Notas de Aula Análise de Sobrevivência

Alisson Rosa

Índice

Prefácio **3**
Convenções 3

1 Introdução **4**
1.1 Breve História 4
1.2 Porque estudar análise de sobrevivência? 4
1.3 Censura 5
1.3.1 Tipos de censura 5

2 Definições básicas **6**
2.1 Função de Sobrevivência 6
2.2 Função taxa de falha 8
2.2.1 Relações 8

3 Layout dos dados **10**
3.1 Layout Básico 10

4 Análise Descritiva **12**
4.1 Estimando a função de sobrevivência 12
4.1.1 Sem dados de censura 12

5 Estimador de Kaplan-Meier **14**
5.1 Outra maneira de layout para os dados 14
5.2 Formalizando 15
5.3 Propriedades do estimador 16
5.3.1 Variância do Estimador 16
5.4 Códigos 17

6 Summary **18**

7 Apêndice **19**

Referências **20**

Prefácio

Notas de Aula do curso de análise de sobrevivência, as duas referências principais vão ser

- Colosimo (2006)

Em termos práticos, os códigos vão ser desenvolvidos usando as linguagens Python e R, portanto assume-se conhecimentos básicos de ao menos umas dessas linguagens para um bom aproveitamento.

Encontrou algum erro? Pode encaminhar uma issue no repósitório ou se preferir pode fazer um pull request, **qualquer** contribuição construtiva é bem vinda.

Convenções

> Blocos dessa maneira indicam definições importantes

Textos dessa cor assinalam concepções relevantes

1 Introdução

Análise de sobrevivência é a área que estuda o **tempo** até acontecer um evento de interesse acontecer. Como por exemplo:

- Tempo até os modelos convergirem
- Tempo até o equipamento falhar duas vezes
- Tempo até o customer não frequentar mais o local
- Tempo até um indivíduo casar-se por ano
- Tempo para o desenvolvimento da Covid-19

Note portanto, que o evento pode encapsular mais de um fato como **falhar duas vezes**

1.1 Breve História

Originalmente, análise de sobrevivência era usada exclusivamente para para estudos de mortalidade em registros estatísticos. Sabe-se que as primeiras análises estatísticas de processos de sobrevivência foram desenvolvidas pelo estatístico John Graunt, por um longo a análise de sobrevivência foi considerada um instrumento analítico para estudos de biomedicina e estudos demográficos, mas assim como estatística em geral alterou-se fortemente nas últimas décadas junto (causa?) com avanços computacionais, a análise de sobrevivência não foi diferente, pois atualmente possuímos um grande poder computacional para desenvolvimento de métodos estatísticos antes inviáveis

1.2 Porque estudar análise de sobrevivência?

Ok, porque não usar usar modelos de regressão em geral para modelar o Tempo (T)? Um ponto fundamental aqui, é que nem sempre o evento de interesse acontece, assim gerando o que é conhecido como censura

1.3 Censura

Quando estamos fazendo uma análise de sobrevivência podemos ter indivíduos que por algum motivo tiveram que sair do estudo e não apresentaram o evento de interesse, portanto os dados desses indivíduos são chamados de **censurados**

> Temos informações sobre o tempo de sobrevivência mas não sabemos exatamente **quando** acontece o evento de interesse.

1.3.1 Tipos de censura

Aqui vamos definir alguns tipos de censuras

Censura Tipo I : O estudo será finalizado após um período pré estabelecido de tempo. Censura Tipo II O estudo será finalizado após ter ocorrido o evento de interesse em um número pré-estabelecido de indivíduos.
Censura Aleatória : Acontece quando o indivíduo é removido do estudo sem ter ocorrido a falha.

2 Definições básicas

Seja $f(t)$ a função densidade de probabilidade da variável aleatória T (tempo até o evento ocorrer). Definimos a função de distribuição acumulada da variável aleatória T como sendo:

$$F(t) = P(T < t) = \int_0^t f(u)du$$

E por consequência $f(t)$ fica definida como:

$$f(t) = \lim_{\Delta t \to 0^+} \frac{P(t < T \leq t + \Delta t)}{\Delta t} = \frac{dF(t)}{dt}$$

2.1 Função de Sobrevivência

A função de sobrevivência é o complemento da função de distribuição acumulada, isso é:
A probabilidade de uma observação não falhar até um tempo t:

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

Pela Figura 2.1 podemos ter um vislumbre que $S(0) = 1$ e $S(\infty) = 0$.
A demonstração de tais fatos fica a cargo do leitor, basta notar que $S(t) = 1 - F(t)$ e usar as propriedades da F.

Por consequência a probabilidade de não sobreviver até um certo tempo t é:

$$1 - S(t)$$

Assim a probabilidade de não sobreviver em um intervalo (t_1, t_2) é dado por:

$$1 - S(t_2) - (1 - S(t_1)) = S(t_1) - S(t_2)$$

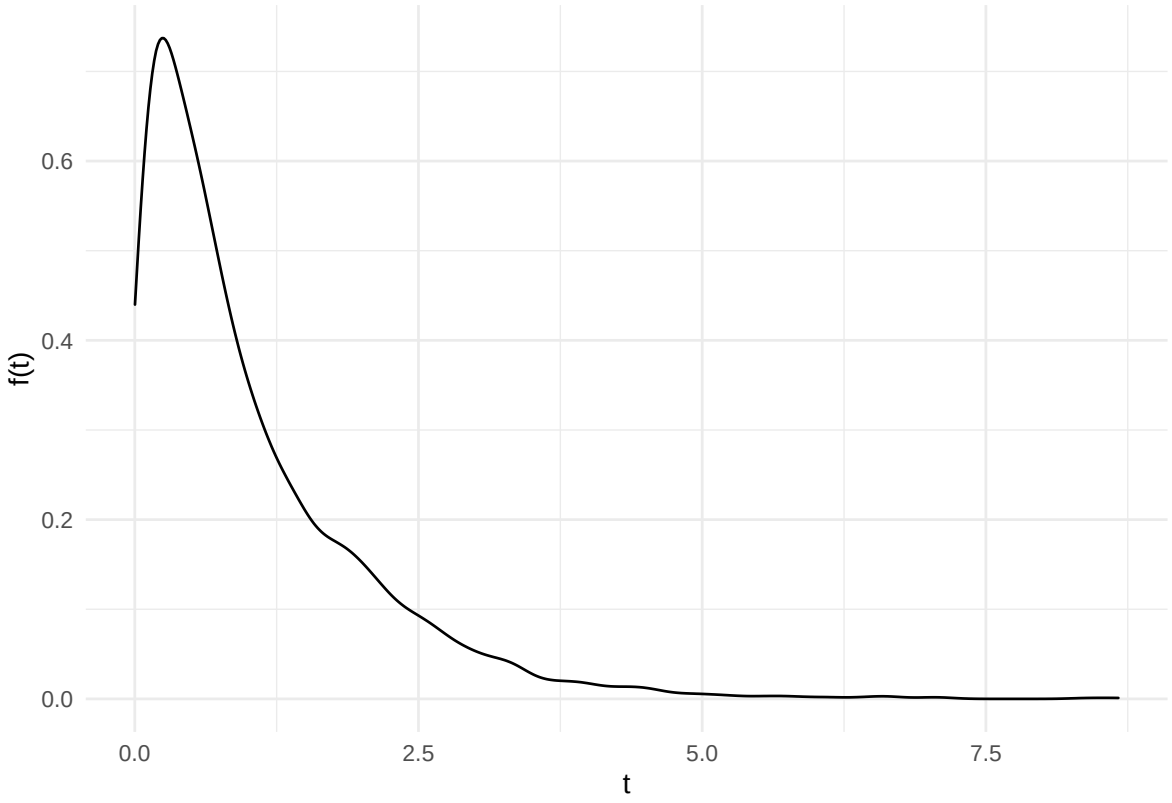

Figura 2.1: Exemplo de uma função de densidade

2.2 Função taxa de falha

Fornece o potencial (taxa de falha) instantâneo do evento **ocorrer**, dado que o indivíduo sobreviveu até o tempo t:

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

Assim estamos interessados em saber qual a probabilidade dele não sobreviver o tempo adicional Δ_t.
Ou seja, temos interesse que o evento ocorra, em termos de interpretação é o **oposto** da função de sobrevivência.

> **Nota**
>
> A taxa de falha **não** é uma probabilidade! Pois quando dividimos por Δ_t obtemos uma probabilidade por unidade de tempo, ou seja a imagem de λ é positiva mas não limitada.

Pelas relações a seguir vemos que a função taxa de falha determina completamente a distribuição de t.

2.2.1 Relações

- $\lambda(t) \geq 0$
  Demonstração
  É trivial pois por definição medidas de probabilidade $\in [0, 1]$ e $\Delta_t \geq 0$ portanto um produto de números positivos
- $\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\left(\log(S(t))\right)$
  Demonstração
  Sabemos que $S(t) = P(T > t)$, substituindo, obtemos:

  $$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{\frac{dF(t)}{dt}}{1 - F(t)} = -\frac{-\frac{dF(t)}{dt}}{1 - F(t)} = -\frac{d}{d_t}\log(S(t))$$

  Multiplica-se e divide-se a expressão por -1, e utiliza-se que a derivada é um operador linear.
- $\text{vmr(t)} = \frac{\int_t^\infty (u - t) f(u) du}{S(t)} = \frac{\int_t^\infty S(u) du}{S(t)}$
  Demonstração

Vamos utilizar integral por partes, usando a fórmula definida no Apêndice. Tomando $f = u - t$ e $g' = f(u)$, obtemos:

$$= \lim_{b \to \infty} (u - t)S(u)]_t^b - \int_t^\infty -S(u)du = \int_t^\infty S(u)du$$

Onde o resultado final é encontrado usando $f(u)du = -\frac{d}{du}S(u)$ e sabendo que $S(t) \to 0$, quando $t \to \infty$

3 Layout dos dados

Nessa seção vamos inserir os alguns dados em duas linguagens de programação, a saber: Python e R

3.1 Layout Básico

Os dados de sobrevivência possuem um layout estabelecido, que é dado a seguir:

Id	T	s_i	X_i	...	X_p
1	t_1	s_1	x_{1i}	...	x_{1p}
2	t_2	s_2	x_{2i}	...	x_{2p}
.	.	.	.	.	.
.	.	.	.	.	.
n	t_n	s_n	x_{ni}	...	x_{np}

R

```
tempo <- c(1, 2, 3, 3, 3, 5, 5, 16, 16, 16, 16, 16, 16, 6, 16, 1, 1, 1, 1, 4, 5, 7, 8, 10,
censura <- c(0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0,
grupo <- c(rep(1, 15), rep(0, 14))
dados <- tempo  |>
            cbind(censura)  |>
            cbind(grupo)
dados  |>
        head()
```

```
     tempo censura grupo
[1,]     1       0     1
[2,]     2       0     1
[3,]     3       1     1
[4,]     3       1     1
[5,]     3       0     1
[6,]     5       0     1
```

Python

```
import pandas as pd
import numpy as np

tempo = [1, 2, 3, 3, 3, 5, 5, 16, 16, 16, 16, 16, 16, 16, 16, 1, 1, 1, 1, 4, 5, 7, 8, 10,
censura = [0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0
grupo = np.concatenate((np.repeat(1, 15), np.repeat(0, 14)))

dados = pd.DataFrame({'tempo':tempo,'censura':censura,'grupo':grupo})
dados.head()
```

```
   tempo  censura  grupo
0      1        0      1
1      2        0      1
2      3        1      1
3      3        1      1
4      3        0      1
```

4 Análise Descritiva

Em Estatística é bastante usual fazer análise descritiva dos dados, como medidas resumo, gráficos e tabelas. Aqui também iremos elaborar análise descritivas, porém com foco nas medidas de sobrevivência e risco.

> **Nota**
>
> O símbolo # aqui é utilizado para indicar a cardinalidade (quantidade de elementos) de um conjunto.
> Por exemplo o conjunto $A = \{a, b, d\}$, possui cardinalidade 3, isso é $\#\text{A} = 3$

4.1 Estimando a função de sobrevivência

Vamos nessa subseção estudar algumas maneiras de estimar a função de sobrevivência $S(t)$

4.1.1 Sem dados de censura

Uma maneira bastante intuitiva para estimarmos $S(t)$ é tomarmos a quantidade de indíviduos que não falharam até o tempo t dividindo pelo total de indíviduos no estudo

$$\hat{S}(t) = \frac{\#\text{Observações que não falharam até t}}{\#\text{Observações}}$$

Na prática, $S(t)$ será obtida por algum estimador, e portanto seu gráfico terá um formato de escada, como dado a seguir:

Pelo Figura 4.1 nota-se que a função mantém-se constante em alguns intervalos e tem um decaimento em alguns pontos específicos, a proxima seção formaliza tal ideia.

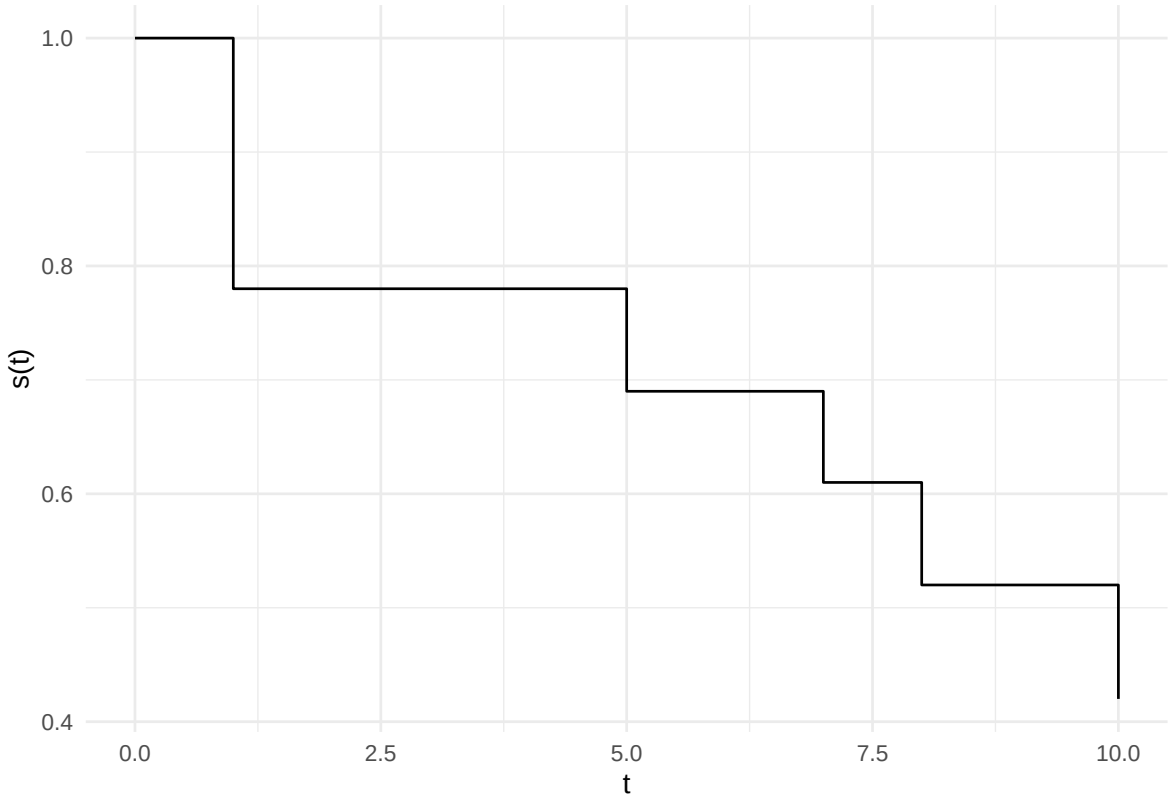

Figura 4.1: Exemplo de uma função de sobrevivência na prática

5 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier também denominado limite-produto é uma adaptação da ideia 'ingênua' que utilizamos na seção anterior. Ele fornece uma maneira simples, mas eficiente de estimar a função de sobrevivência. De forma intuitiva, dividimos o tempo t em uma série de intervalos de acordo com os eventos observados ou dados censurados, após isso calculamos uma sequência de produto de probabilidade condicionais

5.1 Outra maneira de layout para os dados

t ordenados	int	# de falhas	n_j	m_j
0	$[0, t_1)$	0	k	m_1
t_1	$[t_1, t_2)$	d_2	$k-(d_1+m_1)$	m_2
t_2	$[t_2, t_3)$.	.	.	.
.	.	.	.	.
.	.	.	.	.
k	$[t_k, t_{k+\epsilon})$	d_k	$k-\sum_i(d_i+m_i)$	m_n

- **int**: São os intervalos
- **# de falhas**: É o número de falhas naquele intervalo
- n_j : É a quantidade de observações que ainda não falharam ou foram censuradas naquele intervalo (as vezes chamado de indivíduos sob risco)
- c_j: Quantidade de censuras naquele intervalo.

Assim fica fácil ver que no tempo 0 temos 0 falhas, porque como vamos ver a seguir a construção dos intervalos começa a partir do primeiro tempo que acontece um evento, e como temos 0 falhas temos portanto todos (k) os indivíduos sem o evento de interesse

5.2 Formalizando

Sabemos que $S(t) = P(T > t)$, vamos supor que já construímos a tabela e possuímos o tempo 3 e 1, assim queremos calcular, por exemplo:

$$S(3) = P(T > 3)$$

Podemos fazer a seguinte manipulação:

$$S(3) = P(T > 3) = P(T > 1, T > 3) = P(T > 1)P(T > 3|T > 1)$$

> **Nota**
>
> Lembre que $f(X|Y) = \dfrac{f(X,Y)}{f(Y)}$
> E que se A é subconjunto de B, então A $\cap$ B = A (relacione aos intervalos $(1, \infty)$ e $(3, \infty)$

Vamos utilizar o seguinte exemplo para ilustrar a ideia anterior:

t ordenados	int	# de falhas (d_j)	n_j	$\hat{S}(.)$
0	$[0, 1)$	0	14	1
1	$[1, 5)$	3	14	0.78
5	$[5, 7)$	1	9	.
7	$[7, 8)$	1	8	.
8	$[8, 10)$	1	7	.
10	$[10, 16)$	1	6	x_{np}

Para calcular $\hat{S}(1)$, fazemos então:

$$P(T > 1) = P(T > 0, T > 1) = P(T > 0)P(T > 1|T > 0)$$

Sabemos que $P(T > 0) = 1$ como comentado anteriormente. Temos 3 acontecimentos do evento de interesse no tempo $t = 1$ e 14 observações restantes, assim a probabilidade de 'falhar' nesse intervalo é $\dfrac{3}{14}$, portanto a probabilidade de sobreviver é $1 - \dfrac{3}{14}$, substituindo as informações tem-se portanto que $\hat{S}(1) = 0.786$

Para $\hat{S}(5)$, fazemos a mesma decomposição de probabilidade e chegamos em:

$$\hat{S}(5) = P(T > 1)P(T > 5|T > 1)$$

Como calculado anteriormente $P(T > 1) = 1 - \dfrac{3}{14}$ e $P(T > 5|T > 1) = 1 - \dfrac{1}{9}$.

Fica claro então a relação de recursão, pois para o cálculo da estimativa utiliza-se todas as probabilidades calculadas anteriormente, generalizando temos:

$$S(t_j) = (1-q_1)(1-q_2)...(1-q_j) = \prod_{j:t_j<t}\left(1-\frac{d_j}{n_j}\right) = \prod_{j:t_j<t}\left(\frac{n_j-d_j}{n_j}\right)$$

Onde q_j é a probabilidade de uma observação ter o evento de interesse no intervalo $[t_{j-1}, t_j)$ sabendo-se que não teve em t_{j-1}, formalizando tem-se:

$$q_j = P(T \in [t_{j-1}, t_j)]|T > t_{j-1})$$

Assim o estimador reduz-se a estimar os q_j, reescrevendo usando alguns termos já citados temos:

$$\hat{q}_j = \frac{\#\quad \text{de falhas em} \quad t_j}{\#\quad \text{Observações sobre risco}}$$

> **!** Importante
>
> Os métodos construídos anteriormente são ditos **não-paramétricos**, pois para a derivação dos estimadores não se faz pressuposto de distribuição para a variável aleatória T

5.3 Propriedades do estimador

Como temos um estimador pontual, podemos também construir intervalos de confiança para a estimativas.

5.3.1 Variância do Estimador

Ora ora, para calcular a variância do estimador precisamos saber a distribuição do estimador, então aqui vamos nos conter com o estimador da variância do estimador, que é dado por:

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

5.4 Códigos

Python

```
import pandas as pd
import numpy as np
from lifelines import KaplanMeierFitter
#%%
tempo = [1, 1, 1, 1, 4, 5, 7, 8, 10, 10, 12, 16, 16, 16]
falha = [1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0]

dados = pd.DataFrame({'tempo':tempo,'censura':falha})

#%%
KaplanMeierFitter().fit(dados['tempo'],dados['censura']).survival_function_
```

```
          KM_estimate
timeline
0.0          1.000000
1.0          0.785714
4.0          0.785714
5.0          0.698413
7.0          0.611111
8.0          0.523810
10.0         0.436508
12.0         0.436508
16.0         0.436508
```

6 Summary

In summary, this book has no content whatsoever.

7 Apêndice

- Integração por partes

$$\int_a^b fg'dx = fg]_a^b - \int_a^b f'gdx$$

Referências

Colosimo, Suely Ruiz, Enrico Antonio e Giolo. 2006. *Análise de sobrevivência aplicada.* Editora Blucher.