Estimação Não Paramétrica da Função de Sobrevivência

Estimadores Tábua de Vida e Kaplan-Meier

MAE 514 - Introdução à Análise de Sobrevivência e Aplicações

IME-USP

Tópicos

Dados

Estudo para avaliar o tempo livre de obstrução de sondas plásticas e metálicas em pacientes com câncer de fígado

- 35 pacientes
- 17 recebem sondas plásticas e 18 recebem sondas metálicas, via aleatorização
- Evento de interesse: obstrução da sonda (paciente fica *ictérico*
- Origem do tempo de acompanhamento: Implante da sonda

Variáveis

- Tempo entre a colocação da sonda e a primeira obstrução
- delta: indicador de obstrução (variável binária)
- Tipo: Tipo da Sonda (plástica ou metálica)

Dados- Sondas

Tempo	delta	Tipo	Tempo	delta	Tipo
3	1	Plástica	57	1	Metálica
7	0	Plástica	67	0	Metálica
8	1	Plástica	68	0	Metálica
15	1	Plástica	71	1	Metálica
31	1	Plástica	71	1	Metálica
37	1	Plástica	77	1	Metálica
70	0	Plástica	84	1	Metálica
80	1	Plástica	94	1	Metálica
93	1	Plástica	110	0	Metálica
121	1	Plástica	119	1	Metálica
153	0	Plástica	147	1	Metálica
190	1	Plástica	249	0	Metálica
194	1	Plástica	309	1	Metálica
257	1	Plástica	359	1	Metálica
328	1	Plástica	365	0	Metálica
407	0	Plástica	498	0	Metálica
866	1	Plástica	536	1	Metálica
			726	0	Metálica

Análise exploratória

Leitura dos dados no R

```
> sondas<-read.csv("sondas.csv",header=T)
> sondas.plasticas<-subset(sondas,Tipo=="Plástica")
```

Medidas resumo

```
     Tempo           delta             Tipo
 Min.   :  3.0   Min.   :0.0000   Metálica:18
 1st Qu.: 69.0   1st Qu.:0.0000   Plástica:17
 Median :110.0   Median :1.0000
 Mean   :196.2   Mean   :0.6857
 3rd Qu.:283.0   3rd Qu.:1.0000
 Max.   :866.0   Max.   :1.0000
```

Estimação não paramétrica

- Função tipo escada
- Deve satisfazer as propriedades básicas de uma função de sobrevivência teórica:
  - $S(0) = 1$
  - $S(t) \to 0$ quando $t \to +\infty$
  - Função não crescente
  - Contínua à direita

Estimador Tábua de Vida

- O período de realização do experimento é particionado segundo instantes pré-fixados $\xi_1, \ldots, \xi_K$.
- Em cada intervalo $[\xi_{j-1}, \xi_j)$, define-se
  - d_j: número de eventos
  - w_j: número de observações censuradas
  - n_j: número de itens em risco em ξ_{j-1}

$$\widehat{S}_{TV}(t) = \begin{cases} 1, & t \in [\xi_0, \xi_1); \\ \prod_{\ell=1}^{j-1} \left(1 - \frac{d_\ell}{n_\ell - 0{,}5 w_\ell}\right), & t \in [\xi_{j-1}, \xi_j),\ j = 2, \ldots, K+1 \end{cases}$$

Obs.: $\xi_0 = 0$, $\xi_{K+1} = +\infty$.

Obtenção das estimativas TV para $S(t)$ no R

- Várias bibliotecas podem ser utilizadas: `KMsurv`, `demography`, etc.
- Na biblioteca `KMsurv`, usar a função `lifetab`
- A entrada dos dados se dá na forma *atuarial* (isto é, é necessário fornecer os valores de $\{(\xi_j, n_j, d_j, w_j), j = 1, \ldots, K\}$.

Código no R

```
> library(KMsurv)
> xis <- c(0,75,150,300,450,NA)
> nindiv <- c(17,10,7,3,1)
> ncens <- c(2,0,1,1,0)
> neventos <- c(5,3,3,1,1)
> lifetab(xis,nindiv[1],ncens,neventos)
```


Obtenção das estimativas TV para $S(t)$ no R

Código no R

```
> sondas.tv<-lifetab(xis,nindiv[1],ncens,neventos)
> sondas.tv[,c(1:5,8)]
```

Estimativas Tábua de Vida

	nsubs	nlost	nrisk	nevent	surv	se.surv
0-75	17	2	16.0	5	1.0000000	0.0000000
75-150	10	0	10.0	3	0.6875000	0.1158781
150-300	7	1	6.5	3	0.4812500	0.1284732
300-450	3	1	2.5	1	0.2591346	0.1167931
450-NA	1	0	1.0	1	0.1554808	0.1065697

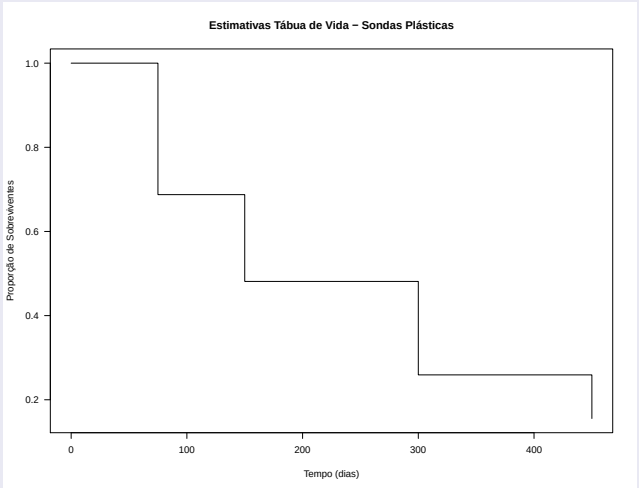
Representação Gráfica - Tábua de Vida

Código no R

```
> plot(xis[1:5], sondas.tv[,5], type="s",
main="Estimativas Tábua de Vida-Sondas Plásticas",
xlab="Tempo (dias)",
ylab="Proporção de Sobreviventes",las=1)
```


Representação Gráfica - Tábua de Vida

Gráfico das Estimativas Tábua de Vida



Estimador Kaplan-Meier

- Obtido ao se aumentar o número de elementos na partição até que ela seja definida pelos instantes de falha observados $t_1^*, \ldots, t_L^*$, em que L é o total dos instantes distintos de falha.
- A cada instantes de falha t_j^*, definimos
  - d_j: total de falhas observadas no instante;
  - n_j: número de indivíduos em risco no instante imeditamente anterior a t_j^* (isto é, $t_j^* - \varepsilon$).

$$\widehat{S}_{KM}(t) = \prod_{j:t_j^* \leq t} \left(\frac{n_j - d_j}{n_j} \right)$$

Obtenção das estimativas KM para $S(t)$ no R

- A forma mais popular é usando-se a função `survfit` da bibliteca `survival`
- Os dados de entrada incluem os tempos efetivamente observados e os valores de δ

Código no R

```
> library(survival)
> km.plasticas<-survfit(Surv(Tempo,delta)~1,
data=sondas.plasticas)
> summary(km.plasticas)
```

Obtenção das estimativas KM para $S(t)$ no R

Estimativas Kaplan-Meier

```
Call: survfit(formula = Surv(Tempo, delta) ~ 1, data = sondas.plasticas)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    3     17       1    0.941  0.0571        0.8357        1.000
    8     15       1    0.878  0.0807        0.7337        1.000
   15     14       1    0.816  0.0963        0.6472        1.000
   31     13       1    0.753  0.1074        0.5693        0.996
   37     12       1    0.690  0.1153        0.4974        0.958
   80     10       1    0.621  0.1227        0.4217        0.915
   93      9       1    0.552  0.1270        0.3518        0.867
  121      8       1    0.483  0.1285        0.2868        0.814
  190      6       1    0.403  0.1299        0.2139        0.758
  194      5       1    0.322  0.1264        0.1492        0.695
  257      4       1    0.242  0.1177        0.0930        0.628
  328      3       1    0.161  0.1024        0.0463        0.560
  866      1       1    0.000     NaN            NA           NA
```

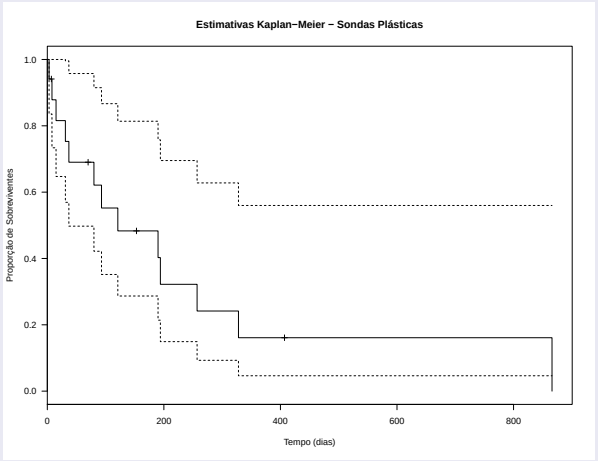

Representação Gráfica - Kaplan-Meier

Código no R

```
> plot(km.plasticas,
main="Estimativas Kaplan-Meier - Sondas Plásticas",
xlab="Tempo (dias)",
ylab="Proporção de Sobreviventes",las=1)
```

Representação Gráfica - Kaplan-Meier

Gráfico das Estimativas Kaplan-Meier



Sondas Plásticas x Metálicas - Comparação Descritiva

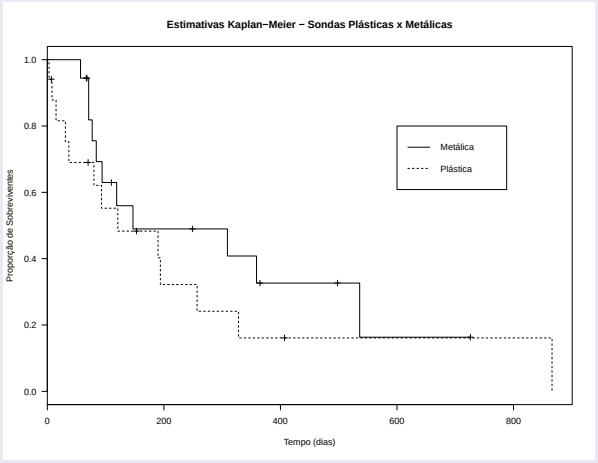
- Usando o estimador Kaplan-Meier

Código no R

```
> km.sondas<-survfit(Surv(Tempo,delta)~Tipo,
data=sondas)
plot(km.sondas,las=1, main="Estimativas
Kaplan-Meier - Sondas Plásticas x Metálicas",
xlab="Tempo (dias)",
ylab="Proporção de Sobreviventes",las=1,lty=c(1,2))
> legend(600,0.8,lty=c(1,2),c("Metálica",
"Plástica"))
```

Representação Gráfica - Kaplan-Meier

Comparação do tipo de sonda



Sondas Plásticas x Metálicas - Medianas

Código no R

```
> km.sondas
```

Estimativas do tempo mediano livre de obstrução

```
Call: survfit(formula = Surv(Tempo, delta) ~ Tipo, data = sondas)

               records n.max n.start events median 0.95LCL 0.95UCL
Tipo=Metálica       18    18      18     11    147      94      NA
Tipo=Plástica       17    17      17     13    121      37      NA
```

Normalidade

Resultado Importante

Para t fixo,

$$\frac{\widehat{S}_{KM}(t) - S(t)}{\sqrt{\text{Var}[\widehat{S}_{KM}(t)]}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

para $n \to +\infty$.

- Inicialmente esse resultado é determinado empiricamente
- Formalmente, depende da teoria de martingais

Intervalo de confiança

A princípio, intervalos de confiança pontuais assintóticos podem ser obtidos por

$$\text{IC}[S(t), \gamma] = \left[\widehat{S}_{KM}(t) - z_{\gamma/2}\sqrt{\text{Var}[\widehat{S}_{KM}(t)]};\ \widehat{S}_{KM}(t) + z_{\gamma/2}\sqrt{\text{Var}[\widehat{S}_{KM}(t)]}\right]$$

- $\gamma \in (0,1)$
- $z_{\gamma/2}$ obtido da distribuição Normal padrão.

Problema: $\text{Var}[\widehat{S}(t)]$ precisa ser estimada.

Fórmula de Greenwood

$$\widehat{\text{Var}}[\widehat{S}_{KM}(t)] \approx [\widehat{S}_{KM}(t)]^2 \sum_{j:t_j^* \leq t} \frac{d_j}{n_j(n_j - d_j)}$$

Passos para obtenção da fórmula de Greenwood (empírico):

- Supõe-se que
  - condicional a n_j, $d_j \sim \text{Bin}(n_j, q_j)$, com

$$q_j = P(\text{Falha em } t_j^* \mid \text{Em risco em } t_j^* - \varepsilon)$$

  - d_j, $j = 1, \ldots, L$ são independentes
- Encontra-se a variância assintótica de $\log \widehat{S}_{KM}(t)$
- Aplicando o método Delta, obtem-se a variância de

$$\widehat{S}_{KM}(t) = e^{\log[\widehat{S}_{KM}(t)]}.$$

Três métodos populares

- Natural (`plain`)
- Utilizando-se a tranformação logarítimica (`log`)
- Utilizandos-se a tranformação “log menos log”(`log-log`)

Método Natural

Usando-se a fórmula de Greenwood, estima-se a variância de $\widehat{S}_{KM}(t)$ e calcula-se

$$\text{IC}[S(t),\gamma] = \left[\widehat{S}_{KM}(t) - z_{\gamma/2}\sqrt{\widehat{\text{Var}}[\widehat{S}_{KM}(t)]};\ \widehat{S}_{KM}(t) + z_{\gamma/2}\sqrt{\widehat{\text{Var}}[\widehat{S}_{KM}(t)]}\right]$$

- Valores maiores do que 1
- Valores menores do que zero

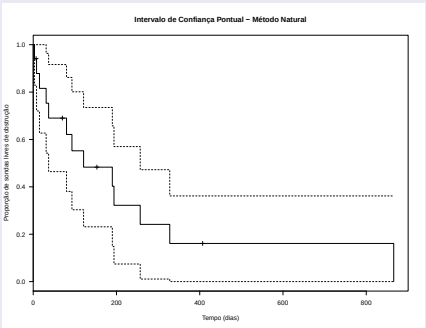
$\Rightarrow$ Truncagem

Método Natural - R

Opção `plain`

```
> plot(survfit(Surv(Tempo,delta)~1,
      data=sondas.plasticas, conf.type="plain")
```

Gráfico das Estimativas Kaplan-Meier



Transformação Logarítimica

Defina $Y_t = \log[\widehat{S}_{KM}(t)]$. Segue pelo *método delta* que

$$\text{Var}(Y_t) = [1/\widehat{S}_{KM}(t)]^2 \text{Var}[\widehat{S}_{KM}(t)]$$

Logo,

$$\text{IC}[\log[S(t)], \gamma] = \left[Y_t - z_{\gamma/2}\sqrt{\widehat{\text{Var}}(Y_t)};\ Y_t + z_{\gamma/2}\sqrt{\widehat{\text{Var}}(Y_t)} \right]$$

e assim,

$$\text{IC}[S(t), \gamma] = \left[e^{Y_t - z_{\gamma/2}\sqrt{\widehat{\text{Var}}(Y_t)}};\ e^{Y_t + z_{\gamma/2}\sqrt{\widehat{\text{Var}}(Y_t)}} \right]$$

com

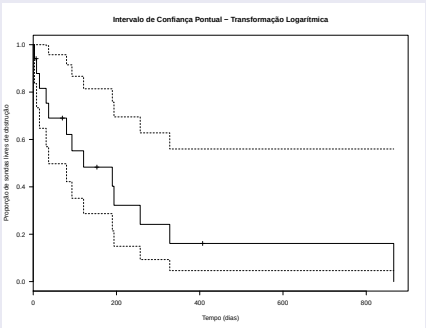
$$\widehat{\text{Var}}(Y_t) = [1/\widehat{S}_{KM}(t)]^2 \widehat{\text{Var}}[\widehat{S}_{KM}(t)]$$

Transformação Logarítimica - R

Opção `log`

```
> plot(survfit(Surv(Tempo,delta)~1,
      data=sondas.plasticas, conf.type="log")
```

Gráfico das Estimativas Kaplan-Meier



Transformação Log-Log

Defina $W_t = \log\{-\log[\widehat{S}_{KM}(t)]\}$. Segue pelo *método delta* que

$$\text{Var}(W_t) = \{1/\log[\widehat{S}_{KM}(t)]\}^2 \text{Var}[\widehat{S}_{KM}(t)]$$

Logo,

$$\text{IC}[\log[S(t)], \gamma] = \left[W_t - z_{\gamma/2}\sqrt{\widehat{\text{Var}}(W_t)};\ W_t + z_{\gamma/2}\sqrt{\widehat{\text{Var}}(W_t)}\right]$$

e assim,

$$\text{IC}[S(t), \gamma] = \left[e^{-e^{W_t + z_{\gamma/2}\sqrt{\widehat{\text{Var}}(W_t)}}};\ e^{-e^{W_t - z_{\gamma/2}\sqrt{\widehat{\text{Var}}(W_t)}}}\right]$$

com

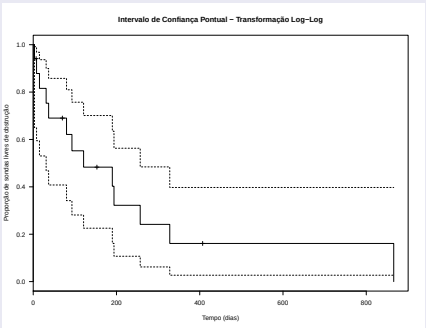
$$\widehat{\text{Var}}(W_t) = [1/\log\{\widehat{S}_{KM}(t)]\}^2 \widehat{\text{Var}}[\widehat{S}_{KM}(t)]$$

Transformação Log-Log - R

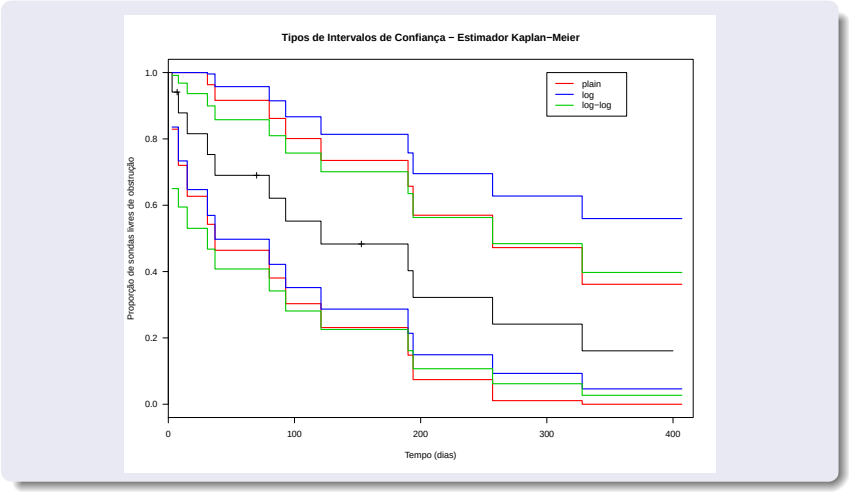
Opção `log`

```
> plot(survfit(Surv(Tempo,delta)~1,
      data=sondas.plasticas, conf.type="log-log")
```

Gráfico das Estimativas Kaplan-Meier



Intervalos de Confiança - Comparação



Comentários

- O IC *natural* precisa, eventualmente, ser truncado.
- O IC *log* é bem assimétrico.
- O IC *log-log* apresenta maior simetria.
- O IC *log-log* tem maior proximidade com o IC *natural*.
- Por omissão, o R inclui automaticamente o IC *log* no gráfico do estimador Kaplan-Meier quando existe apenas um grupo. Para mais de um grupo o programa não apresenta os intervalos.