SARA SILVÉRIO

ANÁLISE DE SOBREVIVÊNCIA COM CENSURA INTERVALAR APLICADA NA GERMINAÇÃO DE SEMENTES DE PITAIA

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Orientador: Sebastião Martins Filho

Coorientadores: Lausanne Soraya de Almeida
Godofredo Quispe Mamani

VIÇOSA - MINAS GERAIS
2024

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
S587a
2024

Silvério, Sara, 2000-
Análise de sobrevivência com censura intervalar aplicada na germinação de sementes de pitaia / Sara Silvério. – Viçosa, MG, 2024.
1 dissertação eletrônica (92 f.): il. (algumas color.).

Inclui apêndice.
Orientador: Sebastião Martins Filho.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Estatística, 2024.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2024.741
Modo de acesso: World Wide Web.

1. Análise de sobrevivência (Biometria). 2. Germinação - Métodos estatísticos. I. Martins Filho, Sebastião, 1961-.
II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 519.546

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

SARA SILVÉRIO

ANÁLISE DE SOBREVIVÊNCIA COM CENSURA INTERVALAR APLICADA NA GERMINAÇÃO DE SEMENTES DE PITAIA

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 16 de agosto de 2024

Assentimento:

Documento assinado digitalmente
gov.br
SARA SILVERIO
Data: 13/11/2024 08:56:52-0300
Verifique em https://validar.iti.gov.br

Sara Silvério
Autora

Documento assinado digitalmente
gov.br
Sebastiao Martins Filho
Data: 13/11/2024 12:02:11-0300
Verifique em https://validar.iti.gov.br

Sebastião Martins Filho
Orientador

AGRADECIMENTOS

Agradeço a Deus, em primeiro lugar, por ter me guiado e dado forças para aproveitar todas as oportunidades que tenho recebido ao longo desta jornada. Sua presença constante tem sido uma bênção em cada etapa.

Aos meus pais, Adilson e Maria Aparecida, que são o alicerce da minha jornada, expresso minha profunda gratidão pelo amor, cuidado e apoio dedicados a mim. Agradeço por me proporcionarem condições de estudar e me dedicar, bem como pelos ensinamentos e pela confiança depositada em mim. Obrigada por acreditarem mais em mim do que eu mesma.

À minha irmã Jhennifer, pela amizade sincera. Sua presença, mesmo à distância, é um apoio incalculável em todos os momentos. Sua força e determinação me inspiram a ser uma mulher e profissional mais forte.

Ao meu namorado Victor, pelo companheirismo, carinho e incentivo em todos os momentos. Obrigada por dividir a vida comigo e por me motivar a ser uma pessoa melhor. Sua calmaria e dedicação são uma inspiração para mim.

Às minhas amigas de graduação e de mestrado, Brenda e Mariana. Vocês foram essenciais durante toda a minha jornada, sempre presentes nas horas boas e ruins, ajudando a superar todos os desafios e tornando a caminhada mais fácil.

Agradeço também a todos os amigos da “Salinha”, pelos momentos de estudo e de descontração. A nossa convivência me proporcionou confiança e alegria durante todo esse processo.

Ao meu orientador, Professor Sebastião Martins Filho, expresso minha sincera gratidão pela sua constante paciência, colaboração e dedicação em me orientar, transmitindo-me valiosos ensinamentos ao longo do caminho.

Ao Laboratório de Análise de Sementes Florestais (LASF), e especialmente à professora Lausanne Soraya de Almeida e à aluna Vívian, por terem gentilmente me acolhido no laboratório, proporcionando-me valioso auxílio em meu desenvolvimento profissional.

Ao doutorando Marciel, pela disposição ao caminhar conosco nesse trabalho, agradeço todo conhecimento compartilhado.

À Universidade Federal de Viçosa, pela oportunidade de realizar a graduação e a pós-graduação que escolhi.

Ao departamento de Estatística e a todos os professores que contribuíram para a minha formação acadêmica e profissional.

Agradeço também à banca examinadora por sua participação na defesa e pelas valiosas contribuições oferecidas para melhoria deste trabalho.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de estudos. O presente trabalho foi realizado com apoio da CAPES – Brasil – Código de Financiamento 001.

Por fim, agradeço a todos que, de forma direta ou indireta, contribuíram para a realização desse trabalho.

“A beleza não está na partida ou na chegada, mas na travessia”.

(Guimarães Rosa)

RESUMO

SILVÉRIO, Sara, M.Sc., Universidade Federal de Viçosa, agosto de 2024. **Análise de sobrevivência com censura intervalar aplicada na germinação de sementes de pitaia**. Orientador: Sebastião Martins Filho. Coorientadores: Lausanne Soraya de Almeida e Godofredo Quispe Mamani.

Devido à presença de dados censurados nos ensaios de germinação de sementes, os métodos convencionais de análise de dados podem não ser a escolha mais apropriada. Uma alternativa recomendada é a aplicação da análise de sobrevivência. Em muitos casos, a determinação exata do momento da falha não é possível devido às observações serem realizadas em intervalos de tempo, caracterizando a censura intervalar. Além disso, pode ocorrer que todas as unidades experimentais sejam observadas nos mesmos intervalos de tempo, caracterizando um caso particular de censura intervalar conhecido como dados grupados. No Capítulo 1, este trabalho teve como objetivo fornecer uma revisão geral da análise de sobrevivência, enfatizando a censura intervalar e abrangendo técnicas não paramétricas, paramétricas e semiparamétricas. O Capítulo 2 teve como objetivo aplicar e avaliar métodos de análise de sobrevivência com censura intervalar, utilizando dados de quatro experimentos de germinação de sementes de pitaia (*Hylocereus* spp.) obtidos da base de dados Mendeley. Nos experimentos foram realizadas contagens semanais durante quatro semanas. As sementes que não germinaram até o final desse período foram consideradas censuradas (não germinadas) à direita. Foram aplicadas as seguintes técnicas: i) o algoritmo EMICM para avaliar o efeito combinado do tempo de armazenamento e do tipo de luz; ii) regressão paramétrica para analisar o efeito conjunto do local e do tempo de armazenamento; iii) regressão semiparamétrica para examinar o impacto do método de extração das sementes; e vi) regressão discreta em dados grupados para avaliar o efeito combinado do armazenamento e da temperatura. As técnicas aplicadas permitiram avaliar esses fatores na germinação de sementes de pitaia. Dessa forma, análise de sobrevivência demonstrou ser uma ferramenta valiosa para lidar com dados censurados intervalares em estudos de germinação, destacando a importância da seleção adequada do método de análise, de acordo com a natureza dos dados e os objetivos do estudo.

Palavras-chave: *Hylocereus* spp.; Armazenamento; Temperatura; Tipos de luz; Extração de semente; Tempo até o evento.

ABSTRACT

SILVÉRIO, Sara, M.Sc., Universidade Federal de Viçosa, August, 2024. **Survival analysis with interval censoring applied to the germination of pitaya seeds**. Adviser: Sebastião Martins Filho. Co-advisers: Lausanne Soraya de Almeida and Godofredo Quispe Mamani.

Due to the presence of censored data in seed germination tests, there may be more appropriate choices than conventional data analysis methods. A recommended alternative is the application of survival analysis. In many cases, it is impossible to observe the exact moment of failure due to observations made at time intervals, resulting in interval censoring. Furthermore, all experimental units may be observed at the same time intervals, representing a specific interval censoring case known as grouped data. In Chapter 1, this work aimed to provide a general review of survival analysis, emphasizing interval censoring and covering non-parametric, parametric, and semi-parametric techniques. Chapter 2 aimed to apply and evaluate survival analysis methods with interval censoring using data from four pitaya seed germination experiments (*Hylocereus spp.*) obtained from the Mendeley database. In the experiments, weekly counts were performed over four weeks. Seeds that did not germinate by the end of this period were considered right-censored (non-germinated). The following techniques were applied: i) the EMICM algorithm to assess the combined effect of storage time and light type; ii) parametric regression to analyze the joint effect of storage location and time; iii) semiparametric regression to examine the impact of seed extraction methods; and iv) discrete regression on grouped data to evaluate the combined effect of storage and temperature. These techniques allowed for an evaluation of these factors in pitaya seed germination. Thus, survival analysis proved to be a valuable tool for handling interval-censored data in germination studies, highlighting the importance of selecting the appropriate analysis method according to the nature of the data and the study's objectives.

Keywords: *Hylocereus* spp.*;* Storage; Temperature; Light types; Seed extraction; Time-to-event.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

A análise de sobrevivência é um conjunto de técnicas e modelos estatísticos usados na análise de experimentos, cuja variável resposta é o tempo até a ocorrência de um evento de interesse (Colosimo e Giolo, 2024). Esse período também é definido como tempo de falha, sendo o termo falha determinada como a ocorrência do evento em questão, neste caso a ocorrência da germinação. Uma característica importante dos dados de sobrevivência é a presença de censuras, que é a observação parcial da resposta (não germinação). As censuras são consideradas observações incompletas e se caracterizam como dados de indivíduos em que a ocorrência do evento não foi verificada por alguma razão.

Na presença de censura, as técnicas estatísticas comumente utilizadas se tornam inviáveis, pois não permitem o uso de observações parciais (Onofri *et al*., 2010; Mcnair *et al*., 2012). Colosimo e Giolo (2024) ressaltam o fato de que todas as observações provenientes de um estudo (completas e parciais) devem ser utilizadas na análise estatística. Esses autores afirmam que, mesmo sendo incompletas, as censuras fornecem informações relevantes sobre o tempo do evento e sua omissão no cálculo das estatísticas de interesse pode acarretar conclusões viesadas. Sendo assim, são necessários métodos estatísticos que possibilitem incorporar na análise a informação contida tanto nas observações completas quanto nas censuras (Duarte *et al*., 2023).

Na análise de sobrevivência as técnicas utilizadas podem ser não-paramétricas, paramétricas e semiparamétricas. A modelagem normalmente considera o tempo exato de falha, mas, em alguns casos, os dados são coletados em intervalos, indicando apenas que o evento ocorreu entre duas visitas, sem especificar o tempo exato, caracterizando censura intervalar. Nesse caso, assumir que dados coletados em intervalos tenham tempo exato de falha pode conduzir a vícios, bem como resultados e conclusões não confiáveis (Bogaerts *et al*., 2017; Colosimo e Giolo, 2024).

Existem estudos que apresentam técnicas para lidar com a censura intervalar dos dados, visando incorporar essa característica específica durante a análise estatística. Turnbull (1976) propõe uma abordagem não paramétrica para estimar a função de distribuição empírica. O modelo de tempo de falha acelerado, uma abordagem paramétrica para dados censurados, é discutido em Rabinowitz *et al*.

(1995). Finkelstein (1986) propõe um modelo de risco proporcional para lidar com dados censurados por intervalo, representando uma abordagem semiparamétrica.

Um caso especial de dados com censura intervalar são os dados de sobrevivência grupados, nos quais as informações estão disponíveis para todas as unidades experimentais no mesmo intervalo de observação (Giolo *et al*., 2009). Este tipo de dado é caracterizado por apresentar muitos empates, isto é, múltiplas ocorrências de falha em um mesmo intervalo de tempo.

Existem três alternativas para trabalhar com esse tipo de dados: (1) utilizar a função de verossimilhança parcial exata no contexto do modelo de taxas proporcionais; (2) utilizar aproximações para a função de verossimilhança parcial no contexto do modelo de taxas de falha proporcionais; (3) utilizar modelos de regressão discretos (Lawless, 2003). A variável resposta é tratada como contínua nas duas primeiras abordagens e como discreta na terceira.

Considerando o que foi mencionado, o objetivo deste trabalho foi realizar uma revisão teórica a respeito dos métodos não paramétricos, paramétricos e semiparamétricos de sobrevivência intervalar, incluindo técnicas para dados grupados, e em seguida aplicar essas técnicas em quatro diferentes experimentos, a fim de investigar os efeitos do método de extração, armazenamento, temperatura e condições de luz na germinação de sementes de pitaia (*Hylocereus* spp.).

Este trabalho está estruturado da seguinte forma: o Capítulo 1 apresenta o referencial teórico, explorando a história da pitaia, suas características e sua importância, além de realizar uma revisão bibliográfica sobre análise de sobrevivência, com ênfase na censura intervalar. No Capítulo 2, é aplicada a análise de sobrevivência com censura intervalar no contexto da germinação de sementes de pitaia, utilizando conjuntos de dados extraídos de um trabalho previamente publicado (Zerpa-Catanho et al., 2019).

Referências

Bogaerts, K., Komarek, A., Lesaffre, E., 2017. **Survival Analysis with Interval-Censored Data: A Practical Approach with Examples in R, SAS, and BUGS**. Boca Raton: CRC Press. https://doi.org/10.1201/9781315116945

Colosimo, E. A.; Giolo, S. R., 2024. **Análise de sobrevivência aplicada**. 2. ed. São Paulo: Editora Blücher. 362p.

Duarte, M. L., Martins Filho, S., Freitas, A. F., Xavier, A., 2023. Rooting of forest species mini-cuttings: an application of non-parametric survival analysis. **New Forests**, *54*(6), 1153-1167. https://doi.org/10.1007/s11056-023-09962-0

Finkelstein, D. M., 1986. A proportional hazards model for interval-censored failure time data. **Biometrics**, 845-854. https://doi.org/10.2307/2530698

Giolo, S. R., Colosimo, E. A., Demétrio, C. G. B., 2009. Different approaches for modeling grouped survival data: A mango tree study. **Journal of agricultural, biological, and environmental statistics**, 14, 154-169. https://doi.org/10.1198/jabes.2009.0010

Lawless, J. F., 2003. **Statistical models and methods for lifetime data**. 2. Ed. New York: John Wiley & Sons.

Mcnair, J. N., Sunkara, A. E., Frobish, D., 2012. How to analyse seed germination data using statistical time-to-event analysis: non-parametric and semi-parametric methods. **Seed Science Research,** 22(2), 77 – 95. https://doi.org/10.1017/S0960258511000547

Onofri, A., Gresta, F., Tei, F., 2010. A new method for the analysis of germination and emergence data of weed species. **Weed Research**, 50(3), p. 187–198. https://doi.org/10.1111/j.1365-3180.2010.00776.x

Rabinowitz, D., Tsiatis, A., Aragon, J., 1995. Regression with interval-censored data. **Biometrika**, 82(3), 501-513. https://doi.org/10.1093/biomet/82.3.501

Turnbull, B. W., 1976. The empirical distribution function with arbitrarily grouped, censored and truncated data. **Journal of the Royal Statistical Society: Series B (Methodological)**, 38(3), 290-295. https://doi.org/10.1111/j.2517-6161.1976.tb01597.x

Zerpa-Catanho, D., Hernández-Pridybailo, A., Madrigal-Ortiz, V., Zúñiga-Centeno, A., Porras-Martínez, C., Jiménez, V. M., Barboza-Barquero, L., 2019. Seed germination of pitaya (*Hylocereus* spp.) as affected by seed extraction method, storage, germination conditions, germination assessment approach and water potential. **Journal of Crop Improvement**, v. 33, n. 3, p. 372-394. https://doi.org/10.1080/15427528.2019.1604457

CAPÍTULO 1: REFERENCIAL TEÓRICO

Neste capítulo, está apresentado o embasamento teórico da produção de pitaia, explorando sua história, características e como alguns fatores ambientais e fisiológicos influenciam sua germinação. Além disso, foi realizada uma revisão bibliográfica sobre a análise de sobrevivência, com ênfase na censura intervalar e a aplicação de análise de sobrevivência em germinação de sementes.

1.1. Germinação de sementes e produção da pitaia

A pitaia é uma fruta pertencente à família *Cactaceae* e sua origem exata é incerta, mas acredita-se que seja das regiões tropicais do México e da América Central (Rojas-Sandoval e Praciak, 2022). Existem, dentro desta família, 35 gêneros com potencial alimentício, destacando-se *Cereus, Leptocereus, Hylocereus, Stenocereus, Selenicereus, Escontria, Myrtillocactos e Opuntia* (Nunes *et al*., 2014). A pitaia é originada de diferentes espécies dos gêneros. No Brasil, algumas espécies comercializadas são *Hylocereus undatus* (pitaia-vermelha-de-polpa-branca), *Hylocereus costaricensis* (pitaia-vermelha-de-polpa-vermelha), *Selenicereus megalanthus* (pitaia-amarela) e *Selenicereus setaceus* (pitaia-do-cerrado), (Junqueira *et al*., 2010).

A pitaia, também conhecida como fruta do dragão, é um fruto tropical adaptado a ambientes áridos e semiáridos (Chen *et al*., 2023), sendo capaz de suportar altas temperaturas, período de estiagem e solos pobres em nutrientes. Os principais produtores de pitaia são os países asiáticos como Vietnã (38,20%), Tailândia (20,22%) e China (10,11%) (Tarte et., 2023).

No Brasil, essa espécie começou a ser cultivada em meados de 1990, inicialmente na região de Catanduva, no estado de São Paulo (Nunes *et al*., 2014). As condições climáticas favoráveis em diversas regiões propiciam um cultivo bem-sucedido dessa fruta exótica. O clima subtropical e tropical dessas áreas oferece um ambiente ideal para o desenvolvimento das pitaias, enquanto a variabilidade de solos permite a adaptação de diversas variedades. Segundo Faleiro (2022), a produção está concentrada nas regiões Sudeste e Sul, com mais de 80% da produção brasileira. São Paulo lidera os estados com maior produção (40%), seguido por Santa Catarina (24%), Minas Gerais (12%) e Pará (10%).

O cultivo de pitaias no Brasil cresce constantemente devido à demanda crescente, tanto no mercado interno quanto no externo. Como resultado, produtores brasileiros têm expandido suas áreas de cultivo e investido em tecnologias para aumentar a produtividade e melhorar a qualidade dos frutos, destacando o potencial econômico e agrícola dessa cultura no país.

A produção de mudas é o primeiro passo para se obter frutos de qualidade e com elevado valor de mercado (Fernandes e Coutinho, 2019). A propagação clonal por meio de estaquia é o modo preferido de reprodução em *Hylocereus* spp., usando estacas de caule (Le Bellec *et al*., 2006) ou cultura in vitro (Hua *et al*., 2015). Entretanto, o uso de sementes é importante para o melhoramento convencional e conservação de recursos genéticos vegetais (Le Bellec *et al*., 2006; Hernández e Salazar, 2012).

A utilização de sementes de alta qualidade fisiológica é indispensável, daí surge a importância de avaliar a capacidade de germinação das sementes. Neste contexto, os testes de germinação são fundamentais para caracterização do potencial dos lotes de sementes (Hampton e TeKrony, 1995), tanto para uso em pesquisa quanto para comercialização. Além disso, para que a germinação ocorra de forma eficiente, diversos fatores ambientais e fisiológicos influenciam diretamente no sucesso desse processo. A compreensão desses fatores é crucial para otimizar as condições de germinação, seja em ambientes naturais ou controlados. Entre os principais aspectos que influenciam a germinação, destacam-se a luz, a temperatura, o tempo de armazenamento e os métodos de extração das sementes.

Ao se tratar da luz, a maioria das espécies de pitaia possui sementes pequenas e com poucas reservas (Ruths *et al*., 2019), que, de acordo com Majerowicz e Peres (2004) tendem a ser fotoblásticas positivas, tornando a luz um regulador crítico da germinação. A luz permite que as sementes detectem sua proximidade com a superfície do solo, o que é crucial para espécies com sementes pequenas, que podem falhar em emergir se germinarem muito profundamente (Fenner e Thompson, 2005).

Quanto ao tempo de armazenamento, Zerpa-Catanho et al. (2019), relatam que as sementes de *Hylocereus* spp. permaneceram viáveis por 12 meses a 5,75 ± 0,08 °C e 62,59 ± 0,73 % de umidade relativa. Da mesma forma, Kataoka et al. (2013) demonstraram que as sementes de *Hylocereus undatus* permaneceram viáveis por 12 meses a 4°C em condições secas. Além disso, Andrade et al., (2005) observaram

que o armazenamento em câmara fria resultou em porcentagens de germinação de sementes de *Hylocereus undatus* significativamente mais altas.

Em relação ao método de extração, Zerpa-Catanho et al. (2019) relataram que a extração enzimática apresenta resultados superiores. Esse efeito é atribuído à degradação dos componentes da parede celular da polpa por meio da enzima pectinase, o que provavelmente amolece o tegumento das sementes e facilita a absorção de água durante a embebição, acelerando assim o processo de germinação. Resultados semelhantes foram observados por Yambe e Takeno (1992), que observaram que enzimas macerantes, como a pectinase, aceleraram e aumentaram a germinação de aquênios de *Rosa multiflora*. Esse efeito foi atribuído à degradação das substâncias cimentantes e paredes celulares, o que separou as células ao longo da sutura e dividiu o pericarpo, removendo a barreira física à germinação.

Quanto a temperatura, diversos estudos apontam que as temperaturas de 20, 25 e 30°C são as mais adequadas para a germinação de pitaia. Lone et al. (2014) verificaram que as temperaturas constantes de 25 e 30°C, bem como a temperatura alternada de 20-30°C, são ideais para a germinação de sementes de *Hylocereus undatus* e do híbrido *Hylocereus undatus x Hylocereus costaricensis*. Além disso, para *Hylocereus polyrhizus*, essas mesmas temperaturas constantes (20, 25 e 30°C) também se mostraram ideais. De forma similar, Ruths et al. (2019) corroboraram com esses achados, concluindo que tanto as temperaturas constantes de 25 e 30°C quanto a alternada de 20-30°C são as mais propícias para a germinação de *Hylocereus undatus* e *Hylocereus polyrhizus*.

Durante o desenvolvimento destes testes de germinação é comum ocorrer o apodrecimento de sementes e/ou chegar ao final deles com sementes que não germinaram. Estas informações, muitas vezes descartadas, são importantes e devem ser colocadas na análise estatística para que se possa fazer conclusões corretas e não viesadas. Uma técnica estatística que leva em consideração estas informações é a Análise de Sobrevivência.

1.2. Análise de sobrevivência

A análise de sobrevivência é uma abordagem empregada em estudos nos quais a variável de interesse é o tempo decorrido desde um ponto definido até a manifestação de um evento específico, denominado "evento de falha".

O aspecto distintivo dessa técnica reside na sua habilidade de considerar observações censuradas durante a análise dos dados, sendo censura o termo utilizado quando um evento de falha não é totalmente observado em um indivíduo. Isso pode ocorrer devido à perda de acompanhamento desse indivíduo por motivos diversos ou pelo fato de o evento de falha simplesmente não ter ocorrido para ele.

A censura pode assumir três formas distintas: censura à esquerda, à direita e intervalar. A censura à esquerda ocorre quando o evento de interesse já ocorreu antes da primeira observação do indivíduo no estudo. A censura à direita ocorre quando o evento de interesse não é observado até o término do estudo; e a censura intervalar ocorre quando não se conhece o momento exato de um evento, mas sabe-se que ele ocorreu dentro de um intervalo de tempo $(L, U]$, onde $L < T \leq U$ (Radke, 2003).

A variável aleatória não-negativa T, que denota o tempo até a ocorrência de uma falha, conforme Colosimo e Giolo (2024), é geralmente especificada em análise de sobrevivência pela sua função de sobrevivência ou pela função de taxa de falha.

A função de sobrevivência desempenha um papel fundamental na análise de estudos de sobrevivência, ela é definida como a probabilidade de uma observação não experimentar a falha até um tempo específico t (George *et al*., 2014), conforme a expressão definida pela Equação 1:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = 1 - \int_0^t f(u)du \qquad (1)$$

Note que $P(T \leq t)$ é a probabilidade de uma observação falhar até um determinado tempo t, sendo a própria função de distribuição acumulada.

A função de taxa de falha ou função de risco $h(t)$ é determinada pela probabilidade instantânea de um indivíduo sofrer o evento em um intervalo de tempo $[\, t, t + \Delta t)$ dado que ele não sofreu o evento até o tempo t dividida pelo comprimento desse intervalo. Assumindo um intervalo muito pequeno, a taxa de falha de T é definida pela Equação 2:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T > t)}{\Delta t} = \frac{S(t) - S(t + \Delta t)}{\Delta t\, S(t)} \qquad (2)$$

A função de risco não representa uma probabilidade, mas sim um indicativo do nível de risco associado à ocorrência do evento. Quanto maior o valor da função de risco, maior é o risco associado ao evento em questão.

Na análise de sobrevivência, diferentes abordagens podem ser utilizadas para estimar a função de sobrevivência e a função de risco, as quais são classificadas em técnicas não paramétricas, paramétricas e semiparamétricas.

1.2.1. Técnicas Não Paramétricas

De acordo com George *et al*. (2014), ao comparar as funções de sobrevivência de dois ou mais grupos, a estimativa não paramétrica de Kaplan-Meier e o teste log-rank são os métodos estatísticos básicos de análise. Por meio deste estimador podemos, ainda, avaliar a adequação do ajuste de modelos que utilizam técnicas paramétricas e semiparamétricas. A estimativa Kaplan-Meier pode ser usada para gerar gráficos das curvas de sobrevivência, enquanto o teste log-rank (Mantel, 1966) e Wilcoxon (Gehan, 1965) podem ser usados para comparar curvas de diferentes grupos, como feito em Duarte *et al*. (2023) e Wijenayake e Hiroshima (2021). A distinção entre os dois testes está no fato de que o teste log-rank atribui o mesmo peso para cada ponto no eixo do tempo, focando em tempos mais longos, ao passo que o teste de Wilcoxon apresenta maior sensibilidade na identificação de diferenças nos estágios iniciais (Emmert-Streib e Dehmer, 2019).

O Estimador de Kaplan-Meier, ou estimador produto-limite, é uma técnica não paramétrica, o que significa que não requer suposições sobre a distribuição dos dados. Ele calcula a probabilidade de sobrevivência em intervalos discretos de tempo, fundamentando-se na relação entre o número de eventos ocorridos e o número de indivíduos sob risco em cada instante (Kaplan e Meier, 1958). As estimativas podem ser então visualizadas em um gráfico de sobrevivência para ilustrar a probabilidade de sobrevivência ao longo do tempo (Emmert-Streib e Dehmer, 2019).

Para comparar duas curvas $S_1(t)$ e $S_2(t)$ um dos testes não paramétricos muito utilizado em análise de sobrevivência é o teste log-rank, no qual testa-se a hipótese nula H_0: $S_1(t) = S_2(t)$. Considerando os tempos de falha distintos da junção das duas amostras: $t_1 < t_2 < \cdots < t_k.$, e supondo que no tempo t_j ocorra d_{ij} falhas, e que no momento t_{j-1}, existiam n_{ij} indivíduos sob risco para $i = 1, 2$ e $j =$

$1, \ldots, k$. Em cada tempo de falha, os dados podem ser representados conforme Tabela 1.1.

Tabela 1.1 - Contagens observadas no tempo t_j para calcular o teste log-rank para comparar duas funções de sobrevivência

	GRUPOS		
	1	2	Totais
Falha	d_{1j}	d_{2j}	d_j
Não Falha	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
Totais	n_{1j}	n_{2j}	n_j

A variável aleatória D_{2j} representa o número de falhas observadas no grupo 2 no tempo t_j. A distribuição de D_{2j}, condicional à falha e censura até o tempo t_j e ao número de falhas no tempo t_j, é uma hipergeométrica, com seu valor esperado $E_{2j} = \frac{n_{2j}d_j}{n_j}$ e variância $(V_j)_2 = \frac{n_{2j}(n_j - n_{2j})d_j(n_j - d_j)}{{n_j}^2(n_j - 1)}$. Dessa forma, a estatística $D_{2j} - E_{2j}$ tem média zero e variância V_{j2}.

Supondo independência das k tabelas de contagens observadas, o teste log-rank para verificar a igualdade de duas funções de sobrevivência é baseado na estatística conforme a Equação 3:

$$T = \frac{\left[\sum_{j=1}^{k}(D_{2j} - E_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2} \tag{3}$$

que possui uma distribuição qui-quadrado com 1 grau de liberdade sob hipótese nula H_0: $S_1(t) = S_2(t)$.

Para melhorar o poder desse teste em relação a diferentes alternativas, foi proposto o teste log-rank ponderado (Tarone e Ware, 1977), que possui sua estatística apresentada na Equação 4:

$$G = \frac{\left[\sum_{j=1}^{k} w_j (D_{2j} - E_{2j})\right]^2}{\sum_{j=1}^{k} w_j^2 (V_j)_2} \quad (4)$$

em que $w_1, \cdots, w_j$ representam os pesos atribuídos. Sob a hipótese nula de que as funções de sobrevivência não diferem, a estatística G segue distribuição qui-quadrado com 1 grau de liberdade.

Pode ser observado que o teste de Wilcoxon e o log-rank são casos específicos do teste log-rank ponderado. O teste log-rank apresentado na equação (3) é derivado ao definir $w_j = 1$, para $j = 1, \ldots, k$. Por outro lado, o teste de Wilcoxon é obtido tomando $w_j = n_j$, ou seja, colocando mais peso na porção inicial do eixo do tempo, pois nos tempos iniciais o número de indivíduos sob risco (n_j) é maior.

Uma generalização mais ampla desses pesos foi proposta por Fleming e Harrington (1991) é denominada $G^{\rho,\gamma}$ $(\rho \geq 0, \gamma \geq 0)$, e definida como $w_j = w_j^{\rho,\gamma} = \left\{\hat{S}(t_{j-1})^{\rho}\right\} \left\{1 - \hat{S}(t_{j-1})^{\gamma}\right\}$. Os parâmetros ρ e γ representam a sensibilidade do teste em relação às diferenças nos riscos de os eventos ocorrerem em momentos diferentes de acompanhamento em um estudo de sobrevivência.

As técnicas não paramétricas destacam-se pela sua aplicação e interpretação simplificadas, mas apresentam a limitação de não permitir a inclusão de covariáveis na análise. É nesse contexto que ganham relevância as abordagens paramétricas e semiparamétricas, as quais oferecem modelos de regressão adequados para acomodar as covariáveis e dados censurados.

1.2.2. Técnicas Paramétricas

A abordagem paramétrica é assim chamada porque atribui uma distribuição de probabilidade específica para o tempo de sobrevivência. Nessa abordagem é possível utilizar modelos de regressão e incluir covariáveis no modelo, para assim estimar o efeito dessas covariáveis sobre a variável resposta. O modelo selecionado assume a distribuição que melhor descreve os dados, modelando a relação entre as covariáveis e a variável aleatória com base na função dessa distribuição e em seus parâmetros.

Entre os modelos paramétricos, a classe denominada tempo de vida acelerado é muito utilizada (Carvalho *et al*., 2011). Os modelos com as distribuições exponencial, Weibull, log-normal e log-logística são os mais utilizados. O modelo de tempo de vida acelerado relaciona linearmente o logaritmo do tempo às covariáveis, como pode ser observado na Equação 5:

$$\ln(T) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \sigma \ln(\varepsilon) \quad (5)$$

em que T é o tempo até o evento; $x_1, \ldots, x_p$ são as covariáveis; $\beta_1, \ldots, \beta_p$ são seus respectivos coeficientes ajustados; ε é o termo de erro que se presume ter uma distribuição paramétrica particular e σ é o parâmetro de escala.

As funções de densidade de probabilidade e funções de sobrevivência para as distribuições citadas podem ser vistas na Tabela 1.2.

Tabela 1.2.22 - Funções de densidade de probabilidade e de sobrevivência para as distribuições Exponencial, Weibull, Log-normal e Log-logística.

Distribuições	Função densidade de Probabilidade[1]	Função de Sobrevivência
Exponencial $\lambda > 0, t \geq 0$	$f(t) = \lambda\, exp\{-\lambda\, t\}$	$S(t\|x) = exp\left\{-\frac{t}{exp(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})}\right\}$
Weibull $\gamma, \lambda > 0, t \geq 0$	$f(t) = \lambda\gamma t^{\gamma-1} exp\{-\lambda t^{\gamma}\}$	$S(t\|x) = exp\left\{-\left(\frac{t}{exp(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})}\right)^{\frac{1}{\sigma}}\right\}$
Log-normal $\sigma > 0, t \geq 0$	$f(t) = \frac{1}{t\sigma\sqrt{2\pi}}\, exp\left\{-\frac{1}{2}\left(\frac{ln(t) - \mu}{\sigma}\right)^2\right\}$	$S(t\|x) = \Phi\left(\frac{-ln(t) + \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}}{\sigma}\right)$
Log-logística $\gamma, \lambda > 0, t \geq 0$	$f(t) = \frac{\gamma t^{\gamma-1}\lambda}{(1 + \lambda t^{\gamma})^2}$	$S(t\|x) = \frac{1}{1 + \left(\frac{t}{exp(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})}\right)^{\frac{1}{\sigma}}}$

[1] $\gamma = \frac{1}{\sigma}$; $\lambda = \frac{1}{\exp(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})}$; ln = Logaritmo natural; Φ = função de distribuição acumulada da distribuição normal padrão.

Além das distribuições pertencentes aos modelos de tempo de vida acelerado, podem ser ajustadas outras distribuições como a normal, logística, etc.

Na prática, a escolha da distribuição paramétrica a ser utilizada envolve a comparação do ajuste do modelo utilizando diversas distribuições distintas, como realizado em Martins Filho *et al*. (2023).

1.2.3. Técnicas Semiparamétricas

O modelo de regressão de Cox, ou modelos de taxas de falhas proporcionais (Cox, 1972), também possibilita a análise de dados em que a variável resposta é o tempo até a ocorrência do evento, incluindo covariáveis no modelo.

O modelo de taxas de falhas proporcionais de Cox, estima o efeito das covariáveis a partir da proporcionalidade entre as funções taxa de falha de dois grupos de indivíduos sem a necessidade de pressupor qualquer distribuição para o tempo de sobrevivência (Carvalho *et al*., 2011). Ele é classificado como uma técnica semiparamétrica pois é composto pelo produto de dois componentes, um não paramétrico, $h_0(t)$, e o outro paramétrico, $\exp(\boldsymbol{\beta}^T\boldsymbol{x})$. Esse modelo pode ser expresso pela Equação 6:

$$h_i(t) = h_0(t)\exp\{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\} \quad (6)$$

Têm-se que $h_i(t)$ é a taxa de falha para o grupo i em função do tempo t, $h_0(t)$ é a função taxa de falha de base e $\boldsymbol{\beta}$ é o vetor de parâmetros associados às covariáveis.

De acordo com Colosimo e Giolo (2024), o modelo de regressão de Cox também é denominado modelo de taxas de falha proporcionais pois a razão das taxas de falha para dois indivíduos quaisquer é constante ao longo do tempo.

Esse modelo é mais flexível se comparado aos modelos paramétricos, e para o termo $h_0(t)$ não é assumida nenhuma distribuição específica, sendo derivada diretamente dos dados (McNair *et al*., 2012), o que torna o modelo mais adequado para lidar com dados complexos. Além disso, os resultados podem ser facilmente interpretados em termos da razão de taxas de falhas proporcionais. Por esses

motivos, segundo Colosimo e Giolo (2024), o modelo de regressão de Cox passou a ser o mais utilizado em estudos clínicos.

Destaca-se a importância da suposição de proporcionalidade para a utilização efetiva da regressão de Cox. Essa suposição postula que a razão das taxas de falha para dois grupos quaisquer não depende do tempo e por isso as taxas de falha são proporcionais.

Assim, é crucial validar essa suposição para garantir a robustez dos resultados da análise (Struthers e Kalbfleisch, 1986). Normalmente, essa validação é conduzida por meio de testes formais, como o teste de correlação de Pearson (ρ) entre os resíduos padronizados de Schoenfeld (1982) e alguma função do tempo $g(t)$ para cada covariável, ou visualmente, por meio de gráficos do log da função taxa de falha acumulada *versus* o tempo. Valores de ρ próximos de zero indicam evidências a favor da não rejeição da suposição de proporcionalidade (Ng'andu, 1997). Curvas não paralelas sinalizam desvios dessa suposição, enquanto situações extremas de violação ocorrem quando as curvas se cruzam (Bogaerts *et al*., 2017).

Seja nos modelos paramétricos ou semiparamétricos, ao empregar abordagens baseadas em modelos probabilísticos, é necessário estimar os parâmetros desses modelos com base nas observações disponíveis, a fim de alcançar o modelo final. Na literatura, são abordados vários métodos para a estimação de parâmetros. Contudo, de acordo com Colosimo e Giolo (2024), quando se lida com dados censurados, o método da máxima verossimilhança emerge como uma escolha apropriada em contextos paramétricos, enquanto a máxima verossimilhança parcial – proposta por Cox (1972) - se destaca na abordagem semiparamétrica.

1.3. Análise de sobrevivência aplicada em germinação de sementes

Em estudos de germinação e emergência de sementes, o tempo até a emergência da radícula ou até que a plântula se torne visível acima do solo ou substrato pode ser uma variável de interesse. Dessa forma, a análise de sobrevivência se destaca como uma ferramenta poderosa para modelar e interpretar esses dados (Scott e Jones, 1982; Onofri *et al*., 2010; McNair *et al*., 2012; Manso *et al*., 2013).

A aplicação dessa abordagem permite considerar a presença de dados censurados, uma vez que o tempo exato de germinação ou emergência pode não ser

conhecido, resultando em dados incompletos. Nesse cenário, a germinação pode ser considerada como um processo qualitativo de resultado binário, onde 1 corresponde à ocorrência do processo de germinação (falha) e 0 à ausência desse evento (censura). Embora a censura seja altamente relevante em estudos de germinação, de acordo com Onofri *et al*. (2010), ela é frequentemente negligenciada e esse fato pode levar a resultados tendenciosos.

No contexto da germinação e emergência das sementes e levando em consideração o esquema de monitoramento empregado Onofri *et al*. (2011) definem a censura à direita quando as sementes ainda são viáveis, mas o tempo de germinação ou emergência é maior do que a duração do experimento; nesse caso, elas devem ser incluídas na análise estatística, pois ainda podem germinar. A censura à esquerda quando, no primeiro momento de avaliação, algumas sementes já germinaram ou emergiram, de modo que só se sabe que o tempo desse evento ocorreu antes da primeira avaliação. E a censura intervalar quando o tempo exato de germinação ou emergência não é conhecido precisamente, mas está entre duas datas sucessivas de monitoramento.

Além disso, Onofri et al. (2011) destacam que a relevância prática dessas formas de censura pode variar de acordo com o tipo de monitoramento utilizado. Em ensaios de germinação de sementes com avaliações regulares e frequentes, como monitoramentos diários, é provável que apenas a censura à direita tenha um impacto significativo nas análises. Em contrapartida, em esquemas de monitoramento mais espaçados e irregulares, tanto a censura à esquerda quanto a censura intervalar devem ser consideradas.

A análise de sobrevivência é uma técnica estatística eficiente para explorar como diferentes fatores afetam o processo de germinação de sementes, aplicando uma modelagem adequada para dados que consideram o tempo até a ocorrência de eventos, como a própria germinação. Entre os fatores frequentemente analisados estão a temperatura (Cristaudo et al., 2014; Solarik et al., 2016), o tempo de armazenamento (Cristaudo et al., 2014), as características físicas das sementes (Barak et al., 2018) e o envelhecimento acelerado (Adegbola e Pérez, 2016; Genna e Pérez, 2016). Além disso, a análise de sobrevivência também permite avaliar a influência da dispersão de sementes por animais (Tang et al., 2012; Harich et al., 2016; Ortega-Flores et al., 2018), ampliando as possibilidades de interpretação sobre as interações entre plantas e animais, bem como a dinâmica reprodutiva das espécies.

Essa modelagem utiliza a função de sobrevivência, que, neste contexto, representa a probabilidade de uma semente não ter germinado até um determinado momento. Para construir um gráfico que ilustre a probabilidade de germinação ao longo do tempo, é necessário recorrer à função de sobrevivência acumulada. A função de risco, por sua vez, descreve a probabilidade instantânea de germinação em cada momento, desde que a semente ainda não tenha germinado

1.4. Censura Intervalar

A censura intervalar é frequentemente encontrada em estudos longitudinais de saúde pública, ensaios clínicos de longo prazo e pesquisas de campo que envolvem coleta de dados em intervalos espaçados. Essa forma de censura pode resultar de limitações logísticas, financeiras ou práticas na observação contínua dos eventos de interesse.

Segundo Lindsey e Ryan (1998), os casos de censura com tempos exatos de falha, censura à esquerda e censura à direita, são casos especiais de dados de sobrevivência intervalar. Quando os tempos de falhas são exatos, observa-se que esse intervalo possui limite inferior e limite superior iguais, ou seja, $L = T = U$. Em situações de censura à direita, o limite inferior do intervalo corresponde à última data de análise, enquanto o limite superior se estende até o infinito, ou seja, $T \in (L, \infty)$, onde L representa o tempo decorrido desde o início do estudo até a última avaliação. Na censura à esquerda, o tempo inferior é zero, e o limite superior é estabelecido como o dia de início do experimento, ou seja, $T \in (0, U]$, onde U representa o tempo decorrido entre o início do estudo até a primeira avaliação.

Uma abordagem alternativa comum para analisar os dados é assumir que o evento ocorreu no final, no início ou no ponto médio de cada intervalo e, em seguida, aplicar métodos para tempo de falha exato. No entanto, esta abordagem pode conduzir a vieses e levar a inferências inválidas, principalmente quando se trata de intervalos de maior comprimento (Rücker e Messerer, 1988; Brookmeyer e Goedert, 1989; Odell *et al*. 1991).

Contudo, dentro da análise de censura intervalar, existem abordagens específicas para estimar a função de sobrevivência, categorizadas em técnicas não paramétricas, paramétricas e semiparamétricas.

1.4.1. Estimador não paramétrico de máxima verossimilhança sob censura intervalar

A busca pelo estimador não paramétrico (NP) da função de sobrevivência sob censura intervalar demanda a criação de um conjunto de intervalos. Esses intervalos são definidos com base em uma sequência de tempos que abrange todos os extremos dos intervalos esquerdo e direito, variando de 0 até o m-ésimo termo. São conhecidos como intervalos de Turnbull, ou regiões de possível massa e são representados por $(p_j\ ,\ q_j]$, para $j\ =\ 1,\ldots,m$. Segundo Anderson-Bergman (2017), foi demonstrado que são esses os intervalos onde se atribui a probabilidade de ocorrência de eventos.

Essa abordagem de identificação de intervalos de possível massa é fundamental para a estimativa não paramétrica da função de sobrevivência em dados censurados por intervalo, permitindo concentrar a análise nos intervalos onde eventos podem ocorrer, o que simplifica o processo de estimativa da função de sobrevivência.

Peto (1973) observou que a probabilidade atribuída a cada um desses intervalos é bem determinada. No entanto, dentro de cada intervalo, não há informações sobre como essa probabilidade é distribuída, podendo afirmar que há a possibilidade de eventos ocorrerem dentro desses intervalos, mas não é possível garantir que eventos de fato ocorrerão neles.

Para determinar as regiões de possíveis massa Peto (1973) e Turnbull (1976) propuseram o seguinte algoritmo. Considere $(L_i\ ,\ U_i]\ (i\ =\ 1,\ldots,n)$, os intervalos de tempo que ocorreram o evento de interesse para n indivíduos analisados. Todos os pontos de extremidade (início ou fim dos intervalos) são ordenados em ordem crescente. Além disso, é registrado se cada ponto é o início (esquerdo) ou o fim (direito) de um intervalo. Os intervalos de possível massa são os intervalos que possuem um ponto de extremidade esquerda seguido imediatamente por um ponto de extremidade direita.

A próxima etapa do processo de estimativa da função de sobrevivência é estimar a probabilidade atribuída a cada um desses intervalos. Define-se então o vetor $s\ =(s_1,\ldots,s_m)^T$, onde cada $\mathrm{s_j}$ é calculado como $\mathrm{s_j} = \mathrm{P}\left(\mathrm{p_j} \leq \mathrm{T} \leq \mathrm{q_j}\right) = \mathrm{S}\left(\mathrm{p_j}\right) - \mathrm{S}\left(\mathrm{q_j}\right)$. Dessa forma, a verossimilhança (Equação 7) depende apenas dos valores em p_j e $q_j\ (j\ =\ 1,\ldots,m)$ e não de como a função evolui entre esses pontos, assim a

busca pelo Estimador de Máxima Verossimilhança (MLE) da função S pode ser restrita a esses intervalos.

Portanto, em um estudo com n indivíduos, para determinar a MLE da função S, é necessário maximizar a função de verossimilhança:

$$L = \prod_{i=1}^{n} \left(\sum_{j=1}^{m} \alpha_{ij} s_j \right) \tag{7}$$

em que α_{ij} é uma variável indicadora cujo valor assume 1, se o intervalo $(L_i\ ,\ U_i]$ contém o intervalo de observação $(p_j\ ,\ q_j]$ e 0 caso contrário.

É importante notar que o NPMLE para a função de sobrevivência corresponde à não ocorrência do evento dentro cada intervalo de Turnbull $(p_j\ ,\ q_j]$, ou seja, a sobrevivência para um tempo t pertencente ao intervalo j de Turnbull é $1 - s_j$, não sendo especificada dentro do intervalo.

No contexto da estimativa não paramétrica da função de sobrevivência, o objetivo é estimar essa função sem assumir uma forma funcional específica para a distribuição dos dados. Isso implica não fazer suposições sobre a forma da função de distribuição de probabilidade. Para alcançar esse objetivo, são utilizados algoritmos iterativos como o algoritmo EM, ICM ou a combinação de ambos (Bogaerts et al., 2017).

O algoritmo convencional é o algoritmo de Turnbull (1976), uma generalização do estimador de Kaplan-Meier, que é uma aplicação do algoritmo EM (Dempster *et al.*, 1977). Este procedimento consiste nos seguintes passos:

1) Os valores de $\mathrm{s_j}$ iniciais são calculados pelos valores estimados para $\mathrm{S(p_j)}$ e $\mathrm{S(q_j)}$ por Kaplan-Meier, considerando os tempos de falha de $\mathrm{p_j}$ e $\mathrm{q_j}$.

2) Estimar o número de eventos ocorridos em $\mathrm{q_j}$ (Equação 8).

$$d_j = \sum_{i=1}^{n} \frac{\alpha_{ij} s_j}{\sum_{k=1}^{m} \alpha_{ik}\, s_k} \tag{8}$$

3) Obter o número estimado em risco no tempo q_j (Equação 9).

$$n_j = \sum_{k=j}^{m} d_k \tag{9}$$

4) Atualizar o estimador produto-limite usando os resultados dos passos 2 e 3 (Equação 10).

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right) \tag{10}$$

Se a estimativa estiver próxima da anterior para todo q_j o procedimento termina, caso contrário deve se repetir os quatro passos anteriores com as novas estimativas.

Conforme Anderson-Bergman, o algoritmo EM demonstrou lentidão em relação ao número de iterações. Como alternativa para acelerar o processo, foi desenvolvido o algoritmo ICM (Groeneboom e Wellner, 1992). Esse algoritmo foi originalmente formulado utilizando processos estocásticos e algoritmos de regressão isotônica, mas aqui será apresentado como uma especificação do método de otimização conhecido como GGP (Bertsekas, 1982; Mangasarian, 1996), com o objetivo de estendê-lo ao modelo semiparamétrico.

O algoritmo ICM iterativamente atualiza as estimativas da função de distribuição acumulada até convergir para uma solução que maximiza a log-verossimilhança dos dados censurados por intervalo conforme definida pela Equação 11:

$$L = \sum_{i=1}^{n} \log[\{1 - F(L_i)\} - \{1 - F(U_i)\}] \tag{11}$$

Para alcançar essa solução, o algoritmo utiliza informações do gradiente e da Hessiana da função de verossimilhança, garantindo que as atualizações preservem a validade da função de distribuição acumulada.

O valor inicial da função de distribuição acumulada F pode ser calculada tratando os dados como censurados à direita. Assim, o primeiro passo consiste em calcular a primeira derivada parcial (gradiente) ∇L e a segunda derivada parcial (hessiano) da função de verossimilhança L em relação à F, para então construir uma matriz diagonal G a partir dos valores negativos da matriz hessiana da função de verossimilhança.

Em seguida, a estimativa da função de distribuição acumulada é atualizada (Equação 12). Uma projeção no intervalo restrito R ponderada por G (Equação 13) é utilizada para garantir que $F^{(m+1)}$ seja novamente uma função de distribuição:

$$F^{(m+1)} = Proj\,[F^{(m)} + \mathrm{G}^{-1}\nabla L\,,\mathrm{G}, R] \tag{12}$$

$$Proj[y, G, R] = \arg\min \sum_{i=1}^{k} (y_i - s_i)^2\, G_{ii} : 0 \leq s_1 \leq \cdots \leq s_k \leq 1 \tag{13}$$

em que $R = \{\, F: 0 \leq F(s_1) \leq \cdots \leq F(s_k) \leq 1 \,\}$ para $s_1, \ldots, s_k$ os pontos em que F pode ter saltos, sendo determinados pelos intervalos de Turnbull.

$Proj$ é apenas um problema de regressão por mínimos quadrados isotônica e, portanto, pode ser eficientemente realizado por alguns algoritmos bem conhecidos, como o algoritmo de violadores adjacentes ao grupo (PAVA) (Robertson, Wright e Dykstra, 1988).

O processo de iteração, desde o cálculo do gradiente e do hessiano, é repetido até que a estimativa convirja para um valor estável. Isso pode ser determinado monitorando a mudança na estimativa entre iterações sucessivas e comparando-a com um critério de convergência predefinido.

Apesar de possuir vantagens, esse algoritmo apresentou um desempenho inferior em casos em que o algoritmo EM tinha um desempenho excelente, especialmente quando havia muitas observações não censuradas.

A combinação dos algoritmos EM e ICM para a formação do EMICM (Wellner e Zhan, 1997) levou a uma melhoria de desempenho. Nesse método, a etapa ICM do

algoritmo busca a NPMLE no conjunto de todas as estimativas autoconsistentes especificadas pelas iterações EM.

Com a finalidade de acelerar este algoritmo, Anderson-Bergman (2017) propôs uma implementação eficiente do algoritmo EMICM. Em seu artigo, ele mostrou que, embora teoricamente semelhante ao algoritmo original, essa nova implementação é consideravelmente mais rápida que as alternativas do EMICM e outros algoritmos concorrentes. Isso permite a análise de conjuntos de dados com várias ordens de magnitude maiores do que era possível anteriormente.

1.4.1.1. Teste de Fleming-Harrington para dados com censura intervalar

O teste de hipótese de Fleming e Harrington é uma versão do teste de log-rank ponderado, utilizado para comparar duas ou mais funções de sobrevivência. Proposto por Fleming e Harrington (1991), este teste é aplicável apenas a dados censurados à direita. Para contornar essa limitação, sabendo que os parâmetros ρ e γ indicam a sensibilidade do teste às variações nos riscos de ocorrência dos eventos, ao longo dos diferentes períodos de acompanhamento, Oller e Gómez, em 2012, propuseram uma extensão desse teste para dados censurados por intervalo.

Sejam $\rho \geq 0, \gamma \geq 0$. Para $j = 1, \ldots, J$, têm-se definido $w_j^{\rho,\gamma}$ (Equação 14) como os pesos que especificam os testes:

$$w_j^{\rho,\gamma} = \hat{S}(p_j)\frac{B\left(1 - \hat{S}(q_j);\, \gamma + 1, \rho\right) - B\left(1 - \hat{S}(p_j);\, \gamma + 1, \rho\right)}{\hat{S}(p_j) - \hat{S}(q_j)} \qquad (14)$$

em que, $B(y; a, b) = \int_0^y x^{a-1}\,(1-x)^{b-1}\,dx$ é uma função beta incompleta, para $y \geq 0, a > 0, b \geq 0$.

Assim como ocorre com dados censurados à direita, é possível desenvolver testes sensíveis para comparar curvas em qualquer período do tempo, avaliado por meio de escolhas apropriadas dos valores de ρ e γ. O parâmetro ρ controla a ênfase dada às diferenças na taxa de falha no início do estudo, enquanto γ destaca variações no meio ou no final (Oller e Langohr, 2017).

Por exemplo, em um ensaio de germinação, para investigar se uma temperatura afeta mais a sobrevivência no início, escolhe-se $\gamma = 0$ e aumenta-se ρ para destacar diferenças mais pronunciadas nesse período. Se for previsto um efeito mais evidente no meio ou no final do estudo, opta-se por $\rho = \gamma > 0$ ou $\rho = 0$, aumentando γ para enfatizar essas fases.

Quando $\rho = 0$ e $\gamma = 0$, o teste adota uma abordagem semelhante ao teste log-rank clássico, que é apropriado quando não há necessidade de ajustar para variações nas taxas de falha ao longo do tempo. Isso permite a comparação de curvas de sobrevivência entre grupos sem considerar diferenças específicas em diferentes fases do acompanhamento. Além disso, quando $\rho = 1$ e $\gamma = 0$, o teste se assemelha ao teste de Prentice-Wilcoxon, que é sensível a diferenças no início das distribuições de sobrevivência (Bogaerts *et al*. 2017).

Essa flexibilidade na escolha dos parâmetros ρ e γ permite que os pesquisadores ajustem os testes de sobrevivência de acordo com as características específicas dos dados e as hipóteses de pesquisa, maximizando assim a capacidade de detectar diferenças significativas entre os grupos estudados.

Como foi apresentado anteriormente para dados sem censura intervalar, a estatística do teste log-rank ponderado é baseada na diferença entre os números de eventos observados e esperados (numerador da Equação 4), sob hipótese nula de que não há diferença entre as duas funções de sobrevivência.

Esses valores são obtidos para cada tempo de falha j da amostra combinada obtida pela fusão dos tempos de sobrevivência de ambos os grupos.

Para comparar três ou mais (k) curvas, pode-se estender esse procedimento e calcular o vetor $\boldsymbol{U} = (U_1, \ldots, U_K)^T$ (Equação 15). Como D_{jk} segue uma distribuição hipergeométrica, pode-se afirmar que $E_{kj} = \frac{n_{jk}}{nj} D_j$. Dessa forma o k-ésimo componente do vetor $\boldsymbol{U}$ é uma diferença ponderada integrada entre as funções de risco estimadas da k-ésima curva e do geral (Equação 16).

$$U_K = \sum_{j=1}^{J} w_j \left(D_{jk} - E_{kj} \right) \tag{15}$$

$$U_K = \sum_{j=1}^{J} w_j \, n_{jk} \left(\frac{D_{jk}}{n_{jk}} - \frac{D_j}{nj} \right) \tag{16}$$

Na comparação de duas ou mais curvas de sobrevivência em dados censurados por intervalo, a estatística do teste log-rank ponderado compara o número de eventos observados e esperados em cada região de suporte da função de sobrevivência, estimada pelo método não paramétrico (NPMLE) na amostra combinada. De forma geral, a estatística do teste utilizado para esses cenários (Fay, 1999) também pode ser definida pelo vetor $\boldsymbol{U} = (U_1, \ldots, U_K)^T$ (equação 16 e equação 17).

Nesse caso, D_{jk} (Equação 17) e n_{jk} (Equação 18) são as estimativas do número total de eventos no k-ésimo grupo ocorrendo no j-ésimo intervalo $(p_j \, , \, q_j]$ e o número em risco imediatamente antes desse intervalo, respectivamente.

$$D_{jk} = N_k \{ \hat{\mathrm{S}}_k(p_j) - \hat{\mathrm{S}}_k(q_j) \} \tag{17}$$

e

$$n_{jk} = N_k \, \hat{\mathrm{S}}_k(p_j) \tag{18}$$

Enquanto D_j (Equação 19) e nj (Equação 20) representam essas mesmas quantidades na amostra combinada referente ao j-ésimo intervalo $(p_j \, , \, q_j]$.

$$D_j = n \, \{ \hat{\mathrm{S}}(p_j) - \hat{\mathrm{S}}(q_j) \} \tag{19}$$

e

$$n_j = n \, \hat{\mathrm{S}}(p_j) \tag{20}$$

1.4.2. Técnica Paramétrica sob censura intervalar

Os modelos paramétricos já apresentados na seção 1.2.2 também são aplicados na censura intervalar. Esses modelos assumem uma distribuição específica para os tempos de sobrevivência e usam parâmetros para descrever essa distribuição. Conforme destacado por Bogaerts *et al*. (2017), sob as premissas adequadas, a função paramétrica de sobrevivência estimada tende a proporcionar resultados mais precisos em comparação com o NPMLE.

Assim, como antes, o método da máxima verossimilhança é utilizado para estimar os parâmetros do modelo especificado. Segundo Sun (2006), a principal vantagem das abordagens paramétricas é que a sua implementação é simples em princípio e a teoria padrão da máxima verossimilhança geralmente se aplica. Nesse caso, para a construção da função de verossimilhança a natureza intervalar dos dados deve ser levada em consideração analisando qual a contribuição de cada indivíduo (Tabela 1.3).

Tabela 1.4.23 - Contribuições de cada indivíduo para a função de verossimilhança na estimação dos parâmetros dos modelos paramétricos sob censura intervalar.

Indivíduo	T	Contribuição
Com tempo exato de falha	$T = t$	$f(t)$
Censurado à direita	$T \in (L, \infty)$	$S(l)$
Censurado à esquerda	$T \in (0, U]$	$1 - S(u)$
Com tempo intervalar	$T \in (L, U]$	$S(l) - S(u)$

Dessa forma, a função de verossimilhança em relação ao vetor de parâmetros ($\boldsymbol{\theta}$) do modelo, para cada indivíduo i, com $T_i \in (L_i, U_i]$ ou $T_i \in (L_i, \infty]$, fica definida pela Equação 21:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [S(l_i|\boldsymbol{x}_i) - S(u_i|\boldsymbol{x}_i)]$$

$$L\,(\boldsymbol{\theta}) = \prod_{i=1}^{n}[S(l_i|\boldsymbol{x}_i) - S(u_i|\boldsymbol{x}_i)]^{\delta_i}\,[S(l_i|\boldsymbol{x}_i)]^{1-\delta_i} \qquad (21)$$

em que $\boldsymbol{x}_i$ é o vetor de covariáveis associado ao i-ésimo indivíduo; δ_i uma variável indicadora que assume 1 se o evento ocorreu no intervalo $(L_i\ ,\ U_i]$, ou assume 0, se não ocorreu até L_i.

Para selecionar o melhor modelo paramétrico, de acordo com Bogaerts *et al* (2017), é possível usar critérios de informação como o AIC - Critério de Informação de Akaike (Akaike, 1973) e o BIC - Critério de Informação Bayesiano (Schwarz, 1978). Tanto o AIC quanto o BIC são medidas que penalizam a complexidade do modelo, o que significa que modelos mais simples e parcimoniosos são favorecidos.

Na prática, pode-se ajustar vários modelos com diferentes distribuições e, em seguida, comparar seus valores de AIC ou BIC. O modelo com a menor dessas estimativas é considerado aquele que melhor se ajusta aos dados (Emiliano *et al*., 2009).

1.4.3. Técnica Semiparamétrica sob censura intervalar

Assim como nos modelos com tempo exato de falha, o modelo semiparamétrico de Cox emerge como uma alternativa para conjuntos de dados com censura intervalar em que não foi possível encontrar um modelo paramétrico que se ajustasse bem.

O modelo de taxas de falhas proporcionais de Cox (Cox, 1972), adaptado para a censura intervalar, mantém a mesma estrutura básica apresentada na seção 1.2.3, com modificações necessárias para lidar com a censura. O ajuste do modelo considera a contribuição de cada indivíduo para a função de verossimilhança, levando em conta a incerteza nos tempos de falha (Finkelstein, 1986).

No contexto intervalar a contribuição de cada indivíduo para a função de verossimilhança é a que está especificada na Tabela 3. Dessa forma, a função de verossimilhança é definida pela Equação 22:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}[S(l_i|\boldsymbol{x_i}) - S(u_i|\boldsymbol{x}_i)]$$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}\left[[S_0(l_i)]^{\exp\{\boldsymbol{\beta^T x_i}\}}\right] - \left[[S_0(u_i)]^{\exp\{\boldsymbol{\beta^T x_i}\}}\right]$$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}\left[[1 - F_0(l_i)]^{\exp\{\boldsymbol{\beta^T x_i}\}}\right] - \left[[1 - F_0(u_i)]^{\exp\{\boldsymbol{\beta^T x_i}\}}\right] \quad (22)$$

em que $S_0(.)$ é a função de sobrevivência de base e $F_0(.)$ é a respectiva função de distribuição acumulada de base.

Para maximizar $\mathcal{L}(\boldsymbol{\theta}) = ln\, L(\boldsymbol{\theta})$ com $\boldsymbol{\theta} = (F_0, \boldsymbol{\beta})$ dessa função de verossimilhança, uma extensão do algoritmo ICM para o modelo de Cox, no contexto de dados censurados por intervalo, foi proposta por Pan (1999).

O algoritmo ICM original, foi desenvolvido para dados intervalares sem covariáveis. Por outro lado, o algoritmo ICM estendido para o modelo de Cox é uma adaptação desse método para incorporar covariáveis no modelo (Bogaerts *et al*., 2017). No contexto do modelo de Cox, o objetivo é maximizar a verossimilhança do modelo, realizando iterações conjuntas para estimar a distribuição acumulada da variável de interesse (equação 23) e os parâmetros de regressão das covariáveis (equação 24).

Da mesma forma que o valor inicial da função de distribuição acumulada F_0 pode ser calculada tratando os dados como censurados à direita, os valores dos parâmetros da regressão também podem ser determinados dessa maneira.

Para o ICM estendido é necessário calcular os gradientes $\nabla_1 L$ e $\nabla_2 L$, da função de verossimilhança L em relação à F_0, e em relação aos parâmetros da regressão. É necessário também calcular as matrizes hessianas para a construção de G_1 e G_2.

Em seguida atualiza-se a estimativa da função de distribuição acumulada e dos parâmetros, através das Equações 23 e 24:

$$F_0{}^{(m+1)} = Proj\,[F_0{}^{(m)} + \alpha^{(m)} \mathrm{G_1}^{-1} \nabla_1 L\,, \mathrm{G_1}, R\,] \quad (23)$$

e

$$\beta^{(m+1)} = \beta^{(m)} + \alpha^{(m)} \mathrm{G}_2{}^{-1} \nabla_2 L \tag{24}$$

$Proj$ é uma operação de projeção que garante que a nova estimativa ${F_0}^{(m+1)}$ seja uma função de distribuição adequada. Têm-se também que $\alpha^{(m)}$ é uma constante e pode ser escolhida utilizando a fórmula $\alpha^{(m)} = \max\{\frac{1}{2^i} : L({F_0}^{m+1}, \beta^{m+1}) > L({F_0}^{m}, \beta^{m}),\ i = 0,1,2,\dots\}$.

1.4.4. Dados Grupados

Os dados grupados são um caso específico de censura intervalar em que todos os indivíduos da amostra são avaliados sempre nos mesmos intervalos de tempo. Esse tipo de dado é caracterizado por possuir muitos empates, ou seja, ocorrem mais de uma falha com o mesmo tempo.

Segundo Colosimo e Giolo (2024), existem na literatura três propostas para trabalhar com dados grupados: (1) utilizar a função de verossimilhança parcial exata no contexto do modelo de taxas de falha proporcionais; (2) utilizar aproximações para a função de verossimilhança parcial no contexto do modelo de taxas de falha proporcionais; (3) utilizar modelos de regressão discretos.

Devido à ocorrência de empates, a ordenação crescente dos dados de tempo de falha torna-se inviável, tornando a estimativa do vetor de parâmetros associados às covariáveis, $\boldsymbol{\beta}$, uma tarefa de extrema complexidade. Para abordar essa questão, foram propostas aproximações para adaptar a função de verossimilhança parcial (Breslow, 1972; Peto, 1972; Efron, 1977; Farewell e Prentice, 1980), permitindo assim a incorporação dos empates no modelo de taxa de falha proporcional.

Quando existe um alto número de empates, a saída é reconhecer a natureza discreta dos tempos de falha, para isso, pode-se assumir que os tempos latentes de falha vêm de um modelo de taxas de falha proporcionais contínuo (Prentice e Gloeckler, 1978), de um modelo de chances proporcionais (Hosmer e Lemeshow, 2000) ou de um modelo log-normal generalizado (Silveira *et al*., 2010).

As aproximações à função de verossimilhança parcial são mais apropriadas quando há poucos empates (Colosimo e Giolo, 2024). No entanto, determinar com precisão a quantidade de empates pode ser uma tarefa desafiadora. Chalita *et al*.

(2002) propuseram uma regra empírica, descrita na Equação 25, que oferece um critério prático para a tomada de decisão.

$$pe = \frac{d - k}{n} \tag{25}$$

em que pe representa a proporção de empates, d é o número de falhas, k é o número de falhas distintas e n é o total de indivíduos.

Se pe for menor que 20% então deve ser utilizado o modelo contínuo com aproximações para a função de verossimilhança parcial. Se pe estiver na faixa de 20% a 25%, ambos os modelos podem ser considerados. Contudo, se pe for superior a 25%, é aconselhável adotar um modelo discreto.

1.4.4.1. Regressão Discreta

Ao agrupar os tempos de vida em k intervalos $I_j = [t_{j-1}, t_j)$, a função de verossimilhança é escrita em termos da probabilidade de falha do i-ésimo indivíduo no intervalo I_j, dado as covariáveis e dado que esse indivíduo estava vivo em t_{j-1}, como definido na Equação 26:

$$p_j(\boldsymbol{x}_i) = P\left[T_i \leq t_j \middle| T_i > t_{j-1}, \boldsymbol{x}_i\right] \tag{26}$$

Como a contribuição de uma observação censurada e de uma observação não censurada para a função de verossimilhança, são, respectivamente $[\{1 - p_1(\boldsymbol{x}_i)\} \ldots \{1 - p_{j-1}(\boldsymbol{x}_i)\}]$ e $[\{1 - p_1(\boldsymbol{x}_i)\} \ldots \{1 - p_{j-1}(\boldsymbol{x}_i)\}]\, p_j(\boldsymbol{x}_i)$, a função de verossimilhança $L(\theta)$ é dada pela Equação 27:

$$L(\theta) = \prod_{j=1}^{k} \prod_{i \in R_j} \{p_j(\boldsymbol{x}_i)\}^{\delta_{ij}} \{1 - p_j(\boldsymbol{x}_i)\}^{(1 - \delta_{ij})} \tag{27}$$

em que R_j é o conjunto de indivíduos em risco em t_{j-1}; e δ_{ij} é uma variável indicadora para o tempo de vida, $\delta_{ij} = 0$, se for censurado e $\delta_{ij} = 1$ caso contrário.

A probabilidade $p_j(\boldsymbol{x}_i)$ na estrutura da regressão discreta, é comumente modelada empregando um modelo de riscos proporcionais ou um modelo de chances proporcionais na função de verossimilhança, resultando em modelos para dados grupados em sobrevivência.

Considerando o modelo de riscos proporcionais, tem-se $p_j(\boldsymbol{x}_i)$ definida pela equação 28:

$$p_j(\boldsymbol{x}_i) = 1 - \gamma_j^{\exp\{\boldsymbol{\beta}^T \boldsymbol{x}_i\}} \tag{28}$$

em que $\gamma_j = \frac{S_0(t_j)}{S_0(t_{j-1})}, j = 1, \ldots, k$ e S_0 é a função de sobrevivência de linha de base.

Ao adotar a abordagem de chances proporcionais, tem-se $p_j(\boldsymbol{x}_i)$ definida pela equação 29:

$$p_j(\boldsymbol{x}_i) = 1 - \{\, 1 + \gamma_j \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i) \,\}^{-1} \tag{29}$$

em que $\gamma_j = \frac{p_j(0)}{1 - p_j(0)}, j = 1, \ldots, k$.

Os modelos de riscos proporcionais e de chances proporcionais podem ser linearizados utilizando diferentes transformações. No modelo de riscos proporcionais, isso é alcançado por meio da função de ligação complemento log-log (cloglog), ou seja, $ln\,[-ln\,(1 - p_j(\boldsymbol{x}_i)]$. Enquanto no modelo de chances proporcionais, é utilizada a função de ligação logito, $ln\left(\frac{p_j(\boldsymbol{x}_i)}{1 - p_j(\boldsymbol{x}_i)}\right)$. Assim, esses modelos podem ser ajustados empregando métodos usuais para modelagem de dados binários.

Referências

Adegbola, Y. U., Pérez, H. E., 2016. Extensive desiccation and aging stress tolerance characterize *Gaillardia pulchella* (Asteraceae) seeds. **HortScience**, 51(2), 159-163. https://doi.org/10.21273/HORTSCI.51.2.159

Akaike, H., 1973. Maximum likelihood identification of Gaussian autoregressive moving average models. **Biometrika**, 60(2), 255-265. https://doi.org/10.1093/biomet/60.2.255

Anderson-Bergman, C., 2017. An efficient implementation of the EMICM algorithm for the interval censored NPMLE. **Journal of Computational and Graphical Statistics**, *26*(2), 463-467. https://doi.org/10.1080/10618600.2016.1208616

Andrade, R. A. D., Oliveira, I. V. D. M., Martins, A. B. G., 2005. Influência da condição e período de armazenamento na germinação de sementes de pitaya vermelha. **Revista Brasileira de Fruticultura**, *27*, 168-170.

Barak, R. S., Lichtenberger, T. M., Wellman-Houde, A., Kramer, A. T., Larkin, D. J., 2018. Cracking the case: Seed traits and phylogeny predict time to germination in prairie restoration species. **Ecology and Evolution**, *8*(11), 5551-5562. https://doi.org/10.1002/ece3.4083

Bogaerts, K., Komarek, A., Lesaffre, E., 2017. **Survival analysis with interval-censored data: a practical approach with examples in R, SAS, and BUGS.** Chapman and Hall/CRC.

Breslow, N. E., 1972. Discussion of Professor Cox's paper. **Journal of the Royal Statistics B**, 34, 216-217.

Brookmeyer, R., Goedert, J. J., 1989. Censoring in an epidemic with an application to hemophilia-associated AIDS. **Biometrics**, 325-335. https://doi.org/10.2307/2532057

Carvalho, M. S., Andreozzi, V. L., Codeço, C. T., Campos, D. P., Barbosa, M. T. S., Shimakura, S. E., 2011. **Análise de sobrevivência: teoria e aplicações em saúde**. SciELO-Editora FIOCRUZ.

Chalita, L. V., Colosimo, E. A., Demétrio, C. G., 2002. Likelihood approximations and discrete models for tied survival data. **Communications in Statistics-Theory and Methods**, 31(7), 1215-1229. https://doi.org/10.1081/STA-120004920

Chen, J., Sabir, I. A., Qin, Y., 2023. From challenges to opportunities: Unveiling the secrets of pitaya through omics studies. **Scientia Horticulturae**, 321, 112357. https://doi.org/10.1016/j.scienta.2023.112357

Colosimo, E. A., Giolo, S. R., 2024. **Análise de sobrevivência aplicada**. 2. ed. São Paulo: Editora Blücher. 362p.

Cox, D. R., 1972. Regression models and life-tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, 34(2), 187-202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Cristaudo, A., Gresta, F., Restuccia, A., Catara, S., Onofri, A., 2016. Germinative response of redroot pigweed (Amaranthus retroflexus L.) to environmental conditions: Is there a seasonal pattern? **Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology**, 150(3), 583-591. https://doi.org/10.1080/11263504.2014.987845

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the royal statistical society: series B (methodological)**, 39(1), 1-22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Duarte, M. L., Martins Filho, S., de Freitas, A. F., Xavier, A., 2023. Rooting of forest species mini-cuttings: an application of non-parametric survival analysis. **New Forests**, 54(6), 1153-1167. https://doi.org/10.1007/s11056-023-09962-0

Efron, B., 1977. The efficiency of Cox's likelihood function for censored data. **Journal of the American statistical Association**, 72, 557-565. https://doi.org/10.1080/01621459.1977.10480613

Emmert-Streib, F., Dehmer, M., 2019. Introduction to survival analysis in practice. **Machine Learning and Knowledge Extraction**, 1(3), 1013-1038. https://doi.org/10.3390/make1030058

Emiliano, P. C., Vivanco, M., Menezes, F. S. D., Avelar, F. G., 2009. Foundations and comparison of information criteria: Akaike and Bayesian. **Revista Brasileira de Biometria**, 27(3), 394-411.

Faleiro, F., 2022. Pitaia: a fruta que está conquistando o Brasil. **Anuário Campo & Negócios Hortifruti**, 11, 97–99. http://www.alice.cnptia.embrapa.br/alice/handle/doc/1152429

Farewell, V. T., Prentice, R. L., 1980. The approximation of partial likelihood with emphasis on case-control studies. **Biometrika**, 67, 273-278. https://doi.org/10.1093/biomet/67.2.273

Fay, M. P., 1999. Comparing several score tests for interval censored data. **Statistics in Medicine**, 18(3), 273-285. https://doi.org/10.1002/(SICI)1097-0258(19990215)18:3<273::AID-SIM19>3.0.CO;2-7

Fenner, M., Thompson, K., 2005. **The ecology of seeds**. Cambridge university press.

Fernandes, A. C., Coutinho, G., 2019. Nitrogénio no desenvolvimento inicial de mudas de pitaya vermelha. **Global Science & Technology**, 12(3).

Fleming, T. R., Harrington, D. P., 1991. **Counting processes and survival analysis** (Vol. 625). John Wiley & Sons, New York.

Finkelstein, D. M., 1986. A proportional hazards model for interval-censored failure time data. **Biometrics**, 845-854. https://doi.org/10.2307/2530698

Gehan, E. A., 1965. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. **Biometrika**, 52(1-2), 203-224.https://doi.org/10.1093/biomet/52.1-2.203

Genna, N. G., Pérez, H. E., 2016. Mass-based germination dynamics of Rudbeckia mollis (Asteraceae) seeds following thermal and ageing stress. **Seed Science Research**, 26(3), 231-244. https://doi.org/10.1017/S0960258516000180

George, B., Seals, S., Aban, I., 2014. Survival analysis and regression models. **Journal of nuclear cardiology**, 21(4), 686-694. https://doi.org/10.1007/s12350-014-9908-2

Groeneboom, P., and Wellner, J. A., 1992. **Information bounds and nonparametric maximum likelihood estimation**. Birkhäuser-Verlag.

Hampton, J. G., Tekrony, D. M., 1995. **Handbook of vigour test methods**. 3rd Edition, The International Seed Testing Association, Zurich (Switzerland).

Harich, F. K., Treydte, A. C., Ogutu, J. O., Roberts, J. E., Savini, C., Bauer, J. M., Savini, T., 2016. Seed dispersal potential of Asian elephants. **Acta Oecologica**, 77, 144-151. https://doi.org/10.1016/j.actao.2016.10.005

Hernández, Y. D. O., Salazar, J. A. C., 2012. Pitahaya (*Hylocereus* spp.): a short review. **Comunicata Scientiae**, *3*(4), 220-237.

Hosmer, D. W., Lemeshow, S., 2000. **Applied Logistic Regression**. John Wiley and Sons, New York, 2nd edition.

Hua, Q., Chen, P., Liu, W., Ma, Y., Liang, R., Wang, L., Qin, Y., 2015. A protocol for rapid in vitro propagation of genetically diverse pitaya. **Plant Cell, Tissue and Organ Culture (PCTOC)**, 120, 741-745. https://doi.org/10.1007/s11240-014-0643-9

Junqueira, K. P., Faleiro, F. G., Junqueira, N. T. V., Bellon, G., Lima, C. A. D., Souza, L. S. D., 2010. Diversidade genética de pitayas nativas do cerrado com base em marcadores RAPD. **Revista Brasileira de Fruticultura**, 32, 819-824. https://doi.org/10.1590/S0100-29452010005000104

Kaplan, E. L., Meier, P., 1958. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, 53(282), 457-481. https://doi.org/10.1080/01621459.1958.10501452

Le Bellec, F., Vaillant, F., Imbert, E., 2006. Pitahaya (*Hylocereus* spp.): a new fruit crop, a market with a future. **Fruits**, *61*(4), 237-250. doi:10.1051/fruits:2006021

Lindsey, J. C., Ryan, L. M, 1998. Methods for interval-censored data. **Statistics in medicine**, 17(2), 219-238. https://doi.org/10.1002/(SICI)1097-0258(19980130)17:2<219::AID-SIM735>3.0.CO;2-O

Lone, A. B., Colombo, R. C., Favetta, V., Takahashi, L. S. A., Faria, R. T., 2014. Temperatura na germinação de sementes de genótipos de pitaya. **Semina: Ciências Agrárias**, *35*(4Supl), 2251-2258. https://doi.org/10.5433/1679-0359.2014v35n4Suplp2251

Majerowicz, N. F., Pesres, L. E. P., 2004. Fotomorfogênese em plantas. In Kerbauy, G.B. **Fisiologia Vegetal**. Rio de Janeiro, RJ. Editora guanabara koogan SA. p. 421-438

Manso, R., Fortin, M., Calama, R., Pardos, M., 2013. Modelling seed germination in forest tree species through survival analysis. The *Pinus pinea* L. case study. **Forest ecology and management**, *289*, 515-524. https://doi.org/10.1016/j.foreco.2012.10.028

Mantel, N., 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. **Cancer Chemother** *Rep*, 50(3), 163-170.

Martins Filho, S., Duarte, M. L., Venzon, M., 2023. Survival Analysis of the Green Lacewing, *Chrysoperla externa* (Hagen) Exposed to Neem-Based Products. **Agriculture**, 13(2), 292. https://doi.org/10.3390/agriculture13020292

McNair, J. N., Sunkara, A., Frobish, D., 2012. How to analyse seed germination data using statistical time-to-event analysis: non-parametric and semi-parametric methods. **Seed Science Research**, 22(2), 77-95. https://doi.org/10.1017/S0960258511000547

Ng'andu, N. H., 1997. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. **Statistics in medicine**, 16(6), 611-626. https://doi.org/10.1002/(SICI)1097-0258(19970330)16:6<611::AID-SIM437>3.0.CO;2-T

Nunes, E. N., de Sousa, A. S. B., de Lucena, C. M., Silva, S. D. M., de Lucena, R. F. P., Alves, C. A. B., Alves, R. E., 2014. Pitaia (*Hylocereus* spp.): Uma revisão para o Brasil. **Gaia Scientia**, v. 8, n. 1, p. 90-98.

Odell, P. M., Anderson, K. M., D'Agostino, R. B., 1992. Maximum Likelihood Estimation for Interval-Censored Data Using a Weibull- Based Accelerated Failure Time Model. **Biometrics**, 48(3), 951-959. https://doi.org/10.2307/2532360

Oller, R., Gómez, G., 2012. A generalized Fleming and Harrington's class of tests for interval-censored data. **Canadian Journal of Statistics**, 40(3), 501-516. https://doi.org/10.1002/cjs.11139

Oller, R., Langohr, K., 2017. FHtest: An R package for the comparison of survival curves with censored data. **Journal of Statistical Software**, 81(15), 1-25. https://doi.org/10.18637/jss.v081.i15

Onofri, A., Gresta, F., Tei, F., 2010. A new method for the analysis of germination and emergence data of weed species. **Weed Research**, 50(3), 187-198. https://doi.org/10.1111/j.1365-3180.2010.00776.x

Onofri, A., Mesgaran, M. B., Tei, F., Cousens, R. D., 2011. The cure model: an improved way to describe seed germination? **Weed Research**, 51(5), 516-524. https://doi.org/10.1111/j.1365-3180.2011.00870.x

Ortega-Flores, M., Maya-Elizarrarás, E., Schondube, J. E., 2018. Effects of Rufous-Backed Robin (*Turdus rufopalliatus*) on brazilian pepper-tree (*Schinus terebinthifolius*) seed germination and dispersal in a subtropical peri-urban environment. **Tropical Conservation Science**, 11, 1940082918761022. https://doi.org/10.1177/1940082918761022

Pan, W., 1999. Extending the iterative convex minorant algorithm to the Cox model for interval-censored data. **Journal of Computational and Graphical Statistics**, 8(1), 109-120. https://doi.org/10.1080/10618600.1999.10474804

Peto, R., 1973. Experimental survival curves for interval-censored data. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, 22(1), 86-91. https://doi.org/10.2307/2346307

Peto, R., 1972. Contribution to the discussion of paper by DR Cox. J. **Journal of the Royal Statistical Society B**, 34, 205-207.

Prentice, R. L., Gloeckler, L. A., 1978. Regression analysis of grouped survival data with application to breast cancer data. **Biometrics**, 57-67. https://doi.org/10.2307/2529588

Radke, B. R., 2003. A demonstration of interval-censored survival analysis**. Preventive veterinary medicine**, 59(4), 241-256. https://doi.org/10.1016/S0167-5877(03)00103-X

Robertson, T., Wright, F. T., Dykstra, R., 1988. **Order restricted statistical inference**. New York: John Wiley & Sons.

Rojas-Sandoval, J., Praciak, A., 2022. *Hylocereus undatus* (dragon fruit). **CABI Compendium**. https://doi.org/10.1079/cabicompendium.27317

Rücker, G., Messerer, D., 1988. Remission duration: an example of interval-censored observations. **Statistics in Medicine**, 7(11), 1139-1145. https://doi.org/10.1002/sim.4780071106

Ruths, R., Silva Bonome, L. T., Tomazi, Y., Siqueira, D. J., Moura, G. S., Lima, C. S. M., 2019. Influência da temperatura e luminosidade na germinação de sementes das espécies: Selenicereus setaceus, *Hylocereus undatus* e *Hylocereus polyrhizus*. **Revista de Ciências Agroveterinárias**, *18*(2), 194-201. https://doi.org/10.5965/223811711812019194

Schoenfeld, D., 1982. Partial residuals for the proportional hazards regression model. **Biometrika**, 69(1), 239-241. https://doi.org/10.1093/biomet/69.1.239

Schwarz, G., 1978. Estimating the dimension of a model. **The annals of statistics**, 461-464.

Scott, S. J., Jones, R. A., 1982. Low temperature seed germination of *Lycopersicon* species evaluated by survival analysis. **Euphytica**, 31, 869-883. https://doi.org/10.1007/BF00039227

Silveira, L. V., Colosimo, E. A., Passos, J. R. D. S., 2010. A Generalized Log-Normal Model for Grouped Survival Data. **Communications in Statistics—Theory and Methods**, 39(15), 2659-2666. https://doi.org/10.1080/03610920903009368

Solarik K. A., Gravel D., Ameztegui A., Bergeron Y., Messier C., 2016. Assessing tree germination resilience to global warming: a manipulative experiment using sugar maple (*Acer saccharum*). **Seed Science Research**, 26(2), 153-164. doi:10.1017/S0960258516000040

Struthers, C. A., Kalbfleisch, J. D., 1986. Misspecified proportional hazard models. **Biometrika**, 73(2), 363-369. https://doi.org/10.1093/biomet/73.2.363

SUN, J. (2006). **The statistical analysis of interval-censored failure time data** (Vol. 3, No. 1). New York: Springer

Tang, Z. H., Xu, J. L., Flanders, J., Ding, X. M., Ma, X. F., Sheng, L. X., Cao, M., 2012. Seed dispersal of *Syzygium oblatum* (Myrtaceae) by two species of fruit bat (*Cynopterus sphinx* and *Rousettus leschenaulti*) in South-West China. **Journal of tropical ecology**, 28(3), 255-261. https://doi.org/10.1017/S0266467412000156

Tarone, R. E., Ware, J., 1977. On distribution-free tests for equality of survival distributions. **Biometrika**, 64(1), 156-160. https://doi.org/10.1093/biomet/64.1.156

Tarte, I., Singh, A., Dar, A. H., Sharma, A., Altaf, A., Sharma, P., 2023. Unfolding the potential of dragon fruit (*Hylocereus* spp.) for value addition: A review. **eFood**, 4(2), e76. https://doi.org/10.1002/efd2.76

Turnbull, B. W., 1976. The empirical distribution function with arbitrarily grouped, censored and truncated data. **Journal of the Royal Statistical Society: Series B (Methodological)**, 38(3),290-295. https://doi.org/10.1111/j.25176161.1976.tb01597.x

Wellner, J. A., Zhan, Y., 1997. A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. **Journal of the American Statistical Association**, 92(439), 945-959. https://doi.org/10.1080/01621459.1997.10474049

Wijenayake, P. R., Hiroshima, T., 2021. Age-based survival analysis of coniferous and broad-leaved trees: a case study of preserved forests in northern Japan. **Forests**, 12(8), 1014. https://doi.org/10.3390/f12081014

Yambe, Y., Takeno, K., 1992. Improvement of rose achene germination by treatment with macerating enzymes. **HortScience**, *27*(9), 1018-1020.

Zerpa-Catanho, D., Hernández-Pridybailo, A., Madrigal-Ortiz, V., Zúñiga-Centeno, A., Porras-Martínez, C., Jiménez, V. M., Barboza-Barquero, L., 2019. Seed germination of pitaya (*Hylocereus* spp.) as affected by seed extraction method, storage, germination conditions, germination assessment approach and water potential. **Journal of Crop Improvement**, v. 33, n. 3, p. 372-394. https://doi.org/10.1080/15427528.2019.1604457

CAPÍTULO 2: ANÁLISE DE SOBREVIVÊNCIA COM CENSURA INTERVALAR APLICADA NA GERMINAÇÃO DE SEMENTES DE PITAIA.

2.1. Resumo

A análise de sobrevivência é recomendada para ensaios de germinação de sementes devido à presença de censura. Em muitos casos, o momento exato da germinação não pode ser determinado devido as observações serem realizadas em intervalos, resultando em censura intervalar. Para lidar com esse tipo de dado, existem técnicas específicas que podem ser empregadas. Assim, esse estudo avaliou o uso de métodos de sobrevivência com censura intervalar na germinação de sementes de pitaia (*Hylocereus* spp.), utilizando também técnicas para dados grupados, um caso particular de censura intervalar. Quatro metodologias foram aplicadas: i) algoritmo EMICM para o efeito combinado do tempo de armazenamento e luz; ii) regressão paramétrica para o efeito conjunto do local e tempo de armazenamento; iii) regressão semiparamétrica para o método de extração; e iv) regressão discreta em dados grupados para o efeito combinado do armazenamento e temperatura. A germinação das sementes foi contabilizada semanalmente por quatro semanas. Como o tempo exato de falha não era conhecido, as sementes foram consideradas censuradas por intervalo e aquelas não germinadas até o fim do experimento foram consideradas censuradas à direita. Os resultados obtidos foram consistentes com as técnicas tradicionais, mas com o diferencial de permitir o estudo detalhado da trajetória das sementes até a germinação, além de incorporar adequadamente as censuras na análise. Embora os dados apresentem poucos tempos de observação e altas taxas de empates, a aplicação dos métodos é valiosa para estabelecer padrões e conclusões sobre a germinação das sementes de pitaia. Dessa forma, esses métodos são particularmente relevantes ao lidar com dados censurados intervalares em estudos de germinação, fornecendo uma análise mais completa e precisa.

Palavras-chave: *Hylocereus* spp., Armazenamento, Temperatura, Tipos de luz, Extração de semente, Tempo até o evento, técnicas tradicionais.

2.2. Introdução

A análise de sobrevivência examina o tempo até a ocorrência de um evento de interesse a partir de um ponto inicial de observação (Silveira *et al*., 2010). Nos estudos sobre germinação de sementes, o evento de interesse é a germinação, e o objetivo é estimar a probabilidade de as sementes germinarem ou não até um determinado momento do estudo. A função de sobrevivência representa a probabilidade de a semente não germinar até o tempo t. Desta forma, a probabilidade de germinação é calculada pela função de distribuição acumulada $F = 1 - S(t)$.

Estudos de germinação de sementes são essenciais na agricultura para avaliar a qualidade fisiológica e o vigor das sementes, além de orientar o planejamento agrícola (Barak *et al*., 2018). Também desempenham um papel crucial na pesquisa e no melhoramento genético de plantas, ao identificar os fatores que afetam o potencial de germinação das sementes e compreender os mecanismos de reparo de DNA que mantêm a viabilidade das sementes (Waterworth *et al*., 2019). No entanto, é comum encontrar sementes que não germinaram ao final do estudo, o que é considerado censura, indicando a não ocorrência de germinação durante o período do experimento.

Métodos convencionais, como ANOVA e regressão, frequentemente desconsideram as censuras em suas análises (Onofri *et al*., 2010), o que pode resultar na subestimação de algumas estimativas, como o tempo médio e mediano de germinação (Scott e Jones, 1990). A análise de sobrevivência, por outro lado, é uma abordagem alternativa capaz de lidar efetivamente com dados censurados.

Existem diversas situações em que as observações são realizadas em datas distantes, e o evento de interesse é detectado apenas durante essas observações, tornando impossível determinar o tempo exato de falha. Esse contexto resulta em um caso específico de censura, chamada censura intervalar, gerando dados conhecidos como dados de sobrevivência intervalar (Colosimo e Giolo, 2024).

Para lidar com esse tipo de dado, existem técnicas não paramétricas, paramétricas e semiparamétricas específicas que podem ser empregadas. Os autores Giolo *et al*., 2009; Gomez *et al*., 2009; Boruvka e Cook, 2015; Anderson-Bergman, 2017a, b; Vaca *et al*., 2021 discutem e aplicam alguns dos métodos estatísticos para lidar com dados censurados por intervalo na análise de sobrevivência.

Um caso particular dos dados de sobrevivência intervalar ocorre quando todas as unidades amostrais compartilham os mesmos tempos de observação. Nesse caso, os dados censurados por intervalo, quando os intervalos são iguais, são denominados dados de sobrevivência grupados (Giolo *et al*., 2009). Uma característica distintiva desses dados é a ocorrência frequente de empates, ou seja, vários tempos de falha no mesmo intervalo. Para lidar com esses empates de forma adequada, é necessário tratar os tempos de falha como discretos e empregar modelos de regressão discreta (Lawless, 2003).

Exemplos de estudos que utilizaram dados grupados incluem pesquisas sobre os tempos de vida utilizando modelos de riscos proporcionais e modelos de chances proporcionais (Colosimo *et al*., 2000; Liciana *et al*., 2006). Outros estudos também exploraram modelos alternativos para dados de sobrevivência grupados, como o modelo log-normal generalizado, proposto por Silveira *et al*. (2010).

Na aplicação dos métodos de censura intervalar, foram utilizados dados de germinação de pitaia. A pitaia, conhecida como "fruta-do-dragão" (Nishikito *et al*., 2023), é uma fruta da família Cactaceae e suas variedades comerciais mais comuns pertencem ao gênero *Hylocereus* (Jalgaonkar *et al*., 2020). A pitaia destaca-se pelo sabor refrescante e adocicado, além de ser rica em água, minerais, vitamina C, gorduras saudáveis, proteínas e fibras (Luu *et al*., 2021). Sua casca e polpa são usadas como corantes e ingredientes naturais nas indústrias alimentícia e farmacêutica (Nur *et al*., 2023). Embora domesticada recentemente, tornou-se popular globalmente devido ao sabor e aparência exóticos, mas muitos aspectos dessa cultura ainda precisam ser estudados (Shah *et al*., 2023).

A propagação clonal por meio de estaquia é o modo preferido de reprodução em *Hylocereus* spp., usando estacas de caule (Le Bellec *et al*., 2006), ou cultura in vitro (Hua *et al*., 2015). Entretanto, o uso de sementes é importante para o melhoramento convencional e conservação de recursos genéticos vegetais (Le Bellec *et al*., 2006; Hernández e Salazar, 2012).

O objetivo deste trabalho, de forma geral, o foi avaliar o uso dos métodos de sobrevivência com censura intervalar para verificar a germinação de sementes de pitaia (*Hylocereus* spp.), mais especificamente, o uso dos métodos não paramétricos, paramétricos, semiparamétricos e dados grupados.

2.3. Materiais e Métodos

Para estudar a aplicação dos métodos de censura intervalar, foram utilizados dados de quatro experimentos publicados por Hernández-Pridybailo (2018) e descritos por Zerpa-Catanho *et al*. (2019). Nos experimentos foi avaliada a germinação de sementes de pitaia de polpa vermelha, das variedades *Orejona* (OR), *San Ignacio* (SI) e *Criollo* (CR).

Para os quatro conjuntos de dados os tempos exatos de germinação das sementes são desconhecidos. Sabe-se apenas que ocorreu entre os intervalos de duas visitas consecutivas, realizadas semanalmente entre o sétimo e o trigésimo dia após o início do experimento. Dessa forma, os dados foram classificados como dados de sobrevivência com censura intervalar e as sementes que não germinaram até o final do experimento foram consideradas censuras à direita.

Realizou-se uma análise preliminar aplicando quatro modelos distintos — não paramétrico, paramétrico, semiparamétrico e regressão discreta para dados agrupados — aos quatro conjuntos de dados obtidos. A partir dos resultados, foi possível identificar o modelo mais adequado para apresentar em cada caso. Entretanto, é válido ressaltar que todas as metodologias podem ser aplicadas a qualquer um dos conjuntos de dados aqui analisados, desde que atendam às condições exigidas por cada método e proporcionem resultados adequados, considerando que todos os dados se referem à germinação com censura intervalar.

2.3.1. Estimador não paramétrico de máxima verossimilhança

Na abordagem não paramétrica não é feita qualquer suposição sobre a distribuição probabilística do tempo de sobrevivência T, essa é a abordagem mais básica para analisar dados de sobrevivência com censura intervalar. Diferentemente do estimador de Kaplan-Meier (Kaplan e Meier, 1958), o estimador da função de sobrevivência não paramétrico de máxima verossimilhança (NPMLE), para dados censurados por intervalo deve ser obtido por um algoritmo iterativo.

Turnbull (Turnbull, 1976) demonstrou que, para estimar a função de sobrevivência sob censura intervalar, pode-se simplificar o problema concentrando a

análise exclusivamente nos intervalos de Turnbull denotados por $(p_j\ ,\ q_j]$, para $j\ =\ 1,\ldots,m$.

As análises foram conduzidas com dados de um experimento que investigou o efeito das luzes branca, vermelha, azul e da ausência de luz na germinação de sementes de pitaia, armazenadas em câmara fria por 12, 13 e 14 meses. Zerpa-Catanho et al. (2019) observaram que o armazenamento em câmara fria por até 12 meses não tem um impacto significativo na germinação das sementes. Com base nessa constatação, os pesquisadores realizaram experimentos a partir desse período, avaliando os efeitos da germinação em intervalos mensais após os 12 meses de armazenamento. Utilizou-se 225 sementes para cada combinação de luz e tempo de armazenamento. A germinação foi avaliada aos 7, 15, 22 e 30 dias, sendo considerada a protrusão da radícula como critério.

Para cada condição de luz em cada tempo de armazenamento, o conjunto possui 225 indivíduos com intervalos de censura $\{(L_i\ ,\ U_i], 1 \leq i \leq 225)\} = \{(1,7], (7,15], (15,22\}, (22,30], (30,\infty\ ]\}$, os intervalos de Turnbull correspondente são dados por $I = \left(p_j\ ,\ q_j\right] =\ \{(1,7], (7,15], (15,22\}, (22,30], (30,\infty\ ]\}$, para $j\ =\ 1,\ldots,5$.

Como a função de sobrevivência estimada permanece constante fora desses intervalos e é indeterminada dentro deles, pode-se definir a probabilidade do j-ésimo intervalo de Turnbull $\mathrm{s_j} = \mathrm{P}\left(\mathrm{p_j} \leq \mathrm{T} \leq \mathrm{q_j}\right) =\ \mathrm{S}\left(\mathrm{p_j}\right) - \mathrm{S}\left(\mathrm{q_j}\right)$. Dessa forma, o NPMLE das funções de sobrevivência para o conjunto de dados se reduz à maximização da seguinte função de verossimilhança:

$$L = \prod_{i=1}^{225}\left(\sum_{j=1}^{5}\alpha_{ij}s_j\right)$$

em que α_{ij} é uma variável indicadora cujo valor assume 1, se o intervalo $(L_i\ ,\ U_i]$ contém o intervalo de observação $(p_j\ ,\ q_j]$ e 0 caso contrário.

Existem diferentes algoritmos para o cálculo do NPMLE. Nesse trabalho ele foi obtido a partir de um procedimento iterativo EMICM proposto por Wellner e Zahn (1997) que combina os algoritmos EM e ICM. Com base nas estimativas das funções

de sobrevivência, foram construídos gráficos exibindo as curvas da função de distribuição acumulada da germinação.

Após estimar as funções de sobrevivência, foi testada a igualdade entre a germinação em duas condições de luz em cada tempo de armazenamento, combinando-os de dois em dois. Para essa análise, foi empregada o vetor das estatísticas não paramétricas do teste de Fleming e Harrington (1991) estendido para dados censurados por intervalo por Oller e Gómez (2012).

Esse teste é uma versão do teste log-rank ponderado (Tarone e Ware, 1977), que se baseia na comparação do número de eventos observados e esperados para cada j-ésima região de suporte da função de sobrevivência estimada pelo método não paramétrico (NPMLE) na amostra combinada dos grupos.

$$U_K = \sum_{j=1}^{J} w_j \left(D_{jk} - E_{kj} \right) = \sum_{j=1}^{J} w_j \left(D_{jk} - D_j \frac{n_{jk}}{n_j} \right)$$

Nesse caso, $D_{jk} = N_k\{\hat{S}_k(p_j) - \hat{S}_k(q_j)\}$ e $n_{jk} = N_k \hat{S}_k(p_j)$ são o número estimado observado de eventos no k-ésimo grupo ocorrendo no j-ésimo intervalo $(p_j , q_j]$ e o número em risco imediatamente antes desse intervalo, respectivamente. Enquanto, $D_j = n\{\hat{S}(p_j) - \hat{S}(q_j)\}$ e $n_j = n\,\hat{S}(p_j)$ representam essas mesmas quantidades na amostra combinada referente ao j-ésimo intervalo $(p_j , q_j]$. O w_j representa o peso dado ao teste.

Considerando $\rho \geq 0, \gamma \geq 0$, têm-se para $j = 1, \ldots, J$:

$$w_j = w_j^{\rho,\gamma} = \hat{S}(p_j) \frac{B\left(1 - \hat{S}(q_j);\, \gamma + 1, \rho\right) - B\left(1 - \hat{S}(p_j);\, \gamma + 1, \rho\right)}{\hat{S}(p_j) - \hat{S}(q_j)} \quad (12)$$

em que $B(y; a, b) = \int_0^y x^{a-1} (1-x)^{b-1}\, dx$ é uma função beta incompleta, para $y \geq 0, a > 0, b \geq 0$.

Esse teste foi implementado utilizando o pacote *FHtest* versão 1.5.1 (Oller e Langohr, 2023) do software R versão 4.4.1 (R Development Core Team, 2024). Nesse caso, os parâmetros do teste Fleming-Harrington foram $\rho = 0$ e $\gamma = 0$, resultando em um teste log-rank.

Na aplicação do algoritmo EMICM foi utilizado o pacote *icenReg* versão 2.0.16 (Anderson-Bergman, 2017b) do *software* R versão 4.4.1 (R Development Core Team, 2024).

2.3.2. Modelo Paramétrico

Os dados desta análise foram obtidos de um experimento que avaliou a germinação de sementes de pitaia armazenadas no escuro por 1, 2, 3, 5, 7, 9 e 12 meses. As sementes foram armazenadas em dois locais: em condições ambientais (22,37 ± 0,18°C e 61,23 ± 2,98% de umidade relativa) e câmara fria (5,75 ± 0,08°C e 62,59 ± 0,73% de umidade relativa). Utilizou-se 350 sementes para cada combinação de tempo e local de armazenamento, com os registros da germinação obtidos visualmente aos 7, 15, 22 e 30 dias, com a protrusão da radícula.

Para analisar esses dados, foi ajustado um modelo de regressão paramétrico que pressupõe uma distribuição de probabilidade para o tempo até a germinação. O modelo considera as covariáveis tempo de armazenamento e o ambiente (câmara fria ou condições ambientais), além da interação entre elas, conforme o preditor linear:

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 armaz + \hat{\beta}_2 amb_{frio} + \hat{\beta}_3 (armaz * amb_{frio})$$

Neste modelo, $armaz$ representa o tempo de armazenamento (em meses), enquanto amb_{frio} é uma variável indicadora para o local de armazenamento, assumindo o valor 1 para sementes armazenadas em câmara fria e 0 para sementes armazenadas em condições ambientais. A interação $armaz * amb_{frio}$ reflete o efeito combinado do local e do tempo de armazenamento sobre o tempo de germinação.

Para selecionar a distribuição que melhor se adequasse aos dados, testou-se as distribuições exponencial, Weibull, log-normal e log-logística, mostradas na Tabela

1.2. Os parâmetros da regressão foram estimados para cada modelo, permitindo avaliar o ajuste de cada distribuição aos dados.

Na estimação dos parâmetros com censura intervalar, é empregado o método da máxima verossimilhança, utilizando a função de verossimilhança em relação aos parâmetros do modelo ($\boldsymbol{\theta}$):

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{350} [S(l_i|\boldsymbol{x}_i) - S(u_i|\boldsymbol{x}_i)]^{\delta_i} \, [S(l_i|\boldsymbol{x}_i)]^{1-\delta_i}$$

em que l_i e u_i representam, respectivamente, os limites inferior e superior dos intervalos de tempo de germinação na i-ésima semente; $\boldsymbol{x}_i$ é o vetor de covariáveis associado a i-ésima semente ($i = 1, \ldots, 350$); δ_i uma variável indicadora que assume 1 se o evento ocorreu no intervalo (L_i , U_i], ou assume 0, caso contrário.

Os modelos paramétricos foram ajustados com a função *survreg* do pacote *survival* versão 3.6-4 (Therneau, 2024) do *software* R versão 4.4.1 (R Development Core Team, 2024). A distribuição foi selecionada com base no valor do AIC - Critério de Informação de Akaike (Akaike, 1974) e do BIC (Critério de Informação Bayesiano (Schwarz, 1978). Com base no modelo selecionado, foram construídos gráficos que exibem as curvas da função de distribuição acumulada da germinação.

2.3.3. Modelo Semiparamétrico

O modelo de taxas de falhas proporcionais, ou riscos proporcionais de Cox (Cox, 1972), é uma abordagem semiparamétrica que visa modelar a função taxa de falha e é a técnica utilizada nesta seção.

Os dados desta análise foram obtidos de um experimento que investigou a germinação de sementes de pitaia utilizando dois métodos de extração: enzimático e manual. O teste de germinação foi conduzido com 150 sementes para cada método, avaliando visualmente a germinação aos 7, 15, 22 e 30 dias, utilizando a protrusão da radícula como critério.

Este modelo pode ser expresso por:

$$h(t) = h_0(t)\exp\{\beta_1 * extrac_{enzim}\}$$

Têm-se que $h(t)$ representa a taxa de falha para o grupo de sementes extraídas enzimaticamente em função do tempo t, e $h_0(t)$ é a função taxa de falha de base, que corresponde à taxa de falha das sementes extraídas manualmente. A variável $extrac_{enzim}$ é indicadora, assumindo o valor 1 para o método de extração enzimático e 0 para o manual.

O modelo não calcula taxas de falha separadas para cada grupo, mas sim a razão da taxa de falha entre eles. Isso significa que a razão das taxas de falha entre as sementes extraídas enzimaticamente e manualmente permanece constante ao longo do tempo, desde que a suposição de proporcionalidade das taxas de falha seja mantida. Portanto, essa pressuposição é essencial para que o modelo de Cox seja aplicado corretamente.

Para estimar os parâmetros desse modelo no contexto intervalar, utilizou-se a seguinte função de log-verossimilhança:

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{150} \log\left[\{1 - F_0(L_i)\}^{\exp\{\beta_1 * extrac_{enzim}\}} - \{1 - F_0(U_i)\}^{\exp\{\beta_1 * extrac_{enzim}\}} \right]$$

Pan (1999) propôs uma extensão do algoritmo ICM para o modelo de Cox. Enquanto, o algoritmo original do ICM foi desenvolvido para o NPMLE para dados intervalares sem covariáveis, a extensão para o modelo de Cox incorpora covariáveis. Este algoritmo visa maximizar a verossimilhança do modelo. Ao contrário do ICM original, o algoritmo estendido realiza iterações conjuntas para estimar tanto a função de distribuição acumulada de linha de base, quanto os parâmetros de regressão.

O algoritmo utilizado possui dois passos distintos para atualizar os parâmetros: um passo de Newton-Raphson condicional para os parâmetros de regressão e um passo ICM para os parâmetros de sobrevivência de base. Em casos de muitas observações não censuradas, um método de ascensão de gradiente é empregado para atualizar os parâmetros de sobrevivência de base. Foram utilizadas 500 amostras *bootstrap* para o cálculo do erro padrão, com o objetivo de aplicar a estatística do teste Wald e realizar inferências sobre o parâmetro do modelo.

Com os resultados dos parâmetros, modelou-se a probabilidade de germinação e construíram-se gráficos que exibem as probabilidades acumuladas ao longo do tempo. Para avaliar a suposição de proporcionalidade de taxas de no modelo de Cox, foi utilizado um gráfico descritivo de diagnóstico que exibe $\ln[-\ln(S_0(t))]$ versus t, em que $S_0(t)$ representa a função de sobrevivência de base.

Esse método estatístico foi aplicado utilizando o pacote *icenReg* versão 2.0.16 (Anderson-Bergman, 2017b) do *software* R versão 4.4.1 (R Development Core Team, 2024).

2.3.4. Regressão Discreta em dados grupados

Os dados grupados geralmente apresentam muitos empates, isto é, múltiplas falhas ocorrem simultaneamente em um mesmo intervalo de tempo, tornando a ordenação dos tempos de falha inviável, o que complica a estimativa dos parâmetros do modelo. Duas abordagens comuns são: usar aproximações para a função de verossimilhança parcial ou modelos de regressão discretos.

Foi utilizada uma regra empírica proposta por Chalita et al. (2002) para auxiliar na decisão sobre a abordagem a ser adotada. Com base na proporção de empates, que ultrapassou 25%, indicou-se a utilização de modelos de regressão discreta. Segundo Hashimoto *et al*., (2011) modelos de regressão discreta para dados de sobrevivência grupados podem modelar efetivamente o tempo de censura e eliminar empates, utilizando funções de ligação logito e complemento log-log.

Os dados foram obtidos de um experimento que avaliou o efeito das temperaturas de 15, 20, 25 e 30 °C armazenadas em câmara fria por 12, 13 e 14 meses. Foram utilizadas 225 sementes para cada tratamento, com a germinação avaliada visualmente aos 7, 15, 22 e 30 dias, utilizando a protrusão da radícula como critério.

Os dados dessa análise estão grupados em 4 intervalos de tempo $I_j = [t_{j-1}, t_j), j = 1, \ldots, 4$ onde $0 = t_0 < \cdots < t_k$ e $t_{k+1} = \infty$. Define-se R_j como o conjunto de sementes sob risco no início do intervalo I_j, ou seja, no tempo t_{j-1}. Para casa semente i $(i = 1, \ldots, 225)$, a variável indicadora $\delta_{ij} = 1$ se o tempo de falha da i-ésima semente ocorrer no intervalo I_j e $\delta_{ij} = 0$, caso contrário.

A probabilidade de germinação da i-ésima semente no intervalo I_j, dado as temperaturas e os tempos de armazenamento, condicionado ao fato de que essa semente não havia germinado em t_{j-1} é representada por $p_j(\boldsymbol{x}_i)$.

A verossimilhança é escrita em função das probabilidades de falha $p_j(\boldsymbol{x}_i)$ de cada semente em cada intervalo, levando em consideração a contribuição de sementes censuradas e não censuradas

$$\prod_{j=1}^{4} \prod_{i \in R_j} \{ p_j(\boldsymbol{x}_i)\}^{\delta_{ij}} \ \{1 - \ p_j(\boldsymbol{x}_i)\}^{(1- \delta_{ij})}$$

Essa probabilidade foi modelada com dois modelos: o de taxas de falhas proporcionais de Cox e o de chances proporcionais. No modelo de taxas de falhas proporcionais de Cox, como proposto por Prentice e Gloeckler (1978), a probabilidade $p_j(\boldsymbol{x}_i)$ é dada por:

$$p_j(\boldsymbol{x}_i) = 1 - \gamma_j^{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}$$

em que $\gamma_j = \frac{S_0(t_j)}{S_0(t_{j-1})}, j = 1, \dots, 4.$

Já no modelo de chances proporcionais, como sugerido por Hosmer e Lemeshow (2000), a $p_j(\boldsymbol{x}_i)$ assume a forma:

$$p_j(\boldsymbol{x}_i) = 1 - \left\{ 1 + \ \gamma_j \exp(\boldsymbol{\beta}^T \boldsymbol{x_i}) \right\}^{-1}$$

em que $\gamma_j = \frac{p_j(0)}{1- p_j(0)}, \ j = 1, \dots, 4.$

Esses modelos foram linearizados, utilizando transformações que permitiram utilizar métodos usuais para a modelagem de resposta binária.

No modelo de riscos proporcionais de Cox, utilizou-se a função de ligação complemento log-log (cloglog):

$$\ln\left[-\ln\left(1-p_j(\boldsymbol{x}_i)\right]=\ \gamma_j^*+\boldsymbol{\beta}^T\boldsymbol{x_i}=\eta_{ij}\right.$$

No modelo de chances proporcionais, esse resultado foi alcançado por meio da função de ligação logito:

$$\ln\left(\frac{p_j(\boldsymbol{x}_i)}{1-\ p_j(\boldsymbol{x}_i)}\right)=\gamma_j^*+\boldsymbol{\beta}^T\boldsymbol{x_i}=\eta_{ij}$$

em que, $\gamma_j^*=\ln\left(\gamma_j\right)$ é o efeito do j-ésimo intervalo de tempo.

Com o modelo ajustado pelo método da máxima verossimilhança, foram construídos gráficos para visualizar as probabilidades de germinação $p_j(\boldsymbol{x}_i)$ em cada temperatura para os três tempos de armazenamento, ao longo dos intervalos de tempo considerados.

Na avaliação do desempenho dos modelos selecionados, utilizou-se a curva ROC (*Receiver Operating Characteristic*) como o critério de validação. A curva ROC, baseada nas medidas de sensibilidade e especificidade, possibilita a análise empírica da precisão e qualidade dos modelos, ao avaliar sua capacidade de discriminar corretamente entre sementes germinadas e não germinadas. A área sob a curva (AUC) foi calculada para representar a eficácia geral dos modelos na classificação dos resultados de germinação.

Para a aplicação da regressão discreta foram utilizando o pacote *icenReg* versão 2.0.16 (Anderson-Bergman, 2017b) do *software* R, versão 4.4.1 (R Development Core Team, 2024).

Para simplificar a apresentação da metodologia, a seguir é fornecido um quadro resumindo os principais aspectos analisados.

Quadro 2.1 - Síntese das covariáveis analisadas em cada abordagem metodológica.

Metodologia	Covariáveis analisadas
Não paramétrica – NPMLE	Efeito conjunto da condição de luz (vermelha, azul, branca e ausência) e do tempo de armazenamento (12, 13 e 14 meses)
Paramétrica	Efeito conjunto do local (condições do ambiente e câmara fria) e do tempo de armazenamento (1,2,3,5,7,9 e 12 meses)
Semiparamétrica	Efeito do método de extração
Regressão Discreta para dados grupados	Efeito conjunto da temperatura (15, 20, 25 e 30°C) e do tempo de armazenamento (12, 13 e 14 meses)

2.4. Resultados

2.4.1 Estimador não paramétrico de máxima verossimilhança

Os percentuais finais de germinação para as diferentes condições de luz e tempos de armazenamentos estão apresentados na Tabela 2.1. Esses resultados foram obtidos por meio do cálculo simples da porcentagem de sementes germinadas em relação ao total de sementes observadas.

Tabela 2.1 – Porcentagem de falhas (germinação) e censuras (não germinação) de sementes de pitaia (*Hylocereus* spp.) em função da condição de luz e do tempo de armazenamento: análise descritiva dos dados.

Luz	12 meses		13 meses		14 meses	
	Falha (%)	Censura (%)	Falha (%)	Censura (%)	Falha (%)	Censura (%)
Vermelha	100,0	00,0	98,2	01,8	100,0	00,0
Azul	100,0	00,0	96,9	03,1	99,1	00,9
Branca	99,6	00,4	98,2	01,8	99,6	00,4
Ausência	78,2	21,8	97,8	02,2	06,2	93,8

A partir dos valores de s_j obtidos pelo algoritmo EMICM ao maximizar a função de verossimilhança, pode-se apresentar as curvas de germinação (Figura 2.1). Nessas curvas estão mostradas a proporção de sementes que germinaram ao longo do tempo sob diferentes condições de luz.

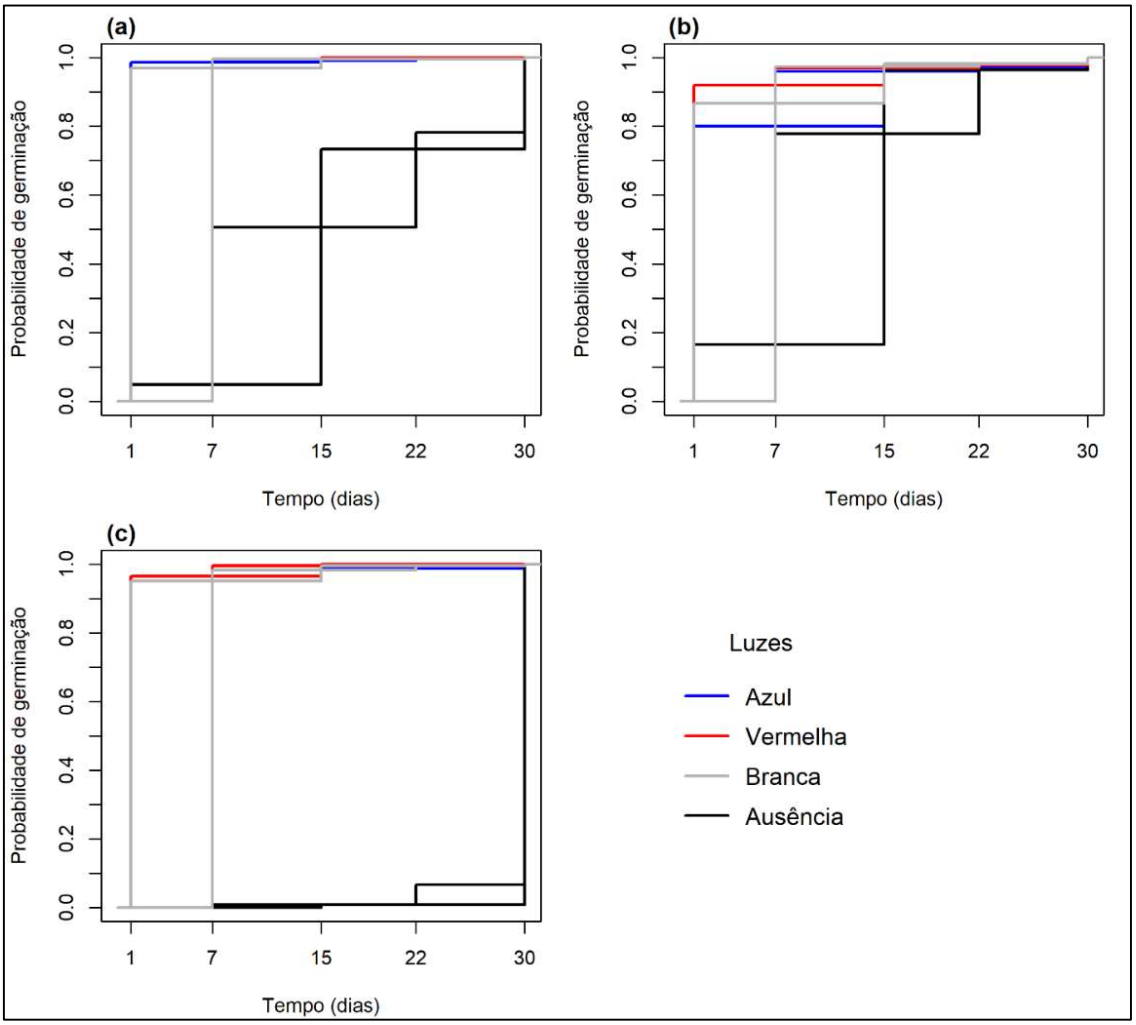

Figura 2.1 – Probabilidades de germinação de sementes de pitaia (*Hylocereus* spp.), estimadas pelo método não paramétrico de censura intervalar, para sementes expostas às luzes azul, vermelha, branca e a ausência de luz em três tempos de armazenamento: (a) 12 meses, (b) 13 meses e (c)14 meses.

É possível notar que as sementes germinaram mais rapidamente sob condições de luz, apresentando altas porcentagens de germinação. De modo contrário, na ausência de luz a germinação foi mais lenta e as porcentagens de germinação menores. Pode ser observado na Tabela 2.2, pelo teste utilizado, que as

curvas de germinação das sementes na ausência de luz diferiram significativamente ($p < 0,05$) dos tipos de luzes estudadas.

Os percentuais de germinação sob condições de luz, independentemente do tipo de luz utilizada, foram altos desde o primeiro intervalo observado (entre 1 e 7 dias), variando aproximadamente de 80% a 98%. Embora mais demorados, os percentuais de germinação na ausência de luz após 12 e 13 meses de armazenamento, foram aproximadamente 78% e 98% ao final do último intervalo de observação. Em contraste, após 14 meses de armazenamento, esse percentual foi de apenas 7%.

De modo geral, não houve diferença significativa na germinação das sementes sob as luzes branca, azul e vermelha, exceto entre as luzes azul e vermelha no 13 mês (Tabela 2.2). Embora o teste tenha mostrado um resultado significativo, observa-se que, no último intervalo de tempo, as curvas para as luzes azul e vermelha ficaram muito próximas, com percentuais de germinação de 96,9% e 98,22%, respectivamente.

.

Tabela 2.2 - Teste de Fleming e Harrington, usado na técnica não paramétrico de censura intervalar, baseado em uma distribuição de permutação, para comparação das curvas de germinação das sementes de pitaia (*Hylocereus* spp.) expostas a diferentes condições de luz e tempos de armazenamento.

Comparações	Estatística	Valor-*p*	Comparações	Estatística	Valor-*p*
Armazenamento de 12 meses					
Azul *vs* Vermelha	-0,90	0,3680	Vermelha *vs* Branca	-0,10	0,9940
Azul *vs* Branca	-1,10	0,2700	Vermelha *vs* Ausência	15,10	**0,0020**
Azul *vs* Ausência	-15,20	**0,0020**	Branca *vs* Ausência	15,00	**0,0020**
Armazenamento de 13 meses					
Azul *vs* Vermelha	3,00	**0,0040**	Vermelha *vs* Branca	-1,30	0,1840
Azul *vs* Branca	1,80	0,0700	Vermelha *vs* Ausência	11,10	**0,0020**
Azul *vs* Ausência	-9,20	**0,0020**	Branca *vs* Ausência	10,80	**0,0020**
Armazenamento de 14 meses					
Azul *vs* Vermelha	1,10	0,2980	Vermelha *vs* Branca	-1,00	0,3700
Azul *vs* Branca	0,10	0,9140	Vermelha *vs* Ausência	20,40	**0,0020**
Azul *vs* Ausência	-20,20	**0,0020**	Branca *vs* Ausência	20,30	**0,0020**

2.4.2. Modelo Paramétrico

Os percentuais finais de germinação para os diferentes tempos e locais de armazenamento estão apresentados na Tabela 2.3. Esses resultados foram obtidos

por meio do cálculo simples da porcentagem de sementes germinadas em relação ao total de sementes observadas.

Tabela 2.3 – Porcentagem de falhas (germinação) e censuras (não germinação) de sementes de pitaia (Hylocereus spp.) em função do local e do tempo de armazenamento: análise descritiva dos dados.

Tempo de armazenamento	Condições ambiente		Câmara fria	
	Falha (%)	Censura (%)	Falha (%)	Censura (%)
1	98,0	02,0	96,3	03,7
2	97,0	03,0	97,0	03,0
3	98,3	01,7	98,7	01,3
5	91,3	08,7	98,8	01,2
7	88,7	11,3	96,3	03,7
9	88,0	12,0	97,7	02,3
12	55,7	44,3	95,7	04,3

Os resultados da análise paramétrica estão apresentados na Tabela 2.4., na qual estão as estimativas dos parâmetros que foram utilizados para determinar a função de sobrevivência $S(t)$ e a curva de germinação.

Tabela 2.4 - Estimativas dos parâmetros, erros padrão (SE), log-verossimilhança, critério de informação de Akaike (AIC) e Critério de Informação Bayesiano (BIC) dos modelos avaliados na germinação de sementes de pitaia (*Hylocereus* spp.), sob diferentes tempos de armazenamento e ambientes.

Parâmetros[1]	Exponencial		Weibull		Log-normal		Log-logístico	
	Estimativa	SE	Estimativa	SE	Estimativa	SE	Estimativa	SE
β_0	1,2409	0,0398	1,3777	0,0333	0,9854	0,03249	0,9733	0,03296
β_1	0,1741	0,0062	0,1606	0,0051	0,1659	0,00446	0,1643	0,00437
β_2	0,1713	0,0563	0,1443	0,0458	0,1517	0,04728	0,1310	0,05004
β_3	-0,1659	0,0086	-0,1518	0,0070	-0,1601	0,00682	-0,1593	0,00715
σ $(scale)$	1,0000	0,0000	0,8060	0,0110	0,6820	0,0045	0,3580	0,0058
Log L	- 4604,40	---	- 4440,90	---	- 3530,90	---	- 3369,30	---
AIC	9216,75	---	8891,76	---	7071,85	---	**6748,55**	---
BIC	9242.74	---	8924.24	---	7104.34	---	**6781.04**	---

[1]Modelo: $\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 armaz + \hat{\beta}_2 amb\_frio + \hat{\beta}_3 armaz * amb\_frio$; $\Pr(\hat{\beta}_3 > z) < 0{,}0001$.

A interação entre tempo e local de armazenamento, foi significativa para todos os modelos estudados, mostrando a dependência desses fatores sobre a germinação. Dentre os modelos analisados, foi selecionado o modelo log-logístico, por apresentar menores valores de AIC e BIC.

O sinal negativo da estimativa da interação indica, em média, que as sementes armazenadas em câmara fria tendem a germinar mais rapidamente em comparação com as armazenadas em temperatura ambiente, especialmente conforme o tempo de armazenamento aumenta. Em outras palavras, o efeito do tempo de armazenamento na porcentagem de germinação é reduzido quando as sementes são armazenadas em câmara fria.

A análise da probabilidade de germinação, utilizando o modelo log-logístico para calcular a função de sobrevivência, revelou que quanto maior o tempo de armazenamento em condição ambiente, menor a probabilidade de germinação das sementes (Figura 2.2 a). Em contrapartida, para as sementes armazenadas em câmara fria, a probabilidade de germinação manteve-se estável ao longo do tempo de armazenamento (Figura 2.2 b).

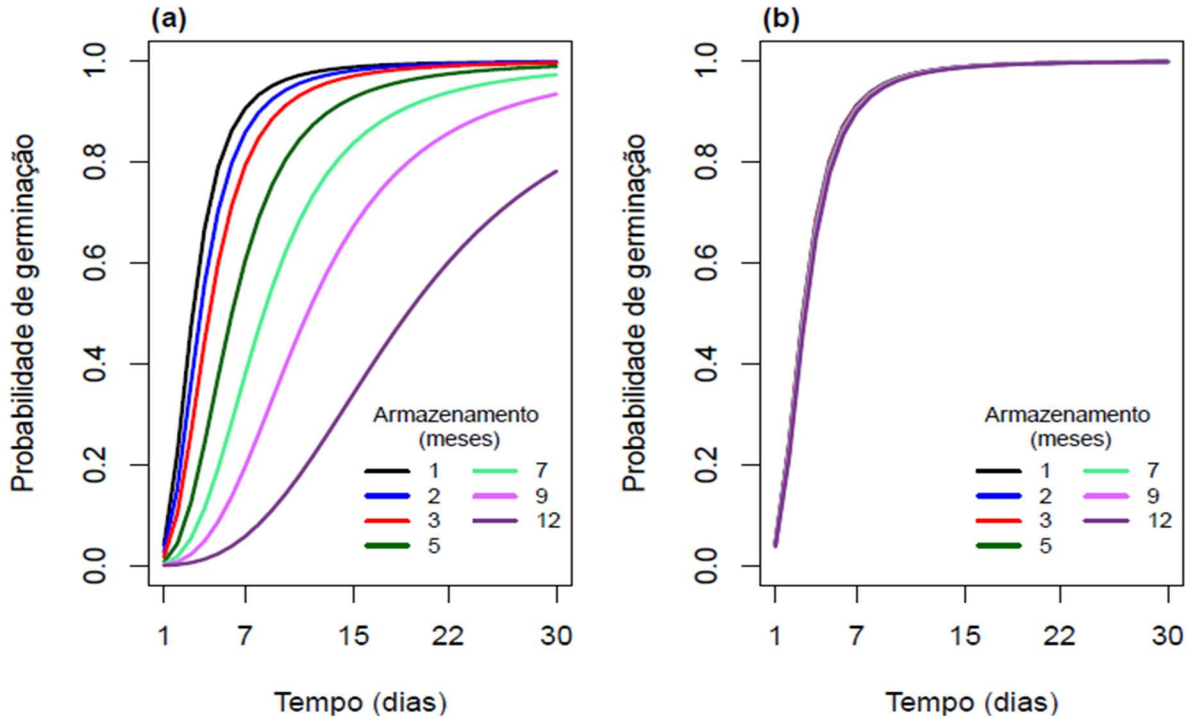

Figura 2.2 – Curvas de germinação estimadas a partir do modelo de regressão log-logístico, para sementes de pitaia (*Hylocereus* spp.), mantidas sob (a) condições ambientes e (b) em câmara fria, por diferentes tempos de armazenamento.

2.4.3. Modelo semiparamétrico

Os percentuais finais de germinação para os métodos de extração estão apresentados na Tabela 2.5. Esses resultados foram obtidos por meio do cálculo simples da porcentagem de sementes germinadas em relação ao total de sementes observadas.

Tabela 2.5 – Porcentagem de falhas (germinação) e censuras (não germinação) de sementes de pitaia (Hylocereus spp.) em função do método de extração: análise descritiva dos dados.

Extração manual		Extração enzimática	
Falha (%)	Censura (%)	Falha (%)	Censura (%)
97,3	02,7	98,0	02,0

Para o estudo dos métodos de extração das sementes, foi ajustado o modelo de Cox, em que a extração manual foi usada como categoria de referência. Os resultados estão apresentados na Tabela 2.6.

Tabela 2.6 - Valores estimados pelo modelo de Riscos Proporcionais de Cox, para a germinação de sementes de pitaia (*Hylocereus* spp.), extraídas pelos métodos manual e enzimático.

Método de Extração	Estimativa (β_1)	RTF[1]	Erro padrão	z	Valor p
Enzimático	0,6772	1,9680	0,3449	1,964	0,0496

[1] Razão de taxas de falha

Com uma razão de taxas de falha associada à extração enzimática de 1,968, exp (0,6772), o risco de germinação das sementes extraídas enzimaticamente, foi aproximadamente duas vezes a das sementes extraídas manualmente. A probabilidade de germinação das sementes extraídas por ambos os métodos pode ser observada na figura 2.3a.

O modelo de Cox assume a proporcionalidade das taxas de falha, essa suposição foi verificada com base nas curvas apresentadas na Figura 2.3b. A análise não encontrou evidências para rejeitar a hipótese de proporcionalidade.

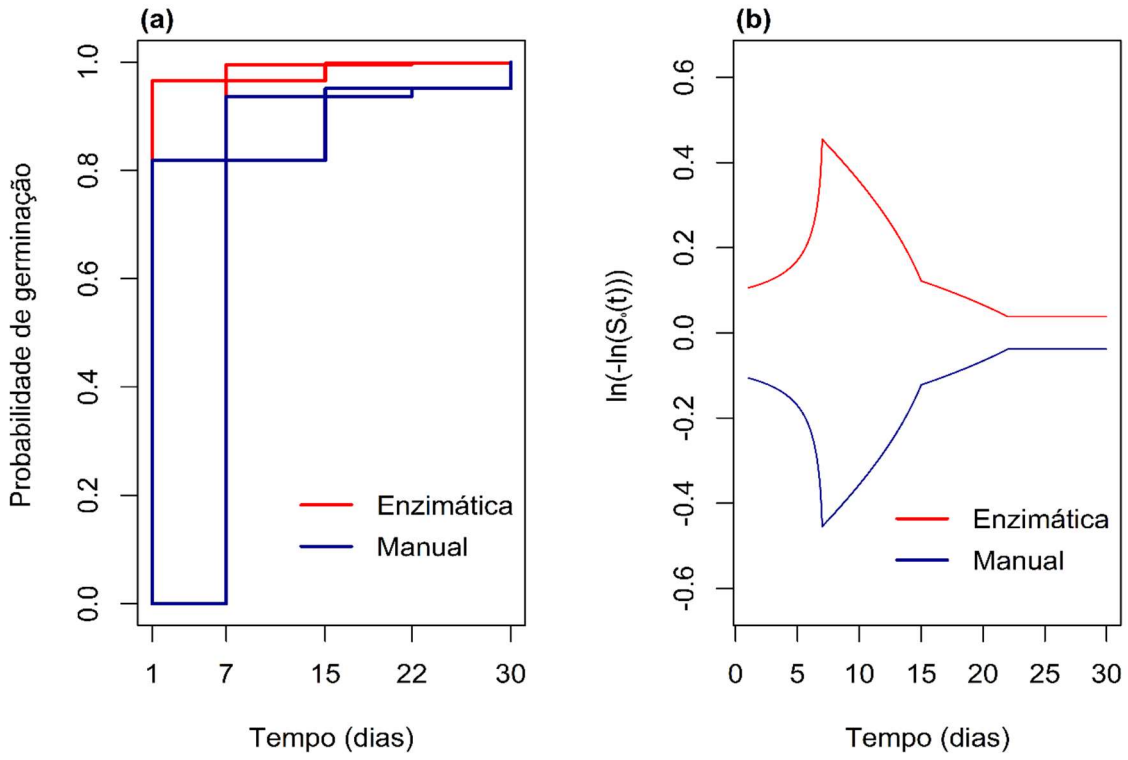

Figura 2.3 – Curvas de germinação estimadas a partir do modelo semiparamétrico de Cox (a) e gráfico de diagnóstico para avaliar a suposição de riscos proporcionais (b), para sementes de pitaia (*Hylocereus* spp.) extraídas pelo método manual e enzimático.

2.4.4. Regressão Discreta

Os percentuais finais de germinação para as diferentes temperaturas e tempos de armazenamento estão apresentados na Tabela 2.7. Esses resultados foram obtidos por meio do cálculo simples da porcentagem de sementes germinadas em relação ao total de sementes observadas.

No estudo utilizando a regressão discreta foram utilizados os modelos de riscos proporcionais de Cox e o modelo de chances proporcionais que apresentaram, respectivamente, valores de AIC iguais a 3240,85 e 2998,17. Desta forma, o modelo selecionado para o estudo da germinação de sementes de pitaia foi o de chances proporcionais, que apresentou menor valor de AIC. Os valores estimados por este modelo estão apresentados na Tabela 2.8, na qual podem ser observados os efeitos significativos *(p* < 0,05) dos intervalos de tempo (γ_i^*) e das interações entre os fatores tempo de armazenamento *versus* temperatura, β_6 *a* β_{11}, com exceção de β_7 e β_{10}.

Tabela 2.7 – Porcentagem de falhas (germinação) e de censuras (não germinação) de sementes de pitaia (*Hylocereus* spp.) em função da temperatura e do tempo de armazenamento: análise descritiva dos dados.

Temperatura	12 meses		13 meses		14 meses	
	Falha (%)	Censura (%)	Falha (%)	Censura (%)	Falha (%)	Censura (%)
15 ºC	96,7	03,3	93,3	06,7	67,1	32,9
20 ºC	98,7	01,3	99,6	00,4	100,0	00,0
25 ºC	100,0	00,0	97,8	02,2	99,1	00,9
30 ºC	97,8	02,2	99,1	00,9	99,1	00,9

Tabela 2.8 - Valores estimados pelo modelo logístico, para a germinação de sementes de pitaia (*Hylocereus* spp.), armazenadas por diferente tempos e temperatura.

Parâmetros	Estimativa	Erro Padrão	z	Valor p
γ_1^*	-1,2703	0,1321	-9,6140	< 0,0001
γ_2^*	0,2363	0,1168	2,0220	0,0432
γ_3^*	1,2698	0,1592	7,9740	< 0,0001
γ_4^*	0,8453	0,2083	4,0590	< 0,0001
β_1	-0,7630	0,1400	-5,4510	< 0,0001
β_2	-1,9111	0,1546	-12,3580	< 0,0001
β_3	3,7867	0,2841	13,3300	< 0,0001
β_4	4,7140	0,4059	11,6140	< 0,0001
β_5	3,5061	0,2596	13,5040	< 0,0001
β_6	1,5532	0,4612	3,3680	0,0008
β_7	-0,3424	0,4709	-0,7270	0,4671
β_8	0,8204	0,3496	2,3470	0,0189
β_9	2,8383	0,4847	5,8560	< 0,0001
β_{10}	0,9246	0,4811	1,9220	0,0546
β_{11}	2,0750	0,3627	5,7210	< 0,0001

$\eta_{li} = \gamma_1^* * int1 + \gamma_2^* * int2 + \gamma_3^* * int3 + \gamma_4^* * int4 + \beta_1 * armaz(13) + \beta_2 * armaz(14) + \beta_3 * temp(20) + \beta_4 * temp(25) + \beta_5 * temp(30) + \beta_6 * armaz(13) * temp(20) + \beta_7 * armaz(13) * temp(25) + \beta_8 * armaz(13) * temp(30) + \beta_9 * armaz(14) * temp(20) + \beta_{10} * armaz(14) * temp(25) + \beta_{11} * armaz(14) * temp(30)$; $\gamma_i^* = \ln(\gamma_i)$ é o efeito do i-ésimo intervalo de tempo.

Na Figura 2.4 está apresentada a curva ROC com a finalidade de avaliar o desempenho do modelo selecionado. Pode ser observado que a área abaixo da curva (AUC) foi de 0,9040, considerando o modelo com um alto desempenho em classificar as sementes em germinadas e não germinadas.

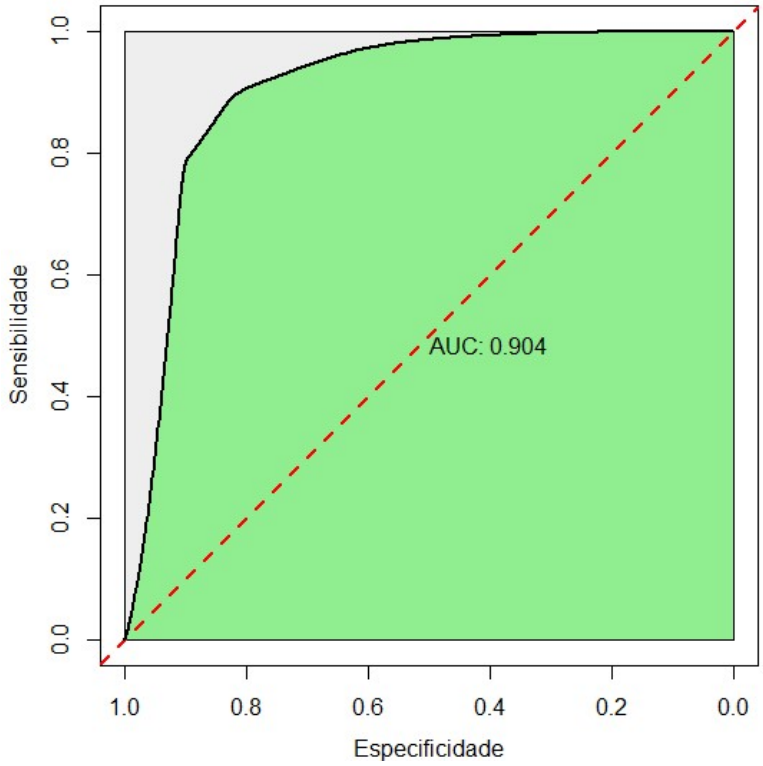

Figura 2.4 – Curva ROC mostrando a área abaixo da curva (AUC), estimada pelo modelo de chances proporcionais, no estudo da germinação de sementes de pitaia (*Hylocereus* spp.), armazenadas por diferente tempos e temperatura.

Na Figura 2.5, observa-se que as temperaturas de 20°C, 25°C e 30°C resultaram em porcentagens de germinação semelhantes e elevadas em todos os tempos de armazenamento. Essas temperaturas alcançaram valores expressivos já no primeiro intervalo de observação (1-7 dias), aproximando-se de uma germinação total no segundo intervalo (7-15 dias). Em contraste, a temperatura de 15°C apresentou porcentagens de germinação inferiores, que foram diminuindo com o aumento do tempo de armazenamento.

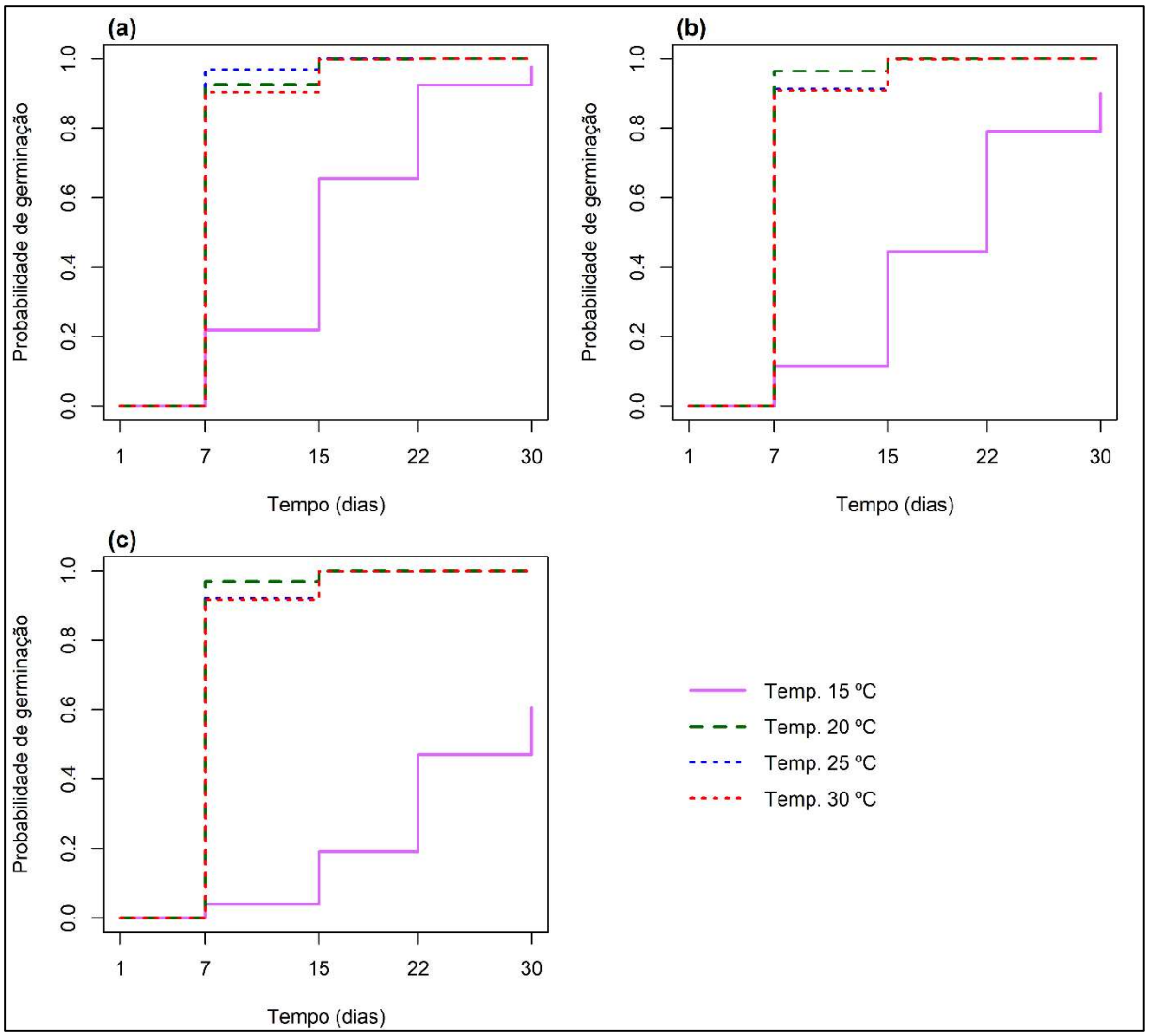

Figura 2.5 – Curvas de germinação de sementes de pitaia (Hylocereus spp.), estimadas pelo modelo de chances proporcionais, sob diferentes temperaturas e por tempos de armazenamento de: (a) 12 meses, (b) 13 meses e (c)14 meses.

2.5. Discussão

Ao investigar o efeito da luz e dos períodos de armazenamento de 12, 13 e 14 meses por meio de uma abordagem não paramétrica, os dados indicam que a ausência de luz retarda a germinação das sementes de pitaia. Nos estudos de Lone *et al*. (2014), as sementes de *Hylocereus undatus*, *Hylocereus polyrhizus*, *Selenicereus megalanthus* e dos híbridos *Hylocereus undatus* x *Hylocereus costaricensis* e *Hylocereus costaricensis* x *Hylocereus undatus* germinaram no escuro. Kataoka *et al*. (2013) sugerem que, apesar do retardamento na ausência de luz, a espécie *Hylocereus undatus* não necessita de luz para germinar, uma vez que sua

porcentagem de germinação foi alta ao final do período avaliado. No presente artigo, o mesmo ocorreu para os tempos de armazenamento de 12 e 13 meses, indicando também que a germinação ocorre mesmo na ausência de luz, mesmo que de forma mais lenta. Em contraste, no tempo de armazenamento de 14 meses, as porcentagens foram significativamente baixas, sugerindo que a capacidade de germinação na ausência de luz diminui com o tempo de armazenamento maior.

Lepiten (2023) relatou que, embora as qualidades da luz fossem comparáveis, a luz vermelha promoveu uma porcentagem de germinação mais alta do que as outras cores. No entanto, de forma geral, este estudo não encontrou diferenças significativas entre as cores de luz na germinação. A diferença observada entre a luz vermelha e a luz azul no 13° mês de armazenamento pode ter sido influenciada pela diferença das porcentagens de germinação no primeiro intervalo observado.

Pelos resultados obtidos na análise paramétrica, nota-se que o tempo de armazenamento em condições ambiente pode contribuir para a perda de viabilidade e vigor das sementes. Isso pode ocorrer porque, segundo Girardi *et al*. (2013), após a maturação fisiológica, inicia-se o processo de deterioração progressiva, reduzindo as porcentagens de germinação. Resultados semelhantes aos deste estudo foram observados por Zerpa-Catanho et al. (2019), embora com algumas variações nas porcentagens de germinação. Em sua pesquisa, ao analisar os efeitos do local e do tempo de armazenamento, as técnicas tradicionais demonstraram porcentagens finais de germinação mais altas em comparação à análise de sobrevivência paramétrica. Essa diferença pode ser explicada pelo fato de que as técnicas tradicionais não consideram adequadamente as censuras à direita, limitando-se a calcular a média das porcentagens de germinação de cada repetição, subestimando as porcentagens de germinação (Scott e Jones, 1990).

Os resultados da modelagem semiparamétrica sugerem que a extração enzimática é uma técnica mais eficiente para promover a germinação das sementes de pitaia, visto que a porcentagem de germinação das sementes submetidas à extração enzimática supera a da extração manual, além de apresentar maior velocidade no processo, alcançando valores superiores já no primeiro intervalo observado. Isso indica que, além de ser mais eficaz, a extração enzimática oferece uma vantagem temporal em relação à extração manual.

Quando comparada à análise tradicional, os valores obtidos pela análise semiparamétrica de sobrevivência para verificar o efeito do método de extração são

próximos, mas não idênticos. Essa diferença pode ser explicada pela limitação da análise tradicional, que não consegue capturar a trajetória completa das sementes ao longo do tempo. A análise de sobrevivência, por outro lado, oferece uma representação mais precisa dos dados, considerando toda a evolução do processo de germinação, inclusive as censuras e as variações temporais que a abordagem tradicional tende a ignorar (Onofri *et al*., 2010).

Os resultados da regressão discreta indicam que a temperatura de 15 ºC prejudica a germinação de *Hylocereus* spp. A temperatura influencia a velocidade de absorção de água e modifica a velocidade das reações químicas que mobilizam ou degradam as reservas armazenadas, além de promover a síntese de substâncias necessárias para o crescimento das plântulas (Bewley e Black, 1994). Segundo Carvalho e Nakagawa (2000), temperaturas baixas podem afetar negativamente a velocidade e a porcentagem de germinação, prejudicando a absorção de água pelas sementes e, consequentemente, reduzindo as reações bioquímicas e fisiológicas que determinam a germinação. Isso explica por que as sementes germinaram em menor quantidade e mais lentamente na temperatura de 15°C.

O tempo de armazenamento ter afetado a germinação das sementes apenas na temperatura de 15°C, reduzindo o percentual e a porcentagem de germinação pode ser explicada por um estresse térmico sofrido pela semente nessa temperatura intensificando os efeitos da degradação das sementes pelo tempo de armazenamento.

Em contraste com a análise tradicional, a análise de sobrevivência não apenas leva em consideração todos os tipos de censura, mas também permite explorar a trajetória completa da germinação. Com isso, a análise de sobrevivência oferece uma compreensão mais profunda de outros fatores, como a velocidade de germinação e vigor das sementes, ao invés de se restringir apenas a momentos específicos de germinação, como pode ocorrer em análises convencionais.

Nesse contexto, com a finalidade de comparar os diferentes grupos de sementes, utilizar um método não paramétrico é uma escolha apropriada, pois evita-se qualquer dependência de suposições paramétricas (McNair *et al*., 2012). Além disso, como apontado por Anderson-Bergman, (2017b) o uso de métodos não paramétricos é frequentemente preferido, especialmente para diagnósticos. As curvas estimadas por esses métodos podem servir como base para técnicas paramétricas ou

semiparamétricas adicionais e podem ser interpretadas de maneira semelhante às curvas de Kaplan-Meier para censura à direita (Gomez *et al*., 2009).

Embora as técnicas não paramétricas sejam valiosas para descrever dados de sobrevivência devido à sua simplicidade e facilidade de aplicação, é importante reconhecer suas limitações. Essas limitações incluem a redução do poder estatístico (McNair *et al*., 2012) e a incapacidade de incorporar covariáveis diretamente na análise (Colosimo e Giolo, 2024). Tal situação pode exigir a divisão da amostra em vários subgrupos, como foi realizado nessa análise. Embora a separação por condições de luz e tempos de armazenamento não tenha complicado a interpretação dos resultados, essa abordagem pode se tornar mais complexa quando o número de subgrupos é grande.

Para incorporar covariáveis e realizar predições, a alternativa é utilizar a abordagem paramétrica. Lindsey e Ryan, (1998) concluíram que métodos paramétricos para censura intervalar são os mais prontamente disponíveis e que seu desempenho é satisfatório, o que se confirma também no presente estudo. Porém, é preciso ter cautela ao utilizar esses métodos pois eles são fortemente influenciados pela escolha do modelo paramétrico, para o qual a inspeção do modelo pode ser difícil (Anderson-Bergman, 2017b). Nesse contexto, foi selecionado o modelo paramétrico que apresentou o melhor ajuste, o qual alinhou-se aos resultados obtidos na análise não paramétrica e às expectativas.

Uma alternativa ao uso dos modelos paramétricos, são os modelos semiparamétricos. O algoritmo proposto por Anderson-Bergman (2017b) demonstrou ser eficiente na estimativa do modelo proporcional de Cox para dados censurados por intervalo. Contudo, a validade desse modelo depende da suposição de proporcionalidade das taxas de falha, que requer que a razão das taxas de falha entre dois indivíduos permaneça constante ao longo do tempo. A simetria das curvas para os grupos de extração enzimática e manual sugere que a razão de risco entre esses grupos é constante ao longo do tempo.

O tempo até a germinação foi medido semanalmente durante quatro semanas, o que resultou em muitos dados empatados. Nesse contexto, a regressão discreta apresentada surge como uma alternativa para abordar esse problema. No entanto, devido à pitaia apresentar rápida germinação, observou-se um alto percentual de germinação já na primeira semana, sugerindo que intervalos de monitoramento mais curtos poderiam ter capturado mais informações durante esse

período crítico. Portanto, recomenda-se que o acompanhamento da germinação das sementes de pitaia seja realizado em intervalos menores, a fim de garantir uma análise mais precisa da dinâmica desse processo. Além disso, considerando que a pitaia não está incluída nos padrões da ISTA (Associação Internacional de Testes de Sementes), é fundamental realizar estudos adicionais para estabelecer um padrão específico para essa espécie.

2.6. Conclusões

Apesar da alta incidência de empates nos dados, a aplicação dos métodos utilizados mostrou-se valiosa para estabelecer padrões e conclusões sobre a germinação das sementes com censura intervalar. Esses métodos se destacam não apenas por considerarem a censura na análise, mas também por permitirem o uso de intervalos de tempo em vez de exigir uma observação exata para cada evento, o que possibilita captar tendências e obter uma visão mais completa do processo de germinação e suas variabilidades. Futuras pesquisas podem explorar esses métodos em contextos distintos, ampliando o conhecimento sobre a germinação e outros fenômenos de interesse.

Referências

Anderson-Bergman, C., 2017a. An efficient implementation of the EMICM algorithm for the interval censored NPMLE. **Journal of Computational and Graphical Statistics**, *26*(2), 463-467. https://doi.org/10.1080/10618600.2016.1208616

Anderson-Bergman, C., 2017b. icenReg: Regression Models for Interval Censored Data in R. **Journal of Statistical Software**, *81*(12), 1–23. https://doi.org/10.18637/jss.v081.i12

Barak, R. S., Lichtenberger, T. M., Wellman-Houde, A., Kramer, A. T., Larkin, D. J., 2018. Cracking the case: Seed traits and phylogeny predict time to germination in prairie restoration species. **Ecology and Evolution**, *8*(11), 5551-5562. https://doi.org/10.1002/ece3.4083

Bewley, J. D., Black, M., 1994. **Seeds: physiology of development and germination**. 2nd ed., Plenum Press: New York, NY, USA

Boruvka, A., Cook, R. J., 2015. A Cox-Aalen Model for Interval-censored Data. **Scandinavian Journal of Statistics**, *42*(2), 414-426. https://doi.org/10.1111/sjos.12113

Carvalho, N. M., Nakagawa, J., 2000. **Sementes: ciência, tecnologia e produção**. 4. ed. rev. ampl. Jaboticabal, SP: FUNEP.

Collett, D., 1991, **Modelling Binary Data**, New York: Chapman & Hall.

Colosimo, E. A., Chalita, L. V., Demétrio, C. G., 2000. Tests of proportional hazards and proportional odds models for grouped survival data. **Biometrics**, *56*(4), 1233-1240. https://doi.org/10.1111/j.0006-341X.2000.01233.x

Colosimo, E. A., Giolo, S. R., 2024. **Análise de sobrevivência aplicada**. 2. ed. São Paulo: Editora Blücher. 362p.

Cox, D. R., 1972. Regression models and life-tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, *34*(2), 187-202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Fleming, T. R., Harrington, D. P., 1991. **Counting processes and survival analysis**. John Wiley & Sons, New York.

Giolo, S. R., Colosimo, E. A., Demétrio, C. G. B., 2009. Different approaches for modeling grouped survival data: A mango tree study. **Journal of agricultural, biological, and environmental statistics**, 14, 154-169. https://doi.org/10.1198/jabes.2009.0010

Girardi, L. B., Bellé, R. A., Lazarotto, M., Michelon, S., Girardi, B. A., Muniz, M. F. B., 2013. Qualidade de sementes de cártamo colhidas em diferentes períodos de maturação. **Revista Acadêmica Ciência Animal**, *11*, 67-73. https://doi.org/10.7213/academica.10.S01.AO08

Gómez, G., Calle, M. L., Oller, R., Langohr, K., 2009. Tutorial on methods for interval-censored data and their implementation in R. **Statistical Modelling**, *9*(4), 259-297. https://doi.org/10.1177/1471082X0900900402

Hashimoto, E. M., Ortega, E. M., Paula, G. A., Barreto, M. L., 2011. Regression models for grouped survival data: estimation and sensitivity analysis. **Computational statistics & data analysis**, *55*(2), 993-1007. https://doi.org/10.1016/j.csda.2010.08.004

Hernández-Pridybailo, A., Zerpa-Catanho, D., Madrigal-Ortiz, V., Zúñiga-Centeno, A., Porras-Martínez, C., Jiménez, V. M., Barboza-Barquero, L.,. Pitaya seed germination research data [dataset]. 22 nov. 2018. Mendeley Data (MD), V1, doi: 10.17632/ddmz8x5spy.1

Hernández, Y. D. O., Salazar, J. A. C., 2012. Pitahaya (*Hylocereus* spp.): a short review. **Comunicata Scientiae**, *3*(4), 220-237.

Hosmer, D. W., Lemeshow, S., 2000. **Applied Logistic Regression**. John Wiley and Sons, New York, 2nd edition

Hua, Q., Chen, P., Liu, W., Ma, Y., Liang, R., Wang, L., Qin, Y., 2015. A protocol for rapid in vitro propagation of genetically diverse pitaya. **Plant Cell, Tissue and Organ Culture (PCTOC)**, *120*, 741-745.

Jalgaonkar, K., Mahawar, M. K., Bibwe, B., Kannaujia, P., 2020. Postharvest Profile, Processing and Waste Utilization of Dragon Fruit (*Hylocereus Spp*.): A Review. **Food Reviews International**, *38*(4), 733–759. https://doi.org/10.1080/87559129.2020.1742152

Kataoka, Fukuda, S., Kozai, N., Beppu, K. and Yonemoto, Y., 2013. Conditions for seed germination in Pitaya. **Acta Hortic**. 975, 281-285. https://doi.org/10.17660/ActaHortic.2013.975.32

Kaplan, E. L., Meier, P., 1958. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, *53*(282), 457-481. https://doi.org/10.1007/978-1-4612-4380-9_25

Lawless, J. F., 2003. **Statistical models and methods for lifetime data**. John Wiley & Sons, New York, 2nd edition.

Le Bellec, F., Vaillant, F., Imbert, E., 2006. Pitahaya (*Hylocereus* spp.): a new fruit crop, a market with a future. **Fruits**, *61*(4), 237-250. https: //doi.org/10.1051/fruits: 2006021

Lepiten, R. C., 2023. Light quality and duration on the germination of dragon fruit (*Hylocereus* spp.) grown out on different potting mixture. **Thai Journal of Agricultural Science**, *56*(1), 16-25

Lindsey, J. C., Ryan, L. M., 1998. Methods for interval-censored data. **Statistics in medicine**, *17*(2), 219-238. https://doi.org/10.1002/(SICI)1097-0258(19980130)17:2<219::AID-SIM735>3.0.CO;2-O

Lone, A. B., Unemoto, L. K., Ferrari, E. A. P., Takahashi, L. S. A., de Faria, R. T., 2014. The effects of light wavelength and intensity on the germination of pitaya seed genotypes. **Australian Journal of Crop Science**, *8*(11), 1475-1480. https://search.informit.org/doi/10.3316/informit.818276748944057

Luu, T. T. H., Le, T. L., Huynh, N., Quintela-Alonso, P., 2021. Dragon fruit: A review of health benefits and nutrients and its sustainable development under climate changes in Vietnam. **Czech Journal of Food Sciences**, *39*(2), 71-94. doi: 10.17221/139/2020-cjfs

McNair, J. N., Sunkara, A., Frobish, D., 2012. How to analyse seed germination data using statistical time-to-event analysis: non-parametric and semi-parametric methods. **Seed Science Research**, *22*(2), 77-95. https://doi.org/10.1017/S0960258511000547

Nishikito, D.F., Borges, A.C.A., Laurindo, L.F., Otoboni, A.M.M.B., Direito, R., Goulart, R.d.A., Nicolau, C.C.T., Fiorini, A.M.R., Sinatora, R.V., Barbalho, S.M., 2023. Anti-Inflammatory, Antioxidant, and Other Health Effects of Dragon Fruit and Potential Delivery Systems for Its Bioactive Compounds. **Pharmaceutics**, 15(1), 159. https://doi.org/10.3390/pharmaceutics15010159

Nur, M. A., Uddin, M. R., Uddin, M. J., Satter, M. A., Amin, M. Z., 2023. Physiochemical and nutritional analysis of the two species of dragon fruits (*Hylocereus* spp.) cultivated in Bangladesh. **South African Journal of Botany**, *155*, 103-109. https://doi.org/10.1016/j.sajb.2023.02.006

Oller, R., Gómez, G., 2012. A generalized Fleming and Harrington's class of tests for interval-censored data. **Canadian Journal of Statistics**, *40*(3), 501-516. https://doi.org/10.1002/cjs.11139

Oller R, Langohr K., 2017. FHtest: An R Package for the Comparison of Survival Curves with Censored Data. **Journal of Statistical Software**, 81(15), 1-25. https://doi.org/10.18637/jss.v081.i15.

Onofri, A., Gresta, F., Tei, F., 2010. A new method for the analysis of germination and emergence data of weed species. **Weed Research**, 50(3), p. 187–198. https://doi.org/10.1111/j.1365-3180.2010.00776.x

Pan, W., 1999. Extending the iterative convex minorant algorithm to the Cox model for interval-censored data. **Journal of Computational and Graphical Statistics**, *8*(1), 109-120. https://doi.org/10.1080/10618600.1999.10474804

Prentice, R. L., Gloeckler, L. A., 1978. Regression analysis of grouped survival data with application to breast cancer data. **Biometrics**, 57-67. https://doi.org/10.2307/2529588

R Core Team, 2024. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria. Available online: <https://www.R-project.org/>.

Schwarz, G., 1978. Estimating the dimension of a model. **The annals of statistics**, 461-464.

Scott, S. J., Jones, R. A., 1990. Generation means analysis of right-censored response-time traits: Low temperature seed germination in tomato. **Euphytica**, v. 48, p. 239-244. https://doi.org/10.1007/BF00023656

Shah, K., Chen, J., Chen, J., Qin, Y., 2023. Pitaya Nutrition, Biology, and Biotechnology: A Review. **International Journal of Molecular Sciences**, v. 24, n. 18, p. 13986. https://doi.org/10.3390/ijms241813986

Silveira, L. V., Colosimo, E. A., Passos, J. R. D. S., 2010. A Generalized Log-Normal Model for Grouped Survival Data. **Communications in Statistics-Theory and Methods**, *39*(15), 2659-2666. https://doi.org/10.1080/03610920903009368

Tarone, R. E., Ware, J., 1977. On distribution-free tests for equality of survival distributions. **Biometrika**, *64*(1), 156-160. https://doi.org/10.1093/biomet/64.1.156

Therneau T., 2024. **A Package for Survival Analysis in R**. R package version 3.5-8. Available online: <https://CRAN.R-project.org/package=survival>.

Turnbull, B. W., 1976. The empirical distribution function with arbitrarily grouped, censored and truncated data. **Journal of the Royal Statistical Society: Series B (Methodological)**,38(3),290-295. https://doi.org/10.1111/j.2517-6161.1976.tb01597.x

Vaca, F. E., Li, K., Gao, X., Zagnoli, K., Wang, H., Haynie, D. L., Fell, J. C., Simons-Morton, B., Romano, E., 2021. Time to licensure for driving among U.S. teens: Survival analysis of interval-censored survey data. **Traffic Injury Prevention**, *22*(6), 431–436. https://doi.org/10.1080/15389588.2021.1939871

Waterworth, W. M., Bray, C. M., West, C. E., 2019. Seeds and the art of genome maintenance. **Frontiers in Plant Science**, 10, 706.

Welchowski T, Berger M, Koehler D, Schmid M., 2022. **discSurv: Discrete Time Survival Analysis**. R package version 2.0.0. Available online: <https://CRAN.R-project.org/package=discSurv>.

Wellner, J. A., Zhan, Y., 1997. A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. **Journal of the American Statistical Association**, *92*(439), 945-959. https://doi.org/10.1080/01621459.1997.10474049

Zerpa-Catanho, D., Hernández-Pridybailo, A., Madrigal-Ortiz, V., Zúñiga-Centeno, A., Porras-Martínez, C., Jiménez, V. M., Barboza-Barquero, L., 2019. Seed germination of pitaya (*Hylocereus* spp.) as affected by seed extraction method, storage, germination conditions, germination assessment approach and water potential. **Journal of Crop Improvement**, v. 33, n. 3, p. 372-394. https://doi.org/10.1080/15427528.2019.1604457

APÊNDICE

Apêndice A - Algoritmos utilizados para as análises

```
## Análise não paramétrica

#Pacotes e Funcoes
 require(openxlsx)
 require(survival)
 require(icenReg)
#Leitura do banco de dados
 dados <- read.xlsx("EfeitoLuz.xlsx")
 head(dados)
#Preparação dos dados
 dados[is.na(dados[,"right"]), "right"]<-"inf"
 dados[is.na(dados[,"left"]), "left"]<-0
 dados$right<-as.numeric(dados$right)
 attach(dados)

#Ajuste dos modelos
#====================================
#1-Seleção do ARMAZ=12
#====================================
 dados12 <- subset(dados, armaz==12)
 dados12
 fit12<-ic_np(cbind(left, right) ~ luz, maxIter = 10000,
       tol = 10^-10, data = dados12)
 fit12$scurves
 azul12 <- fit12$scurves[[1]]
 escuro12 <- fit12$scurves[[2]]
 vermelha12 <- fit12$scurves[[3]]
 branca12 <- fit12$scurves[[4]]
 dadosvazios <- data.frame(x=numeric(0), y=numeric(0))

#Curvas com as probabilidades de germinação para o tempo de
armazenamento de 12 meses
 par(mfrow=c(2,2))
 color = c("blue","black","red","seagreen4")
 lty=c(1,1,2,1)
 plot(dadosvazios, xlab="Time (days)", xlim=c(1,30),
      ylab="Seed germination probability",ylim=range(c(0,1)),
      xaxt="n",yaxt="n")
 axis(1, at=c(1,8,15,22,30))
 axis(2, at=seq(0,1,by=0.1))
 lines(azul12, fun = "cdf", lty=1, col="blue", lwd = 2.5)
 lines(escuro12, fun = "cdf", lty=1, col="black", lwd = 2.5)
 lines(branca12, fun = "cdf", lty=2, col="seagreen4", lwd = 2.5)
 lines(vermelha12, fun = "cdf", lty=3, col="red", lwd = 2.5)
 mtext("(a)",side=3,adj=-0.05,at=0.3,line=0.2, font=2, cex=1.0)
```

```
#Teste logrank para comparacao das curvas
 require(FHtest)

#Azul x Ausência
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados12,luz       %in%
          c("blue","darkness")), alternative="different")

#Azul x Vermelha
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados12,luz       %in%
          c("blue","red")), alternative="different")

#Azul x Branca
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados12,luz       %in%
          c("blue","white")), alternative="different")

#Ausência x Vermelha
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados12,luz       %in%
          c("darkness","red")), alternative="different")

#Ausência x Branca
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          method="exact.mc",data=subset(dados12,luz        %in%
          c("darkness","white")), alternative="different")

#Vermelha x Branca
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados12,luz       %in%
          c("red","white")), alternative="different")

#====================================
#1-Seleção do ARMAZ=13
#====================================
 dados13 <- subset(dados, armaz==13)
 dados13
 fit13<-ic_np(cbind(left, right) ~ luz, maxIter = 10000,
       tol = 10^-10, data = dados13)
 fit13$scurves
 azul13 <- fit13$scurves[[1]]
 escuro13 <- fit13$scurves[[2]]
 vermelha13 <- fit13$scurves[[3]]
 branca13 <- fit13$scurves[[4]]
 dadosvazios <- data.frame(x=numeric(0), y=numeric(0))

#Curvas com as probabilidades de germinação para o tempo de
armazenamento de 13 meses
 color = c("blue","black","red","seagreen4")
 lty=c(1,1,2,1)
```

```
 plot(dadosvazios, xlab="Time (days)", xlim=c(1,30),
     ylab="Seed germination probability",ylim=range(c(0,1)),
     xaxt="n",yaxt="n")
 axis(1, at=c(1,8,15,22,30))
 axis(2, at=seq(0,1,by=0.1))
 lines(escuro13, fun = "cdf", lty=1, col="black", lwd = 2.5)
 lines(branca13, fun = "cdf", lty=1, col="seagreen4", lwd = 2.5)
 lines(vermelha13, fun = "cdf", lty=1, col="red", lwd = 2.5)
 lines(azul13, fun = "cdf", lty=2, col="blue", lwd = 2.5)
 mtext("(b)",side=3,adj=-0.05,at=0.3,line=0.2, font=2, cex=1.0)

#Teste logrank para comparacao das curvas
 require(FHtest)

#Azul x Ausência
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          method="exact.mc",data=subset(dados13,luz        %in%
          c("blue","darkness")), alternative="different")

#Azul x Vermelha
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados13,luz       %in%
          c("blue","red")), alternative="different")

#Azul x Branca
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados13,luz       %in%
          c("blue","white")), alternative="different")

#Ausência x Vermelha
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados13,luz       %in%
          c("darkness","red")), alternative="different")

#Ausência x Branca
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
           ,method="exact.mc",data=subset(dados13,luz      %in%
           c("darkness","white")), alternative="different")

#Vermelha x Branca
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados13,luz       %in%
          c("red","white")), alternative="different")

#====================================
#1-Seleção do ARMAZ=14
#====================================
 dados14 <- subset(dados, armaz==14)
 dados14
 fit14<-ic_np(cbind(left, right) ~ luz, maxIter = 10000,
       tol = 10^-10, data = dados14)
```

```
 fit12$scurves
 azul14 <- fit14$scurves[[1]]
 escuro14 <- fit14$scurves[[2]]
 vermelha14 <- fit14$scurves[[3]]
 branca14 <- fit14$scurves[[4]]
 dadosvazios <- data.frame(x=numeric(0), y=numeric(0))

#Curvas com as probabilidades de germinação para o tempo de
armazenamento de 14 meses
 color = c("blue","black","red","seagreen4")
 lty=c(1,1,2,1)
 plot(dadosvazios, xlab="Time (days)", xlim=c(1,30),
     ylab="Seed germination probability",ylim=range(c(0,1)),
     xaxt="n",yaxt="n")
 axis(1, at=c(1,8,15,22,30))
 axis(2, at=seq(0,1,by=0.1))
 lines(escuro14, fun = "cdf", lty=1, col="black", lwd = 2.5)
 lines(branca14, fun = "cdf", lty=1, col="seagreen4", lwd = 2.5)
 lines(vermelha14, fun = "cdf", lty=1, col="red", lwd = 2.5)
 lines(azul14, fun = "cdf", lty=2, col="blue", lwd = 2.5)
 mtext("(c)",side=3,adj=-0.05,at=0.3,line=0.2, font=2, cex=1.0)

#Teste logrank para comparacao das curvas
 require(FHtest)

#Azul x Ausência
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados14,luz       %in%
          c("blue","darkness")), alternative="different")

#Azul x Vermelha
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados14,luz       %in%
          c("blue","red")), alternative="different")

#Azul x Branca
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados14,luz       %in%
          c("blue","white")), alternative="different")

#Ausência x Vermelha
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados14,luz       %in%
          c("darkness","red")), alternative="different")

#Ausência x Branca
FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          ,method="exact.mc",data=subset(dados14,luz       %in%
          c("darkness","white")), alternative="different")
```

```
#Vermelha x Branca

FHtesticp(Surv(left,right,type="interval2")~luz,rho=0,lambda=0
          method="exact.mc",data=subset(dados14,luz        %in%
          c("red","white")), alternative="different")

#Legenda
 Lty<-c(1,1,1,1)
 Color = c("seagreen4","blue","red","black")
 plot(1:15, 1:15, type = "n", ann = FALSE, axes = FALSE, bg =
"green",
      xaxt = "n", yaxt = "n")
 legend("left",  lty=Lty,  lwd=2,bty="n",ncol=1,cex=1.0,  col  =
Color,
        x.intersp=1.0,y.intersp=1.4,seg.len=2.0,
        c("Branca","Azul","Vermelha","Ausência"))

## Análise paramétrica

#Pacotes e Funcoes
 require(openxlsx)
 require(survival)
#Leitura do banco de dados
 dados <- read.xlsx("EfeitoArmazenamento.xlsx")
 head(dados)
#Preparação dos dados
 dados$right[is.na(dados$right)] <- Inf
 dados$lugar<- factor(dados$lugar,levels=c("ambiente", "cold"))
 attach(dados)
#Ajuste dos modelos e cálculo dos erros padrão
#Exponencial
 fit1<-
survreg(Surv(left,right,type="interval2")~month*lugar,dados,
              dist="exponential")
 summary(fit1)
 AIC1<-AIC(fit1);AIC1
 BIC1<-BIC(fit1);BIC1
 vb <- vcov(fit1);vb #ou ajust2$var
 grad <- fit1$scale;grad #ja em exponencial
 vb2 <- vcov(fit1)[5,5];vb2
 vG <- grad %% vb2 %% grad;vG
 sqrt(vG) # Std. Error do scale
#Weibull
 fit2<-
survreg(Surv(left,right,type="interval2")~month*lugar,dados,
              dist="weibull")
 summary(fit2)
 AIC2<-AIC(fit2);AIC2
 BIC2<-BIC(fit2);BIC2
 vb <- vcov(fit2);vb #ou ajust2$var
 grad <- fit2$scale;grad #ja em exponencial
```

```
 vb2 <- vcov(fit2)[5,5];vb2
 vG <- grad %% vb2 %% grad;vG
 sqrt(vG) # Std. Error do scale
#Lognormal
 fit3<-
survreg(Surv(left,right,type="interval2")~month*lugar,dados,
               dist="lognorm")
 summary(fit3)
 AIC3<-AIC(fit3);AIC3
 BIC3<-BIC(fit3);BIC3

 vb <- vcov(fit3);vb #ou ajust2$var
 grad <- fit3$scale;grad #ja em exponencial
 vb2 <- vcov(fit3)[5,5];vb2
 vG <- grad %% vb2 %% grad;vG
 sqrt(vG) # Std. Error do scale
#Loglogistico
 fit4<-
survreg(Surv(left,right,type="interval2")~month*lugar,dados,
               dist="loglogistic")
 summary(fit4)
 AIC4<-AIC(fit4);AIC4
 BIC4<-BIC(fit4);BIC4
 vb <- vcov(fit4);vb #ou ajust2$var
 grad <- fit4$scale;grad #ja em exponencial
 vb2 <- vcov(fit4)[5,5];vb2
 vG <- grad %% vb2 %% grad;vG
 sqrt(vG) # Std. Error do scale
#Tabela de AIC e BIC
 dist <- c("Exponencial", "Weibull", "LogNormal",
           "Log-Logistica","Logistica","Gaussiana")
 AIC <- as.numeric(c(AIC1,AIC2,AIC3,AIC4,AIC5, AIC6))
 BIC <- as.numeric(c(BIC1,BIC2,BIC3,BIC4,BIC5,BIC6))
 resultados <- cbind(AIC,BIC)
 rownames(resultados) <- dist
 colnames(resultados) <- c("AIC","BIC")
 print(resultados, digits = 7)
 min(AIC)
 min(BIC)

#Ajuste do modelo sem a interação para efeito de comparação
 fit4a <- survreg(Surv(left,right, type =
          "interval2")~month+lugar,
          data = dados, dist = "loglogistic")
 summary(fit4a)
 AIC(fit4a)
 anova(fit4a,fit4)
# h0 = não há diferença real entre os modelos
# Interacao continuou no modelo

#Sobrevivencia considerando o modelo Log-Logistico
```

```
  summary(fit4)
  sigma<-fit4$scale;sigma
  tempoE<-1:30;tempoE

#====================================
#1-Seleção da condição ambiente
#====================================
 month <- 1
 lugarcold <- 0
  etaamb1<-exp(fit4$coefficients[1] +
        fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
        fit4$coefficients[4]*month*lugarcold)
  etaamb1
  stAamb1<- 1/(1+(tempoE/etaamb1)^(1/sigma));stAamb1
  FtAamb1<-1-stAamb1;FtAamb1

 month <- 2
 lugarcold <- 0
  etaamb2<-exp(fit4$coefficients[1] +
        fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
        fit4$coefficients[4]*month*lugarcold)
  etaamb2
  stAamb2<- 1/(1+(tempoE/etaamb2)^(1/sigma));stAamb2
  FtAamb2<-1-stAamb2;FtAamb2

 month <- 3
 lugarcold <- 0
  etaamb3<-exp(fit4$coefficients[1] +
        fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
        fit4$coefficients[4]*month*lugarcold)
  etaamb3
  stAamb3<- 1/(1+(tempoE/etaamb3)^(1/sigma));stAamb3
  FtAamb3<-1-stAamb3;FtAamb3

 month <- 5
 lugarcold <- 0
  etaamb5<-exp(fit4$coefficients[1] +
        fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
        fit4$coefficients[4]*month*lugarcold)
  etaamb5
  stAamb5<- 1/(1+(tempoE/etaamb5)^(1/sigma));stAamb5
  FtAamb5<-1-stAamb5;FtAamb5

 month <- 7
 lugarcold <- 0
  etaamb7<-exp(fit4$coefficients[1] +
        fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
        fit4$coefficients[4]*month*lugarcold)
  etaamb7
  stAamb7<- 1/(1+(tempoE/etaamb7)^(1/sigma));stAamb7
  FtAamb7<-1-stAamb7;FtAamb7
```

```
 month <- 9
 lugarcold <- 0
  etaamb9<-exp(fit4$coefficients[1] +
        fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
        fit4$coefficients[4]*month*lugarcold)
  etaamb9
  stAamb9<- 1/(1+(tempoE/etaamb9)^(1/sigma));stAamb9
  FtAamb9<-1-stAamb9;FtAamb9

 month <- 12
 lugarcold <- 0
  etaamb12<-exp(fit4$coefficients[1] +
        fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
        fit4$coefficients[4]*month*lugarcold)
  etaamb12
  stAamb12<- 1/(1+(tempoE/etaamb12)^(1/sigma));stAamb12
  FtAamb12<-1-stAamb12;FtAamb12

#Curvas com as probabilidades de germinação para condição
ambiente
 par(mfrow=c(1,2))
 color = c("black","blue","red","seagreen4", "limegreen",
           "pink", "purple")
 lty=c(1,1,1,1,1,1,1)
 plot(tempoE, tempoE*0, pch="", xlim=range(c(1,30)),
     ylim=range(c(0,1)),      xlab="Time      (days)",ylab="Seed
     germination probability",xaxt="n", yaxt="n")
 axis(1, at = c(1,8,15,22,30))
 axis(2, at = seq(0, 1, by = 0.1))
 lines(tempoE,FtAamb1,lty=1,col="black",lwd =2.5)
 lines(tempoE,FtAamb2,lty=1,col="blue",lwd =2.5)
 lines(tempoE,FtAamb3,col="red",lwd =2.5,lty=1)
 lines(tempoE,FtAamb5,col="seagreen4",lwd =2.5,lty=1)
 lines(tempoE,FtAamb7,col="limegreen",lwd =2.5,lty=1)
 lines(tempoE,FtAamb9,col="pink",lwd =2.5,lty=1)
 lines(tempoE,FtAamb12,col="purple",lwd =2.5,lty=1)
 legend(20,0.35,lty=lty,col=color,lwd=2,c("1","2","3","5","7",
          "9","12"), bty="n", y.intersp = 1.2,  ncol=2, title =
          "Armazenamento (meses)")
 mtext("(a)",side=3,adj=-0.05,at=0.3,line=0.2, font=2, cex=1.0)

#=======================================
#1-Seleção da condição de câmara fria
#=======================================
 month <- 1
 lugarcold <- 1
  etacold1<-exp(fit4$coefficients[1] +
        fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
        fit4$coefficients[4]*month*lugarcold)
  etacold1
```

```
 stAcold1<- 1/(1+(tempoE/etacold1)^(1/sigma));stAcold1
 FtAcold1<-1-stAcold1;FtAcold1

month <- 2
lugarcold <- 1
 etacold2<-exp(fit4$coefficients[1] +
       fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
       fit4$coefficients[4]*month*lugarcold)
 etacold2
 stAcold2<- 1/(1+(tempoE/etacold2)^(1/sigma));stAcold2
 FtAcold2<-1-stAcold2;FtAcold2

month <- 3
lugarcold <- 1
 etacold3<-exp(fit4$coefficients[1] +
       fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
       fit4$coefficients[4]*month*lugarcold)
 etacold3
 stAcold3<- 1/(1+(tempoE/etacold3)^(1/sigma));stAcold3
 FtAcold3<-1-stAcold3;FtAcold3

month <- 5
lugarcold <- 1
 etacold5<-exp(fit4$coefficients[1] +
       fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
       fit4$coefficients[4]*month*lugarcold)
 etacold5
 stAcold5<- 1/(1+(tempoE/etacold5)^(1/sigma));stAcold5
 FtAcold5<-1-stAcold5;FtAcold5

month <- 7
lugarcold <- 1
 etacold7<-exp(fit4$coefficients[1] +
       fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
       fit4$coefficients[4]*month*lugarcold)
 etacold7
 stAcold7<- 1/(1+(tempoE/etacold7)^(1/sigma));stAcold7
 FtAcold7<-1-stAcold7;FtAcold7

month <- 9
lugarcold <- 1
 etacold9<-exp(fit4$coefficients[1] +
       fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
       fit4$coefficients[4]*month*lugarcold)
 etacold9
 stAcold9<- 1/(1+(tempoE/etacold9)^(1/sigma));stAcold9
 FtAcold9<-1-stAcold9;FtAcold9

month <- 12
lugarcold <- 1
 etacold12<-exp(fit4$coefficients[1] +
```

```
        fit4$coef[2]*month + fit4$coefficients[3]*lugarcold +
        fit4$coefficients[4]*month*lugarcold)
  etacold12
  stAcold12<- 1/(1+(tempoE/etacold12)^(1/sigma));stAcold12
  FtAcold12<-1-stAcold12;FtAcold12

#Curvas com as probabilidades de germinação para condição câmara
fria
 color = c("black","blue","red","seagreen4", "limegreen",
           "pink", "purple")
 lty=c(1,1,1,1,1,1,1)
 plot(tempoE, tempoE*0, pch="", xlim=range(c(1,30)),
          ylim=range(c(0,1)),   xlab="Time   (days)",ylab="Seed
          germination probability", xaxt="n", yaxt="n")
 axis(1, at = c(1,8,15,22,30))
 axis(2, at = seq(0, 1, by = 0.1))
 lines(tempoE,FtAcold1,lty=1,col="black",lwd =2.5)
 lines(tempoE,FtAcold2,lty=1,col="blue",lwd =2.5)
 lines(tempoE,FtAcold3,col="red",lwd =2.5,lty=1)
 lines(tempoE,FtAcold5,col="seagreen4",lwd =2.5,lty=1)
 lines(tempoE,FtAcold7,col="limegreen",lwd =2.5,lty=1)
 lines(tempoE,FtAcold9,col="pink",lwd =2.5,lty=1)
 lines(tempoE,FtAcold12,col="purple",lwd =2.5,lty=1)
 legend(20,0.35,lty=lty,col=color,lwd=2,c("1","2","3","5","7",
          "9","12"), bty="n",y.intersp = 1.2,  ncol=2, title =
          "Armazenamento (meses)")
 mtext("(b)",side=3,adj=-0.05,at=0.3,line=0.2, font=2, cex=1.0)

## Análise semiparamétrica

#Pacotes e Funcoes
 require(openxlsx)
 require(survival)
 require(icenReg)
#Leitura do banco de dados
 dados <- read.xlsx("EfeitoExtração.xlsx")
 head(dados)
#Preparação dos dados
 dados$right[is.na(dados$right)] <- Inf
 dados$extrac< factor(dados$extrac,  levels=c("manual",
              "enzymatic"))
 attach(dados)
#Usando doParallel para diminuir o tempo de processamento
 ptm <- proc.time()
 library(doParallel)
 myCluster <- makeCluster(4)
 registerDoParallel(myCluster)

#Ajuste dos modelos - risco proporcional (Cox PH)
 fit_ph <- ic_sp(cbind(left, right) ~ extrac, model = 'ph',B =
     c(0, 1), bs_samples = 500, data = dados,useMCores = TRUE)
```

```
 stopCluster(myCluster)
 proc.time() - ptm
 summary(fit_ph)
 newdata <- data.frame(extrac = c("manual","enzymatic") )
 newdata
 rownames(newdata) <- c("Manual","Enzymatic")
 getSCurves(fit_ph,newdata )

#Curvas com as probabilidades de germinação
 Color = c("blue","red")
 Lty=c(1,1)
 plot(fit_ph, fun="cdf", newdata,xlab = "Time(days)",
     xlim= c(0,30), xaxt= "n",yaxt="n", ylab = "Seed germination
     probability", col=Color, lwd=2, plot_legend= F)
 axis(1, at = c(1,8,15,22,30))
 axis(2, at = seq(0,1, by = 0.1))
 legend(21,0.35, lty= Lty,col=Color, lwd=2, c("Manual",
     "Enzimática"), bty="n",y.intersp = 1.3)

#Gráficos para diagnóstico
 Color1 = c("red","green")
 Lty=c(1,1)
 diag_covar(fit_ph, "extrac", col = Color1, xlab = "Time
             (days)", main = "", lgdLocation = NULL)
 legend(22,-1.2, lty= Lty, col= Color1, lwd= 2, c("Enzymatic",
          "Manual"), bty="n",y.intersp = 1.3)

## Análise para dados grupados

#Pacotes e Funcoes
 require(openxlsx)
 require(tidyverse)
 require(survival)
 require(ROCR)
 require(pROC)
#Para o ajuste do modelo de riscos proporcionais deve ser usado
o arquivo intervs obtido pelo SAS
#Leitura do banco de dados
 dados<-read.xlsx("INTERVS.xlsx")
 dim(dados)
 head(dados)
#Preparação dos dados
 dados$armaz<-factor(dados$armaz,levels=c("12","13","14"))
 dados$temper< factor(dados$temper, levels=
                    c("15","20","25","30"))
 attach(dados)
#Ajuste dos modelos
 fit2a<-glm(y~-1+int1+int2+int3+int4+factor(armaz, levels=
          14:12)+as.factor(temper),family=
          binomial(link="logit"))
 summary(fit2a)
```

```
#Ajuste do modelo sem a interação para efeito de comparação
 fit2b<-glm(y~-1+int1+int2+int3+int4+
       factor(armaz,levels=14:12)+as.factor(temper)+
       factor(armaz,levels=12:14)*as.factor(temper),
       family=binomial(link="logit"))
 summary(fit2b)
 anova(fit2a,fit2b,test="Chisq")

# CURVA ROC - fit2b
 roc2=plot.roc(y, fitted(fit2b))
 plot(smooth(roc2,method="density"), xlab="Especificidade",
      ylab="Sensibilidade",
      print.auc=TRUE,
      auc.polygon=TRUE,
      grud=c(0.1,0.2),
      grid.col=c("gray","red"),
      identity.lty=2,identity.lwd=2,identity.col="red",
      max.auc.polygon=TRUE,
      auc.polygon.col="lightgreen",
      print.thres=F)
# Selecionado o Modelo Logistico
 summary(fit2b)
 cf<-fit2b$coefficients[1:4];
 gi<-exp(cf);gi
 cf1<-fit2b$coefficients[5:18];cf1

#=================================
#Seleção do ARMAZ=12
#=================================
#Armazenamento de 12 meses com temperatura de 15°C
 qi1<-(1/(1+gi))
 SA12T15<-qi1
 for(i in 1:3){
   SA12T15[i+1]<-prod(qi1[1:(i+1)])}
 SA12T15<-c(1,SA12T15)
 FA12T15<-1-SA12T15;FA12T15
#Armazenamento de 12 meses com temperatura de 20°C
 qi2<-(1/(1+gi*exp(cf1[4])))
 SA12T20<-qi2
 for(i in 1:3){
   SA12T20[i+1]<-prod(qi2[1:(i+1)])}
 SA12T20<-c(1,SA12T20)
 FA12T20<-1-SA12T20;FA12T20
#Armazenamento de 12 meses com temperatura de 25°C
 qi3<-(1/(1+gi*exp(cf1[5])))
 SA12T25<-qi3
 for(i in 1:3){
   SA12T25[i+1]<-prod(qi3[1:(i+1)])}
 SA12T25<-c(1,SA12T25)
 FA12T25<-1-SA12T25;FA12T25
```

```
#Armazenamento de 12 meses com temperatura de 30°C
 qi4<-(1/(1+gi*exp(cf1[6])))
 SA12T30<-qi4
 for(i in 1:3){
   SA12T30[i+1]<-prod(qi4[1:(i+1)])}
 SA12T30<-c(1,SA12T30)
 FA12T30<-1-SA12T30;FA12T30
#Curvas com as probabilidades de germinação para o tempo de
armazenamento de 12 meses
#tiff("FigRegDiscreta.tiff", width=200,    height=180,
     units="mm", res=600)
 par(mar=c(4.0,4.0,1.5,1)+0.1)
 par(mfrow=c(2,2))
 t<-c(1,7,15,22,30)
 cbind(t,FA12T15,FA12T20,FA12T25,FA12T30)
 color = c("mediumorchid1","darkgreen","blue","red")
 plot(t,FA12T15, type="s", lty=1, ylim=range(c(0,1)),
     xaxt="n",xlab="Tempo (dias)", col="mediumorchid1", lwd=2,
     ylab="Probabilidade de germinação")
 axis(1,at=c(1,7,15,22,30))
 lines(t,FA12T20, type="s",lty=2,lwd=2,col="darkgreen")
 lines(t,FA12T25, type="s",lty=3,lwd=2,col="blue")
 lines(t,FA12T30, type="s",lty=3,lwd=2,col="red")
 mtext("(a)",side=3,adj=0.1,at=0.3,line=0.2, font=2, cex=1.0)

#================================
#Seleção do ARMAZ=13
#================================
#Armazenamento de 13 meses com temperatura de 15°C
 qi5<-(1/(1+gi*exp(cf1[2])))
 SA13T15<-qi5
 for(i in 1:3){
   SA13T15[i+1]<-prod(qi5[1:(i+1)])}
 SA13T15<-c(1,SA13T15)
 FA13T15<-1-SA13T15;FA13T15
#Armazenamento de 13 meses com temperatura de 20°C
 qi6<-(1/(1+gi*exp(cf1[2]+cf1[4]+cf1[9])))
 SA13T20<-qi6
 for(i in 1:3){
   SA13T20[i+1]<-prod(qi6[1:(i+1)])}
 SA13T20<-c(1,SA13T20)
 FA13T20<-1-SA13T20
#Armazenamento de 13 meses com temperatura de 25°C
 qi7<-(1/(1+gi*exp(cf1[2]+cf1[5]+cf1[10])))
 SA13T25<-qi7
 for(i in 1:3){
   SA13T25[i+1]<-prod(qi7[1:(i+1)])}
 SA13T25<-c(1,SA13T25)
 FA13T25<-1-SA13T25

#Armazenamento de 13 meses com temperatura de 30°C
```

```
 qi8<-(1/(1+gi*exp(cf1[2]+cf1[6]+cf1[11])))
 SA13T30<-qi8
 for(i in 1:3){
   SA13T30[i+1]<-prod(qi8[1:(i+1)])}
 SA13T30<-c(1,SA13T30)
 FA13T30<-1-SA13T30

#Curvas com as probabilidades de germinação para o tempo de
armazenamento de 13 meses
 cbind(t,FA13T15,FA13T20,FA13T25,FA13T30)
 color = c("mediumorchid1","darkgreen","blue","red")
 plot(t,FA13T15, type="s", lty=1, ylim= range(c(0,1)),
     xaxt="n",  xlab="Tempo  (dias)",  col=  "mediumorchid1",
     lwd=2,ylab= "Probabilidade de germinação")
 axis(1,at=c(1,7,15,22,30))
 lines(t,FA13T20, type="s",lty=2,lwd=2,col="darkgreen")
 lines(t,FA13T25, type="s",lty=3,lwd=2,col="blue")
 lines(t,FA13T30, type="s",lty=3,lwd=2,col="red")
 mtext("(b)",side=3,adj=0.1,at=0.3,line=0.2, font=2, cex=1.0)

#==================================
#Seleção do ARMAZ=14
#==================================
#Armazenamento de 14 meses com temperatura de 15°C
 qi9<-(1/(1+gi*exp(cf1[1])))
 SA14T15<-qi9
 for(i in 1:3){
   SA14T15[i+1]<-prod(qi9[1:(i+1)])}
 SA14T15<-c(1,SA14T15)
 FA14T15<-1-SA14T15
#Armazenamento de 14 meses com temperatura de 20°C
 qi10<-(1/(1+gi*exp(cf1[1]+cf1[4]+cf1[12])))
 SA14T20<-qi10
 for(i in 1:3){
   SA14T20[i+1]<-prod(qi10[1:(i+1)])}
 SA14T20<-c(1,SA14T20)
 FA14T20<-1-SA14T20
#Armazenamento de 14 meses com temperatura de 25°C
 qi11<-(1/(1+gi*exp(cf1[1]+cf1[5]+cf1[13])))
 SA14T25<-qi11
 for(i in 1:3){
   SA14T25[i+1]<-prod(qi11[1:(i+1)])}
 SA14T25<-c(1,SA14T25)
 FA14T25<-1-SA14T25
#Armazenamento de 14 meses com temperatura de 30°C
 qi12<-(1/(1+gi*exp(cf1[1]+cf1[6]+cf1[14])))
 SA14T30<-qi12
 for(i in 1:3){
   SA14T30[i+1]<-prod(qi12[1:(i+1)])}
 SA14T30<-c(1,SA14T30)
 FA14T30<-1-SA14T30
```

```
#Curvas com as probabilidades de germinação para o tempo de
armazenamento de 14 meses
 cbind(t,FA14T15,FA14T20,FA14T25,FA14T30)
 color = c("mediumorchid1","darkgreen","blue","red")
 plot(t,FA14T15, type="s", lty=1, ylim= range(c(0,1)), xaxt="n",
     xlab= "Tempo (dias)", col="mediumorchid1", lwd=2, ylab=
     "Probabilidade de germinação")
 axis(1,at=c(1,7,15,22,30))
 lines(t,FA14T20, type="s",lty=2,lwd=2,col="darkgreen")
 lines(t,FA14T25, type="s",lty=3,lwd=2,col="blue")
 lines(t,FA14T30, type="s",lty=3,lwd=2,col="red")
 mtext("(c)",side=3,adj=0.1,at=0.3,line=0.2, font=2, cex=1.0)

#Legenda
 Lty<-c(1,2,3,3)
 Color = c("mediumorchid1","darkgreen","blue","red")
 plot(1:15, 1:15, type = "n", ann = FALSE, axes = FALSE, bg =
     "green", xaxt = "n", yaxt = "n")
 legend("left", lty=Lty, lwd=2,bty="n",ncol=1,cex=1.0, col =
          Color,       x.intersp=1.1,y.intersp=1.5,seg.len=3.0,
          c("Temp. 15 °C", "Temp. 20 °C", "Temp. 25 °C", "Temp.
          30 °C"))
#dev.off()
```