Análise de Sobrevivência

Iniciação Ciêntifica - PIBIC 2024 (UFPA)

Breno Cauã Rodrigues da Silva

Invalid Date

Índice

# Prefácio

## Resumo

## Abstract

# 1. Conceitos Básicos e Exemplos

## 1.1 Introdução

O primeiro capítulo do livro de Enrico Antônio Colosimo e Suely Ruiz Giolo tem como objetivo apresentar alguns *conceitos* e *fundamentos* de uma das áreas da Estatística e Análise de Dados que mais cresceram nas últimas duas décadas do século passado. Esse crescimento foi impulsionado pelo desenvolvimento e avanço de técnicas, juntamente com o progresso computacional.

Na Análise de Sobrevivência, a variável resposta é, geralmente, o tempo até a ocorrência de um determinado evento. De forma mais precisa, trata-se de uma técnica estatística utilizada para modelar e entender o tempo até que ocorra um evento de interesse, denominado **tempo de falha**. Para um entendimento inicial, Colosimo e Giolo dão os seguintes exemplos: tempo até a morte de um paciente, tempo até a cura ou até a recidiva de uma doença.

Uma questão que pode surgir é: por que não usar outras técnicas estatísticas? O uso de outras abordagens não é adequado para dados de sobrevivência devido à característica desses dados, que é a presença de **censura**. De forma simples, censura refere-se à observação parcial da resposta, o que ocorre quando o acompanhamento do paciente é interrompido por alguma razão. Sendo um conceito chave na análise de sobrevivência, podemos defini-la como a situação em que o tempo de falha real não é conhecido, apenas que ele excede certo ponto.

## 1.2 Tempo de Falha

Em Análise de Sobrevivência, é essencial definir alguns pontos fundamentais para o estudo. O primeiro deles é o tempo de início do estudo, que deve ser definido com precisão, garantindo que os indivíduos sejam comparáveis na origem do estudo, diferenciando-se apenas nas medidas das covariáveis. Existem muitas alternativas para definir o tempo inicial. Geralmente, esse tempo é o tempo real ou “de relógio”. Porém, em outras áreas, como a Engenharia, outras medidas podem ser utilizadas. Colosimo e Giolo fornecem exemplos como número de ciclos, quilometragem de um carro ou qualquer outra medida de carga.

Outro ponto importante relacionado ao Tempo de Falha é a definição do evento de interesse. Normalmente, esses eventos correspondem a situações indesejáveis, por isso são chamados de falhas. A definição da falha deve ser clara e precisa. Destaca-se um trecho do livro:

\*“Em algumas situações, a definição de falha já é clara, como morte ou recidiva, mas em outras pode assumir termos ambíguos. Por exemplo, fabricantes de produtos alimentícios desejam saber o tempo de vida de seus produtos expostos em balcões frigoríficos de supermercados. O tempo de falha vai do momento de exposição (chegada ao supermercado) até o produto se tornar ‘inapropriado para consumo’. Esse evento deve ser claramente definido antes do início do estudo. Por exemplo, o produto é considerado inapropriado para consumo quando atinge uma concentração específica de microrganismos por\* de área.”

## 1.3 Censura

Frequentemente, estudos clínicos que assumem a resposta como uma variável temporal são prospectivos e de longa duração. Mesmo sendo longos, esses estudos costumam terminar antes que todos os indivíduos venham a falhar.

Uma característica comum a esses estudos é a presença de **censura**, ou seja, observações incompletas ou parciais. É importante ressaltar que, mesmo censuradas, essas observações fornecem informações valiosas sobre o tempo de vida dos pacientes. Colosimo e Giolo destacam a importância de manter os dados censurados na análise:

*“Ressalta-se que, mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser incluídos na análise estatística. Duas razões justificam esse procedimento: (i) mesmo sendo incompletas, as observações censuradas fornecem informações sobre o tempo de vida dos pacientes; (ii) a exclusão das censuras no cálculo das estatísticas pode levar a conclusões enviesadas.”*

São apresentados três tipos principais de censura:

* **Censura Tipo I:** O estudo é encerrado após um período de tempo pré-estabelecido.
* **Censura Tipo II:** O estudo termina quando o evento de interesse ocorre em um número específico de indivíduos.
* **Censura Aleatória:** Ocorre quando um paciente é retirado do estudo antes do evento.

No livro, a Figura 1.1 ilustra esses tipos de censura, todos conhecidos como censura à direita, pois o evento ocorre após o tempo registrado. Contudo, outros tipos de censura, como à esquerda e intervalar, também são possíveis.

Censura à esquerda ocorre quando o evento já aconteceu antes da observação. Um exemplo do livro é um estudo sobre a idade em que as crianças aprendem a ler em determinada comunidade:

*“Quando os pesquisadores começaram a pesquisa, algumas crianças já sabiam ler e não se lembravam com que idade isso ocorreu, caracterizando observações censuradas à esquerda.”*

No mesmo estudo, há censura à direita para crianças que não sabiam ler quando os dados foram coletados. Neste caso, os tempos de vida são considerados duplamente censurados (Turnbull, 1974).

De forma geral, a censura intervalar ocorre em estudos com visitas periódicas espaçadas, onde só se sabe que a falha ocorreu dentro de um intervalo de tempo. Quando o tempo de falha é impreciso, é dito que ele pertence a um intervalo . Esses dados são conhecidos como sobrevivência intervalar ou dados de censura intervalar. Note que tempos exatos de falha, sejam censura à direita ou à esquerda, são casos especiais de sobrevivência intervalar com . Em particular, para censura à direita e para censura à esquerda (Lindsey et al., 1998). veja a nota a seguir, que enfatiza um trecho que merece atenção no livro.

**Nota:** *“A presença de censura traz desafios para a análise estatística. A censura do Tipo II é, em princípio, mais tratável que os outros tipos, mas para situações simples, que raramente ocorrem em estudos clínicos (Lawless, 1982). Na prática, utiliza-se resultados assintóticos para a análise dos dados de sobrevivência.”*

## 1.4 Dados Truncados

Truncamento é uma característica de alguns estudos de sobrevivência que, muitas vezes, é confundida com censura. Ele ocorre quando certos indivíduos são excluídos do estudo devido a uma condição específica. Nesses casos, os pacientes só são incluídos no acompanhamento após passarem por um determinado evento, em vez de serem acompanhados desde o início.

## 1.5 Representação dos Dados de Sobrevivência

Seja uma amostra aleatória de tamanho , o -ésimo indivíduo no estudo é representado, em geral, pelo par , onde é o tempo de falha ou censura, indicado pela variável binária , definida como:

Portanto, a variável resposta em análise de sobrevivência é representada por duas colunas no conjunto de dados.

Se o estudo também incluiu covariáveis, os dados são representados por . Caso a censura seja intervalar, a representação é .

**Nota:** A Seção 1.5 do livro apresenta exemplos de *Dados de Sobrevivência*.

## 1.6 Especificando o Tempo de Sobrevivência

Seja , uma variável aleatória (va) que, na maioria dos casos é contínua, representa o tempo de falha, assim, . Tal variável é geralmente pela sua *função risco* ou pela *função de taxa de falha* (ou risco). Tais funções, e outras relacionadas, são usados ao decorrer do processo de análise de dados de sobrevivência. A seguir, algumas definições.

### 1.6.1 Função de Sobrevivência

Esta é uma das principais funções probabilísticas usadas em análise de sobrevivência. A função sobrevivência é definida como a probabilidade de uma observação não falhar até certo ponto , ou seja a probabilidade de uma observação sobreviver ao tempo . Em probabilidade, isso pode ser escrito como:

uma conclusão a qual podemos chegar, é que a probabilidade de uma observação não sobreviver até o tempo , é a acumulada até o ponto , logo,

### 1.6.2 Função de Taxa de Falha ou de Risco

A probabilidade da falha ocorrer em um intervalo de tempo pode ser expressa em termos da função de sobrevivência como:

A taxa de falha no intervalo é definida como a probabilidade de que a falha ocorra neste intervalo, dado que não ocorreu antes de , dividida pelo comprimento do intervalo. Assim, a taxa de falha no intervalo é expressa por

De forma geral, redefinindo o intervalo como a expressão assume a seguinte forma:

Assumindo bem pequeno, representa a taxa de falha instantânea no tempo condicional à sobrevivência até o tempo . Observe que as taxas de falha são números positivos, mas sem limite superior. A função de taxa de falha 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 é, então, definida como:

A Figura 1.3, do livro, mostra três funções de taxa de falha. A função crescente indica que a taxa de falha do paciente aumenta com o transcorrer do tempo. Este comportamento mostra um efeito gradual de envelhecimento. 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.

Sabe-se, ainda, que a taxa de falha para o tempo de vida de seres humanos é uma combinação das curvas apresentadas na Figura 1.3 em diferentes períodos de tempo. Ela é conhecida como *curva da banheira* e tem uma taxa de falha decrescente no período inicial, representando a mortalidade infantil, constante na faixa intermediária e crescente na porção final. Uma representação desta curva é mostrada na Figura 1.4, do livro.

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.

### 1.6.3 Função de Taxa de Falha Acumulada

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

A função de taxa de falha acumulada, , não têm uma interpretação direta, mas pode ser útil na avaliação da função de maior interesse que é a função de taxa de falha, . Isto acontece essencialmente na estimação não-paramétrica em que apresenta um estimador com propriedades ótimas e é difícil de ser estimada.

### 1.6.4 Tempo Médio e Vida Média Residual

Outras duas quantidades de interesse em análise de sobrevivência são: o tempo médio de via e a vida média residual. A primeira é obtida pela área sob a função de sobrevivência. Isto é,

Já a vida média residual é definida condicional a um certo tempo de vida . Ou seja, para indivíduos com idade está quantidade mede o tempo médio restante de vida e é, então, a área sob a curva de sobrevivência à direita do tempo dividida por . Isto é,

sendo a função densidade de . Observe que .

## 1.7 Relações entre as Funções

Para uma variável aleatória contínua e não-negativa, tem-se, em termos das funções definidas anteriormente, algumas relações matemáticas importantes entre elas, a saber:

e

Tais relações mostram que o conhecimento de uma das funções, por exemplo , implica no conhecimento das demais, isto é, , , e . Outras relações envolvendo estas funções são as seguintes:

e

# 2. Técnicas Não Paramétricas

## 2.1 Introdução

O segundo capítulo do livro que está sendo usado como um dos livros-base, apresenta as técnicas não-paramétricas utilizadas para a análise de dados de sobrevivência. Essas técnicas são empregadas quando não se faz suposições sobre a forma específica da distribuição dos tempos de falha, sendo particularmente úteis para dados censurados.

## 2.2 O Estimador de Kaplan-Meier

Proposto em 1958 por Edward L. Kaplan e Paul Meier. É um estimador não-paramétrico utilizado para estimar a função de sobrevivência, . Tal estimador também é chamado de de . O Estimador de Kaplan-Meier é uma adaptação a empiríca que, na ausência de censura nos dados, é definida como:

é uma função que tem uma formato gráfico de escada com degraus nos tempos observados de falha de tamanho , onde é o tamanho amostral.

O processo utilizado até se obter a estimativa de Kaplan-Meier é um processo passo a passo, em que o próximo passo depende do anterior. De forma suscetível, para qualquer , pode ser escrito em termos de probabilidades condicionais. Suponha que existam pacientes no estudo e falhas distintas nos tempos . Considerando uma função discreta com probabilidade maior que zero somente nos tempos de falha , , tem-se que:

em que é a probabilidade de um indivíduo morrer no intervalo sabendo que ele não morreu até e considerando . Ou seja, pode se escrever como:

para .

A expressão geral do estimador de Kaplan-Meier pode ser apresentada após estas considerações preliminares, Formalmente, considere:

* , os tempos distintos e ordenados de falha;
* o número de falhas em , ;
* o número de indivíduos sob risco em , ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a .

Com isso, pode-se definir o estimador de Kaplan-Meier como:

De forma intuitiva, por assim dizer, a [Equação 2.3](#eq-ESTKaplanMeier) é proveniente da [Equação 2.1](#eq-DecomposeSt), sendo está, uma decomposição de em termos ’s. Assim, a [Equação 2.3](#eq-ESTKaplanMeier) é justificada se os ’s forem estimados por , que em palavras está expresso na [Equação 2.2](#eq-FormProbQj). No artigo original de 1958, Kaplan e Meier provam que a [Equação 2.3](#eq-ESTKaplanMeier) é um *estimador de máxima verossimilhança* para . Seguindo certos passos, é possível provar que que é um estimador de máxima verossimilhança de . Supondo que observações falham no tempo tempo , para , e observações são censuradas no intervalo , nos tempos . A probabilidade de falha no tempo é, então,

com , . Por outro lado, a contribuição para a função de verossimilhança de um tempo de sobrevivência censurado em para , é:

A função de verossimilhança pode, então, ser escrita como:

Com isso, é possível provar que que maximiza é exatamente a expressão dada pela [Equação 2.3](#eq-ESTKaplanMeier).

### 2.2.1 Propriedades do Estimador de Kaplan-Meier

Como um estimador de máxima verossimilhança, o estimador de Kaplan-Meier têm interessantes propriedades. As principais são:

* É não-viciado para grandes amostras;
* É fracamente consistente;
* Converge assintoticamente para um processo gaussiano.

A consistência e normalidade assintótica de foram provadas sob certas condições de regularidade, por Breslow e Crowley (1974) e Meier (1975) e, no artigo original Kaplan e Meier (1958) mostram que é o estimador de máxima verossimilhança, como já dito.

### 2.2.2 Variância do Estimador de Kaplan-Meier

Para que se possa construir intervalos de confiança e testar hipóteses para , se faz necessário ter conhecimento quanto variabilidade e precisão do estimador de Kaplan-Meier. Este estimador, assim como outros, está sujeito a variações que devem ser descritas em termos de estimações intervalares. A expressão assintótica do estimador de Kaplan-Meier é dada pela [Equação 2.4](#eq-VarKaplanMeier).

A expressão dada na [Equação 2.4](#eq-VarKaplanMeier), é conhecida como fórmula de Greenwood e pode ser obtida a partir de propriedades do estimador de máxima verossimilhança. Os detalhes da obtenção da ([Equação 2.4](#eq-VarKaplanMeier) estão disponíveis em Kalbfleisch e Prentice (1980, pag. 12-14).

Como , para um fixo, tem distribuição assintóticamente Normal. O intervalo de confiança com % de confiança para é expresso por:

Vale salientar que para valores extremos de , este intervalo de confiança pode apresentar limites que não condizem com a teoria de probabilidades. Para solucionar tal problema, aplica-se uma transformação em como, por exemplo, . Esta transformação foi sugerida por Kalbfleisch e Prentice (1980), tendo sua variância estimada por:

## 2.3 Outros Estimadores Não Parâmetricos

Texto a ser preenchido…

### 2.3.1 Estimador de Nelson-Aalen

Texto a ser preenchido…

### 2.3.2 Estimador da Tabela de Vida ou Atuarial

Texto a ser preenchido…

# 3. Técnicas Paramétricas - Modelos Probabilísticos

## 3.1 Introdução

## 3.2 Modelo Exponencial

### 3.2.1 Distribuição Exponencial

### 3.2.2 Simulações

Simularemos uma amostra aleatória proveniente de uma distribuição exponencial.

# -------------  
# [1] Simulação  
# -------------  
  
# Lambda verdadeiro  
lambda <- 0.5  
  
# Definindo semente para reprodutibilidade  
set.seed(123)  
  
# Definindo o tamanho da amostra  
n <- 1000  
  
# Simulando  
survival\_times <- rexp(n, rate = lambda)

Após simulação, é exibido o histogrma dos dados na [Figura 3.1](#fig-histExp)

# -------------------------  
# [2] Visualização do dados  
# -------------------------  
  
library(ggplot2)  
  
df <- data.frame(TIME = survival\_times)  
  
ggplot(data = df, aes(x = TIME)) +  
 geom\_histogram(bins = 20, fill = "gray", color = "black") +  
 labs(x = "Dados Simulados", y = "Frequência") +  
 theme\_minimal()

|  |
| --- |
| Figura 3.1: Histograma dos dados simulados a partir de uma Distribuição Exponecial |

## 3.3 Modelo Weibull

### 3.3.1 Distribuição Weibull