Especialização em *Data Science* e Estatística Aplicada

Módulo IV - Análise de Sobrevivência

Prof. Dr. Eder Angelo Milani

Goiânia, 2025

IME
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
FEN
FACULDADE DE ENFERMAGEM
UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Conteúdo Programático

- Conceitos básicos.
- Técnicas não-paramétricas.
- Modelos probabilísticos em análise de sobrevivência.
- Modelos de regressão paramétrico.
- Modelo semiparamétrico de riscos proporcionais de Cox.
- Métodos para verificação do modelo ajustado.
- Modelo de Cox estratificado.

Conteúdo - Aula 1

1. Conceitos básicos
   - Introdução
   - Censura
   - Principais funções

2. Técnicas não-paramétricas
   - Introdução
   - O estimador de Kaplan-Meier
   - Comparação de curvas de sobrevivência

Introdução

Tempo de falha

Em Análise de sobrevivência, a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse. Este tempo é denominado **tempo de falha**, podendo ser o tempo até a morte do paciente, bem como até a cura ou recidiva de uma doença.

Exemplo 1

- O tempo até o aparecimento do efeito colateral de um medicamento;
- O tempo até o diagnóstico de uma doença;
- O tempo até a cura de um determinado tumor;
- O tempo até uma lâmpada queimar.

Introdução

A teoria de análise de sobrevivência pode ser aplicada à

- área médica;
- pesquisa industrial;
- área financeira;
- área educacional;
- entre outras.

Introdução

O **tempo de falha** precisa estar bem definido, assim como a escala de medida.

Exemplo 2

- Quando o evento de interesse é a morte devido a um tipo de tumor, geralmente o tempo se inicia com o diagnóstico da doença e se finda no momento da ocorrência do óbito. A escala utilizada pode ser dias, meses ou anos.

Introdução

O termo análise de sobrevivência refere-se basicamente a situações médicas envolvendo dados censurados. Entretanto, condições similares ocorrem em outras áreas em que se usam as mesmas técnicas de análise de dados.

Em engenharia, são comuns os estudos em que produtos ou componentes são colocados sob teste para se estimar características relacionadas aos seus tempos de vida, tais como o tempo médio ou a probabilidade de um certo produto durar mais que 5 anos, por exemplo. Os engenheiros denominam esta **área de confiabilidade**.

Censura

Censura

A principal característica de dados de sobrevivência é a presença de **censura**, que é a observação parcial da resposta. Isto se refere a situações em que, por alguma razão, o acompanhamento do paciente foi interrompido, seja porque o paciente mudou de cidade, o estudo terminou para a análise dos dados ou, o paciente morreu de causa diferente da estudada.

A censura ocorre quando o tempo até o evento de interesse não é observado, ou quando o estudo, ou experimento, deve ser encerrado e ainda existem itens/indivíduos funcionando/vivos, cujos tempos de falha, obviamente, ainda não foram observados. **A presença de observações incompletas ou parciais são denominadas censura**.

Censura

Ressalta-se o fato de que, **mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser usados na análise estatística**. Duas razões justificam tal procedimento:

- mesmo sendo incompletas, as observações censuradas fornecem informações sobre o tempo de vida de pacientes;
- a omissão das censuras no cálculo das estatísticas de interesse pode acarretar conclusões viciadas.

Censura

Três razões usuais para ocorrer censura:

- O evento não ocorre para um determinado indivíduo antes do final do estudo;
- Perdeu-se o acompanhamento (*follow-up*) do indivíduo;
- O indivíduo é retirado do estudo, por exemplo porque morreu de outra causa ou porque interrompeu o tratamento.

Tipos de censura

Tipos de censura:

- **Censura Tipo I** ocorrem naqueles estudos que ao serem finalizados após um período pré-estabelecido de tempo registram, em seu término, alguns indivíduos que ainda não apresentaram o evento de interesse.
- **Censura Tipo II** resultam de estudos os quais são finalizados após a ocorrência do evento de interesse em um número pré-estabelecido de indivíduos.
- **Censura do tipo aleatória** é quando um paciente é retirado no decorrer do estudo sem ter ocorrido a falha, ou também, por exemplo, se o paciente morrer por uma razão diferente da estudada.

Tipos de censura

Tipos de censura (continuação):

- **Censura à direita** é quando o tempo de ocorrência do desfecho está à direita do tempo registrado.
- **Censura à esquerda** ocorre quando o tempo registrado é maior do que o tempo de falha. Isto é, o evento de interesse já aconteceu quando o indivíduo foi observado.
- **A censura intervalar** ocorre quando não se sabe o tempo exato da ocorrência da falha, no entanto, sabe-se que ela ocorreu dentro de um intervalo de tempo conhecido.

Exercício

Qual o tipo de censura representado na Figura 1?

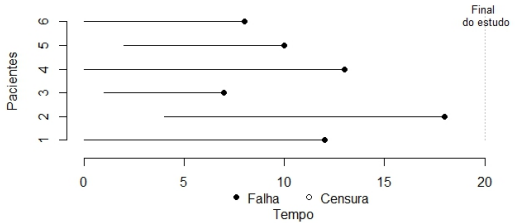

Figura 1: Representação gráfica do tempo de 6 pacientes

Exercício

Qual o tipo de censura representado na Figura 1?

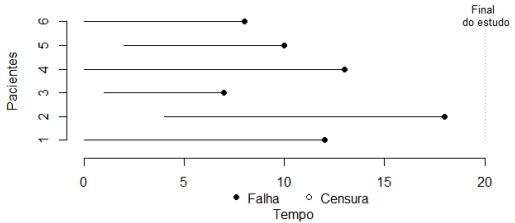

Figura 1: Representação gráfica do tempo de 6 pacientes

Resposta: Não é apresentado nenhum tipo de censura na figura.

Exercício

Qual o tipo de censura representado na Figura 2?

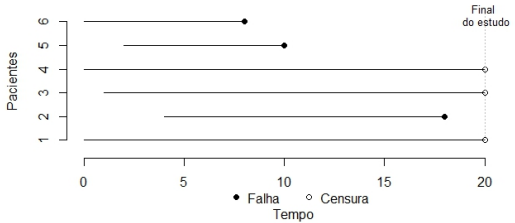

Figura 2: Representação gráfica do tempo de 6 pacientes

Exercício

Qual o tipo de censura representado na Figura 2?

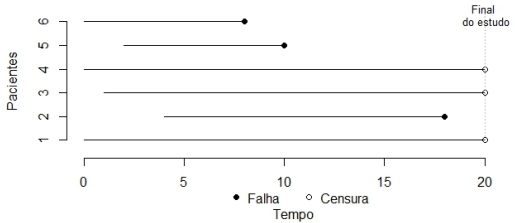

Figura 2: Representação gráfica do tempo de 6 pacientes

Resposta: Censura tipo I.

Exercício

Qual o tipo de censura representado na Figura 3?

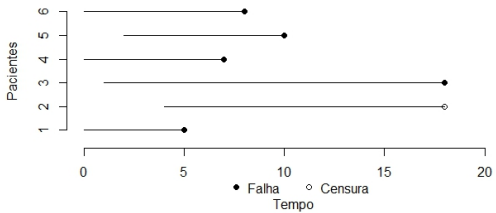

Figura 3: Representação gráfica do tempo de 6 pacientes

Exercício

Qual o tipo de censura representado na Figura 3?

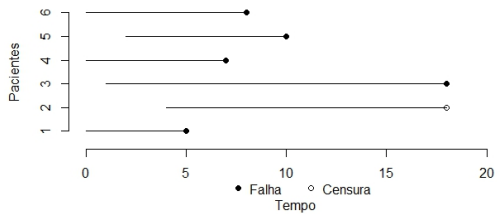

Figura 3: Representação gráfica do tempo de 6 pacientes

Resposta: Censura tipo II.

Exercício

Qual o tipo de censura representado na Figura 4?

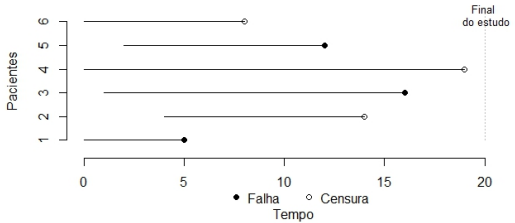

Figura 4: Representação gráfica do tempo de 6 pacientes

Exercício

Qual o tipo de censura representado na Figura 4?

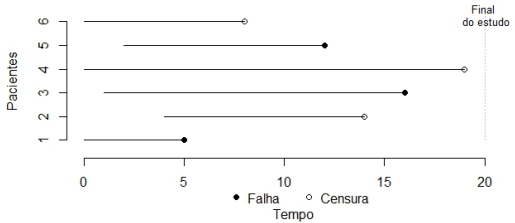

Figura 4: Representação gráfica do tempo de 6 pacientes

Resposta: Censura do tipo aleatória.

Exercício

Qual o tipo de censura representado na Figura 5?

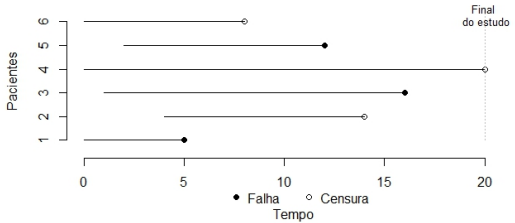

Figura 5: Representação gráfica do tempo de 6 pacientes

Exercício

Qual o tipo de censura representado na Figura 5?

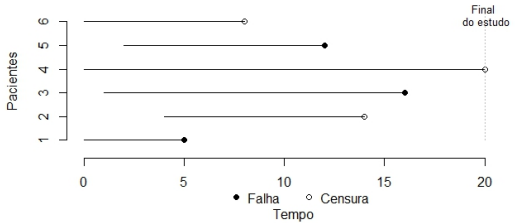

Figura 5: Representação gráfica do tempo de 6 pacientes

Resposta: Censura do tipo à direita.

Censura à direita

O curso será voltado a dados de sobrevivência com censura à direita, que é a situação encontrada com mais frequência em estudos, tanto na área médica quanto em engenharia.

Desta forma, quando for simplesmente mencionado a palavra **censura entende-se por censura à direita**.

Censura aleatória

Uma representação simples do mecanismo de **censura aleatória** é feita usando duas variáveis aleatórias. Considere T uma variável aleatória representando o tempo de falha e C, uma outra variável aleatória independente de T, representando o tempo de censura associado a este paciente. O que se observa para este paciente é, portanto,

$$t = min(T, C),$$

e

$$\delta = \begin{cases} 1, & \text{se } T \leq C \\ 0, & \text{se } T > C. \end{cases}$$

Censura na prática

No conjunto de dados tem uma variável (coluna) indicando a censura?

Censura na prática

No conjunto de dados tem uma variável (coluna) indicando a censura?
Resposta: Em geral, não.

Como identificar se um conjunto de dados tem ou não tem censura?

Censura na prática

No conjunto de dados tem uma variável (coluna) indicando a censura?
Resposta: Em geral, não.

Como identificar se um conjunto de dados tem ou não tem censura?
Resposta: Tem que analisar o problema a ser respondido e o conjunto de dados.

Censura na prática - Exemplo

Problema: Analisar o tempo do diagnóstico até a morte devido ao câncer de pulmão.

Conjunto de dados: Analisar as variáveis que dão início e que finaliza a contagem do tempo.

Para exemplificar analisamos o dicionário dos dados disponibilizados pela FOSP - Fundação Oncocentro do Estado de São Paulo (https://fosp.saude.sp.gov.br).

Censura na prática - Exemplo

Em resumo, as principais variáveis são:

- identificação do tipo de câncer: **TOPOGRUP = c34 - CID C34** - “neoplasia maligna dos brônquios e dos pulmões”
- inicio da contagem do tempo: **DTDIAG - data do diagnóstico**
- a que finaliza a contagem do tempo: **DTULTINFO - data da última informação**
- determinação se o tempo é de falha ou censurado: **ULTINFO - última informação do paciente**

Censura na prática - Exemplo

Vamos então construir a variável que indica se o tempo é de falha ou de censura (δ). Na sequência, vamos selecionar algumas variáveis que serão posteriormente utilizadas para explicar o evento de interesse.

Aplicação no *software* **R.**

Função de sobrevivência

A variável aleatória não-negativa T, usualmente contínua, que representa o tempo de falha, é geralmente especificada em análise de sobrevivência pela sua função de sobrevivência ou pela função de taxa de falha.

Função de sobrevivência

A função de sobrevivência, uma das principais funções probabilísticas usadas para descrever estudos de sobrevivência, é definida como **a probabilidade de uma observação não falhar até um certo tempo** t, ou seja, a probabilidade de uma observação sobreviver ao tempo t. Em termos probabilísticos, isto é escrito como:

$$S(t) = P(T \geq t) = 1 - F(t).$$

Função de sobrevivência

Note que

- $S(0) = 1$;
- $\lim_{t \to \infty} S(t) = 0$;
- a função $S(t)$ é não crescente.

Função de sobrevivência - Exemplo

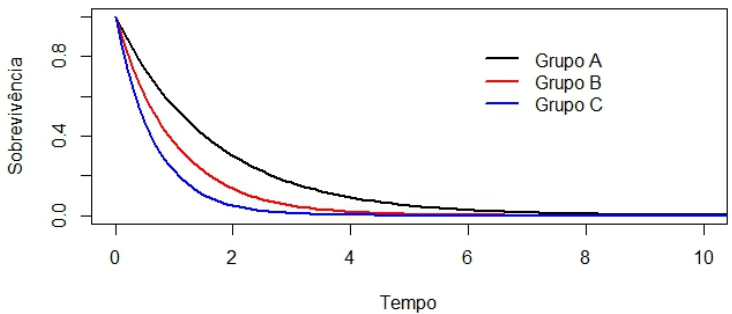

Figura 6: Várias curvas da função de sobrevivência

Função de sobrevivência - Exemplo

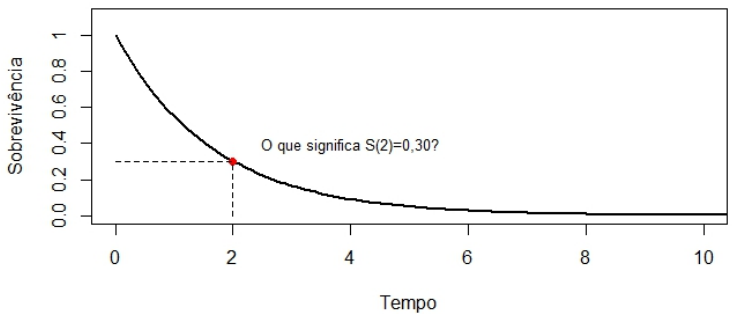

Figura 7: Uma curva da função de sobrevivência

Função de Taxa de Falha

Função taxa de falha

A função de taxa de falha $\lambda(t)$, representa a taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t, é bastante útil para descrever a distribuição do tempo de vida de pacientes. Ela descreve a forma em que a taxa instantânea de falha muda com o tempo.

A função de taxa de falha de T é, então, definida por

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

Observe que as taxas de falha são números positivos, mas sem limite superior.

Função de Taxa de Falha

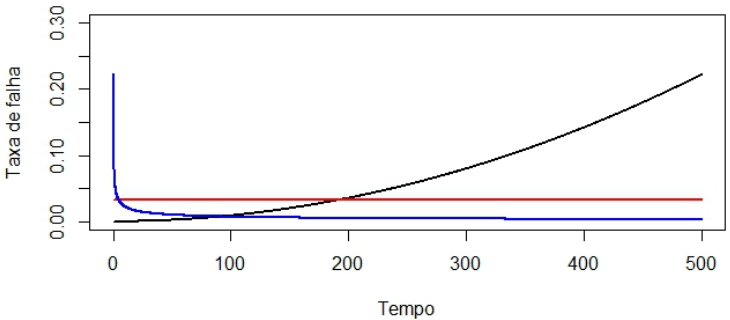

Figura 8: Exemplos da função taxa de falha

Função de Taxa de Falha

- A função crescente indica que a taxa de falha do paciente aumenta com o transcorrer do tempo.
- A função constante indica que a taxa de falha não se altera com o passar do tempo.
- A função decrescente mostra que a taxa de falha diminui à medida que o tempo passa.

Função de Taxa de Falha

Observação
A função de taxa de falha é mais informativa do que a função de sobrevivência. Diferentes funções de sobrevivência podem ter formas semelhantes, enquanto as respectivas funções de taxa de falha podem diferir drasticamente. Desta forma, a modelagem da função de taxa de falha é um importante método para dados de sobrevivência.

A função taxa de falha também é comumente chamada de função de risco.

Função de Taxa de Falha Acumulada

Outra função útil em análise de dados de sobrevivência é a função de taxa de falha acumulada. Esta função, como o próprio nome sugere, fornece a taxa de falha acumulada do indivíduo e é definida por

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

Ou seja, é a área embaixo da curva da função taxa de falha.

Outras funções importantes

Outras três quantidades de interesse em análise de sobrevivência são: o tempo médio de vida, o tempo de vida mediano e a vida média residual.

Tempo médio de vida

O tempo médio de vida (ou em inglês “mean time to failure = MTTF”) é obtido pela área sob a função de sobrevivência, isto é $t_m = MTTF = \int_0^\infty S(t)dt$

Tempo de vida mediano

O tempo de vida mediano é o tempo que satisfaz $S(t_{md}) = 0,5$

Outras funções importantes

Vida média residual

Já a vida média residual é definida condicional a um certo tempo de vida t. Ou seja, para indivíduos com idade t esta quantidade mede o tempo médio restante de vida e é, então, a área sob a curva de sobrevivência à direita do tempo t dividida por $S(t)$, ou seja, $vmr(t) = \frac{\int_t^\infty S(u)du}{S(t)}$. Note que $vmr(0) = t_m$.

Introdução

Os objetivos da análise estatística de dados de sobrevivência na área da saúde estão geralmente relacionados com a identificação de fatores prognósticos para uma doença ou com a comparação de tratamentos em um ensaio clínico, controlado por outros fatores.

Por mais complexo que seja o estudo, respostas às perguntas de interesse são obtidas a partir de um conjunto de dados de sobrevivência, em que o passo inicial consiste na realização de uma análise descritiva dos dados.

Introdução

A presença de observações censuradas é um problema para as técnicas convencionais de análise descritiva, envolvendo média, desvio-padrão e técnicas gráficas, como histograma, box-plot, entre outras.

Tais problemas podem ser ilustrados por meio de uma situação bem simples em que há interesse na construção de um histograma. Se a amostra não contiver observações censuradas, a construção do histograma consiste na divisão do eixo do tempo em um certo número de intervalos e, em seguida, conta-se o número de ocorrências de falhas em cada intervalo. Entretanto, **quando existem censuras, não é possível construir um histograma, pois não se conhece a frequência exata associada a cada intervalo**.

Introdução

Nos textos básicos de estatística, uma análise descritiva consiste, essencialmente, em encontrar medidas de tendência central e de variabilidade.

Devido à presença de censuras nos dados de sobrevivência, **o principal componente da análise descritiva envolvendo dados de tempo de vida é a função de sobrevivência**.

Desse modo, é necessário estimar está função e, a partir dela, estimar as estatísticas de interesse.

O estimador de Kaplan-Meier

O estimador de Kaplan-Meier para a função de sobrevivência é o mais utilizado em estudos clínicos e vem ganhando cada vez mais espaço em estudos de confiabilidade.

O estimador não-paramétrico de Kaplan-Meier (ou estimador limite-produto), proposto por Kaplan e Meier (1958) para estimar a função de sobrevivência. Ele é uma adaptação da função de sobrevivência empírica que, na ausência de censuras, é definida como:

$$\widehat{S}(t) = \frac{\text{nº de observações que não falharam até o tempo t}}{\text{nº de observações no estudo}},$$

sendo que $\widehat{S}(t)$ é uma função escada com degraus nos tempos observados de falha de tamanho $1/n$, em que n é o tamanho da amostra. Se existirem empates em um certo tempo t, o tamanho do degrau fica multiplicado pelo número de empates.

O estimador de Kaplan-Meier

Considere

- $t_1 < t_2 < \ldots < t_k$, os k tempos distintos e ordenados de falha,
- d_j o número de falhas em t_j, $j = 1, \ldots, k$ e
- n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

O estimador de Kaplan-Meier é definido como:

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

O estimador de Kaplan-Meier

Aplicação no *software* **R.**

Comparação de curvas de sobrevivência

Os dados apresentados anteriormente, foram obtidos para investigar o efeito da variável sexo no tratamento de câncer CID C34. Isto significa que o **objetivo principal do estudo é comparar o tempo de sobrevida em ambos os sexos**.

A seguir o teste *logrank* é apresentado. Este teste é muito utilizado em análise de sobrevivência e é particularmente apropriado quando a razão das funções de taxa de falha dos grupos a serem comparados é aproximadamente constante.

Teste *logrank*

Considere o teste de hipóteses de igualdade de duas funções de sobrevivência $S_1(t)$ e $S_2(t)$. **A hipótese nula considerada é a de igualdade das curvas**, isto é

$$H_0 : S_1(t) = S_2(t).$$

Ao analisar o p-valor tem-se duas possíveis conclusão para o teste, que são:

(i) **p-valor** $<$ α (nível de significância) - existe evidências para **rejeitar** a hipótese de igualdade das curvas de sobrevivência;

(ii) **p-valor** $>$ α (nível de significância) - existe evidências para **não rejeitar** a hipótese de igualdade das curvas de sobrevivência.

Teste *logrank*

Aplicação no *software* **R.**

Especialização em *Data Science* e Estatística Aplicada

Módulo IV - Análise de Sobrevivência

Prof. Dr. Eder Angelo Milani

edermilani@ufg.br

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

FEN FACULDADE DE ENFERMAGEM

UFG UNIVERSIDADE FEDERAL DE GOIÁS