EPI10 - Análise de Sobrevivência

Estimação da curva de sobrevivência

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Faculdade de Medicina
Programa de Pós-Graduação em Epidemiologia

Porto Alegre, 2022

Introdução

Introdução

- Os objetivos de uma análise estatística envolvendo dados de sobrevivência geralmente estão relacionados, em medicina ou epidemiologia, à **identificação de fatores associados** para uma certa doença **(ou desfecho de interesse)** ou à **comparação de tratamentos** em um estudo clínico enquanto controlado por outros fatores.
- Por mais complexo que seja o estudo, as respostas às perguntas de interesse são dadas a partir de um **conjunto de dados de sobrevivência**, e o passo inicial de qualquer análise estatística consiste em uma **descrição dos dados**.
- A presença de observações censuradas é, contudo, um problema para as técnicas convencionais de análise descritiva.

Introdução

- Os problemas gerados por observações censuradas podem ser ilustrados numa situação bem simples em que se tenha interesse na construção de um histograma.
  - Se a amostra não contiver observações censuradas, a construção do histograma consiste na divisão do eixo do tempo em um certo número de intervalos e, em seguida, conta-se o número de ocorrências de falhas em cada intervalo.
  - Entretanto, quando existem censuras, não é possível construir um histograma, pois não se conhece a frequência exata associada a cada intervalo.

Introdução

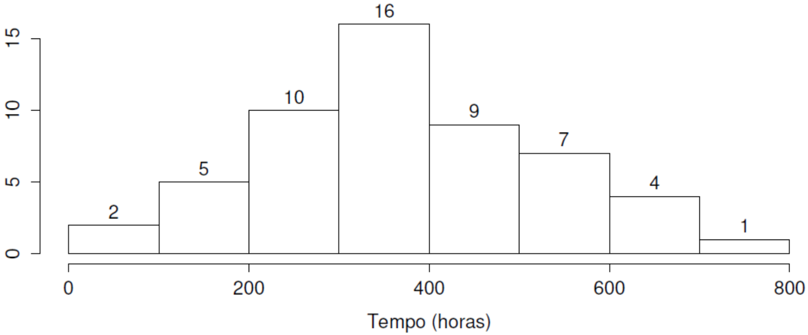
- Nos textos básicos de estatística, uma análise descritiva consiste essencialmente em encontrar medidas de tendância central e variabilidade.
- Como a presença de censuras invalida este tipo de tratamento aos dados de sobrevivência, o **principal componente da análise descritiva** envolvendo dados de tempo de vida é a **função de sobrevivência**, ou seja, $S(t) = \Pr(T > t)$[1].
- Nesta situação, o procedimento inicial é encontrar uma **estimativa para esta função de sobrevivência** e então, a partir dela, estimar as estatísticas de interesse que usualmente são o **tempo médio** ou **mediano**, alguns percentis ou certas frações de falhas em tempos fixos de acompanhamento.

[1] Note que se T é uma variável contínua, $S(t) = \Pr(T > t) = \Pr(T \geq t)$.

Estimação na ausência de censura

Estimação na ausência de censura

- O histograma a seguir representa a distribuição dos tempos até um certo evento de um grupo de 54 indivíduos, em que o evento ocorreu para todos os elementos do grupo **(observações não censuradas)**.



Estimação na ausência de censura

- A probabilidade de sobrevivência no tempo $t = 400$ horas é estimada por

$$\widehat{S}(t) = \frac{\#\text{indivíduos que não experimentaram o evento até o tempo } t = 400}{\#\text{indivíduos no estudo}}$$
$$= \frac{21}{54} = 0,39.$$

- Este número significa que 39% destes indivíduos sobrevivem mais que 400 horas.

Estimação na ausência de censura

- De forma geral, temos

$$\widehat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} I(T_i \geq t),$$

em que $I(\cdot)$ é uma **função indicadora**, e $I(A) = 1$, se o evento A ocorre, e $I(A) = 0$, se o evento A não ocorre.

Estimação na ausência de censura

Exemplo

Intervalos	Frequência	Sobrevivência
[0,100)	2	
[100,200)	5	
[200,300)	10	
[300,400)	16	
[400,500)	9	
[500,600)	7	
[600,700)	4	
[700,800)	1	

Estimação na ausência de censura

- A partir destes resultados, informações importantes sobre o tempo de vida dos indivíduos em estudo podem ser obtidas.
- Também podemos utilizar estas estimativas para estimar a função de taxa de falha em um determinado intervalo:

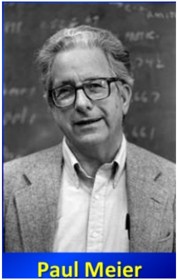
$$\widehat{\lambda}([400, 500)) = \frac{\widehat{S}(400) - \widehat{S}(500)}{(500 - 400)\widehat{S}(400)} = \frac{0,39 - 0,22}{(100)0,39} = 0,0044/\text{hora}.$$

O Estimador de Kaplan-Meier

O Estimador de Kaplan-Meier

- Na presença de censuras, o estimador de **Kaplan-Meier**, também conhecido como o **estimador limite-produto**, fornece uma comparação gráfica que de maneira apropriada leva em conta as observações censuradas.

Estimador de Kaplan-Meier[2]



- O artigo de Kaplan e Meier começa em 1952 quando Paul Meier, então na *Johns Hopkins University*, encontrou o artigo de Greenwood na duração do câncer.
- Um ano após, no *Bell Laboratories*, Kaplan se interessou tempos de vida de componentes da rede de telefones.

- Kaplan e Meier trabalharam independentemente e submeteram seus respectivos trabalhos para o *Journal of American Statistical Association*.
- O JASA os encorajou a submeterem um trabalho conjunto.
- Kaplan and Meier levaram mais 4 anos para resolverem diferenças em suas abordagens e publicaram um método que se tornou a abordagem não paramétrica padrão na análise de tempos de vida com observações censurdas.

[2]Kaplan, E. L., Meier, P., Nonparanietric estimation from incomplete observations. *JASA* **53**: 457-81, 1958.

O Estimador de Kaplan-Meier

- O estimador de Kaplan-Meier é uma adaptação da função de sobrevivência empírica que, na ausência de censuras, é definida como:

$$\widehat{S}(t) = \frac{\#\text{observações que não experimentaram o evento até o tempo } t}{\#\text{observações no estudo}}$$

$$= \frac{1}{n}\sum_{i=1}^{n} I(T_i \geq t).$$

O Estimador de Kaplan-Meier

- $\widehat{S}(t)$ é uma **função escada** com degraus nos **tempos observados de falha** de tamanho $1/n$, em que n é o tamanho da amostra.
- Se exitirem empates em um certo tempo t, o tamanho do degrau fica multiplicado pelo número de empates.
- O estimador de Kaplan-Meier, na sua construção, considera tantos intervalos de tempo quantos forem o **número de falhas distintas**.
- Os limites dos intervalos de tempo são os tempos de falha da amostra.

O Estimador de Kaplan-Meier

Relembrando: exemplo estudo de hepatite

Grupo	Tempo de sobrevivência em semanas
Controle	1+, 2+, 3, 3, 3+, 5+, 5+, 16+, 16+, 16+, 16+, 16+, 16+, 16+, 16+
Esteroide	1, 1, 1, 1+, 4+, 5, 7, 8, 10, 10+, 12+, 16+, 16+, 16+

- **Pergunta:** quantos tempos de falha distintos são observados no grupo “Esteroide”?

O Estimador de Kaplan-Meier

- Ideia intuitiva do estimador: reescrever $S(t) = \Pr(T \geq t)$ em termos de probabilidades condicionais. Seja $t_0 < t_1$, então

$$S(t_1) = \Pr(T \geq t_0)\Pr(T \geq t_1 | T \geq t_0).$$

- No exemplo,

$$S(5) = \Pr(T \geq 5) = \Pr(T \geq 1, T \geq 5) = \Pr(T \geq 1)\Pr(T \geq 5 | T \geq 1).$$

- Assim, para o indivíduo sobreviver por 5 semanas, ele vai precisar sobreviver, em um primeiro passo, à primeira semana e depois sobreviver à quinta semana, sabendo-se que ele sobreviveu à primeira semana.

O Estimador de Kaplan-Meier

- Os tempos **1** e **5** foram tomados por serem os dois primeiros tempos distintos de falha nos dados do **grupo esteroide**.
- Os passos são gerados a partir de intervalos definidos pela ordenação dos tempos de forma que cada um deles começa em um tempo de falha observado e termina no próximo tempo de falha.

Tempos ordenados (t_j)	Intervalos
0	[0,1)
1	[1,5)
5	[5,7)
7	[7,8)
8	[8,10)
10	[10,16)

O Estimador de Kaplan-Meier

- Todos os indivíduos estavam vivos em $t = 0$ e se mantêm até a primeira morte que ocorre em $t = 1$ semana.
  - Então a estimativa de $S(t)$ deve ser 1 neste intervalo compreendido entre 0 e 1 semana.
- No valor correspondente a 1 semana, a **estimativa deve cair** devido a três mortes que ocorrem neste tempo.
  - No segundo intervalo, $[1, 5)$, existem então 14 indivíduos que estavam vivos **(sob risco; $n_2 = 14$)** antes de $t = 1$ e 3 morrem $(d_2 = 3)$.
- Desta forma, a estimativa da probabilidade condicional de morte neste intervalo é 3/14 e a probabilidade de sobreviver é 1 - 3/14. Isto pode ser escrito como

$$\widehat{S}(1) = \widehat{\Pr}(T \geq 0)\widehat{\Pr}(T \geq 1 | T \geq 0) = (1) \times (1 - 3/14) = 11/14 = 0,786.$$

O Estimador de Kaplan-Meier

Exercício

1. Utilizando a mesma ideia, calcule $\widehat{S}(5)$.

O Estimador de Kaplan-Meier

Exercício

2. Utilizando a mesma ideia, complete a tabela.

Tempos ordenados (t_j)	Intervalos	d_j	n_j	$\widehat{S}(t_j+)$
0	[0,1)	0	14	1
1	[1,5)	3	14	0.786
5	[5,7)	1	9	
7	[7,8)	1	8	
8	[8,10)	1	7	
10	[10,16)	1	6	

O Estimador de Kaplan-Meier

Exercício

3. Faça o gráfico para $\widehat{S}(t)$.

O Estimador de Kaplan-Meier

- Se considerermos uma amostra de tamanho n e k $(k \leq n)$ falhas distintas $t_1 < t_2 < \ldots < t_k$, podemos reescrever $S(t)$ para qualquer t_j observado:

$$S(t_j) = (1 - q_1)(1 - q_2)\ldots(1 - q_j),$$

em que $q_j = \Pr(T \in [t_{j-1}, t_j) | T \geq t_{j-1}])$.

O Estimador de Kaplan-Meier

- O estimador de Kaplan-Meier estima estas probabilidades por

$$\hat{q}_j = \frac{\text{nº de eventos em } t_j}{\text{nº de observações sob risco em } t_j-}, j = 1, \ldots, k.$$

O Estimador de Kaplan-Meier

- Formulação equivalente
  - $t_1 < t_2 < \ldots < t_k$ são k tempos distintos de falha
  - d_j é o número de falhas em t_j, $j = 1, 2, \ldots, k$
  - n_j é o número de observações sob risco em t_j (não falharam e nem censuraram até o instante imediatamente anterior a t_j)

$$\widehat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right).$$

O Estimador de Kaplan-Meier

- O estimador de Kaplan-Meier é
  - uma função do tempo tipo escada
  - não muda entre tempos de eventos
  - não muda em tempos de censura

O Estimador de Kaplan-Meier

- Uma estimativa de $S(t)$ está sujeita a variações amostrais que devem ser descritas em termos de **estimações intervalares**.
- Para tal, precisamos de uma expressão para a **variância de $\widehat{S}(t)$**.

Fórmula de Greenwood

- A estimativa da **variância assintótica** do estimador de Kaplan-Meier é dada por:

$$\widehat{\text{Var}}\,[\widehat{S}(t)] = [\widehat{S}(t)]^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)},$$

que é conhecida como a **fórmula de Greenwood**.

O Estimador de Kaplan-Meier

- A estimativa da variância de $\widehat{S}(5)$, para o exemplo considerado, é, então, dada por:

$$\widehat{\text{Var}}\left(\widehat{S}(5)\right) = (0,698)^2 \left[\frac{3}{(14)(11)} + \frac{1}{(9)(8)}\right] = 0,0163.$$

O Estimador de Kaplan-Meier

Intervalo de confiança para $S(t)$

- Para **grandes amostras**, $\widehat{S}(t)$, para t fixo, tem uma distribuição aproximadamente normal.
- Assim, um intervalo de $100(1-\alpha)\%$ de confiança para $S(t)$ é dado por

$$\widehat{S}(t) \pm z_{\alpha/2}\sqrt{\widehat{\text{Var}}\,(\widehat{S}(t))},$$

em que $\alpha/2$ é o $\alpha/2$-percentil da distribuição normal padrão.

- O intervalo de 95% de confiança para $S(5)$ é $0,698 \pm 1,96\sqrt{0,0163} = (0,45; 0,95)$.

O Estimador de Kaplan-Meier

IC para $S(t)$: comentários e alternativas

- Este intervalo pode apresentar **limite inferior negativo** e **limite superior maior que 1**.
  - $\widehat{U}(t) = \log[-\log \widehat{S}(t)]$
  - $\widehat{Var}(\widehat{U}(t)) = \dfrac{\sum_{j:t_j<t} \frac{d_j}{n_j(n_j-d_j)}}{[\log \widehat{S}(t)]^2}$
  - Intervalo aproximado de $100(1-\alpha)\%$ de confiança para $S(t)$

$$[\widehat{S}(t)]^{\exp\left\{\pm z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{U}(t))}\right\}}.$$

Exemplo computacional

Estudo de Hepatite

```
# install.packages("survival")

# Carregando pacote
library(survival)

# Dados do exemplo

head(df.hep)
```

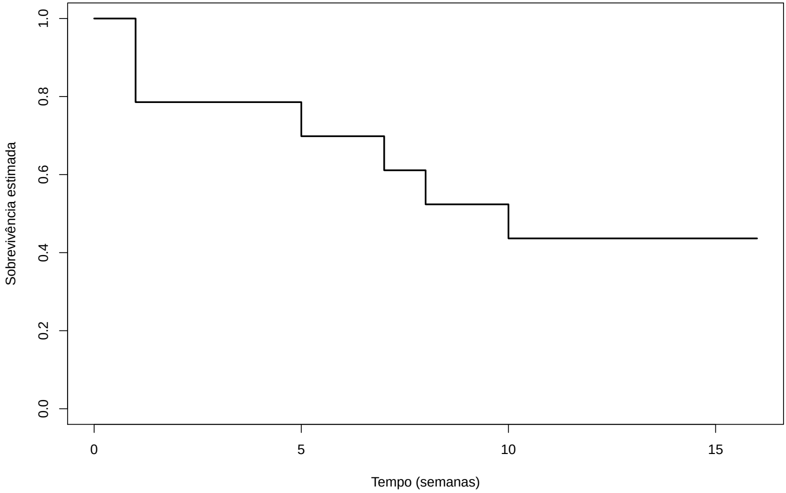
```
##   tempo cens    grupo
## 1     1    0 Controle
## 2     2    0 Controle
## 3     3    1 Controle
## 4     3    1 Controle
## 5     3    0 Controle
## 6     5    0 Controle
```

Estudo de Hepatite

```
ekm <- survfit(Surv(time = tempo, event = cens) ~ 1,
               data = df.hep,
               subset = grupo == "Esteroide",
               conf.type = "log-log")

plot(ekm, conf.int = FALSE,
     lwd = 2, xlab = "Tempo (semanas)",
     ylab = "Sobrevivência estimada")
```

Estudo de Hepatite



Estudo de Hepatite

```
summary(ekm)
```

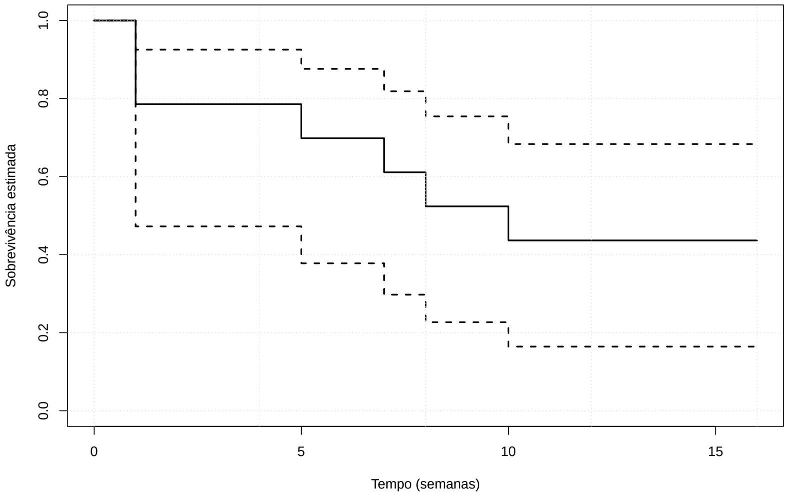
```
## Call: survfit(formula = Surv(time = tempo, event = cens) ~ 1, data = df.hep,
##     subset = grupo == "Esteroide", conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     14       3    0.786   0.110        0.472        0.925
##     5      9       1    0.698   0.128        0.378        0.876
##     7      8       1    0.611   0.138        0.298        0.819
##     8      7       1    0.524   0.143        0.227        0.754
##    10      6       1    0.437   0.144        0.164        0.683
```

Estudo de Hepatite

Intervalo de confiança

```
plot(ekm, conf.int = TRUE,
     lwd = 2, xlab = "Tempo (semanas)",
     ylab = "Sobrevivência estimada")
abline(h = seq(0, 1, by = 0.2),
       v = seq(0, 16, by = 4),
       col = "lightgrey", lty = 3)
```

Estudo de Hepatite



Estudo de Hepatite

Estratificando por grupos de tratamentos

```
ekm <- survfit(Surv(time = tempo, event = cens) ~ grupo,
               data = df.hep,
               conf.type = "log-log")

summary(ekm)
```

```
## Call: survfit(formula = Surv(time = tempo, event = cens) ~ grupo, data = df.hep,
##     conf.type = "log-log")
##
##                 grupo=Controle
##         time       n.risk      n.event     survival      std.err lower 95% CI upper 95%
##        3.000       13.000        2.000        0.846        0.100        0.512        0.9
##
##                 grupo=Esteroide
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     14       3    0.786   0.110        0.472        0.925
##     5      9       1    0.698   0.128        0.378        0.876
##     7      8       1    0.611   0.138        0.298        0.819
##     8      7       1    0.524   0.143        0.227        0.754
##    10      6       1    0.437   0.144        0.164        0.683
```

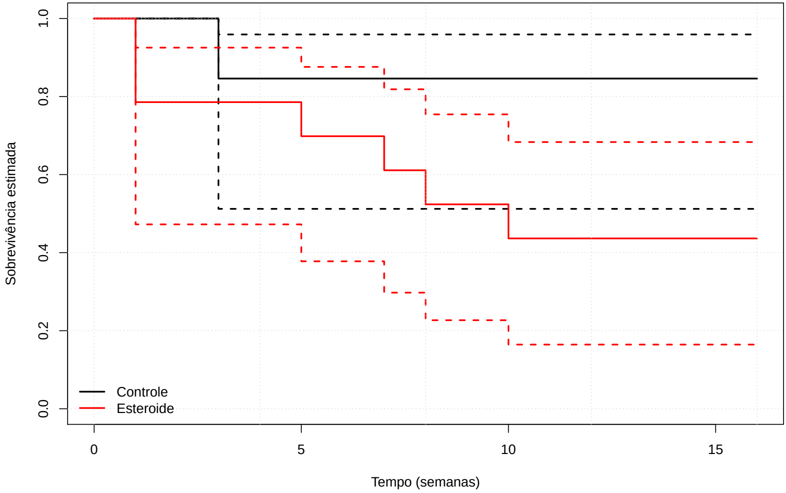
Estudo de Hepatite

```
plot(ekm, conf.int = TRUE,
     col = c("black", "red"),
     lwd = 2, xlab = "Tempo (semanas)",
     ylab = "Sobrevivência estimada")

abline(h = seq(0, 1, by = 0.2),
       v = seq(0, 16, by = 4),
       col = "lightgrey", lty = 3)

legend("bottomleft",
       c("Controle", "Esteroide"),
       col = c("black", "red"),
       lwd = 2, bty = "n")
```

Estudo de Hepatite



Para casa

1. Leia o capítulo 2 do livro **Análise de sobrevivência aplicada**[3].
2. Leia os capítulo 4 do livro **Análise de sobrevivência: teoria e aplicações em saúde**[4].

---

[3] Colosimo, E. A. e Giolo, S. R. **Análise de sobrevivência aplicada**, Blucher, 2006.

[4] Carvalho, M. S., Andreozzi, V. L., Codeço, C. T., Campos, D. P., Barbosa, M. T. S. e Shimakura, E. S. **Análise de sobrevivência: teoria e aplicações em saúde**, 2ª ed. Editora Fiocruz, 2011.

Próxima aula

- Comparação de curvas de sobrevivência.

Por hoje é só!

Bons estudos!