UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE BACHARELADO EM ESTATÍSTICA

MARIA KAROLINA DE FARIAS RAMOS

ANÁLISE DE SOBREVIVÊNCIA DE MULHERES DO ESTADO DE PERNAMBUCO DIAGNOSTICADAS COM CÂNCER DE MAMA

CAMPINA GRANDE - PB
2022

MARIA KAROLINA DE FARIAS RAMOS

ANÁLISE DE SOBREVIVÊNCIA DE MULHERES DO ESTADO DE PERNAMBUCO DIAGNOSTICADAS COM CÂNCER DE MAMA

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Tiago Almeida de Oliveira.

CAMPINA GRANDE - PB
2022

R175a Ramos, Maria Karolina de Farias.

Análise de sobrevivência de mulheres do estado de Pernambuco diagnosticadas com câncer de mama [manuscrito] / Maria Karolina de Farias Ramos. - 2022.

33 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2022.

"Orientação : Prof. Dr. Tiago Almeida de Oliveira , Coordenação do Curso de Estatística - CCT."

1. Neoplasia mamária. 2. Kaplan-Meier. 3. Curva de sobrevivência. I. Título

21. ed. CDD 616.994 49

Elaborada por Deivid A. M. Neves - CRB - 15/899 BC/UEPB

MARIA KAROLINA DE FARIAS RAMOS

ANÁLISE DE SOBREVIVÊNCIA DE MULHERES DO ESTADO DE PERNAMBUCO DIAGNOSTICADAS COM CÂNCER DE MAMA

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Trabalho aprovado em 01 de dezembro de 2022.

BANCA EXAMINADORA

Prof. Dr. Tiago Almeida de Oliveira
(Orientador)
Universidade Estadual da Paraíba (UEPB)

Prof. Me. Cleanderson Romualdo Fidelis
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Marcelino Alves Rosa de Pascoa
Universidade Federal de Mato Grosso (UFMT)

Dedico este trabalho ao meu Senhor Deus e minha amada família por estarem presentes em todo momento nessa trajetória acadêmica.

AGRADECIMENTOS

Acima de tudo agradeço a Deus por atender minhas orações durante todo esse período de faculdade para que tudo desse certo.

Aos meus pais Inácio Ramos e Maria do Carmo Farias por serem minha fonte de inspiração e estarem sempre presentes me apoiando nos momentos que mais precisei. E por mostrar que a Educação é uma das mais importantes formas de se construir um futuro promissor.

Aos meus irmãos (melhores amigos) Izaias, Izabel, Daniel, Moisés, Bernadete, Ezequiel, Beatriz e Euflaudizia pelo incentivo e companheirismo de sempre. Vocês são incríveis demais.

A minha "Família Ramos" pelas ótimas conversas e momentos de distração.

Aos meus avós maternos Sebastião Gonçalves e Terezinha Farias por sempre me abençoar ao sair de casa deixando meus dias mais felizes e motivadores.

Aos meus amigos e colegas de classe da turma "2017.2", em especial, Clevia Bento, Joseferson Barreto, Emanuela Rodrigues e Fernanda Lima pelo compartilhamento de conhecimentos e apoio durante todo o curso.

Ao professor e orientador Dr. Tiago Almeida de Oliveira por aceitar o convite de me guiar nesse trabalho repassando ensinamentos e instruções fundamentais.

Aos membros da banca, professor Me. Cleanderson Romualdo Fidelis por ministrar a disciplina Análise de Sobrevivência de forma clara e objetiva me despertando interesse de buscar mais conhecimentos sobre a mesma e ao professor Dr. Marcelino Alves Rosa de Pascoa que disponibilizou seu tempo participando desse momento importante para mim.

À Prefeitura Municipal de Parari, na pessoa do Secretário de Transportes André por toda assistência.

À Universidade Estadual da Paraíba que desde março de 2018 tem feito parte da minha vida, me oferecendo uma oportunidade incrível de poder concluir o curso superior de Bacharelado em Estatística.

A todos que contribuíram de uma forma ou de outra para que eu pudesse concluir mais uma etapa da minha vida.

“Deem graças ao Senhor, porque Ele é bom.
O seu amor dura para sempre!”
(Salmos 136:1)

RESUMO

A análise de sobrevivência é uma área da estatística cujo objetivo é avaliar o tempo até a ocorrência de um evento de interesse. Esse é um dos ramos da estatística mais procurados nos últimos anos. Dentre as áreas exploradas, o setor clínico tem se destacado com análises de fatores que levam a remissão. Desse modo, o objetivo deste trabalho é avaliar fatores de risco que afetam o tempo de mulheres diagnosticadas por câncer de mama vir a óbito por esse motivo. Trata-se de uma população de 2.337 mulheres do estado de Pernambuco diagnosticadas com câncer de mama entre os anos de 1996 a 2017. Cerca de 7,06% das observações foram censuradas, sendo essas, observações parciais da resposta ocorrendo por diversos motivos, dentre esses, a perda de informações ou a não ocorrência do falecimento por câncer de mama. E 92,94% obtiveram a falha, ou seja, faleceram por câncer de mama. Foi utilizado o estimador de Kaplan-Meier como técnica não-paramétrica para uma análise descritiva inicial dos dados. Por esta, foi possível concluir que o risco de vir a óbito por câncer de mama cresce com o passar dos meses e caso a mulher seja diagnosticada com um tumor em metástase e independente disso possua uma idade numa faixa etária de 81 anos ou mais, o risco é ainda maior em relação às demais categorias de ambas. O teste de *logrank* apontou que apenas entre os grupos: faixas etárias "21 a 40 $\times$ 41 a 60", meios de diagnósticos "h. do tumor primário $\times$ citologia" e meios de diagnósticos "h. da metástase $\times$ clínico" não houve diferenças significativas entre as curvas de sobrevivência. O modelo que melhor representou os dados para um ajuste geral foi o modelo paramétrico pela distribuição exponencial. As análises gráficas dos resíduos de Cox-Snell e a comparação de curvas de sobrevivências evidenciaram isto, além da comprovação pelo teste da razão de verossimilhança. Já para os modelos de regressão paramétricos ajustados pelas covariáveis extensão, faixa e meio de diagnóstico, a distribuição weibull foi a melhor ajustada.

Palavras-chaves: Censura. Neoplasia Mamária. Kaplan-Meier.

ABSTRACT

Survival analysis is an area of statistics whose objective is to evaluate the time until the occurrence of an event of interest. This is one of the most sought-after branches of statistics in recent years. Among the areas explored, the clinical sector has stood out with analyzes of factors that lead to remission. Thus, the objective of this study is to evaluate risk factors that affect the time that women diagnosed with breast cancer die for this reason. This is a population of 2.337 women from the state of Pernambuco diagnosed with breast cancer between the years 1996 to 2017. About 7,06% of the observations were censored, these being partial observations of the response occurring for several reasons, among them, the loss of information or the non-occurrence of death from breast cancer. And 92,94% failed, that is, they died of breast cancer. The Kaplan-Meier estimator was used as a non-parametric technique for an initial descriptive analysis of the data. For this reason, it was possible to conclude that the risk of dying from breast cancer grows over the months and if the woman is diagnosed with a tumor in metastasis and, regardless of this, is aged 81 years or older, the risk is even greater in relation to the other categories of both. The logrank test showed that only between the groups: age groups "21 to 40 $\times$ 41 to 60", means of diagnosis "h. of the primary tumor $\times$ cytology" and means of diagnosis "h. of metastasis $\times$ clinical" there were no significant differences between the survival curves. The model that best represented the data for a general adjustment was the parametric model by the exponential distribution. The graphical analyzes of the Cox-Snell residuals and the comparison of survival curves showed this, in addition to confirmation by the likelihood ratio test. As for the parametric regression models adjusted by the covariates extension, range and means of diagnosis, the weibull distribution was the best fit.

Keywords: Censorship. Breast Neoplasm. Kaplan-Meier.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Também conhecido como neoplasia mamária, o câncer de mama é uma doença causada pela multiplicação de células anormais na mama (INCA, 2019). Este está presente tanto em homens quanto em mulheres, porém, considerado raro em homens (apenas 1% dos casos), é o tumor mais comum entre as mulheres após o câncer de pele e o que causa mais morte (INCA, 2022).

Estima-se que em 2020 no Brasil o câncer de mama em mulheres foi o mais incidente, com aproximadamente 88.492 casos, representando 41,3% de todos os tipos de câncer. O Brasil é considerado o país onde mais ocorrem óbitos por essa causa, registrando 20.725, ou seja, 24,7% dentre os tipos de câncer (IARC, 2020).

Os sintomas mais comuns identificados são, na maioria dos casos, por meio de nódulos palpáveis, secreção aquosa ou sanguinolenta pelo mamilo, dores regionais e edema (inchaço) na pele. Segundo Buzaid, Maluf e Gagliato (2020), em relação ao câncer de mama localizado (precoce), geralmente a mulher não apresenta nenhum sintoma, podendo notar por vezes a presença de um nódulo indolor. Já o tumor metastático tem um poder mais agressivo ocupando grande parte da mama, este, podendo haver dores e infiltrações na pele e nos músculos peitorais.

Um fator de risco é algo que tem demonstrado ser de grande influência no que se trata ao surgimento do câncer, e na neoplasia mamária não é diferente. Quanto mais desses fatores a pessoa possuir, mais provável será de desenvolver a doença. Buzaid, Maluf e Gagliato (2020) relatam que a síndrome genética que é uma deficiência genética rara conhecida como mutação em dois genes é um fator dos portadores desse gene que aumenta o risco de desenvolver a doença em mais de 50%. Dentre outras causas, podemos citar também: a presença da doença em familiares próximos, mais especificamente, de primeiro grau; o diagnóstico em pessoas mais idosas; consumo de bebidas alcoólicas e, etc.

A melhor maneira de evitar o diagnóstico positivo para esta classe de câncer é a prevenção. E além dos cuidados sob o controle de estilo de vida social, bem como a alimentação saudável e a prática de exercícios físicos, Thuler (2003) destaca duas estratégias: a educação para promover o diagnóstico precoce e o rastreamento. Sendo que o rastreamento, através da mamografia, corresponde a uma diminuição em 25% da mortalidade por câncer de mama.

Dentre vários tipos histológicos desta doença, destacam-se alguns: carcinoma ductal (mais frequente), carcinoma lobular infiltrativo, carcinoma adenoide-cístico (mais raro), carcinoma mucinoso e carcinoma metaplásico (BUZAID; MALUF; GAGLIATO, 2020). Desses, o carcinoma ductal e o carcinoma lobular infiltrativo são respectivamente os tipos mais comuns. A neoplasia mamária pode ser diagnosticada por exames de imagens como ultrassons, mamografias e também por exames clínicos. É importante destacar que a confirmação diagnóstica é exclusivamente obtida mediante a biópsia.

Ferraz e Filho (2017), em seu estudo, mostraram que a estimativa de sobrevivência através do estimador de Kaplan-Meier resultou em 79,7% nos primeiros cinco anos após o

diagnóstico, 68,9% aos dez anos e 60,8% ao final do estudo. Nesse estudo foram diagnosticadas 524 mulheres com câncer de mama no período de 1993 a 1995 com idades entre 25 e 93 anos, sendo que aproximadamente 64% das mulheres possuíam 50 anos ou mais.

Em Dias, Martins e Gradim (2018), um estudo envolvendo 62 mulheres num período de cinco anos apontou que 61,29% tiveram câncer de mama possuindo de 50 a 69 anos e 35,49% foram detectadas por metástase. A estimativa por Kaplan-Meier foi de 80% de sobrevivência após o diagnóstico. A análise mostrou que dado que a paciente possui metástase, a probabilidade de sobreviver é significativamente menor.

Diante dessas informações, destaca-se a importância de tornar público os principais fatores que colocam em risco a vida de mulheres diagnosticadas com câncer de mama. E consequentemente, diante desse conhecimento, auxiliar as autoridades responsáveis (neste caso, área da saúde) na tomada de decisão no que tange ao tratamento e precauções relativos à mulheres com diagnóstico positivo para o câncer de mama.

2 MATERIAL E MÉTODOS

Nesta seção serão apresentados brevemente os recursos para elaboração deste trabalho. Neste sentido, falaremos da obtenção e descrição do conjunto de dados utilizado, dos métodos estatísticos que compõem as análises do nosso estudo e a utilidade que cada um contribui para o desenvolvimento do trabalho.

2.1 Material

Para a análise, foi utilizado um banco de dados de acesso livre e gratuito do Instituto Nacional de Câncer (INCA[1]) referindo-se a mulheres do estado de Pernambuco diagnosticadas com câncer de mama. O estudo é iniciado em 1996 quando ocorre o primeiro diagnóstico e finalizado em 2017. O conjunto de dados é formado de 2.337 observações e 5 variáveis: faixa etária, meio de diagnóstico, extensão, tempo em meses e status.

O tempo em meses foi definido pela diferença da data de óbito e a data de diagnóstico. Os tempos considerados censura à esquerda foram excluídos do banco de dados, sendo então trabalhado apenas o mecanismo de censura à direita, com o tipo de censura aleatória ocorrendo quando a paciente sai do estudo sem ter ocorrido a falha, ou seja, é retirada do estudo antes de um exato período (MARTINS; WERNER, 2010).

O ambiente computacional estatístico utilizado para a análise dos dados foi o *software R* Team (2022), um programa de código livre e gratuito. O pacote utilizado para tratar os dados de sobrevivência foi o *survival* (THERNEAU et al., 2022). As variáveis utilizadas estão descritas na Tabela 1.

Tabela 1 – Classificação das variáveis utilizadas para os dados de câncer de mama.

Variável	**Classificação**
Faixa etária	21 a 40; 41 a 60; 61 a 80; 81 ou mais
Meio de diagnóstico	Histologia do tumor primário; histologia da metástase; SDO; pesquisa; citologia; clínico; sem informações
Extensão	Localizado; metástase; sem informações
Tempo em meses	De 1 a 229
Status	Censura = 0; óbito = 1

Fonte: Produzida pelo autor (2022).

Na variável “Status”, a falha que designa que o indivíduo faleceu por câncer de mama será representada por “1” e censura; indivíduo não faleceu por câncer de mama ou não se sabe informação será representada por “0”.

[1] https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/registros/base-populacional

2.2 Métodos

Para o conhecimento dos dados, foi realizada uma análise descritiva do ponto de vista geral e também voltado a conceitos em análise de sobrevivência. A partir disso, foram abordadas técnicas de estimação por modelo não-paramétrico utilizando o estimador de Kaplan-Meier para estimar as funções de sobrevivência e o teste para comparação de curvas, logrank. Em seguida, modelos paramétricos, com o objetivo também de estimar funções de sobrevivências, porém, considerando distribuições de probabilidades. Para estas técnicas paramétricas, utilizou-se o teste de razão de verossimilhança para verificar a adequação dos modelos. Por fim, para verificar a possível influência de variáveis em relação ao tempo de falha foram aplicado os modelos de regressão paramétricos.

2.3 Conceitos Básicos de Análise de Sobrevivência

A análise de sobrevivência consiste em uma área da Estatística que visa analisar o tempo até a ocorrência de determinado evento, tempo este que representa o evento de interesse. Esse tempo T é ainda designado por **tempo de falha**. Segundo Colosimo e Giolo (2006), além dos tempos de falhas, os dados de sobrevivência são compostos também, na maioria das vezes, pelas censuras (termo abordado na subseção *2.3.1*). Com essas duas informações, formamos então a variável resposta.

As técnicas da análise de sobrevivência podem ser aplicadas as mais variadas áreas do conhecimento, tal como na saúde, avaliando o tempo de cura de pacientes diagnosticados com determinado tipo de doença; na indústria avaliando o tempo até o defeito de um equipamento; nas ciências sociais estimando o número de crianças nascidas, divórcios registrados e, etc.

2.3.1 Censura

A principal característica dos dados de sobrevivência é a presença de censuras (PEREIRA; VIVANCO, 2003). A censura é uma observação parcial da resposta, podendo ocorrer por diversos motivos. Nos casos clínicos, por exemplo: na perda de acompanhamento do paciente; falecimento por outro motivo diferente do estudado; o paciente chega ao fim do estudo sem que o mesmo apresente o evento de interesse.

2.3.1.1 Tipos de Censura

Nos estudos clínicos há três tipos de censuras que são mais comuns:

- Censura do tipo I - O estudo termina após um tempo pré-estabelecido.
- Censura do tipo II - O estudo é encerrado após ocorrer uma quantidade pré-estabelecida de falhas no evento de interesse.

- Censura aleatória - Ocorre com frequência na área médica; o indivíduo sai do estudo sem ter ocorrido o evento de interesse.

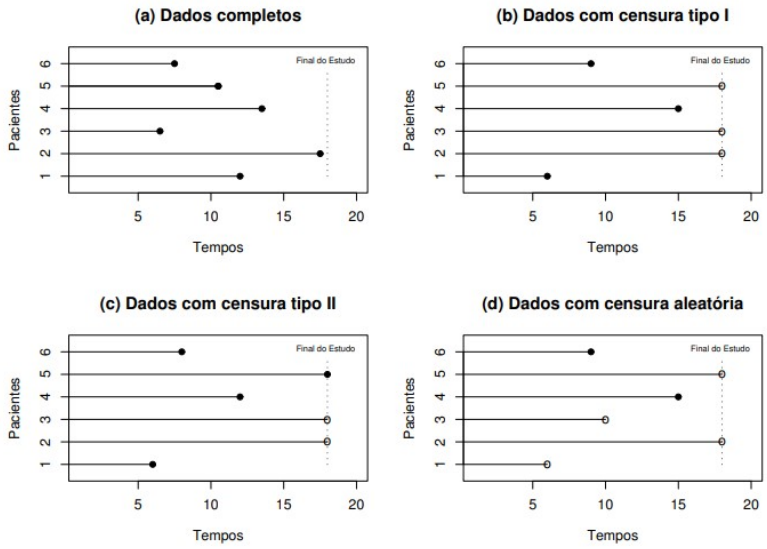
Segundo Strapasson (2007), o mecanismo de censura pode ser classificado em censura à direita, censura à esquerda e censura intervalar. A censura à direita é a mais utilizada na qual o tempo de ocorrência do evento de interesse está à direita do tempo registrado. A censura à esquerda é aquela em que o indivíduo ou objeto já experimenta o evento de interesse no início do estudo. E a censura intervalar é aquela em que não se sabe o tempo exato em que a falha ocorreu, sabe-se apenas que se deu em um intervalo de tempo.

Para a análise dos dados de sobrevivência, os tempos dos indivíduos observados t_i sendo t o tempo e i $(i = 1, \ldots, n)$ os indivíduos observados, a variável indicadora de falha ou censura δ_i é representada da seguinte forma:

$$\delta_i = \begin{cases} 1, & \text{se } t_i \text{ é um tempo de falha} \\ 0, & \text{se } t_i \text{ é um tempo de censura} \end{cases}$$

A Figura 1 representa o mecanismo de censura à direita em que (a) todos os pacientes experimentaram o evento antes do final do estudo; (b) alguns pacientes não experimentaram o evento até o final do estudo; (c) o estudo foi finalizado após a ocorrência de um número pré-estabelecido de falhas e (d) o acompanhamento de alguns pacientes foi interrompido por alguma razão e alguns pacientes não experimentaram o evento até o final do estudo.

Figura 1 – Representação dos mecanismos de censura em que (•) indica a falha e (○) indica censura.



Fonte: (COLOSIMO; GIOLO, 2006)

2.3.2 Função de Sobrevivência

A função de sobrevivência é uma das funções mais utilizadas no estudo da análise de sobrevivência. Ela é definida como a probabilidade de um evento não ocorrer acima de um tempo t, ou seja, é a probabilidade de uma observação sobreviver além de t. É definida por:

$$S(t) = P(T > t).$$

Note que, a função de sobrevivência pode ser obtida em termos da função de distribuição acumulada. Nesse contexto, a função de distribuição acumulada pode ser entendida como a probabilidade de uma observação não sobreviver ao tempo t. A função é definida por:

$$F(t) = 1 - S(t).$$

2.3.3 Função de Taxa de Falha ou de Risco

A função taxa de falha ou de risco $\lambda(t)$ representa a taxa instantânea do indivíduo sofrer o evento num intervalo de tempo $(t, t+\Delta t)$, dado que ele sobreviveu até o tempo t. Seja T uma variável aleatória que corresponde o tempo até a ocorrência de um evento, a função é então definida como:

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

2.3.4 Função de Taxa de Falha Acumulada

Outra função importante é a função de taxa de falha acumulada, ela fornece a soma de todas as taxas de falhas $\lambda(u)$ dos indivíduos até o tempo t, propriamente dita, a taxa de falha acumulada. Esta é uma função que não possui uma interpretação direta, mas fornece informação no que se refere à função taxa de falha. É dada por:

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

2.4 Técnicas Não-Paramétricas

Assim como nas análises básicas de estatística, na análise de sobrevivência a estatística descritiva é fundamental para obter informações de medidas de tendência central como o tempo médio, mediano e medidas de variabilidade tal como o desvio padrão e variância. Na análise de sobrevivência, portanto, o fato de existir censuras torna esse tipo de estudo inadequado (tendencioso), a partir disso surgiu a necessidade de estudar dados na presença de censuras.

As técnicas não-paramétricas são ideais para esse tipo de estudo. Têm por finalidade estimar a função de sobrevivência sem utilizar nenhuma distribuição de probabilidade. Há dois

estimadores não-paramétricos mais utilizados: Kaplan-Meier e Nelson-Aalen. Neste trabalho veremos apenas o estimador de Kaplan-Meier.

2.4.1 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier foi proposto por Kaplan e Meier (1958), é o estimador mais utilizado em estudos clínicos. Esse modelo é aplicado por ser não viciado para amostras grandes e também por permitir a estimativa no tempo mesmo possuindo casos censurados. Pelo estimador de Kaplan-Meier é possível comparar os tempos distintos de falhas através da curva de sobrevivência. Sua função de sobrevivência estimada é uma função "escada" pelo qual os "degraus" correspondem aos tempos distintos de falhas observados. Considerando:

- $t_1 < t_2 < \ldots < t_k$, 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 por:

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

A seguir, Breslow e Crowley (1974) destacam algumas propriedades desse estimador:

- é fracamente consistente;
- converge assintoticamente para um processo gaussiano e
- é estimador de máxima verossimilhança de $S(t)$.

2.4.2 Comparação de Curvas de Sobrevivência

A comparação de curvas de sobrevivência tem sido muito procurada principalmente na área médica. Estudos como este possibilitam a comparação entre diversas categorias de uma única variável utilizando as curvas de sobrevivência de uma técnica não-paramétrica. Na área médica, como foi mencionado, o interesse principal é avaliar se dois ou mais tratamentos são estatisticamente iguais.

2.4.2.1 Logrank

O teste de logrank proposto por Mantel (1966) foi o mais utilizado nesse estudo para comparar curvas de sobrevivência. Este teste é particularmente apropriado quando a razão das funções de risco dos grupos a ser comparados é aproximadamente constante.

Hipóteses para o teste:

$$\begin{cases} H_0: \text{ não há diferença entre as curvas de sobrevivência.} \\ H_1: \text{ há diferença entre as curvas de sobrevivência.} \end{cases}$$

Suponha uma comparação entre duas curvas de sobrevivência $S_1(t)$ e $S_2(t)$. Considere ainda como $t_1 < t_2 < \ldots < t_k$ sendo os tempos distintos de falha obtidos pela combinação de duas amostras, d_j o número de falhas, n_j o número de indivíduos sob risco inferior a d_j na amostra combinada e respectivamente d_{ij} e d_{ij} na amostra $i; i = 1,2$ e $j = 1,\ldots,k$. A estatística de teste logrank é dada por:

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2},$$

sendo que para cada tempo distinto, d_{2j} representa o número observado de falhas no grupo dois, $w_{2j} = n_{2j}d_j n_j^{-1}$ corresponde a média de falhas para o grupo dois e $(V_j)_2$ as variâncias para o mesmo grupo. O teste tem uma distribuição qui-quadrado com 1 grau de liberdade e é baseado na hipótese nula, ou seja, $H_0 : S_1(t) = S_2(t)$.

2.5 Modelos Paramétricos

O modelo paramétrico é uma alternativa de estimação quando se considera distribuições de probabilidades. Dentre vários motivos para aplicação, o principal deles é entender melhor o comportamento de um determinado fenômeno por meio de uma relação estabelecida entre variáveis que são passíveis de exercer determinada influência sob outra, seja esta de impacto negativo ou positivo.

Considerando que para esse trabalho, o evento de interesse é representar o tempo até a falha, nas subseções *2.5.1*, *2.5.2* e *2.5.3* serão apresentadas três distribuições mais apropriadas para esse caso: exponencial, weibull e log-normal respectivamente. A escolha do modelo deve ser feita com base na distribuição de probabilidade que apresenta melhor adequabilidade em relação aos dados que estão sendo tratados, caso contrário, os resultados das análises não serão confiáveis.

2.5.1 Distribuição Exponencial

A distribuição exponencial é o modelo mais simples para descrever o tempo de falha, e tem como característica principal a função taxa de falha constante. Isto significa que todos os indivíduos ou objetos que não falharam, o risco se torna o mesmo em todo tempo. Essa propriedade é conhecida como falta de memória da distribuição exponencial. Considerando T a variável aleatória de tempo até a falha, algumas funções importantes são destacadas abaixo:

- Função densidade de probabilidade

$$f(t) = \frac{1}{\alpha} exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, \quad t \geq 0, \quad \alpha \geq 0.$$

- Função de sobrevivência

$$S(t) = exp\left\{-\left(\frac{t}{\alpha}\right)\right\}.$$

- Função taxa de falha

$$\lambda(t) = \frac{1}{\alpha} \quad \text{para} \quad t \geq 0,$$

sendo α o tempo médio de vida.

2.5.2 Distribuição Weibull

A distribuição weibull é bastante utilizada por biomédicos devido à propriedade da sua função de taxa de falha ser monótona, ou seja, esta função é constante, crescente ou decrescente. Considerando T a variável aleatória com distribuição de weibull do tempo até a falha, tem-se:

- Função densidade de probabilidade

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{-1} exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}, \quad t \geq 0,$$

em que γ é o parâmetro de forma e α o de escala.

- Função de sobrevivência

$$S(t) = exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}.$$

- Função taxa de falha

$$\lambda(t) = \frac{\gamma}{\alpha^{\gamma}} t^{-1},$$

sendo t, α e $\lambda \geq 0$.

2.5.3 Distribuição Log-Normal

A distribuição log-normal está atrelada aos tempos de vida de produtos e também é usada para determinar o tempo de vida de pacientes com determinada doença. As taxas de falha crescem, atingem um valor máximo e logo depois decrescem. Considerando T a variável aleatória com distribuição log-normal, tem-se:

- Função densidade de probabilidade

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} exp\left\{-\frac{1}{2}\left(\frac{log(t)-\mu}{\sigma}\right)^2\right\}, \quad t \geq 0,$$

em que o parâmetro μ é a média do logaritmo do tempo de falha e σ é o desvio-padrão.

- Função de sobrevivência

$$S(t) = \phi\left(\frac{-log(t)+\mu}{\sigma}\right),$$

sendo $\phi(\cdot)$ a função de distribuição acumulada de uma normal padrão.

- Função taxa de falha

$$\lambda(t) = \frac{f(t)}{S(t)}.$$

2.5.4 Modelos Paramétricos de Regressão

Nesta subseção serão apresentados três modelos paramétricos de regressão das distribuições citadas nas subseções *2.5.1*, *2.5.2* e *2.5.3* respectivamente. Estes são utilizados para estimar o tempo de sobrevivência de objetos ou pessoas em estudo. Considere β' como sendo os parâmetros a serem estimados e x a covariável. Considere ainda que T segue a distribuição associada. Os modelos são dados da seguinte forma:

$$\text{Exponencial: } T = exp\{\beta_0 + \beta_1 x\} + \varepsilon,$$

$$\text{Weibull: } Y = log(T) = \beta_0 + \beta_1 x + \sigma\nu$$

e

$$\text{Log-normal: } Y = log(T) = \beta_0 + \beta_1 x + \mu\sigma.$$

2.5.5 Adequação dos Modelos Ajustados

Em geral, é de suma importância a verificação da qualidade dos ajustes dos modelos utilizados. Por meio desta, pode-se observar a ocorrência de falhas que eventualmente são capazes de tornar o modelo inútil para os dados estudados e, além disso, essa verificação tem também a característica de nos auxiliar na melhor escolha do modelo. Para os modelos paramétricos de regressão, um meio de verificar esse fato é através de uma análise de resíduos. E para isso, Cox e Snell (1968) empregaram uma técnica gráfica essencial onde é possível examinar o ajuste geral do modelo como é o caso aplicado neste trabalho.

Os resíduos são determinados por:

$$\hat{e_i} = \hat{\Lambda}(t_i \mid x_i),$$

em que $\hat{\Lambda}(\cdot)$ é a função de risco acumulado obtida do modelo ajustado.

Para os modelos vistos na subseção *2.5.4*, os resíduos de Cox-Snell são definidos por:

$$\text{Exponencial: } \hat{e_i} = \left[t_i exp\{-x_i'\hat{\beta}\} \right],$$

$$\text{Weibull: } \hat{e_i} = \left[t_i exp\{-x_i'\hat{\beta}\} \right]^{\hat{\gamma}}$$

e

$$\text{Log-normal: } \hat{e_i} = -log\left[1 - \phi\left(\frac{log(t_i) - x_i'\hat{\beta}}{\hat{\sigma}} \right) \right].$$

Conforme Lawless (1982), os resíduos seguem uma distribuição exponencial caso o modelo seja adequado, isto é, apresentem estimativas que mais se aproximam dos verdadeiros valores. Nesse contexto, as estimativas serão comparadas com base nas mesmas de Kaplan-Meier.

Para a seleção dos modelos de regressão paramétricos ajustados por covariáveis foi utilizado o Critério de Informação de Akaike (AIC). O AIC é uma das formas mais utilizadas para verificar a qualidade de um modelo. Esta é uma métrica interpretada de maneira comparativa, ou seja, quanto menor for essa métrica em relação as demais, mais adequado o modelo está.

2.5.6 Estimação dos Parâmetros dos Modelos

O procedimento de estimação dos parâmetros dos modelos é feito através de estatísticas inferenciais. É um método pelo qual são estimados parâmetros de certas distribuições de probabilidades em torno uma amostra representativa, sendo que esses passem a representar dados populacionais (que possuem parâmetros desconhecidos) de onde foi retirada a amostra. Para este caso, será utilizado o método de máxima verossimilhança.

2.5.6.1 Método da Máxima Verossimilhança

Segundo Colosimo e Giolo (2006), o método de máxima verossimilhança consiste em obter estimadores para os parâmetros, mais detalhadamente, consiste em obter estimativas mais verossímeis (verdadeiras) dentro de uma amostra para o parâmetro populacional desconhecido. Supondo uma amostra de observações $t_i, \ldots, t_n$ não censuradas de certa população e que essa população é caracterizada pela função densidade $f(t) = (1/\alpha)exp(-t/\alpha)$ significa que as observações vêm de uma distribuição exponencial com parâmetro α a ser estimado. A função de verossimilhança para um parâmetro genérico θ desta população é dada por:

$$L(\theta) = \prod_{i=1}^{n} f(t_i; \theta),$$

em que θ é o valor procurado que maximize a função $L(\theta)$.

Para cada tipo de censura específica é obtida uma função de verossimilhança distinta. Para esse caso, de censura à direita, que engloba as censuras do tipo aleatória, tipo I e tipo II, a expressão para todos os mecanismos de censura é dada por:

$$L(\theta) = \prod_{i=1}^{r} f(t_i; \theta) \prod_{i=r+1}^{n} S(t_i; \theta),$$

onde as observações são divididas em dois conjuntos: as r primeiras ordenadas que são as não-censuradas $(1, 2, \ldots, r)$ e as $n - r$ seguintes são as censuradas $(r+1, r+2, \ldots, n)$.

2.5.6.2 Teste da Razão de Verossimilhanças

O Teste da Razão de Verossimilhança (TRV) pode ser usado para verificar a adequação do modelo, visando comparar modelos através dos valores do logaritmo da função de verossimilhança maximizada. Portanto, testamos as seguintes hipóteses:

$$\begin{cases} H_0: \text{ o modelo de interesse é adequado.} \\ H_1: \text{ o modelo de interesse não é adequado.} \end{cases}$$

Colosimo e Giolo (2006) pontuam, que no contexto de análise de sobrevivência, esse teste é usualmente realizado utilizando a distribuição gama generalizada que apresenta o modelo exponencial, weibull, log-normal, dentre outros modelos como casos particulares. A estatística de teste é dada por:

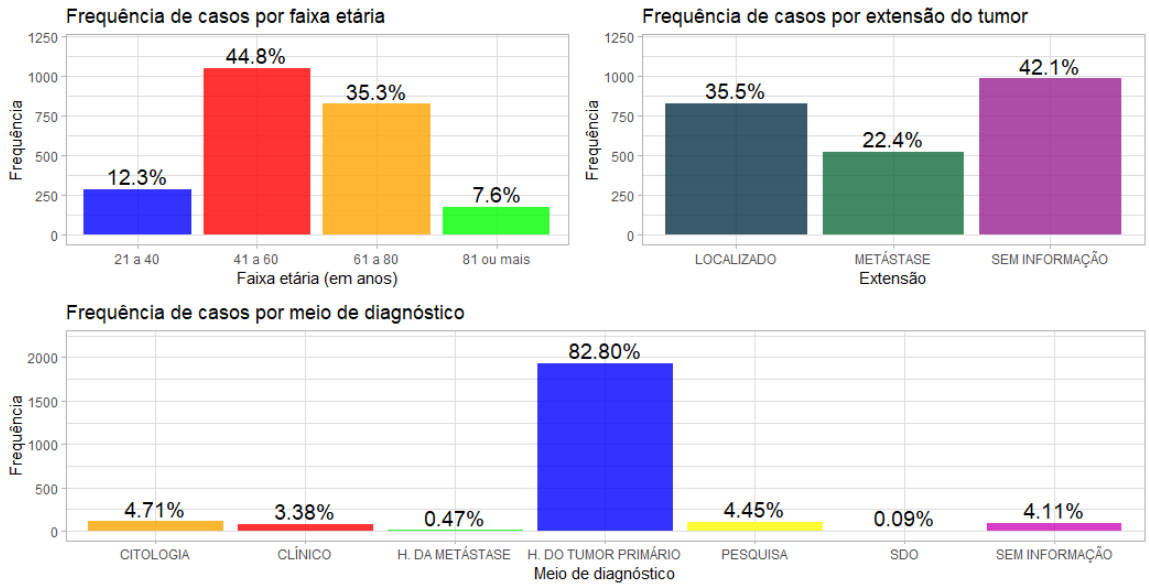
$$TRV = -2log\left[\frac{L(\hat{\theta}_M)}{L(\hat{\theta}_G)}\right] = 2[logL(\hat{\theta}_G) - L(\hat{\theta}_M)],$$

onde $logL(\hat{\theta}_G)$ é o logaritmo de verossimilhança do modelo generalizado e $logL(\hat{\theta}_M)$ o logaritmo de verossimilhança dos modelos de interesse (exponencial, weibull e log-normal). Esse teste segue uma distribuição qui-quadrado com os graus de liberdade definidos pela diferença entre a quantidade de parâmetros das distribuições comparadas.

3 RESULTADOS E DISCUSSÃO

Os dados foram observados no período de 1996 a 2017. Os resultados foram obtidos de uma amostra de 2.337 mulheres diagnosticadas com câncer de mama residentes do estado de Pernambuco. Cerca de 93% delas faleceram por câncer de mama. A princípio foi realizada uma análise descritiva dos dados (Figura 2), onde foi possível constatar que aproximadamente 12% das mulheres diagnosticadas estão numa faixa etária de 21 a 40 anos, 45% de 41 a 60 anos, 35% de 61 a 80 anos e 8% de 81 ou mais anos. Considerando a gravidade do tumor, cerca de 35% foram diagnosticadas no grau de tumor localizado (precoce); 22% no grau mais avançado do câncer, a metástase e não se sabe informação de 42%. Verificou-se também que aproximadamente 83% foram diagnosticadas pela histologia do tumor primário.

Figura 2 – Frequência de casos de câncer de mama por faixa etária, extensão do tumor e meio de diagnóstico respectivamente.



Fonte: Produzida pelo autor (2022).

Apresentado na Tabela 2 foi verificado o percentual de censuras e óbitos para algumas covariáveis em análise. Em relação a covariável extensão das pessoas diagnosticadas com câncer de mama que possuíam um tumor localizado, 89,02% faleceram de câncer de mama e aquelas que possuíam um tumor em metástase, 94,65% faleceram pelo mesmo motivo. Na covariável faixa etária, 95,52% das mulheres que tinham idades de 41 a 60 anos falharam e 88,14% das que tinham 81 anos ou mais falharam. No que se refere ao meio de diagnóstico, todas as mulheres que foram diagnosticadas por meio de análise das células (citologia), obtiveram o evento de interesse.

Através das estimativas de Kaplan-Meier, a probabilidade de sobreviver tende a cair variando de 95% a zero. Foi observado, com auxílio do *software R*, que no primeiro mês 105

Tabela 2 – Percentual de censuras e óbitos para as covariáveis em estudo dentro de cada categoria.

Variável	**Classificação**	**Censuras + Óbitos**	**Censuras (em %)**	**Óbitos (em %)**
Extensão	1: Localizado	829	10,98%	89,02%
	2: Metástase	523	5,35%	94,65%
	3: Sem informação	985	4,67%	95,33%
Faixa etária	1: 21 a 40	288	3,12%	96,87%
	2: 41 a 60	1.048	4,48%	95,52%
	3: 61 a 80	824	10,68%	89,32%
	4: 81 ou mais	177	11,86%	88,14%
Meio de diagnóstico	1: Citologia	110	0%	100%
	2: Clínico	79	2,53%	97,47%
	3: Histologia da metástase	11	9,09%	90,91%
	4: Histologia do tumor primário	1.935	7,70%	92,30%
	5: Pesquisa	104	9,62%	90,38%
	6: Somente por declaração de óbito - SDO	2	0%	100%
	7: Sem informação	96	3,13%	96,87%

Fonte: Produzida pelo autor (2022).

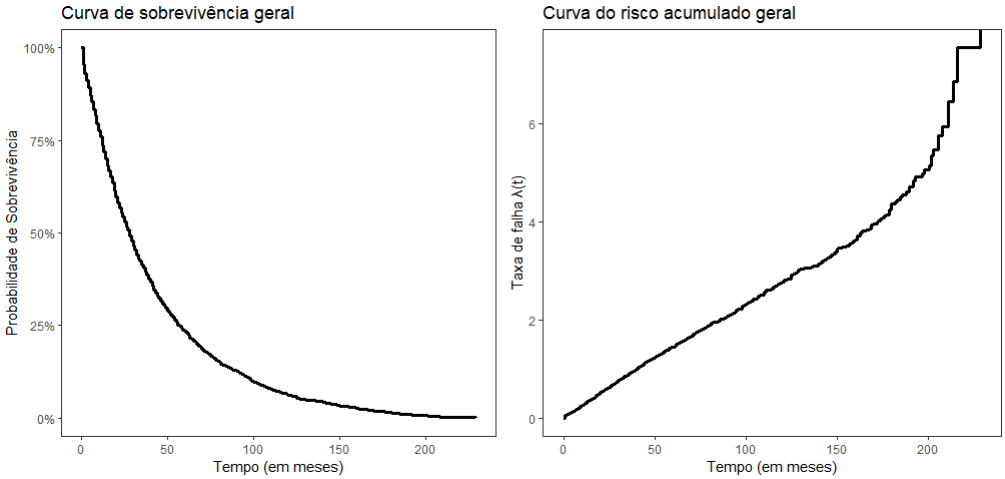
pessoas faleceram por câncer mamário. O tempo mediano de sobrevivência baseado nessa estimativa de Kaplan-Meier foi de 28 meses, indicando que 50% daqueles que falharam, faleceram antes do 28º mês e 50% depois. Nesse 28º mês somam-se 1.115 pacientes sob risco com 0,49 de probabilidade de sobrevivência. Pode-se observar que, o risco de ocorrência da falha, isto é, o risco de morte, aumenta consideravelmente ao decorrer do tempo. A Figura 3 ilustra tais características.

Fazendo a análise das covariáveis, algumas informações importantes foram observadas. As pacientes têm maior probabilidade de sobreviver quando o tumor é detectado precocemente. No 25º mês de acompanhamento, por exemplo, a probabilidade de sobreviver é de 0,61 quando a doença é localizada diferente da metástase (0,44 de probabilidade). O gráfico "Curvas de sobrevivência por extensão" na Figura 4 mostra que a linha de cor rosa representando o tumor localizado sempre está superior às outras. Em relação ao risco, quanto mais alto estiver a curva, maior o risco de morte por câncer de mama.

Para a covariável faixa etária, a probabilidade de sobreviver possuindo de 21 a 40 anos foi maior até o 28º mês. A partir desse mês, pessoas com faixa etária de 41 a 60 anos tiveram uma probabilidade de sobrevivência maior comparada às demais curvas. O risco é superior para as pessoas mais idosas.

Na maioria do tempo, pessoas diagnosticadas através da pesquisa tiveram uma probabilidade de sobrevivência superior, como ilustra o gráfico "Curvas de sobrevivência por meio do diagnóstico". As curvas de sobrevida da histologia da metástase e SDO não seguem um

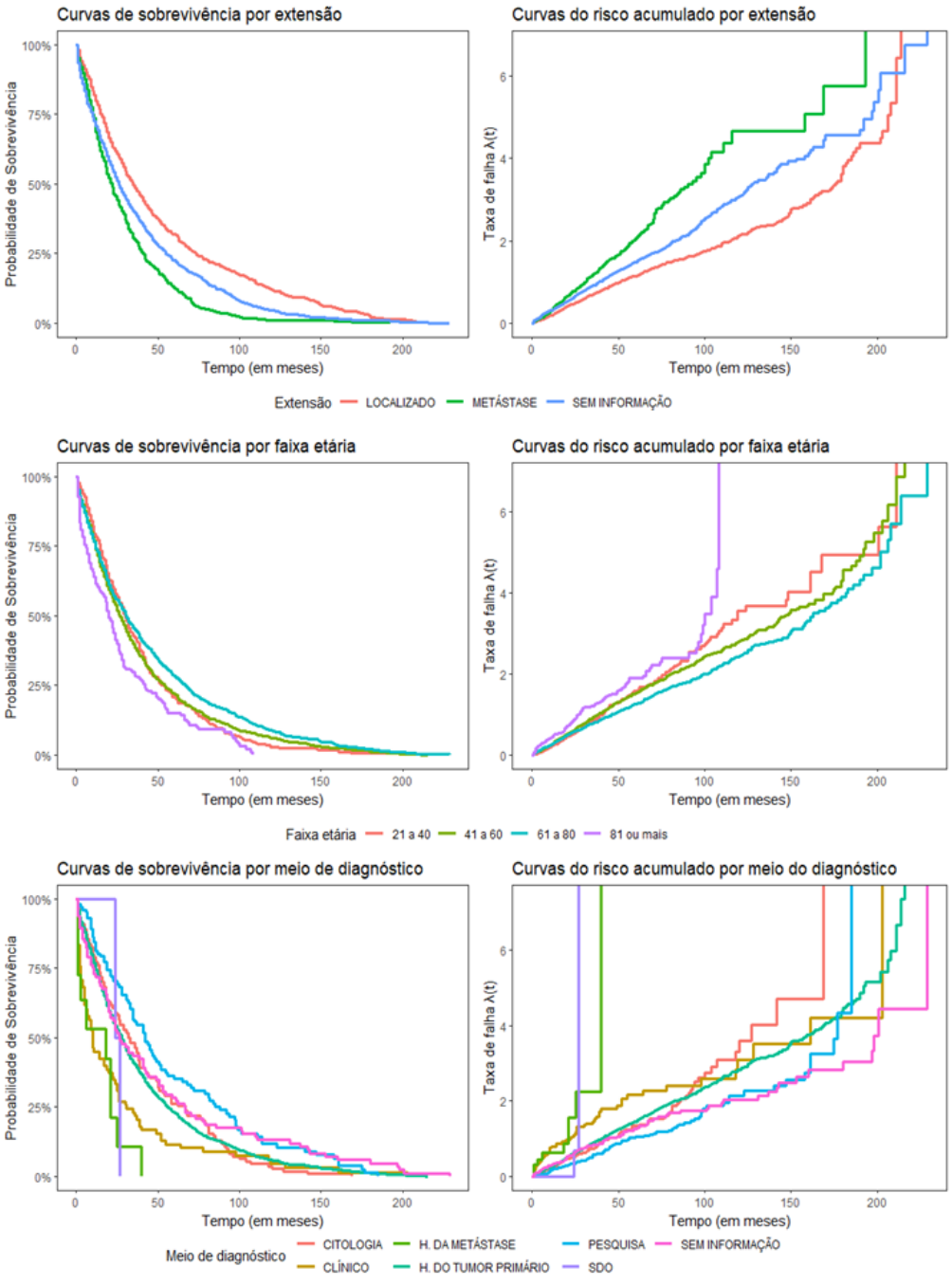
Figura 3 – Curva de sobrevivência e risco de morte (Kaplan-Meier) para as mulheres diagnosticadas com câncer de mama.



Fonte: Produzida pelo autor (2022).

comportamento exponencial devido os tempos distintos de falha ocorrerem nos primeiros meses e, além disso, a quantidade desses tempos distintos são baixas. Até o 40º mês o risco de obter o evento de interesse oscila entre pessoas diagnosticadas pela histologia da metástase e exames clínicos. Do 40º mês até aproximadamente os cem primeiros meses de acompanhamento, o risco foi maior para quem obteve um diagnóstico clínico. Seguido do diagnóstico por citologia até o 169º mês.

Figura 4 – Curva de sobrevivência e risco de morte (Kaplan-Meier) por covariáveis para as mulheres diagnosticadas com câncer de mama.



Fonte: Produzida pelo autor (2022).

Em seguida foi aplicado o teste de *logrank* com objetivo de comparar as curvas de sobrevivência para os grupos de cada covariável. O teste foi aplicado em três covariáveis: extensão, faixa etária e meio de diagnóstico, respectivamente, sendo que para cada uma, as comparações foram feitas em uma combinação de dois a dois grupos. Nesse teste, não foram considerados os grupos classificados como "sem informação" por não apresentarem uma categoria específica, assim como também o grupo "SDO" da covariável meio de diagnóstico por ter um número de

informações muito baixo (apenas duas informações).

Tabela 3 – Teste *logrank* para a covariável extensão.

Grupos	**P-valor**
Localizado × metástase	< 0,001

Fonte: Produzida pelo autor (2022).

Tabela 4 – Teste *logrank* para a covariável faixa etária.

Grupos	**P-valor**
21 a 40 × 41 a 60	0,900
21 a 40 × 61 a 80	0,020
21 a 40 × 81 ou mais	0,001
41 a 60 × 61 a 80	0,003
41 a 60 × 81 ou mais	< 0,001
61 a 80 × 81 ou mais	< 0,001

Fonte: Produzida pelo autor (2022).

Tabela 5 – Teste *logrank* para a covariável meio de diagnóstico.

Grupos	**P-valor**
H. do tumor primário × pesquisa	0,001
H. do tumor primário × citologia	0,900
H. do tumor primário × h. da metástase	0,001
H. do tumor primário × clínico	< 0,001
Pesquisa × citologia	0,005
Pesquisa × h. da metástase	< 0,001
Pesquisa × clínico	< 0,001
Citologia × h. da metástase	0,001
Citologia × clínico	0,006
H. da metástase × clínico	0,300

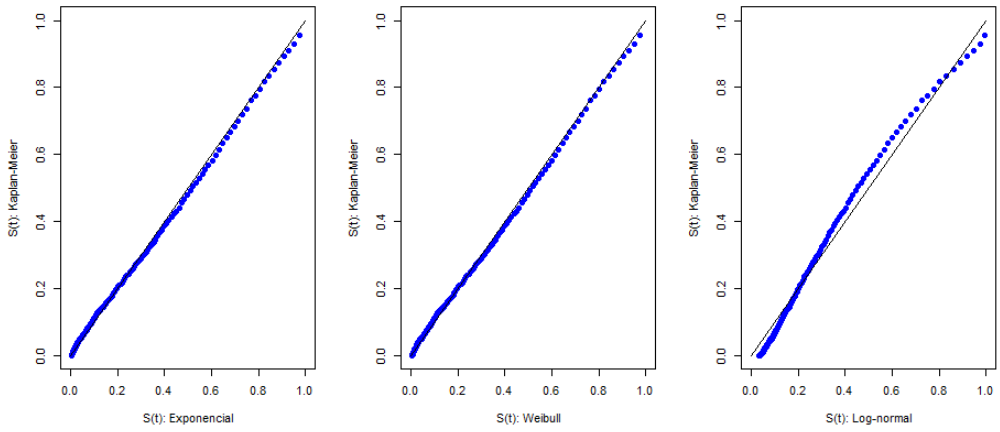
Fonte: Produzida pelo autor (2022).

Considerando as hipóteses H_0 : não há diferença entre as curvas de sobrevivência e H_1 : há diferença entre as curvas de sobrevivência, foi verificado na Tabela 3, Tabela 4 e Tabela 5 que apenas três grupos não rejeitaram a hipótese nula (p-valor > 0,05) indicando ao nível de 5% de significância que não há diferença significativa entre as curvas de sobrevivência desses grupos. Os demais grupos rejeitam H_0, portanto, há diferença significativa nas curvas dos mesmos.

Para a escolha do melhor modelo utilizou-se inicialmente duas análises gráficas para comparar as estimativas de sobrevivência de Kaplan-Meier com as estimativas dos modelos probabilísticos. Pelos resíduos de Cox-Snell (Figura 5) tanto a distribuição exponencial como a

weibull se ajustaram consideravelmente bem, pois em ambas, os pontos estão bem sobrepostos sob a reta de regressão $y = x$.

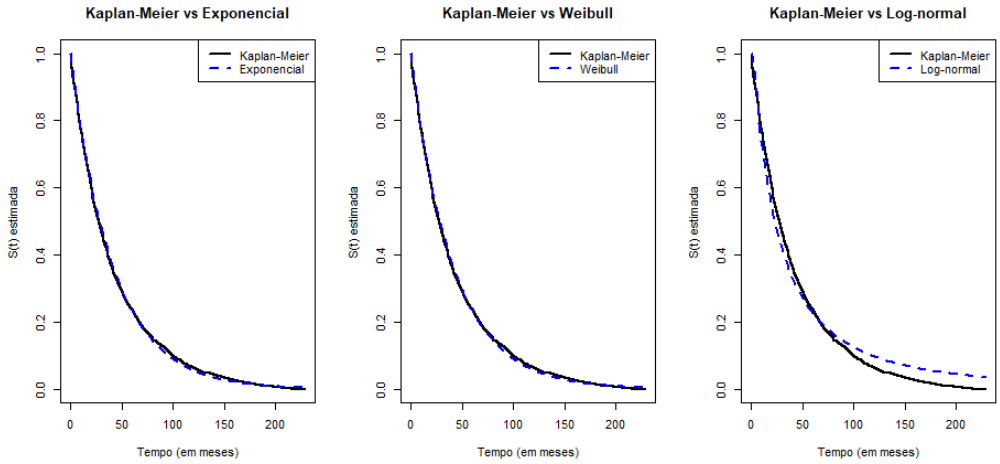
Figura 5 – Análise de resíduos de Cox-Snell das sobrevivências estimadas pelo modelo de regressão exponencial, weibull e log-normal versus as estimativas de Kaplan-Meier.



Fonte: Produzida pelo autor (2022).

As curvas de sobrevivência ajustadas pelas três distribuições de probabilidades junto às estimativas de Kaplan-Meier confirmam na Figura 6 que também a distribuição exponencial e weibull são as mais adequadas para os dados em estudo, pois a curva azul pontilhada representa bem a curva de Kaplan-Meier.

Figura 6 – Gráfico das sobrevivências estimadas por Kaplan-Meier versus às sobrevivências estimadas pelo modelo exponencial, weibull e log-normal.



Fonte: Produzida pelo autor (2022).

Para a confirmação dos resultados das análises gráficas foi realizado o teste da razão de verossimilhança com as seguintes hipóteses: i) o modelo exponencial é adequado; ii) o modelo weibull é adequado; iii) o modelo log-normal é adequado. O objetivo desse teste é identificar qual distribuição descreve melhor os dados tão bem quanto a gama generalizada.

Tabela 6 – Logaritmo da função de verossimilhança e resultados dos TRV para o modelo geral.

Modelo	**log(L(θ))**	**TRV**	**P-valor**
Gama generalizada	$log(L(\gamma,k,\alpha)) = -10259,68$	-	1,00
Exponencial	$log(L(\alpha)) = -10261$	2,78	0,2486
Weibull	$log(L(\gamma,\alpha)) = -10261$	2,57	0,1089
Log-normal	$log(L(\mu,\sigma)) = -10376$	233,1	$\cong 0$

Fonte: Produzida pelo autor (2022).

O teste na Tabela 6 indica que o modelo exponencial é o mais adequado para o estudo desses dados por apresentar um maior p-valor. Portanto, podem-se tirar conclusões no que se refere ao tempo de falha sob o modelo exponencial. O tempo médio de vida, por exemplo, é o próprio parâmetro da distribuição: $\alpha = 41,44$ meses; segundo Colosimo e Giolo (2006), o tempo mediano pode ser obtido pelo percentil: $t_p = -\alpha log(1-p)$, sendo $p = 0,5$ o índice de percentil mediano. Com isso, $t_{0,5} = 28$, logo, 50% das mulheres faleceram em até 28 meses após diagnosticadas por câncer de mama e 50% faleceram após 28 meses; a probabilidade de sobreviver por câncer de mama até o 24º mês é de 0,56, isto é, $S(24) = 0,56$.

Além das curvas de Kaplan-Meier (Figura 4) fornecer estatísticas descritivas na perspectiva de conceitos em análise de sobrevivência, também nos fornece informações de variáveis que possivelmente influenciam no tempo até a ocorrência do evento de interesse.

Foram considerados, assim como no ajuste paramétrico geral, os três modelos de regressão (exponencial, weibull e log-normal) em função dessas covariáveis: extensão, faixa etária e meio de diagnóstico. A Tabela 7 apresenta os resultados dos TRV e p-valores como também o AIC para esse ajuste.

Tabela 7 – Logaritmo da função de verossimilhança, resultados dos TRV e AIC para o modelo ajustado por covariáveis.

Modelo	**log(L(θ))**	**TRV**	**P-valor**	**AIC**
Gama generalizada	$log(L(\gamma,k,\alpha)) = -10180,54$	-	1,00	-
Exponencial	$log(L(\alpha)) = -10182,27$	3,47	0,18	20388,55
Weibull	$log(L(\gamma,\alpha)) = -10180,66$	0,24	0,88	20387,33
Log-normal	$log(L(\mu,\sigma)) = -10299,47$	237,86	0	20624,94

Fonte: Produzida pelo autor (2022).

É possível observar no teste da Tabela 7 que o modelo weibull é o mais adequado, pois além de possuir um maior p-valor (0,88), este é muito próximo do p-valor do modelo de

referência comparado, a gama generalizada. Nota-se também que o modelo weibull apresenta um menor AIC. No entanto, podemos tirar algumas conclusões sobre o risco de morte por câncer de mama através das estimativas do modelo de regressão weibull.

Os coeficientes estimados $(\hat{\beta}_i)$ para cada categoria das covariáveis estão descritos na Tabela 8. Coeficientes negativos contribuem para o aumento do risco e positivos contribuem para a redução.

Tabela 8 – Modelo de regressão paramétrico ajustado por covariáveis através da distribuição weibull.

Covariáveis	**Estimativas** $(\hat{\beta}_i)$	**Exp(estimativas)**
Extensão (metástase)	-0,553	0,575
Extensão (sem informação)	-0,306	0,736
Faixa etária (41 a 60)	-0,019	0,981
Faixa etária (61 a 80)	0,105	1,111
Faixa etária (81 ou mais)	-0,339	0,712
Meio de diag. (clínico)	-0,567	0,567
Meio de diag. (h. da metástase)	-0,957	0,384
Meio de diag. (h. do tumor primário)	-0,099	0,906
Meio de diag. (pesquisa)	0,182	1,199
Meio de diag. (SDO)	-0,318	0,727
Meio de diag. (sem informação)	0,156	1,169

Fonte: Produzida pelo autor (2022).

Estar numa extensão de metástase mostrou ser um fator de risco, aumentando em 0,575 (57,52%) vezes mais a chance de um indivíduo vir a óbito em relação aos indivíduos com extensão localizada. Quando não se sabe a informação da extensão também é um fator de risco de morte, aumentando em 0,736 vezes a taxa de falha.

Comparando-se em relação a covariável faixa etária numa categoria de 21 a 40 anos, verificamos que pertencer à faixa etária de 81 anos ou mais mostrou ser um fator de risco em relação à morte por câncer de mama, aumentando o risco de morte em cerca de 0,712 vezes mais.

Ser diagnosticado tanto por meio clínico como pela histologia da metástase apresentaram ser fatores de risco de obter a falha. Por meio clínico, portanto, a taxa de falha é de 0,567 vezes mais, aumentando em 56,72%. Já diagnosticado pela histologia da metástase o risco é de 0,384 mais, crescendo em 38,40%. Isto, comparando-se em relação a covariável meio do diagnóstico por citologia.

4 CONCLUSÃO

Por meio das estimativas de Kaplan-Meier foi possível concluir que o risco falecer por câncer de mama aumenta com o passar dos meses. Dado que a mulher contém um tumor em metástase, o risco é maior em relação ao tumor localizado, como esperado. Em relação à faixa etária, as pacientes mais idosas (81 anos ou mais) possuem um risco maior de falecer por câncer de mama, diferente daquelas que possuem de 61 a 80 anos, que porventura o risco é menor. Para a covariável meio de diagnóstico, o risco oscila com base dos tempos distintos de falha.

O teste de *logrank* para as covariáveis extensão, faixa etária e meio de diagnóstico mostraram que ao nível de 5% de significância, em apenas três grupos não houveram diferenças significativas entre as curvas de sobrevivência. São eles: faixa etária "21 a 40 $\times$ 41 a 60", meios de diagnóstico "h. do tumor primário $\times$ citologia" e meios de diagnóstico "h. da metástase $\times$ clínico".

O modelo que melhor descreveu os dados para um ajuste geral foi o modelo paramétrico utilizando a distribuição exponencial (não descartando as estimativas pelo modelo weibull). As análises gráficas dos resíduos de Cox-Snell e a comparação de curvas de sobrevivências evidenciaram isto, além da verificação pelo teste da razão de verossimilhança. Já para os modelos de regressão paramétricos, o modelo weibull foi o mais adequado, pois obteve um maior p-valor (0,88) em relação aos demais assim como o menor AIC (20387,33).

Com tudo, verificamos o quanto o estudo da análise de sobrevivência tem sido importante ao trazer informações que poderão servir como base para tomada de decisões. Através deste, foram identificados os principais fatores de risco que afetam o tempo de vida de mulheres diagnosticadas com câncer de mama tal como as probabilidades de sobrevivência.

REFERÊNCIAS

BRESLOW, N.; CROWLEY, J. A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, JSTOR, p. 437–453, 1974. Citado na página 17.

BUZAID, A. C.; MALUF, F. C.; GAGLIATO, D. de M. *Vencer o Câncer de Mama*. 2. ed. São Paulo: Dentrix, 2020. Citado na página 11.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. 1. ed. São Paulo: Blucher, 2006. Citado 4 vezes nas páginas 14, 15, 22 e 29.

COX, D. R.; SNELL, E. J. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 30, n. 2, p. 248–275, 1968. Citado na página 21.

DIAS, J. F.; MARTINS, N. S.; GRADIM, C. V. C. Análise de sobrevida de mulheres com câncer de mama. *Revista de Enfermagem UFPE On Line*, 2018. Citado na página 12.

FERRAZ, R. de O.; FILHO, D. de C. M. Análise de sobrevivência de mulheres com câncer de mama: modelos de riscos competitivos. *Ciência Saúde Coletiva [online]*, v. 21, n. 11, 2017. Citado na página 11.

IARC, I. A. F. R. O. C. *Cancer today*. Lyon: França, 2020. Citado na página 11.

INCA, I. N. de C. *Câncer de mama: vamos falar sobre isso?* 5. ed. Rio de Janeiro, 2019. Citado na página 11.

INCA, I. N. de C. *Câncer de mama: vamos falar sobre isso?* 7. ed. Rio de Janeiro, 2022. Citado na página 11.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53, p. 457–481, 1958. Citado na página 17.

LAWLESS, J. F. *Statistical Models and Methods for Lifetime Data*. 2. ed. New York: John Wiley and Sons, 1982. Citado na página 21.

MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports*, v. 50, p. 163–170, 1966. Citado na página 17.

MARTINS, V. L. M.; WERNER, L. Análise não paramétrica de falhas ao longo do calendário para alto-falantes. *Produto Produção*, v. 11, n. 3, 2010. Citado na página 13.

PEREIRA, P. J.; VIVANCO, M. J. F. Viabilidade da aplicação de mecanismos de censura tipo i e aleatória em dados entomológicos. *Ciência e Agrotecnologia*, Editora da Universidade Federal de Lavras, v. 27, n. 2, 2003. Citado na página 14.

STRAPASSON, E. *Comparação de Modelos com Censura Intervalar em Análise de Sobreviência*. Tese (Doutorado em Agronomia) — Universidade de São Paulo - Escola Superior de Agricultura "Luiz de Queiroz", 2007. Citado na página 15.

TEAM, R. C. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: <http://www.R-project.org>. Citado na página 13.

THERNEAU, T. M. et al. *survival: Survival Analysis*. [S.l.], 2022. R package version 3.4-0. Disponível em: <https://cran.r-project.org/web/packages/survival>. Citado na página 13.

THULER, L. C. Considerações sobre a prevenção do câncer de mama feminino. *Revista Brasileira de Cancerologia*, v. 49, n. 4, p. 227–238, 2003. Citado na página 11.