ANÁLISE DE SOBREVIVÊNCIA
Técnicas não-paramétricas

Rivert Oliveira – DEEST – ICEB - UFOP

Roteiro

- Introdução
- Estimação na ausência de censura
- O estimador de Kaplan-Meier
- O estimador de Nelson-Aalen
- O estimador de Tabela de Vida ou Atuarial
- Estimação de quantidades básicas
- Comparação de curvas de sobrevivência

INTRODUÇÃO

Introdução

- No caso de análise de sobrevivência a função de sobrevivência é o principal componente para descrever os dados

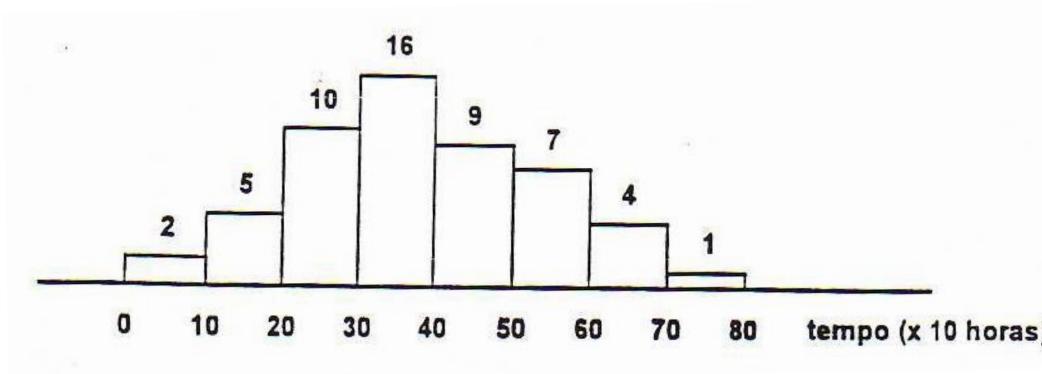
- A partir da estimação da função de sobrevivência é possível obter quantidades de interesse como o tempo médio ou mediano de vida, percentis, frações de falha etc.

- A presença de censura dificulta ou impede o uso de técnicas usuais para estatística descritiva

ESTIMAÇÃO NA AUSÊNCIA DE CENSURA

Estimação na ausência de censura

- Figura mostrando a distribuição dos tempos de falha de um certo conjunto de itens amostrais (todos falharam).



- **Na ausência de censura** a função taxa de falha no intervalo $[400, 500)$ é estimada por

$$\hat{\lambda}([400,500)) = \frac{P(400 \le T < 500|T \ge 400)}{500 - 400} = \boxed{P(400 \le T < 500|T \ge 400)}\,\frac{1}{100} \Rightarrow$$

$$\hat{\lambda}([400,500)) = \boxed{\frac{\#\, falhas\ no\ intervalo[400,500)}{\#\, de\ itens\ sob\ risco\ em\ t = 400}}\,\frac{1}{100} = \boxed{\frac{9}{21} = 0{,}429/100hs}$$

- A taxa de falha é de $42{,}9\%$ durante o período de 100 horas compreendido entre 400 e 500 horas. Ou seja, se 100 itens amostrais sobreviverem além de 400 horas, espera-se que 57 sobrevivam mais 100 horas.

Obs: note que a taxa estimada não é instantânea, portanto grosseira

A Função de Sobrevivência e Quantidades Relacionadas

- **Na ausência de censura** a probabilidade de sobrevivência no tempo $t = 400$ horas é estimada por

$$\hat{S}(400) = \frac{\#\ itens\ sob\ risco\ em\ t = 400}{\#\ de\ total\ de\ itens\ no\ estudo} = \frac{21}{54} = 0{,}389$$

- $38{,}9\%$ dos itens sobrevive além de 400 horas.

- A probabilidade de um item sobreviver além de 400 horas é de $0{,}389$

- Observe que a taxa de falha poderia ter sido estimada por

$$\hat{\lambda}([400, 500)) = \frac{S(400) - S(500)}{(500 - 400)S(400)} = \frac{0{,}389 - 0{,}222}{100 \cdot 0{,}389} = 0{,}0043/h$$

A Função de Sobrevivência e Quantidades Relacionadas

- Observa-se que a taxa de falha parece ser crescente.

Intervalo {li-ls}	Taxa de falha(%/hora) [li,ls)	Sobrevivência (%) P(T≥l_i)
$[0 - 100)$	$0{,}037$	$100{,}0$
$[100 - 200)$	$0{,}096$	$96{,}3$
$[200 - 300)$	$0{,}213$	$87{,}0$
$[300 - 400)$	$0{,}432$	$68{,}5$
$[400 - 500)$	$0{,}429$	$38{,}9$
$[500 - 600)$	$0{,}583$	$22{,}2$
$[600 - 700)$	$0{,}800$	$9{,}3$
$[700 - 800)$	1	$1{,}9$
$[800 - \infty)$	0	0

O ESTIMADOR DE KAPLAN-MEIER

Estimador de Kaplan-Meier

- A função de sobrevivência é uma função escada. Ela pode ser vista conforme o seguinte:
  - Considere $t_j's, j = 1, \ldots, k$ falhas distintas como limites de intervalos em $\mathbb{R}$. Convencione $t_0 = 0$.
  - A probabilidade de um item sobreviver além do limite em t_2 é igual à probabilidade do item sobreviver a t_2 e t_1.
- Matematicamente, pela regra do produto temos:

$$S(t_2) = P(T \geq t_2, T \geq t_1) = P(T \geq t_2 | T \geq t_1) P(T \geq t_1)$$
$$S(t_2) = [1 - P(t_1 \leq T < t_2 | T \geq t_1)][1 - P(t_0 \leq T < t_1 | T \geq t_0)]$$
$$S(t_2) = (1 - q_2)(1 - q_1)$$

- Por indução $S(t_j) = (1 - q_1)(1 - q_2) \ldots (1 - q_j), j = 1, \ldots, k$
- Onde $q_j = P(t_{j-1} \leq T < t_j | T \geq t_{j-1})$,

Estimador de Kaplan-Meier

- O estimador de Kaplan-Meier de R(t) é definido como:

$$\hat{S}(t) = \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{2}\right) \dots \left(1 - \frac{d_j}{n_j}\right) = \prod_{j:t_j<t}\left(1 - \frac{d_j}{n_j}\right), \text{ j} = 1, \dots, k$$

k é o número total de falhas distintas

$$\hat{q}_j = \frac{\#\ falhas\ em\ t_j}{\#\ de\ itens\ sob\ risco\ em\ t_j -}$$

- A tabela de sobrevivência deve ser organizada com as seguintes informações:
  - t_j: tempos de falha distintos ordenados (do menor para o maior). Os limites dos invervalos são os tempos de falha!
  - d_j: número de falhas no tempo t_j
  - n_j: número de itens sob risco (não falhou e não foi censurado) até o tempo $t_j -$
  - $\hat{S}(t)$ estimativa da função de confiabilidade além do tempo t_j

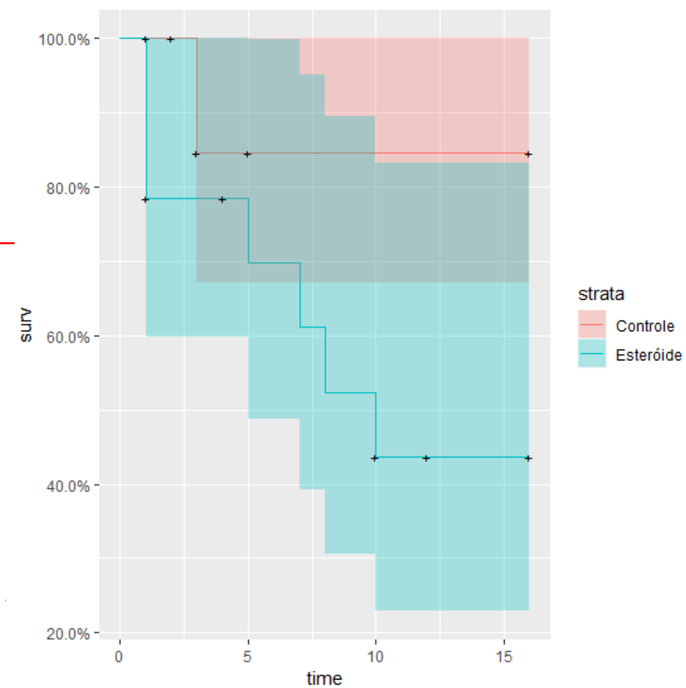
Estimador de Kaplan-Meier

- Tempos, em semanas, observados no estudo de hepatite i $(i = 1, \dots, n = 29)$. Os pacientes foram acompanhados por 16 semanas ou até a morte (evento de interesse). O estudo foi um ensaio clínico aleatorizado (tratamentos: controle placebo, esteróide)

```
Call: survfit(formula = Surv(tempos, censura) ~ grupo)

                grupo=Controle
       time          n.risk        n.event       survival        std.err lower 95% CI
      3.000          13.000          2.000          0.846          0.100        0.671
upper 95% CI
      1.000

                grupo=Esteróide
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     14       3    0.786   0.110        0.598        1.000
    5      9       1    0.698   0.128        0.488        0.999
    7      8       1    0.611   0.138        0.392        0.952
    8      7       1    0.524   0.143        0.306        0.896
   10      6       1    0.437   0.144        0.229        0.832
```

$$\hat{S}_{est}(5+) = \lim_{\Delta_t \to 0^+} \hat{S}_{est}(5+\Delta_t) = (1-\hat{q}_1)(1-\hat{q}_2)$$

$$\hat{S}_{est}(5+) = \left(1-\frac{3}{14}\right)\left(1-\frac{1}{9}\right) = 0{,}698$$



Estimador de Kaplan-Meier

- Propriedades do estimador de Kaplan-Meier:
  - Não viciado para amostras grandes
  - É fracamente consistente
  - Converge assintoticamente para um processo gaussiano
  - É estimador de máxima verossimilhança de $S(t)$
- Para construir intervalos de confiança e testar hipóteses o estimador da variância de $S(t)$ é dado por

$$\widehat{Var}\left[\hat{S}(t)\right] = \left[\hat{S}(t)_{obs}\right]^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)}$$

- No exemplo
  - $\widehat{Var}\left[\hat{S}_{\text{est}}(5)\right] = [0{,}698]^2 \left[\frac{3}{14(14-3)} + \frac{1}{9(9-1)}\right] = 0{,}0163$

Estimador de Kaplan-Meier

- Como $\hat{S}(t)$, para t fixo, tem distribuição assintótica Normal, um IC de $100(1-\alpha)\%$ de confiança para $S(t)$ é dado por

$$IC[S(t), 100(1-\alpha)\%] = \hat{S}(t) \mp |z_{\alpha/2}| \sqrt{\widehat{Var}[\hat{S}(t)]}$$

$z_{\alpha/2}$ é obtido da distribuição normal padrão

- No exemplo, para $IC[S(t), 95\%]$ temos

$$IC[S(t), 95\%] = 0{,}698 \mp 1{,}96\sqrt{0{,}0163} = (0{,}45; 0{,}95)$$

Estimador de Kaplan-Meier

- Para valores extremos de t o intervalo anterior pode apresentar valores fora do conjunto $[0,1]$.
- Kalbfleish e Prentice (1980) sugerem a transformação

$$\hat{\mathrm{U}}(t) = \log\{-\log[\hat{\mathrm{S}}(t)]\}$$

cuja variância assintótica é dada por

$$\widehat{Var}[\hat{\mathrm{U}}(t)] = \frac{\sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)}}{\left[\sum_{j:t_j<t} \log\left(\frac{n_j - d_j}{n_j}\right)\right]^2} = \frac{\sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)}}{\left[\log[\hat{\mathrm{S}}(t)]\right]^2}$$

Estimador de Kaplan-Meier

- Neste caso o $IC[S(t), 100(1-\alpha)\%]$ fica:

$$IC[S(t), 100(1-\alpha)\%] = \hat{S}(t)^{exp\left\{\mp z_{\alpha/2}\sqrt{\widehat{Var}[\hat{U}(t)]}\right\}}$$

- Outras transformações já foram propostas, como por exemplo:

$$\hat{U}(t) = \log\{\hat{S}(t)\}$$

O ESTIMADOR DE NELSON-AALEN

Estimador de Nelson-Aalen

- O estimador de Nelson-Aalen se baseia na função de risco (taxa de falha instantânea) acumulado, isto é:

$$S(t) = \exp\{-\Lambda(t)\}$$

- O estimador da função de risco e sua variância são dados por:
  - $\widetilde{\Lambda}(t) = \sum_{j:t_j<t}\left(\frac{d_j}{n_j}\right)$ e $\widehat{Var}\left[\widetilde{\Lambda}(t)\right] = \sum_{j:t_j<t}\left(\frac{d_j}{n_j^2}\right)$

- Os estimadores da função de sobrevivência e da variância da mesma ficam:
  - $\tilde{S}\,(t) = \exp\{-\widetilde{\Lambda}(t)\}$ e $\widehat{Var}\left[\widetilde{\Lambda}(t)\right] = \left[\tilde{S}(t)_{obs}\right]^2 \sum_{j:t_j<t}\left(\frac{d_j}{n_j^2}\right)$

O ESTIMADOR DE TABELA DE VIDA OU ATUARIAL

Estimador de Tabela de Vida ou Atuarial

- Suponha que o eixo do tempo seja dividido em s intervalos definidos pelos pontos de corte $t_1, \ldots, t_s$. Isto é, $I_j = [t_{j-1}, t_j)$ para $j = 1, \ldots, s$, em que $t_0 = 0$ e $t_s = +\infty$. O estimador da tabela de vida é o mesmo de KM, exceto por q_j, o qual é dado por

$$\hat{q}_j = \frac{\#\, de\ falhas\ no\ intervalo\ [t_{j-1}, t_j)}{[\#sob\ risco\ em\ t_{j-1}] \boxed{- 1/2\, [\#censuras\ em\ [t_{j-1}, t_j)]}}$$

- O estimador fica:

$$\hat{S}(t) = \prod_{l=1}^{j} (1 - \hat{q}_{l-1}), \mathrm{t} \in I_j,$$

com $j = 1, \ldots, s$ e $\hat{q}_0 = 0$.

Estimador de Tabela de Vida ou Atuarial

- A variância assintótica estimada para $\hat{S}(t)$ é dada por:

$$\widehat{Var}\left[\hat{S}(t)\right] \cong \left[\hat{S}(t)_{obs}\right]^2 \sum_{l=1}^{j} \frac{\hat{q}_l}{n_l(1-\hat{q}_l)}, t \in \mathrm{I}_j$$

com $j = 1, \ldots, s$

Estimador de Tabela de Vida ou Atuarial

- Estimativas da tabela de vida para o grupo esteróide

Intervalo I_j	# sob risco	# de falhas	# de censuras	$\hat{q}_j$	$(1-\hat{q}_j)$	$\hat{S}(t)$
[0,5)	14	3	2	0,231	0,769	$1-\hat{q}_{j-1} = 1-\hat{q}_0 = 1-0 = 1$
[5,10)	9	3	0	0,333	0,667	0,769
[10,15)	6	1	2	0,200	0,8	0,513
[15,16)	3	0	3	0	1	0,410

- $$\hat{S}(10) = \underbrace{\underbrace{(1-0)}_{\hat{S}(t\in[0,5))}\left(1-\frac{3}{14-\frac{1}{2}2}\right)}_{\hat{S}(t\in[5,10))}\left(1-\frac{3}{9}\right) = 0,513$$

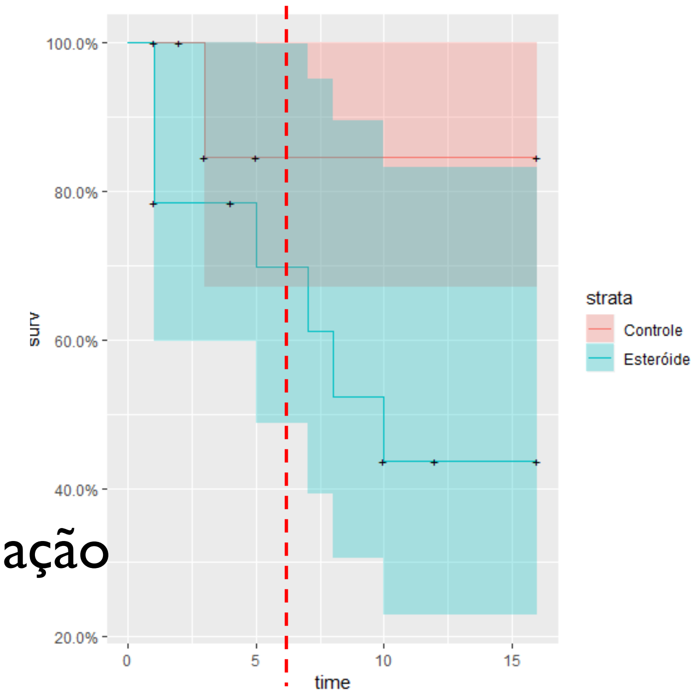
ESTIMAÇÃO DE QUANTIDADES BÁSICAS

ESTIMAÇÃO DE QUANTIDADES BÁSICAS SOBREVIVÊNCIA

- A curva de Kalan- Meier fornece quantidades básicas, por exemplo, a probabilidade de um paciente tratado com esteroides sobreviver a 6 semanas de tratamento.
  - $\hat{S}(6|Esteróide) = 0{,}698$

grupo=Esteróide

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	14	3	0.786	0.110	0.598	1.000
5	9	1	0.698	0.128	0.488	0.999
7	8	1	0.611	0.138	0.392	0.952
8	7	1	0.524	0.143	0.306	0.896
10	6	1	0.437	0.144	0.229	0.832



- Contudo, deve ser preferida uma interpolação
  - $\dfrac{7-5}{0{,}611-0{,}698} = \dfrac{6-5}{\hat{S}(6|Esteróide)-0{,}698} \Rightarrow$
  - $\hat{S}(6|Esteróide) = 0{,}655$

$$\widehat{Var}\left[\hat{S}(t)\right] = \left[\hat{S}(t)_{obs}\right]^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)}$$

ESTIMAÇÃO DE QUANTIDADES BÁSICAS PERCENTIL

- Analogamente, podemos obter percentis. Por exemplo, para o percentil $50\% \ (t_p = t_{50})$, a mediana, temos

grupo=Esteróide

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	14	3	0.786	0.110	0.598	1.000
5	9	1	0.698	0.128	0.488	0.999
7	8	1	0.611	0.138	0.392	0.952
8	7	1	0.524	0.143	0.306	0.896
10	6	1	0.437	0.144	0.229	0.832

  - $\frac{10-8}{0{,}437-0{,}524} = \frac{\hat{t}_{50}-8}{0{,}5-0{,}524} \Rightarrow \hat{t}_{50} = 8{,}55$
- A variância assintótica do estimador de percentis $(\hat{t}_{50})$ é expressa por
  - $\text{Var}(\hat{t}_p) = \frac{Var[\hat{S}(\hat{t}_p)]}{[f(\hat{t}_p)]^2}$
  - Contudo a difícil obtenção de $f(\hat{t}_p)$ é um fator limitador para obtenção da estimativa da variância, usualmente obtida da inversão da região de rejeição de um teste de hipótese para a mediana (Brookmeyer e Crowley, 1982)

ESTIMAÇÃO DE QUANTIDADES BÁSICAS TEMPO MÉDIO DE VIDA

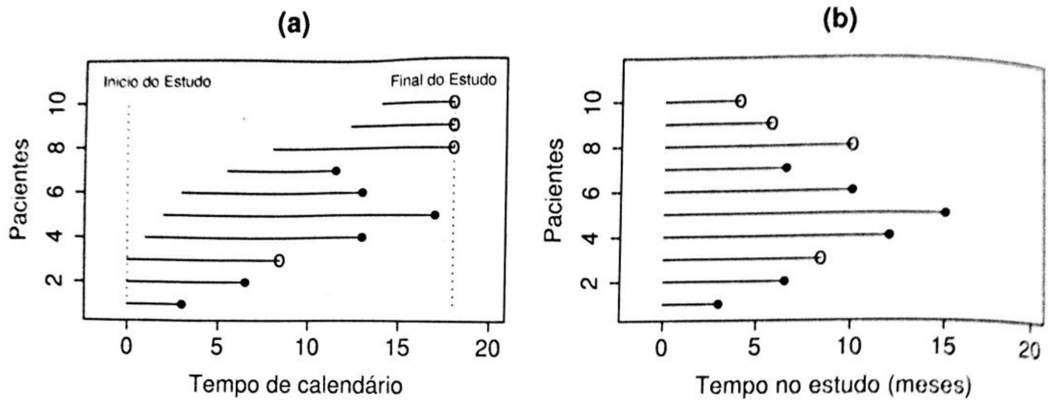
- O estimador tempo médio de vida $(\hat{t}_m)$ é dado pela área sob a curva de sobrevivência
  - $\hat{t}_m = t_1 + \sum_{j=1}^{k-1} \hat{S}(t_j)(t_{j+1} - t_j)$
  - em que $t_1 < \cdots < t_k$ são os tempos distintos e ordenados de falha
  - Obs: o tempo médio de vida é subestimado caso o maior tempo de falha observado seja uma censura
- Sua variância assintótica é dada por:
  - $\widehat{Var}(\hat{t}_m) = \frac{r}{r-1}\left[\sum_{j=1}^{r-1} \frac{A_j^2}{n_j(n_j-d_j)}\right]$
  - $A_j = \hat{S}(t_j)(t_{j+1} - t_j) + \cdots + \hat{S}(t_{r-1})(t_r - t_{r-1})$
  - r é o número de falhas (não falhas distintas!)

ESTIMAÇÃO DE QUANTIDADES BÁSICAS TEMPO MÉDIO DE VIDA RESIDUAL

- O estimador tempo médio de vida residual $(\widehat{vmr})$ é dado por
  - $\widehat{vmr}(t) = \dfrac{\acute{a}rea\ sob\ a\ curva\hat{S}(t)\ \grave{a}\ direita\ de\ t}{\hat{S}(t)}$
  - Obs: o valor médio residual apresenta limitações similares às do estimador do tempo médio de vida

ESTIMAÇÃO DE QUANTIDADES BÁSICAS

- Tempo até a reincidência de tumor sólido em $i = 1, \dots, n = 10$ pacientes (Lee, 1980). O tempo total do estudo foi de 14 meses (não o tempo de acompanhamento).



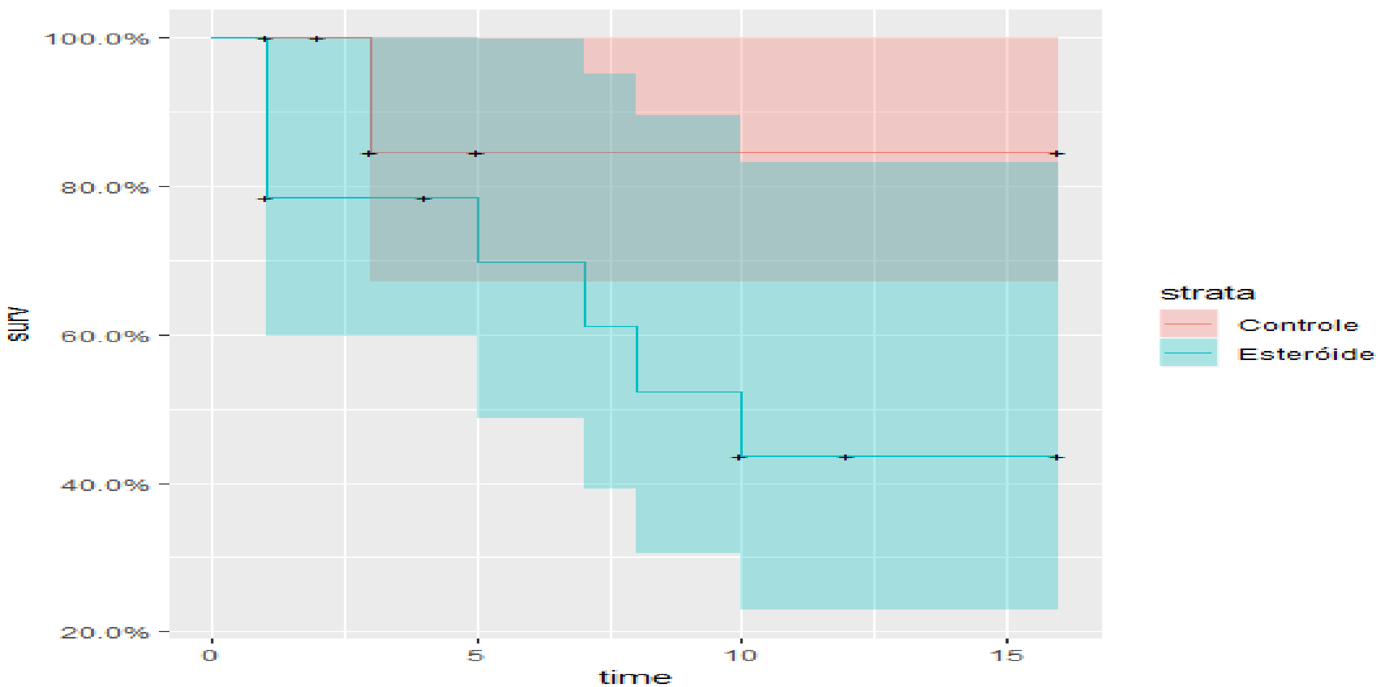

```
Call: survfit(formula = Surv(tempos, censura) ~ 1, conf.type = "plain")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  3.0     10       1    0.900  0.0949        0.714        1.000
  6.5      7       2    0.643  0.1679        0.314        0.972
 10.0      4       1    0.482  0.1877        0.114        0.850
 12.0      2       1    0.241  0.1946        0.000        0.622
 15.0      1       1    0.000     NaN          NaN          NaN
```

- $\hat{t}_{50} = 9{,}6$, $\hat{t}_m = 10{,}1$, $\widehat{Var}(\hat{t}_m) = 2{,}33$ e $\widehat{vmr}(10) = 3{,}5$

COMPARAÇÃO DE CURVAS DE SOBREVIVÊNCIA

COMPARAÇÃO DE CURVAS DE SOBREVIVÊNCIA

- Dados de hepatite: a sobrevivência é diferente entre os grupos "Controle" e "Esteróide"? Para responder a esta pergunta os testes comuns são:
  - Logrank (Mantel, 1966):
  - Wilcoxon (generalizado)



COMPARAÇÃO DE CURVAS DE SOBREVIVÊNCIA

- Teste LOGRANK
  - Sejam duas funções de sobrevivência $S_1(t)$ e $S_2(t)$.
  - Sejam $t_1 < t_2 < \cdots < t_k$ os distintos tempo de falha da amostra combinada.
  - Suponha que em t_j aconteçam d_j falhas e que n_j indivíduos estejam sob risco em um tempo imediatamente inferior a t_j na amostra combinada e, respectivamente, d_{ij} e n_{jj} na amostra i: $i = 1,2$ e $j = 1, \dots, k$.
  - Em cada tempo de falha t_j, os dados podem ser dispostos na tabela de contingência

	Grupos		
	1	2	
Falha	d_{1j}	d_{2j}	d_j
Não Falha	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
	n_{1j}	n_{2j}	n_j

COMPARAÇÃO DE CURVAS DE SOBREVIVÊNCIA

- Teste LOGRANK
  - Condicional na experiência de falha e censura até o tempo t_j (fixando marginais de coluna) e ao número de falhas no tempo t_j (fixando marginais de linha), a distribuição de d_{2j} é hipergeométrica com média
$$\omega_{2j} = n_{2j} d_j n_j^{-1}$$
  - e variância
$$(V_j)_2 = n_{2j}(n_j - n_{2j}) d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1}$$
  - Se as k tabelas de contingência forem independentes um teste aproximado para a hipótese $H_0: S_1(t) = S_2(t)$ usa a seguinte estatística de teste:

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - \omega_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2} \overset{H_0}{\approx} \chi_1^2$$

COMPARAÇÃO DE CURVAS DE SOBREVIVÊNCIA

- Dados de hepatite

$$T = \frac{\left[\sum_{j=1}^{k}\left(d_{2j} - \omega_{2j}\right)\right]^2}{\sum_{j=1}^{k}\left(V_j\right)_2} = 3{,}67$$

```
-----.
survdiff(formula = Surv(tempos, censura) ~ grupo, rho = 0)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
grupo=Controle  15        2     4.81      1.64      3.67
grupo=Esteróide 14        7     4.19      1.89      3.67

 Chisq= 3.7  on 1 degrees of freedom, p= 0.06
```

A diferença na sobrevivência entre os grupos é significativa a um nível de significância $\alpha = 0,1$.

COMPARAÇÃO DE CURVAS DE SOBREVIVÊNCIA

- Exemplo: Dados de Malária
- Estudo: clínico aleatorizado
- Tempo de estudo: 30 dias
- Resposta: tempo decorrido desde a infecção até a morte do camundongo
- Grupos:
  - Grupo I: imunizado para malária 30 dias antes da infeção por malária e esquistossomose
  - Grupo II: infectado por malária
  - Grupo III: infectado por malária e esquistossomose

Grupos	Tempos de sobrevivência
I	7, 8, 8, 8, 8, 12, 12, 17, 18, 22, 30+, 30+, 30+, 30+, 30+, 30 +
II	8, 8, 9, 10, 10, 14, 15, 15, 18, 19, 21, 22, 22, 23, 25
III	8, 8, 8, 8, 8, 8, 9, 10, 10, 10, 11, 17, 19

COMPARAÇÃO DE CURVAS DE SOBREVIVÊNCIA

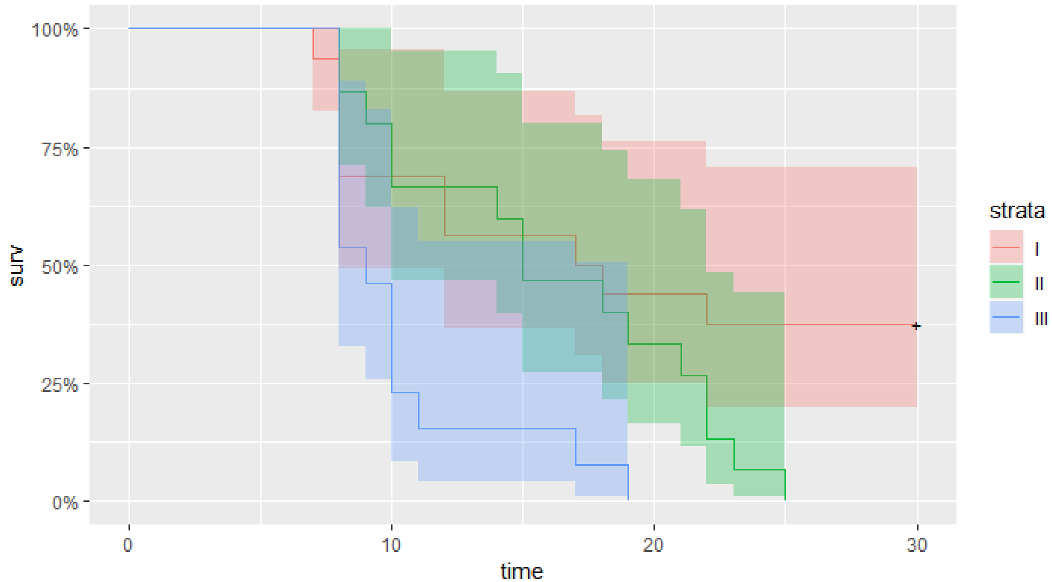
- LOGRANK Generalizado
- Teste: $H_0: S_1(t) = \cdots = S_g(t)$

$$T = \nu' V^{-1} \nu \overset{H_0}{\sim} \chi^2_{g-1}$$

- em que $\nu = \sum_j^k \nu_j$ e $V = \sum_j^k V_j$ com
  - $\nu_j' = \left(d_{2j} - \omega_{2j}, \dots, d_{gj} - \omega_{gj}\right)$
  - $\left(V_j\right)_{ii} = n_{ij}\left(n_j - n_{ij}\right)d_j\left(n_j - d_j\right)n_j^{-2}\left(n_j - 1\right)^{-1}$ e $\left(V_j\right)_{il} = -n_{ij}n_{lj}d_j\left(n_j - d_j\right)n_j^{-2}\left(n_j - 1\right)^{-1}$
  - Para k tabelas de contingência, $i, l = 2, \dots, g$ grupos

COMPARAÇÃO DE CURVAS DE SOBREVIVÊNCIA

- Exemplo: Dados de Malária



Comparação	T	Valor-p	α
$\boldsymbol{H_0: S_1(t) = S_2(t) = S_3(t)}$	12,6	**0, 002**	0,05
$H_0: S_1(t) = S_2(t)$	2,5	0,114	0,0017
$\boldsymbol{H_0: S_1(t) = S_3(t)}$	7,9	**0, 005**	0,0017
$\boldsymbol{H_0: S_2(t) = S_3(t)}$	8	**0, 005**	0,0017

COMPARAÇÃO DE CURVAS DE SOBREVIVÊNCIA

- Outros testes:
- Teste: $H_0: S_1(t) = S_2(t)$

$$T = \frac{\left[\sum_{j=1}^{k} u_j \left(d_{2j} - \omega_{2j}\right)\right]^2}{\sum_{j=1}^{k} u_j \left(V_j\right)_2}$$

Teste	T	Valor-p	u_j	Efeito
Logrank	3,67	0,055	1	Detecção de diferenças considerando todos os tempos
Wilcoxon	3,19	0,074	n_j	Reforça detecção de diferenças para tempos menores
Tarone-Ware	3,43	0,064	$\sqrt{n_j}$	Meio termo dos testes acima

É estatisticamente significativa a diferença na sobrevivência entre os grupos!