Análise de Sobrevida

Silvia Emiko Shimakura
Marilia Sá Carvalho
Valeska Andreozzi

Bibliografia

- Kleinbaum, D. & Klein, M. *Survival analysis: a self-learning text*. Springer, 1997.
- Therneau, T.M.& Grambsch, P.M. *Modelling survival data:extending the Cox model*. Springer, 2000.
- Carvalho, M.S.&Andreozzi, V.L.& Codeço, C.T.& Barbosa, M.T.S.& Shimakura, S.E. *Análise de sobrevida: Teoria e aplicações em saúde*. Editora Fiocruz, 2005.

Variável resposta de interesse

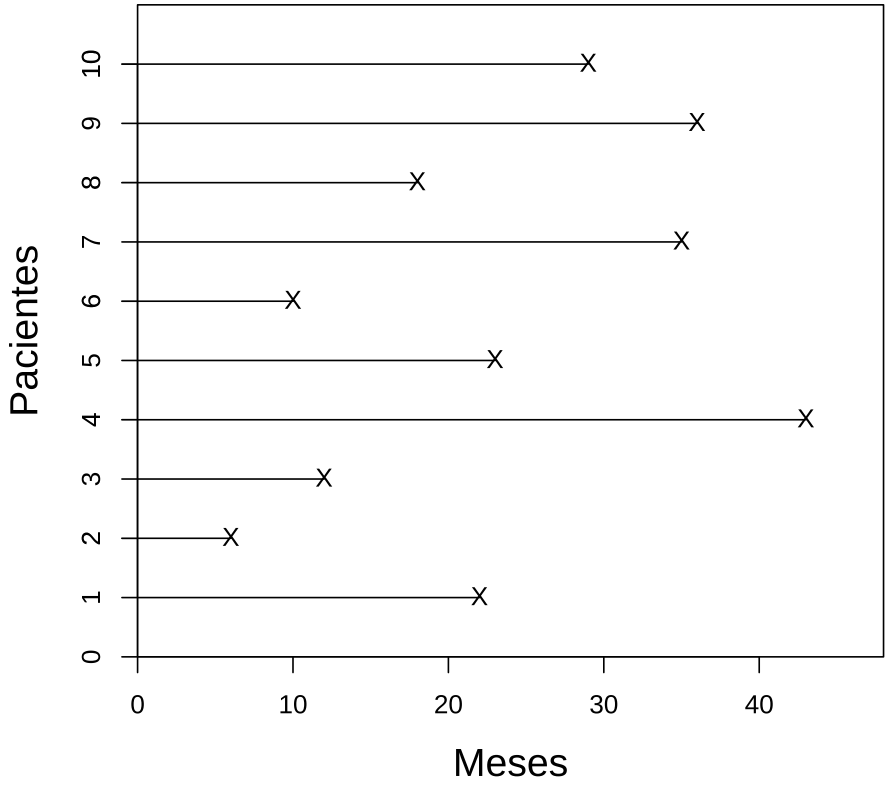
Tempo até...

Variável resposta de interesse

Tempo até...

- óbito
- transplante
- doença
- cura

Representando o tempo



Cada linha representa a trajetória de um paciente e o símbolo **X** indica a ocorrência do evento ou falha.

O que torna a análise de sobrevida distinta?

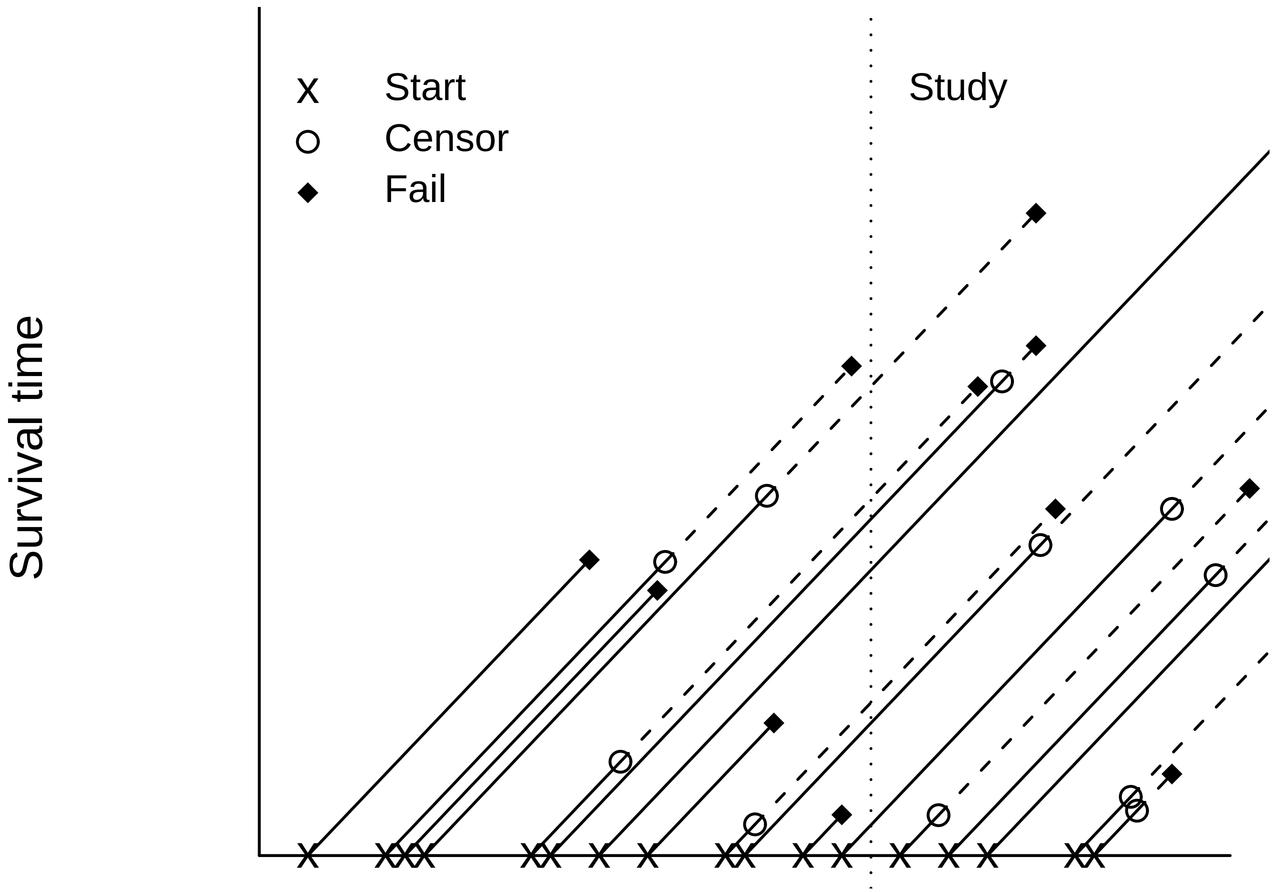
1. Respostas não negativas
2. Dados incompletos

Informação incompleta

- óbito por outras causas – morte do paciente por causas externas;
- término do estudo;
- perda de contato – mudança de residência;
- recusa em continuar participando;
- mudança de procedimento;
- abandono devido a efeitos adversos de tratamento (!!!);
- desconhecimento da data de início – em pacientes HIV+ com data de infecção desconhecida;
- dados truncados – prevalentes.

Censura e **Truncamento**

Diagrama de Lexis



Mecanismos de censura

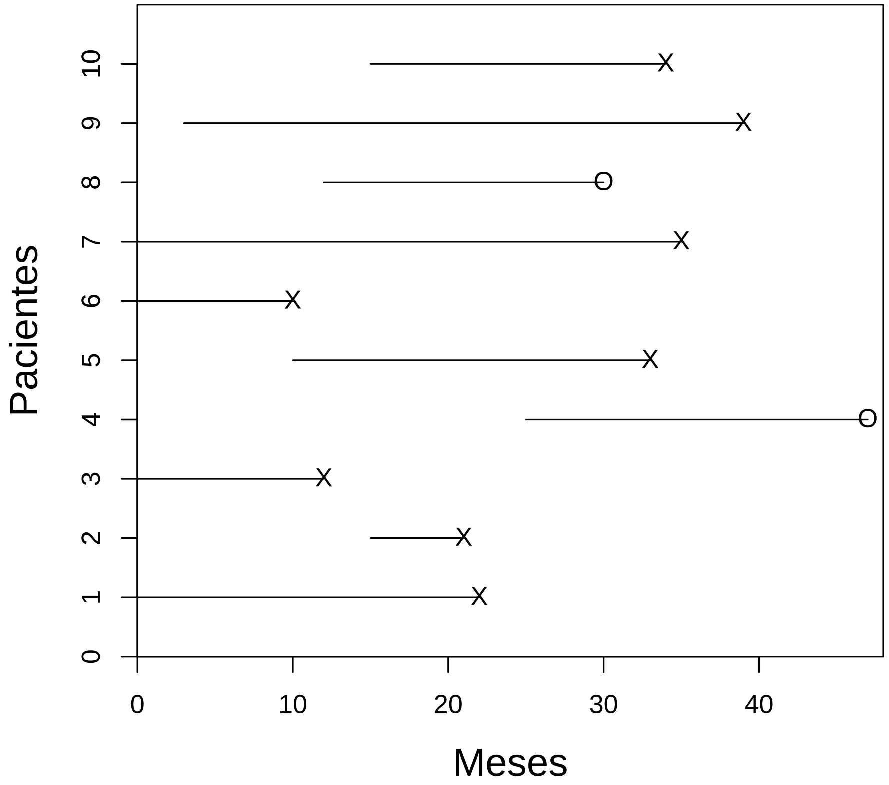
- Censura à direita: evento de interesse não é observado até o término do estudo.
- Censura à esquerda: evento de interesse já ocorreu antes do sujeito ser observado no estudo.
- Censura intervalar: sabe-se somente que o evento de interesse ocorreu dentro de algum intervalo.

Truncamento

- À esquerda: indivíduo que sofre desfecho antes do início do estudo é desconhecido.

- À direita: só são incluídos indivíduos que sofreram o evento.

Coorte aberta com censura à direita



Trajetórias individuais de pacientes com censura e com diferentes tempos de entrada em observação.

Registro do tempo

Tempo de observação de pacientes de uma coorte aberta.

Paciente	Tempo* inicial (I)	Tempo* final (F)	Tempo* T (final - inicial)	Censura (C)
1	0	22	22	1
2	15	21	6	1
3	0	12	12	1
4	25	47	22	0
5	10	33	23	1
6	0	10	10	1
7	0	35	35	1
8	12	30	18	0
9	3	39	36	1
10	15	34	19	1

*Tempo calendário em meses

Processo de contagem

O par (T_i, C_i) é substituído por $(N_i(t), Y_i(t))$, onde:

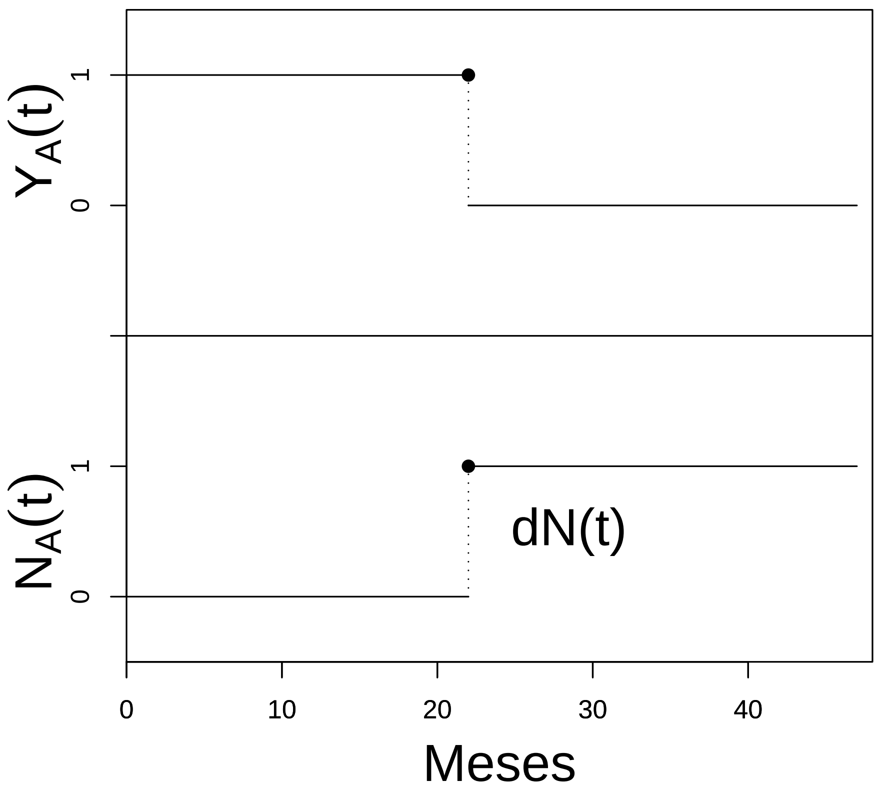
- $N_i(t)$ é o número de eventos observados em $[0, t]$
- $Y_i(t) = 1$, se o indivíduo i está sob observação e em risco no instante t
- $Y_i(t) = 0$, se o indivíduo i não está em risco.

Processo de contagem

Formalmente:

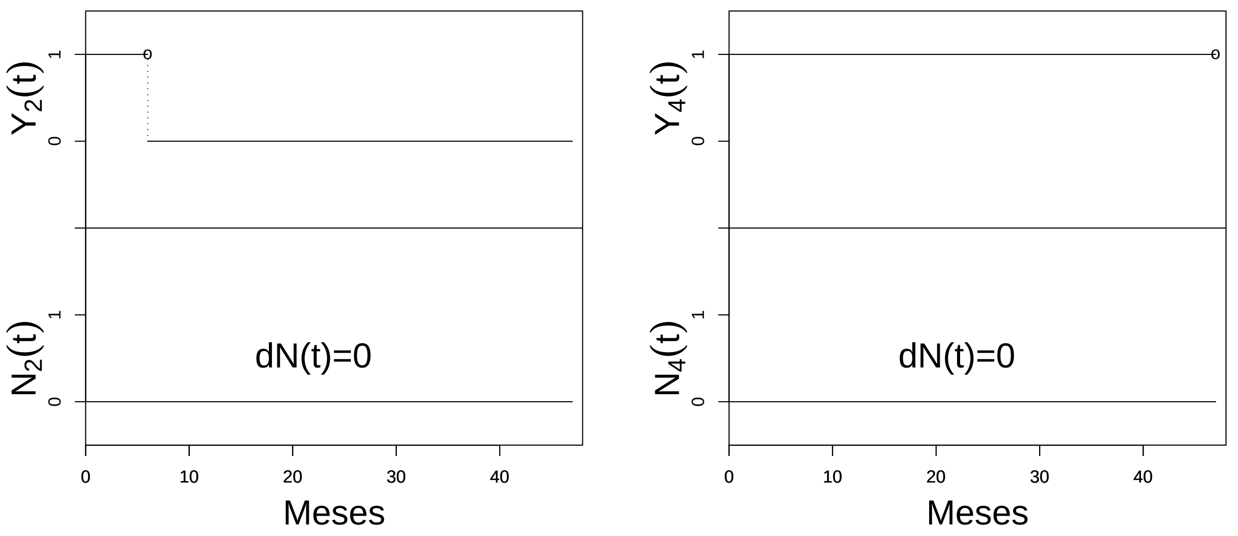
- um processo de contagem é um processo estocástico $N(t)$ com $t > 0$, de tal forma que $N(0) = 0$ e $N(t) < \infty$;
- a trajetória de $N(t)$ é contínua à direita a partir de uma função escada com saltos de tamanho igual a um;
- a análise de sobrevida pode ser pensada como um processo de contagem onde $N(t)$ é o número de eventos observados até o tempo t e $dN_i(t)$ é a diferença entre a contagem de eventos até o instante t e a contagem no momento imediatamente anterior a t.

Graficamente



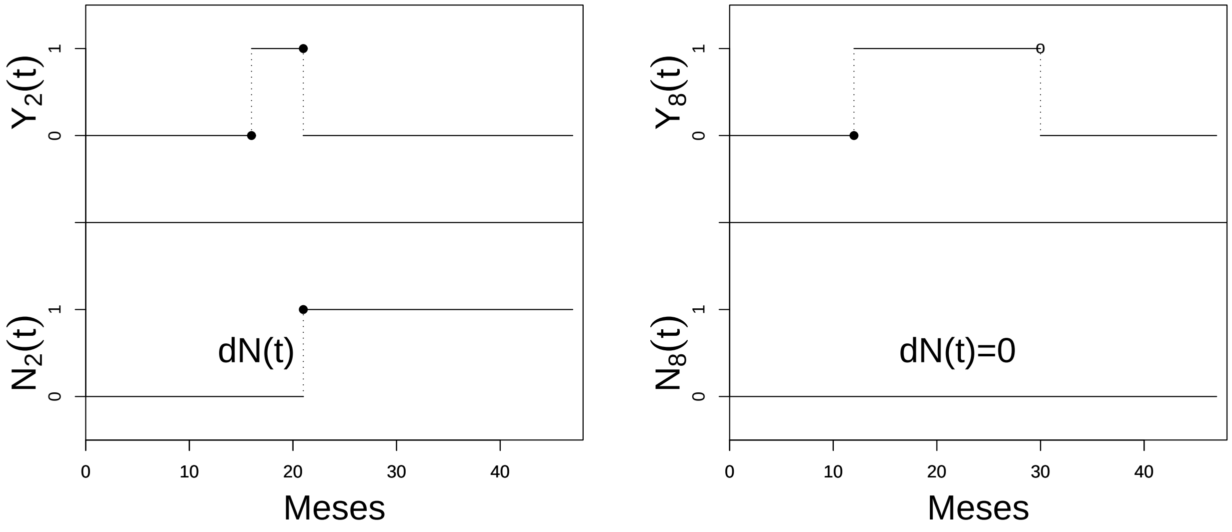
Paciente A: Diagnosticado no mês zero, acompanhado até o mês 22. A ocorrência do evento é assinalada pelo sinal ●

Graficamente



Trajetória de dois pacientes censurados. No primeiro quadro ocorre censura aos 6 meses; no segundo ocorre censura ao término do estudo.

Graficamente



Trajetória de dois pacientes censurados que entraram na coorte ao longo do estudo.

Qual o ganho?

- Mudança no valor de covariável
- Evento múltiplos
- Dados prevalentes

Organização dos dados

id	tempo (T)	censura (C)	sexo	idade
1	30	0	F	54
2	14	1	F	34
3	23	1	M	65
4	11	1	F	45
5	12	0	M	44

Organização dos dados

id	inicio (I)	fim (F)	censura (C)	sexo	idade
1	0	30	0	F	54
2	5	19	1	F	34
3	3	26	1	M	65
...	0	11	1	F	45
n	4	16	0	M	44

Tempo de Sobrevida no R

- O R aceita os dois formatos de registro do tempo de sobrevida.
- O comando *Surv()* tem como função combinar, em uma única variável, a informação referente ao tempo de sobrevivência de cada indivíduo e a informação a respeito do status do paciente.
  - Status = 1 (um), se ocorreu o evento
  - Status = 0 (zero) se o tempo foi censurado
- require(survival)
  - Surv(tempo,status)
  - Surv(inicio,fim,status)

EXEMPLO: Banco de dados ipec

- Os dados de uma amostra de 193 indivíduos diagnosticados com AIDS provenientes de uma coorte aberta com 1591 pacientes HIV positivos atendidos entre 1986 e 2000 no Instituto de Pesquisa Clínica Evandro Chagas/Fiocruz.
- O tempo de sobrevivência é definido como o tempo entre o diagnóstico de Aids e o óbito ou censura.
- `ipec<-read.table('ipec.csv',header=TRUE,sep=';')`

EXEMPLO: Dados do ipec no R

Dados referentes aos primeiros 6 indivíduos do banco:

```
> ipec[1:6,]
  id  ini  fim tempo status sexo escola idade risco acompan
1  1 1243 2095   852      1    M      3    34     0       1
2  2 2800 2923   123      1    M      2    38     6       1
3  3 1250 2395  1145      1    M     NA    32     0       1
4  4 1915 4670  2755      0    M     NA    43     6       0
5  5 2653 4770  2117      0    M     NA    40     0       1
6  6    3  332   329      0    M     NA    34     0       1
  tratam doenca propcp
1      1      4      3
2      0      7      4
3      1      3      4
4      1     10      4
5      1      5      4
6      0      7      0
```

EXEMPLO: Dados do ipec no R

- Primeiro dia de estudo (chamado dia 1) foi 20 de outubro de 1987.
- O tempo de entrada e saída de cada indivíduo foi contado como número de dias a partir do dia 1.
- Todas as censuras são classificadas como censuras à direita.
- Fazer uma análise exploratória dos dados no R.

EXEMPLO: Dados do ipec no R

```
> table(ipec$sexo)
  F   M
 49 144

> summary(ipec$idade)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   30.00   35.00   36.55   43.00   68.00

> table(ipec$status)
  0   1
103  90
```

EXEMPLO: Dados do ipec no R

```
> require(survival)
> Surv(ipec$tempo,ipec$status)[1:6]
[1]  852   123  1145  2755+ 2117+  329+

> Surv(ipec$ini,ipec$fim,ipec$status)[1:6]
[1] (1243,2095 ] (2800,2923 ] (1250,2395 ] (1915,4670+]
[5] (2653,4770+] (   3, 332+]
```

Funções de sobrevida

Funções de sobrevida

Seja T uma variável aleatória contínua e positiva.

- Distribuição de probabilidade e densidade de probabilidade

$$F(t) = P(T \leq t) \qquad\qquad f(t) = \frac{\partial F(t)}{\partial t}$$

- Sobrevida

$$S(t) = P(T > t) = 1 - F(t)$$

Funções de sobrevida

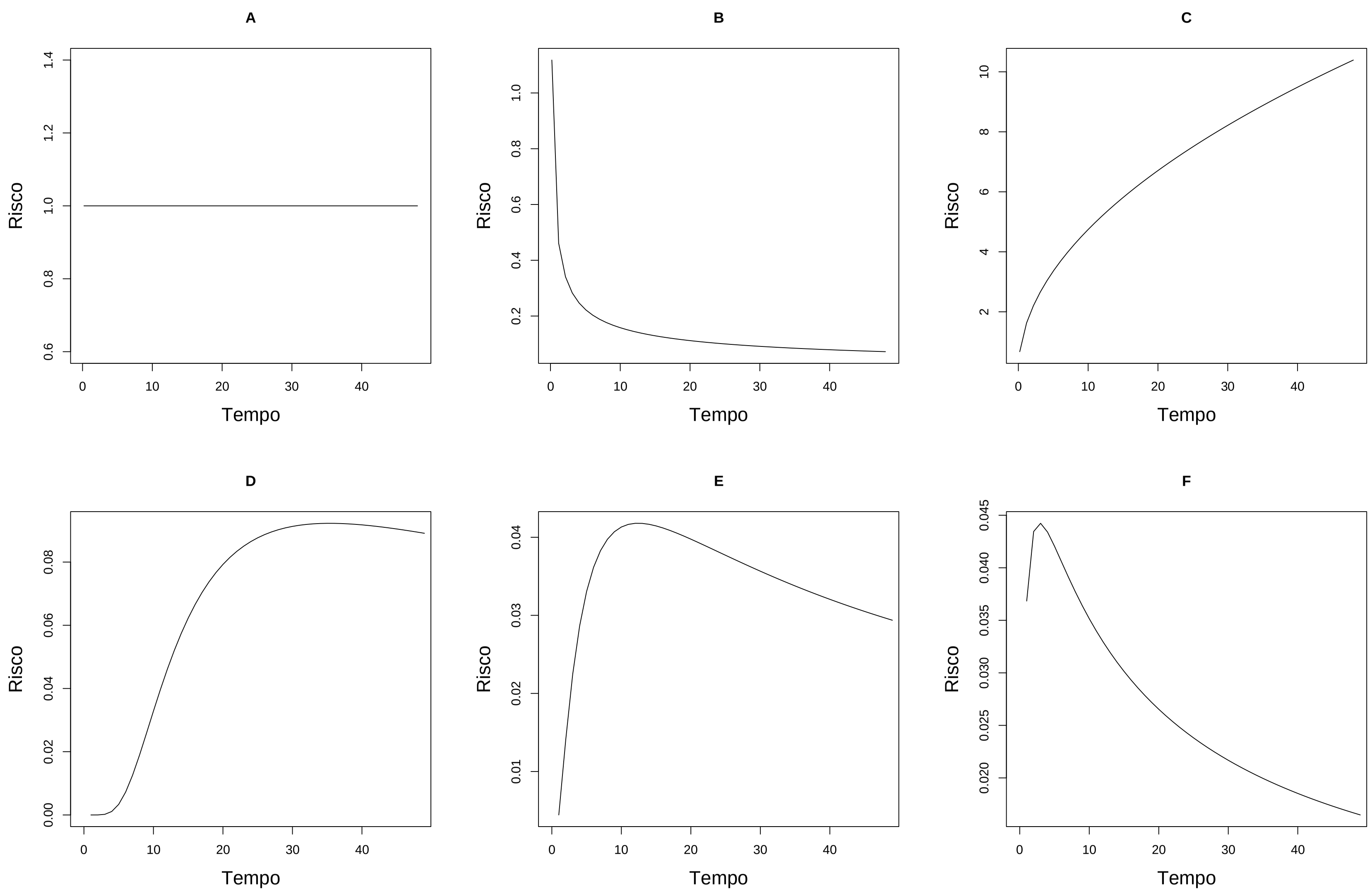
- Risco $\lambda(t)$ –> probabilidade instantânea de um indivíduo sofrer o evento em um intervalo de tempo t e $t+\epsilon$ dado que ele sobreviveu até o tempo t.

$$\lambda(t) = \lim_{\epsilon \longrightarrow \infty} \frac{Pr((t < T < t + \epsilon)|T > t)}{\epsilon}$$

- Risco Acumulado

$$\Lambda(t) = \int_0^t \lambda(u)du$$

Comportamento do Risco



Função de risco com diversos formatos.

Relação entre as funções básicas de sobrevida

$$\boxed{\begin{aligned} S(t) &= 1 - F(t) \\ \lambda(t) &= -\frac{d\ln(S(t))}{dt} \\ \lambda(t) &= \frac{f(t)}{S(t)} \\ \lambda(t) &= \frac{f(t)}{1 - F(t)} \\ \Lambda(t) &= -\ln(S(t)) \end{aligned}}$$

Estimação Não-Paramétrica

Estimação Não-Paramétrica

- estimadores de sobrevida e risco
- Kaplan-Meier e Nelson Aalen
- intervalos de confiança
- Kaplan-Meier estratificado
- testes de Log-Rank e Peto

Incorporando a censura
Sem suposições sobre a distribuição do tempo

Kaplan-Meier

- Seja $t_1 < t_2 < \cdots < t_m$ os tempos onde ocorreram os eventos;
- $Y_i(t) = 1$ se a pessoa i está em risco no tempo t e 0 caso contrário.
- $R(t_i)$ é o total de pessoas a risco no tempo t_i.
- A cada tempo t_i em que houver um evento, a probabilidade de sobrevivência será o número dos que sobreviveram até aquele tempo ($R(t_i) - N(t_i)$) sobre os que estavam em risco naquele tempo ($R(t_i)$).

$$\hat{S}_{KM}(t) = \prod_{t_i \leq t} \frac{R(t_i) - N(t_i)}{R(t_i)}$$

Da sobrevida ao risco

$$\hat{\Lambda}_{KM}(t) = -\ln \hat{S}_{KM}(t)$$

Logo.... pode-se estimar qualquer das funções.

Estimador de Nelson-Aalen

$$\hat{\Lambda}_{NA}(t) = \sum_{t_i \leq t} \frac{N(t_i)}{R(t_i)}$$

Melhor para amostras muito pequenas

Intervalos de confiança

Variância do estimador Kaplan-Meier para a sobrevida
Estimador de Greenwood

$$Var(\hat{S}_{KM}(t)) = (\hat{S}_{KM}(t))^2 \sum_{t_i \leq t} \frac{N(t_i)}{R(t_i)(R(t_i) - N(t_i))}$$

Intervalos de confiança

Assumindo erro α, o intervalo fica assim:

$$\left[\hat{S}_{KM}(t) - z_{\alpha/2}\sqrt{Var(\hat{S}_{KM}(t))}; \hat{S}_{KM}(t) + z_{\alpha/2}\sqrt{Var(\hat{S}_{KM}(t))}\right]$$

Entretanto, este intervalo permite valores negativos e maiores do que 1, o que é incompatível com distribuição de probabilidade.

Intervalos de confiança

Construindo intervalo simétrico para o risco – $\ln \Lambda(t) = \ln(-\ln S(t))$ – pode-se obter um intervalo assimétrico para $S(t)$, porém sempre positivo e menor do que 1.

$$[l_i; l_s] = \left[\ln(\hat{\Lambda}_{KM}(t)) - z_{\alpha/2} dp; \ln(\hat{\Lambda}_{KM}(t)) + z_{\alpha/2} dp\right]$$

onde dp é o desvio padrão e dado por:

$$dp = \sqrt{\frac{\sum_{t_i \leq t} \frac{N(t_i)}{R(t_i)(R(t_i)-N(t_i))}}{\left\{\sum_{t_i \leq t} \ln\left[\frac{R(t_i)-N(t_i)}{N(t_i)}\right]\right\}^2}}$$

no R

- Criando o objeto sobrevida (tempo, censura):

```
> Surv(tempo,status)
#variável status=1 indica evento, 0 censura
```

```
16 18 21+ 21 22 25+ 29 35 37 39 40 50+ 52 54 60 80+ 80 81+ 83 84 85+
```

- Kaplan-Meier

```
> KM <- survfit(Surv(tempo,status), data = ipec90)
> summary(KM)
> plot(KM)
```

- Nelson-Aalen

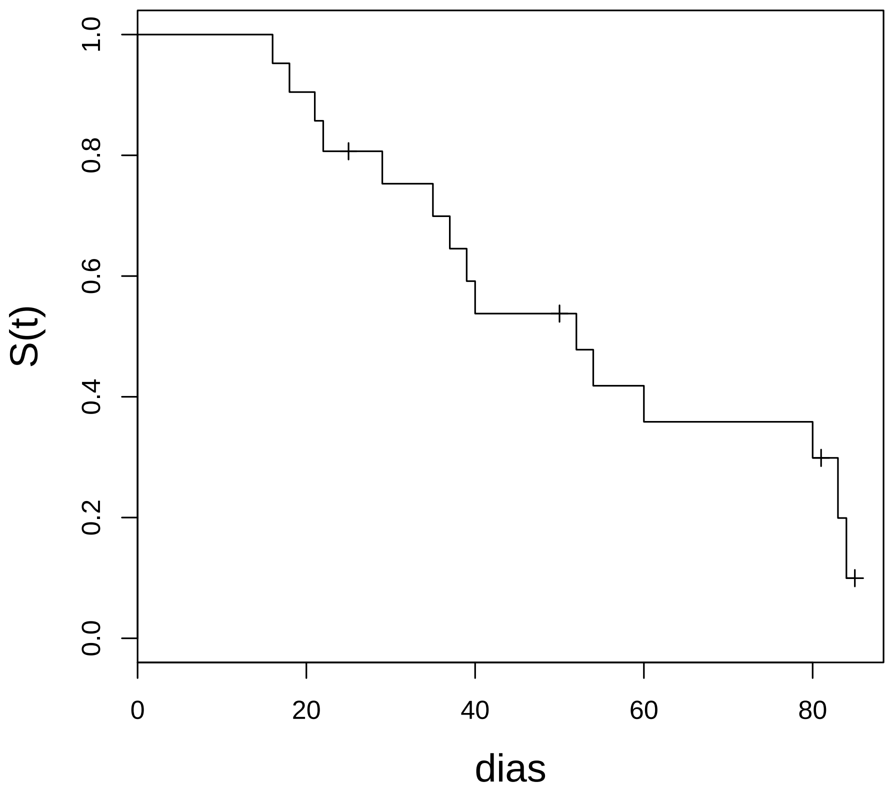
```
> sob.NA <- survfit(coxph(y~1, data = ipec90))
> sob.NA
> summary(sob.NA)
```

Saídas do R – summary(KM)

time	n.risk	n.event	survival	std.err	lower95%CI	upper95%CI
16	21	1	0.9524	0.0465	0.8655	1.000
18	20	1	0.9048	0.0641	0.7875	1.000
21	19	1	0.8571	0.0764	0.7198	1.000
22	17	1	0.8067	0.0869	0.6531	0.996
29	15	1	0.7529	0.0963	0.5859	0.968
35	14	1	0.6992	0.1034	0.5232	0.934
37	13	1	0.6454	0.1085	0.4642	0.897
39	12	1	0.5916	0.1120	0.4082	0.857
40	11	1	0.5378	0.1140	0.3550	0.815
52	9	1	0.4781	0.1160	0.2972	0.769
54	8	1	0.4183	0.1158	0.2431	0.720
60	7	1	0.3585	0.1137	0.1926	0.667
80	6	1	0.2988	0.1093	0.1459	0.612
83	3	1	0.1992	0.1092	0.0680	0.583
84	2	1	0.0996	0.0891	0.0172	0.575

Saídas do R – plot(KM)

Função de sobrevida dos pacientes com aids, utilizando o estimador produto Kaplan-Meier.
Os símbolos $+$ localizam as censuras.



Kaplan-Meier estratificado

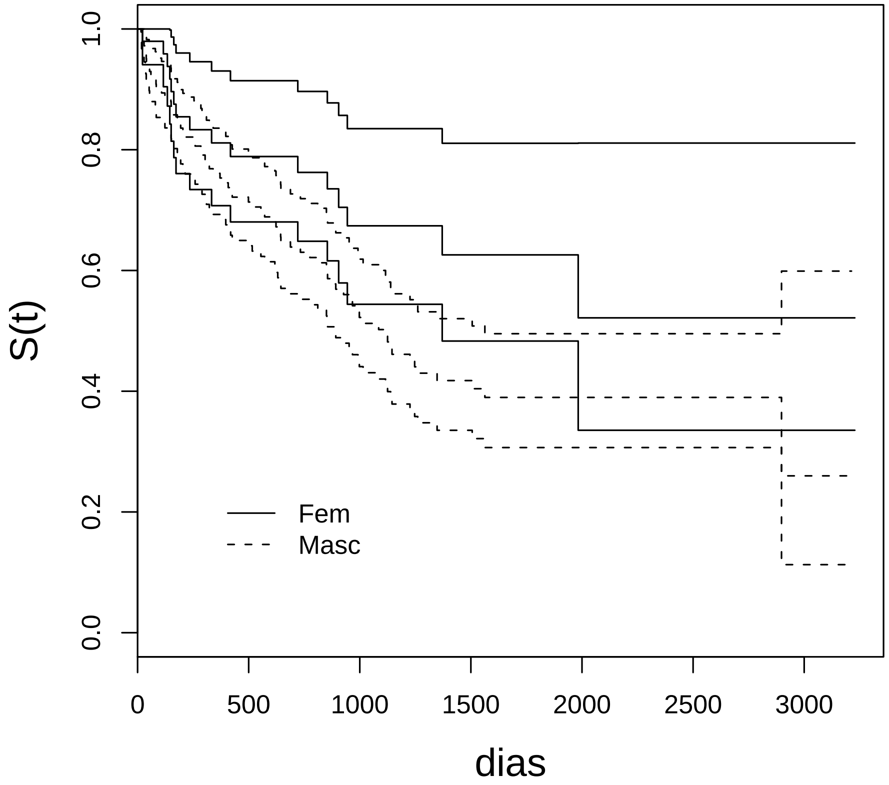
- A sobrevivência é estimada separadamente para cada estrato, utilizando Kaplan-Meier.
- no R

```
> survaids <- survfit(Surv(tempo,status)~ sexo, data = ipec)
> survaids

Call: survfit(formula = resp ~ sexo, data = ipec)

          n events rmean se(rmean) median 0.95LCL 0.95UCL
sexo=F   49     16   2096       229    Inf    1371     Inf
sexo=M  144     74   1581       122   1116     887    1563
```

Gráfico sobrevida estratificada



Curvas de sobrevida de pacientes com aids, estratificado por sexo. Estimação por Kaplan-Meier, com intervalo de confiança de 95%.

Testes

Hipótese nula: não há diferença entre estratos

$$H_0 : \lambda_1(t) = \lambda_2(t) = \cdots = \lambda_k(t)$$

Log-rank (ou Mantel-Haenszel)

Distribuição esperada de eventos igual em todos os estratos:

$$e_k(t) = N(t)\frac{R_k(t)}{R(t)}$$

Estatística de teste log-rank para dois estratos $(k = 2)$:

$$\text{Log-rank} = \frac{(N_1 - E_1)^2}{Var(N_1 - E_1)}$$

com $N_1 =$ ao total de eventos observados no estrato 1 e $E_1 =$ ao total de eventos esperados no estrato 1.

Teste log-rank

A variância, que entra no cálculo como um fator de padronização, tem a fórmula (para $k = 2$):

$$Var(N_1 - E_1) = v_i$$

em que

$$v_i = \sum_{t_i} \frac{R_1(t_i)[R(t_i) - R_1(t_i)]N(t_i)[R(t_i) - N(t_i)]}{R(t_i)^2[R(t_i) - 1]}.$$

A estatística log-rank, sob a hipótese nula, segue uma distribuição χ^2 , com $k - 1$ graus de liberdade.

Teste de Peto

Dá maior peso às diferenças (ou semelhanças), no início da curva, onde se concentra a maior parte dos dados e por isso é mais informativa. Usa um ponderador $S(t)$ no estimador.

$$\text{Peto} = \frac{(N_1 - E_1)^2}{Var(N_1 - E_1)}$$

sendo que

$$N_1 - E_1 = \sum S(t_i)(N_1(t_i) - E_1(t_i))$$

$$Var(N_1 - E_1) = \sum (S(t_i))^2 v_i$$

Também a estatística Peto segue aproximadamente uma distribuição χ^2 com $k - 1$ graus de liberdade.

no R

```
> survdiff(Surv(tempo,status)~sexo, data=ipec,rho=0)

Call:
survdiff(formula = Surv(tempo, status) ~ sexo, data = ipec, rho = 0)

           N Observed Expected (O-E)^2/E (O-E)^2/V
sexo=F    49       16     24.5      2.93      4.03
sexo=M   144       74     65.5      1.09      4.03

 Chisq= 4  on 1 degrees of freedom, p= 0.0447
```

O argumento `rho` determina o tipo de teste a ser realizado. Para log-rank, use `rho = 0` (*default*). Para o teste Peto, use `rho = 1`.

no R

```
> survdiff(Surv(tempo,status)~sexo, data=ipec,rho=1)

Call:
survdiff(formula = Surv(tempo, status) ~ sexo, data = ipec, rho = 1)

           N Observed Expected (O-E)^2/E (O-E)^2/V
sexo=F  49     12.1     18.2     2.011      3.54
sexo=M 144     55.1     49.0     0.746      3.54

 Chisq= 3.5  on 1 degrees of freedom, p= 0.0598
```

Modelagem Paramétrica

Distribuições Paramétricas

- Distribuições estatísticas para modelar as funções de sobrevida:
  - Exponencial
  - Weibull
  - Lognormal
- verossimilhança para dados censurados
- modelo paramétrico no R

Distribuição Exponencial

Se a variável T possui uma distribuição exponencial,

- Densidade de probabilidade:
  $f(t) = \alpha \exp(-\alpha t), \qquad \alpha > 0$

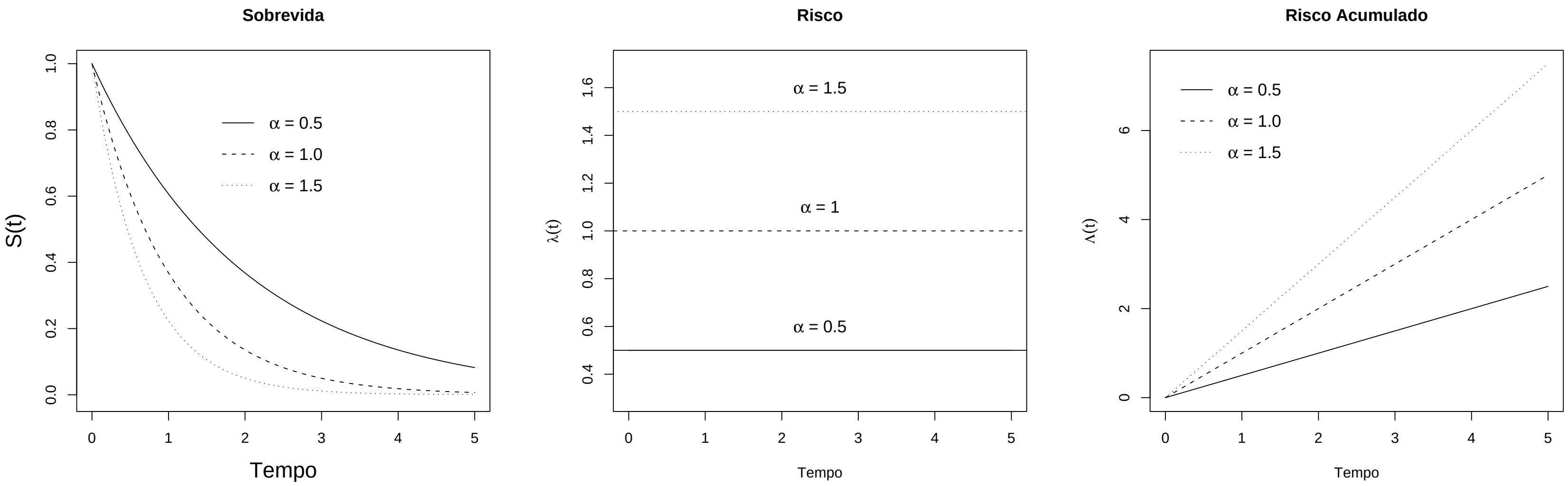
- Função de sobrevida: $S(t) = \exp(-\alpha t)$

- A função risco é constante para todo o tempo de observação t, ou seja: $\lambda(t) = \frac{f(t)}{S(t)} = \alpha = \text{constante}$

- A função de risco acumulado é uma função linear no tempo e é dada por: $\Lambda(t) = -\ln S(t) = \alpha t$

Algumas exponenciais

Função de sobrevida, de risco e de risco acumulado para a distribuição exponencial considerando diferentes valores de α



A distribuição exponencial é conhecida como distribuição exponencial padrão quando $\alpha = 1$.

Interpretando risco exponencial

- média e a variância: $\bar{T} = \frac{1}{\alpha}$ e $var(T) = \frac{1}{\alpha^2}$
- quanto maior o risco, menor o tempo médio de sobrevida e menor a variabilidade deste em torno da média
- como a distribuição do tempo de sobrevida T é assimétrica, usa-se mais o tempo mediano
- o modelo exponencial é matematicamente simples, mas a suposição de risco constante no tempo é pouco plausível
- é aplicável quando o tempo de acompanhamento é curto o suficiente para que o risco naquele período possa ser considerado constante (por ex., o risco de óbito de crianças entre dois e cinco anos, independente da causa, pode ser considerado constante neste intervalo)

Distribuição Weibull

- permite variação do risco no tempo
- é uma generalização da distribuição exponencial, sendo a densidade $f(t) = \gamma\alpha^\gamma t^{\gamma-1}\exp(-(\alpha t)^\gamma)$ e a sobrevida $S(t) = \exp(-(\alpha t)^\gamma)$ ($\alpha > 0$ e $\gamma > 0$)
- o parâmetro γ determina a forma da função de risco sendo chamado de parâmetro de forma:

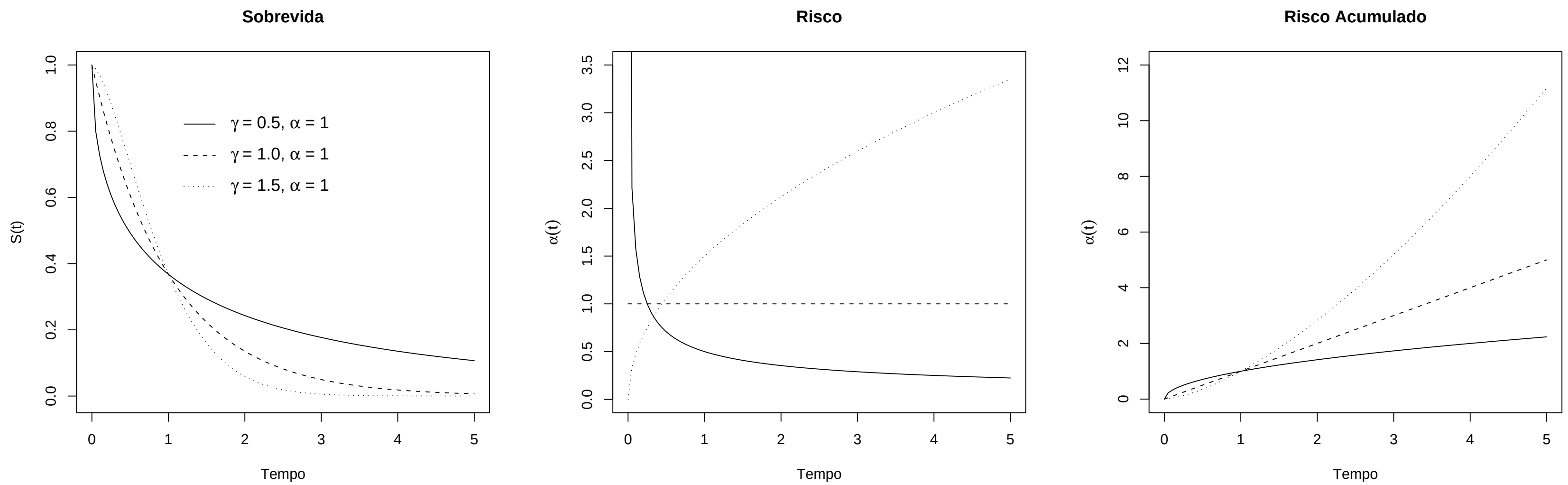
  $\gamma < 1$ função de risco decrescente

  $\gamma > 1$ função de risco crescente

  $\gamma = 1$ função de risco constante (equivalente ao modelo exponencial)
- o parâmetro α determina a escala da distribuição
- a função de risco é: $\Lambda(t) = -\ln S(t) = (\alpha t)^{\gamma-1}$

Algumas Weibull

Função de sobrevida, de risco e de risco acumulado com parâmetro escala $\alpha = 1$ e diferentes valores do parâmetro de forma γ



Estimação

- Para estimar os parâmetros das funções básicas de sobrevida assumindo uma distribuição para a variável T utiliza-se o método da máxima verossimilhança, adaptado para a ocorrência da censura.
- Esta função de verossimilhança supõe que os tempos em que há censura são independentes dos tempos de ocorrência do evento, ou seja, as censuras são não informativas.
- Quando há censura o tempo de sobrevida é maior do que o tempo observado t_+ e $S(t_+) = P(T > t_+)$.
- A função de verossimilhança modificada é: $L = \prod_{i \in F} f(t_i) \prod_{i \in C} S(t_{i+})$.

Regressão Paramétrica

- A inclusão de covariáveis segue a forma utilizada em modelos lineares generalizados, podendo ser contínuas – pressão sanguínea, idade, dosagens bioquímicas – ou categóricas – gênero, tratamento, comportamentos.
- O objetivo é de estimar o efeito de covariáveis, $x_1, x_2, \cdots, x_p$, sobre a variável resposta Y.

Modelo Exponencial

As funções de risco e sobrevida para o modelo exponencial ficam assim:

$$
\begin{aligned}
\lambda(t|\boldsymbol{x}) &= \alpha(\boldsymbol{x}) = \exp(\boldsymbol{x}\boldsymbol{\beta}) \\
S(t|\boldsymbol{x}) &= \exp(-\alpha(\boldsymbol{x})t) = \exp(-\exp(\boldsymbol{x}\boldsymbol{\beta})t)
\end{aligned}
$$

Exemplo

Assumindo que o risco de morrer é constante ao longo do tempo, pode-se estimar o efeito da $idade$ na sobrevida e no risco de 6.805 pacientes em diálise acompanhados durante um ano (1.603 morreram) através do modelo exponencial:

$$\lambda(t|idade) = \exp(\beta_0 + idade \times \beta_1)$$

Os parâmetros estimados são: $\beta_0 = -6,14$ e $\beta_1 = 0,04$, ou seja, para cada ano a mais de vida o risco aumenta de $exp(0,04) = 1,04$.

Pode-se comparar o risco constante de morte no tempo, entre dois indivíduos submetidos à diálise, um com 70 anos e outro com 30, substituindo as estimativas dos parâmetros $\boldsymbol{\beta}$:

$$\frac{\lambda(t|x_1 = 70)}{\lambda(t|x_1 = 30)} = \frac{\exp(\beta_0 + 70\beta_1)}{\exp(\beta_0 + 30\beta_1)} = \frac{0,035}{0,007} = 4,95$$

Modelo Weibull

O tempo T segue uma distribuição de Weibull e o parâmetro de escala α depende das covariáveis.

Neste caso são estimados os parâmetros:

- β_0 – cujo exponencial representa o risco médio, quando todas as covariáveis são zero;
- β_1 – cujo exponencial é a parcela de variação no tempo de sobrevida devido à idade do paciente;
- γ – a forma da função de risco ao longo do tempo.

Avaliação do modelo

- Teste de Wald – testa a hipótese nula H_0 de que o parâmetro β da regressão é igual a zero.

- Razão de Verossimilhança (análise de deviance) – estatística global do ajuste e comparação de modelos

- Comparar um modelo com distribuição exponencial e outro com distribuição Weibull equivale a testar a hipótese nula de que o parâmetro de forma, γ, da distribuição Weibull é igual a 1.

Exemplo

```
> survreg(formula=Surv(tempo,status)~1, data=dialise,
        dist='exponential')

Call:
survreg(formula=Surv(tempo, status)~1, data=dialise,
        dist="exponential")
Coefficients:
(Intercept)
   4.096059
Scale fixed at 1
Loglik(model)= -8169   Loglik(intercept only)= -8169
n= 6805
```

Exemplo

```
> survreg(formula=Surv(tempo,status)~1, data=dialise,
        dist='weibull')

Call:
survreg(formula=Surv(tempo, status)~1, data=dialise,
        dist="weibull")
Coefficients:
(Intercept)
   4.388833
Scale= 1.257539
Loglik(model)= -8104.2   Loglik(intercept only)= -8104.2
n= 6805
```

Exemplo

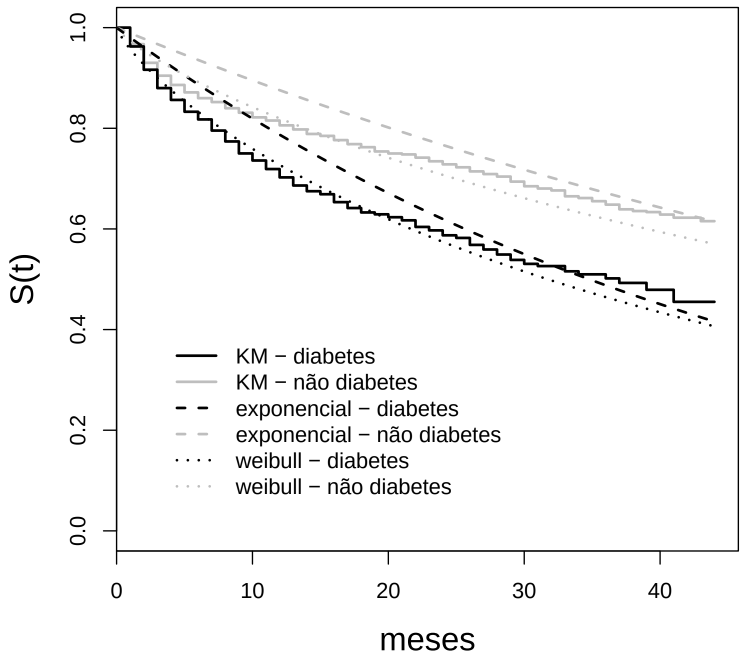
Comparando:
$D = 2(L_{weibull} - L_{exponencial}) = 2(-8104,2 - (-8169)) = 129,6$

Como D segue uma distribuição χ^2 com 1 g.l., $p = 0$, ou seja, rejeitamos a hipótese nula de que $\gamma = 1$.

Isto é, o modelo de Weibull, com $\gamma = 0,795$ é melhor do que o modelo exponencial.

Análise Gráfica

Comparar a curva do Kaplan-Meier com as estimadas parametricamente. Quanto mais próximo o modelo paramétrico estiver da curva do Kaplan-Meier, melhor.



As três curvas em cinza referem-se aos paciente sem diabetes e as três curvas pretas aos pacientes com diabetes.