Revisão Inicial

Reposição de aulas

MAE 514 - Introdução à Análise de Sobrevivência e Aplicações

IME-USP

A análise de sobrevivência se ocupa do estudo do tempo necessário até a ocorrência de eventos de interesse

- Tempo de sobrevivência de pacientes
- Tempo de remissão de sintomas de uma doença
- Tempo entre empréstimo e *default* de empresas do setor terciário
- Tempo em que um cliente permanece fiel a certo produto
- Tempo de desemprego de certa categoria profissional
- etc...

Formalização

Tempo de Sobrevivência

T: variável aleatória não-negativa

Caracterizada por

$$S(t) = P(T > t) \text{ ou } \alpha(t) = \lim_{\Delta t \to 0} \frac{P(T \in (t,\ t + \Delta t] \mid T > t)}{\Delta t}$$

$\alpha(t)$: Função de taxa de falha ou função de risco.
Recorde que

- $S(t) = e^{-\int_0^t \alpha(s)ds}$
- $\alpha(t) = f(t)/S(t)$
- $f(t) = \alpha(t)e^{-\int_0^t \alpha(s)ds}$

$\Rightarrow$ equivalência nas quantidades que regem a v.a. T.

Observações Incompletas

- Censuras à direita
  - Tipo I
  - Tipo II
  - Aleatória (Tipo III)
- Censuras à esquerda
- Truncagem (direita ou esquerda)

Quantidades observáveis

$$Z = \min(T, C)$$
$$\delta = I(T \leq C)$$

T: tempo de falha; C tempo de censura

Estimação Não Paramétrica de $S(t)$

Estimador Kaplan-Meier

$$\widehat{S}(t) = \prod_{k:\tau_k \leq t} \left(\frac{n_k - d_k}{n_k} \right)^{\delta_i}$$

- τ_k: instantes **observados** de falha ou censura, $k = 1, \ldots, L$
- n_k: total de indivíduos *em risco* em τ_k
- d_k: total de falhas em τ_k

Estimador Produto-Limite: Produto de estimativas de probabilidades condicionais, tomadas em uma partição definida a partir dos instantes de falha observados

Exemplo: Dados de Remissão (Leucemia)

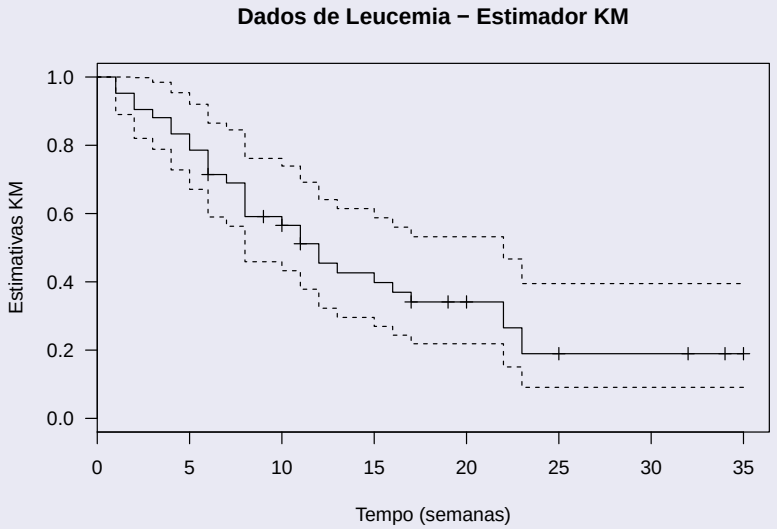
grupo	tempo	delta	log(wbc)		grupo	tempo	delta	log(wbc)
1	6	1	2.31		0	1	1	2.80
1	6	1	4.06		0	1	1	5.00
1	6	1	3.28		0	2	1	4.91
1	7	1	4.43		0	2	1	4.48
1	10	1	2.96		0	3	1	4.01
1	13	1	2.88		0	4	1	4.36
1	16	1	3.60		0	4	1	2.42
1	22	1	2.32		0	5	1	3.49
1	23	1	2.57		0	5	1	3.97
1	6	0	3.20		0	8	1	3.52
1	9	0	2.80		0	8	1	3.05
1	10	0	2.70		0	8	1	2.32
1	11	0	2.60		0	8	1	3.26
1	17	0	2.16		0	11	1	3.49
1	19	0	2.05		0	11	1	2.12
1	20	0	2.01		0	12	1	1.50
1	25	0	1.78		0	12	1	3.06
1	32	0	2.20		0	15	1	2.30
1	32	0	2.53		0	17	1	2.95
1	34	0	1.47		0	22	1	2.73
1	35	0	1.45		0	23	1	1.97

Exemplo: Dados de Remissão (Leucemia)

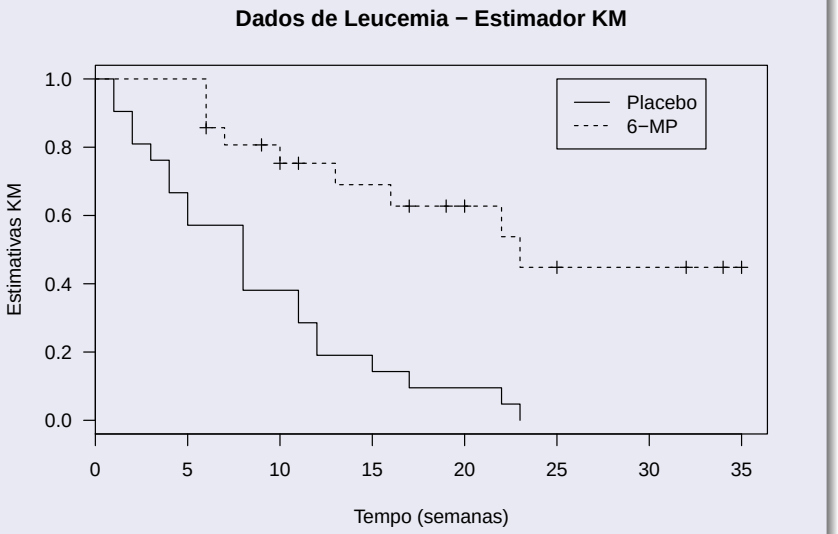
Código R

```
tempo <- c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,
           19,20,25,32,32,34,35,1,1,2,2,3,4,4,
           5,5,8,8,8,8,11,11,12,12,15,17,22,23)
delta <- c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,
           0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,
           1,1,1,1,1,1,1,1)
grupo <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
           1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,
           0,0,0,0,0,0,0,0)
library(survival)
grafico <- survfit(Surv(tempo,delta)~1)
plot(grafico,main="Dados de Leucemia - Estimador KM",
     xlab="Tempo (semanas)", ylab="Estimativas KM")
```


Exemplo: Dados de Remissão (Leucemia)



Exemplo: Dados de Remissão (Leucemia)



Comparando Grupos - Abordagem Não Paramétrica

Exemplo: Remissão de Sintomas (Leucemia)
Objetivo: Comparação de dois grupos

Droga 6-MP X Placebo

Supondo

- T_{1i}: Tempo livre de sintomas para paciente i recebendo droga 6-MP

$$T_{1i} \sim S_1(t)$$

- T_{2i}: Tempo livre de sintomas para paciente i recebendo placebo

$$T_{2i} \sim S_2(t)$$

6-MP sem efeito $\Leftrightarrow S_1(t) = S_2(t),\ \forall t \geq 0.$

Comparando Grupos - Abordagem Não Paramétrica

Duas amostras de duas populações:

- População 1: $\{(Z_{1i}, \delta_{1i}),\ i = 1, \ldots, n_1\}$
- População 2: $\{(Z_{2i}, \delta_{2i}),\ i = 1, \ldots, n_2\}$

$$Z_{ki} = \min(T_{ki}, C_{ki})$$
$$\delta_{ki} = I(T_{ki} \leq C_{ki}),$$

$k = 1, 2$, com $T_{ki} \sim S_k(t)$.

Objetivo: Testar

$$H_0 : S_1(t) = S_2(t),\ \forall t \geq 0$$

versus

$$H_1 : S_1(t) \neq S_2(t), \text{para algum } t \geq 0.$$

Comparando Grupos - Abordagem Não Paramétrica

Para $\tau_1, \ldots, \tau_L$, instantes observados de falha, defina para o grupo $k(= 1, 2)$:

- n_{ki}: número de unidades *em risco* no instante imediatamente anterior a τ_i
- d_{ki}: número de eventos observados no instante τ_i

Testes de Postos Lineares

$$Q_W^2 = \frac{[\sum_{\ell=1}^{L} w_\ell (d_{1\ell} - e_{1\ell})]^2}{\sum_{\ell=1}^{L} w_\ell^2 \left[\frac{n_{1\ell} n_{2\ell} d_\ell (n_\ell - d_\ell)}{n_\ell^2 (n_\ell - 1)}\right]}$$

- $w_\ell = 1,\ \ell = 1, \ldots L \Rightarrow$ teste log-rank
- $w_\ell =$estimador Kaplan-Meier em $t_\ell \Rightarrow$ teste Gehan-Wilcoxon

Comparando Grupos - Não Paramétrica - R

Família G-Rho de Harrington & Fleming (1982)

$$w_\ell = [\widehat{S}_{KM}(t_{\ell-1})]^\rho$$

- $\rho = 0 \Rightarrow$ log-rank
- $\rho = 1 \Rightarrow$ Gehan-Wilcoxon (Peto-Prentice)

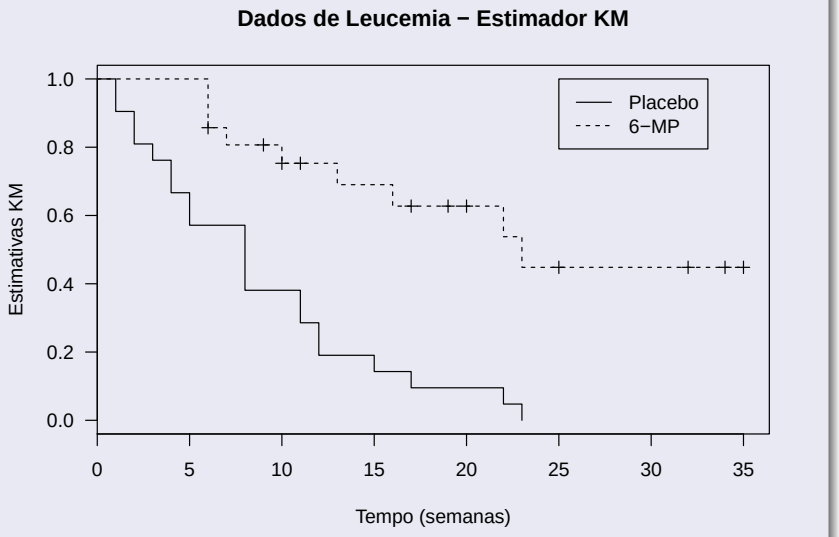
log-rank

```
survdiff(Surv(tempo,delta)~grupo,rho=0)
```

Gehan-Wilcoxon

```
survdiff(Surv(tempo,delta)~grupo,rho=1)
```


Comparando Grupos - Não Paramétrica - Leucemia



Comparando Grupos - Não Paramétrica - Leucemia

log-rank

```
Call:
survdiff(formula = Surv(tempo, delta) ~ grupo, rho = 0)

         N Observed Expected (O-E)^2/E (O-E)^2/V
grupo=0 21       21     10.7      9.77      16.8
grupo=1 21        9     19.3      5.46      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

Gehan-Wilcoxon

```
Call:
survdiff(formula = Surv(tempo, delta) ~ grupo, rho = 1)

         N Observed Expected (O-E)^2/E (O-E)^2/V
grupo=0 21    14.55     7.68      6.16      14.5
grupo=1 21     5.12    12.00      3.94      14.5

 Chisq= 14.5  on 1 degrees of freedom, p= 0.000143
```


Modelagem Paramétrica

Modelos paramétricos são definidos em termos de $\alpha(t)$, por exemplo,

- $\alpha(t) = \theta \Rightarrow$ modelo exponencial
- $\alpha(t) = \rho\theta^\rho t^{\rho-1} \Rightarrow$ modelo Weibull
- $\alpha(t) = \lambda^r t^{r-1} e^{-\lambda t}/\Gamma(\lambda t, r) \Rightarrow$ modelo Gama
- etc

Complexidade versus Aplicabilidade

Classes de modelos:

- Modelos de vida acelerada ou modelos de locação-escala
- Modelos de Risco proporcional

Modelagem Paramétrica - Estimação

Dados:

- $\{(Z_i, \delta_i),\ i = 1, \ldots, n\}$, $Z_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$
- $T_i \sim F_\theta$, uma família de distribuições indexada por θ

Suposições:

1. Censura independente
2. Censura não-informativa

Verossimilhança

$$L(\theta) = \prod_{i=1}^{n} [f_\theta(z_i)]^{\delta_i} [S_\theta(z_i)]^{1-\delta_i}$$

Maximização através de métodos iterativos - Newton-Raphson

Modelagem Paramétrica - Testes de hipóteses

Supondo

$$T_i \sim F_\theta, \quad \theta \in \mathbb{R}^p$$

Hipóteses

$$H_o : \theta = \theta_0 \text{ versus } H_a : \theta \neq \theta_0$$

1. Teste de Wald (W)
2. Teste Escore ou Rao (R)
3. Teste da razão de verossimilhanças (Λ)

Distribuição assintótica

$$W, R, \Lambda \xrightarrow{\mathcal{D}} \chi^2_p$$

- Os testes são assintoticamente equivalentes
- Λ é mais poderoso (pequenas amostras)
- R não necessita do EMV para avaliação

Modelo de Regressão Paramétrico

Dados: $\{(Z_i, \delta_i, \boldsymbol{X}_i),\ i = 1, \ldots, n\}$ com

- Z_i e δ_i como antes
- $\boldsymbol{X}_i^T = (X_{1i}, \ldots, X_{pi})$ vetor de covariáveis

Possíveis abordagens:

Modelo de vida acelerada

$$S(t \mid \boldsymbol{X}_i) = S_0(\Psi(\boldsymbol{X}_i)t), \text{ usualmente } \Psi(\boldsymbol{X}_i) = e^{\boldsymbol{X}_i^T \beta}$$

item Modelo de locação escala

$$Y_i = \log Z_i = \boldsymbol{X}_i^T \beta + \sigma\varepsilon, \text{ com } \varepsilon \sim F_o$$

Modelo de taxas de falha proporcionais

$$\alpha(t \mid \boldsymbol{X}_i) = \alpha_o(t) \exp\{\boldsymbol{X}_i^T \beta\}$$

Modelo de Regressão Paramétrico - Estimação

Semelhante ao caso para populações homogêneas:

Verossimilhança

$$L(\beta) = \prod_{i=1}^{n} [f_\theta(z_i \mid \beta)]^{\delta_i} [S_\theta(z_i \mid \beta)]^{1-\delta_i}$$

- Necessita método numérico - Newton-Raphson
- β: Efeitos das covariáveis
- θ: Parâmetros de Perturbação

Exemplo: Modelo de Regressão Weibull - Vida Acelerada

$$S(t \mid \boldsymbol{X}_i) = e^{-[\Psi(\mathbf{X_i})t]^\rho} \text{ e } f(t \mid \boldsymbol{X}_i) = \rho[\Psi(\boldsymbol{X}_i)]^\rho t^{\rho-1} e^{-[\Psi(\mathbf{X_i})t]^\rho}$$

com $\Psi(\boldsymbol{X}_i) = e^{\boldsymbol{X}_i^T \beta}$.

- β: Parâmetros de regressão
- ρ: Parâmetro de perturbação

Note que $\rho = 1 \Rightarrow$ modelo Exponencial.

Modelos Paramétricos: Observações

- Equivalência entre modelos de vida acelerada e de locação-escala
- Modelos Exponencial e Weibull
  - São modelos de riscos proporcionais
  - São modelos de locação escala
- Modelos de riscos proporcionais:
  - Risco relativo não varia com o tempo
  - Interpretação dos parâmetros mais simples
- Programas estatísticos:
  - Consideram modelos de locação-escala
  - Biblioteca `survival` no R: função `survreg`

Exemplo: Dados de Remissão (Leucemia)

grupo	tempo	delta	log(wbc)		grupo	tempo	delta	log(wbc)
1	6	1	2.31		0	1	1	2.80
1	6	1	4.06		0	1	1	5.00
1	6	1	3.28		0	2	1	4.91
1	7	1	4.43		0	2	1	4.48
1	10	1	2.96		0	3	1	4.01
1	13	1	2.88		0	4	1	4.36
1	16	1	3.60		0	4	1	2.42
1	22	1	2.32		0	5	1	3.49
1	23	1	2.57		0	5	1	3.97
1	6	0	3.20		0	8	1	3.52
1	9	0	2.80		0	8	1	3.05
1	10	0	2.70		0	8	1	2.32
1	11	0	2.60		0	8	1	3.26
1	17	0	2.16		0	11	1	3.49
1	19	0	2.05		0	11	1	2.12
1	20	0	2.01		0	12	1	1.50
1	25	0	1.78		0	12	1	3.06
1	32	0	2.20		0	15	1	2.30
1	32	0	2.53		0	17	1	2.95
1	34	0	1.47		0	22	1	2.73
1	35	0	1.45		0	23	1	1.97

Exemplo: Dados de Remissão (Leucemia)

Código R - Distribuição Weibull

```
logwbc <- c(2.31,4.06,3.28,4.43,2.96,2.88,
            3.60,2.32,2.57,3.20,2.80,2.70,
            2.60,2.16,2.05,2.01,1.78,2.20,
            2.53,1.47,1.45,2.80,5.00,4.91,
            4.48,4.01,4.36,2.42,3.49,3.97,
            3.52,3.05,2.32,3.26,3.49,2.12,
            1.50,3.06,2.30,2.95,2.73,1.97)
dados <- data.frame(tempo,delta,grupo,logwbc)
regreparam <- survreg(Surv(tempo,delta)~grupo+
                      logwbc,data=dados)
summary(regreparam)
```


Exemplo: Dados de Remissão (Leucemia)

Código R - Distribuição Weibull

```
Call:
survreg(formula = Surv(tempo, delta) ~ grupo + logwbc, data = dados)
             Value Std. Error      z        p
(Intercept)  4.740      0.368 12.87 6.62e-38
grupo        0.659      0.189  3.49 4.90e-04
logwbc      -0.807      0.108 -7.45 9.68e-14
Log(scale)  -0.793      0.142 -5.58 2.45e-08

Scale= 0.452

Weibull distribution
Loglik(model)= -90.1   Loglik(intercept only)= -116.4
Chisq= 52.68 on 2 degrees of freedom, p= 3.6e-12
Number of Newton-Raphson Iterations: 7
n= 42
```

Exemplo: Dados de Remissão (Leucemia)

Código R - Distribuição Gama

```
logwbc <- c(2.31,4.06,3.28,4.43,2.96,2.88,
            3.60,2.32,2.57,3.20,2.80,2.70,
            2.60,2.16,2.05,2.01,1.78,2.20,
            2.53,1.47,1.45,2.80,5.00,4.91,
            4.48,4.01,4.36,2.42,3.49,3.97,
            3.52,3.05,2.32,3.26,3.49,2.12,
            1.50,3.06,2.30,2.95,2.73,1.97)
dados <- data.frame(tempo,delta,grupo,logwbc)
regreparam <- survreg(Surv(tempo,delta)~grupo+
                      logwbc,dist="lognormal",
                      data=dados)
summary(regreparam)
```


Exemplo: Dados de Remissão (Leucemia)

Código R - Distribuição Gama

```
Call:
survreg(formula = Surv(tempo, delta) ~ grupo + logwbc, data = dados,
    dist = "lognormal")
             Value Std. Error     z        p
(Intercept)  4.135      0.422  9.79 1.22e-22
grupo        0.848      0.232  3.66 2.52e-04
logwbc      -0.716      0.123 -5.80 6.55e-09
Log(scale)  -0.437      0.129 -3.39 7.00e-04

Scale= 0.646

Log Normal distribution
Loglik(model)= -93.9   Loglik(intercept only)= -115.4
Chisq= 42.9 on 2 degrees of freedom, p= 4.8e-10
Number of Newton-Raphson Iterations: 6
n= 42
```


Tópicos Abordados Nesta Revisão

- Objeto de estudo da Análise de Sobrevivência
- Caracterização do Tempo de Sobrevivência
- Censuras
- Estimação não paramétrica da função de sobrevivência
- Comparação de grupos - Não Paramétrica
- Modelagem paramétrica
- Modelos de regressão paramétricos