \prod_{k=1}^{t} \frac{S(k)}{S(k-1)}.$$

Visto que $S(0) = 1 - p(0)$ e $h(0) = p(0)$, tem-se que a função de sobrevivência pode ser obtida a partir da função de risco por meio da expressão[(1)]

$$\begin{aligned} S(t) &= [1 - h(0)] \prod_{k=1}^{t} \frac{S(k)}{p(k) + S(k)} = [1 - h(0)] \prod_{k=1}^{t} \left[1 - \frac{p(k)}{p(k) + S(k)} \right] = [1 - h(0)] \prod_{k=1}^{t} [1 - h(k)] \\ &= \prod_{k=0}^{t} [1 - h(k)], \quad t = 0,1,2,\ldots \end{aligned} \quad (2.23)$$

De (2.21) e (2.23) é possível escrever a distribuição de probabilidades em termos da função de risco por meio da expressão:

$$p(t) = \frac{h(t)}{1 - h(t)} \prod_{k=0}^{t} [1 - h(k)], \; t = 0, 1, 2, \ldots \quad (2.24)$$

---

(1) Note que, diferente de uma variável contínua, no caso discreto $S(t) \neq \exp\{-H(t)\} = \exp\left\{-\sum_{k=0}^{t} h(k)\right\}$. Por essa razão, Roy e Gupta (1992) propuseram uma definição alternativa da função de risco para variáveis aleatórias discretas: $h'(t) = \log\left(\frac{S(t-1)}{S(t)}\right)$. Rinne (2008) descreve *h*'(t) como uma pseudo-função de risco de forma diferenciá-la de *h*(*t*) definida por (2.18). Ao contrário de *h*(*t*), a função de risco *h*'(*t*) satisfaz $S(t) = \exp\{-H'(t)\} = \exp\left\{-\sum_{k=0}^{t} h'(k)\right\}$.

2.2.6. Momentos e Vida Média Residual

Se T é uma variável aleatória discreta não-negativa, então o seu r-ésimo momento, se existir, pode ser obtido por:

$$E(T^r) = \sum_{t=0}^{\infty} t^r \, p(t) \text{, para todo } r \geq 1. \tag{2.25}$$

Em particular, o primeiro e segundo momento da variável aleatória T podem ser obtidos em termos da função de sobrevivência por

$$E(T) = \sum_{t=0}^{\infty} S(t) \tag{2.26}$$

e

$$E(T^2) = E(T) + 2\sum_{t=1}^{\infty} tS(t) \, . \tag{2.27}$$

A função Vida Média Residual da variável aleatória discreta T é dada por

$$\nu(t) = E[T - t \mid T \geq t] = \frac{1}{p(t) + S(t)} \sum_{k=t}^{\infty} S(k) \text{, } t = 0, 1, 2, ... \tag{2.28}$$

Note que, $\nu(0) = E(T)$.

2.2.7. Distribuições discretas análogas às distribuições de variáveis aleatórias contínuas não-negativas.

Distribuições de variáveis aleatórias contínuas podem ser usados para gerar modelos discretos análogos. Um procedimento simples é agrupar os tempos contínuos em intervalos unitários. Seja X uma variável aleatória contínua não negativa com função de distribuição acumulada $F_X(x)$. A variável aleatória discreta pode ser obtida por $T=[X]$, em que $[X]$ representa "a parte inteira de X", isto é, o maior inteiro menor ou igual a X. Desta forma, a distribuição de probabilidades da variável aleatória discreta T pode ser escrita como

$$P(T = t) = P(t \leq X < t + 1) = F_X(t + 1) - F_X(t) \text{, } \quad t = 0, 1, 2, ... \tag{2.29}$$

Nota: Existem outras formas de obter distribuições discretas a partir de variáveis aleatórias contínuas. Stein and Dattero (1984) propuseram uma distribuição discreta que imita a forma da função de risco da variável continua e Padgett e Spurrier (1985) propõe obter distribuições discretas por meio da função de risco e da vida média residual de variáveis contínuas (Rinne, 2008).

2.3. Exercícios

1. Seja T uma variável aleatória com distribuição Exponencial com média λ, isto é,

$$f(t) = \frac{1}{\lambda} e^{-\frac{t}{\lambda}}, \qquad t \geq 0 .$$

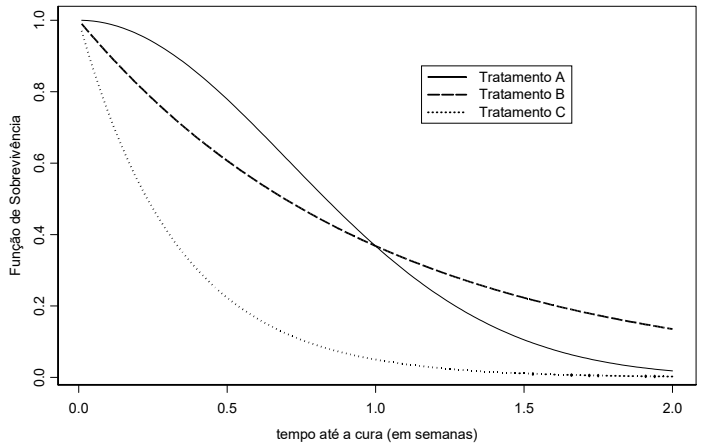
i) Determine as funções $S(t)$, $h(t)$, $H(t)$ e $\nu(t)$.
ii) Encontre o p-ésimo quantil, da variável aleatória T.
iii) Mostre que a variável discreta análoga à T, segundo (2.29) é a distribuição Geométrica.

2. Seja T uma variável aleatória discreta que assume valores inteiros não negativos. Mostre que

$$E(T) = \sum_{t=0}^{\infty} S(t) \quad \text{e} \quad E(T^2) = E(T) + 2\sum_{t=1}^{\infty} tS(t) .$$

3. Considere que o tempo de vida (em anos) de um equipamento eletrônico pode ser representado por uma variável aleatória T com função de risco $h(t) = 2t$, $t \geq 0$.
i) Determine a função densidade de probabilidades de T.
ii) Qual a probabilidade desse equipamento viver por mais de 1 ano?
iii) Qual a mediana do tempo de vida desse equipamento?
iv) Calcule a vida média residual de um equipamento após 1 ano de vida.
v) Encontre a distribuição de probabilidades, a função de sobrevivência e a função de risco da variável discreta análoga à T, segundo (2.29).

4. Considere três tratamentos, A, B e C, para o combate de uma doença. O tempo (em semanas) até a cura é representado por uma variável aleatória T, que depende de qual tratamento foi utilizado. A figura abaixo apresenta a Função de Sobrevivência para os três tratamentos.



i) Qual dos três tratamentos é o mais eficiente?
ii) Se o objetivo é a cura da doença em menos de uma semana, o tratamento A ou B deve ser preferido?
iii) Se o objetivo é garantir a cura da doença em no máximo duas semanas, o tratamento A ou B deve ser preferido?

2.4. Referências

Colosimo, E.A; Giolo, S.R. *Análise de sobrevivência aplicada*. São Paulo: Edgard Blucher Ltda, 369p, 2006.

James, B.R. *Probabilidade: um curso em nível intermediário*. Rio de Janeiro: IMPA, 4 ed., 299p, 2015.

Meyer, P.L. *Probabilidade - Aplicações à Estatística*, Livros Técnicos e Científicos Editora S.A, Rio de Janeiro, 2 ed., 444p,1983.

Padgett, D.J; Spurrier, J.D. Discrete failure models, *IEEE Transactions on Reliability*, vol. 34, n. 3, p. 253–256, 1985.

Rinne, H. *The Weibull distribution: a handbook,* Chapman and Hall/CRC, 1 ed., 808p, 2008.

Roy, D.; Gupta, R.P. Classifications of discrete lives, *Microelectronics and Reliability*, 32, 1459–1473, 1992.

Stein, W.E.; Dattero, R. A new discrete Weibull distribution, *IEEE Transactions on Reliability*, vol. 33, n. 2, p. 196–197, 1984.

APENDICE A

Tabela. Definições e relações das principais funções utilizadas para descrever o tempo de sobrevivência.

	Variável aleatória contínua	**Variável aleatória discreta**
Função (densidade) de probabilidade	$f(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t}$	$p(t) = P(T = t)$
Função de distribuição	$F(t) = P[T \le t] = \int_0^t f(t)dt$	$F(t) = P[T \le t] = \sum_{k=0}^{t} p(k)$
Função de sobrevivência	$S(t) = P[T > t] = \int_t^\infty f(t)dt$	$S(t) = P[T > t] = \sum_{k=t+1}^{\infty} p(k)$
Função de risco	$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$	$h(t) = P(T = t \mid T \ge t)$
Função de risco acumulado	$H(t) = \int_0^t h(u)du$	$H(t) = \sum_{k=0}^{t} h(k)$
Vida media residual E[T–t|T≥t]	$\nu(t) = \frac{1}{S(t)} \int_t^\infty S(u)du$	$\nu(t) = \frac{1}{p(t) + S(t)} \sum_{k=t}^{\infty} S(k)$
Algumas relações entre as funções	$h(t) = \frac{f(t)}{S(t)}$	$h(t) = \frac{p(t)}{p(t) + S(t)}$
	$h(t) = -\frac{d}{dt} \log S(t)$	$h(t) = \begin{cases} 1 - S(0), & se\ t = 0 \\ 1 - \frac{S(t)}{S(t-1)}, & se\ t = 1, 2, \ldots \end{cases}$
	$H(t) = -\log S(t)$	$H(t) = [1 - S(0)] + \sum_{k=1}^{t} \left(1 - \frac{S(k)}{S(k-1)}\right)$
	$S(t) = \exp\{-H(t)\}$	$S(t) = \prod_{k=0}^{t} [1 - h(k)]$
	$f(t) = h(t)S(t)$	$p(t) = \frac{h(t)}{1 - h(t)} S(t)$
	$f(t) = -\frac{d}{dt} S(t)$	$p(t) = \begin{cases} 1 - S(0), & se\ t = 0 \\ S(t-1) - S(t), & se\ t = 1, 2, \ldots \end{cases}$
	$f(t) = h(t) \exp\left\{-\int_0^t h(u)du\right\}$	$p(t) = \frac{h(t)}{1 - h(t)} \prod_{k=0}^{t} [1 - h(k)]$

3. MÉTODOS NÃO-PARAMÉTRICOS DE ESTIMAÇÃO

Do ponto de vista prático, tem-se inicialmente o interesse na estimação da função densidade de probabilidades $f(t)$, da função de sobrevivência, $S(t)$ e da função de risco, $h(t)$. Estas funções podem ser estimadas diretamente a partir dos dados amostrais por meio de procedimentos não-paramétricos que serão apresentados a seguir.

3.1. Estimação na ausência de censura

O procedimento de estimação de dados não censurados será ilustrado a partir do seguinte exemplo:

Exemplo 1: *Tempo de sobrevivência (em meses) de n=20 pacientes após receberem transplante de rim.*

2	4	5	7	11	11	12	13	15	19
22	23	25	25	29	32	37	39	44	55

3.1.1. Estimação da Função Densidade de Probabilidades

Quando a variável de interesse é contínua, a função densidade de probabilidades, $f(t)$, pode ser estimada a partir dos dados amostrais por meio da expressão

$$\hat{f}(t) = \frac{\text{nº de indivíduos que experimentaram o evento de interesse no intervalo } [t, t+\Delta t)}{\Delta t \times (\text{nº total de indivíduos})}. \quad (3.1)$$

Aqui, Δt representa a amplitude do intervalo.

Considerando os dados do Exemplo 1 e 6 intervalos com amplitude Δt =10 (amplitudes iguais), tem-se por meio de (3.1) que a estimativa de $f(t)$ no intervalo [0,10) é dado por

$$\hat{f}(t) = \frac{4}{10 \times 20} = 0{,}02 \quad \text{para } 0 \le t < 10.$$

E a estimativa de $f(t)$ no intervalo [10,20) é dado por

$$\hat{f}(t) = \frac{6}{10 \times 20} = 0{,}03 \quad \text{para } 10 \le t < 20.$$

De forma análoga é possível obter as estimativas de f(t) para os demais intervalos. Os resultados podem ser vistos na Tabela 3.1 e Figura 3.1.

Tabela 3.1 : Estimativa não paramétrica de $f(t)$ para os dados do Exemplo 1.

Intervalo	Número de mortes no intervalo	Estimativa de $f(t)$
[0 ; 10)	4	0,020
[10 ; 20)	6	0,030
[20 ; 30)	5	0,025
[30 ; 40)	3	0,015
[40 ; 50)	1	0,005
[50 ; 60)	1	0,005

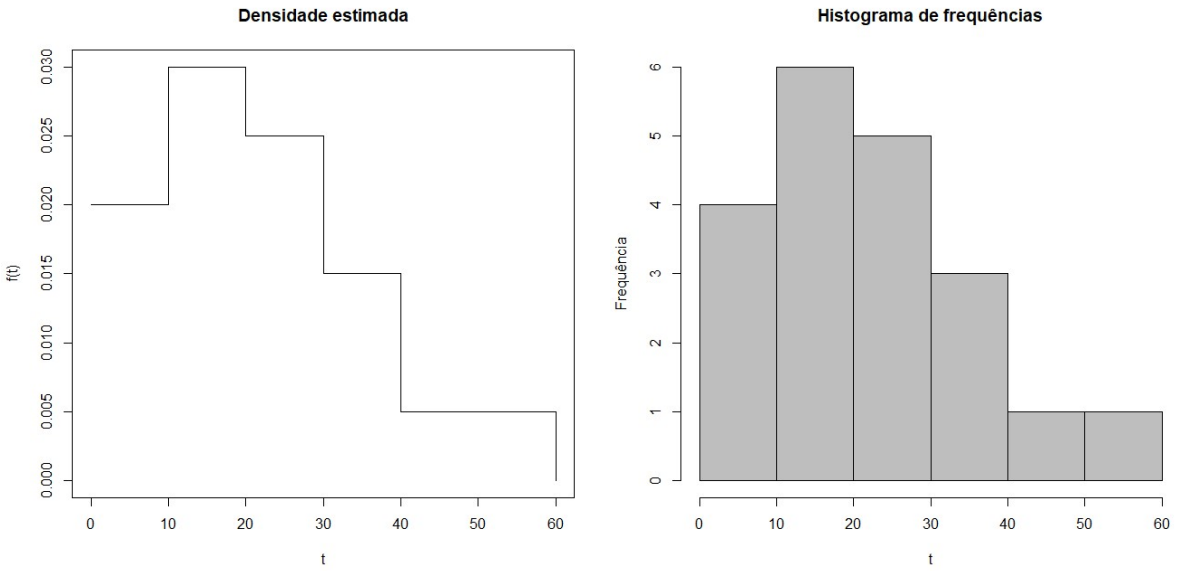


Figura 3.1 : Estimativa não paramétrica de $f(t)$.

Notas:
(i) O valor de Δt pode ser definido da mesma forma que a amplitude de classes nas Tabelas de Frequências.
(ii) Na prática a forma da densidade de probabilidades é, usualmente, estimada por meio de um Histograma de frequências (brutas ou relativas). Note na Figura 3.1 que a densidade estimada e o histograma de frequências diferem apenas de uma constante normalizadora.
(iii) Se a variável de interesse é discreta, a estimativa da distribuição (função) de probabilidades é dada pela frequência relativa de cada valor observado na amostra, isto é

$$\hat{p}(t) = \frac{\text{nº de indivíduos que experimentaram o evento de interesse no tempo } t}{\text{nº total de indivíduos}}, \; t = 0,1,2,...$$

3.1.2. Estimação da Função de Sobrevivência

A função de sobrevivência, $S(t)$, pode ser estimada, a partir dos dados amostrais, como a proporção de indivíduos que não falharam até o tempo t. Seu estimador é dado por:

$$\hat{S}(t) = \frac{\text{nº de indivíduos com tempo de sobrevivência} > t}{\text{nº total de indivíduos}} \text{, para todo } t \geq 0. \qquad (3.2)$$

Em Análise de Sobrevivência, pode ser interessante ordenar os valores (distintos) observados na amostra. Seja $t_{(j)}$ o j-ésimo valor ordenado da amostra observada., j=1,2,....k. Considerando que, $t_{(0)} = 0$ e $t_{(k+1)} = \infty$, tem-se que

$$0 = t_{(0)} < \min\{t_1,t_2,\ldots,t_n\} = t_{(1)} < t_{(2)} < \ldots < t_{(k)} = \max\{t_1,t_2,\ldots,t_n\} < t_{(k+1)} = \infty.$$

Note que $k \leq n$ (no caso em que a amostra observada não apresenta empates, tem-se que $k = n$), sendo n o tamanho da amostra.

Desta forma, o estimador da função de sobrevivência definido em (3.2) pode ser representado por:

$$\hat{S}(t) = \frac{n_j - d_j}{n}, \text{ para } t_{(j)} \leq t < t_{(j+1)}, \; j = 0,1,\ldots,k. \qquad (3.3)$$

Aqui, n_j é o número de indivíduos que estão sob risco no o tempo $t_{(j)}$ (inclusive) e d_j representa o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=0, 1, 2, ..., k.

A função de sobrevivência (3.2) é uma função escada com degraus nos tempos de falha observados de tamanho d_j/n.

Para os dados do Exemplo 1, a estimativa da função de sobrevivência no ponto t=2 é dado por

$$\hat{S}(2) = \frac{\#(t > 2)}{n} = \frac{19}{20} = 0{,}95.$$

Note neste exemplo que $t_{(1)}$=2, $t_{(2)}$=4, n_1=20 e d_1=1. Desta forma, a estimativa de $S(t)$ para qualquer ponto no intervalo [2,4) é dada por

$$\hat{S}(t) = \frac{n_1 - d_1}{n} = \frac{19}{20} = 0{,}95, \text{ para } 2 \leq t < 4.$$

Realizando o mesmo cálculo, obtemos a estimativa de $S(t)$ para os demais intervalos. Os resultados podem ser vistos na Tabela 3.2 e Figura 3.2.

Tabela 3.2: Estimativas da função de sobrevivência para os dados Exemplo 1.

j	**Tempo** (t_j)	**Intervalo** $[t_j , t_{j+1})$	**nº de indivíduos sob risco** (n_j)	**nº de mortes no tempo** t_j (d_j)	$\hat{S}(t)$ $\left(\frac{n_j - d_j}{n}\right)$
0	0	[0,2)	20	0	1
1	2	[2,4)	20	1	0,95
2	4	[4, 5)	19	1	0,90
3	5	[5, 7)	18	1	0,85
4	7	[7, 11)	17	1	0,80
5	11	[11, 12)	16	2	0,70
6	12	[12, 13)	14	1	0,65
7	13	[13, 15)	13	1	0,60
8	15	[15, 19)	12	1	0,55
9	19	[19, 22)	11	1	0,50
10	22	[22, 23)	10	1	0,45
11	23	[23, 25)	9	1	0,40
12	25	[25, 29)	8	2	0,30
13	29	[29, 32)	6	1	0,25
14	32	[32, 37)	5	1	0,20
15	37	[37, 39)	4	1	0,15
16	39	[39, 44)	3	1	0,10
17	44	[44, 55)	2	1	0,05
18	55	[55, ∞)	1	1	0

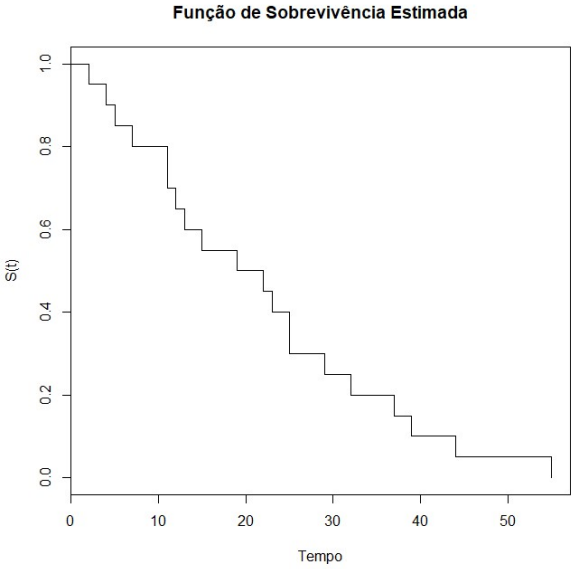


Figura 3.2: Função de Sobrevivência estimada para os dados do Exemplo 1.

3.2. Estimação na presença de censura

As estimativas apresentadas na Seção 3.1 não permitem a presença de observações censuradas, comuns em dados de sobrevivência. Será descrito nesta seção como as funções sobrevivência e de risco acumulado podem ser estimadas na presença de censura à direita. Serão apresentados os estimadores de Kaplan-Meier (Kaplan e Meier, 1958) e o estimador de Nelson-Aalen (Nelson, 1972; Aalen, 1978).

Para a obtenção dos estimadores citados, considere que é o $t_{(j)}$ o j-ésimo valor não censurado, distinto e ordenado da amostra, j=1,2,....k. Considerando que, $t_{(0)} = 0$ e $t_{(k+1)} = \infty$, tem-se que

$$0 = t_{(0)} < t_{(1)} < t_{(2)} < ... < t_{(k)} < t_{(k+1)} = \infty.$$

Note que $k \leq n$ (no caso em que a amostra observada não apresenta censuras e empates, tem-se que $k = n$), sendo n o tamanho da amostra.

3.2.1. Estimador de Kaplan Meier

O estimador não-paramétrico de Kaplan é um estimador da função de sobrevivência bastante popular e o mesmo se encontra presente nos principais softwares estatísticos. Este estimador é também conhecido na literatura como estimador limite-produto e é definido como (Kaplan e Meyer, 1958):

$$\hat{S}_{KM}(t) = \prod_{j:t_{(j)} \leq t} \left[1 - \frac{d_j}{n_j} \right]. \tag{3.4}$$

Aqui, n_j é o número de indivíduos que estão sob risco no o tempo $t_{(j)}$ (inclusive) e d_j representa o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=1,2,...,k.

O estimador de Kaplan-Meier (3.4) é uma adaptação da estimativa da função de sobrevivência empírica apresentado por (3.2). Sua intuição é baseada numa sequência de passos em que o próximo depende do anterior. Note que, para um indivíduo sobreviver até o tempo $t_{(j)}$, o mesmo teve que ter sobrevivido até o tempo $t_{(j-1)}$, isto é,

$$\begin{aligned}
S(t_{(j)}) &= P(T > t_{(j)}) \\
&= P(T > t_{(j-1)}, T > t_{(j)}) \\
&= P(T > t_{(j-1)}) P(T > t_{(j)} \mid T > t_{(j-1)}) \\
&= S(t_{(j-1)}) \left[1 - P(T \leq t_{(j)}) \mid T > t_{(j-1)}) \right] \\
&= S(t_{(j-1)}) \left[1 - P(t_{(j-1)} < T \leq t_{(j)}) \mid T > t_{(j-1)}) \right] \\
&= S(t_{(j-1)}) \left[1 - P(t_{(j-1)} < T \leq t_{(j)}) \mid T \geq t_{(j)}) \right], \quad j = 1,2,...,k.
\end{aligned} \tag{3.5}$$

A partir de (3.5) tem-se que a função de sobrevivência pode ser expressa por

$$S(t) = S(t_{(0)}) \prod_{j:t_{(j)} \leq t} \left[1 - P(t_{(j-1)} < T \leq t_{(j)}) \mid T \geq t_{(j)}) \right], \quad \text{para } t \geq 0. \tag{3.6}$$

Visto que $S(t_{(0)}) = 1$, tem-se que obter a estimativa de Kaplan-Meier da função de sobrevivência se reduz em estimar $q_j = P(t_{(j-1)} < T \leq t_{(j)}) \mid T \geq t_{(j)})$, isto é, a probabilidade de um indivíduo experimentar o evento de interesse no intervalo $(t_{(j-1)}, t_{(j)}]$, sabendo que o mesmo estava sob risco em $t_{(j)}$. O estimador de q_j, j = 1, 2, ..., k, é dado por

$$\hat{q}_j = \frac{\text{nº de indivíduos que experimentaram o evento de interesse em } (t_{(j-1)},\ t_{(j)}]}{\text{nº indivíduos sob risco em } t_{(j)}} = \frac{d_j}{n_j}. \qquad (3.7)$$

Logo, de (3.6) e (3.7), chega-se ao estimador de Kaplan-Meier para a função de sobrevivência.

Na ausência de censuras, o estimador de Kaplan-Meier da função de sobrevivência se reduz ao estimador empírico da função de sobrevivência expresso por (3.2) e (3.3).

Como visto em (2.9), a função de sobrevivência de uma variável contínua pode ser expressa por em termos da função taxa de falha acumulada por meio da relação: $S(t) = \exp\{-H(t)\}$. Assim, o estimador de Kaplan-Meier da função de risco acumulado é dado por

$$\hat{H}_{KM}(t) = -\log \hat{S}_{KM}(t).\,, \qquad (3.8)$$

em que $\hat{S}_{KM}(t)$ é o estimador de Kaplan-Meier da função de sobrevivência, expresso por (3.4).

Exemplo 2: *A Tabela 3.3 apresenta os tempos de sobrevivência (em meses) de mulheres submetidas a mastectomia simples ou radical. O início do estudo para cada mulher é caracterizado pelo mês da cirurgia, entretanto, o tempo final do estudo é comum para todas. São considerados censurados os tempos de sobrevivência das mulheres que no final do estudo não tinham falecido ou tinham abandonado o estudo. O tempo de sobrevivência de cada mulher foi classificado de acordo com o seguinte critério: seções dos tumores extraídos de cada mulher foram tratados com um marcador (lectina HPA) e classificados de acordo com o contraste apresentado (negativo ou positivo). O objetivo do estudo é verificar se o marcador pode ser utilizado para predizer o tempo de sobrevivência de mulheres que apresentam câncer de mama*.

Tabela 3.3: Tempos de sobrevivência (em meses) de mulheres submetidas a mastectomia simples ou radical.

Contraste	**Tempos de sobrevivência**
Negativo	23, 47, 69, 70^+, 71^+, 100^+, 101^+, 148, 181, 198^+, 208^+, 212^+, 224^+
Positivo	5, 8, 10, 13, 18, 24, 26, 26, 31, 35, 40, 41, 48, 50, 59, 61, 68, 71, 76^+, 105^+, 107^+, 109^+, 113, 116^+, 118, 143, 154^+, 162^+, 188^+, 212^+, 217^+, 225^+

+: observações censuradas

Tem-se, para os dados da Tabela 3.3 – Contraste Negativo, que $t_{(1)}$=23, $t_{(2)}$=47, n_1=13 e d_1=1. Desta forma, a estimativa de Kaplan-Meier para $S(t)$ para qualquer ponto no intervalo [23, 47) é dada por

$$\hat{S}_{KM}(t) = 1 - \frac{d_1}{n_1} = 1 - \frac{1}{13} = 0{,}9231 \quad \text{para } 23 \le t < 47.$$

O valor de $\hat{S}_{KM}(t)$ para o próximo intervalo, em que $t_{(2)}$=47, $t_{(3)}$=69, n_2=12 e d_2=1, é dado por

$$\hat{S}_{KM}(t) = \left[1 - \frac{d_1}{n_1}\right]\left[1 - \frac{d_2}{n_2}\right] = \left[1 - \frac{1}{13}\right]\left[1 - \frac{1}{12}\right] = 0{,}8462 \quad \text{para } 47 \le t < 69.$$

Realizando o mesmo cálculo, pode-se obter a estimativa de Kaplan-Meier para $S(t)$ para os demais intervalos e também para o grupo Contraste Positivo. Esses resultados, assim como a estimativa da função de risco acumulado (3.8), podem ser vistos na Tabela 3.4 e Figura 3.3.

Tabela 3.4: Estimativas de Kaplan-Meier para os dados do Exemplo 2.

Contraste Negativo						Contraste Positivo					
Tempo $(t_{(j)})$	**Intervalo** $[t_{(j)}, t_{(j+1)})$	d_j	n_j	$\hat{S}_{KM}(t)$	$\hat{H}_{KM}(t_{(j)})$	**Tempo** $(t_{(j)})$	**Intervalo** $[t_{(j)}, t_{(j+1)})$	d_j	n_j	$\hat{S}_{KM}(t)$	$\hat{H}_{KM}(t)$
0	[0, 23)	0	13	1	0	0	[0, 5)	0	32	1	0
23	[23, 47)	1	13	0,9231	0,0800	5	[5, 8)	1	32	0,9688	0,0317
47	[47, 69)	1	12	0,8462	0,1671	8	[8, 10)	1	31	0,9375	0,0645
69	[69, 148)	1	11	0,7692	0,2624	10	[10, 13)	1	30	0,9063	0,0984
148	[148, 181)	1	6	0,6410	0,4447	13	[13, 18)	1	29	0,8750	0,1335
181	[181, ∞)	1	5	0,5128	0,6678	18	[18, 24)	1	28	0,8438	0,1699
						24	[24, 26)	1	27	0,8125	0,2076
						26	[26, 31)	2	26	0,7500	0,2877
						31	[31, 35)	1	24	0,7188	0,3302
						35	[35, 40)	1	23	0,6875	0,3747
						40	[40, 41)	1	22	0,6563	0,4212
						41	[41, 48)	1	21	0,6250	0,4700
						48	[48, 50)	1	20	0,5938	0,5213
						50	[50, 59)	1	19	0,5625	0,5754
						59	[59, 61)	1	18	0,5313	0,6325
						61	[61, 68)	1	17	0,5000	0,6931
						68	[68, 71)	1	16	0,4688	0,7577
						71	[71, 113)	1	15	0,4375	0,8267
						113	[113, 118)	1	10	0,3938	0,9320
						118	[118, 143)	1	8	0,3445	1,0656
						143	[143, ∞)	1	7	0,2953	1,2197

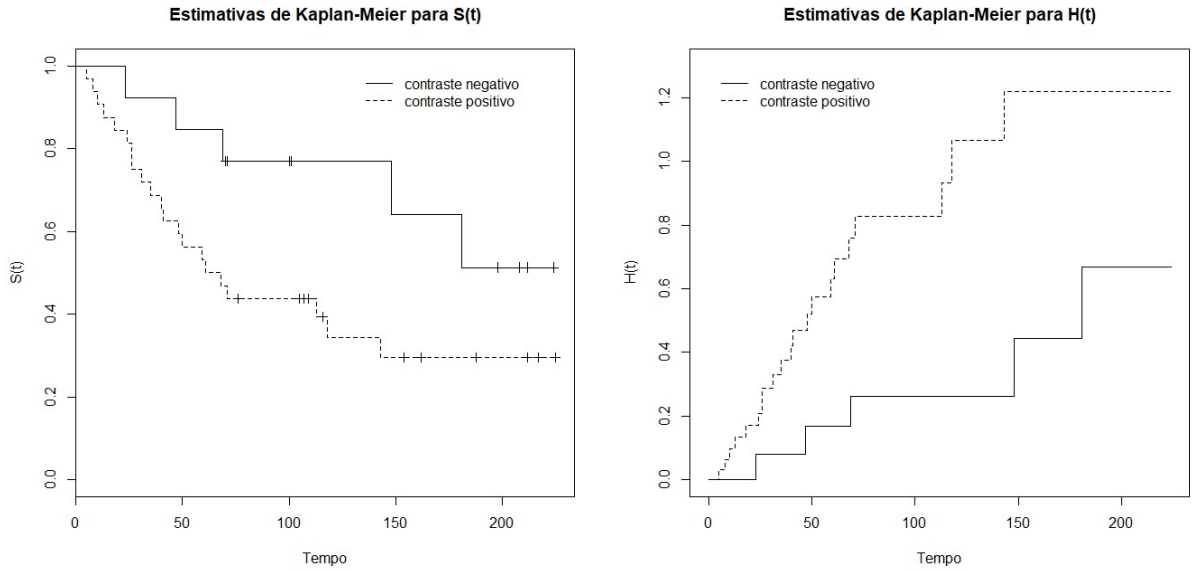


Figura 3.3: Estimativas de Kaplan-Meier para as funções de sobrevivência e risco acumulado para os dados do Exemplo 2.

3.2.1.1. Intervalos de confiança de *S*(*t*)

O estimador de Kaplan-Meier, como qualquer outro estimador, está sujeito a variações amostrais que devem ser descritas em temos de sua variância. Serão descritos aqui três procedimentos de estimação intervalar referente ao estimador de Kaplan-Meier da função de sobrevivência: i) o primeiro procedimento usa como base a função de sobrevivência para a estimativa intervalar, resultando em um intervalo simétrico; ii) o segundo procedimento apresentado obtém o intervalo de confiança a partir da função de risco acumulado; e iii) o último procedimento se baseia no logaritmo da função de risco acumulado, por meio da transformação "log-log" da função de sobrevivência.

Intervalo simétrico

A variância assintótica do estimador de Kaplan-Meier da Função de Sobrevivência pode ser estimada pela Fórmula de Greenwood (Greenwood, 1926):

$$\hat{Var}\left(\hat{S}_{KM}(t)\right) = \hat{S}^2_{KM}(t) \sum_{j:t_{(j)}\le t} \frac{d_j}{n_j(n_j - d_j)}. \qquad (3.9)$$

Aqui, $\hat{S}_{KM}(t)$ é o estimador de Kaplan-Meier da função de sobrevivência, expresso por (3.4), n_j é o número de indivíduos que estão sob risco no o tempo $t_{(j)}$ (inclusive) e d_j representa o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=1,2,...,k.

Com o a estimativa da variância do estimador de Kaplan-Meier para a função de sobrevivência, é possível obter um intervalo de confiança assintótico para o verdadeiro valor de $S(t)$. Este intervalo de confiança é obtido pelo fato de $\hat{S}_{KM}(t)$ convergir assintoticamente para a uma distribuição normal com média $S(t)$ e variância estimada por (3.9) (Colosimo e Giolo, 2006).

Desta forma, um intervalo de $100(1-\alpha)\%$ de confiança para $S(t)$, para todo $t \ge 0$ fixo, é dado por:

$$\underset{100(1-\alpha)\%}{I.C.}\left[S(t)\right]: \left[\hat{S}_{KM}(t) - z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{S}_{KM}(t)\right)}\ ;\ \hat{S}_{KM}(t) + z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{S}_{KM}(t)\right)}\right], \qquad (3.10)$$

em que $\hat{S}_{KM}(t)$ é o estimador de Kaplan-Meier da função de sobrevivência, expresso por (3.4) e $z_{(1-\alpha/2)}$ é o quantil $\left(1-\frac{\alpha}{2}\right)$ da distribuição normal padrão.

Para os dados da Tabela 3.3 – Contraste Negativo tem-se, por exemplo, que variância da estimativa de Kaplan-Meier da função de sobrevivência em t=47 é estimada por

$$\hat{Var}\left(\hat{S}_{KM}(47)\right) = (0{,}8462)^2\left(\frac{1}{13(12)} + \frac{1}{12(11)}\right) = 0{,}01001.$$

Assim, um intervalo de 95% de confiança para $S(t)$ em t=47 é dado por:

$$\underset{95\%}{I.C.}\left[S(47)\right]: \left[\hat{S}_{KM}(47) - z_{(0{,}975)}\sqrt{\hat{Var}\left(\hat{S}_{KM}(47)\right)}\ ;\ \hat{S}_{KM}(47) + z_{(0{,}975)}\sqrt{\hat{Var}\left(\hat{S}_{KM}(47)\right)}\right]$$

$$\underset{95\%}{I.C.}\left[S(47)\right]: \left[0{,}8462 - 1{,}96\sqrt{0{,}01001}\ ;\ 0{,}8462 - 1{,}96\sqrt{0{,}01001}\right]$$

$$\underset{95\%}{I.C.}\left[S(47)\right]: \left[0{,}650\ ;1\right].$$

Note que no cálculo do intervalo de confiança simétrico o limite superior do mesmo ultrapassou o valor 1 e, portanto, o mesmo foi truncado em 1. Realizando o mesmo cálculo, pode-se obter os intervalos de confiança de $S(t)$ para os demais valores de t. Os intervalos de confiança para a função de sobrevivência do grupo Contraste Negativo são apresentados na Tabela 3.5 e Figura 3.4.

Nota: O R fornece os intervalos simétricos por meio do comando "survfit" da biblioteca *survival*. Para obter os intervalos simétricos para a função de sobrevivência (expressão 3.10), basta usar a extensão *conf.type="plain"*.
Ex: km<- survfit(Surv(tempo,censura)~1,type="kaplan-meier",conf.type= "plain")

Tabela 3.5: Estimativas de Kaplan-Meier e intervalos de confiança simétricos para a função de sobrevivência. Dados do Exemplo 2 – Contraste Negativo.

j	**Tempo** $t_{(j)}$	**Intervalo** $[t_{(j)}, t_{(j+1)})$	d_j	n_j	$\hat{S}_{KM}(t)$	$Vâr\left(\hat{S}_{KM}(t)\right)$	I.C. (95%) simétrico
0	0	[0, 23)	0	13	1	---	---
1	23	[23, 47)	1	13	0,9231	0,0055	[0,778 ; 1]
2	47	[47, 69)	1	12	0,8462	0,0100	[0,650 ; 1]
3	69	[69, 148)	1	11	0,7692	0,0137	[0,540 ; 0,998]
4	148	[148, 181)	1	6	0,6410	0,0232	[0,343 ; 0,939]
5	181	[181, ∞)	1	5	0,5128	0,0280	[0,185 ; 0,841]

A função de sobrevivência é estimada com precisão decrescente (variância crescente) na medida em que t cresce. Isto ocorre devido ao fato do número de indivíduos em risco diminuir ao longo do tempo (os indivíduos vão experimentando o evento de interesse ou vão censurando com o avanço do tempo) e pequenas variações nas observações podem acarretar grandes variações no cálculo da estimativa de $S(t)$.

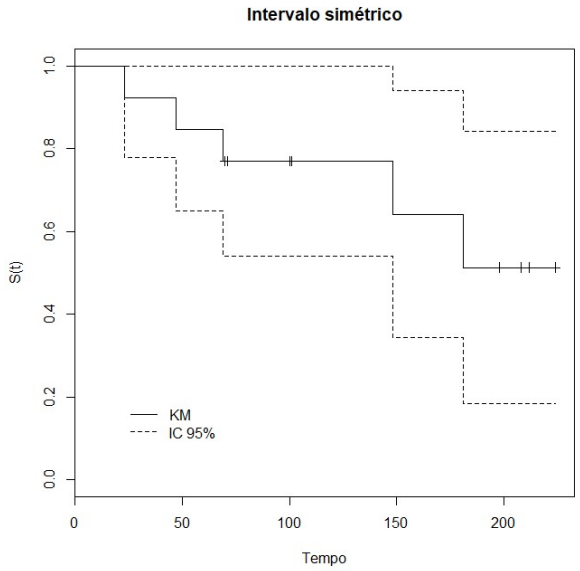


Figura 3.4: Estimativa de Kaplan-Meier para as funções de sobrevivência e seu respectivo intervalo de confiança simétrico para os dados do Exemplo 2 – Contraste Negativo.

Intervalo com base na função taxa de falha acumulada (log)

Uma forma alternativa de construir um intervalo de confiança para a função de sobrevivência é construí-lo com base na estimativa do risco acumulado. Como visto em (2.9), a função de risco acumulado pode ser expressa em termos da função de sobrevivência por meio da relação: $H(t) = -\log S(t)$. Desta forma, a estimativa de da variância do estimador de Kaplan-Meier para a função de risco acumulado (expresso por 3.8), é dada por:

$$\hat{Var}\left(\hat{H}_{KM}(t)\right) = \sum_{j:t_{(j)}\le t} \frac{d_j}{n_j(n_j - d_j)}. \qquad (3.11)$$

Aqui, n_j é o número de indivíduos que estão sob risco no o tempo $t_{(j)}$ (inclusive) e d_j representa o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=1,2,...,k.

Assim, para todo t fixo, um intervalo $100(1-\alpha)\%$ de confiança para $H(t)$ é dado por:

$$\underset{100(1-\alpha)\%}{I.C.}\left[H(t)\right]: \left[\hat{H}_{KM}(t) - z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{H}_{KM}(t)\right)}\,;\, \hat{H}_{KM}(t) + z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{H}_{KM}(t)\right)}\right], \qquad (3.12)$$

que resulta no seguinte intervalo de $100(1-\alpha)\%$ de confiança para $S(t)$:

$$\underset{100(1-\alpha)\%}{I.C.}\left[S(t)\right]: \left[e^{-z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{H}_{KM}(t)\right)}}\hat{S}_{KM}(t)\;;\; e^{z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{H}_{KM}(t)\right)}}\hat{S}_{KM}(t)\right], \qquad (3.13)$$

em que $\hat{S}_{KM}(t)$ é o estimador de Kaplan-Meier da função de sobrevivência, expresso por (3.4), $\hat{H}_{KM}(t)$ é o estimador de Kaplan-Meier da função de risco acumulado expresso por (3.8) e $z_{(1-\alpha/2)}$ é o quantil $\left(1-\frac{\alpha}{2}\right)$ da distribuição normal padrão.

Para os dados da Tabela 3.3 – Contraste Negativo tem-se, por exemplo, que variância da estimativa de Kaplan-Meier da função $H(t)$ em t=47 é estimada por

$$\hat{Var}\left(\hat{H}_{KM}(47)\right) = \left(\frac{1}{13(12)} + \frac{1}{12(11)}\right) = 0{,}01399$$

Assim, um intervalo de 95% de confiança para S(t) em t=47 é dado por:

$$\underset{95\%}{I.C.}\left[S(t)\right]: \left[e^{-z_{(0,975)}\sqrt{\hat{Var}\left(\hat{H}_{KM}(t)\right)}}\hat{S}_{KM}(t)\;;\; e^{z_{(0,975)}\sqrt{\hat{Var}\left(\hat{H}_{KM}(t)\right)}}\hat{S}_{KM}(t)\right]$$

$$\underset{95\%}{I.C.}\left[S(t)\right]: \left[e^{-1{,}96\sqrt{0{,}01399}}\,0{,}8462\;;\; e^{-1{,}96\sqrt{0{,}01399}}\,0{,}8462\right]$$

$$\underset{95\%}{I.C.}\left[S(t)\right]: \left[0{,}671\;;\; 1\right]$$

Note que no cálculo do intervalo de confiança (3.13), o limite superior do mesmo ultrapassou o valor 1 e, portanto, o mesmo foi truncado em 1. Realizando o mesmo cálculo, pode-se obter os intervalos de confiança de $S(t)$ para os demais valores de t. Os intervalos de confiança (3.13) para a função de sobrevivência do grupo Contraste Negativo são apresentados na Tabela 3.6 e Figura 3.5.

Nota: O R fornece os intervalos com base em $H(t)$ por meio do comando "survfit" da biblioteca *survival*. Para obter os intervalos de confiança (3.13) para a função de sobrevivência, basta usar a extensão *conf.type="log"*.
Obs: O comando "survfit" tem como padrão o intervalo tipo "log", sendo neste caso, desnecessário usar a extensão acima.
Ex: km<- survfit(Surv(tempo,censura)~1,type="kaplan-meier",conf.type= "log")
Ou

```
km<- survfit(Surv(tempo,censura)~1,type="kaplan-meier")
```

Tabela 3.6: Estimativas de Kaplan-Meier e intervalos de confiança da transformação "log" para a função de sobrevivência. Dados do Exemplo 2 – Contraste Negativo.

j	**Tempo** $t_{(j)}$	**Intervalo** $[t_{(j)}, t_{(j+1)})$	d_j	n_j	$\hat{S}_{KM}(t)$	$\hat{H}_{KM}(t)$	$\hat{Var}\left(\hat{H}_{KM}(t)\right)$	I.C. 95% p/ $S(t)$ "log"
0	0	[0, 23)	0	13	1	0	---	---
1	23	[23, 47)	1	13	0,9231	0,0800	0,0064	[0,789 ; 1]
2	47	[47, 69)	1	12	0,8462	0,1671	0,0140	[0,671 ; 1]
3	69	[69, 148)	1	11	0,7692	0,2624	0,0231	[0,571 ; 1]
4	148	[148, 181)	1	6	0,6410	0,4447	0,0564	[0,402 ; 1]
5	181	[181, ∞)	1	5	0,5128	0,6678	0,1064	[0,271 ; 0,972]

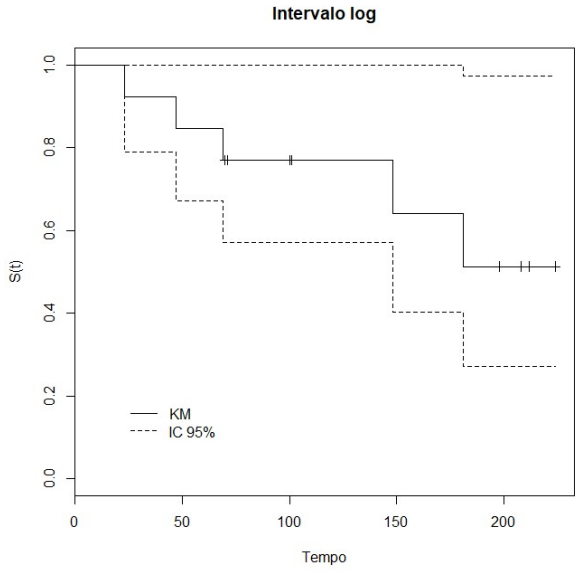


Figura 3.5: Estimativa de Kaplan-Meier para as funções de sobrevivência e seu respectivo intervalo de confiança da transformação "log" para os dados do Exemplo 2 – Contraste Negativo.

Intervalo log-log

Note que o limite superior dos intervalos de confiança (3.10) e (3.13) foram truncados em 1, visto que a função de sobrevivência é limitada no intervalo [0,1]. Este é um inconveniente em considerar a normalidade assintótica do estimador de $S(t)$ e $H(t)$. O problema ocorre para valores extremos ou próximos de zero de t. Quando a estimativa da função de sobrevivência está próxima de 1 ou 0 o intervalo de confiança simétrico da função de sobrevivência pode apresentar limite inferior negativo ou limite superior maior do que 1. Neste último caso, o intervalo da transformação log (baseado na função de risco acumulado) pode também apresentar limite superior maior do que 1. Para evitar este problema, um procedimento alternativo consiste em aplicar nova transformação na função de sobrevivência.

Kalbfleisch e Prentice (2002) sugeriram aplicar o logaritmo na função de risco acumulado, ou a transformação "log-log" na função de sobrevivência de Kaplan-Meier, isto é:

$$\hat{U}_{KM}(t) = \log\left\{-\log \hat{S}_{KM}(t)\right\}. \tag{3.14}$$

A variância de $\hat{U}_{KM}(t)$ é dada por:

$$\hat{Var}\left(\hat{U}_{KM}(t)\right) = \frac{1}{\left[\log \hat{S}_{KM}(t)\right]^2} \sum_{j:t_{(j)}\le t} \frac{d_j}{n_j(n_j - d_j)}, \tag{3.15}$$

Aqui, $\hat{S}_{KM}(t)$ é o estimador de Kaplan-Meier da função de sobrevivência, expresso por (3.4), n_j é o número de indivíduos que estão sob risco no o tempo $t_{(j)}$ (inclusive) e d_j representa o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=1,2,...,k.

Assim, para todo t fixo, um intervalo 100(1 – α)% de confiança para $U(t)$ é dado por:

$$\underset{100(1-\alpha)\%}{I.C.}\left[U(t)\right]: \left[\hat{U}_{KM}(t) - z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{U}_{KM}(t)\right)}\ ;\ \hat{U}_{KM}(t) + z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{U}_{KM}(t)\right)}\right], \tag{3.16}$$

que resulta no seguinte intervalo de 100(1 – α)% de confiança para $S(t)$:

$$\underset{100(1-\alpha)\%}{I.C.}\left[S(t)\right]: \left[\left(\hat{S}_{KM}(t)\right)^{\exp\left\{z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{U}_{KM}(t)\right)}\right\}}\ ;\ \left(\hat{S}_{KM}(t)\right)^{\exp\left\{-z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{U}_{KM}(t)\right)}\right\}}\right], \tag{3.17}$$

em que $\hat{S}_{KM}(t)$ é o estimador de Kaplan-Meier da função de sobrevivência, expresso por (3.4), $\hat{U}_{KM}(t)$ é o estimador de Kaplan-Meier da transformação log-log da função de sobrevivência expresso por (3.14) e $z_{(1-\alpha/2)}$ é o quantil $\left(1-\frac{\alpha}{2}\right)$ da distribuição normal padrão.

Para os dados da Tabela 3.3 – Contraste Negativo tem-se, por exemplo, que variância de $\hat{U}_{KM}(t)$ em t=47 é estimada por

$$\hat{Var}\left(\hat{U}_{KM}(47)\right) = \frac{1}{\left[\log(0{,}8462)\right]^2}\left(\frac{1}{13(12)} + \frac{1}{12(11)}\right) = 0{,}5015\,.$$

Assim, um intervalo de 95% de confiança para S(t) em t=14 é dado por

$$I.C._{95\%}[S(47)]: \left[\left(\hat{S}_{KM}(47)\right)^{\exp\left\{z_{(0,975)}\sqrt{\hat{Var}\left(\hat{U}_{KM}(47)\right)}\right\}} ; \left(\hat{S}_{KM}(47)\right)^{\exp\left\{-z_{(0,975)}\sqrt{\hat{Var}\left(\hat{U}_{KM}(47)\right)}\right\}}\right]$$

$$I.C._{95\%}[S(47)]: \left[(0{,}8462)^{\exp\left\{1{,}96\sqrt{0{,}5015}\right\}} ; (0{,}8462)^{\exp\left\{-1{,}96\sqrt{0{,}5015}\right\}}\right]$$

$$I.C._{95\%}[S(47)]: [0{,}512 ; 0{,}959]$$

Realizando o mesmo cálculo, pode-se obter os intervalos de confiança de $S(t)$ para os demais valores de t. Os intervalos de confiança para a função de sobrevivência do grupo Contraste Negativo são apresentados na Tabela 3.7 e Figura 3.6.

Nota: O R fornece os intervalos log-log por meio do comando "survfit" da biblioteca *survival*. Para obter os intervalos log-log para a função de sobrevivência (expressão 3.17), basta usar a extensão *conf.type="log-log"*.
Ex: km<- survfit(Surv(tempo,censura)~1,type="kaplan-meier",conf.type= "log-log")

Tabela 3.7: Estimativas de Kaplan-Meier e intervalos de confiança log-log para a função de sobrevivência. Dados do Exemplo 2 – Contraste Negativo.

j	Tempo $t_{(j)}$	Intervalo $[t_{(j)}, t_{(j+1)})$	d_j	n_j	$\hat{S}_{KM}(t)$	$\hat{Var}\left(\hat{U}_{KM}(t)\right)$	I.C. (95%) "log-log"
0	0	[0, 23)	0	13	1	---	---
1	23	[23, 47)	1	13	0,9231	1,0012	[0,566 ; 0,989]
2	47	[47, 69)	1	12	0,8462	0,5015	[0,512 ; 0,959]
3	69	[69, 148)	1	11	0,7692	0,3351	[0,442 ; 0,919]
4	148	[148, 181)	1	6	0,6410	0,2853	[0,282 ; 0,855]
5	181	[181, ∞)	1	5	0,5128	0,2385	[0,176 ; 0,774]

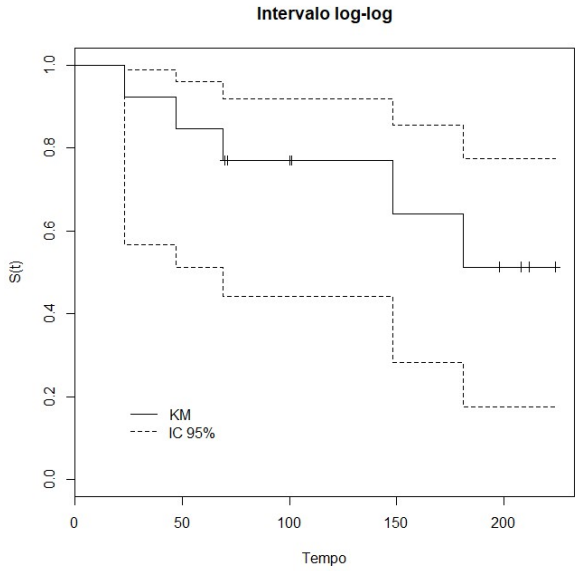


Figura 3.6: Estimativa de Kaplan-Meier para as funções de sobrevivência e seu respectivo intervalo de confiança log-log para os dados do Exemplo 2 – Contraste Negativo.

3.2.2. Estimador de Nelson-Aalen

O estimador de Nelson-Aalen da função de sobrevivência se baseia na relação entre a função de sobrevivência e a função de risco acumulado.

Nelson (1972) propôs um estimador para a função de risco acumulado, $H(t)$, cujas propriedades assintóticas foram provadas por Aalen (1978). Este estimador é denominado na literatura por Nelson-Aalen e é definido por

$$\hat{H}_{NA}(t) = \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j}, \tag{3.18}$$

Aqui, n_j é o número de indivíduos que estão sob risco no o tempo $t_{(j)}$ (inclusive) e d_j representa o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=1,2,...,k.

Como visto em (2.9), a função de sobrevivência de uma variável contínua pode ser expressa por em termos da função taxa de falha acumulada por meio da relação: $S(t) = \exp\{-H(t)\}$. Assim, o estimador de Nelson-Aalen da função de sobrevivência é dado por

$$\hat{S}_{NA}(t) = \exp\left\{-\hat{H}_{NA}(t)\right\} = \exp\left\{-\sum_{j:t_{(j)} \le t} \frac{d_j}{n_j}\right\}. \tag{3.19}$$

Tem-se, para os dados da Tabela 3.3 – Contraste Negativo, que $t_{(1)}$=23, $t_{(2)}$=47, n_1=13 e d_1=1. Desta forma, as estimativas de Nelson-Aalen para $H(t)$ e $S(t)$ para qualquer ponto no intervalo [23, 44) são dadas, respectivamente, por:

$$\hat{H}_{NA}(t) = \frac{d_1}{n_1} = \frac{1}{13} = 0{,}0769, \text{ para } 23 \le t < 47.$$

e

$$\hat{S}_{NA}(t) = \exp\left\{-\frac{1}{13}\right\} = 0{,}9260, \text{ para } 23 \le t < 47.$$

Os valores de $\hat{H}_{NA}(t)$ e $\hat{S}_{NA}(t)$ para próximo intervalo, em que $t_{(2)}$=47, $t_{(3)}$=69, n_2=12 e d_2=1 são dados, respectivamente por:

$$\hat{H}_{NA}(t) = \frac{d_1}{n_1} + \frac{d_2}{n_2} = \frac{1}{13} + \frac{1}{12} = 0{,}1603, \text{ para } 47 \le t < 69.$$

e

$$\hat{S}_{NA}(t) = \exp\left\{-\frac{1}{13} - \frac{1}{12}\right\} = 0{,}8519, \text{ para } 47 \le t < 69.$$

Realizando o mesmo cálculo, pode-se obter a estimativa de Nelson-Aalen para $H(t)$ e $S(t)$ para os demais intervalos e também para o grupo Contraste Positivo. Esses resultados, assim como a estimativa da função de risco acumulado (3.8), podem ser vistos na Tabela 3.8 e Figura 3.7.

Tabela 3.8: Estimativas de Nelson-Aalen para os dados do Exemplo 2.

Contraste Negativo						Contraste Positivo					
Tempo ($t_{(j)}$)	**Intervalo $[t_{(j)}, t_{(j+1)})$**	d_j	n_j	$\hat{S}_{NA}(t)$	$\hat{H}_{NA}(t)$	**Tempo ($t_{(j)}$)**	**Intervalo $[t_{(j)}, t_{(j+1)})$**	d_j	n_j	$\hat{S}_{NA}(t)$	$\hat{H}_{NA}(t)$
0	[0, 23)	0	13	1	0	0	[0, 5)	0	32	1	0
23	[23, 47)	1	13	0,9260	0,0769	5	[5, 8)	1	32	0,9692	0,0312
47	[47, 69)	1	12	0,8519	0,1603	8	[8, 10)	1	31	0,9385	0,0635
69	[69, 148)	1	11	0,7779	0,2512	10	[10, 13)	1	30	0,9077	0,0968
148	[148, 181)	1	6	0,6585	0,4178	13	[13, 18)	1	29	0,8769	0,1313
181	[181, ∞)	1	5	0,5391	0,6178	18	[18, 24)	1	28	0,8462	0,1670
						24	[24, 26)	1	27	0,8154	0,2041
						26	[26, 31)	2	26	0,7550	0,2810
						31	[31, 35)	1	24	0,7242	0,3227
						35	[35, 40)	1	23	0,6934	0,3661
						40	[40, 41)	1	22	0,6626	0,4116
						41	[41, 48)	1	21	0,6318	0,4592
						48	[48, 50)	1	20	0,6010	0,5092
						50	[50, 59)	1	19	0,5702	0,5618
						59	[59, 61)	1	18	0,5393	0,6174
						61	[61, 68)	1	17	0,5085	0,6762
						68	[68, 71)	1	16	0,4777	0,7387
						71	[71, 113)	1	15	0,4469	0,8054
						113	[113, 118)	1	10	0,4044	0,9054
						118	[118, 143)	1	8	0,3569	1,0304
						143	[143, ∞)	1	7	0,3094	1,1733

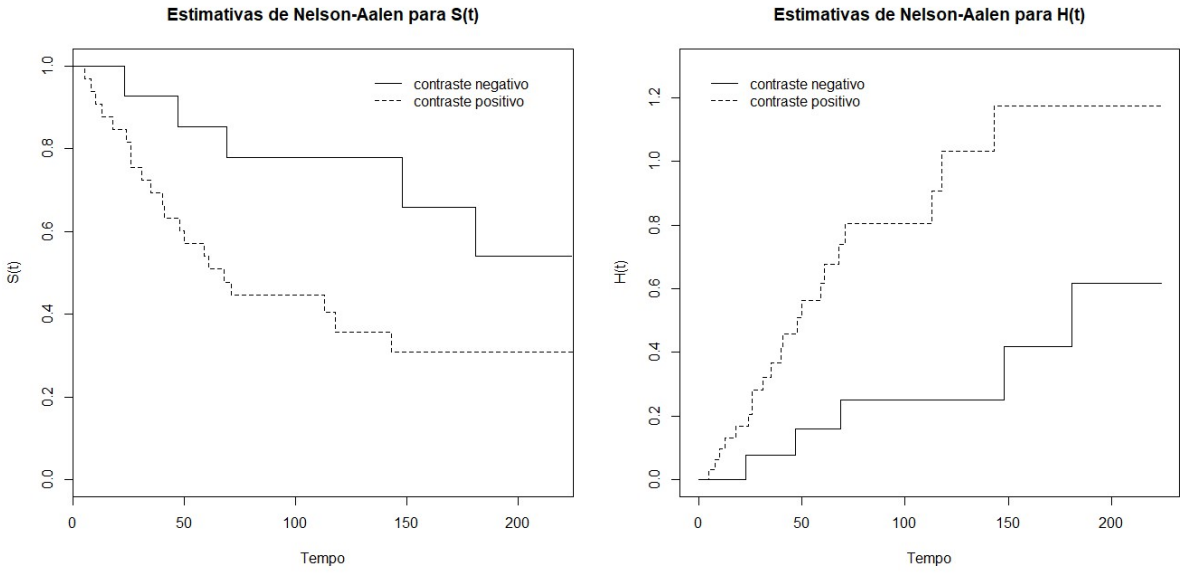


Figura 3.7: Estimativas de Nelson-Aalen para as funções de sobrevivência e risco acumulado para os dados do Exemplo 2.

3.2.2.1. Intervalos de confiança de *S*(*t*)

Assim como na Seção 3.2.1.2, serão descritos aqui três procedimentos de estimação intervalar referente ao estimador de Nelson-Aalen: i) o intervalo simétrico; ii) o intervalo tipo log e; o intervalo log-log.

Intervalo simétrico

A variância assintótica do estimador da função de sobrevivência Nelson-Aalen é dada por (Aalen e Johansen, 1978):

$$\hat{Var}\left(\hat{S}_{NA}(t)\right) = \hat{S}^2_{NA}(t) \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j^2}. \tag{3.20}$$

Aqui, $\hat{S}_{NA}(t)$ é o estimador de Nelson-Aalen da função de sobrevivência, expresso por (3.19), n_j é o número de indivíduos que estão sob risco no o tempo $t_{(j)}$ (inclusive) e d_j representa o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, $j = 1, 2, ..., k$.

Assim, para todo t fixo, um intervalo de $100(1 - \alpha)\%$ de confiança para $S(t)$, é dado por:

$$\underset{100(1-\alpha)\%}{I.C.}\left[S(t)\right]: \left[\hat{S}_{NA}(t) - z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{S}_{NA}(t)\right)}\,;\, \hat{S}_{NA}(t) + z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{S}_{NA}(t)\right)}\right], \tag{3.21}$$

em $\hat{S}_{NA}(t)$ é o estimador de Nelson-Aalen da função sobrevivência expresso por (3.19) e $z_{(1-\alpha/2)}$ é o quantil $\left(1 - \frac{\alpha}{2}\right)$ da distribuição normal padrão.

Como visto em (2.9), a função de sobrevivência de uma variável contínua pode ser expressa por em termos da função taxa de falha acumulada por meio da relação: $S(t) = \exp\{-H(t)\}$. Assim, o estimador de Nelson-Aalen da função de sobrevivência é dado por

$$\hat{S}_{NA}(t) = \exp\left\{-\hat{H}_{NA}(t)\right\} = \exp\left\{-\sum_{j:t_{(j)} \le t} \frac{d_j}{n_j}\right\}. \tag{3.22}$$

Para os dados da Tabela 3.3 – Contraste Negativo tem-se, por exemplo, que variância da estimativa de Nelson-Aalen da função de sobrevivência em t=47 é estimada por

$$\hat{Var}\left(\hat{S}_{NA}(47)\right) = (0{,}8519)^2\left(\frac{1}{13^2} + \frac{1}{12^2}\right) = 0{,}0093\,.$$

Assim, um intervalo de 95% de confiança para $S(t)$ em t=47 é dado por:

$$\underset{95\%}{I.C.}\left[S(47)\right]: \left[\hat{S}_{NA}(47) - z_{(0{,}975)}\sqrt{\hat{Var}\left(\hat{S}_{NA}(47)\right)}\,;\, \hat{S}_{NA}(47) + z_{(0{,}975)}\sqrt{\hat{Var}\left(\hat{S}_{NA}(47)\right)}\right]$$

$$\underset{95\%}{I.C.}\left[S(47)\right]: \left[0{,}8519 - 1{,}96\sqrt{0{,}0093}\,;\, 0{,}88519 - 1{,}96\sqrt{0{,}0093}\right]$$

$$\underset{95\%}{I.C.}\left[S(47)\right]: \left[0{,}663;1\right].$$

Note que no cálculo do intervalo de confiança simétrico o limite superior do mesmo ultrapassou o valor 1 e, portanto, o mesmo foi truncado em 1. Realizando o mesmo cálculo, pode-se

obter os intervalos de confiança de $S(t)$ para os demais valores de t. Os intervalos de confiança para a função de sobrevivência do grupo Contraste Negativo são apresentados na Tabela 3.9 e Figura 3.8.

Nota: O R fornece os intervalos simétricos por meio do comando "survfit" da biblioteca *survival*. Para obter os intervalos simétricos para a função de sobrevivência (expressão 3.21), basta usar a extensão *conf.type="plain"*.
Ex: na<- survfit(coxph(Surv(tempo,censura)~1,method="breslow"),conf.type= "plain")

Tabela 3.9: Estimativas de Nelson-Aalen e intervalos de confiança simétricos para a função de sobrevivência. Dados do Exemplo 2 – Contraste Negativo.

j	Tempo $t_{(j)}$	Intervalo $[t_{(j)}, t_{(j+1)})$	d_j	n_j	$\hat{S}_{NA}(t)$	$Vâr\left(\hat{S}_{NA}(t)\right)$	I.C. (95%) simétrico
0	0	[0, 23)	0	13	1	---	---
1	23	[23, 47)	1	13	0,9260	0,0051	[0,786 ; 1]
2	47	[47, 69)	1	12	0,8519	0,0093	[0,663 ; 1]
3	69	[69, 148)	1	11	0,7779	0,0128	[0,556 ; 0,999]
4	148	[148, 181)	1	6	0,6585	0,0212	[0,373 ; 0,944]
5	181	[181, ∞)	1	5	0,5391	0,0258	[0,224 ; 0,854]

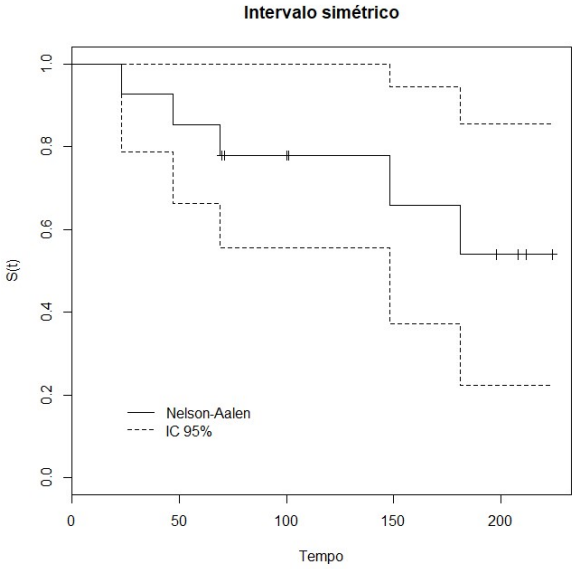


Figura 3.8: Estimativa de Nelson-Aalen para as funções de sobrevivência e seu respectivo intervalo de confiança simétrico para os dados do Exemplo 2 – Contraste Negativo.

Intervalo com base na função taxa de falha acumulada (log)

Como visto em (2.9), a função de risco acumulado pode ser expressa em termos da função de sobrevivência por meio da relação: $H(t) = -\log S(t)$.

A variância assintótica do estimador da função de risco acumulado de Nelson-Aalen é dada por (Aalen, 1978):

$$\hat{Var}\left(\hat{H}_{NA}(t)\right) = \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j^2}. \qquad (3.23)$$

Aqui, $\hat{H}_{NA}(t)$ é o estimador de Nelson-Aalen da função de risco acumulado, expresso por (3.18), n_j é o número de indivíduos que estão sob risco no o tempo $t_{(j)}$ (inclusive) e d_j representa o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=1,2,...,k.

Assim, para todo t fixo, um intervalo de $100(1-\alpha)\%$ de confiança para $H(t)$, é dado por:

$$\underset{100(1-\alpha)\%}{I.C.}\left[H(t)\right]: \left[\hat{H}_{NA}(t) - z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{H}_{NA}(t)\right)}\ ;\ \hat{H}_{NA}(t) - z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{H}_{NA}(t)\right)}\right] \qquad (3.24)$$

que resulta no seguinte intervalo de $100(1-\alpha)\%$ de confiança para $S(t)$:

$$\underset{100(1-\alpha)\%}{I.C.}\left[S(t)\right]: \left[e^{-z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{H}_{NA}(t)\right)}}\hat{S}_{NA}(t)\ ;\ e^{z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{H}_{NA}(t)\right)}}\hat{S}_{NA}(t)\right], \qquad (3.25)$$

em que $\hat{S}_{NA}(t)$ é o estimador de Nelson-Aalen da função de sobrevivência, expresso por (3.19), $\hat{H}_{NA}(t)$ é o estimador de Nelson-Aalen da função de risco acumulado expresso por (3.18) e $z_{(1-\alpha/2)}$ é o quantil $\left(1-\frac{\alpha}{2}\right)$ da distribuição normal padrão.

Para os dados da Tabela 3.3 – Contraste Negativo tem-se, por exemplo, que variância da estimativa de Nelson-Aalen da função $H(t)$ em t=47 é estimada por

$$\hat{Var}\left(\hat{H}_{NA}(47)\right) = \left(\frac{1}{13^2} + \frac{1}{12^2}\right) = 0{,}0129\,.$$

Assim, um intervalo de 95% de confiança para S(t) em t=47 é dado por:

$$\underset{95\%}{I.C.}\left[S(t)\right]: \left[e^{-z_{(0,975)}\sqrt{\hat{Var}\left(\hat{H}_{NA}(t)\right)}}\hat{S}_{NA}(t)\ ;\ e^{z_{(0,975)}\sqrt{\hat{Var}\left(\hat{H}_{NA}(t)\right)}}\hat{S}_{NA}(t)\right]$$

$$\underset{95\%}{I.C.}\left[S(t)\right]: \left[e^{-1{,}96\sqrt{0{,}0129}}\,0{,}8519\ ;\ e^{1{,}96\sqrt{0{,}0129}}\,0{,}8519\right]$$

$$\underset{95\%}{I.C.}\left[S(t)\right]: \left[0{,}682\ ;\ 1\right].$$

Note que no cálculo do intervalo de confiança (3.25), o limite superior do mesmo ultrapassou o valor 1 e, portanto, o mesmo foi truncado em 1. Realizando o mesmo cálculo, pode-se obter os intervalos de confiança de $S(t)$ para os demais valores de t. Os intervalos de confiança (3.25) para a função de sobrevivência do grupo Contraste Negativo são apresentados na Tabela 3.10 e Figura 3.9.

Nota: O R fornece os intervalos com base em $H(t)$ por meio do comando "survfit" da biblioteca *survival*. Para obter os intervalos de confiança (3.25) para a função de sobrevivência, basta usar a extensão *conf.type="log"*.
Obs: O comando "survfit" tem como padrão o intervalo tipo "log", sendo neste caso, desnecessário usar a extensão acima.
Ex: na<- survfit(coxph(Surv(tempo,censura)~1,method="breslow"),conf.type= "log")
ou
na<- survfit(coxph(Surv(tempo,censura)~1,method="breslow"))

Tabela 3.10: Estimativas de Nelson-Aalen e intervalos de confiança da transformação "log" para a função de sobrevivência. Dados do Exemplo 2 – Contraste Negativo.

j	**Tempo** $t_{(j)}$	**Intervalo** $[t_{(j)}, t_{(j+1)})$	d_j	n_j	$\hat{S}_{NA}(t)$	$\hat{H}_{NA}(t)$	$Vâr\left(\hat{H}_{NA}(t)\right)$	I.C. 95% p/ $S(t)$ "log"
0	0	[0, 23)	0	13	1	0	---	---
1	23	[23, 47)	1	13	0,9260	0,0769	0,0059	[0,796 ; 1]
2	47	[47, 69)	1	12	0,8519	0,1603	0,0129	[0,682 ; 1]
3	69	[69, 148)	1	11	0,7779	0,2512	0,0211	[0,585 ; 1]
4	148	[148, 181)	1	6	0,6585	0,4178	0,0489	[0,427 ; 1]
5	181	[181, ∞)	1	5	0,5391	0,6178	0,0889	[0,301 ; 0,967]

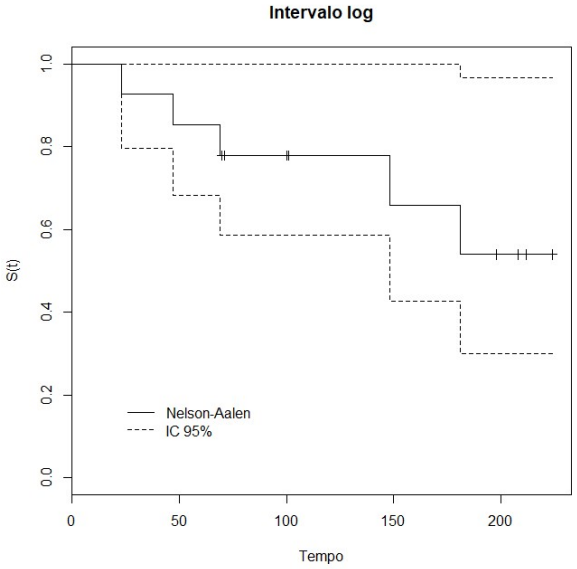


Figura 3.9: Estimativa de Nelson-Aalen para as funções de sobrevivência e seu respectivo intervalo de confiança da transformação "log" para os dados do Exemplo 2 – Contraste Negativo.

Intervalo log-log

Note que o limite superior dos intervalos de confiança (3.21) e (3.25) foram truncados em 1, visto que a função de sobrevivência é limitada no intervalo [0,1]. Como visto anteriormente, na Seção 3.2.1.1, esse inconveniente pode ser evitado aplicando-se a transformação "log-log" na função de sobrevivência. Assim, considerando que $\hat{U}_{NA}(t) = \log\left\{-\log \hat{S}_{NA}(t)\right\}$, tem-se que a variância de $\hat{U}_{NA}(t)$ é dada por:

$$\hat{Var}\left(\hat{U}_{NA}(t)\right) = \frac{1}{\left[\log \hat{S}_{NA}(t)\right]^2} \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j^2}. \qquad (3.26)$$

Aqui, $\hat{S}_{NA}(t)$ é o estimador de Nelson-Aalen da função de sobrevivência, expresso por (3.19), n_j é o número de indivíduos que estão sob risco no o tempo $t_{(j)}$ (inclusive) e d_j representa o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=1, 2, ..., k.

Assim, para todo t fixo, um intervalo 100(1 – α)% de confiança para $U(t)$ é dado por:

$$\underset{100(1-\alpha)\%}{I.C.}\left[U(t)\right]: \left[\hat{U}_{NA}(t) - z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{U}_{NA}(t)\right)}\,;\, \hat{U}_{NA}(t) + z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{U}_{NA}(t)\right)}\right], \qquad (3.27)$$

que resulta no seguinte intervalo de 100(1 – α)% de confiança para $S(t)$:

$$\underset{100(1-\alpha)\%}{I.C.}\left[S(t)\right]: \left[\left(\hat{S}_{NA}(t)\right)^{\exp\left\{z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{U}_{NA}(t)\right)}\right\}}\,;\, \left(\hat{S}_{NA}(t)\right)^{\exp\left\{-z_{(1-\alpha/2)}\sqrt{\hat{Var}\left(\hat{U}_{NA}(t)\right)}\right\}}\right], \qquad (3.28)$$

em que $\hat{S}_{NA}(t)$ é o estimador de Nelson-Aalen da função de sobrevivência, expresso por (3.19), $\hat{U}_{NA}(t)$ é o estimador de Nelson-Aalen da transformação log-log da função de sobrevivência e $z_{(1-\alpha/2)}$ é o quantil $\left(1-\frac{\alpha}{2}\right)$ da distribuição normal padrão.

Para os dados da Tabela 3.3 – Contraste Negativo tem-se, por exemplo, que variância de $\hat{U}_{NA}(t)$ em t=47 é estimada por

$$\hat{Var}\left(\hat{U}_{NA}(47)\right) = \frac{1}{\left[\log(0{,}8519)\right]^2}\left(\frac{1}{13^2} + \frac{1}{12^2}\right) = 0{,}5006\,.$$

Assim, um intervalo de 95% de confiança para S(t) em t=14 é dado por

$$\underset{95\%}{I.C.}\left[S(47)\right]: \left[\left(\hat{S}_{NA}(47)\right)^{\exp\left\{z_{(0,975)}\sqrt{\hat{Var}\left(\hat{U}_{NA}(47)\right)}\right\}}\,;\, \left(\hat{S}_{NA}(47)\right)^{\exp\left\{-z_{(0,975)}\sqrt{\hat{Var}\left(\hat{U}_{NA}(47)\right)}\right\}}\right]$$

$$\underset{95\%}{I.C.}\left[S(47)\right]: \left[(0{,}8519)^{\exp\left\{1{,}96\sqrt{0{,}5006}\right\}}\,;\, (0{,}8519)^{\exp\left\{-1{,}96\sqrt{0{,}5006}\right\}}\right]$$

$$\underset{95\%}{I.C.}\left[S(47)\right]: \left[0{,}527\,;\, 0{,}961\right]$$

Realizando o mesmo cálculo, pode-se obter os intervalos de confiança de $S(t)$ para os demais valores de t. Os intervalos de confiança para a função de sobrevivência do grupo Contraste Negativo são apresentados na Tabela 3.11 e Figura 3.10.

Nota: O R fornece os intervalos log-log por meio do comando "survfit" da biblioteca *survival*. Para obter os intervalos log-log para a função de sobrevivência (expressão 3.28), basta usar a extensão *conf.type="log-log"*.
Ex: na<- survfit(coxph(Surv(tempo,censura)~1,method="breslow"),conf.type= "log-log")

Tabela 3.11: Estimativas de Nelson-Aalen e intervalos de confiança log-log para a função de sobrevivência. Dados do Exemplo 2 – Contraste Negativo.

j	**Tempo** $t_{(j)}$	**Intervalo** $[t_{(j)}, t_{(j+1)})$	d_j	n_j	$\hat{S}_{NA}(t)$	$Vâr\left(\hat{U}_{NA}(t)\right)$	I.C. (95%) log-log
0	0	[0, 23)	0	13	1	---	---
1	23	[23, 47)	1	13	0,9260	1,0000	[0,579 ; 0,989]
2	47	[47, 69)	1	12	0,8519	0,5006	[0,527 ; 0,961]
3	69	[69, 148)	1	11	0,7779	0,3349	[0,458 ; 0,922]
4	148	[148, 181)	1	6	0,6585	0,2801	[0,308 ; 0,862]
5	181	[181, ∞)	1	5	0,5391	0,2329	[0,204 ; 0,787]

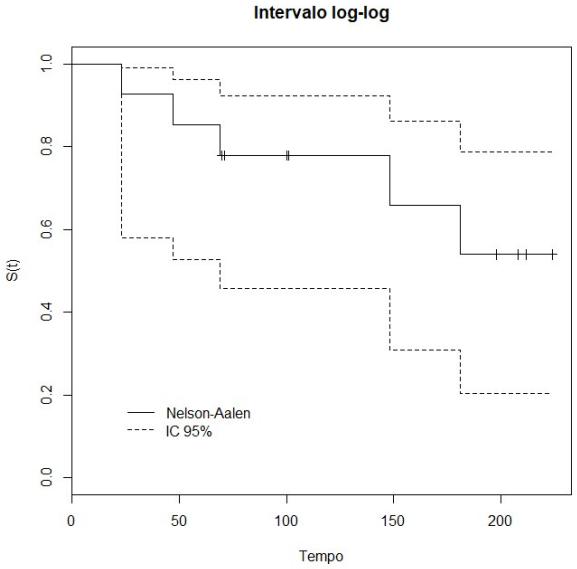


Figura 3.10: Estimativa de Nelson-Aalen para as funções de sobrevivência e seu respectivo intervalo de confiança log-log para os dados do Exemplo 2 – Contraste Negativo.

3.3. Estimação de quantidades básicas

A partir da função de sobrevivência é possível obter algumas medidas de interesse como os quantis do tempo de sobrevivência, o tempo de vida médio e a vida média residual. Procedimentos para estimar essas quantidades por meio das funções de sobrevivência de Kaplan-Meier e Nelson-Aalen são descritos a seguir.

3.3.1. Quantis do tempo de sobrevivência

Como visto em (2.3), o p-ésimo do tempo de sobrevivência, t_p, pode ser determinado a partir da função de sobrevivência por meio da expressão: $t_p = S^{-1}(1-p)$. Assim, o valor de t_p pode ser obtido por:

$$t_p = \inf\{t : S(t) \leq 1-p\}, \text{ para } 0 < p < 1. \tag{3.29}$$

Visto que a estimativa não paramétrica da função de sobrevivência é uma função escada, uma alternativa é obter a estimativa de t_p por meio de uma interpolação linear. No entanto, um problema pode ocorrer quando o maior tempo amostrado for uma censura. Como consequência, a função de sobrevivência empírica não atinge o ponto zero, impossibilitando o cálculo de quantis próximos a 1. Note que isto ocorre na função de sobrevivência de Nelson-Aalen, que também não atinge o valor zero, mesmo quando não há observações censuradas. Uma alternativa neste caso é limitar a estimativa empírica da função de sobrevivência ao maior tempo observado, isto é, considerar que $S(t) = 0$ para $t \geq \max\{t_1,t_2,...,t_n\}$.

Assim, um estimador do p-ésimo quantil, t_p, pode ser dado por:

$$\hat{t}_p = t_{(u)} + \frac{[t_{(u+1)} - t_{(u)}][\hat{S}(t_{(u)}) - (1-p)]}{[\hat{S}(t_{(u)}) - \hat{S}(t_{(u+1)})]}, \text{ para } 0 < p < 1. \tag{3.30}$$

Em (3.30), $[t_{(u)}; t_{(u+1)})$ é o intervalo que contém o p-ésimo quantil, isto é, o intervalo que satisfaz $\hat{S}(t_{(u+1)}) < 1-p \leq \hat{S}(t_{(u)})$ e $\hat{S}(t_{(u)})$ é a estimativa empírica da função de sobrevivência em $t_{(u)}$, u=0,1,..., k+1. Aqui, $t_{(k+1)}$=max$\{t_1,t_2,...,t_n\}$ e $\hat{S}(t_{(k+1)}) = 0$.

Note que estimar os quantis (3.30) é equivalente a conectar por retas as estimativas de $\hat{S}(t_{(j)})$, ao invés da forma de escada. É importante destacar que as estimativas dadas por (3.30) ficam subestimadas quando $1 - p < S(t_{(k)})$. Nesses casos, essas estimativas devem ser interpretadas com bastante cautela.

A variância assintótica do estimador dos quantis, $\hat{t}_p$, pode ser expressa por:

$$\hat{Var}(\hat{t}_p) = \frac{\hat{Var}(\hat{S}(\hat{t}_p))}{[f(\hat{t}_p)]^2} = \left[\frac{\hat{S}(t_{(u)}) - \hat{S}(t_{(u+1)})}{[t_{(u+1)} - t_{(u)}]}\right]^{-2} \hat{Var}(\hat{S}(\hat{t}_p)), \tag{3.31}$$

que resulta no seguinte intervalo de $100(1-\alpha)\%$ de confiança para t_p :

$$\underset{100(1-\alpha)\%}{I.C.}[t_p]: \left[\hat{t}_p - z_{(1-\alpha/2)}\sqrt{\hat{Var}(\hat{t}_p)}\,;\, \hat{t}_p + z_{(1-\alpha/2)}\sqrt{\hat{Var}(\hat{t}_p)}\right], \tag{3.32}$$

em que $\hat{t}_p$ é a estimativa do p-ésimo quantil do tempo de sobrevivência e $z_{(1-\alpha/2)}$ é o quantil $\left(1-\frac{\alpha}{2}\right)$ da distribuição normal padrão.

Para os dados da Tabela 3.3 – Contraste Positivo tem-se, considerando a função de sobrevivência de Nelson-Aalen (Tabela 3.8), que o primeiro quartil ($t_{0,25}$) está no intervalo

$[t_{(u)} = 26; t_{(u+1)} = 31)$, pois $\hat{S}_{NA}(31) < 1 - p \leq \hat{S}(26)$. Assim, segundo (3.30), a estimativa do primeiro quartil é dada por:

$$\hat{t}_{0,25} = 26 + \frac{[31-26][0{,}7550 - 0{,}75]}{[0{,}7550 - 0{,}7242]} = 26{,}81\,.$$

O valor de $\hat{Var}\left(\hat{S}_{NA}(\hat{t}_{0,25})\right)$ é dado por

$$\hat{Var}\left(\hat{S}_{NA}(26{,}81)\right) = \hat{S}^2_{NA}(26{,}81) \sum_{j:t_{(j)} \leq 26{,}81} \frac{d_j}{n_j^2} = 0{,}7550^2\left(\tfrac{1}{32^2} + \tfrac{1}{31^2} + \tfrac{1}{30^2} + \tfrac{1}{29^2} + \tfrac{1}{28^2} + \tfrac{1}{27^2} + \tfrac{2}{26^2}\right) = 0{,}005656\,,$$

e, portanto, tem-se que

$$\hat{Var}\left(\hat{t}_{0,25}\right) = \left[\frac{0{,}7550 - 0{,}7242}{[31-26]}\right]^{-2} 0{,}005656 = 149{,}05\,.$$

Assim, um intervalo de 95% de confiança para $t_{0,25}$ é dado por

$$I.C._{95\%}\left[t_{0,25}\right]: \left[\hat{t}_{0,25} - z_{(0,975)}\sqrt{\hat{Var}\left(\hat{t}_{0,25}\right)}\,;\, \hat{t}_{0,25} + z_{(0,975)}\sqrt{\hat{Var}\left(\hat{t}_{0,25}\right)}\right]$$

$$I.C._{95\%}\left[t_{0,25}\right]: \left[26{,}81 - 1{,}96\sqrt{149{,}05}\,;\, 26{,}81 + 1{,}96\sqrt{149{,}05}\right]$$

$$I.C._{95\%}\left[t_{0,25}\right]: \left[2{,}88\,;\, 50{,}74\right].$$

Realizando os mesmos cálculos, tem-se que a estimativa da mediana ($t_{0,5}$) do tempo de sobrevivência de mulheres que apresentam contraste positivo (e seu respectivo intervalo de 95% de confiança) é de 62,93 meses (IC 95%: [23,89 ; 101,97]).

3.3.2. Média do tempo de sobrevivência

Como visto em (2.13), o tempo médio de sobrevivência é representado pela área sob a função de sobrevivência. Visto que a estimativa não paramétrica da função de sobrevivência é uma função escada, esta área é simplesmente a soma de áreas de retângulos com base $(t_{(j+1)}-t_{(j)})$ e altura $\hat{S}(t_{(j)})$, j=1,2,...,k. No entanto, um problema pode ocorrer quando o maior tempo amostrado for uma censura pois, neste caso, $\hat{S}(t_{(k)}) \neq 0$ e o último retângulo terá área infinita (isto ocorre na função de sobrevivência de Nelson-Aalen, que também não atinge o valor zero, mesmo quando não há observações censuradas). Uma alternativa neste caso é limitar a estimativa empírica da função de sobrevivência ao maior tempo observado, isto é, considerar que $S(t) = 0$ para $t \geq \max\{t_1,t_2,...,t_n\}$.

Assim, um estimador do tempo médio de sobrevivência, $\overline{T}$, pode ser dado por:

$$\overline{T} = \sum_{j=0}^{k} (t_{(j+1)} - t_{(j)})\hat{S}(t_{(j)}). \tag{3.33}$$

Em (3.33), $\hat{S}(t_{(j)})$ é a estimativa empírica da função de sobrevivência em $t_{(j)}$, j = 0, 1, 2,..., k e $t_{(k+1)}=\max\{t_1,t_2,...,t_n\}$.

A variância da estimativa de $\overline{T}$ com base no estimador de Kaplan-Meier é dada por (Kaplan-Meier, 1958)

$$\hat{Var}(\overline{T}_{KM}) = \frac{d}{d-1}\sum_{j=1}^{k}\left(\left[\sum_{l=j}^{k}(t_{(l+1)} - t_{(l)})\hat{S}_{KM}(t_{(l)})\right]^2 \frac{d_j}{n_j(n_j - d_j)}\right), \tag{3.34}$$

e a variância da estimativa de $\overline{T}$ com base no estimador de Nelson-Aalen é dada por

$$\hat{Var}(\overline{T}_{NA}) = \frac{d}{d-1}\sum_{j=1}^{k}\left(\left[\sum_{l=j}^{k}(t_{(l+1)} - t_{(l)})\hat{S}_{NA}(t_{(l)})\right]^2 \frac{d_j}{n_j^2}\right). \tag{3.35}$$

Em (3.34) e (3.35), n_j é o número de indivíduos que estão sob risco no o tempo $t_{(j)}$ (inclusive), d_j representa o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=1,2,...,k+1, com $t_{(k+1)}=\max\{t_1,t_2,...,t_n\}$ e $d = \sum_{j=1}^{k} d_j$ é o número total de observações não censuradas na amostra. Note que o termo $A_j = \sum_{l=j}^{k}(t_{(l+1)} - t_{(l)})\hat{S}(t_{(l)})$ representa a área sob $\hat{S}(t)$ à direita de $t_{(j)}$, quando $\hat{S}(t_{(k+1)}) = 0$.

Assim, um intervalo (assintótico) de $(1 - \alpha)\times 100\%$ de confiança para o tempo médio de sobrevivência pode ser dado por

$$\underset{(1-\alpha)\times 100\%}{IC\ p/\ média}: \left[\overline{T} - z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\overline{T})}\ ;\ \overline{T} + z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\overline{T})}\right], \tag{3.36}$$

em que $z_{(1-\alpha/2)}$ é o quantil $\left(1-\frac{\alpha}{2}\right)$ da distribuição Normal Padrão. Em (3.36), $\overline{T}$ e $\sqrt{\hat{Var}(\overline{T})}$ podem ser obtidos com base no estimador da função de sobrevivência de Kaplan-Meier ou de Nelson-Aalen.

Na ausência de censuras, ao considerar a função de sobrevivência de Kaplan-Meier, as expressões (3.33) e (3.34) se reduzem a $\overline{T} = \frac{1}{n}\sum_{i=1}^{n} t_i$ e $\hat{Var}(\overline{T}_{KM}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(T_i - \overline{T})^2$,

respectivamente. No entanto, isso não ocorre quando se considera a função de sobrevivência de Nelson-Aalen.

Para os dados da Tabela 3.3 – Contraste Negativo tem-se, considerando a função de sobrevivência de Nelson-Aalen (Tabela 3.8), que $\overline{T} = 170{,}33$ e $\hat{Var}(\overline{T}_{NA}) = 510{,}20$. A Tabela 3.12 apresenta alguns elementos para o cálculo de $\overline{T}$ e $\sqrt{\hat{Var}(\overline{T})}$.

Tabela 3.12: Obtenção do tempo médio de sobrevivência a partir do estimador de Nelson-Aalen. Dados do Exemplo 2 – Contraste Negativo.

j	**Tempo** $t_{(j)}$	**Intervalo** $[t_{(j)}, t_{(j+1)})$	d_j	n_j	$\hat{S}_{NA}(t)$	$[t_{(j+1)} - t_{(j)}] \times (\hat{S}_{NA}(t))$	A_j^*
0	0	[0, 23)	0	13	1	23,0000	170,3304
1	23	[23, 47)	1	13	0,9260	22,2240	147,3304
2	47	[47, 69)	1	12	0,8519	18,7418	125,1074
3	69	[69, 148)	1	11	0,7779	61,4541	106,3650
4	148	[148, 181)	1	6	0,6585	21,7305	44,9114
5	181	[181, 224)	1	5	0,5391	23,1813	23,1818
			d=5			$\overline{T}$ =170,33	

* $A_j = \sum_{l=j}^{k} (t_{(l+1)} - t_{(l)}) \hat{S}_{NA}(t_{(l)})$

Assim, um intervalo de 95% de confiança para a média do tempo de sobrevivência (μ) de mulheres que apresentam contraste negativo é dado por

$$I.C._{95\%}[\mu]: \left[\overline{T} - z_{(0,975)}\sqrt{\hat{Var}(\overline{T})}\ ;\ \overline{T} + z_{(0,975)}\sqrt{\hat{Var}(\overline{T})}\right]$$

$$I.C._{95\%}[\mu]: \left[170{,}33 - 1{,}96\sqrt{510{,}20}\ ;\ 170{,}33 + 1{,}96\sqrt{510{,}20}\right]$$

$$I.C._{95\%}[\mu]: \left[126{,}06\ ;\ 214{,}60\right].$$

Realizando os mesmos cálculos, tem-se que a estimativa da média do tempo de sobrevivência de mulheres que apresentam contraste positivo (e seu respectivo intervalo de 95% de confiança) é de 106,18 meses (IC 95%: [72,03 ; 140,34]).

3.3.3. Vida média residual

A vida média residual no tempo *t* é a expectativa média de vida de um indivíduo que não falhou até *t* e, como visto na equação (2.15), a mesma pode ser representada pela área sob a curva de sobrevivência à direita do ponto *t*, dividido pelo valor da função de sobrevivência neste ponto. Assim, uma estimativa da função vida média residual pode ser dada por:

$$\hat{\nu}(t) = \frac{1}{\hat{S}(t)}\left[(t_{(u)} - t)\hat{S}(t) + \sum_{j:t_{(j)}\geq t}(t_{(j+1)} - t_{(j)})\hat{S}(t_{(j)})\right], \text{ para todo } t \geq 0. \quad (3.37)$$

Em (3.37), $\hat{S}(.)$ é a estimativa da função de sobrevivência, $t_{(k+1)}=\max\{t_1,t_2,...,t_n\}$ e $u = \inf\left\{j : t_{(j)} \geq t\right\}$. Em particular, tem-se que $\hat{\nu}(0) = \overline{T}$.

Para os dados da Tabela 3.3 – Contraste Negativo tem-se, para o cálculo de $\hat{\nu}(100)$ que $u = \inf\left\{j : t_{(j)} \geq 100\right\} = 4$. Assim, considerando a função de sobrevivência de Nelson-Aalen (Tabela 3.8), tem-se a partir de (3.37) que

$$\begin{aligned}\hat{\nu}(100) &= \frac{1}{\hat{S}_{NA}(100)}\left[(t_{(4)} - 100)\hat{S}_{NA}(100) + (t_{(5)} - t_{(4)})\hat{S}_{NA}(t_{(4)}) + (t_{(6)} - t_{(5)})\hat{S}_{NA}(t_{(5)})\right] \\ &= \frac{1}{0,7779}\left[(148-100)0,7779 + (181-148)0,6585 + (224-181)0,5391\right] \\ &= 105,73.\end{aligned}$$

Note que a função vida média residual é definida por $\nu(t) = E(T - t \mid T \geq t)$. Desta forma, uma alternativa para obter $\hat{\nu}(t)$ é estimar a média condicional $E(T \mid T \geq t)$ e subtrair o valor *t*, isto é,

$$\hat{\nu}(t) = \hat{E}(T \mid T \geq t) - t \text{, para todo } t \geq 0. \quad (3.38)$$

Em (3.38), o valor $\hat{E}(T \mid T \geq t)$ é a estimativa da média do tempo de sobrevivência dos tempos maiores ou iguais a *t*, que pode ser obtida a partir da expressão (3.33) e da estimativa da função de sobrevivência (obtidas por Kaplan-Meier ou Nelson-Aalen) excluindo-se as observações menores do que *t*.

Visto que $Var\left(\hat{\nu}(t)\right) = Var\left(\hat{E}(T \mid T \geq t)\right)$, tem-se que a variância $\hat{\nu}(t)$ pode, também, ser estimada por meio das expressões (3.34) e (3.35) excluindo da amostra as observações menores do que *t*.

Tomando o exemplo anterior, para obter a vida média residual em *t*=100, exclui-se todos as observações menores do que 100. Assim, as observações grupo contraste negativo será reduzida aos valores

$$100^+, 101^+, 148, 181, 198^+, 208^+, 212^+, 224^+.$$

A Tabela 3.13 apresenta a função de sobrevivência de Nelson-Aalen para a amostra reduzida.

Tabela 3.13: Obtenção do tempo médio de sobrevivência a partir do estimador de Nelson-Aalen. Dados reduzidos do Exemplo 2 – Contraste Negativo.

j	**Tempo** $t_{(j)}$	**Intervalo** $[t_{(j)}, t_{(j+1)})$	d_j	n_j	$\hat{S}_{NA}(t)$	$[t_{(j+1)} - t_{(j)}] \times (\hat{S}_{NA}(t))$	A_j^*
0	0	[0, 148)	0	8	1	148	205,7346
1	148	[148, 181)	1	6	0,8465	27,9339	57,7346
2	181	[181, 224]	1	5	0,6930	29,8008	29,8008
			d=2		$\hat{E}(T \mid T \geq 100) = 205,73$		

$A_j = \sum_{l=j}^{k} (t_{(l+1)} - t_{(l)}) \hat{S}_{NA}(t_{(l)})$

Tem-se que $\hat{E}(T \mid T \geq 100) = 205,73$ e, portanto, $\hat{\nu}(100) = 205,73 - 100 = 105,73$. Além disso, tem-se que

$$\hat{Var}(\hat{\nu}(100)) = \hat{Var}\left(\hat{E}(T \mid T \geq 100)\right) = \frac{2}{2-1}\left([57,7346]^2 \frac{1}{6^2} + [29,8007]^2 \frac{1}{5^2} \right) = 265,2295 .$$

Assim, um intervalo de 95% de confiança para vida média residual de mulheres que apresentam contraste negativo e que sobreviveram até o tempo t=100 é dado por

$$I.C._{95\%}[\nu(100)]: \left[\hat{\nu}(100) - z_{(0,975)} \hat{Var}(\hat{\nu}(100)) \; ; \; \hat{\nu}(100) + z_{(0,975)} \hat{Var}(\hat{\nu}(100))\right]$$

$$I.C._{95\%}[\nu(100)]: \left[105,73 - 1,96\sqrt{256,23} \; ; \; 105,73 - 1,96\sqrt{256,23}\right]$$

$$I.C._{95\%}[\nu(100)]: \left[74,36 \; ; \; 137,11\right].$$

Realizando o mesmo cálculo, pode-se obter a vida média residual e seu respectivo intervalo de confiança para os demais valores de t. A Figura 3.11 presenta a estimativa da função vida média residual para os dados do Exemplo 2.

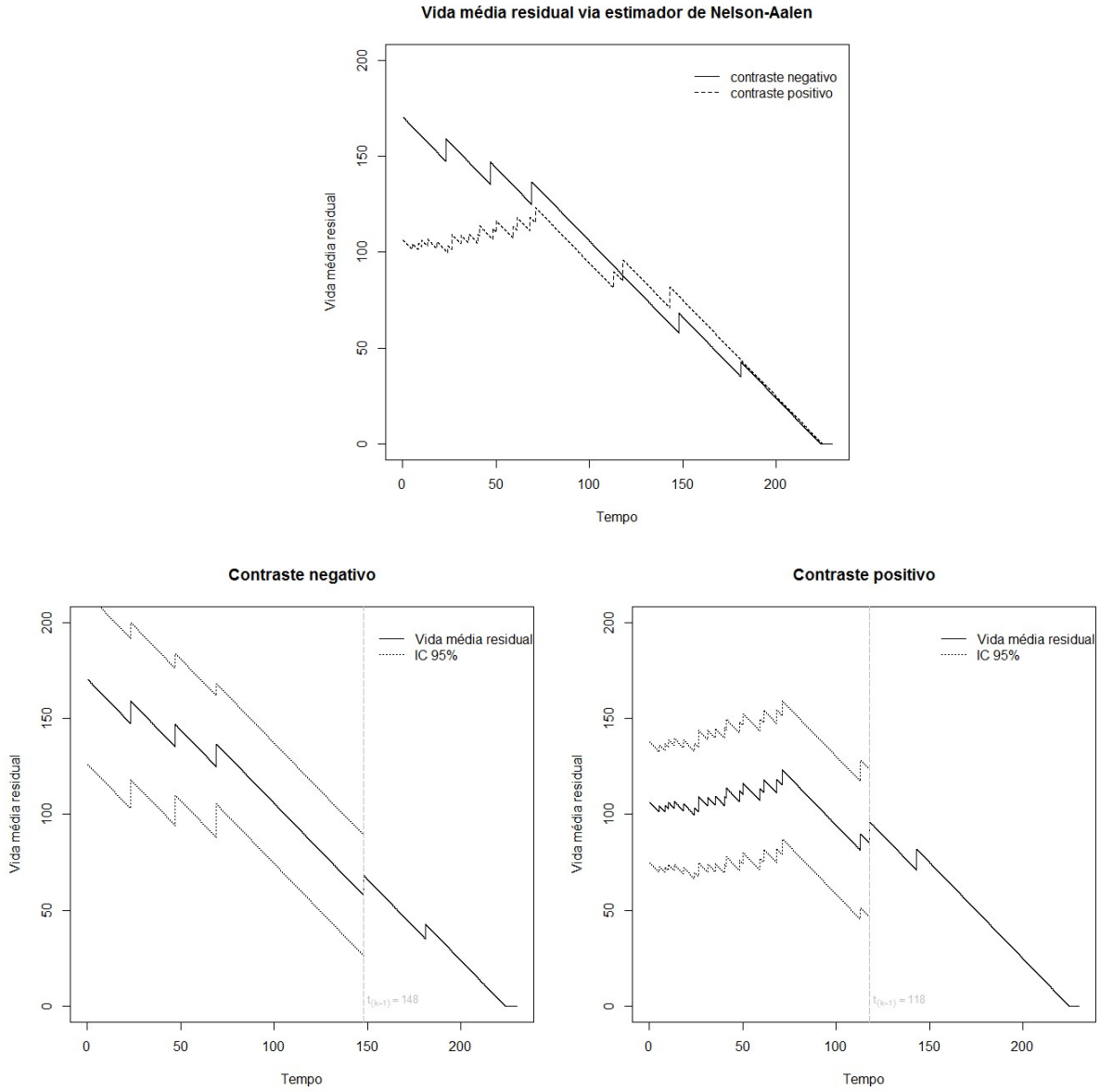


Figura 3.11: Estimativas da função vida média residual para os dados do Exemplo 2.

3.4. Comparação de curvas de sobrevivência

Embora seja possível comparar grupos de indivíduos visualmente por meio dos gráficos das Funções de Sobrevivência estimadas, é desejável que se tenha um teste estatístico para apoiar decisões sobre a igualdade ou não entre duas ou mais curvas de sobrevivência. No caso particular de dois grupos, A e B, o teste terá as seguintes hipóteses.

$H_0: S_a(t) = S_b(t)$ Não existe diferença entre as curvas de sobrevivência
$H_1: S_a(t) \neq S_b(t)$ Existe diferença entre as curvas de sobrevivência

Os testes mais comuns para a comparação de duas ou mais funções de sobrevivência podem ser vistas como generalizações para dados censurados de testes não paramétricos conhecidos.

3.4.1. Teste de Mantel-Haenszel

Proposto em 1959, o teste de Mantel-Haenszel compara conjuntos de proporções em tabelas de contingência estratificadas (Mantel e Haenszel, 1959).

Considere dois grupos de indivíduos, A e B, a serem comparados. Para cada $t_{(j)}$ fixo tem-se a seguinte tabela de contingência:

	Grupo A	Grupo B	
Experimentou o evento de interesse	d_{Aj}	d_{Bj}	d_j
Não experimentou o evento de interesse	$n_{Aj} - d_{Aj}$	$n_{Bj} - d_{Bj}$	$n_j - d_j$
	n_{Aj}	n_{Bj}	n_j

Aqui, n_j representa o número total de indivíduos sob risco no tempo $t_{(j)}$, n_{A_j} (n_{B_j}) representa o número de indivíduos do Grupo A (B) sob risco no tempo $t_{(j)}$, d_{A_j} (d_{B_j}) representa o número de indivíduos do Grupo A (B) que experimentaram o evento de interesse e d_j representa o número total de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=1,2,...,k.

Em uma primeira etapa deve-se combinar e ordenar os dados dos dois grupos e calcular os números esperados de eventos de interesse no Grupo A, em um determinado tempo $t_{(j)}$, isto é,

$$E_{A_j} = \frac{d_j n_{A_j}}{n_j}, \tag{3.39}$$

A partir de E_{A_j}, obtém-se os números totais esperados de eventos de interesse nos dois grupos (A e B), dados por,

$$E_A = \sum_{j=1}^{k} E_{A_j} \quad \text{e} \quad E_B = d - E_A, \tag{3.40}$$

em que $d = \sum_{j=1}^{k} d_j$ é o número total de indivíduos que experimentaram o evento de interesse considerando os dois grupos.

A estatística para testar a hipótese de igualdade entre duas funções de sobrevivência é dada por

$$U_{MH} = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B} \approx \chi^2_{(1)} \tag{3.41}$$

que segue (aproximadamente) uma distribuição Qui-quadrado com 1 grau de liberdade. Aqui, O_A e O_B representam os números totais observados de indivíduos que experimentaram o evento de interesse em cada grupo.

Assim, o nível descritivo (valor-p) do teste de Mantel-Haenszel é dado por

$$p = P[\chi^2_{(1)} > U_{MH}], \tag{3.42}$$

e a hipótese de igualdade das duas curvas de sobrevivência (hipótese H_0) é rejeitada se $p < \alpha$, em que α é o nível de significância do teste.

OBS: O teste de Mantel-Haenszel pode ser facilmente estendido para comparar r (r>2) curvas de sobrevivência. Neste caso, os dados podem ser rearranjados em uma tabela de contingência $2\times r$ e a estatística do teste tem (aproximadamente) uma distribuição Qui-quadrado com r – 1 graus de liberdade.

Exemplo 3. Teste do Esforço: *Os dados a seguir, são referentes a um estudo sobre o esforço aplicado a dez pacientes normais (Grupo A) e dez severamente obesos (Grupo B). Cada paciente foi submetido a um esforço por até 12 minutos com a possibilidade de desistir antes caso não suportasse mais. Os tempos (em minutos) até a desistência foram:*

Grupo A	3	6	8	9	9^+	10	11	12^+	12^+	12^+
Grupo B	1	1	2	3	4	5	5^+	7	9	12^+

+ indica que houve falhas no equipamento e o teste foi interrompido ou o final do tempo de teste foi atingido e foram considerados tempos censurados.

O gráfico da função de sobrevivência de Kaplan-Meier para os dois grupos é apresentado pela Figura 3.12.

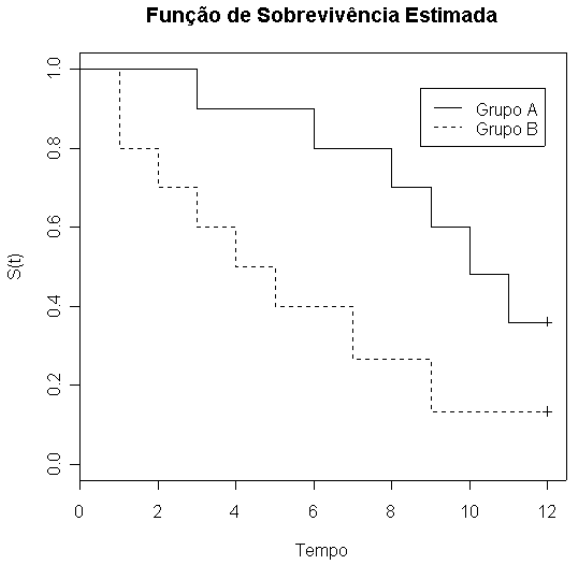


Figura 3.12: Estimativa de Kaplan-Meier para dos dados do Exemplo 3.

É possível observar, a partir da Figura 3.12, que o Grupo A parece suportar por mais tempo o esforço que o Grupo B. Para verificar se realmente essa diferença é significativa, será realizado a seguir o teste de Mantel-Haenszel para a comparação das duas curvas de sobrevivência.

A Tabela 3.14 apresenta os resultados dos cálculos dos números esperados de desistências no Grupo A.

Tabela 3.14: Número esperado de desistências do Grupo A.

$t_{(j)}$	n_j	n_{Aj}	d_j	d_{Aj}	$E_{A_j} = \frac{d_j n_{A_j}}{n_j}$
1	20	10	2	0	1,0000
2	18	10	1	0	0,5556
3	17	10	2	1	1,1765
4	15	9	1	0	0,6000
5	14	9	1	0	0,6429
6	12	9	1	1	0,7500
7	11	8	1	0	0,7273
8	10	8	1	1	0,8000
9	9	7	2	1	1,5556
10	6	5	1	1	0,8333
11	5	4	1	1	0,8000
			d=14	**O_A = 6**	**E_A = 9,4412**

Neste exemplo, tem-se que E_A=9,4412; E_B=14–9,4412=4,5588; O_A=6 e O_B=8. Assim,

$$U_{MH} = \frac{(6-9{,}4412)^2}{(9{,}4412)} + \frac{(8-4{,}5588)^2}{4{,}5588} = 3{,}8519\,,$$

que resulta em

$$p = P[\chi^2_{(1)} > 3{,}8519] = 0{,}0497.$$

A um nível de significância de 5%, pode-se dizer que as curvas de sobrevivência dos Grupos A e B são diferentes. Assim, é possível concluir que os pacientes obesos suportam o esforço por menos tempo que os pacientes com pesos normais.

3.4.2. Teste *logrank*

O teste *logrank* é o teste mais utilizado para comparar duas ou mais curvas de sobrevivência (Colosimo e Giolo, 2006). Apresentado por Mantel em 1966 (Mantel, 1966), a fórmula do teste *logrank* é obtida de forma similar à do teste de Mantel-Haenszel apresentado na seção anterior.

Considere dois grupos de indivíduos, A e B, a serem comparados. Assim como no teste de Mantel-Haenszel tem-se, para cada $t_{(j)}$, a seguinte tabela de contingência.

	Grupo A	Grupo B	
Experimentou o evento de interesse	d_{Aj}	d_{Bj}	d_j
Não experimentou o evento de interesse	$n_{Aj} - d_{Aj}$	$n_{Bj} - d_{Bj}$	$n_j - d_j$
	n_{Aj}	n_{Bj}	n_j

Aqui, n_j representa o número total de indivíduos sob risco no tempo $t_{(j)}$, n_{A_j} (n_{B_j}) representa o número de indivíduos do Grupo A (B) sob risco no tempo $t_{(j)}$, d_{A_j} (d_{B_j}) representa o número de indivíduos do Grupo A (B) que experimentaram o evento de interesse e d_j representa o número total de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=1,2,...,k.

Condicional ao número de indivíduos sob risco (fixando as marginais de coluna) e ao número de eventos de interesse (fixando as marginais de linha), a distribuição de d_{Aj} é uma Hipergeométrica com **média** E_{Aj} e **variância** V_{Aj}, em que

$$E_{A_j} = \frac{d_j n_{A_j}}{n_j} \quad \text{e} \quad V_{A_j} = \frac{d_j n_{A_j}(n_j - n_{A_j})(n_j - d_j)}{n_j^2(n_j - 1)}. \tag{3.43}$$

Assim, tem-se que $(d_{Aj} - E_{Aj})$ tem média zero e variância V_{Aj}. Assumindo que as k tabelas de contingência são independentes, um teste aproximado para a igualdade das duas funções de sobrevivência pode ser baseado na estatística (Mantel, 1966):

$$U_{\text{logrank}} = \frac{\left[\sum_{j=1}^{k}\left(d_{A_j} - E_{A_j}\right)\right]^2}{\sum_{j=1}^{k} V_{A_j}} \approx \chi^2_{(1)}, \tag{3.44}$$

que, sob a hipótese de igualdade das duas funções de sobrevivência, tem aproximadamente uma distribuição Qui-Quadrado com 1 grau de liberdade.

Assim, o nível descritivo (valor-p) do teste *logrank* é dado por

$$p = P[\chi^2_{(1)} > U_{\text{logrank}}], \tag{3.45}$$

e a hipótese de igualdade das duas curvas de sobrevivência (hipótese H_0) é rejeitada se $p < \alpha$, em que α é o nível de significância do teste.

OBS: O teste *logrank* pode ser estendido para comparar r (r>2) curvas de sobrevivência. Neste caso, os dados podem ser rearranjados em uma tabela de contingência 2×r, a estatística do teste é construída com base na distribuição Hipergeométrica multivariada e terá (aproximadamente) uma distribuição Qui-quadrado com r – 1 graus de liberdade.

Para os dados do Exemplo 3 (Teste do Esforço) tem-se os seguintes resultados para o teste *logrank* (Tabela 3.15):

Tabela 3.15: Dados para o cálculo da estatística do teste *logrank* (Exemplo 3).

$t_{(j)}$	n_j	n_{Aj}	d_j	d_{Aj}	$E_{A_j} = \frac{d_j n_{A_j}}{n_j}$	$d_{Aj} - E_{Aj}$	$V_{A_j} = \frac{d_j n_{A_j}(n_j - n_{A_j})(n_j - d_j)}{n_j^2(n_j - 1)}$
1	20	10	2	0	1,0000	-1,0000	0,4737
2	18	10	1	0	0,5556	-0,5556	0,2469
3	17	10	2	1	1,1765	-0,1765	0,4542
4	15	9	1	0	0,6000	-0,6000	0,2400
5	14	9	1	0	0,6429	-0,6429	0,2296
6	12	9	1	1	0,7500	0,2500	0,1875
7	11	8	1	0	0,7273	-0,7273	0,1983
8	10	8	1	1	0,8000	0,2000	0,1600
9	9	7	2	1	1,5556	-0,5556	0,3025
10	6	5	1	1	0,8333	0,1667	0,1389
11	5	4	1	1	0,8000	0,2000	0,1600
			14	**6**		**-3,4410**	**2,7915**

Assim,

$$U_{\text{logrank}} = \frac{\left[\sum_{j=1}^{k}\left(d_{A_j} - E_{A_j}\right)\right]^2}{\sum_{j=1}^{k} V_{A_j}} = \frac{(-3{,}4410)^2}{2{,}7915} = 4{,}2416,$$

que resulta em

$$p = P[\chi^2_{(1)} > 4{,}2416] = 0{,}0394.$$

A um nível de significância de 5%, pode-se dizer que as curvas de sobrevivência dos Grupos A e B são diferentes. Assim, é possível concluir que os pacientes obesos suportam o esforço por menos tempo que os pacientes com pesos normais.

3.4.3. Teste de Wilcoxon

Os testes de Mantel-Haenszel e *logrank* propostos nas seções anteriores consideram o mesmo peso para as k tabelas de contingência, independentemente do número de indivíduos sob risco, n_j. No início do estudo todos os indivíduos estão sob risco e esse número diminui com o passar do tempo (os indivíduos vão experimentando o evento de interesse ou vão censurando). Assim, esses testes são indicados em situações onde os riscos têm um comportamento proporcional (os riscos relativos das tabelas de contingencia são iguais).

Uma alternativa para comparar funções de sobrevivência quando não há proporcionalidade dos riscos é uma adaptação do conhecido teste não paramétrico de Wilcoxon para dados não censurados (Gehan, 1965). Este teste, ao contrário do teste de Mantel-Haenszel e *logrank*, considera um peso maior nas comparações das tabelas de contingência para tempos mais recente (quando mais indivíduos ainda estão sob risco). A estatística do teste de Wilcoxon é dada por (Gehan, 1965):

$$U_{wilcoxon} = \frac{\left[\sum_{j=1}^{k} n_j\left(d_{A_j} - E_{A_j}\right)\right]^2}{\sum_{j=1}^{k} n_j^2 V_{A_j}} \approx \chi^2_{(1)}, \qquad (3.46)$$

que tem (aproximadamente) uma distribuição Qui-Quadrado com 1 grau de liberdade. Em (3.46), n_j representa o número total de indivíduos sob risco no tempo $t_{(j)}$, n_{A_j} (n_{B_j}) representa o número de indivíduos do Grupo A (B) sob risco no tempo $t_{(j)}$, d_{A_j} (d_{B_j}) representa o número de indivíduos do Grupo A (B) que experimentaram o evento de interesse e d_j representa o número total de indivíduos que experimentara o evento de interesse no tempo $t_{(j)}$, j=1,2,...,k. Assim como no teste *logrank*, E_{A_j} e V_{A_j} são expressos por (3.43).

Assim, o nível descritivo (valor-p) do teste de Wilcoxon é dado por

$$p = P[\chi^2_{(1)} > U_{wilcoxon}], \qquad (3.47)$$

e a hipótese de igualdade das duas curvas de sobrevivência (hipótese H_0) é rejeitada se $p < \alpha$, em que α é o nível de significância do teste.

Para os dados do Exemplo 3 (Teste do Esforço) tem-se os seguintes resultados para o teste de Wilcoxon (Tabela 3.16):

Tabela 3.16: Dados para o cálculo da estatística do teste de Wilcoxon (Exemplo 3).

$t_{(j)}$	n_j	n_{Aj}	d_j	d_{Aj}	$E_{A_j} = \frac{d_j n_{A_j}}{n_j}$	$n_j(d_{Aj} - E_{Aj})$	$n_j^2 V_{A_j} = \frac{d_j n_{A_j}(n_j - n_{A_j})(n_j - d_j)}{(n_j - 1)}$
1	20	10	2	0	1,0000	-20	189,47
2	18	10	1	0	0,5556	-10	80,00
3	17	10	2	1	1,1765	-3	131,25
4	15	9	1	0	0,6000	-9	54,00
5	14	9	1	0	0,6429	-9	45,00
6	12	9	1	1	0,7500	3	27,00
7	11	8	1	0	0,7273	-8	24,00
8	10	8	1	1	0,8000	2	16,00
9	9	7	2	1	1,5556	-5	24,50
10	6	5	1	1	0,8333	1	5,00
11	5	4	1	1	0,8000	1	4,00
			14	**6**		**-57**	**600,22**

Assim,

$$U_{wilcoxon} = \frac{\left[\sum_{j=1}^{k} n_j \left(d_{A_j} - E_{A_j}\right)\right]^2}{\sum_{j=1}^{k} n_j^2 V_{A_j}} = \frac{(-57)^2}{600{,}22} = 5{,}4130\,,$$

que resulta em

$$p = P[\chi^2_{(1)} > 5{,}4130] = 0{,}0200\,.$$

A um nível de significância de 5%, pode-se dizer que as curvas de sobrevivência dos Grupos A e B são diferentes. Assim, é possível concluir que os pacientes obesos suportam o esforço por menos tempo que os pacientes com pesos normais.

Exemplo 4. *Para ilustrar a diferença entre o teste de Wilcoxon e logrank, considere o seguinte conjunto de dados de sobrevivência de dois grupos (A e B).*

Grupo A	0	1	1	2	2	3	4	5	5	6	7	9	10	11	12	13	15	18	19	25	26	32	32	37^+	38^+	40^+	46^+	53^+	67	127^+
Grupo B	1	1	2	3	2	3	4	5	6	7	8	9^+	11	25	25^+	25^+	25^+	26	27^+	29^+	32^+	32^+	32^+	32^+	40^+	48^+	60^+	89^+	90^+	112^+

As funções de sobrevivência de Kaplan-Meier para os dois grupos são dadas pela Figura 3.13.

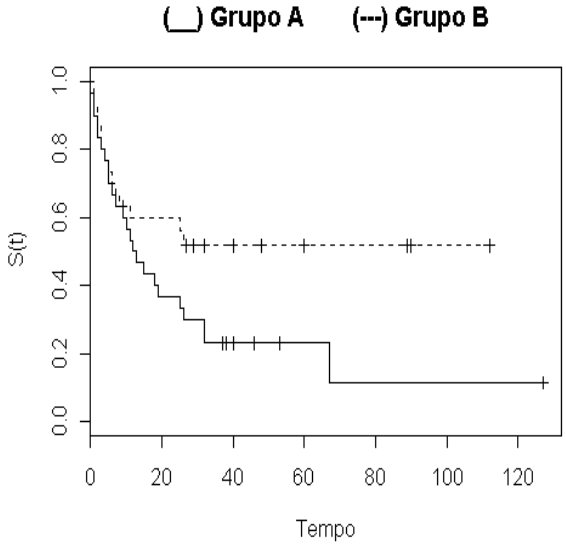


Figura 3.14: Estimativa de Kaplan-Meier para dos dados do Exemplo 4.

Neste exemplo tem-se que:

	p
Logrank	0,0498
Wilcoxon	0,1887

É possível notar que as duas curvas de sobrevivência similares até t=10, e somente a partir deste tempo que as funções do Grupo A e B divergem. Considerando um nível de significância de 5%, o teste de Wilcoxon não rejeita a hipótese de igualdade das curvas de sobrevivência. Já o teste *logrank* indica que as curvas de sobrevivência são diferentes. Isto ocorre porque o teste de Wilcoxon considera um peso maior para os tempos recentes (onde as funções de sobrevivência são similares).

3.4.4. Família de testes

Existe outro teste de comparação de curvas de sobrevivência definido como Família de Testes. A Família de testes é baseada na seguinte estatística

$$U_{FT} = \frac{\left[\sum_{j=1}^{k} u_j \left(d_{A_j} - E_{A_j}\right)\right]^2}{\sum_{j=1}^{k} u_j^2 V_{A_j}} \approx \chi^2_{(1)}, \qquad (3.48)$$

que tem (aproximadamente) uma distribuição Qui-Quadrado com 1 grau de liberdade. Em (3.48), u_j representa os pesos que especificam os testes, n_j representa o número total de indivíduos sob risco no tempo $t_{(j)}$, n_{A_j} (n_{B_j}) representa o número de indivíduos do Grupo A (B) sob risco no tempo $t_{(j)}$, d_{A_j} (d_{B_j}) representa o número de indivíduos do Grupo A (B) que experimentaram o evento de interesse e d_j representa o número total de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=1,2,...,k. Assim como no teste de Wilcoxon e *logrank*, E_{A_j} e V_{A_j} são expressos por (3.43).

Assim, o nível descritivo (valor-p) da Família de testes é dado por

$$p = P[\chi^2_{(1)} > U_{FT}], \tag{3.49}$$

e a hipótese de igualdade das duas curvas de sobrevivência (hipótese H_0) é rejeitada se $p < \alpha$, em que α é o nível de significância do teste.

.

A família de testes engloba uma grande variedade de testes de acordo com a escolha dos pesos u_j na expressão (3.48). Dentre outros, pode-se destacar:

i) o teste *logrank* (Mantel, 1966), quando $u_j = 1$;
ii) o teste de Wilcoxon (Gehan, 1965), quando $u_j = n_j$;
iii) o teste de Tarone-Ware (Tarone e Ware, 1977), quando $u_j = \sqrt{n_j}$; e
iv) a família de pesos de Harrington-Fleming (Harrington e Fleming, 1982), quando $u_j = \left[S_{KM}(t_{(j-1)})\right]^{\rho}$, $-\infty < \rho < \infty$.

A escolha dos pesos u_j direciona o tipo de comparação das funções de sobrevivência. O teste *logrank* é obtido quando $u_j = 1$, isto é, considera o mesmo peso para todo o eixo do tempo. Já o teste de Wilcoxon é obtido com $u_j = n_j$ e, portanto, considera pesos menores para tempos mais avançados (pois a medida que o tempo avança, o número de indivíduos sob risco diminui), reforçando o enfoque nas diferenças nos tempos mais recentes. Já o teste de Tarone-Ware se encontra numa situação intermediária entre os testes *logrank* e Wilcoxon.

A família de testes de Harrington-Fleming define os pesos em termos da função de sobrevivência no tempo anterior ($u_j = \left[S_{KM}(t_{(j-1)})\right]^{\rho}$). Desta forma, o teste pode considerar pesos iguais para todos os tempos, quando $\rho = 0$ (se reduz ao teste *logrank*); maior peso para os tempos mais recentes quando $\rho > 0$; e maior peso para os tempos mais avançados quando $\rho < 0$. Em particular, quando $\rho = 1$, o peso é a função de sobrevivência de Kaplan-Meier no tempo de falha anterior e, neste caso tem-se um teste similar ao de Wilcoxon (Colosimo e Giolo, 2006).

Para os dados do Exemplo 3 (Teste do Esforço) tem-se os seguintes resultados dos testes de comparação das curvas de sobrevivência dos Grupos A e B.

Teste	**Estatística (*U*)**	***p***
Mantel-Haenzel	3,8519	0,0497
Logrank	4,2416	0,0394
Wilcoxon	5,4130	0,0200
Tarone e Ware	5,0544	0,0246
Harrington-Fleming (ρ=1)	5,2500	0,0219

Nota: O R, para a comparação de curvas de sobrevivência, utiliza a *família de pesos de Harrington-Fleming*. Para a realização do teste utiliza-se o comando "survdiff", que possui como opção a escolha de ρ (rho). O comando survdiff tem como *default* rho=0 (teste *logrank*).

```
Ex: survdiff(Surv(tempo,censura)~grupo,rho=0)     # teste logrank
    survdiff(Surv(tempo,censura)~grupo,rho=1)     # similar ao teste de Wilcoxon
```

3.5. Exercícios

1. Considere os dados de um estudo com pacientes de leucemia quanto à introdução de um novo tratamento com a droga *6*-mercaptopurina.

Tabela. Tempos de Remissão (semanas) de pacientes com leucemia.

Tratamento	6*	6	6	6	7	9*	10*	10	11*	13	16	17*	19*	20*	22	23	25*	32*	32*	35*	35*
Controle	1	1	2	2	3	4	4	5	5	8	8	8	8	11	11	12	12	15	17	22	23

*: censura

i) Calcule as estimativas de $S(t)$ e $H(t)$.
ii) Calcule os intervalos de confiança da função de sobrevivência (intervalos simétricos e log-log);
iii) Calcule os intervalos de confiança de $H(t)$;
iv) Qual é estimativa da função de sobrevivência (com seu respectivo intervalo de confiança simétrico) para o tempo t=14?
v) Calcule as estimativas pontuais e intervalares da média e da mediana do tempo de remissão dos pacientes dos grupos Tratamento e Controle.
vi) Calcule a estimativa da vida média residual em t=20 semanas e seu respetivo intervalo de 95% de confiança.
Nota: Resolver os itens acima considerando o estimador de Kaplan-Meier e Nelson-Aalen.

2. Mostre que:
i) Se $\hat{S}(t) \sim N\left(S(t); Vâr[\hat{S}(t)]\right)$, então um intervalo de 100(1 – α)% de confiança para $S(t)$, é dado por: $\underset{100(1-\alpha)\%}{I.C.}\left[S(t)\right]$: $\left[\hat{S}(t) - z_{(1-\alpha/2)}\sqrt{Vâr\left(\hat{S}(t)\right)}\ ;\ \hat{S}(t) + z_{(1-\alpha/2)}\sqrt{Vâr\left(\hat{S}(t)\right)}\right]$;

ii) Se $\hat{H}(t) \sim N\left(H(t); Vâr[\hat{H}(t)]\right)$, então um intervalo de 100(1 – α)% de confiança para $S(t)$, é dado por: $\underset{100(1-\alpha)\%}{I.C.}\left[S(t)\right]$: $\left[e^{-z_{(1-\alpha/2)}\sqrt{Vâr\left(\hat{H}(t)\right)}}\hat{S}(t)\ ;\ e^{z_{(1-\alpha/2)}\sqrt{Vâr\left(\hat{H}(t)\right)}}\hat{S}(t)\right]$;

iii) Se $\hat{U}(t) = \log\left\{-\log\hat{S}(t)\right\}$ e $\hat{U}(t) \sim N\left(U(t); Vâr[\hat{U}(t)]\right)$, então um intervalo de 100(1 – α)% de confiança para $S(t)$, é dado por: $\underset{100(1-\alpha)\%}{I.C.}\left[S(t)\right]$: $\left[\left(\hat{S}(t)\right)^{\exp\left\{z_{(1-\alpha/2)}\sqrt{Vâr\left(\hat{U}(t)\right)}\right\}}\ ;\ \left(\hat{S}(t)\right)^{\exp\left\{-z_{(1-\alpha/2)}\sqrt{Vâr\left(\hat{U}(t)\right)}\right\}}\right]$.

3. A tabela abaixo apresenta os tempos de sobrevivência (em horas) de um tipo de filme utilizado como isolante elétrico. Os tempos até a falha são observados em 4 diferentes níveis de estresse: 5, 7, 10, e 15 quilovolts-ampères.

Tabela. Tempos de resistência de filmes isolantes

Voltagem	N	Tempo (em horas) até a falha
5	10	71,31 8482 8559 8762 9026 9034 9104 9104,25* 9104,25* 9104,25*
7	15	50,25 87,75 87,76 87,77 92,9 92,91 95,96 108,3 108,3 117,9 123,9 124,3 129,7 135,6 135,6
10	10	15,17 19,87 20,18 21,5 21,88 22,23 23,02 23,9 28,17 29,7
15	9	2,4 2,42 3,17 3,75 4,65 4,95 6,23 6,68 7,3

*: censura

No R: Compare os grupos usando o teste *logrank* e Harrington-Fleming com ρ = 1. Apresente uma alternativa para realizar testes de comparações múltiplas para identificar quais grupos diferem significativamente.

3.5. Referências

Aalen, O.O. Nonparametric inference for a family of counting processes. *Annals of Statistics*, vol. 6, p.701-726, 1978.

Aalen, O.O.; Johansen, S. An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, vol.5, p.141-150, 1978.

Colosimo, E.A.; Giolo, S.R. *Análise de sobrevivência aplicada*. São Paulo: Edgard Blucher Ltda, 1 ed., 369p, 2006.

Gehan, E.A. A generalized Wilcoxon test for comparing arbitrarity singly-censored samples. *Biometrika*, vol. 52, p.203-223, 1965.

Greenwood, M. *The natural duration of cancer*. Reports on Public Health and Medical Subjects. London: Her Majesty's Stationery Office, vol. 33, p.1–26, 1926.

Harrington D.P.; Fleming, T.R. A class of rank test procedures for censored survival data. *Biometrika*, vol.69, p.133-143, 1982.

Kalbfleisch, J.D.; Prentice, R.L. *The statistical analysis of failure time data*. John Wiley and Sons, New York, 2 ed., 462p, 2002.

Kaplan, E.L.; Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, vol. 53, p. 457-481, 1958.

Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, vol. 50, p.163-170, 1966.

Mantel, N; Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.*, vol.22, p.719-748, 1959.

Nelson. W. Theory and applications of hazard plotting for censored failure data. *Technometrics*, vol. 14, p.945-965, 1972.

Tarone, R.E.; Ware, J.H. On distribution-free tests for equality for survival distributions. *Biometrika*, vol. 64, p.156-160, 1977.

APENDICE B

Tabela. Estimadores das principais funções e quantidades básicas utilizadas para descrever o tempo de sobrevivência.

	Kaplan-Meier	**Nelson-Aalen**
Função de sobrevivência	$\hat{S}_{KM}(t) = \prod_{j:t_{(j)} \le t} \left[1 - \frac{d_j}{n_j} \right]$	$\hat{S}_{NA}(t) = \exp\left\{ - \hat{H}_{NA}(t) \right\}$
Função de risco acumulado	$\hat{H}_{KM}(t) = -\log \hat{S}_{KM}(t)$	$\hat{H}_{NA}(t) = \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j}$
Log-log sobrevivência	$\hat{U}_{KM}(t) = \log \hat{H}_{KM}(t) = \log\left\{ - \log \hat{S}_{KM}(t) \right\}$	$\hat{U}_{NA}(t) = \log \hat{H}_{NA}(t) = \log\left\{ - \log \hat{S}_{NA}(t) \right\}$
Variâncias das funções	$\hat{Var}\left(\hat{S}_{KM}(t)\right) = \hat{S}^2_{KM}(t) \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j (n_j - d_j)}$ $\hat{Var}\left(\hat{H}_{KM}(t)\right) = \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j (n_j - d_j)}$ $\hat{Var}\left(\hat{U}_{KM}(t)\right) = \frac{1}{\left[\log \hat{S}_{KM}(t)\right]^2} \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j (n_j - d_j)}$	$\hat{Var}\left(\hat{S}_{NA}(t)\right) = \hat{S}^2_{NA}(t) \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j^2}$ $\hat{Var}\left(\hat{H}_{NA}(t)\right) = \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j^2}$ $\hat{Var}\left(\hat{U}_{NA}(t)\right) = \frac{1}{\left[\log \hat{S}_{NA}(t)\right]^2} \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j^2}$
Intervalos de (1-α)×100% de conf. das funções	***S*(*t*) simétrico:** $\left[\hat{S}_{KM}(t) \pm z_{(1-\alpha/2)} \sqrt{\hat{Var}\left(\hat{S}_{KM}(t)\right)} \right]$ ***S*(*t*) log:** $\left[e^{\pm z_{(1-\alpha/2)} \sqrt{\hat{Var}\left(\hat{H}_{KM}(t)\right)}} \hat{S}_{KM}(t) \right]$ ***S*(*t*) log-log:** $\left[\left(\hat{S}_{KM}(t)\right)^{\exp\left\{ \pm z_{(1-\alpha/2)} \sqrt{\hat{Var}\left(\hat{U}_{KM}(t)\right)} \right\}} \right]$ ***H*(*t*) simétrico:** $\left[\hat{H}_{KM}(t) \pm z_{(1-\alpha/2)} \sqrt{\hat{Var}\left(\hat{H}_{KM}(t)\right)} \right]$ ***H*(*t*) log:** $\left[e^{\pm z_{(1-\alpha/2)} \sqrt{\hat{Var}\left(\hat{U}_{KM}(t)\right)}} \hat{H}_{KM}(t) \right]$	
p-ésimo quantil	$\hat{t}_p = t_{(u)} + \frac{[t_{(u+1)} - t_{(u)}][\hat{S}(t_{(u)}) - (1-p)]}{[\hat{S}(t_{(u)}) - \hat{S}(t_{(u+1)})]}$	
Média	$\overline{T} = \sum_{j=0}^{k} (t_{(j+1)} - t_{(j)}) \hat{S}(t_{(j)})$	
Vida média residual	$\hat{v}(t) = \frac{1}{\hat{S}(t)} \left[(t_{(u)} - t) \hat{S}(t) + \sum_{j:t_{(j)} \ge t} (t_{(j+1)} - t_{(j)}) \hat{S}(t_{(j)}) \right]$	
Variância das quantidades básicas	$\hat{Var}\left(\hat{t}_p\right) = \left[\frac{\hat{S}(t_{(u)}) - \hat{S}(t_{(u+1)})}{[t_{(u+1)} - t_{(u)}]} \right]^{-2} \hat{Var}\left(\hat{S}(\hat{t}_p)\right)$	
	$Var(\overline{T}_{KM}) = \frac{d}{d-1} \sum_{j=1}^{k} \left(\left[\sum_{l=j}^{k} (t_{(l+1)} - t_{(l)}) \hat{S}_{KM}(t_{(l)}) \right]^2 \frac{d_j}{n_j (n_j - d_j)} \right)$	$Var(\overline{T}_{NA}) = \frac{d}{d-1} \sum_{j=1}^{k} \left(\left[\sum_{l=j}^{k} (t_{(l+1)} - t_{(l)}) \hat{S}_{NA}(t_{(l)}) \right]^2 \frac{d_j}{n_j^2} \right)$
Intervalos de (1-α)×100% de conf. das quantidades básicas	***p*-ésimo quantil:** $\left[\hat{t}_p - z_{(1-\alpha/2)} \sqrt{\hat{Var}\left(\hat{t}_p\right)} \; ; \; \hat{t}_p + z_{(1-\alpha/2)} \sqrt{\hat{Var}\left(\hat{t}_p\right)} \right]$ **média:** $\left[\overline{T} - z_{(1-\frac{\alpha}{2})} \sqrt{Var(\overline{T})} \; ; \; \overline{T} + z_{(1-\frac{\alpha}{2})} \sqrt{Var(\overline{T})} \right]$	

4. ESTIMAÇÃO PARAMÉTRICA DE ALGUMAS DISTRIBUIÇÕES BÁSICAS DE SOBREVIVÊNCIA

O comportamento do tempo de sobrevivência pode estar associado à várias causas físicas, geralmente difíceis de serem representadas matematicamente. Neste contexto, apresentamos neste capítulo a análise paramétrica de várias distribuições de probabilidades que são usualmente utilizadas para modelar, de forma plausível, dados relacionados ao tempo até a ocorrência de um evento de interesse.

ESTIMAÇÃO PARAMÉTRICA

Apresentamos aqui os procedimentos de estimação de máxima verossimilhança para os parâmetros de algumas distribuições básicas de sobrevivência.

Considere a situação em que temos disponível uma amostra aleatória $t_1, t_2, ..., t_n$ de tempos de sobrevivência e os valores de uma variável indicadora δ_i, onde

$$\delta_i = \begin{cases} 0, & \text{se } t_i \text{ foi censurado} \\ 1, & \text{se } t_i \text{ foi observado} \end{cases} \qquad i = 1, 2, ..., n.$$

Note que cada elemento da amostra contribui para a função de verossimilhança da seguinte forma:

$$\underset{\text{para a função } L(\theta)}{\text{contrib. do elemento i da amostra}} = \begin{cases} f(t_i \mid \theta), & \text{se } t_i \text{ foi observado} \\ S(t_i \mid \theta), & \text{se } t_i \text{ foi censurado} \end{cases} \qquad i = 1, 2, ..., n.$$

Baseado na amostra $(t_1, \delta_1), (t_2, \delta_2), ..., (t_n, \delta_n)$, e considerando que os t_i's são independentes e provenientes da mesma distribuição de probabilidades indexada pelo parâmetro θ, a função de verossimilhança é genericamente escrita na forma:

$$L(\theta) \propto \prod_{i=1}^{n} [f(t_i|\theta)]^{\delta_i} [S(t_i|\theta)]^{1-\delta_i}$$

Os Estimadores de Máxima Verossimilhança (EMV) de θ é o valor $\hat{\theta}$ que maximiza a função $L(\theta)$. Detalhes sobre este procedimento são apresentados na próxima seção em que descreveremos a distribuição Exponencial

DISTRIBUIÇÃO EXPONENCIAL

A distribuição Exponencial é uma das mais simples e importantes distribuições de probabilidades utilizadas na modelagem de dados que representam o tempo até a ocorrência de algum evento de interesse. A mesma tem sido utilizada intensivamente na literatura de sobrevivência e confiabilidade assim como a distribuição Normal é utilizada em outras áreas da Estatística.

Uma variável aleatória não negativa e contínua, T, tem distribuição Exponencial com parâmetro λ ($\lambda>0$) se sua Função Densidade de Probabilidades for escrita da forma:

$$f(t) = \frac{1}{\lambda} e^{-\frac{t}{\lambda}}, \quad t \geq 0.$$

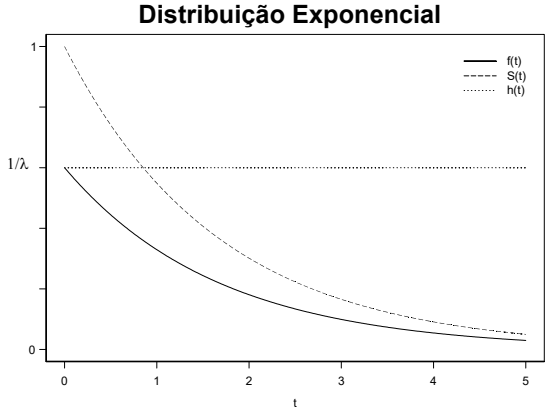
Notação: T ~ Exponencial(λ).

OBS: É comum encontrarmos na literatura diferentes parametrizações da distribuição Exponencial, porém a parametrização apresentada aqui é interessante pelo fato do parâmetro representar a média da variável aleatória.

É fácil mostrar que se T ~ Exponencial(λ), então:

i) $E[T] = \lambda$ e $Var[T] = \lambda^2$;
ii) $t_p = -\lambda\log(1 - p)$;
iii) $S(t) = e^{-\frac{t}{\lambda}}, \quad t \geq 0$;
iv) $h(t) = \frac{1}{\lambda}, \quad t \geq 0$.

Nota: Note que a Função de Risco, h(t), da distribuição Exponencial não depende do tempo (é constante ao longo do tempo), assim, a distribuição Exponencial é a única distribuição contínua (univariada) que apresenta a propriedade de falta de memória.



Pela propriedade da falta de memória da distribuição Exponencial, o que acarreta em risco constante, a mesma é adequada para situações onde a morte de um indivíduo, ou a falha de um equipamento independe da idade do mesmo, ou seja, a probabilidade de termos uma sobrevivência ou funcionamento adicional não se altera, dado que ainda não foi observado morte ou falha. Isto significa que, por exemplo, um equipamento com 1000 horas de funcionamento tem o mesmo risco de falha que um equipamento recém fabricado.

Considere que $t_1, t_2, ..., t_n$ é uma amostra aleatória de uma variável aleatória T, onde T~Exponencial(λ) e $\delta_1, \delta_2, ..., \delta_n$ os respectivos indicadores de censuras. Desta forma, a função de verossimilhança é dada por:

$$L(\lambda) \propto \prod_{i=1}^{n} \left[h(t_i \mid \theta)\right]^{\delta_i} S(t_i \mid \theta) = \prod_{i=1}^{n} \lambda^{-\delta_i} e^{-\frac{t_i}{\lambda}} = \lambda^{-\left(\sum_{i=1}^{n}\delta_i\right)} e^{\left(-\frac{1}{\lambda}\sum_{i=1}^{n}t_i\right)}$$

Aplicando o logaritmo de L(λ) temos

$$\ell(\lambda) = \ln L(\lambda) = -\ln(\lambda)\sum_{i=1}^{n}\delta_i - \frac{1}{\lambda}\sum_{i=1}^{n}t_i$$

Derivando $\ell(\lambda)$ com relação a λ e igualando a zero, temos como solução o estimador de máxima verossimilhança de λ, dado por

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} \delta_i}.$$

Note que em situações em que temos a ausência de observações censuradas o estimador de máxima verossimilhança é dado pela média dos tempos de vidas observados.

No caso particular de se trabalhar com a **censura do Tipo II**, temos que $r = \sum_{i=1}^{n} \delta_i$ **é fixo**. Assim, utilizando o fato de que

$$\frac{2r\hat{\lambda}}{\lambda} \sim \chi^2_{2r},$$

um intervalo de 100(1 – α)% de confiança para λ é obtido a partir de

$$P\left(\chi^2_{(2r;\alpha/2)} \leq \frac{2r\hat{\lambda}}{\lambda} \leq \chi^2_{(2r;1-\alpha/2)} \right) = 1-\alpha$$

onde $\chi^2_{(2r;\alpha/2)}$ é o quantil $\alpha/2$ de uma distribuição Qui-Quadrado com 2r graus de liberdade.

Desta forma, temos o seguinte intervalo de confiança (exato) para o parâmetro λ:

$$\underset{100\times(1-\alpha)\%}{I.C.(\lambda)}: \left[\frac{2r\hat{\lambda}}{\chi^2_{(2r;1-\alpha/2)}} ; \frac{2r\hat{\lambda}}{\chi^2_{(2r;\alpha/2)}} \right]$$

Quando grandes amostras são disponíveis, podemos utilizar a normalidade assintótica dos estimadores de máxima verossimilhança para obter um intervalo de confiança para λ, ou seja

$$\frac{\hat{\lambda} - E(\hat{\lambda})}{\sqrt{\text{var}(\hat{\lambda})}} = \frac{\hat{\lambda} - \lambda}{\sqrt{I_o^{-1}(\lambda)}} \overset{n\to\infty}{\sim} N(0,1)$$

onde $I_o(\lambda)$ é a *Informação de Fisher Observada*, definida por $I_o(\lambda) = -\left(\frac{d^2\ell(\lambda)}{(d\lambda)^2} \right)\Big|_{\lambda=\hat{\lambda}}$. Note que os estimadores de máxima verossimilhança são assintoticamente não-viciados, isto é, $E(\hat{\lambda}) \underset{n\to\infty}{\approx} \lambda$.

Para o modelo Exponencial temos

$$I_o(\lambda) = -\left(\frac{d^2\ell(\lambda)}{(d\lambda)^2} \right)\Big|_{\lambda=\hat{\lambda}} = -\left(\frac{\sum_{i=1}^{n}\delta_i}{\lambda^2} - \frac{2\sum_{i=1}^{n} t_i}{\lambda^3} \right)\Bigg|_{\lambda=\hat{\lambda}} = \frac{\sum_{i=1}^{n}\delta_i}{\hat{\lambda}^2}, \quad \text{e assim,} \quad \frac{\hat{\lambda}-\lambda}{\sqrt{\frac{\hat{\lambda}^2}{\sum_{i=1}^{n}\delta_i}}} \sim N(0,1).$$

Logo um intervalo de confiança (assintótico) para o parâmetro λ é dado por:

$$\underset{100\times(1-\alpha)\%}{I.C.(\lambda)}: \left[\hat{\lambda} - z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\frac{\hat{\lambda}^2}{\sum_{i=1}^{n}\delta_i}} \; ; \; \hat{\lambda} + z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\frac{\hat{\lambda}^2}{\sum_{i=1}^{n}\delta_i}} \right],$$

onde $z_{\left(1-\frac{\alpha}{2}\right)}$ é o quantil $\left(1-\frac{\alpha}{2}\right)$ de uma distribuição Normal padrão.

Pela propriedade de invariância dos estimadores de máxima verossimilhança, temos que os intervalos de confiança, exato e assintótico, para a Função de Sobrevivência S(t) são dados, respectivamente por (visto que a função de sobrevivência é uma função monótona decrescente).

Intervalo exato: $$\underset{100\times(1-\alpha)\%}{I.C.\big(S(t)\big)}: \left[e^{-\frac{t\chi^2_{(2r;1-\alpha/2)}}{2r\hat{\lambda}}} ; e^{-\frac{t\chi^2_{(2r;\alpha/2)}}{2r\hat{\lambda}}} \right]$$

Intervalo assintótico: $$\underset{100\times(1-\alpha)\%}{I.C.\big(S(t)\big)}: \left[e^{-t\left(\hat{\lambda}-z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\hat{\lambda}^2}{\sum_{i=1}^{n}\delta_i}}\right)^{-1}} ; e^{-t\left(\hat{\lambda}+z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\hat{\lambda}^2}{\sum_{i=1}^{n}\delta_i}}\right)^{-1}} \right]$$

Exemplo 4.1. *Considere os dados do tempo até a falha (em horas) de um certo equipamento eletrônico. Ajuste o modelo exponencial para esta amostra.*

1	1	2	2	3	3	3	4	5^+	6	6	6	8^+	9	10
12^+	12	12	12	13	13	16	18	19	22	22	26	29	39^+	42^+

+: censura

Para este exemplo temos: n = 30, $\sum_{i=1}^{30} t_i = 376$ e $\sum_{i=1}^{30} \delta_i = 25$. Que resulta em $\hat{\lambda} = \frac{376}{25} = 15{,}04$ e

$$\underset{95\%}{I.C.(\lambda)}: \left[15{,}04 - 1{,}96\sqrt{\frac{(15{,}04)^2}{25}} ; 15{,}04 - 1{,}96\sqrt{\frac{(15{,}04)^2}{25}} \right] = [9{,}1443 ; 20{,}9357]$$

Assim pelos dados da amostra, podemos dizer que uma estimativa para o tempo médio de vida destes equipamentos eletrônicos é de 15,04 horas (I.C. 95% [9,14 ; 20,94]).

As estimativas da Função de Sobrevivência são apresentadas na tabela e figura abaixo:

Tabela: Estimativas das Funções de Sobrevivência para os dados do Exemplo 4.1.

Kaplan-Meier		**Modelo Exponencial**		
Tempo	$\hat{S}_{KM}(t+0)$	$\hat{S}(t)$	**IC 95% (lim inferior)**	**IC 95% (lim superior)**
1	0,9333	0,9357	0,8964	0,9534
2	0,8667	0,8755	0,8036	0,9089
3	0,7667	0,8192	0,7203	0,8665
4	0,7333	0,7665	0,6457	0,8261
6	0,6286	0,6710	0,5188	0,7508
9	0,5916	0,5497	0,3737	0,6506
10	0,5546	0,5143	0,3350	0,6202
12	0,4437	0,4503	0,2692	0,5637
13	0,3630	0,4213	0,2413	0,5374
16	0,3227	0,3451	0,1738	0,4657
18	0,2824	0,3022	0,1397	0,4233
19	0,2420	0,2827	0,1252	0,4035
22	0,1613	0,2316	0,0902	0,3496
26	0,1210	0,1775	0,0582	0,2888
29	0,0807	0,1454	0,0419	0,2503

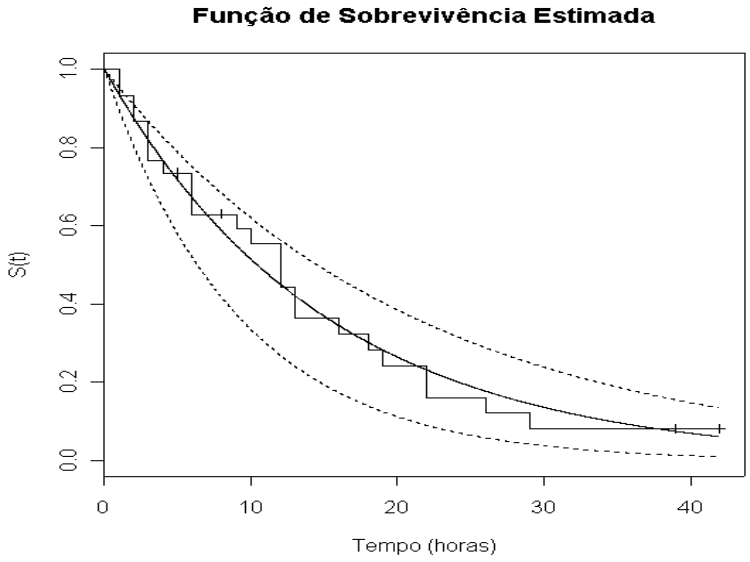


Figura: Gráfico das estimativas de S(t) para os dados do Exemplo 4.1.
A função escada é a estimativa de Kaplan-Meier
(___) estimativa obtida pelo modelo Exponencial
(----) I.C. 95% de S(t) pelo modelo Exponencial

DISTRIBUIÇÃO WEIBULL

A distribuição Weibull foi proposta originalmente por Wallodi Weibull em 1951 e desde então, devido em grande parte à sua simplicidade e flexibilidade, tem sido uma das distribuições de probabilidades mais utilizadas na modelagem de dados biomédicos como também industriais. Sua função de densidade, dotada de dois parâmetros positivos é definida por:

$$f(t) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \exp\left\{-\left(\frac{t}{\lambda}\right)^{\alpha}\right\}, \quad t \geq 0$$

onde α, $\lambda > 0$ são os parâmetros de forma e escala, respectivamente.

Notação: T ~ Weibull(α,λ).

A distribuição Weibull é uma extensão da distribuição Exponencial. Se tomarmos α=1, obtemos a distribuição Exponencial como caso particular.

É comum encontrarmos na literatura a distribuição Weibull escrita sob diferentes parametrizações, porém, a parametrização apresentada aqui permite uma interpretação direta do parâmetro de escala: λ é aproximadamente o percentil 63% da distribuição da variável aleatória T.

Se T ~ Weibull(α,λ), então:

i) $E[T] = \lambda\Gamma\left(1+\frac{1}{\alpha}\right)$ e $Var[T] = \lambda^2\left(\Gamma\left(1+\frac{2}{\alpha}\right) - \left[\Gamma\left(1+\frac{1}{\alpha}\right)\right]^2\right)$,

onde $\Gamma(.)$ denota a função gama, definida por: $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$;

ii) $t_p = \lambda\left[-\ln(1-p)\right]^{\frac{1}{\alpha}}$;

iii) $S(t) = \exp\left\{-\left(\frac{t}{\lambda}\right)^{\alpha}\right\}, \quad t \geq 0$;

iv) $h(t) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1}, \quad t \geq 0$.

Uma característica importante da distribuição Weibull na modelagem de tempos de sobrevivência está relacionada à sua flexibilidade em acomodar diferentes formas de funções de risco. Diferentemente da distribuição Exponencial cuja função de risco é constante, na distribuição Weibull para o parâmetro de forma $\alpha<1$ temos funções de risco monótonas decrescentes, para $\alpha>1$ as funções de risco são monótonas crescentes e para $\alpha=1$ temos a distribuição Exponencial com função de risco constante. A figura abaixo apresenta as funções densidade e de risco para alguns valores do parâmetro de forma da distribuição Weibull.

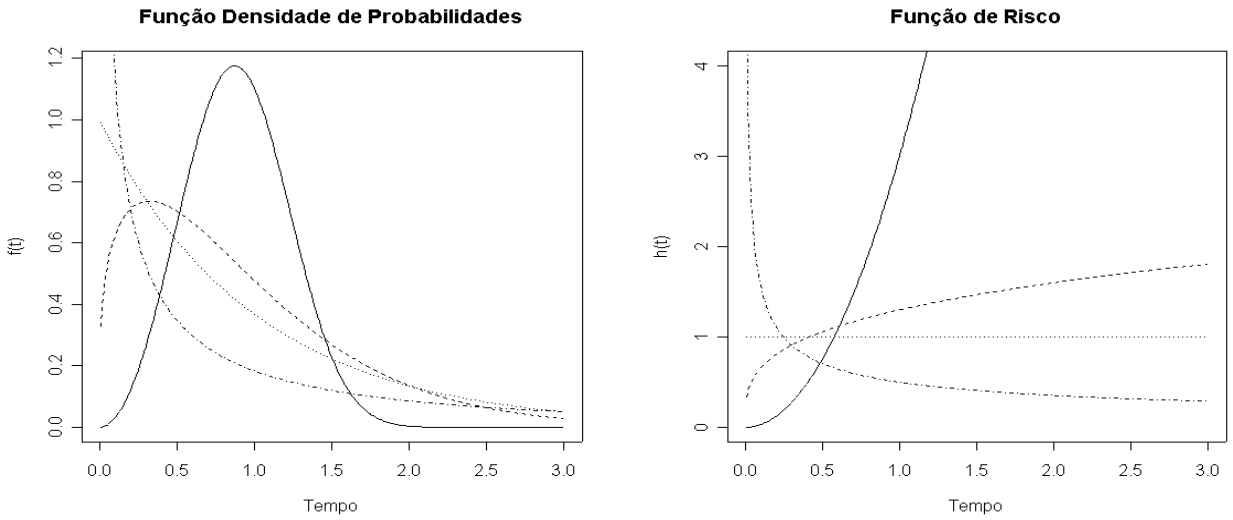


Figura: Funções densidade e de risco para diferentes valores do parâmetro de forma da distribuição Weibull. (___) $\alpha = 3$; (----) $\alpha = 1{,}5$; (.....) $\alpha = 1$; (-·-·-·) $\alpha = 0{,}5$.

Para a distribuição Weibull com parâmetros α e λ, considerando uma amostra aleatória $t_1,t_2,...,t_n$ e a variável indicadora de censura δ_i, onde $\delta_i=1$ se t_i é exatamente observado ou $\delta_i=0$ se t_i é censurado à direita, a função de verossimilhança é escrita na forma

$$L(\alpha,\lambda)=\prod_{i=1}^{n}\left[f(t_i)\right]^{\delta_i}\left[S(t_i)\right]^{1-\delta_i}=\prod_{i=1}^{n}\left[\frac{\alpha}{\lambda}\left(\frac{t_i}{\lambda}\right)^{\alpha-1}\exp\left\{-\left(\frac{t_i}{\lambda}\right)^{\alpha}\right\}\right]^{\delta_i}\left[\exp\left\{-\left(\frac{t_i}{\lambda}\right)^{\alpha}\right\}\right]^{1-\delta_i}$$

E o logaritmo da verossimilhança é dado por

$$\log L(\alpha,\lambda)=\sum_{i=1}^{n}\left[\delta_i\log f(t_i)+(1-\delta_i)\log S(t_i)\right]$$

Os estimadores de máxima verossimilhança dos parâmetros α e λ são obtidos resolvendo iterativamente o seguinte sistema de equações não lineares (que não tem uma solução explícita).

$$\frac{\partial\, log\, L\,(\alpha,\lambda))}{\partial\alpha}=\alpha^{-1}\sum_{i=1}^{n}\delta_i+\sum_{i=1}^{n}\delta_i\, ln\left(\frac{t_i}{\lambda}\right)-\left[\sum_{i=1}^{n}\left(\frac{t_i}{\lambda}\right)^{\alpha} ln\left(\frac{t_i}{\lambda}\right)\right]=0$$

$$\frac{\partial\, log\, L\,(\alpha,\lambda))}{\partial\lambda}=-\frac{\alpha}{\lambda}\sum_{i=1}^{n}\delta_i+\frac{\alpha}{\lambda}\sum_{i=1}^{n}\left(\frac{t_i}{\lambda}\right)^{\alpha}=0$$

Resolvendo o sistema em λ, temos como solução

$$\hat{\lambda} = \left(\frac{\sum_{i=1}^{n} t_i^{\alpha}}{\sum_{i=1}^{n} \delta_i} \right)^{\frac{1}{\alpha}}$$

que é o estimador de máxima verossimilhança de λ, para α conhecido.

Para obtermos $\hat{\alpha}$, resolvemos numericamente em α a primeira equação do sistema. Um método que pode ser utilizado é o de Newton-Raphson.

No software R, as estimativas de máxima verossimilhança podem ser obtidas por meio da rotina "nlm" (non-linear minimization), uma vez especificadas as funções de verossimilhança ou log-verossimilhança.

O exemplo a seguir mostrará como obter as estimativas dos parâmetros do modelo Weibull usando o software R.

Exemplo 4.2. *Considere os dados obtidos sobre o tempo de resistência (em horas) de 15 filmes isolantes.*

50,25	87,75	87,76	87,77	92,90
92,91	95,96	108,30	108,30	117,90
123,90	124,30	129,70	135,60	135,60

Obs: dados com ausência de censuras.

Para obter as estimativas dos parâmetros do modelo Weibull para este conjunto de dados usaremos os seguintes comandos no R:

```
like.weibull<-function(parametro,tempo,delta){
  L1<-dweibull(tempo,parametro[1],parametro[2],log=TRUE)
  L2<-pweibull(tempo,parametro[1],parametro[2],log.p=TRUE,lower.tail=FALSE)
  -sum(L1*delta + L2*(1-delta) ) }

tempo<-c(50.25,87.75,87.76,87.77,92.9,92.91,95.96,108.3,108.3,117.9,123.9,124.3,129.7,135.6,135.6)
delta<-rep(1,15)
chute.inicial<-c(1,1)
emv<-nlm(like.weibull,chute.inicial,hessian=TRUE,tempo=tempo,delta=delta)
emv
solve(emv$hessian)      # fornece a matriz de variância-covariância de α e λ

# Obtenção das estimativas de S(t), via Kaplan-Meier e modelo Weibull.

library(survival)
dados<-Surv(tempo,delta)
km<-survfit(dados~1)
plot(km,conf.int=F,main="Função de Sobrevivência Estimada",xlab="Tempo",ylab="S(t)")
t<-(1:14000)/100
estimativa<-emv$estimate
s.weibull<-pweibull(t,estimativa[1],estimativa[2],lower.tail=F)
points(t,s.weibull,type="l",lty=2)
legend(10,0.2,c("___  Kaplan-Meier","------ Weibull"))
```

A inversa da matriz hessiana é a matriz de variância-covariância dos estimadores de α e λ, isto é,

$$\text{hessiana}^{-1} = \begin{bmatrix} \text{Var}(\hat{\alpha}) & \text{Cov}(\hat{\alpha},\hat{\lambda}) \\ & \text{Var}(\hat{\lambda}) \end{bmatrix}$$

Com as variâncias dos estimadores, podemos construir os intervalos (assintóticos) para os parâmetros do modelo Weibull. Cabe lembrar que no modelo Weibull estamos trabalhando com dois parâmetros, α e λ, portanto o correto seria construir uma *região de confiança* para os parâmetros. Mas usualmente o que se faz é construir os intervalos de confiança para cada parâmetro. Na hora de calcular os intervalos é importante "ajustar" o coeficiente de confiança, visto que tempos dois intervalos (simultâneos). Assim, os intervalos de $100(1-\beta)\%$ de confiança para os parâmetros do modelo Weibull são dados por

$$\underset{100\times(1-\beta)\%}{I.C.(\alpha)}: \left[\hat{\alpha} - z_{\left(1-\frac{\beta^*}{2}\right)} \sqrt{\text{var}(\hat{\alpha})} \; ; \; \hat{\alpha} + z_{\left(1-\frac{\beta^*}{2}\right)} \sqrt{\text{var}(\hat{\alpha})} \right]$$

e

$$\underset{100\times(1-\beta)\%}{I.C.(\lambda)}: \left[\hat{\lambda} - z_{\left(1-\frac{\beta^*}{2}\right)} \sqrt{\text{var}(\hat{\lambda})} \; ; \; \hat{\lambda} + z_{\left(1-\frac{\beta^*}{2}\right)} \sqrt{\text{var}(\hat{\lambda})} \right], \qquad \text{onde } (1-\beta^*) = \sqrt{1-\beta}\,.$$

Os resultados do Exemplo 4.2 são dados por:

$$\hat{\alpha} = 5{,}6492 \quad \text{Var}(\hat{\alpha}) = 1{,}4070$$
$$\hat{\lambda} = 114{,}0556 \quad \text{Var}(\hat{\lambda}) = 29{,}9472$$

Assim, a estimativa da média e da variância do tempo de resistência deste filme são dadas por:

$$\hat{E}[T] = \hat{\lambda}\Gamma\left(1+\tfrac{1}{\hat{\alpha}}\right) = 114{,}0556\,\Gamma\left(1+\tfrac{1}{5{,}6492}\right) = 105{,}4559$$

e

$$\hat{\text{Var}}[T] = \lambda^2\left(\Gamma\left(1+\tfrac{2}{\alpha}\right) - \left[\Gamma\left(1+\tfrac{1}{\alpha}\right)\right]^2\right) = 466{,}5313$$

Logo, um intervalo de 95% de confiança (assintótico) para o tempo médio de resistência pode ser dado por

$$\underset{95\%}{\text{I.C.}}: \left[\hat{E}[T] \pm z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\hat{\text{Var}}[T]/\textstyle\sum \delta_i}\;\right] = [105{,}4559 \pm 1{,}96\sqrt{466{,}5313/15} = [94{,}53\,;\,116{,}37]$$

Para construir um intervalo (simultâneo) de 95% de confiança para os parâmetros, temos que ajustar o coeficiente de confiança:

$(1 - \beta) = 0{,}95 \quad \Rightarrow \quad (1 - \beta^*) = (0{,}95)^{0{,}5} = 0{,}9747$

Assim, $z_{\left(1-\frac{\beta^*}{2}\right)} = z_{(0{,}9874)} = 2{,}238$. Logo,

$$\underset{95\%(\text{simultâneo})}{\text{I.C.}(\alpha)} : \left[5{,}6492 - 2{,}238\sqrt{1{,}407}\ ;\ 5{,}6492 + 2{,}238\sqrt{1{,}407}\right] = [2{,}9946\ ;\ 8{,}3038]$$

$$\underset{95\%(\text{simultâneo})}{\text{I.C.}(\lambda)} : \left[114{,}0556 - 2{,}238\sqrt{29{,}9472}\ ;\ 114{,}0556 + 2{,}238\sqrt{29{,}9472}\right] = [101{,}8084\ ;\ 126{,}3028]$$

Para valores grandes de $\text{Corr}(\hat{\alpha}, \hat{\lambda})$ os intervalos simultâneos não são indicados (visto que ele supõe independência entre os parâmetros), sendo neste caso o mais adequado construir uma Região de Confiança.

Neste exemplo, $Cov(\hat{\alpha}, \hat{\lambda}) = 1{,}9668$, resultando em $Corr(\hat{\alpha}, \hat{\lambda}) = \dfrac{1{,}9668}{\sqrt{1{,}407}\sqrt{29{,}9472}} = 0{,}303$.

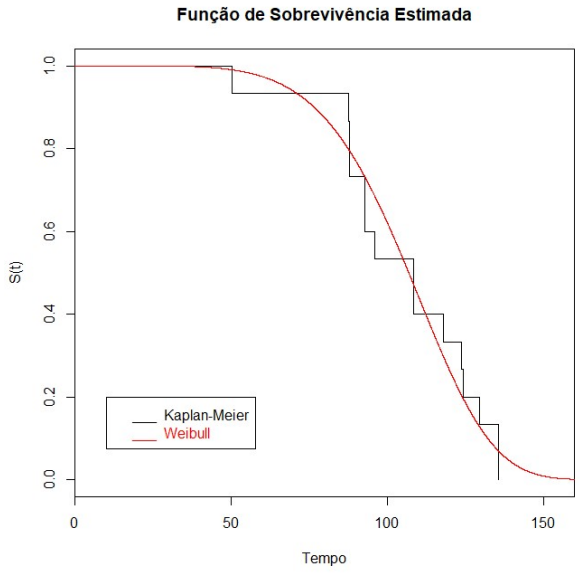
Neste caso, os intervalos apresentados podem ser considerados.

Suponha que $\theta = (\alpha, \lambda)$, como no modelo Weibull. Algumas vezes o interesse é estimar uma função do parâmetro $g(\alpha, \lambda)$. Como por exemplo a função de sobrevivência $S(t; \alpha, \lambda)$.

Pela propriedade de invariância dos estimadores de sobrevivência, o estimador pontual da função de sobrevivência no tempo t é dado por

$$\hat{S}(t; \alpha, \lambda) = S(t; \hat{\alpha}, \hat{\lambda}) = \exp\left\{-\left(\frac{t}{\hat{\lambda}}\right)^{\hat{\alpha}}\right\}, \quad t \geq 0$$

Assim, a estimativa da Função de Sobrevivência para os dados do Exemplo 4.2 é apresentada pela figura abaixo



Se além de estimar $g(\alpha,\lambda)$, o interesse é construir um intervalo de confiança, é necessário obter uma estimativa para o erro padrão de $\hat{g}(\alpha,\lambda)$. Isto pode ser feito a partir do *método delta*.

Para o caso univariado, considere que θ é um escalar e que já interesse em obter $Var(g(\hat{\theta}))$. Expandindo-se $g(\hat{\theta})$ em torno de $E(\hat{\theta}) \cong \theta$ e ignorando-se os termos superiores ao de primeira ordem tem-se que

$$\boldsymbol{g(\hat{\theta}) \cong g(\theta) + (\hat{\theta} - \theta)\left(\frac{dg(\theta)}{d\theta}\right)}$$

E assim,

$$\boldsymbol{\widehat{Var}[g(\hat{\theta})] \cong Var(\hat{\theta})\left(\frac{dg(\theta)}{d\theta}\Big|_{\theta=\hat{\theta}}\right)^2}.$$

Para o caso bivariado, em que $\theta = (\alpha,\lambda)$ e que há interesse em $g(\alpha,\lambda)$, a versão multivariada do método delta resulta em

$$\begin{aligned}\boldsymbol{\widehat{Var}[g(\hat{\alpha},\hat{\lambda})] \cong Var(\hat{\alpha})}&\boldsymbol{\left(\frac{\partial g(\alpha,\lambda)}{\partial \alpha}\Big|_{\substack{\alpha=\hat{\alpha}\\ \lambda=\hat{\lambda}}}\right)^2}\\ &\boldsymbol{+2Cov(\hat{\alpha},\hat{\lambda})\left(\frac{\partial g(\alpha,\lambda)}{\partial \alpha}\Big|_{\substack{\alpha=\hat{\alpha}\\ \lambda=\hat{\lambda}}}\right)\left(\frac{\partial g(\alpha,\lambda)}{\partial \lambda}\Big|_{\substack{\alpha=\hat{\alpha}\\ \lambda=\hat{\lambda}}}\right)}\\ &\boldsymbol{+Var(\hat{\lambda})\left(\frac{\partial g(\alpha,\lambda)}{\partial \lambda}\Big|_{\substack{\alpha=\hat{\alpha}\\ \lambda=\hat{\lambda}}}\right)^2}\end{aligned}$$

Assim, um intervalo de confiança (assintótico) para uma função $g(\alpha,\lambda)$é dado por:

$$I.C._{100\times(1-\beta)\%}\left(g(\alpha,\lambda)\right): \left[\, g(\hat{\alpha},\hat{\lambda}) \pm z_{\left(1-\frac{\beta}{2}\right)}\sqrt{Var\left[g(\hat{\alpha},\hat{\lambda})\right]}\right],$$

onde $z_{\left(1-\frac{\beta}{2}\right)}$ é o quantil $\left(1-\frac{\beta}{2}\right)$ de uma distribuição Normal padrão.

O software R, tem implementado o método delta por meio do comando *deltavar* da biblioteca "emdbook".

O exemplo a seguir mostrará como obter os intervalos de confiança da função de sobrevivência do modelo Weibull usando o software R.

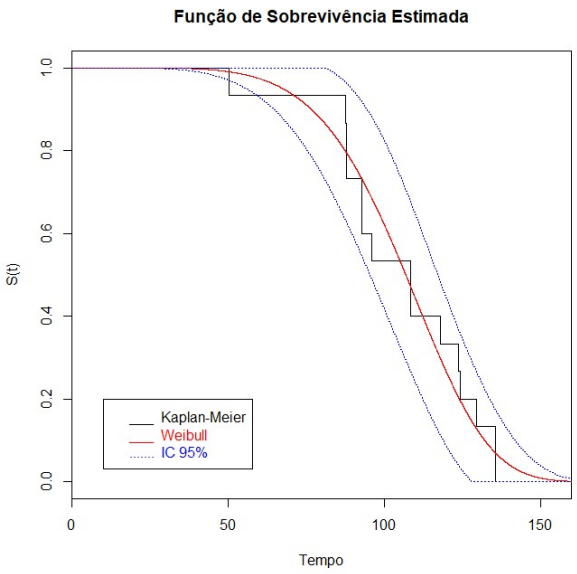
```
require(emdbook)
#### calculo da variância de S(t)
var.sob<-numeric(length(t))
for (i in 1:length(t)) {
  x<-t[i]
  var.sob[i]<-deltavar(fun = exp(-1*(x/lambda)^alpha),Sigma=Sigma,
meanval=c(alpha=estimativa[1],lambda=estimativa[2]))
}

#### Construindo IC 95%
LI.sob<-s.weibull - qnorm(0.975)*var.sob^.5
LS.sob<-s.weibull + qnorm(0.975)*var.sob^.5

### Intervalo simétrico (truncando em 0 e 1)
LI.sob[which(LI.sob<0)]<-0
LS.sob[which(LS.sob>1)]<-1

points(t,LI.sob,type="l",lty=3,col=4)
points(t,LS.sob,type="l",lty=3,col=4)
```

A estimativa da Função de Sobrevivência para os dados do Exemplo 4.2 com seu respectivo intervalo de 95% de confiança é apresentada pela figura abaixo.



DISTRIBUIÇÃO LOG-NORMAL

Assim como a distribuição Weibull, uma outra distribuição de probabilidade bastante utilizada na modelagem de dados de sobrevivência é a distribuição Log-Normal. Dizemos que uma variável aleatória T tem distribuição Log-Normal com parâmetros μ e σ, se Y=log(T) tem uma distribuição Normal. A função densidade de T é escrita na forma:
Sua função de densidade, dotada de dois parâmetros positivos é definida por:

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

onde $-\infty<\mu<\infty$ e $\sigma>0$ são os parâmetros de escala e forma, respectivamente.
Notação: T ~ Log-Normal(μ,σ^2).

Se T ~ Log-Normal(μ,σ^2) então:

i) $E[T] = \exp\left\{\mu + \frac{\sigma^2}{2}\right\}$ e $Var[T] = \exp\{2\mu + \sigma^2\}(\exp\{\sigma^2\} - 1)$;

ii) $t_p = \exp\{\mu + \sigma z_p\}$,

onde z_p é o p-ésimo quantil da distribuição Normal Padrão;

iii) $S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right), \quad t > 0$,

onde Φ(.) é a Função Distribuição Acumulada da Normal Padrão;

iv) $h(t) = \dfrac{\frac{1}{\sigma t\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right\}}{1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)}, \quad t > 0$

Um problema de se trabalhar com a distribuição Log-Normal está relacionado ao fato da mesma não apresentar funções de sobrevivência e de risco analiticamente explícitas. Ambas as funções são escritas em termos da função distribuição acumulada da distribuição Normal Padrão.

Uma característica importante da distribuição Log-Normal na modelagem de tempos de sobrevivência está relacionada ao fato dela permitir acomodar funções de risco unimodais, o que pode ser adequado em algumas situações práticas. A figura abaixo apresenta as funções densidade e de risco para alguns valores do parâmetro de forma da distribuição Log-Normal.

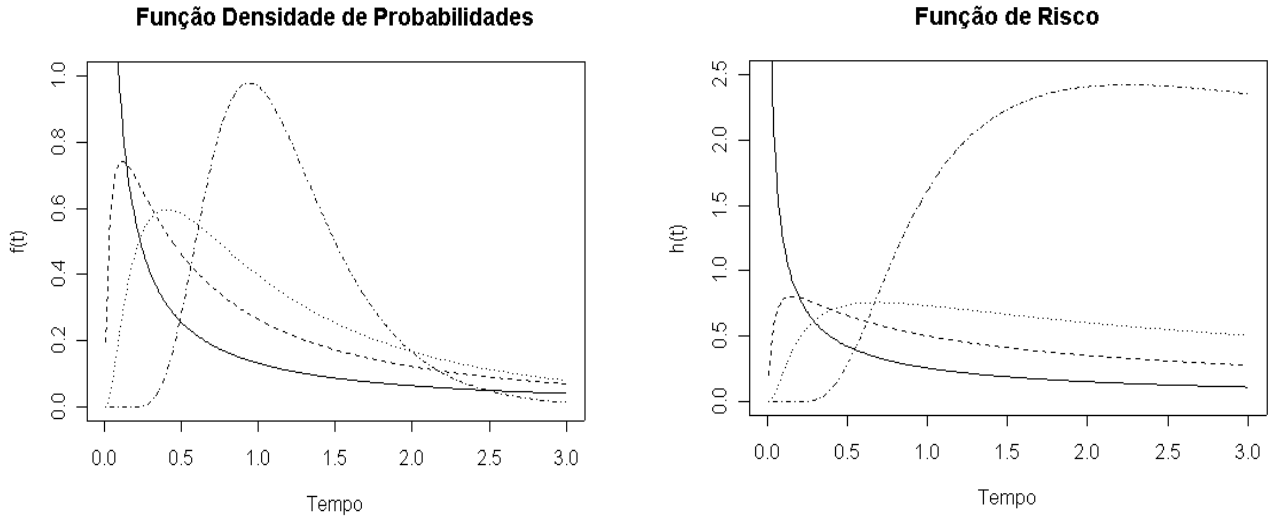


Figura: Funções densidade e de risco para diferentes valores do parâmetro de forma da distribuição Log-Normal. (___) σ = 3; (----) σ = 1,5; (.....) σ = 1; (-·-·-·) σ = 0,4.

Para a distribuição Log-Normal com parâmetros μ e σ, considerando uma amostra aleatória $t_1,t_2,...,t_n$ e a variável indicadora de censura δ_i, onde $\delta_i=1$ se t_i é exatamente observado ou $\delta_i=0$ se t_i é censurado à direita, a função de verossimilhança é escrita na forma

$$L(\mu,\sigma)=\prod_{i=1}^{n}\left[\frac{1}{\sigma t_i\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{\log t_i-\mu}{\sigma}\right)^2\right\}\right]^{\delta_i}\left[1-\Phi\left(\frac{\log t_i-\mu}{\sigma}\right)\right]^{1-\delta_i}$$

$$=\exp\left\{-\frac{1}{\sigma\sqrt{2\pi}}\sum_{i=1}^{n}\delta_i\log t_i-\frac{1}{2}\sum_{i=1}^{n}\delta_i\left(\frac{\log t_i-\mu}{\sigma}\right)^2+\sum_{i=1}^{n}(1-\delta_i)\ln\left[1-\Phi\left(\frac{\log t_i-\mu}{\sigma}\right)\right]\right\}$$

Como no modelo Weibull, os estimadores dos parâmetros μ e σ não podem ser encontrados em sua forma explícita.

Para obtermos $\hat{\mu}$ e $\hat{\sigma}$ maximizaremos a função de verossimilhança numericamente utilizando o método de Newton-Raphson.

O exemplo a seguir mostrará como obter as estimativas dos parâmetros do modelo Log-Normal usando o software R (comando *nlm*).

Exemplo 4.3. *Considere a seguinte amostra de tempos de sobrevivência.*

0,1; 0,3$^+$; 0,3; 0,4; 0,4; 0,5$^+$; 0,5; 0,6; 0,6; 0,7; 0,8; 1,0$^+$; 1,0; 1,1; 1,2$^+$;
1,3; 1,8$^+$; 2,6; 2,9; 4,3; 5,6; 6,6; 10,7; 11,9; 12,5; 26,1; 65,0; 86,9; 90,5; 120,6

+: censura

Para obter as estimativas dos parâmetros do modelo Log-Normal para este conjunto de dados usaremos os seguintes comandos no R:

```
like.lognormal<-function(parametro,tempo,delta){
  L1<-dlnorm(tempo,parametro[1],parametro[2],log=TRUE)
  L2<-plnorm(tempo,parametro[1],parametro[2],log.p=TRUE,lower.tail=FALSE)
  -sum(L1*delta + L2*(1-delta) ) }
tempo<-c(0.1,0.3,0.3,0.4, ...,90.5,120.6)
delta<-c(1,0,1,....,1,1)
chute.inicial<-c(1,1)
emv<-nlm(like.lognormal,chute.inicial,hessian=TRUE,tempo=tempo,delta=delta)
emv
solve(emv$hessian)      # fornece a matriz de variância-covariância de μ e σ
## Obtenção das estimativas de S(t), via Kaplan-Meier e modelo Log-Normal.
library(survival)
dados<-Surv(tempo,delta)
km<-survfit(dados~1)
plot(km,conf.int=F,main="Função de Sobrevivência Estimada",xlab="Tempo",ylab="S(t)")
t<-(1:14000)/100
estimativa<-emv$estimate
s.lognormal<-plnorm(t,estimativa[1],estimativa[2],lower.tail=F)
points(t,s.lognormal,type="l",lty=2)
legend(70,0.9,c("___  Kaplan-Meier","------ Log-Normal"))
```

Os resultados do Exemplo 4.3 são dados por:

$\hat{\mu}=1{,}2864 \qquad \mathrm{Var}(\hat{\mu})=0{,}1386$

$\hat{\sigma}=1{,}9316 \qquad \mathrm{Var}(\hat{\sigma})=0{,}0716$

Assim, as estimativas da média e da variância do tempo de resistência deste filme são dadas por:

$$E[T] = \exp\left\{\mu + \frac{\sigma^2}{2}\right\} = \exp\left\{1{,}2864 + \frac{(1{,}9316)^2}{2}\right\} = 23{,}38$$

e

$$Var[T] = \exp\left\{2(1{,}2864) + (1{,}9316)^2\right\}\left(\exp\left\{(1{,}9316)^2\right\} - 1\right) = 22263{,}38$$

Logo, um intervalo de 95% de confiança (assintótico) para o tempo médio de resistência pode ser dado por

$$\underset{95\%}{I.C.} : \left[\hat{E}[T] \pm z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\hat{Var}[T]/\sum \delta_i}\ \right] = [23{,}38 \pm 1{,}96\sqrt{22263{,}38/25} = [0\ ;\ 81{,}87]$$

Para construir um intervalo (simultâneo) de 95% de confiança para os parâmetros, temos que ajustar o coeficiente de confiança (como foi feito na distribuição Weibull):

$$(1 - \beta) = 0{,}95 \quad \Rightarrow \quad (1 - \beta^*) = (0{,}95)^{0{,}5} = 0{,}9747$$

Assim, $z_{\left(1-\frac{\beta^*}{2}\right)} = z_{(0{,}9874)} = 2{,}238$. Logo,

$$\underset{95\%(\,simultâneo\,)}{I.C.}(\mu): \left[1{,}2864 \pm 2{,}238\sqrt{0{,}1386}\right] = [0{,}4532\ ;\ 2{,}1196]$$

e

$$\underset{95\%(\,simultâneo\,)}{I.C.}(\sigma): \left[1{,}9316 \pm 2{,}238\sqrt{0{,}0716}\right] = [1{,}3328\ ;\ 2{,}5304]$$

Para valores grandes de $\mathrm{Corr}(\hat{\mu}, \hat{\sigma})$ os intervalos simultâneos não são indicados (visto que ele supõe independência entre os parâmetros), sendo neste caso o mais adequado construir uma Região de Confiança.

Neste exemplo, $\mathrm{Cov}(\hat{\mu}, \hat{\sigma}) = 0{,}0092$, resultando em $Corr(\hat{\mu}, \hat{\sigma}) = \dfrac{0{,}0092}{\sqrt{0{,}1386}\sqrt{0{,}0716}} = 0{,}0924$.

Neste caso, os intervalos apresentados podem ser considerados.

Assim como no modelo Weibull, a variância da estimativa da função de sobrevivência do modelo log-normal pode ser obtida pelo método delta, calculada pelo comando *deltavar* da biblioteca "emdbook" do R.

O exemplo a seguir mostrará como obter os intervalos de confiança da função de sobrevivência do modelo log-normal usando o software R.

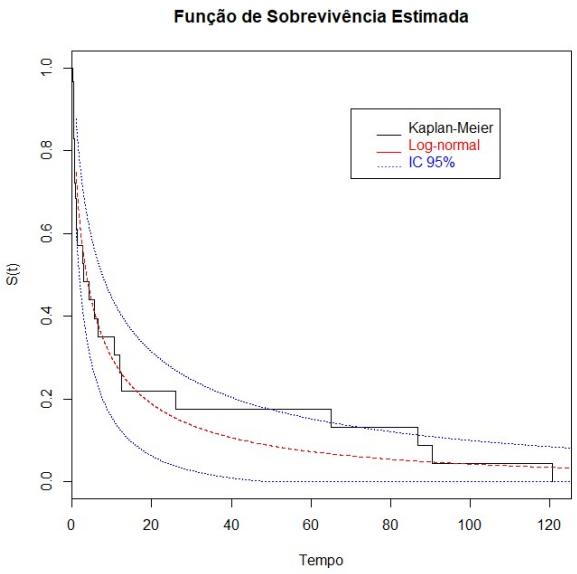
```
require(emdbook)
#### calculo da variância de S(t)
var.sob<-numeric(length(t))
for (i in 1:length(t)) {
  x<-t[i]
  var.sob[i]<-deltavar(fun = 1-pnorm((log(x)-mu)/sigma),Sigma=Sigma,
meanval=c(mu=estimativa[1],sigma=estimativa[2]))
}

#### Construindo IC 95%
LI.sob<-s.lognormal - qnorm(0.975)*var.sob^.5
LS.sob<-s.lognormal + qnorm(0.975)*var.sob^.5

### Intervalo simétrico (truncando em 0 e 1)
LI.sob[which(LI.sob<0)]<-0
LS.sob[which(LS.sob>1)]<-1

points(t,LI.sob,type="l",lty=3,col=4)
points(t,LS.sob,type="l",lty=3,col=4)
```

A estimativa da Função de Sobrevivência para os dados do Exemplo 4.3 com seu respectivo intervalo de 95% de confiança é apresentada pela figura abaixo.



DISTRIBUIÇÃO LOG-LOGÍSTICA

A distribuição Log-Logística é outra distribuição de sobrevivência que tem sido bastante utilizada na prática.

Dizemos que uma variável aleatória não negativa T segue uma distribuição Log-Logística com parâmetros μ e β, se seu logaritmo, Y = log(T), tem distribuição logística. A função densidade de T é escrita na forma:

$$f(t) = \frac{\beta\left(\frac{t}{\mu}\right)^{\beta-1}}{\mu\left[1+\left(\frac{t}{\mu}\right)^{\beta}\right]^2}, \quad t>0\,.$$

onde μ>0 e β>0 são os parâmetros de escala e forma, respectivamente.

Notação: T ~ Log-Logística(μ,β).

Se T ~ Logística(μ,β), então:

i) $E[T] = \frac{\mu\left(\frac{\pi}{\beta}\right)}{sen\left(\frac{\pi}{\beta}\right)}$, se β>1;

ii) $Var[T] = \mu^2\left(\frac{2\left(\frac{\pi}{\beta}\right)}{sen\left(\frac{2\pi}{\beta}\right)} - \frac{\left(\frac{\pi}{\beta}\right)^2}{\left[sen\left(\frac{\pi}{\beta}\right)\right]^2}\right)$, se β>2.

As Funções de Sobrevivência e de Risco, e o percentil da distribuição Log-Logística são dadas respectivamente por:

$$S(t) = \frac{1}{1+\left(\frac{t}{\mu}\right)^{\beta}}, \quad t>0\,, \qquad h(t) = \frac{\beta\left(\frac{t}{\mu}\right)^{\beta-1}}{\mu\left[1+\left(\frac{t}{\mu}\right)^{\beta}\right]}, \quad t>0 \qquad \text{e} \qquad t_p = \mu\left[\frac{p}{1-p}\right]^{\frac{1}{\beta}}$$

Assim como a Log-Normal, a distribuição Log-Logística também permite acomodar funções de risco unimodais. Entretanto, uma vantagem da distribuição Log-Logística sobre a distribuição Log-Normal relaciona-se ao fato da primeira apresentar expressões explícitas tanto para a Função de Sobrevivência quanto para a Função de Risco. A figura abaixo apresenta as funções densidade e de risco para alguns valores do parâmetro de forma da distribuição Log-Normal.

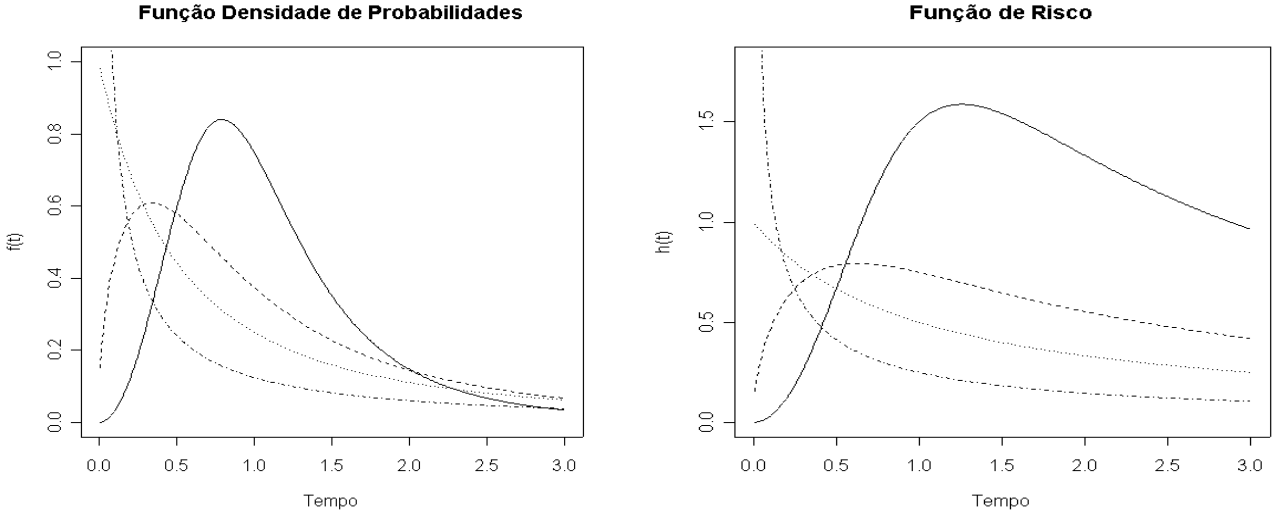


Figura: Funções densidade e de risco para diferentes valores do parâmetro de forma da distribuição Log-Logística. (___) β = 3; (----) β = 1,5; (.....) β = 1; (-·-·-) β = 0,5.

Para a distribuição Log-Logística com parâmetros μ e β, considerando uma amostra aleatória $t_1,t_2,...,t_n$ e a variável indicadora de censura δ_i, onde $\delta_i=1$ se t_i é exatamente observado ou $\delta_i=0$ se t_i é censurado à direita, a função de verossimilhança é escrita na forma

$$L(\mu,\beta)=\prod_{i=1}^{n}\left[\frac{\beta\left(\frac{t}{\mu}\right)^{\beta-1}}{\mu\left[1+\left(\frac{t}{\mu}\right)^{\beta}\right]}\right]^{\delta_i}\left[\frac{1}{1+\left(\frac{t}{\mu}\right)^{\beta}}\right]^{1-\delta_i}$$

Assim como no modelo Weibull e Log-Normal, os estimadores dos parâmetros μ e β não podem ser encontrados em sua forma explícita.

Para obtermos $\hat{\mu}$ e $\hat{\beta}$ maximizaremos a função de verossimilhança numericamente.

O exemplo a seguir mostrará como obter as estimativas dos parâmetros do modelo Log-Normal usando o software R (comando *nlm*).

Exemplo 4.4. *Ajuste o modelo Log-Logístico para o seguinte conjunto de dados:*

1,3; 1,4; 1,5; 1,6; 1,7; 1,7; 1,8; 1,9$^+$; 2,2; 2,2; 2,3; 2,3$^+$; 2,4$^+$; 2,5; 2,7;
2,7; 2,8; 2,8; 2,9; 3,2; 3,6; 3,6; 3,7; 3,8; 3,9; 4,0; 4,2; 4,3; 7,4; 11,6

+: censura

Para obter as estimativas dos parâmetros do modelo Log-Logístico para este conjunto de dados usaremos os seguintes comandos no R:

```
dens<-function(t,mu,beta){                          # sub-funções que realizarão o cálculo
(beta*(t/mu)^(beta-1)) / (  mu* (1+(t/mu)^beta)^2  ) }   # da função densidade

sobrev<-function(t,mu,beta){                        # sub-funções que realizarão o cálculo
(1+(t/mu)^beta)^-1}                                 # da função de sobrevivência

like.loglogist<-function(parametro,tempo,delta){
  L1<-log(dens(tempo,parametro[1],parametro[2]))
  L2<-log(sobrev(tempo,parametro[1],parametro[2]))
  -sum(L1*delta + L2*(1-delta) ) }
tempo<-c(1.3,1.4,1.5, ..., 7.4,11.6)
delta<-c(1,1,1,....,1,1)
chute.inicial<-c(1,1)
emv<-nlm(like.loglogist,chute.inicial,hessian=TRUE,tempo=tempo,delta=delta)
emv
solve(emv$hessian)      # fornece a matriz de variância-covariância de α e λ

## Obtenção das estimativas de S(t), via Kaplan-Meier e modelo Log-Logística.
library(survival)
dados<-Surv(tempo,delta)
km<-survfit(dados~1)
plot(km,conf.int=F,main="Função de Sobrevivência Estimada",xlab="Tempo",ylab="S(t)")
m<-max(tempo)*100
t<-(1:m)/100
estimativa<-emv$estimate
s.loglogistica<-sobrev(t,estimativa[1],estimativa[2])
points(t,s.loglogistica,type="l",lty=2)
legend(4,0.95,c("___  Kaplan-Meier","------ Log-Logística"))
```

Os resultados do Exemplo 4.4 são dados por:

$$\hat{\mu} = 2{,}8308 \qquad Var(\hat{\mu}) = 0{,}0603$$
$$\hat{\beta} = 3{,}7505 \qquad Var(\hat{\beta}) = 0{,}3569$$

Para construir um intervalo (simultâneo) de 95% de confiança para os parâmetros, temos que ajustar o coeficiente de confiança (como foi feito na distribuição Weibull e log-normal):

$$(1-\alpha) = 0{,}95 \quad \Rightarrow \quad (1-\alpha^*) = (0{,}95)^{0{,}5} = 0{,}9747$$

Assim, $z_{\left(1-\frac{\alpha^*}{2}\right)} = z_{(0{,}9874)} = 2{,}238$. Logo,

$$I.C.(\mu)_{95\%(simultâneo)}: \left[2{,}8308 \pm 2{,}238\sqrt{0{,}0603}\right] = [2{,}2812 \ ; \ 3{,}3804]$$

e

$$I.C.(\beta)_{95\%(simultâneo)}: \left[3{,}7505 \pm 2{,}238\sqrt{0{,}3569}\right] = [2{,}4135 \ ; \ 5{,}0875]$$

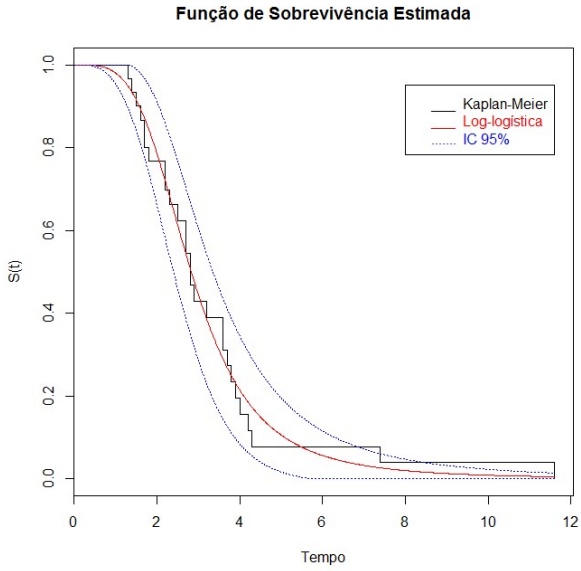
Neste exemplo, $\text{Cov}(\hat{\mu}, \hat{\beta}) = -0{,}0117$, resultando em $Corr(\hat{\mu}, \hat{\beta}) = \dfrac{-0{,}0072}{\sqrt{0{,}0603}\sqrt{0{,}3568}} = -0{,}0491$, mostrando que, neste caso, os intervalos apresentados podem ser considerados.

Assim como no modelo Weibull e log-normal, a variância da estimativa da função de sobrevivência do modelo log-logístico pode ser obtida pelo método delta e pode ser obtida pelo método delta, calculada pelo comando *deltavar* da biblioteca "emdbook" do R.

O exemplo a seguir mostrará como obter os intervalos de confiança da função de sobrevivência do modelo log-logístico usando o software R.

```
require(emdbook)
#### calculo da variância de S(t)
var.sob<-numeric(length(t))
for (i in 1:length(t)) {
  x<-t[i]
  var.sob[i]<-deltavar(fun = (1+(x/mu)^beta)^-
1,Sigma=Sigma,meanval=c(mu=estimativa[1],beta=estimativa[2]))
}
#### Construindo IC 95%
LI.sob<-s.loglogistica - qnorm(0.975)*var.sob^.5
LS.sob<-s.loglogistica + qnorm(0.975)*var.sob^.5
### Intervalo simétrico (truncando em 0 e 1)
LI.sob[which(LI.sob<0)]<-0
LS.sob[which(LS.sob>1)]<-1
points(t,LI.sob,type="l",lty=3,col=4)
points(t,LS.sob,type="l",lty=3,col=4)
```

A estimativa da Função de Sobrevivência para os dados do Exemplo 4.4 com seu respectivo intervalo de 95% de confiança é apresentada pela a seguir abaixo.



DETERMINAÇÃO EMPÍRICA DA FORMA DA FUNÇÃO DE RISCO

Uma vez que os modelos apresentados anteriormente são concorrentes entre si, para ajustar um determinado conjunto de tempos de sobrevivência, e que os mesmos apresentam diferentes formas de funções de risco, torna-se necessário a utilização de algumas metodologias para selecionar o modelo mais apropriado, mesmo antes de qualquer ajuste.

Em muitas aplicações existe informação qualitativa e, muitas vezes, estrutural a respeito do fenômeno em questão, que pode ser utilizada na determinação empírica da forma da função de risco. Informações estruturais estão diretamente vinculadas ao conhecimento do pesquisador sobre o fenômeno, enquanto que informações qualitativas podem ser extraídas por meio de uma análise gráfica. Neste contexto, um gráfico conhecido como Gráfico do Tempo Total em Teste (Curva TTT) é de grande utilidade. Este gráfico foi inicialmente proposto por Aarset em 1987 e é construído a partir das quantidades

$$G\left(\frac{r}{n}\right) = \frac{\sum_{i=1}^{r} T_{i:n} + (n-r)T_{r:n}}{\sum_{i=1}^{n} T_i} \qquad \text{versus} \qquad A = \frac{r}{n}$$

onde r = 1, 2, ..., n e $T_{i:n}$, i = 1, 2, ..., n são as estatísticas de ordem da amostra (amostra organizada em um rol crescente).

A figura a seguir apresenta algumas curvas TTT, que ilustram as várias formas que podem ser observadas para uma função de risco.

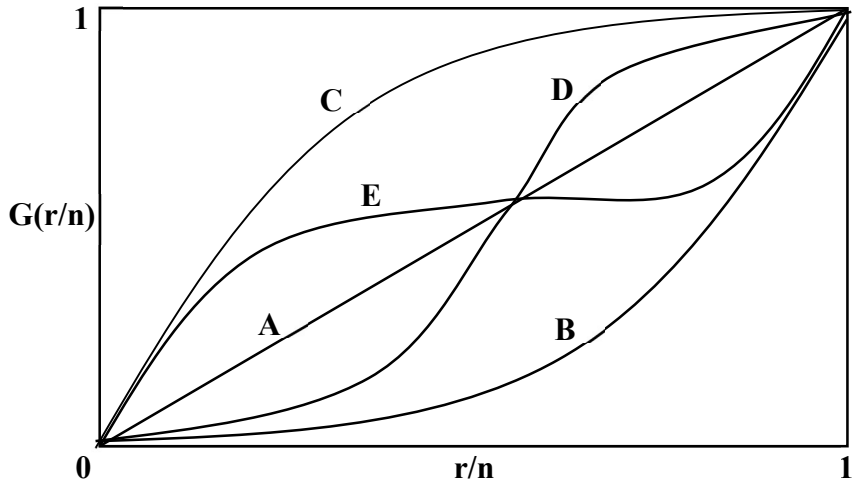


Figura: Gráfico ilustrativo de alguns TTT plots.

Se uma reta diagonal é observada **(A)**, uma função de risco constante é indicada (modelo Exponencial). Se a curva é convexa **(B)** ou côncava **(C)**, a função de risco é monotonicamente decrescente ou crescente, respectivamente (modelo Weibull). Se a curva é côncava e então convexa **(E)**, a função de risco é unimodal (Log-Normal e Log-Logística), e no caso reverso **(D)** a função de risco tem a forma de "U". Várias regiões côncavas e convexas direcionam para o ajuste de uma função de risco multimodal. Essas curvas (riscos multimodais) podem ser ajustadas através de distribuições de múltiplos riscos ou distribuição de misturas.

Caso tenhamos informações sobre covariáveis para cada indivíduo e uma quantidade significativa de indivíduos em cada nível ou combinação destas covariáveis, a curva TTT pode ser construída considerando cada nível da covariável ou combinação das mesmas, separadamente.

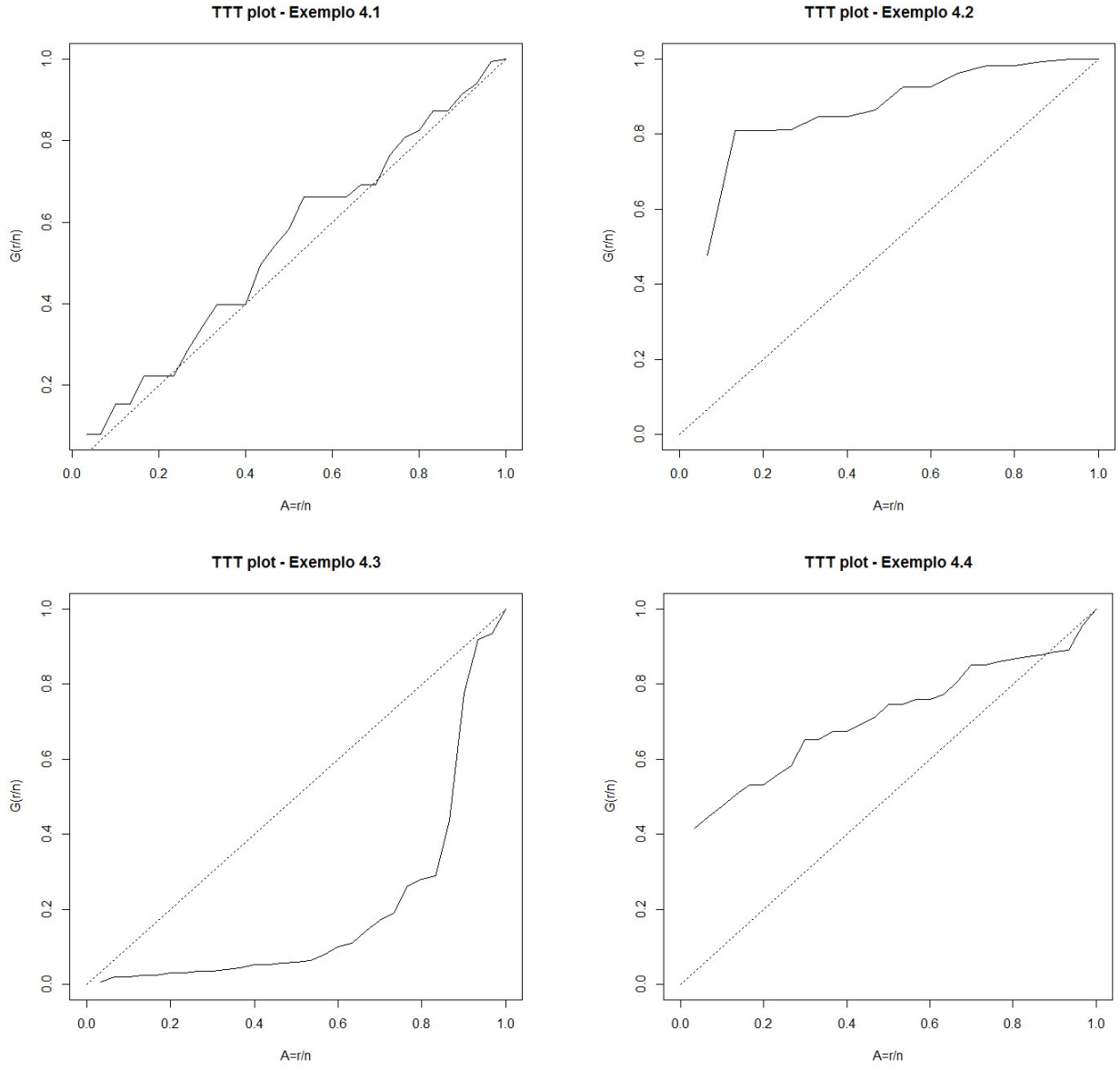
Tabela. Para os dados do Exemplo 4.1 Valores da Curva TTT para os dados do Exemplo 4.1.

$T_{i:n}$	A=r/n	G(r/n)	$T_{i:n}$	A=r/n	G(r/n)	$T_{i:n}$	A=r/n	G(r/n)
1	1/30	0,080	6	11/30	0,399	13	21/30	0,691
1	2/30	0,080	6	12/30	0,399	16	22/30	0,763
2	3/30	0,154	8^+	13/30	0,495	18	23/30	0,806
2	4/30	0,154	9	14/30	0,540	19	24/30	0,824
3	4/30	0,223	10	14/30	0,582	22	24/30	0,872
3	6/30	0,223	12^+	16/30	0,662	22	26/30	0,872
3	7/30	0,223	12	17/30	0,662	26	27/30	0,915
4	8/30	0,285	12	18/30	0,662	29	28/30	0,939
5^+	9/30	0,343	12	19/30	0,662	39^+	29/30	0,992
6	10/30	0,399	13	20/30	0,691	42^+	1	1

No R, podemos construir o TTT plot por meio dos seguintes comandos:

```
tempo<-c(1,1,2,2,3,3,3,4,5,6,6,6,8,9,10,12,12,12,12,13,13,16,18,19,22,22,26,29,39,42)
t<-sort(tempo)
n<-length(t)
A<-(1:n)/n
G<-numeric()
for (i in 1:n){G[i]<- (sum(t[1:i]) + (n-i)*t[i])/sum(t)}
plot(A,G,type="l",main="TTT plot para os dados do Exemplo 5", xlab="A=r/n", ylab="G(r/n)")
points(c(0,1),c(0,1),type="l",lty=3)
```

A seguir são apresentados os TTT-plots para os dados dos Exemplos 4.1, 4.2, 4.3 e 4.4.



Para os dados do Exemplo 4.1, podemos notar que a curva TTT apresenta um comportamento linear (Caso A), indicando que a função de risco é constante. Neste caso o modelo Exponencial é um candidato para ajuste do conjunto de dados. Já a curva TTT para os dados do Exemplo 4.2 é côncava (Caso C), indicando que a função de risco é (monótona) crescente. Aqui, o modelo Weibull com o parâmetro de forma maior que 1 é um forte candidato para o ajuste da amostra. No Exemplo 4.3, a curva TTT é convexa, que indica um possível comportamento decrescente para a função de risco. É importante lembrar que a distribuição Log-Normal (e a Log-Logística também) acomoda funções de riscos decrescentes, além das unimodais. Portanto, para os dados do Exemplo 4.3, os modelos candidatos para ajuste são: Weibull, Log-Normal, e Log-Logística, sendo que na ocasião a distribuição utilizada foi a Log-Normal. A curva TTT para os dados do Exemplo 4.4 apresenta um comportamento inicialmente côncavo e então (levemente) convexo (Caso E), indicando que a função de risco é unimodal. Neste caso, modelos candidatos são a Log-Normal e a Log-Logística, sendo que a segunda foi preferida para realizar o ajuste.

NOTA: É importante observar que as censuras **não** são consideradas no momento da construção da curva TTT, o que pode induzir a erros na interpretação e de escolha do modelo nos casos em que o número de censuras é grande.

DETERMINAÇÃO EMPÍRICA DA FUNÇÃO H(t)

Como visto anteriormente, a curva TTT ignora as censuras nos dados, o que pode resultar erros na escolha do modelo nos casos em que o número de censuras é grande. Uma alternativa neste caso é observar o comportamento da Função de Risco Acumulada, H(t). Desta forma, podemos ter uma idéia do comportamento da função de risco observando o gráfico $\hat{H}(t)$, que deve apresentar as mesmas características de H(t). A estimativa da função de risco acumulado, $\hat{H}(t)$, pode ser obtida por meio dos estimadores de Kaplam-Meier ou Nelson-Aalen. A interpretação dos gráficos de $\hat{H}(t)$ pode ser feita da seguinte forma:

Se uma reta diagonal (não necessariamente a reta Y=X) é observada **(A)**, uma função de risco constante é indicada (modelo Exponencial). Se a curva é convexa **(B)**, a função de risco é monotonicamente crescente. Se a curva é côncava **(C)**, a função de risco é monotonicamente decrescente. Se a curva é convexa e então côncava **(D)**, a função de risco é unimodal (Log-Normal e Log-Logística), e no caso reverso **(E)** a função de risco tem a forma de "U". Várias regiões côncavas e convexas direcionam para o ajuste de uma função de risco multimodal. Essas curvas (riscos multimodais) podem ser ajustadas através de distribuições de múltiplos riscos ou distribuição de misturas.

IMPORTANTE: Note que a interpretação da Função de Risco Acumulada, H(t) é diferente (inverso) da interpretação da curva TTT.

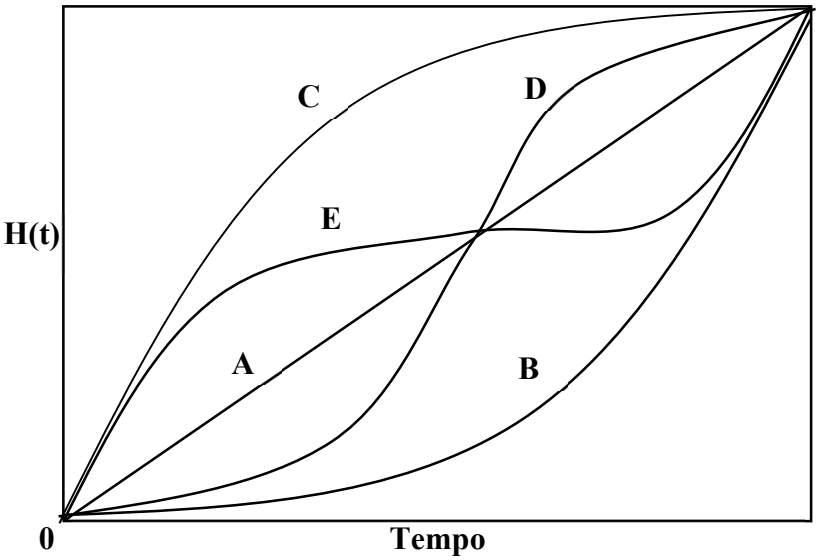
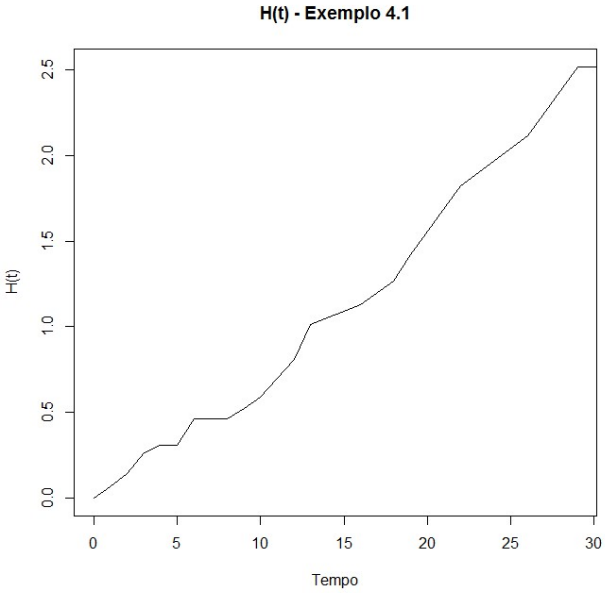
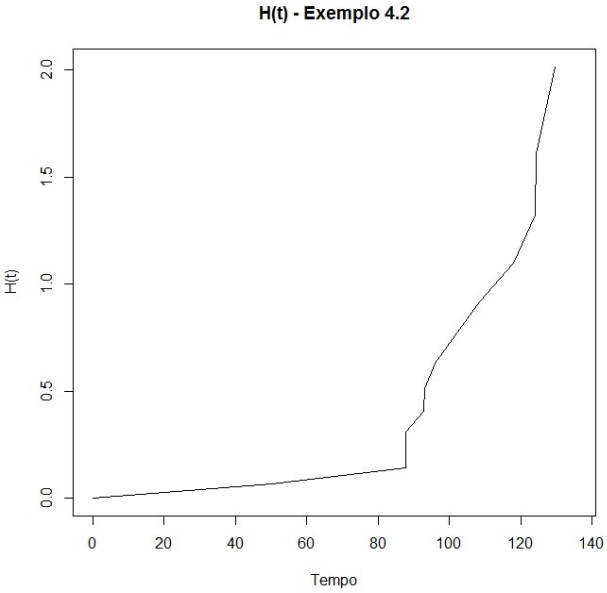


Figura: Gráfico ilustrativo de algumas funções de risco acumuladas.

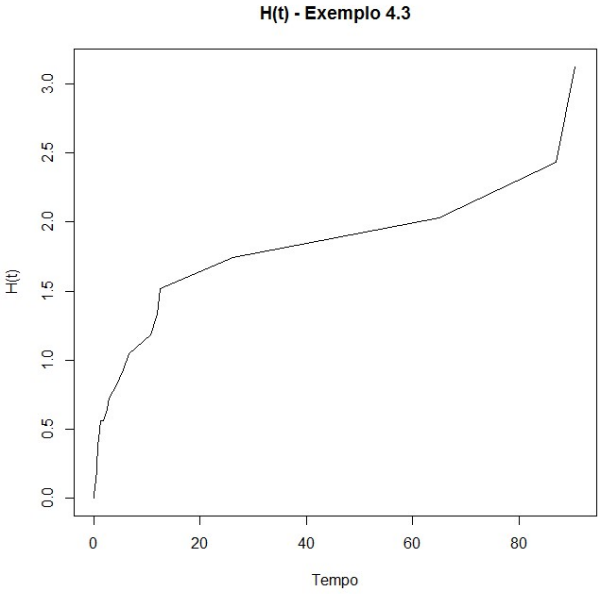
A seguir são apresentados os gráficos de $\hat{H}(t)$ para os dados dos Exemplos 4.1, 4.2, 4.3 e 4.4.



linear (Caso A) – **risco constante**
Candidato: Exponencial

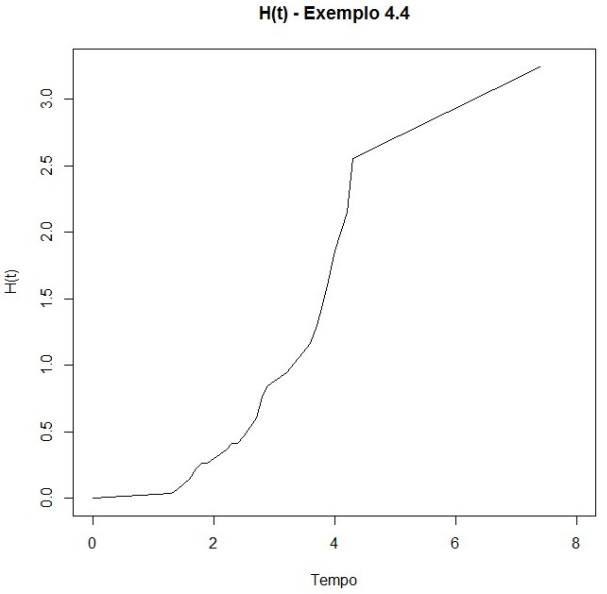


Convexo (Caso B) – **risco crescente**
Candidato: Weibull



Côncavo e então convexo (Caso D)
Risco em forma de "U"
Ou
Côncavo (Caso C) – **risco decrescente**
Note que excluindo o último ponto desse gráfico notamos nitidamente um comportamento côncavo.

Candidatos: Weibull, Log-Normal, Log-Logística e modelos de riscos múltiplos.



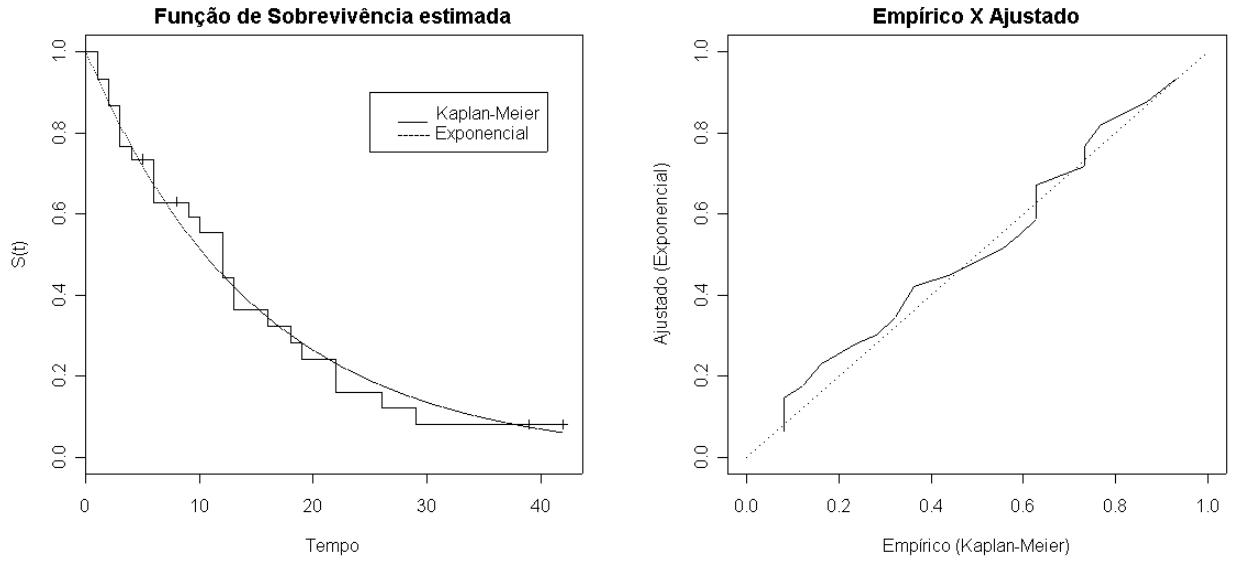
Convexo e então côncavo (Caso E)
Risco unimodal
Ou
Convexo (Caso B) – **risco crescente**
Note que excluindo o último ponto desse gráfico notamos nitidamente um comportamento convexo.

Candidatos: Weibull, Log-Normal e Log-Logística.

VERIFICAÇÃO DO AJUSTE DO MODELO

Depois de selecionado o modelo e realizado o ajuste, é interessante verificarmos a validade desse ajuste. Essa verificação pode ser feita através de procedimentos gráficos ou por meio de testes de hipóteses.

Verificação Gráfica. A verificação do ajuste pode ser feita através dos gráficos da função de sobrevivência estimada. A verificação é feita através da comparação das estimativas empíricas (Kaplan-Meier) e ajustadas pelo modelo. Dois tipos de gráficos podem ser considerados: um contendo as duas funções de sobrevivência estimadas (empírica e ajustada) e outro gráfico da estimativa empírica versus estimação ajustada. As figuras abaixo apresentam a verificação gráfica do ajuste do modelo Exponencial para os dados do Exemplo 4.1.



Podemos observar, através dos gráficos apresentados, que o modelo Exponencial apresentou um bom ajuste para os dados do Exemplo 4.1. Note que no gráfico "Empírico X Ajustado", quanto melhor é o ajuste, mais próximo a curva se aproxima a uma reta diagonal.

Testes de Hipóteses. Uma maneira formal de verificarmos a validade do ajuste de um determinado modelo é dado através de um teste de hipóteses. Aqui testaremos a hipótese de um modelo se ajustar a um determinado conjunto de dados. Para tanto, utilizaremos o teste de Kolmogorov-Smirnov.

Assim, estaremos testando as seguintes hipóteses:

H_0: $S_{ajustado}(t) = S_{km}(t)$
H_a: $S_{ajustado}(t) \neq S_{km}(t)$

A estatística do teste de Kolmogorov-Smirnov é dada por:

$$KS = \max_{t:\ t \text{ não censurado}} \left| \hat{S}_{km}(t) - \hat{S}_{ajustado}(t) \right|.$$

Assim, **rejeitamos a hipótese H_0** (invalidando o ajuste do modelo), sob um nível de significância α, **se** a estatística **KS for maior que o quantil α to teste de Kolmogorov**. A tabela a seguir apresenta os quantis do teste de Kolmogorov.

Tabela: Quantis do teste de Kolmogorov-Smirnov.

α	**0,80**	**0,90**	**0,95**	**0,98**	**0,99**		**0,80**	**0,90**	**0,95**	**0,98**	**0,99**
n' = 1	,900	,950	,975	,990	,995	n' = 21	,226	,259	,287	,321	,344
2	,684	,776	,842	,900	,929	22	,221	,253	,281	,314	,337
3	,565	,636	,708	,785	,829	23	,216	,247	,275	,307	,330
4	,493	,565	,624	,689	,734	24	,212	,242	,269	,301	,323
5	,447	,509	,563	,627	,669	25	,208	,238	,264	,295	,317
6	,410	,468	,519	,577	,617	26	,204	,233	,259	,290	,311
7	,381	,436	,483	,538	,576	27	,200	,229	,254	,284	,305
8	,358	,410	,454	,507	,542	28	,197	,225	,250	,279	,300
9	,339	,387	,430	,480	,513	29	,193	,221	,246	,275	,295
10	,323	,369	,409	,457	,489	30	,190	,218	,242	,270	,290
11	,308	,352	,391	,437	,468	31	,187	,214	,238	,266	,285
12	,296	,338	,375	,419	,449	32	,184	,211	,234	,262	,281
13	,285	,325	,361	,404	,432	33	,182	,208	,231	,258	,277
14	,275	,314	,349	,390	,418	34	,179	,205	,227	,254	,273
15	,266	,304	,338	,377	,404	35	,177	,202	,224	,251	,269
16	,258	,295	,327	,366	,392	36	,174	,199	,221	,247	,265
17	,250	,286	,318	,355	,381	37	,172	,196	,218	,244	,262
18	,244	,279	,309	,346	,371	38	,170	,194	,215	,241	,258
19	,237	,271	,301	,337	,361	39	,168	,191	,213	,238	,255
20	,232	,265	,294	,329	,352	40	,165	,189	,210	,235	,252
						Aproximação para n' > 40	$\frac{1,07}{\sqrt{n'}}$	$\frac{1,22}{\sqrt{n'}}$	$\frac{1,36}{\sqrt{n'}}$	$\frac{1,52}{\sqrt{n'}}$	$\frac{1,63}{\sqrt{n'}}$

Nota: Uma melhor aproximação pode ser alcançada para n' > 40 se $\left(n'+\sqrt{n'/10}\right)^{0,5}$ é usado no lugar de $\sqrt{n'}$ no denominador

Importante: No teste de Kolmogorov, **n'** é o número de observações não censuradas na amostra, isto é $n' = \sum_{i=1}^{n} \delta_i$.

Para os dados do Exemplo 4.1, temos que a estatística do teste de Kolmogorov é dada por:

$$KS = \underset{t:\ t \text{ não censurado}}{Max} \left| \hat{S}_{km}(t) - \hat{S}_{ajustado}(t) \right| = 0,0703$$

Temos também que $n' = \sum_{i=1}^{n} \delta_i = 25$

Para um nível de significância de α = 5% e n'=25, temos que o quantil do teste é 0,264.

Logo, como **KS = 0,0703 < 0,264**, não rejeitamos a hipótese de o modelo Exponencial apresentar um bom ajuste para os dados do Exemplo 4.1. Ou seja, o modelo Exponencial é adequado para o ajuste dos dados do Exemplo 4.1.

EXERCÍCIOS

1. Sejam $T_1,T_2,...,T_n$ uma amostra aleatória de uma distribuição exponencial com média λ, isto é

$$f(t) = \frac{1}{\lambda} e^{-\frac{t}{\lambda}}, \quad \lambda>0, t \geq 0.$$

i) Obtenha o valor de $Var(\hat{S}(t))$ pelo método delta.
ii) Com base na variância obtida no item i) e os dados do exemplo 4.1, calcule o intervalo de confiança da função de sobrevivência em t=1.
iii) Obtenha o intervalo de confiança do tipo "log" para a média de T, isto é, o intervalo construído a partir da transformação $\phi=\log(\lambda)$;
iv) com base no intervalo obtido em iii), obtenha o intervalo da função de sobrevivência para um valor t, $t \geq 0$.

2. Encontre os estimadores de máxima verossimilhança e a Informação de Fisher observada de cada um dos modelos abaixo (considere a presença de dados censurados):
i) T~Geometrica(θ); $p(t) = \theta(1-\theta)^t$, $0<\theta<1$, t=0,1,2,3...
ii) T~Weibull(2, λ);
iii) T ~ Pareto(α,1);

3. Considere uma variável aleatória T que possui a seguinte Função Densidade de Probabilidades:

$$f(t) = \frac{\alpha\beta^{\alpha}}{(t+\beta)^{\alpha+1}}, \quad t \geq 0$$

Obs: Note que T ~ Pareto(α, β).

i) mostre que a Função de Sobrevivência e Risco da Pareto são dadas por

$$S(t) = \left(\frac{\beta}{t+\beta}\right)^{\alpha} \qquad \text{e} \qquad h(t) = \frac{\alpha}{(t+\beta)}.$$

ii) monte a função de verossimilhança da Pareto e elabore um algoritmo no R que forneça as estimativas de máxima verossimilhança dos parâmetros;

4. Para cada um dos conjuntos de dados a seguir:

i) determine as estimativas de S(t) pelo método de Kaplan-Meier;
ii) construa a curva TTT e o gráfico empírico de H(t) e aponte quais são os modelos candidatos para o ajuste;
iii) ajuste os modelos: Exponencial, Weibull, Log-Normal, Log-Logística e Pareto;
iv) verifique a validade de cada um dos modelos através dos gráficos de verificação de ajuste e também pelo teste de Kolmogorov-Smirnov;
v) apresente os algoritmos utilizados no R para a resolução dos itens anteriores;
vi) escolha qual é o melhor modelo para ajustar o conjunto de dados. Justifique a resposta.

A. Considere os dados de um estudo com pacientes de leucemia quanto à introdução de um novo tratamento com a droga *6*-mercaptopurina.

Tabela. Tempos de Remissão (semanas) de pacientes com leucemia (Grupo controle).

Tempo	1	1	2	2	3	4	4	5	5	8	8	8	8	11	11	12	12	15	17	22	23

Obs: dados sem censuras

B (Exemplo 2, pág. 14). Tempos de sobrevivência (em meses) de mulheres submetidas a mastectomia simples ou radical (contraste positivo)

Tempo	5, 8, 10, 13, 18, 24, 26, 26, 31, 35, 40, 41, 48, 50, 59, 61, 68, 71, 76^{+}, 105^{+}, 107^{+}, 109^{+}, 113, 116^{+}, 118, 143, 154^{+}, 162^{+}, 188^{+}, 212^{+}, 217^{+}, 225^{+}

+: observações censuradas

C. A tabela abaixo apresenta os tempos de sobrevivência (em horas) de certo equipamento eletrônico.

0,01	0,06	$0,06^{+}$	0,07	0,12	0,18	0,19	0,41	0,57	0,83	$0,83^{+}$	1,31	1,42	1,64	1,79
2,77	2,78	3,03	3,04	3,84	5,14	6,21	6,30	$10,76^{+}$	11,52	14,06	27,27	29,94	38,09	$186,73^{+}$

+: observações censuradas