

# **Análise de Sobrevivência de Mulheres com Câncer de Mama Via Estimador de Kaplan-Meier e Teste Log-Rank**

Pedro Frazão, Núbia Almeida

Universidade Federal Fluminense

Seminários de Iniciação Científica do IME

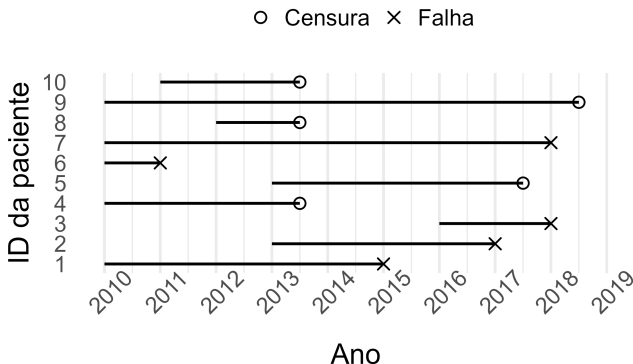
03 de dezembro de 2025

# Objetivos

- Geral: identificar atributos que influenciam a sobrevivência de mulheres em tratamento contra câncer de mama
- Específicos:
  - 1 Calcular a proporção de óbito entre as pacientes
  - 2 Descrever as curvas de sobrevida de forma geral e de acordo com características pessoais, da doença e do tratamento.
  - 3 Testar se as curvas estratificadas de acordo com essas características são iguais.

# Dados e desenho do estudo

- Os dados são provenientes do INCA (Integrador - RHC)
- Amostra de aproximadamente 36.000 mulheres que iniciaram seu tratamento entre 2010 e 2019.
- Desenho do estudo = Coorte aberta



# Função de sobrevivência

- $T$  a variável aleatória que denota o tempo até o óbito por câncer de mama
- A função de sobrevivência é definida por:

$$S(t) = P(T > t), \quad t \in \mathbb{R}$$

ou seja, é a probabilidade de uma paciente sobreviver por mais do que um tempo especificado.

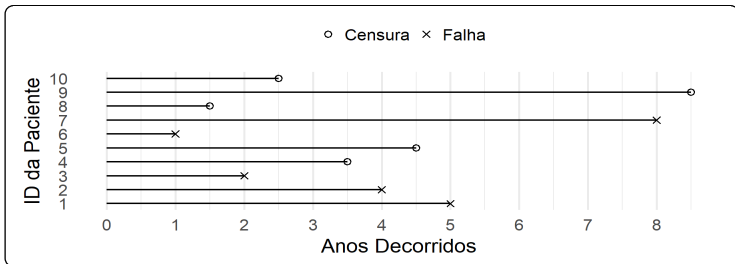
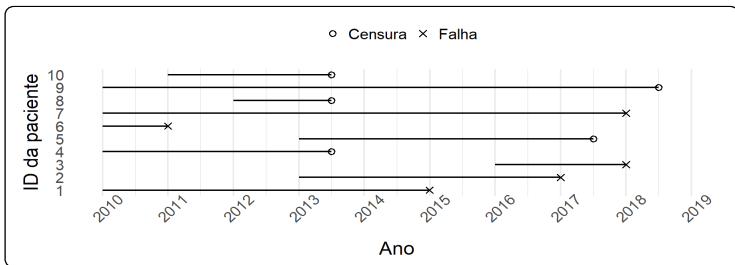
# Estimador de Kaplan-Meier

- Para estimação da função de sobrevivência, será considerado o estimador

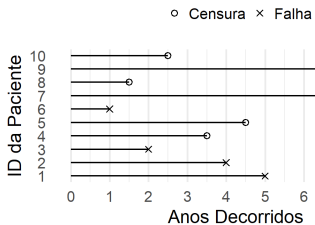
$$\hat{S}(t) = \prod_{j|t_j \leq t} \frac{n_j - d_j}{n_j}$$

- $t_1 < t_2 < \dots < t_k$ : tempos de falha distintos da amostra
- $d_j$ : quantidade de falhas em  $t_j$
- $c_j$ : quantidade de censuras que ocorreram no intervalo  $[t_j, t_{j+1})$
- $n_j$ : número de pacientes em risco num tempo imediatamente anterior a  $t_j$

# Exemplo didático...

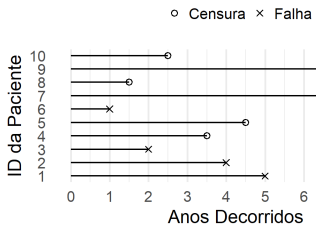


# Exemplo didático...



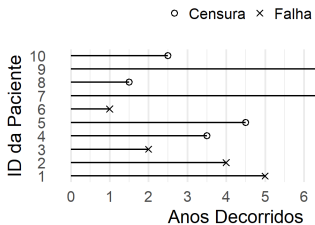
j	1	2	3	4	5
$t_j$					
$n_j$					
$d_j$					
$c_j$					

# Exemplo didático...



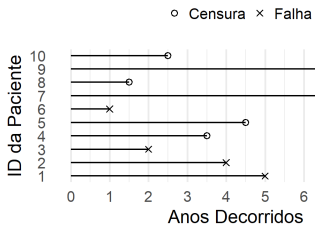
j	1	2	3	4	5
$t_j$	1				
$n_j$					
$d_j$					
$c_j$					

# Exemplo didático...



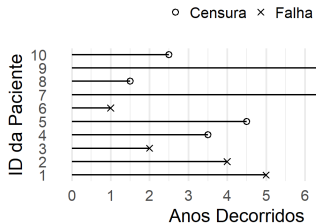
j	1	2	3	4	5
$t_j$	1				
$n_j$	10				
$d_j$					
$c_j$					

# Exemplo didático...



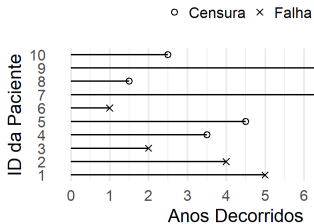
j	1	2	3	4	5
$t_j$	1				
$n_j$	10				
$d_j$	1				
$c_j$					

# Exemplo didático...



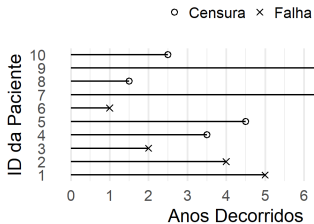
j	1	2	3	4	5
$t_j$	1				
$n_j$	10				
$d_j$	1				
$c_j$	1				

# Exemplo didático...



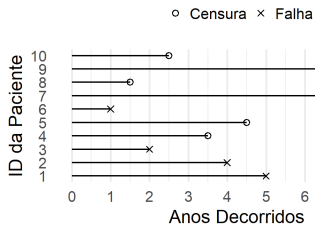
j	1	2	3	4	5
$t_j$	1	2			
$n_j$	10				
$d_j$	1				
$c_j$	1				

# Exemplo didático...



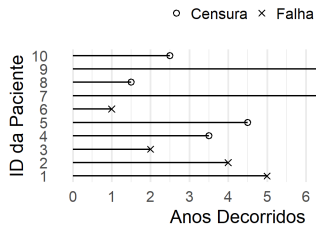
j	1	2	3	4	5
$t_j$	1	2			
$n_j$	10	8			
$d_j$	1				
$c_j$	1				

# Exemplo didático...



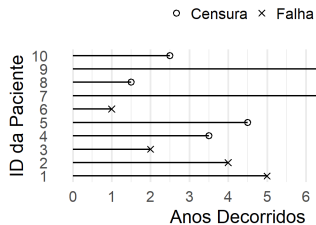
j	1	2	3	4	5
$t_j$	1	2			
$n_j$	10	8			
$d_j$	1	1			
$c_j$	1				

# Exemplo didático...



j	1	2	3	4	5
$t_j$	1	2			
$n_j$	10	8			
$d_j$	1	1			
$c_j$	1	2			

# Exemplo didático...



j	1	2	3	4	5
$t_j$	1	2	4	5	8
$n_j$	10	8	5	3	2
$d_j$	1	1	1	1	1
$c_j$	1	2	1	0	1

## Exemplo didático...

$j$	1	2	3	4	5
$t_j$	1	2	4	5	8
$n_j$	10	8	5	3	2
$d_j$	1	1	1	1	1
$c_j$	1	2	1	0	1

## Exemplo didático...

$j$	1	2	3	4	5
$t_j$	1	2	4	5	8
$n_j$	10	8	5	3	2
$d_j$	1	1	1	1	1
$c_j$	1	2	1	0	1

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

## Exemplo didático...

$j$	1	2	3	4	5
$t_j$	1	2	4	5	8
$n_j$	10	8	5	3	2
$d_j$	1	1	1	1	1
$c_j$	1	2	1	0	1

$$\hat{S}(1) = P(T > 1) = \frac{10 - 1}{10}$$

## Exemplo didático...

$j$	1	2	3	4	5
$t_j$	1	2	4	5	8
$n_j$	10	8	5	3	2
$d_j$	1	1	1	1	1
$c_j$	1	2	1	0	1

$$\hat{S}(1) = P(T > 1) = \frac{10 - 1}{10} = 0,900$$

## Exemplo didático...

$j$	1	2	3	4	5
$t_j$	1	2	4	5	8
$n_j$	10	8	5	3	2
$d_j$	1	1	1	1	1
$c_j$	1	2	1	0	1

$$\hat{S}(1) = P(T > 1) = \frac{10 - 1}{10} = 0,900$$

$$\hat{S}(2)$$

## Exemplo didático...

$j$	1	2	3	4	5
$t_j$	1	2	4	5	8
$n_j$	10	8	5	3	2
$d_j$	1	1	1	1	1
$c_j$	1	2	1	0	1

$$\hat{S}(1) = P(T > 1) = \frac{10 - 1}{10} = 0,900$$

$$\hat{S}(2) = P(T > 2)$$

## Exemplo didático...

$j$	1	2	3	4	5
$t_j$	1	2	4	5	8
$n_j$	10	8	5	3	2
$d_j$	1	1	1	1	1
$c_j$	1	2	1	0	1

$$\hat{S}(1) = P(T > 1) = \frac{10 - 1}{10} = 0,900$$

$$\hat{S}(2) = P(T > 2) = P(T > 2 \mid T > 1) \cdot P(T > 1)$$

## Exemplo didático...

$j$	1	2	3	4	5
$t_j$	1	2	4	5	8
$n_j$	10	8	5	3	2
$d_j$	1	1	1	1	1
$c_j$	1	2	1	0	1

$$\hat{S}(1) = P(T > 1) = \frac{10 - 1}{10} = 0,900$$

$$\hat{S}(2) = P(T > 2) = P(T > 2 \mid T > 1) \cdot P(T > 1) = \frac{8 - 1}{8} \cdot \frac{10 - 1}{10}$$

## Exemplo didático...

$j$	1	2	3	4	5
$t_j$	1	2	4	5	8
$n_j$	10	8	5	3	2
$d_j$	1	1	1	1	1
$c_j$	1	2	1	0	1

$$\hat{S}(1) = P(T > 1) = \frac{10 - 1}{10} = 0,900$$

$$\hat{S}(2) = P(T > 2) = P(T > 2 \mid T > 1) \cdot P(T > 1) = \frac{8 - 1}{8} \cdot \frac{10 - 1}{10} = 0,787$$

## Exemplo didático...

$j$	1	2	3	4	5
$t_j$	1	2	4	5	8
$n_j$	10	8	5	3	2
$d_j$	1	1	1	1	1
$c_j$	1	2	1	0	1

$$\hat{S}(1) = P(T > 1) = \frac{10 - 1}{10} = 0,900$$

$$\hat{S}(2) = P(T > 2) = P(T > 2 \mid T > 1) \cdot P(T > 1) = \frac{8 - 1}{8} \cdot \frac{10 - 1}{10} = 0,787$$

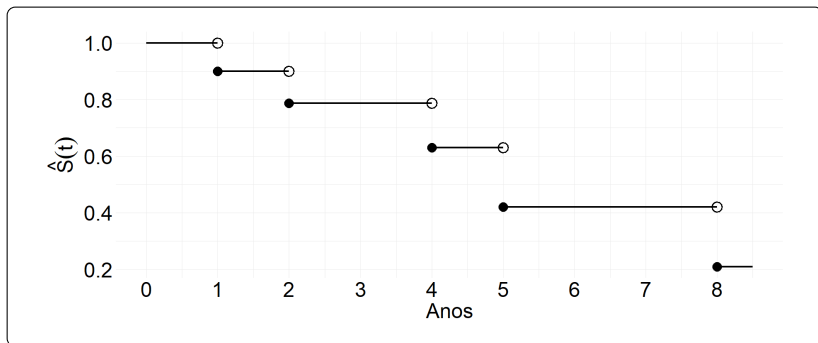
$$\hat{S}(4) = P(T > 4) = P(T > 4 \mid T > 2) \cdot P(T > 2) = \frac{5 - 1}{5} \cdot 0,787 = 0,630$$

## Exemplo didático...

$t_j$	$n_j$	$d_j$	$\hat{S}(t_j)$
1	10	1	$\frac{10-1}{10} = 0,900$
2	8	1	$\frac{8-1}{8} \cdot \frac{10-1}{10} = 0,787$
4	5	1	$\frac{5-1}{5} \cdot \frac{8-1}{8} \cdot \frac{10-1}{10} = 0,630$
5	3	1	$\frac{3-1}{3} \cdot \frac{5-1}{5} \cdot \frac{8-1}{8} \cdot \frac{10-1}{10} = 0,420$
8	2	1	$\frac{2-1}{2} \cdot \frac{3-1}{3} \cdot \frac{5-1}{5} \cdot \frac{8-1}{8} \cdot \frac{10-1}{10} = 0,210$

$$\hat{S}(t) = \prod_{j|t_j \leq t} \frac{n_j - d_j}{n_j}$$

## Exemplo didático...



# Teste Log-rank

- Hipóteses:

$$H_0 : S_1(t) = S_2(t) = \cdots = S_r(t)$$

$$H_1 : \exists u, v \mid S_u(t) \neq S_v(t)$$

# Teste Log-rank

- Hipóteses:

$$H_0 : S_1(t) = S_2(t) = \cdots = S_r(t)$$

$$H_1 : \exists u, v \mid S_u(t) \neq S_v(t)$$

- Para cada tempo de falha  $t_j$ ,  $j = 1, \dots, k$ :

	Estrato 1	Estrato 2	...	Estrato r	Total
Falhas	$d_{1j}$	$d_{2j}$	...	$d_{rj}$	$d_j$
Sob risco	$n_{1j}$	$n_{2j}$	...	$n_{rj}$	$n_j$
Proporção	$\frac{d_{1j}}{n_{1j}}$	$\frac{d_{2j}}{n_{2j}}$	...	$\frac{d_{rj}}{n_{rj}}$	$\frac{d_j}{n_j}$

# Teste Log-rank

- Hipóteses:

$$H_0 : S_1(t) = S_2(t) = \dots = S_r(t)$$

$$H_1 : \exists u, v \mid S_u(t) \neq S_v(t)$$

- Para cada tempo de falha  $t_j$ ,  $j = 1, \dots, k$ :

	Estrato 1	Estrato 2	...	Estrato r	Total
Falhas	$d_{1j}$	$d_{2j}$	...	$d_{rj}$	$d_j$
Sob risco	$n_{1j}$	$n_{2j}$	...	$n_{rj}$	$n_j$
Proporção	$\frac{d_{1j}}{n_{1j}}$	$\frac{d_{2j}}{n_{2j}}$	...	$\frac{d_{rj}}{n_{rj}}$	$\frac{d_j}{n_j}$
Falhas esperadas	$n_{1j} \frac{d_j}{n_j}$	$n_{2j} \frac{d_j}{n_j}$	...	$n_{rj} \frac{d_j}{n_j}$	$d_j$

# Teste Log-rank

- Nº de falhas esperadas sob  $H_0$ :

$$e_{ij} = n_{ij} \left( \frac{d_j}{n_j} \right)$$

- Desvios em  $t_j$ :

$$w'_j = (d_{1j} - e_{1j}, \dots, d_{rj} - e_{rj})$$

- Desvios ao longo de todo estudo:

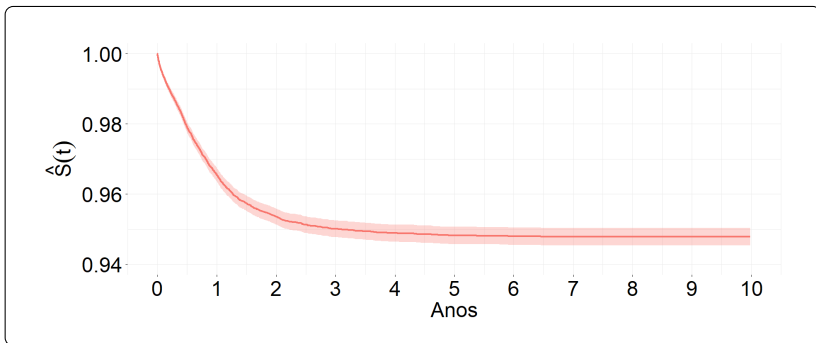
$$w = \sum_{j=1}^k w_j = \left( \sum_{j=1}^k (d_{1j} - e_{1j}), \dots, \sum_{j=1}^k (d_{rj} - e_{rj}) \right)$$

- Estatística de teste:

$$w' W w \xrightarrow[n \rightarrow \infty]{d} \chi^2_{r-1}$$

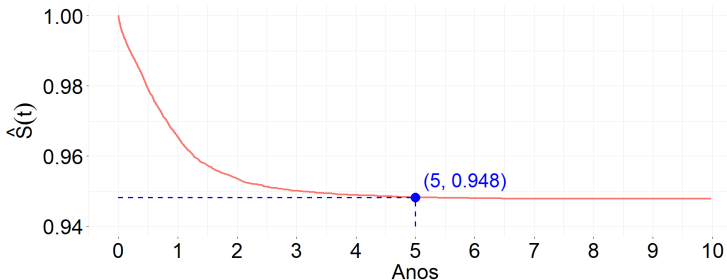
# Função de sobrevivência geral

- Percentual de pacientes que foram a óbito: 4,9% (1787 mulheres)



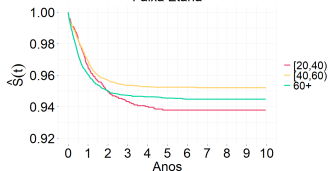
# Função de sobrevivência geral

- Obtendo a probabilidade de sobreviver por mais do que 5 anos...

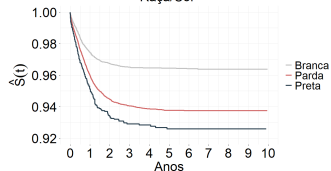


# Características da paciente

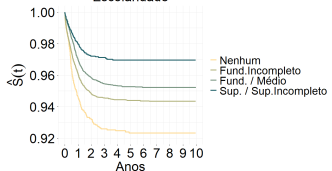
Faixa Etária



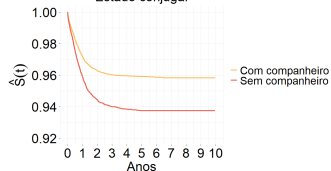
Raça/Cor



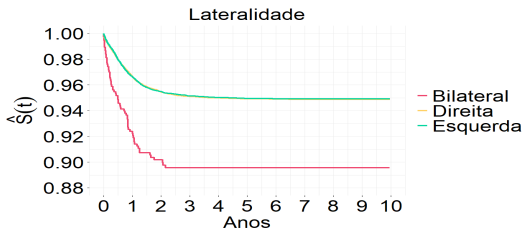
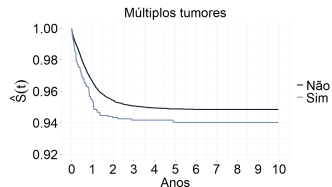
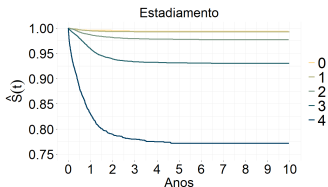
Escolaridade



Estado conjugal

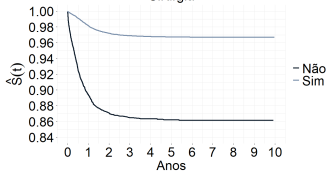


# Características da doença

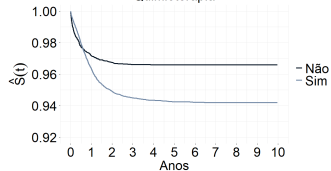


# Características do tratamento

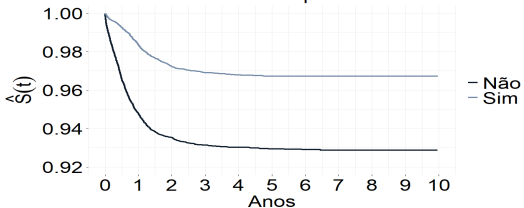
Cirurgia



Quimioterapia



Radioterapia

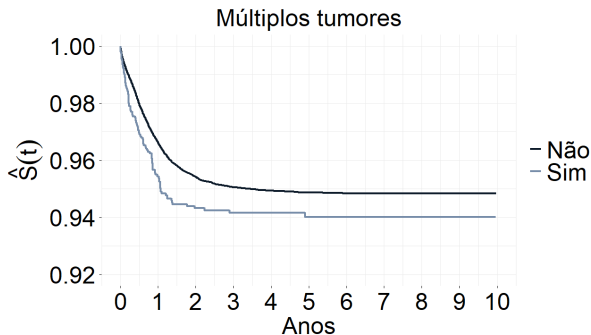


## Conclusão e pontos a meditar...

- Todas as variáveis analisadas, a princípio, exercem algum tipo de influência na sobrevivência das mulheres (valor- $p < 5\%$ )

## Conclusão e pontos a meditar...

- Todas as variáveis analisadas, a princípio, exercem algum tipo de influência na sobrevivência das mulheres (valor- $p < 5\%$ )
- Percebe-se que em alguns casos há estratos com curvas de sobrevivência próximas.



## Conclusão e pontos a meditar...

- Todas as variáveis analisadas, a princípio, exercem algum tipo de influência na sobrevivência das mulheres (valor- $p < 5\%$ )
- Percebe-se que em alguns casos há estratos com curvas de sobrevivência próximas.
- Supusemos que as mulheres sem data de óbito preenchida estavam vivas.

## Conclusão e pontos a meditar...

- Todas as variáveis analisadas, a princípio, exercem algum tipo de influência na sobrevivência das mulheres (valor- $p < 5\%$ )
- Percebe-se que em alguns casos há estratos com curvas de sobrevivência próximas.
- Supusemos que as mulheres sem data de óbito preenchida estavam vivas.
- O banco de dados apresentou algumas inconsistências quanto a ordenação do acontecimento de eventos.

## Conclusão e pontos a meditar...

- Todas as variáveis analisadas, a princípio, exercem algum tipo de influência na sobrevivência das mulheres (valor- $p < 5\%$ )
- Percebe-se que em alguns casos há estratos com curvas de sobrevivência próximas.
- Supusemos que as mulheres sem data de óbito preenchida estavam vivas.
- O banco de dados apresentou algumas inconsistências quanto a ordenação do acontecimento de eventos.
- Variáveis importantes com baixo grau de preenchimento.

## Conclusão e pontos a meditar...

- Todas as variáveis analisadas, a princípio, exercem algum tipo de influência na sobrevivência das mulheres ( $\text{valor-}p < 5\%$ )
- Percebe-se que em alguns casos há estratos com curvas de sobrevivência próximas.
- Supusemos que as mulheres sem data de óbito preenchida estavam vivas.
- O banco de dados apresentou algumas inconsistências quanto a ordenação do acontecimento de eventos.
- Variáveis importantes com baixo grau de preenchimento.
- Houve dúvidas em relação a verdadeira causa dos óbitos.

## Trabalhos futuros...

- Utilização de outro banco de dados

## Trabalhos futuros...

- Utilização de outro banco de dados
- Tentar ajustar modelos que quantificam os efeitos de um fator no risco de óbito, considerada a presença das demais variáveis.

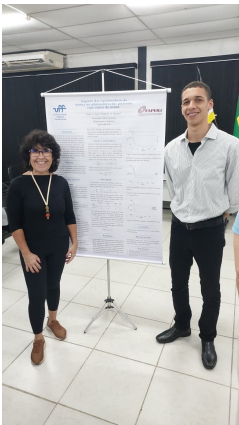
# Experiência de IC

- Altamente agregadora para minha formação.
- Aprendi (e ainda estou aprendendo) a fazer e comunicar ciência.

# Experiência de IC - MGEST



# Experiência de IC - SIMMA



## Referências

- **KAPLAN, E. L.; MEIER, P.** Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, v. 53, n. 282, p. 457-481, jun. 1958.
- **KALBFLEISCH, J. D.; PRENTICE, R. L.** *The Statistical Analysis of Failure Time Data*. 2. ed. Nova York: Wiley, 2002.



# Agradecimentos



Este trabalho contou com o apoio da Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), por meio de bolsa de iniciação científica.

## Contato

# Obrigado!

-  **Email:** [pefrazao@id.uff.br](mailto:pefrazao@id.uff.br)
-  **GitHub:** [github.com/FrazaoPe](https://github.com/FrazaoPe)

Dúvidas? Sugestões? Vamos conversar!