Universidade de Brasília
Departamento de Estatística

Estudo da Evasão nos Cursos de Licenciatura em Matemática da Universidade de Brasília

Otávio Alves Cavalcante

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Otávio Alves Cavalcante

Estudo da Evasão nos Cursos de Licenciatura em Matemática da Universidade de Brasília

Orientadora: Profª. Drª. Juliana Betini Fachini Gomes

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Agradecimentos

Agradeço primeiramente aos meus pais, Mércia e Mário, meus pilares que sempre me apoiaram e me proporcionaram a melhor educação possível. Sem vocês, eu não teria chegado até aqui.

Agradeço à minha orientadora Juliana, por ter aceitado esse desafio comigo, sempre com um sorriso no rosto e disposta a ajudar, tirando dúvidas e me tranquilizando em momentos de ansiedade. Você é um exemplo de professora, quase uma mãe para os alunos.

Agradeço aos meus queridos amigos que ganhei de presente ao longo dessa jornada, em especial a Fabiana, a Jéssica, a Bruna, o Carlo e a Thais. Todos sabem que este não é um curso fácil, e nós sempre nos apoiamos uns nos outros e caminhamos juntos até aqui. Sozinho eu não teria conseguido.

A todos os professores do Departamento de Estatística, em especial ao George e à Maria Teresa, e à professora Cátia Regina do Departamento de Matemática, pela destreza e empenho na arte de ensinar e de proporcionar o nível mais alto de aprendizado aos seus alunos.

Por fim, agradeço a todos que direta ou indiretamente me apoiaram nessa trajetória. Foram anos com muitos altos e baixos, momentos de estresse, mas também de perseverança. Deixo também o meu apreço por essa universidade maravilhosa chamada UnB. Aqui sem dúvida eu vivi os melhores anos da minha vida, tive muitos ensinamentos, conquistei amizades e cresci como pessoa. A universidade não forma apenas estudantes, constrói também cidadãos, e sempre a defenderei com unhas e dentes.

Resumo

A evasão de estudantes em cursos de graduação traz perdas irreparáveis ao crescimento e desenvolvimento do país. Para solucionar essa questão, é preciso entender as causas que levam esses estudantes a evadir de seus cursos. O presente estudo tem como propósito identificar e analisar fatores que influenciam alunos de licenciatura diurna e noturna no âmbito do Departamento de Matemática da Universidade de Brasília a cometer evasão. A base de dados utilizada nas análises é composta por 335 alunos da licenciatura diurna e 371 alunos da licenciatura noturna, que ingressaram nos respectivos cursos entre os períodos do 1º semestre 2014 e do 2º semestre de 2019. A metodologia de análise de sobrevivência foi usada com o objetivo de avaliar o tempo de evasão dos alunos. Os modelos de regressão log-normal propostos identificaram a influência das covariáveis "Forma de ingresso", "Índice de Rendimento Acadêmico"(IRA) do aluno e "Cursou verão"para a licenciatura diurna e, na noturna, todas essas variáveis exceto Forma de ingresso. Nesse aspecto, o aproveitamento destes resultados torna-se útil ao subsidiar discussões acerca do assunto e incentivar futuras análises por parte de pesquisadores, gestores e membros da comunidade acadêmica.

Palavras-chave: Evasão; Falha; Censura; Licenciatura; Log-normal; Modelo de Regressão.

Abstract

The dropout of students in undergraduate courses brings irreparable losses to the country's growth and development. To resolve this issue, it is necessary to understand the causes that lead these students to drop out of their courses. The present study aims to identify and analyze factors that influence day and night undergraduate students within the Mathematics Department of the University of Brasília to drop out. The database used in the analyzes is composed of 335 daytime undergraduate students and 371 nighttime undergraduate students, who entered the respective courses between the periods of the 1^{st} semester 2014 and the 2^{nd} semester of 2019. The survival analysis methodology was used with the objective of evaluating the students' dropout time. The proposed log-normal regression models identified the influence of the covariates "Method of admission", "Academic performance Index"of the student and "Study summer"for daytime teaching and, for nighttime, all these variables except Form of admission. In this aspect, the use of these results becomes useful to support discussions on the subject and encourage future analyzes by researchers, managers and members of the academic community.

Keywords: School dropout; Failure; Censoring; Graduation; Log-normal; Regression Model.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

A problemática a ser abordada neste trabalho é de grande relevância em termos da educação no Brasil. A evasão de estudantes em cursos de graduação traz prejuízos imensuráveis ao crescimento e desenvolvimento do país. Nesse aspecto, a evasão, especialmente em cursos das áreas de setores como engenharia e ciências naturais, impacta diretamente na geração de inovações tecnológicas e de produtividade, deixando o Brasil em desvantagem em comparação aos outros países. (SACCARO; FRANÇA; JACINTO, 2019)

Para Filho et al. (2007), as perdas de estudantes que iniciam mas não concluem seus cursos incluem desperdícios sociais, acadêmicos e econômicos. No setor público, são recursos públicos investidos sem o devido retorno. No setor privado, é uma importante perda de receitas. Ainda segundo Filho et al. (2007), o setor privado investe entre 2% e 6% de suas receitas com marketing, visando atrair novos ingressantes no ensino superior. Em contrapartida, não se vê um esforço por parte das instituições privadas para mantê-los em seus cursos.

Por meio de testes qui-quadrado, Felder et al. (1993) avaliaram a relação de diversas variáveis e concluíram que alunos residentes em áreas urbanas e que participam de atividades extracurriculares por até 12 horas semanais têm menores chances de evadir. Ou seja, para as instituições de ensino, conhecer o perfil dos seus alunos e promover atividades que os aproximem do ambiente acadêmico é essencial para evitar esse problema.

Outro ponto a ser levado em consideração quando falamos em evasão nos cursos superiores é a base que estes ingressantes carregam do ensino médio. É razoável imaginar que uma base sólida garanta melhores condições destes estudantes permanecerem no curso. Nesse sentido, para Stinebrickner e Stinebrickner (2014), Saccaro, França e Jacinto (2019), além de possuir uma boa educação básica, os ingressantes precisam ter o entendimento acerca do nível de preparo exigido para aquele curso que estejam pleiteando. Muitas vezes o curso é escolhido sem ao menos uma pesquisa prévia a respeito. Isso ajudaria para que o aluno elevasse seu esforço nos primeiros semestres do curso, fato que contribuiria para a redução das taxas de evasão.

A questão da evasão em cursos superiores é um fato preocupante em nossa sociedade. São cidadãos que deixam de contribuir para crescimento do país por não terem a possibilidade de concluir um curso superior, ou que migram de um curso para o outro por falta de conhecimento prévio do que vão enfrentar, gerando perda de recursos e causando decepção nos discentes. Para solucionar essa questão, é preciso entender as causas que levam esses estudantes a cometer a evasão. Assim, a contribuição que estudos sobre o assunto trazem é fundamental para ao menos amenizar o problema.

Nessa questão, este trabalho pretende contribuir para o esclarecimento do problema, bem como incentivar o estudo do tema por parte dos acadêmicos e pesquisadores. Para isso, será estudado a evasão nos cursos de Licenciatura em Matemática da Universidade de Brasília por meio de técnicas de análise de sobrevivência.

2 Objetivos

2.1 Objetivo Geral

Este trabalho busca investigar os fatores que contribuem para a evasão nos cursos de Licenciatura em Matemática da Universidade de Brasília. Para isso, pretende-se utilizar dados fornecidos pela própria UnB a fim de servir como suporte para as análises a serem realizadas posteriormente.

2.2 Objetivos Específicos

- Comparar a evasão entre as Licenciaturas Diurna e Noturna.
- Identificar possíveis causas da evasão.
- Mensurar até que ponto fatores como idade, sexo, ser ou não ser cotista afetam as taxas de evasão em Matemática.

3 Revisão de Literatura

Este capítulo foi baseado no livro de Colosimo e Giolo (2006) e se dedica a introduzir os principais conceitos e técnicas que serão abordadas ao longo de todo o Trabalho de Conclusão de Curso.

3.1 Evasão Escolar

Antes de realizar qualquer tipo de estudo ou análise acerca do tema, é importante a definição do que vem a ser evasão. Segundo Santos e Albuquerque (2019), a evasão se caracteriza quando o estudante que estava matriculado no início do ano t deixa de estar matriculado no ano $t+1$. Em outras palavras, a evasão acontece quando o estudante conclui um período letivo (ano, semestre, bimestre), porém não retorna no período seguinte para dar prosseguimento aos estudos.

Por outro lado, na situação de abandono escolar, o estudante inicia o período letivo e no decorrer desse período o indivíduo deixa de comparecer ao ambiente de ensino, seja ele a escola ou a faculdade. Dessa forma, o aluno continua matriculado mas abandona os estudos. Comumente esses dois conceitos geram confusão na maioria dos casos, por isso é importante fazer essa distinção.

3.2 Conceitos Básicos em Análise de Sobrevivência

A Análise de Sobrevivência é um ramo da estatística que lida com dados em que a variável resposta é o tempo até a ocorrência do evento de interesse, denominado **tempo de falha**. Por isso, é fundamental definir três elementos que constituem o tempo de falha: o tempo inicial, a escala de medida e o evento de interesse (falha). No entanto, dados desta natureza poderiam ser tratados com outras técnicas estatísticas clássicas, como a análise de regressão linear por exemplo. O grande diferencial da Análise de Sobrevivência é a presença de **censura**, isto é, a perda de informação decorrente do fato de que, por algum motivo, não ser possível observar o evento de interesse, o que resulta em informações parciais ou incompletas. Entretanto, não se deve remover estas observações do estudo, uma vez que, ainda que incompletos, estes dados contém informações sobre o tempo de vida das observações. Além disso, a omissão de dados censurados acarretaria em estimativas viesadas.

3.2.1 Tempo de Falha

O termo **falha** surgiu em meio a estudos de análise de confiabilidade, nos quais o interesse é modelar o tempo até a ocorrência da falha em um equipamento.

Como já mencionado, o tempo de falha compreende três elementos. O primeiro deles, o tempo de início do estudo, deve ser claramente especificado para que os indivíduos possam ser comparados sob a mesma linha temporal. Em um estudo clínico aleatorizado, por exemplo, a origem natural do estudo seria a data da aleatorização.

A escala de medida a ser adotada é quase sempre o tempo real. Deve-se definir se esse tempo será medido em anos, meses, semestres, horas, etc.

Já o evento de interesse é, em sua maioria, um evento indesejado, tal qual a evasão. A importância de se definir o evento de interesse é evitar interpretações dúbias do que venha a ser uma falha. Em alguns casos a definição de falha já é óbvia por si só, como no caso de morte. Entretanto, em outras situações como no caso do presente estudo, existem diversas formas de se cometer evasão, seja por abandono do curso, seja por mudança de curso, por exemplo. Por isso estes eventos devem ser claramente definidos antes do estudo.

3.2.2 Censura

A presença de censura é inerente ao contexto da Análise de Sobrevivência. Dados censurados consistem em informações incompletas ou parciais decorrentes do fato de não se ter observado o evento de interesse antes do término do estudo. Nessa situação, temos três tipos de censura:

- Censura à Direita

  Nesse tipo de censura, o tempo de ocorrência do evento está à direita do tempo registrado. Podemos ainda considerar três subtipos de censura à direita:

  - Censura do Tipo I: Ocorre quando o estudo é finalizado após um período pré-estabelecido de tempo (t_f). Ao final de t_f, uma ou mais observações em estudo não falharam.
  - Censura do Tipo II: Ocorre quando o estudo é finalizado após a ocorrência do evento de interesse em um número k fixo de indivíduos. O número k de falhas deve ser estabelecido antes do início do estudo.
  - Censura Aleatória: Bastante comum na área médica, esse tipo de censura se caracteriza por conter os casos em que as observações não experimentam o evento

de interesse por motivos não controláveis. Pode ocorrer, por exemplo, em um estudo em que o paciente morre por causas distintas das causas estudadas.

A Figura 1 apresenta estes mecanismos de censura, em que "●"representa a falha e "○"representa a censura. Assim, no caso (a) temos dados em que não há a presença de censura, no caso (b) o estudo foi finalizado após um período fixado de tempo e alguns indivíduos não falharam, no caso (c) o estudo foi finalizado após a ocorrência de um número de falhas e no caso (d) alguns indivíduos foram retirados do estudo por motivos não controláveis.

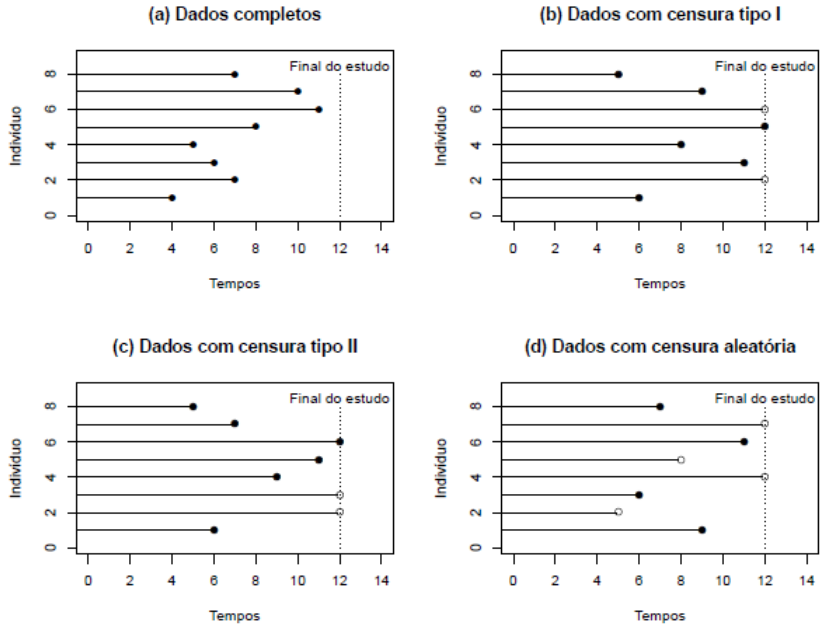


Figura 1: Ilustração de alguns mecanismos de censura. Fonte: Colosimo e Giolo (2006), adaptado por Santos (2017)

- Censura à Esquerda

  Ocorre quando o tempo registrado é maior que o tempo de falha. Em outras palavras, quando o evento de interesse já aconteceu no momento em que o indivíduo foi observado.

- Censura Intervalar

  A censura intervalar engloba os casos em que os estudos acompanham os indivíduos em visitas periódicas. Nesse caso, só é conhecido que o evento de interesse ocorreu em um certo intervalo de tempo $T \in (L, U]$. Vale ressaltar que censuras à direita e à esquerda são casos particulares de dados com censura intervalar, com $U = \infty$ para censuras à direita e $L = 0$ para censuras à esquerda (COOPER et al., 1998).

3.2.3 Representação dos Dados

Em Análise de Sobrevivência os dados da variável resposta associada a cada indivíduo $i(i = 1, ..., n)$ são representados pelo par (t_i, δ_i), sendo t_i o tempo de falha ou de censura e δ_i a variável indicadora de falha ou censura, ou seja,

$$\delta_i = \begin{cases} 1 & \text{se } t_i \text{ é tempo de falha} \\ 0 & \text{se } t_i \text{ é tempo de censura.} \end{cases} \tag{3.2.1}$$

Quando são consideradas variáveis explicativas no i-ésimo indivíduo do estudo, como por exemplo x_i =(sexo, idade, tratamento recebido), os dados são representados por $(t_i, \delta_i, \mathbf{x_i})$. No caso ainda de dados de sobrevivência intervalar, temos a representação $(l_i, u_i, \delta_i, \mathbf{x_i})$, em que l_i e u_i são, respectivamente, os limites inferior e superior do intervalo observado para o i-ésimo indivíduo (COLOSIMO; GIOLO, 2006).

3.3 Funções do Tempo de Sobrevivência

A distribuição do tempo de sobrevivência é geralmente caracterizada por três funções: (1) a função densidade de probabilidade, (2) a função de sobrevivência, e (3) a função de risco ou taxa de falha. Essas funções são matematicamente equivalentes, isto é, a partir de uma delas pode-se deduzir as demais.

Essas três funções servem de suporte para ilustrar o comportamento dos dados sob diferentes aspectos. Um dos interesses em Análise de Sobrevivência é estimar uma ou mais dessas funções a partir de uma amostra, e consequentemente realizar inferências a respeito do padrão de sobrevivência na população.

Seja T a variável aleatória não-negativa que representa o tempo de falha.

3.3.1 Função Densidade de Probabilidade

A função de probabilidade é definida como a probabilidade de um indivíduo experimentar o evento de interesse em um intervalo de tempo $[t, t + \Delta t)$ por unidade de comprimento do intervalo (Δt), ou simplesmente a probabilidade de falha em um curto intervalo por unidade de tempo. No caso contínuo, Lee e Wang (2003) definem $f(t)$ como sendo:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, \tag{3.3.1}$$

sendo $f(t) \geq 0$ para todo t e a área abaixo da curva de $f(t)$ igual a 1.

3.3.2 Função de Sobrevivência

A função de sobrevivência, denotada por $S(t)$, é definida como a probabilidade de um indivíduo não falhar até um certo tempo t, ou seja, a probabilidade de um indivíduo sobreviver ao tempo t:

$$S(t) = P(T \geq t) = \int_t^{\infty} f(x)\, dx. \tag{3.3.2}$$

A partir disso, da definição de distribuição acumulada, tem-se que a probabilidade de um indivíduo não sobreviver ao tempo t é $F(t) = 1 - S(t) = 1 - P(T \geq t)$.

Aqui $S(t)$ é uma função monótona, decrescente e geralmente contínua com a seguinte propriedade:

$$S(t) = \begin{cases} 1 & \text{para } t = 0 \\ 0 & \text{para } t = \infty, \end{cases}$$

ou seja, a probabilidade de sobreviver pelo menos ao tempo zero é 1 e a de sobreviver a um tempo infinito é zero (LEE; WANG, 2003). Dizemos que esse tipo de comportamento é típico da função de sobrevivência própria. Já em outras situações, a função de sobrevivência pode ter o seguinte comportamento:

$$\lim_{t \to 0} S(t) = 1 \quad \text{e} \quad \lim_{t \to \infty} S(t) = p,$$

em que p é uma probabilidade. Neste caso, a função de sobrevivência é considerada imprópria.

3.3.3 Função de Risco

As funções $f(t)$ e $S(t)$ fornecem duas formas matematicamente equivalentes de especificar a distribuição de uma variável aleatória contínua e não negativa. Uma outra função equivalente e de suma importância no contexto da Análise de Sobrevivência é a função de risco, ou taxa de falha, definida por (COX; OAKES, 1984):

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{3.3.3}$$

Colosimo e Giolo (2006) assumem que a taxa de falha no intervalo $[t_1, t_2)$ é definida

como a probabilidade de que a falha ocorra neste intervalo, dado que não ocorreu antes de t_1, dividida pelo comprimento do intervalo. Assim, a taxa de falha no intervalo $[t_,, t_2)$ é expressa por:

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}. \tag{3.3.4}$$

De modo geral, redefinindo o intervalo como $[t, t+\Delta t)$, a expressão (3.3.4) assume a seguinte forma

$$h(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}.$$

Assumindo Δt bem pequeno, $h(t)$ representa a taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t.

Podemos ainda obter a função de risco em termos da função de sobrevivência e da função de probabilidade, ou seja,

$$h(t) = \frac{f(t)}{S(t)}. \tag{3.3.5}$$

Colosimo e Giolo (2006) também avaliam que a modelagem da função de risco é um método importante pois pode conter forma crescente, decrescente, constante ou não monótona, sendo portanto uma função mais informativa se comparada à função de sobrevivência, no sentido de que diferentes funções de sobrevivência podem ter formas semelhantes.

Outra função bastante abordada no contexto de Análise de Sobrevivência é a função de risco acumulada. Com um nome bastante sugestivo, essa função fornece a taxa de falha acumulada do indivíduo, e pode ser definida como:

$$H(t) = \int_0^t h(u)\, du.$$

Embora não tenha uma interpretação direta, a função de risco acumulada pode ser útil na avaliação da função de risco, especialmente na estimação não paramétrica em que $h(t)$ é difícil de ser estimada e $H(t)$ apresenta um estimador com propriedades ótimas. Uma análise gráfica por meio da estimação dessa função também ajudará na escolha do melhor modelo que descreva os dados, fato que será abordado mais adiante.

3.3.4 Relações entre as Funções

As funções (3.3.1), (3.3.2) e (3.3.3) possuem relações matematicamente equivalentes entre si. Estas relações implicam na possibilidade de se obter todas as funções definidas anteriormente se tivermos em posse de apenas uma. Uma delas, definida em (3.3.5), ainda pode ser estendida da seguinte forma:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\left(\log S(t)\right).$$

Outras relações importantes incluem:

$$H(t) = -\log S(t),$$

$$S(t) = \exp\{-H(t)\} = \exp\left\{-\int_0^t h(u)\,du\right\}.$$

Definidas as principais funções e suas respectivas equivalências, o próximo passo é obter técnicas para a estimação dessas funções a partir de um banco de dados de sobrevivência.

3.4 Técnicas Não-Paramétricas

A estatística não-paramétrica se caracteriza por abranger técnicas que não dependem de dados com uma distribuição de probabilidade associada. A análise descritiva tradicional inclui encontrar medidas de tendência central e de variabilidade, como a média e a variância por exemplo. No entanto, uma vez que a presença de censura inviabiliza a estimação destes parâmetros, precisamos de técnicas não-paramétricas a fim de estimar a função de sobrevivência e, a partir dela, estimar as estatísticas de interesse, como o tempo médio e mediano. Essas técnicas formam a base da estatística descritiva no contexto da análise de sobrevivência.

3.4.1 Estimador de Kaplan-Meier

Proposto por Kaplan e Meier (1958), este é um estimador não viesado (para grandes amostras) de máxima verossimilhança não-paramétrico da função de sobrevivência $S(t)$. Na ausência de censura, o estimador de Kaplan-Meier é definido por:

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

em que $\hat{S}(t)$ é uma função do tipo escada com degraus nos tempos observados de falha de tamanho $1/n$, sendo n o tamanho da amostra. Caso hajam empates em um certo tempo t, o tamanho do degrau é multiplicado pelo número de empates (COLOSIMO; GIOLO, 2006). Para obter as estimativas da função de sobrevivência, deve-se ordenar de forma crescente os dados.

Contudo, na prática, o conjunto de dados amostrais em sua maioria apresenta dados censurados, o que requer o uso de técnicas mais refinadas para as análises futuras. A obtenção da estimativa de Kaplan-Meier requer um procedimento sucessivo em que cada passo depende do antecessor.

Assim, na presença de censura, o estimador de Kaplan-Meier pode ser generalizado da seguinte forma:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right), \tag{3.4.1}$$

desde que:

- $t_1 < t_2 < ... < t_k$ sejam os k tempos distintos e ordenados de falha,
- Considere tantos intervalos de tempo quantos forem o número de falhas distintas. Os limites dos intervalos de tempo são os tempos de falha da amostra.
- d_j seja o número de falhas em t_j, $j = 1, ..., k$, e
- n_j seja o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e que não foram censurados até o instante imediatamente anterior a t_j.

Breslow e Crowley (1974) provam, sob certas condições de regularidade, que $\hat{S}(t)$ é fracamente consistente e converge assintoticamente para um processo gaussiano. Logo, $\hat{S}(t) \sim N(S(t), Var(\hat{S}(t)))$. Já Kaplan e Meier (1958) demonstram que $\hat{S}(t)$ é um estimador de máxima verossimilhança de $S(t)$.

3.4.2 Curva do Tempo Total em Teste

Graficamente, a função de risco vista em (3.3.3) pode assumir formas distintas em sua representação. Nesse aspecto, torna-se necessário identificar o modelo mais adequado

para a variável T. Com isso, surgiu o gráfico do tempo total em teste, mais comumente chamado de gráfico TTT. Proposto por Aarset (1987), a construção do gráfico é obtida da seguinte forma:

$$G(r/n) = \frac{[(\sum_{i=1}^{r} T_{i:n}) + (n-r)T_{r:n}]}{(\sum_{i=1}^{n} T_i)},$$

por r/n, em que $r = 1, ..., n$ e $T_{i:n}, i = 1, ..., n$ são as estatísticas de ordem da amostra.

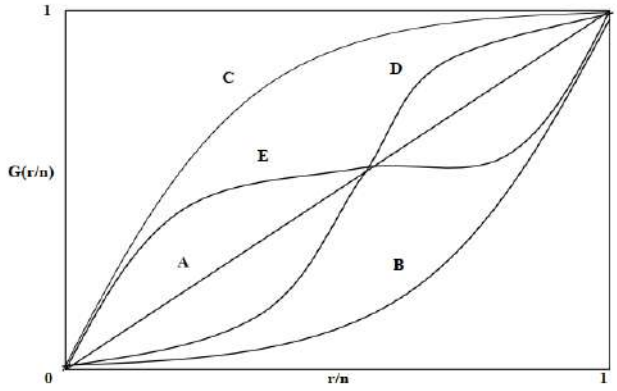


Figura 2: Ilustração de algumas formas da função de risco e curvas TTT.

A Figura 2 nos permite distinguir diferentes comportamentos da curva TTT, os quais são:

- Reta diagonal (A): A função de risco constante é apropriada.
- Curva convexa (B): A função de risco é monotonicamente decrescente.
- Curva côncava (C): A função de risco é monotonicamente crescente.
- Curva convexa e depois côncava (D): A função de risco tem forma de **U** (típica para o tempo de vida de pessoas).
- Curva côncava e depois convexa (E): A função de risco é unimodal.

3.4.3 Gráfico da Função de Risco Acumulada

O gráfico de $\hat{H}(t)$ surge como uma alternativa ao gráfico TTT, sendo mais indicado em casos que o número de censuras é grande. Podemos encontrá-lo por meio do estimador de Kaplan-Meier. Sua interpretação é o inverso da curva TTT, ou seja:

- Reta diagonal (não necessariamente a reta $y = x$) (A): A função de risco constante é apropriada.
- Curva convexa (B): A função de risco é monotonicamente crescente.
- Curva côncava (C): A função de risco é monotonicamente decrescente.
- Curva convexa e depois côncava (D): A função de risco é unimodal.
- Curva côncava e depois convexa (E): A função de risco tem forma de **U**.

3.4.4 Teste de *logRank*

O teste de *logRank* é o mais famoso teste para comparar curvas de sobrevivência. A estatística do teste é a diferença entre o número de falhas observado em cada nível da variável e uma quantidade que pode ser entendida como o valor esperado de falhas sob a hipótese nula, a de não existência de diferença entre as curvas de sobrevivência.

Considere o índice i variando entre 1 e r, em que $r > 2$. Desse modo, os dados podem ser arranjados em forma de uma tabela de contingência $2 \times r$ com d_{ij} falhas e $n_{ij} - d_{ij}$ sobreviventes na coluna i. A tabela de contingência passaria a ter r colunas em vez de simplesmente duas.

Condicional à experiência de falha e censura até o tempo t_j e o número de falhas no tempo t_j, a distribuição de $d_{2j}, ..., d_{rj}$ é uma distribuição hipergeométrica multivariada, isto é,

$$\frac{\prod_{i=1}^{r} \binom{n_{ij}}{d_{ij}}}{\binom{n_j}{d_j}}.$$

A média de d_{ij} é $w_{ij} = n_{ij} d_j n_j^{-1}$, bem como a variância de d_{ij} e a covariância de d_{ij} e d_{lj} são, respectivamente,

$$(V_j)_{ii} = n_{ij}(n_j - n_{ij}) d_{ij}(n_j - d_j) n_j^{-2}(n_j - 1)^{-1}$$

e

$$(V_j)_{il} = -n_{ij} n_{lj} d_j (n_j - d_j) n_j^{-2}(n_j - 1)^{-1}.$$

A estatística $v_j^{'} = (d_{2j} - w_{2j}, ..., d_{rj} - w_{rj})$ tem média zero e matriz de variância e covariância V_j de dimensão $r - 1$, com $(V_j)_{ii},\ i = 2, ..., r$ na diagonal principal e os elementos $(V_j)_{il},\ l = 2, ..., r$ fora da diagonal principal.

Dessa forma, pode-se formar a estatística v, somando sobre todos os tempo distintos de falha, isto é,

$$v = \sum_{j}^{k} v_j,$$

com v um vetor de dimensão $(r-1) \times 1$, cujos elementos são as diferenças entre os totais observados e esperados de falha.

Ao considerar a suposição de que as k tabelas de contingência são independentes, a variância da estatística v será $V = V_1 + V_2 + ... + V_k$. Um teste aproximado para a igualdade das r funções de sobrevivência pode ser baseado na estatística:

$$T = v^{'} V^{-1} v,$$

que sob H_0 tem distribuição χ^2 com $r-1$ graus de liberdade para amostras grandes. Os graus de liberdade são $r-1$ e não r porque os elementos de v somam zero. (COLOSIMO; GIOLO, 2006)

3.5 Modelos Probabilísticos

Modelos probabilísticos para o tempo de falha são distribuições de probabilidade que compõem a análise estatística de dados de sobrevivência. São modelos paramétricos que vêm sendo amplamente utilizados por possuírem uma boa adequação a diversas situações práticas. Algumas dessas principais distribuições são a Weibull e a Log-logística.

3.5.1 Distribuição de Weibull

A distribuição de Weibull (WEIBULL, 1954) ganhou popularidade por apresentar uma variedade de formas para a função de risco de acordo com o valor do parâmetro de forma γ. Todas elas com a propriedade de apresentar uma função de risco monótona, isto é, a função é decrescente se $\gamma < 1$, crescente se $\gamma > 1$ e constante se $\gamma = 1$.

- Densidade de Probabilidade

Seja T uma variável aleatória com distribuição de Weibull. Sua função de densidade de probabilidade é dada por:

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

em que γ e α são ambos positivos e α possui mesma unidade de medida de t. O parâmetro γ não tem unidade.

- Função de Sobrevivência

Já a função de sobrevivência da distribuição de Weibull é dada por:

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

para t, α e $\gamma \geq 0$.

- Função de Risco

Enquanto que a função de risco é representada por:

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

para t, α e $\gamma \geq 0$.

As funções da distribuição de Weibull assumem diferentes formas de acordo com o valor dos parâmetros γ e α, como mostra a Figura 3 abaixo:

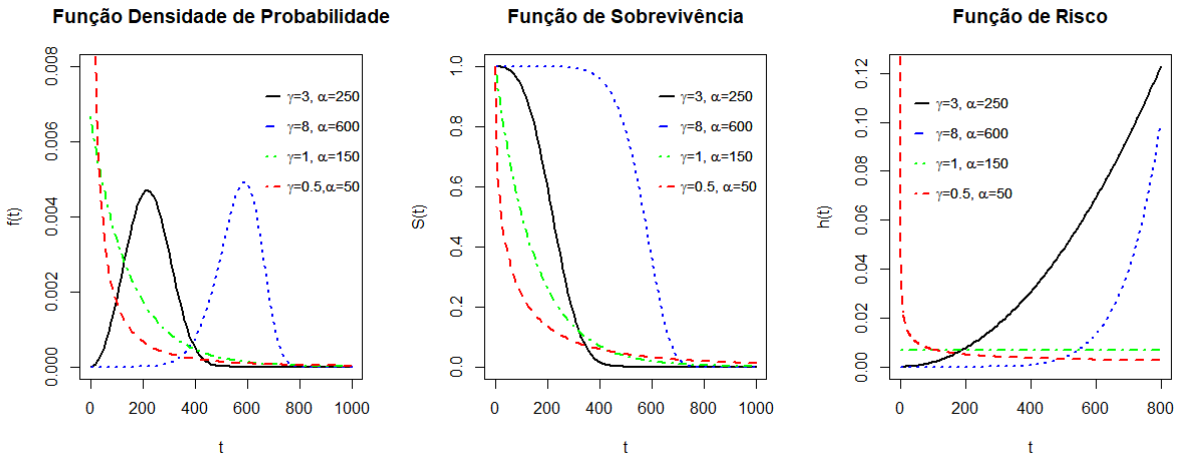


Figura 3: Ilustração da densidade de probabilidade, função de sobrevivência e de risco da distribuição Weibull para diferentes valores de γ e α.

A Figura 3 acima evidencia o fato da função de risco ser constante para $\gamma = 1$, estritamente crescente para $\gamma > 1$ e estritamente decrescente para $\gamma < 1$.

Os percentis são dados por:

$$t_p = \alpha \left[-\log(1-p)\right]^{1/\gamma}.$$

As expressões tanto para a média quanto para a variância da distribuição de Weibull incluem o uso da função gama, ou seja,

$$E[T] = \alpha\Gamma[1 + (1/\gamma)],$$

$$Var[T] = \alpha^2 \left[\Gamma[1 + (2/\gamma)] - \Gamma[1 + (1/\gamma)]^2\right],$$

sendo $\Gamma(r) = (r-1)!$ para r inteiro. (COLOSIMO; GIOLO, 2006)

3.5.2 Distribuição Log-logística

A vantagem dessa distribuição é a de apresentar uma forma analítica explícita para as funções de sobrevivência e de risco. Assim, se T possui distribuição Log-logística, suas funções são representadas por:

- Densidade de Probabilidade

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \left[1 + (t/\alpha)^\gamma\right]^{-2},$$

em que $t > 0$, $\alpha > 0$ é o parâmetro de escala e $\gamma > 0$ o parâmetro de forma.

- Função de Sobrevivência

Sua função de sobrevivência é expressa da seguinte forma:

$$S(t) = \frac{1}{1 + (t/\alpha)^\gamma}, \quad t > 0.$$

- Função de Risco

E a função de risco é dada por:

$$h(t) = \frac{\gamma(t/\alpha)^{\gamma-1}}{\alpha[1 + (t/\alpha)^\gamma]}, \quad t > 0.$$

A função de risco $h(t)$ também apresenta formas unimodais ($\gamma > 1$) e decrescente ($\gamma \leq 1$), conforme a Figura 4 abaixo:

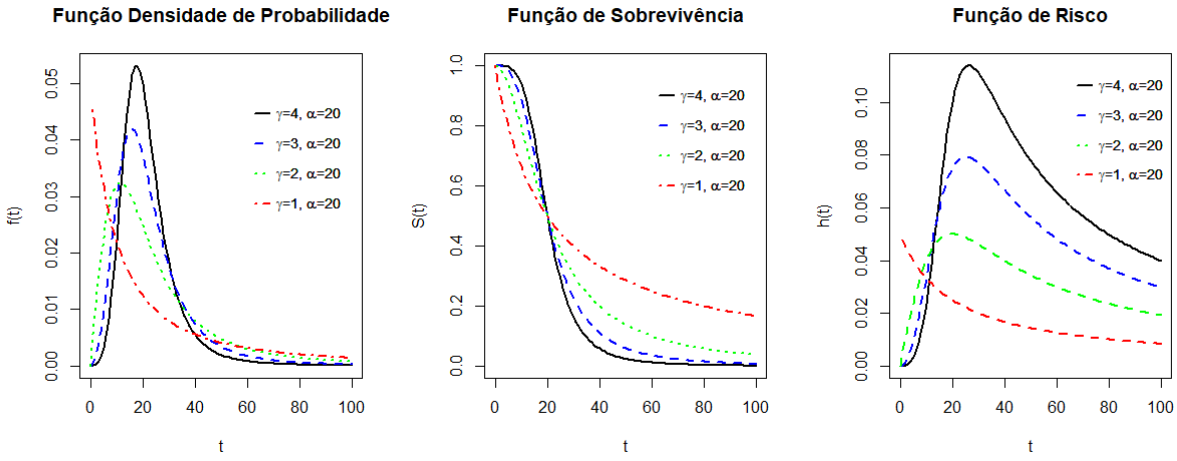


Figura 4: Ilustração da densidade de probabilidade, função de sobrevivência e de risco da distribuição Log-logística para $\alpha = 20$ e diferentes valores de γ.

Fica claro, pela Figura 4, que tanto a densidade de probabilidade quanto a função de risco apresentam um comportamento estritamente decrescente quando $\gamma = 1$. Já para $\gamma > 1$, essas duas funções são unimodais, ou seja, possuem comportamento crescente até determinado ponto, atingem o seu pico e depois apresentam uma forma decrescente.

Além disso, se T tem distribuição Log-logística com parâmetros α e γ, então a variável $T = \log(T)$ tem distribuição logística com parâmetros $-\infty < \mu < \infty$ e $\sigma > 0$, em que $\gamma = 1/\sigma$ e $\alpha = \exp(\mu)$.

3.5.3 Distribuição Log-normal

A distribuição log-normal, assim como a Weibull, é amplamente utilizada para descrever tempos de vida de produtos e indivíduos. Logo, se T possui distribuição log-normal, suas funções são representadas por:

- Densidade de Probabilidade

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{log(t) - \mu^2}{\sigma^2}\right)\right\}, \tag{3.5.1}$$

em que $t \geq 0$, μ é a média do logaritmo do tempo de falha e σ é o desvio-padrão.

- Função de Sobrevivência e Função de Risco

As funções de sobrevivência e de risco de uma variável log-normal não possuem forma analítica explícita. Sendo assim, são espressas respectivamente por:

$$S(t) = \Phi\left(\frac{-log(t) + \mu}{\sigma}\right) \quad \text{e} \quad h(t) = \frac{f(t)}{S(t)},$$

em que $\Phi(.)$ é a função de distribuição acumulada de uma normal padrão.

As funções taxa de falha não são monótonas como nas distribuições de Weibull. Elas primeiro crescem, atingem seu pico e depois decrescem, conforme mostrado na figrua abaixo.

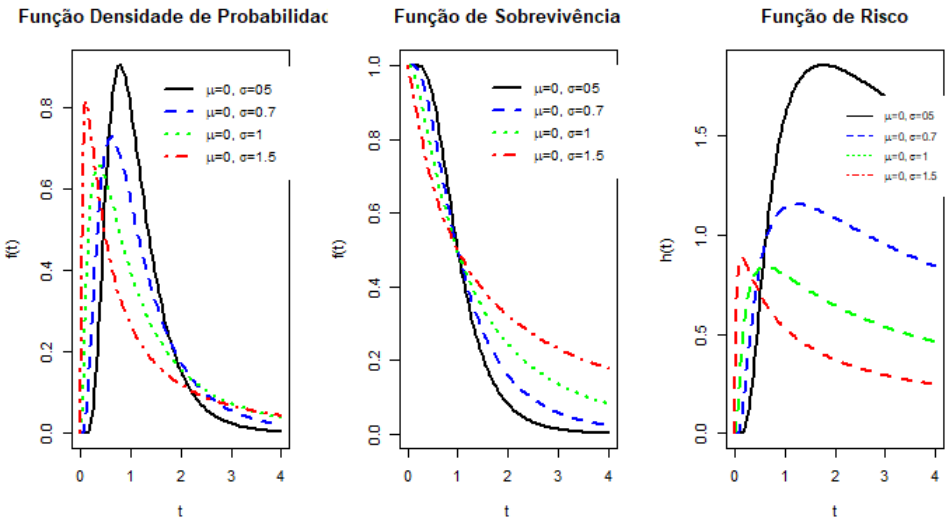


Figura 5: Ilustração da densidade de probabilidade, função de sobrevivência e de risco da distribuição log-normal para $\mu = 0$ e diferentes valores de σ

3.6 Estimação dos Parâmetros

As distribuições Weibull e Log-logística apresentadas anteriormente possuem ambas dois parâmetros (α e γ) que precisam ser estimados a partir das observações amostrais. Um método bastante difundido em cursos básicos de estatística é o método de mínimos quadrados. Contudo, essa técnica não é capaz de incorporar dados censurados no processo de estimação, por isso não seria apropriado utilizá-la em estudos de sobrevivência. Uma outro opção que surge para a estimação dos parâmetros é justamente o método de máxima verossimilhança, mencionado anteriormente para tratar do estimador de máxima verossimilhança $\hat{S}(t)$, apresentado em (3.4.1), para a função de sobrevivência $S(t)$.

3.6.1 Método de Máxima Verossimilhança

A ideia do método da máxima verossimilhança é encontrar, dentre todas as distribuições definidas pelas possíveis combinações de parâmetros, a distribuição que tenha maior possibilidade de ter gerado os dados daquela amostra.

Assim, considere uma amostra de observações $t_1, t_2, ..., t_n$ em que todas são não-censuradas. Seja ainda $f(t)$ a função densidade de probabilidade que caracteriza a população. Define-se a função de verossimilhança para um vetor de parâmetros $\boldsymbol{\theta}$ qualquer como:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i, \boldsymbol{\theta}). \tag{3.6.1}$$

Note que a função L nesse caso é uma função do parâmetro $\boldsymbol{\theta}$, e por isso é preciso evidenciar essa dependência.

O próximo passo é encontrar o valor de $\boldsymbol{\theta}$ que maximize a função $L(\boldsymbol{\theta})$. Porém, a função (3.6.1) nos fornece a indicação de que a contribuição de cada observação não-censurada é a sua função de densidade, quando na verdade essas observações somente nos informam que o tempo de falha é maior que o tempo de censurado observado e, portanto, que a sua contribuição para $L(\boldsymbol{\theta})$ é a sua função de sobrevivência $S(t)$. As observações podem então ser divididas em dois conjuntos, as r primeiras são as não censuradas $(1, 2, ..., r)$, e as $n-r$ restantes são as censuradas $(r+1, r+2, ..., n)$. (COLOSIMO; GIOLO, 2006)

Então, temos a seguinte função de verossimilhança:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{r} f(t_i, \boldsymbol{\theta}) \prod_{i=r+1}^{n} S(t_i; \boldsymbol{\theta}), \tag{3.6.2}$$

ou de forma equivalente,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [f(t_i, \boldsymbol{\theta})]^{\delta_i} [S(t_i; \boldsymbol{\theta})]^{1-\delta_i} = \prod_{i=1}^{n} \left[\underbrace{\frac{f(t_i, \boldsymbol{\theta})}{S(t_i; \boldsymbol{\theta})}}_{h(t_i; \boldsymbol{\theta})} \right]^{\delta_i} S(t_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} [h(t_i; \boldsymbol{\theta})]^{\delta_i} S(t_i; \boldsymbol{\theta}), \tag{3.6.3}$$

em que δ_i é a variável indicadora de falha ou censura definida em (3.2.1). As expressões (3.6.2) e (3.6.3) no entanto só são válidas para censurar do tipo I e II e supondo que o mecanismo de censura não carrega informações sobre os parâmetros. Também é válida para censura do tipo aleatória.

Como o próximo passo é justamente derivar a função de verossimilhança, por questões de otimização dos cálculos, é sempre mais conveniente trabalhar com a log-verossimilhança. Desse modo, aplica-se o log na expressão (3.6.3) e obtemos o seguinte:

$$l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \delta_i \log[f(t_i, \boldsymbol{\theta})] + (1 - \delta_i) \sum_{i=1}^{n} \log[S(t_i; \boldsymbol{\theta})].$$

Portanto, os estimadores de máxima verossimilhança são os valores de $\boldsymbol{\theta}$ que maximizam $L(\boldsymbol{\theta})$ ou equivalentemente $\log(L(\boldsymbol{\theta}))$, e são encontrados resolvendo-se o sistema de equações:

$$U(\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0.$$

A solução deste sistema para um conjunto de dados específico deve ser obtida por meio de métodos numéricos. O método de Newton-Raphson é um dos mais utilizados nesse sentido. Uma forma alternativa e até mais rápida para se obter a solução desse sistema é por meio de métodos computacionais. O *software* R possui um pacote específico para este tipo de problema, no qual pode-se encontrar a função *optim*, que será usada neste trabalho.

3.6.2 Intervalo de Confiança para os Parâmetros

Uma vez obtidas as estimativas pontuais, o método de máxima verossimilhança também permite obter estimativas intervalares, ou seja, a construção de intervalos de confiança para os parâmetros. Essas estimativas são possíveis graças a um conjunto de propriedades para grandes amostras destes estimadores. Cox e Hinkley (1974) provaram essas propriedades, sendo a principal delas a que diz respeito à precisão do estimador de máxima verossimilhança:

$$Var(\hat{\boldsymbol{\theta}}) \approx - \left[E(I_F(\boldsymbol{\theta})\right]^{-1},$$

ou seja, sob certas condições de regularidade, a matriz de variância e covariância é aproximadamente o negativo da inversa da informação de Fisher.

Em situações em que é difícil ou impossível de calcular a esperança, usa-se a matriz de informação observada $-[I_F(\boldsymbol{\theta}]^{-1}$, avaliada em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$

A estimativa para o erro padrão de $\hat{\boldsymbol{\theta}}$ é necessária para se obter os intervalos de confiança, isto é, para $\sqrt{Var(\hat{\boldsymbol{\theta}})}$. Quando θ é um escalar, um intervalo aproximado de $100\%(1 - \alpha)$ de confiança para θ é dado por:

$$\hat{\theta} \pm z_{\alpha/2} \sqrt{\widehat{Var}(\hat{\theta})}.$$

Já se $\boldsymbol{\theta}$ for um vetor de parâmetros, basta obter uma estimativa para o erro padrão a partir da matriz de variância e covariância $Var(\hat{\theta})$ e construir um intervalo de confiança para cada parâmetro separadamente.

3.7 Seleção de Modelos

Incluem métodos que permitem escolher um modelo que seja o mais parcimonioso em detrimento de outros. Há métodos para comparar modelos que são encaixados, como há métodos que comparam modelos de diferentes distribuições.

3.7.1 Teste da Razão de Verossimilhança

Os métodos gráficos são amplamente utilizados como critérios para a seleção de modelos. No entanto, trata-se de uma avaliação subjetiva que pode mudar entre diferentes analistas. Por isso, um teste de hipótese busca fornecer uma conclusão direta e objetiva, que não envolva o componente subjetivo na interpretação.

Dentre esses testes, destaca-se o Teste da Razão de Verossimilhança, ou simplesmente TRV. Este teste é utilizado apenas na comparação de modelos encaixados, isto é, quando existe um modelo principal tal que os modelos de interesse de comparação são casos particulares do modelo principal. Possui as seguintes hipóteses:

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

O teste considera dois ajustes: primeiro o modelo principal e a obtenção do logaritmo de sua função de verossimilhança ($logL(\hat{\theta}_G)$). Depois, o modelo de interesse e a obtenção do logaritmo da sua função de verossimilhança ($logL(\hat{\theta}_M)$). Desse modo, calcula-se a estatística do teste da seguinte maneira:

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

que, sob H_0, possui distribuição qui-quadrado com graus de liberdade igual a diferença do número de parâmetros dos modelos comparados.

3.8 Critérios de Informação

Outras medidas para classificar e selecionar modelos são os critérios de informação de Akaike e Bayesiano.

3.8.1 Critério de Akaike - AIC

Estimativa baseada no logaritmo da função de verossimilhança no ponto de máximo, acrescido de uma penalidade associada ao número de parâmetros. A função de penalidade tem a função de corrigir um viés proveniente da comparação de modelos com diferentes números de parâmetros. Sua medida é definida por:

$$AIC = -2\ log(L(\hat{\theta})) + 2p$$

,

em que p é número de parâmetros estimados no modelo. Anderson e Burnham (2004) recomendam o uso do AIC apenas quando $n/p \geq 40$.

O modelo escolhido deve apresentar menor valor de AIC dentre todos os modelos considerados para determinado problema.

3.8.2 Critério de Akaike Corrigido - AICc

Utilizado para pequenas amostras ($n/p < 40$), o AICc é dado por:

$$AICc = AIC + \frac{2p(p+1)}{n-p-1}$$

.

O modelo escolhido deve apresentar menor valor de AICc dentre todos os modelos considerados para determinado problema.

3.8.3 Critério de Informação Bayesiano - BIC

O critério de Informação Bayesiano (BIC) é definido por:

$$BIC = -2\ log(L(\hat{\theta})) + p \times log(n)$$

.

O modelo escolhido deve apresentar menor valor de BIC dentre todos os modelos considerados para determinado problema.

O critério BIC penaliza mais modelos com maior número de parâmetros do que o critério AIC. Tendendo dessa forma a selecionar modelos com um número menor de parâmetros.

3.9 Adequação do Modelo

Como há a presença de censura, os resíduos não seguem a distribuição normal e são assimétricos, portanto sendo necessários técnicas especias para analisar os resíduos.

3.9.1 Resíduos de Cox-Snell

Em análise de sobrevivência, os resíduos mais utilizados são os de Cox-Snell, sendo definidos por:

$$\hat{e}_i = \hat{H}(t_i|x_i),$$

em que $\hat{H}(.)$ é a função de risco acumulada obtida do modelo ajustado e x é o vetor de covariáveis.

Os resíduos $\hat{e}_i$ vêm de uma população homogênea e devem seguir uma distribuição exponencial padrão. Assim, o gráfico de $\hat{e}_i$ versus $\hat{H}(\hat{e}_i)$ deve ser aproximadamente uma reta.

4 Metodologia

4.1 Base de Dados

O banco de dados que servirá de base para este estudo foi obtido por meio da Secretaria de Tecnologia da Informação da UnB.

Como em 2020 houve uma migração do sistema SIGRA para o sistema SIGAA na UnB, foi disponibilizado três bancos de dados: o primeiro referente aos dados do SIGAA, onde encontram-se os alunos ativos nas duas licenciaturas (diurna e noturna) e também aqueles que saíram do curso a partir do 1º semestre de 2020, seja por ter se formado seja por outro motivo. Já os outros dois bancos compreendem os alunos do antigo sistema SIGRA, sendo um banco para a licenciatura diurna e o outro para a noturna. Vale destacar que as bases de ambos os sistemas não se misturam, ou seja, um mesmo aluno não pode estar na base do SIGAA e na base do SIGRA simultaneamente.

4.2 Variáveis

As três bases de dados possuem as mesmas 32 variáveis, a saber:

1. Sistema;
2. Aluno;
3. Id pessoa;
4. ira;
5. genero;
6. nascimento;
7. endereço;
8. cep;
9. estado nascimento;
10. sistema cotas;
11. cota;
12. raça;

13. Escola;
14. chamada ingressou UnB;
15. ano conclusão 2 grau;
16. curso;
17. período ingresso UnB;
18. período ingresso curso;
19. forma ingresso curso;
20. período saída curso;
21. forma saída curso;
22. período cursou disciplina;
23. modalidade disciplina;
24. media semestre aluno;
25. min cred para formatura;
26. créditos no período;
27. total créditos cursados aluno;
28. créditos aprovados no período;
29. código disciplina;
30. nome disciplina;
31. créditos disciplina;
32. menção disciplina;

Além disso, a base do SIGAA possui 26.277 linhas, enquanto que as bases do SIGRA para o diurno e para o noturno possuem 39.599 e 53.419 linhas, respectivamente. Salienta-se que essa quantidade não corresponde ao número de alunos, visto que cada linha corresponde na verdade a uma disciplina cursada pelo aluno. Por isso será necessário realizar uma filtragem, a fim de se obter uma base de dados onde cada linha seja única para cada aluno para que se possa realizar os procedimentos de modelagem mais adiante.

4.3 Filtragem e Junção das Bases

Antes de fazer a filtragem para se obter apenas uma linha por aluno, alguns outros filtros precisam ser feitos para deixar a base pronta para as análises.

Primeiramente realizou-se a junção das três bases de dados em uma só, a fim de facilitar o processo.

Em seguida foi preciso filtrar o período de ingresso no curso. A ideia é manter na base alunos sob o mesmo currículo ou com currículos muito semelhantes, a fim de não influenciar nos resultados. Como no 1º semestre de 2014 houve uma mudança substancial nos currículos da matemática, esse foi o período de corte escolhido para fazer o filtro, ou seja, apenas os ingressantes no curso a partir de 2014 permaneceram na base de dados.

Outro tema relevante nas discussões do trabalho foi a atipicidade do período de pandemia (do 1/2020 em diante). Optou-se então por desconsiderar os alunos que ingressaram e também saíram do curso entre 2020 e 2022. Considerando que nesse período a universidade flexibilizou critérios que influenciam diretamente na evasão, como a possibilidade de trancar disciplinas a qualquer momento do semestre e o não jubilamento de alunos nesse período, por exemplo. Além é claro de o calendário ter sido totalmente alterado, considerando que não teve aula entre março e agosto de 2020.

Por fim, pelas mesmas razões citadas no parágrafo anterior, optou-se também por desconsiderar as disciplinas cursadas pelos alunos ativos no curso durante o período de pandemia.

4.4 Criação de Variáveis

Algumas variáveis importantes no processo de análise e modelagem precisaram ser criadas a partir de variáveis já existentes na base de dados.

4.4.1 Total de Trancamentos

Essa variável foi criada com o intuito de contar quantas disciplinas o aluno trancou durante o período em que esteve ativo no curso. A partir da variável Menção disciplina, foi somado para cada aluno o número de menções TR e TJ, que correspondem a trancamento e trancamento justificado, respectivamente.

4.4.2 Taxa de Reprovação

A taxa de reprovação é uma variável quantitativa que varia de 0 a 1 e avalia a proporção de disciplinas reprovadas pelo aluno. Ela é construída também a partir da variável Menção disciplina, da seguinte maneira: o conjunto de menções {SR, II, MI} corresponde às disciplinas reprovadas pelo aluno, enquanto o conjunto {SR, II, MI, MM, MS, SS} corresponde às devidamente cursadas pelo aluno. Com isso, a variável Taxa de reprovação é construída dividindo o total de disciplinas reprovadas pelo total de disciplinas cursadas pelo aluno ao longo do curso.

4.4.3 Cursou Verão

Cursou verão é uma variável categórica binária que recebe valor 0 caso o aluno não tenham cursado nenhuma disciplina no verão, e 1 caso tenha cursado. Sua construção foi possível a partir da variável Período cursou disciplina.

4.4.4 Status

Essa variável é primordial em estudos de análise de sobrevivência pois vai indicar se o aluno sofreu um tempo de falha ou de censura, conforme mostrado na equação (3.2.1).

Sua construção foi feita a partir da variável Forma saída curso, sendo os níveis dessa variável classificados em falha ou censura de acordo com os critérios citados por Santos e Albuquerque (2019), em que será considerado o conceito de evasão quando o aluno é desvinculado da matrícula do curso de graduação por qualquer razão que seja. Assim, a Tabela 1 apresenta as formas de saída existentes no banco e suas respectivas classificações quanto à variável Status.

Tabela 1: Formas de saída e a variável Status

Forma de saída	Status
Ativo	Censura
Formatura	Censura
Desligamento - não cumpriu condição	Falha
Desligamento - Abandono	Falha
Desligamento Voluntário	Falha
Mudança de Curso	Falha
Mudança de Habilitação	Falha
Mudança de Turno	Falha
Novo Vestibular	Falha
Reprovou 3 vezes na mesma disciplina obrigatória	Falha

4.4.5 Tempo

A variável Tempo será medida em semestres, e indicará se o aluno sofreu um tempo de falha ou de censura. Sua construção foi feita a partir das variáveis Período saída curso e Período ingresso curso da seguinte maneira:

$$Tempo = (\text{periodo saida curso} - \text{periodo ingresso curso}) \times 2 + 1$$

No caso das censuras do tipo Ativo, considerou-se como período de saída o último período da variável Período cursou disciplina, isto é, o período mais recente no qual o aluno ativo cursou alguma disciplina.

Como as variáveis de período são dadas em anos e o tempo de falha é dado em semestres, é preciso multiplicar essa subtração por 2, já que um ano compreende dois semestres. Além disso, a interpretação do tempo de falha é a seguinte: Se um aluno ingressou no 1/2019 e evadiu no 2/2019, quer dizer que ele evadiu no segundo semestre cursado. Por isso é necessário acrescentar 1 a essa subtração. Isso evita também que tenhamos tempo de falha/censura igual a zero.

Nesse aspecto, esse cuidado em relação ao tempo zero se deve ao fato de que, por se tratar de dados discretos, o aluno que ingressou e evadiu no mesmo semestre poderia ser tratado como evasão no tempo 0 ou no tempo 1. Contudo, tratá-lo como zero implicaria dizer que esse aluno evadiu no momento exato 0, o que não ocorre. Na verdade, mesmo que esse aluno tenha evadido no mesmo semestre que ingressou, ele possui alguma convivência

com o ambiente acadêmico, fazendo mais sentido portanto caracterizá-lo como tempo 1.

4.4.6 Idade

Essa variável caracteriza a idade do aluno ao ingressar no curso. Foi construída a partir da data de nascimento e do período que ingressou no curso, sendo dada em anos completos.

4.5 Reclassificação de Variáveis

A variável Forma de ingresso compreende a forma como o aluno ingressou no curso. Nela, o ingresso pelo Vestibular corresponde a cerca de 37% do total, pelo Programa de Avaliação Seriada (PAS) corresponde a 30% e pelo Sistema de Seleção Unificada (Sisu) a 22%, ou seja, juntas essas três principais formas de ingresso somam quase 90% do total de alunos.

Além destas três, existem mais outras 9 categorias, como a transferência obrigatória, mudança de curso, dupla habilitação entre outras. Contudo, juntas elas somam cerca de 11% do total. Por isso decidiu-se agrupá-las em uma nova categoria denominada "outras formas de ingresso". Assim, a nova classificação da variável ficou da seguinte maneira:

Tabela 2: Reclassificação da variável Forma de ingresso

Forma de ingresso	Frequência relativa
Vestibular	37%
PAS	30%
Sisu	22%
Outras formas de ingresso	11%

4.6 Removendo Duplicidades

Após toda a rodada de filtragens e criação de variáveis na base de dados, chegou a hora de remover as linhas duplicadas e deixar apenas uma linha para cada aluno. Feito isso, a base de dados foi reduzida a 706 observações.

4.7 Divisão da Base de Dados

Como o interesse do estudo é modelar o tempo de falha das duas licenciaturas do curso de Matemática, o último passo no tratamento da base de dados é justamente dividir essa base em duas: uma base para a licenciatura diurna e a outra para a licenciatura noturna. Essa divisão visa facilitar as etapas seguintes de análise descritiva e modelagem dos dados. Assim, a base de dados da licenciatura diurna contém 335 observações, enquanto a base da licenciatura noturna possui 371 observações.

4.8 Análise dos Dados

A análise dos dados começará pela análise descritiva, por meio da qual se avaliará o comportamento das covariáveis, fazendo uso de métodos gráficos e tabelas de frequência. No caso das variáveis categóricas, gráficos de barras serão ferramentas importantes para este propósito, bem como o uso de histogramas e gráficos *boxplot* no caso de variáveis quantitativas. Além disso, gráficos de colunas justapostas auxiliam na compreensão do comportamento das variáveis de acordo com os níveis da variável Status (falha e censura).

Depois será feita uma análise descritiva utilizando recursos e técnicas de análise de sobrevivência, tais como gráficos com as curvas de sobrevivência, construídas a partir do estimador de Kaplan-Meier, e a função de risco acumulada para encontrar a melhor distribuição de probabilidade que modela os dados.

4.9 Modelagem

A parte da modelagem começará com a escolha da distribuição de probabilidade que irá modelar os dados em questão. Para isso, métodos gráficos e os critérios de informação definidos na subseção 3.8 serão usados na escolha dessa distribuição.

Em seguida se dará o processo de seleção de variáveis, que será conduzido seguindo o método derivado de Collett (1994).

4.10 Modelo de Regressão

Há muitos problemas em que é razoável esperar que uma variável resposta T possam ser explicada em termos de duas ou mais covariáveis. Sob a ótica da análise de sobrevivência, a variável resposta T é o tempo de sobrevivência do indivíduo, e o interesse é construir um modelo de regressão a partir dessas covariáveis.

Seja $\boldsymbol{x}^T = (1, x_1, ..., x_p)$ um vetor de covariáveis e g uma função de ligação. Dado um conjunto de p variáveis, o vetor de parâmetros θ é definido como:

$$\theta = g(\boldsymbol{x}^T \beta),$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ é o vetor de coeficientes de regressão.

Considere ainda que T seja uma variável aleatória com distribuição log-normal definida na equação (3.5.1).

Assim, tomando o parâmetro μ como $\mu = \boldsymbol{x}^T\boldsymbol{\beta}$, a função de ligação para a distribuição log-normal torna-se a função identidade $I(.)$. Logo, o modelo de regressão log-normal é definido por:

$$f(t|x) = \frac{1}{t\sigma\sqrt{2\pi}} exp\left\{\frac{-[log(t) - \boldsymbol{x}^T\boldsymbol{\beta}]^2}{2\sigma^2}\right\}.$$

Portanto a função de sobrevivência é dada por:

$$S(t) = \Phi\left(\frac{-log(t) + \boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right),$$

e a função de risco é dada por:

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

A estimação dos parâmetros do modelo de regressão log-normal seguirá o método da máxima verossimilhança descrito na subseção 3.6.1.

Para realizar todas as análises estatísticas pertinentes e calcular as estimativas para o modelo, será utilziado o *software* estatístico livre R.

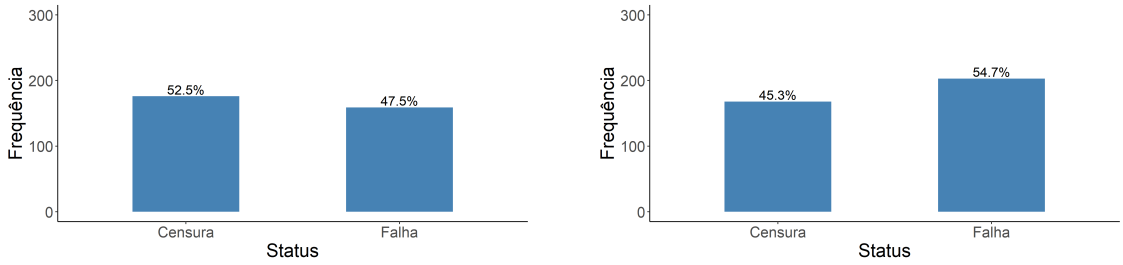
5 Resultados

5.1 Análise Descritiva

Antes de iniciar o processo de modelagem dos dados, é fundamental conhecê-los mais a fundo. Para tal, uma análise exploratória dos dados em questão nos permite enxergar padrões, detalhes, valores atípicos e até mesmo correlação entre variáveis, fatos que ajudarão no procedimento de modelagem mais adiante. Análises gráficas e testes de hipóteses serão utilizados para este fim.

5.1.1 Status

Esta é uma variável binária que indica se o i-ésimo aluno falhou ou foi censurado, conforme mostrado na seção 3.2. A Figura 6 abaixo apresenta os gráficos de barra, tanto para a licenciatura diurna quanto para a noturna.



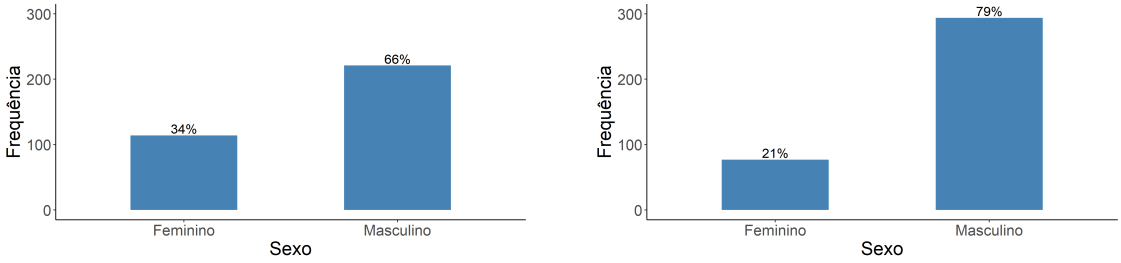
(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 6: Gráficos de barras para a variável Status

Apesar de apresentar um certo equilíbrio quanto ao número de falhas e censuras, o banco da licenciatura diurna possui mais censuras (52,5%) do que falhas (47,5%). Já para a licenciatura noturna essa situação se inverte, apresentando mais falhas (54,7%) do que censuras (45,3%).

5.1.2 Sexo

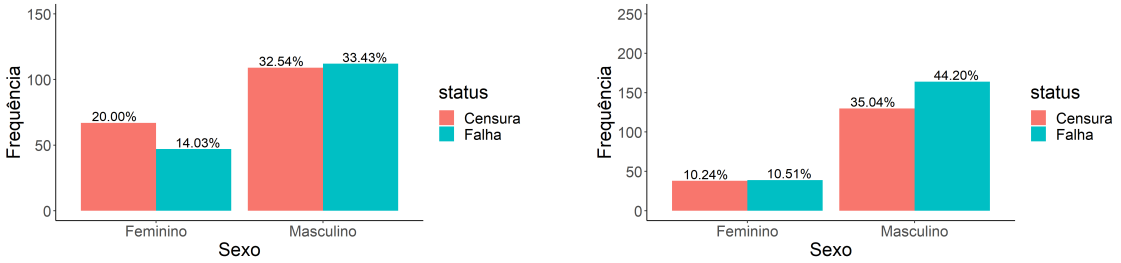
Esta variável, assim como o Status, é uma variável binária e indica o sexo do aluno.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 7: Gráficos de barras para a variável Sexo

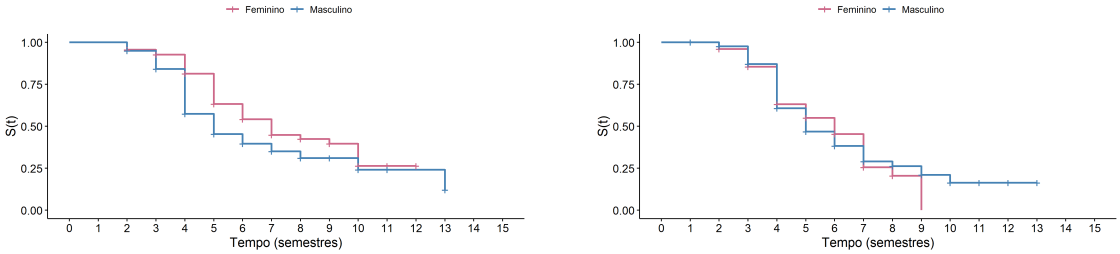
E pelos gráficos acima fica nítido que o curso de licenciatura em matemática, tanto diurno quanto noturno, é majoritariamente masculino. Ao observar a licenciatura noturna, essa diferença é ainda mais acentuada, sendo 79% dos alunos homens, contra apenas 21% de mulheres. Por ser um curso das ciências exatas, esse fato evidencia a carência de mulheres na ciência e o quanto ainda precisa ser feito para a igualdade de gêneros nessa área.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 8: Gráficos de barras para as variáveis Sexo vs Status

Já a Figura 8 mostra a distribuição das falhas e das censuras de acordo com o sexo. Nota-se que a dinâmica para o sexo feminino se altera quando analisamos a licenciatura diurna e notura. Na diurna, a maioria das mulheres são censura, ou seja, não experimentaram o evento de interesse (evasão), enquanto que na noturna essa quantidade é bastante equilibrada. Já para os homens, a maioria é composta de falhas, sendo essa diferença mais significativa no banco da licenciatura noturna.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 9: Gráficos das curvas de sobrevivência por Sexo

Examinando agora as curvas de sobrevivência, parece existir uma diferença na probabilidade de sobrevivência entre homens e mulheres na licenciatura diurna. Já na noturna, essa diferença não parece muito significativa, uma vez que as curvas para ambos os sexos estão muito próximas.

Um teste de *logRank* pode ser útil a fim de se obter uma constatação mais precisa acerca da diferença entre as curvas de sobrevivência das variáveis regressoras categóricas. As hipóteses do teste são as seguintes:

$$\begin{cases} H_0: & \text{Não existe diferença entre as curvas de sobrevivência} \\ H_1: & \text{Existe diferença entre as curvas de sobrevivência} \end{cases} \tag{5.1.1}$$

Os resultados do teste são apresentados na Tabela 3 abaixo.

Tabela 3: Teste de *logRank* para a variável Sexo

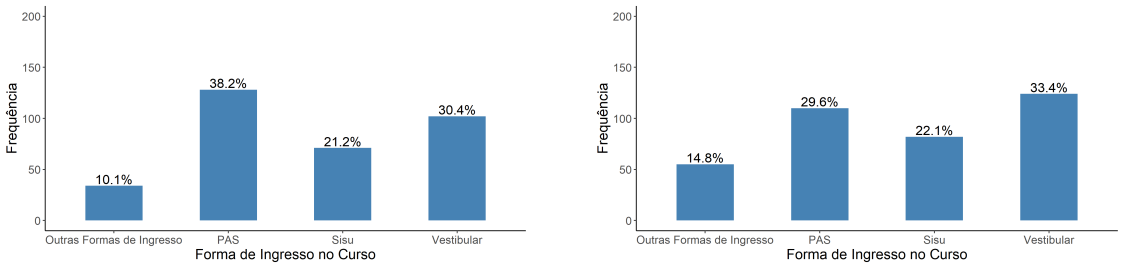
Licenciatura	Estatística do teste	g.l.	p-valor
Diurna	6,19	1	0,01
Noturna	0,01	1	0,90

Assim, considerando um vível de significância de 5%, temos evidências para rejeitar a hipótese nula no caso da licenciatura diurna, ou seja, a probabilidade de sobrevivência para o sexo feminino é superior ao sexo masculino, como notado na Figura 9 (a). No entanto, para a licenciatura noturna, a hipótese nula não é rejeitada, isto é, não há evidências estatísticas para dizer que as curvas de sobrevivência são diferentes.

5.1.3 Forma de Ingresso

A variável Forma de ingresso foi otimizada para que considerasse apenas 4 categorias, aglutinando todas as formas de ingresso que não fossem Vestibular, PAS ou Sisu

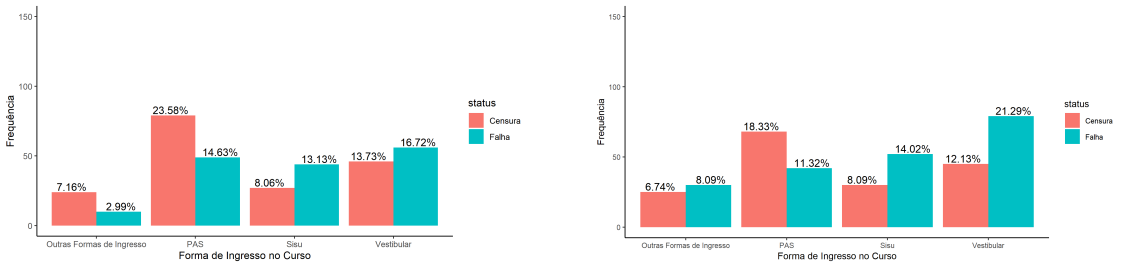
na categoria "Outras formas de ingresso".



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 10: Gráficos de barras para a variável Forma de ingresso

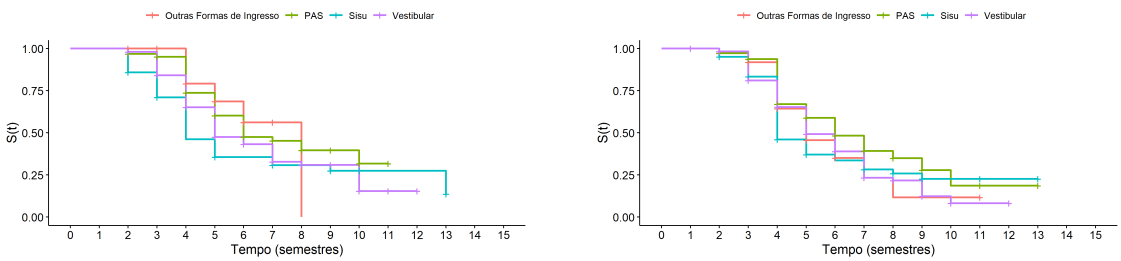
Com isso, a distribuição dessa variável possui o PAS com a maior frequência relativa de ingressantes na licenciatura diurna, enquanto na noturna essa posição é ocupada pelos ingressantes via vestibular. Outras formas de ingresso são minoria em ambas as licenciaturas.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 11: Gráficos de barras para as variáveis Forma de ingresso vs Status

Ao observar a distribuição das falhas e das censuras de acordo com a forma de ingresso, nota-se que a proporção de falhas é maior no caso do Sisu e do Vestibular em ambas as licenciaturas. Já para o PAS, a proporção de censuras é maior em ambos os bancos.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 12: Gráficos das curvas de sobrevivência a variável Forma de ingresso

No caso da licenciatura noturna, as curvas de sobrevivência acima mostram um comportamento semelhante até o 3º semestre. Já a partir do 4º semestre a probabilidade de sobrevivência para os alunos que ingressaram pelo PAS parece ser maior. Contudo, o comportamento geral das curvas não parece muito distinto e aparente seguir um mesmo padrão, mesmo que para algumas formas de ingresso a probabilidade de sobrevivência pareça ser menor.

Já no caso da licenciatura diurna o comportamento não parece ser semelhante. Há uma diferença grande entre as curvas do Sisu e do PAS, por exemplo, enquanto que para o vestibular, a probabilidade de sobrevivência parece ser maior em relação ao Sisu, e menor em relação ao PAS e a Outras formas de ingresso.

Para dar uma resposta mais precisa sobre as diferenças entre as curvas de sobrevivência, o teste de *logRank* novamente se faz presente, desta vez com as seguintes hipóteses:

$$\begin{cases} H_0: & \text{Não existe diferença entre as curvas de sobrevivência} \\ H_1: & \text{Existe ao menos uma diferença entre as curvas de sobrevivência} \end{cases} \tag{5.1.2}$$

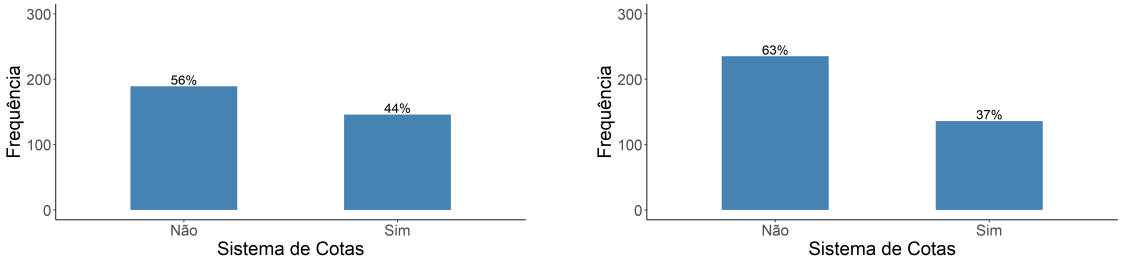
E pelos resultados da Tabela 4, rejeitamos a hipótese nula no caso da licenciatura diurna, e não rejeitamos no caso da noturna.

Tabela 4: Teste de *logRank* para a variável Forma de ingresso

Licenciatura	Estatística do teste	g.l.	p-valor
Diurna	2,404	3	0,006
Noturna	0,051	3	0,2

5.1.4 Sistema de Cotas

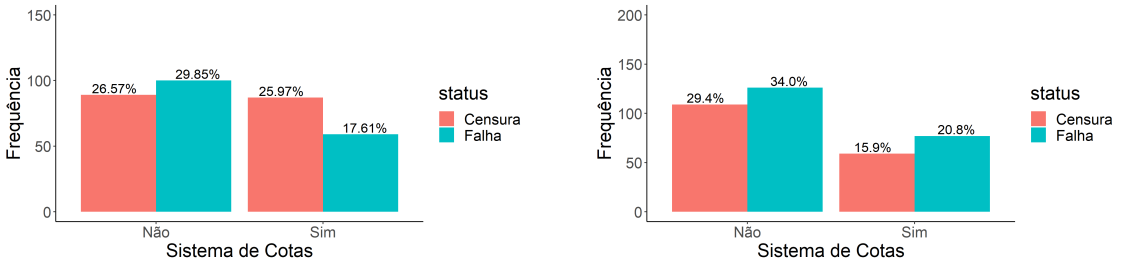
Essa variável determina se o aluno ingressou no curso por algum tipo de cota (escola pública, candidato negro e outras). Se o aluno ingressou por algum tipo de cota, é atribuído Sim, e se não ingressou por qualquer tipo de cota, recebe Não.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 13: Gráficos de barras para a variável Sistema de cotas

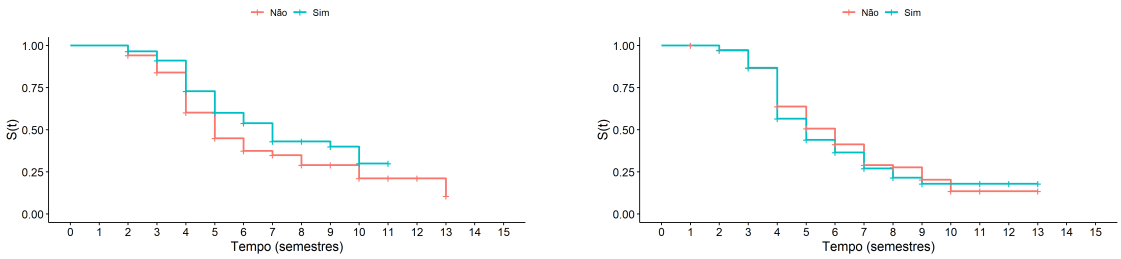
A Figura 13 mostra que os não cotistas são maioria em ambos os bancos, com uma ligeira diferença para mais no caso da licenciatura noturna, em que os cotistas representam apenas 37% dos alunos, enquanto os não cotistas representam 63%. No caso da licenciatura diurna os cotistas são 44% do total enquanto os não cotistas são 56%.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 14: Gráficos de barras para as variáveis Sistema de cotas vs Status

Observa-se que as falhas são maioria no caso dos não cotistas, tanto para o diurno quanto para o noturno. Já no caso dos cotistas, a distribuição é diferente em ambos os bancos. No diurno, as censuras são maioria, enquanto no noturno as falhas que são maioria.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 15: Gráficos das curvas de sobrevivência para a variável Sistema de cotas

Quanto às curvas de sobrevivência, os gráficos acima revelam uma diferença maior no diurno, indicando que os cotistas possuem uma probabilidade de sobrevivência maior do que os não cotistas. Já na licenciatura noturna essa diferença não parece significativa, já que as curvas são muito semelhantes.

Mais uma vez o teste de *logRank* será útil para fornecer uma resposta com maior exatidão. As hipóteses do teste são as mesmas fornecidas em (5.1.1).

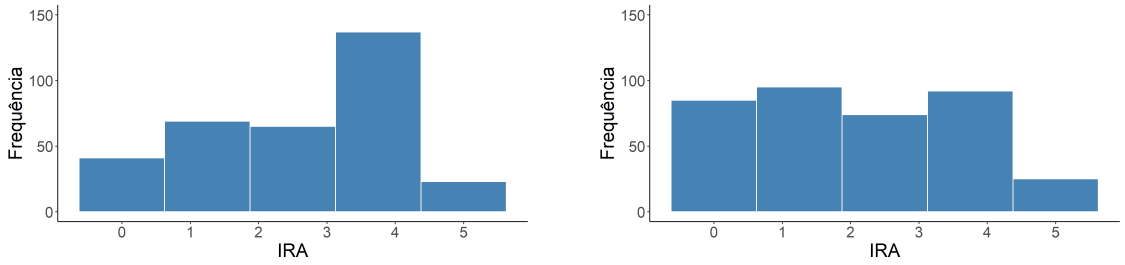
Tabela 5: Teste de *logRank* para a variável Sistema de cotas

Licenciatura	Estatística do teste	g.l.	p-valor
Diurna	5,86	1	0,02
Noturna	0,62	1	0,40

Pela Tabela 5, a hipótese nula é rejeitada a 5% de significância para a licenciatura diurna. Já na noturna a hipótese nula não é rejeitada. Ou seja, temos evidências estatísticas para dizer que as curvas de sobrevivência são diferentes na licenciatura diurna.

5.1.5 Índice de Rendimento Acadêmico (IRA)

O IRA é uma medida utilizada pela Universidade de Brasília que permite mensurar o rendimento acadêmico do aluno. Essa medida leva em conta uma série de fatores, como as menções dos alunos nas disciplinas, trancamentos, número de créditos cursados entre outros. Varia de 0 a 5, sendo 5 o nível máximo que um aluno pode alcançar.

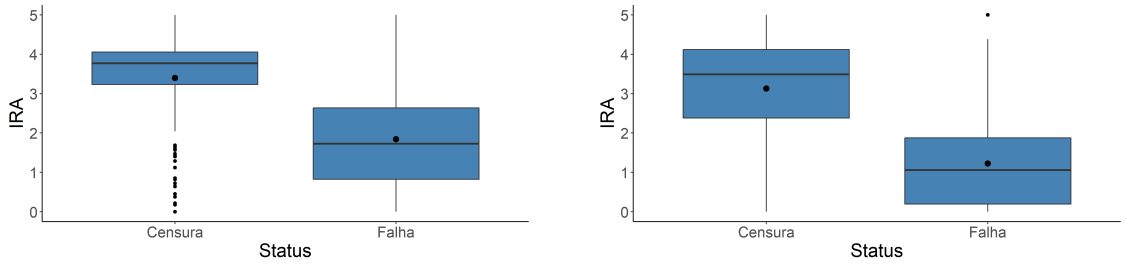


(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 16: Histogramas para a variável IRA

A Figura 16 acima apresenta os respectivos histogramas para as licenciaturas diurna e noturna. Para a diurna, há uma concentração maior de alunos com o IRA entre o 3 e o 4, enquanto aqueles com IRA superior a 4 são minoria. Já para a noturna a distribuição parece mais homogênea, com exceção da última categoria, onde os alunos com ira entre 4 e 5 também são minoria. Nota-se também uma proximidade entre os alunos com IRA entre 1 e 2 e aqueles entre 3 e 4. A quantidade de valores entre 0 e 1

também chama a atenção na licenciatura noturna.



(a) Licenciatura Diurna (b) Licenciatura Noturna

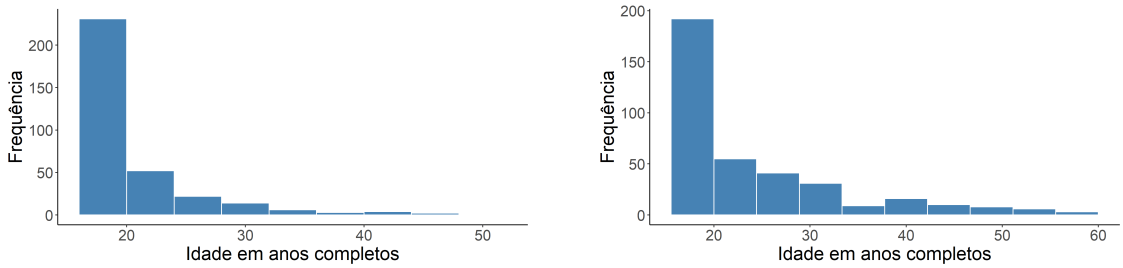
Figura 17: Gráficos *boxplot* para as variáveis IRA vs Status

O gráfico *boxplot* acima indica a relação do IRA com a variável Status. Nota-se primeiramente a presença de muitos valores atípicos na licenciatura diurna para o caso das censuras. Isso pode ser devido ao fato de que, pelo gráfico, a distribuição do IRA dos dados censurados parece ser mais homogênea, já que a caixa é mais achatada se comparada com as falhas. Com isso, o gráfico pode ser facilmente afetado por valores discrepantes, indicados pelas bolas pretas. Por outro lado, no caso das falhas a caixa é mais alongada, indicando que a variância do IRA nesse caso é maior. Nota-se também que a média está mais próxima da mediana no caso das falhas.

Já para o noturno, a média do ira é claramente superior para os tempos de censura em relação aos tempos de falha, assim como ocorre no diurno. Porém, no caso do noturno a presença de *outliers* se restringe a apenas um, justamente o valor máximo do ira, 5. As variâncias de ambos os tempos são próximas, já que as caixas possuem comprimentos semelhantes. A média do tempo de censura está mais distante da mediana em relação ao tempo de falha, indicando uma distribuição assimétrica nesse caso.

5.1.6 Idade

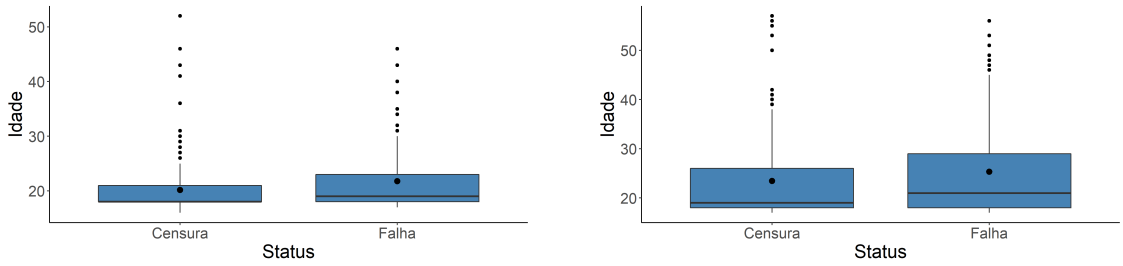
A variável idade também é uma das variáveis construídas a partir de outras existentes no banco, e expressa a idade do aluno ao ingressar no curso.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 18: Histogramas para a variável Idade

Os gráficos acima apresentam um comportamento bastante semelhante. Em geral, os alunos ingressam no curso antes do 20 anos, valendo tanto para o diurno quanto para o noturno. Contudo, no noturno há mais alunos mais velhos, com uma quantidade considerável de alunos em torno dos 40 anos ou mais. Isso pode ser reflexo do horário, já que pessoas mais velhas costumam trabalhar durante o dia.



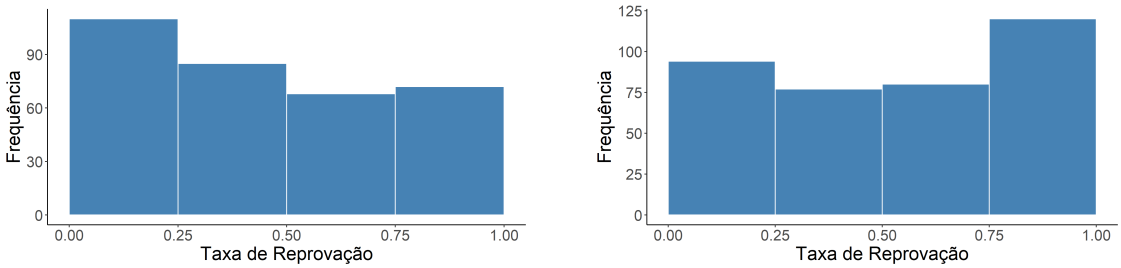
(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 19: Gráficos *boxplot* para as variáveis Idade vs Status

Se examinarmos a idade de acordo com a variável Status, a Figura 19 mostra que, no diurno, a idade mediana é a mesma do 1º quartil para o tempo de censura. Já no noturno, a dispersão das idades é maior em relação ao diurno. Além disso, 75% dos alunos ingressaram com idade até aproximadamente 30 anos. A forte presença de valores atípicos também se faz presente em ambas as licenciaturas, tanto para tempos de falha quanto de censura.

5.1.7 Taxa de Reprovação

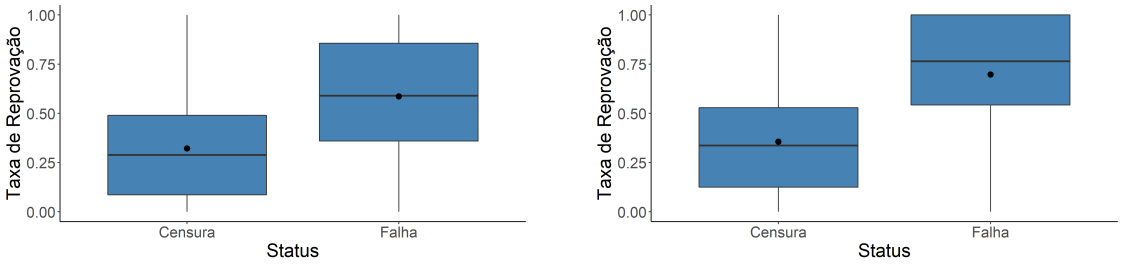
Essa variável varia de 0 a 1 e mensura a proporção de créditos, dentre todos os cursados, nos quais o aluno obteve uma menção que ocasiona reprovação (SR, II ou MI).



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 20: Histogramas para a variável Taxa de reprovação

Pelo gráfico acima, na licenciatura diurna a distribuição dos alunos se concentra em taxas de reprovação entre 0 e 0,25, enquanto na licenciatura noturna a concentração é maior nas taxas mais elevadas (entre 0,75 e 1). Isso chama a atenção pois aponta para uma proporção de reprovações bem maior no noturno se comparado ao diurno.



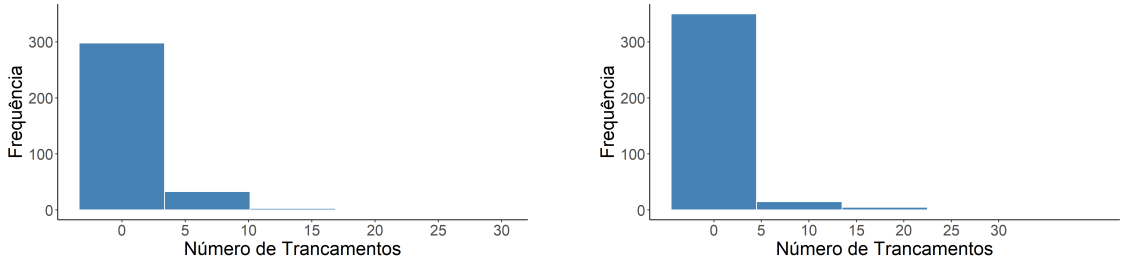
(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 21: Gráficos *boxplot* para as variáveis Taxa de reprovação vs Status

Quanto à taxa de reprovação de acordo com o Status, é nítido que para os tempos de falha as taxas de reprovação são maiores em relação aos tempos de censura, em ambos os bancos. No caso do noturno, essa diferença é ainda mais acentuada, visto que o 3º quartil do tempo de censura corresponde ao 1º quartil do tempo de falha, isto é, 75% dos alunos censurados possuem uma taxa até aproximadamente 0,50, enquanto 75% dos alunos com falha possuem uma taxa acima dessa proporção.

5.1.8 Total de Trancamentos

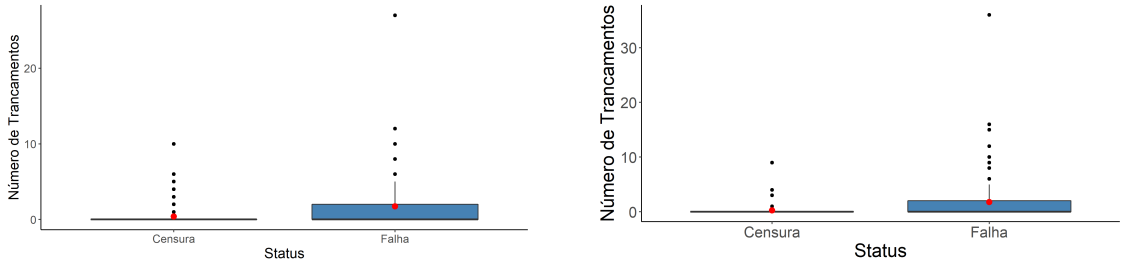
O Total de trancamentos contabiliza a quantidade de disciplinas trancadas pelo aluno no período estudado.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 22: Histogramas para a variável Total de trancamentos

Os gráficos acima são bastante semelhantes. Em geral, tanto na licenciatura diurna quanto na noturna, o número de trancamentos dos alunos se mantém próximo de zero.



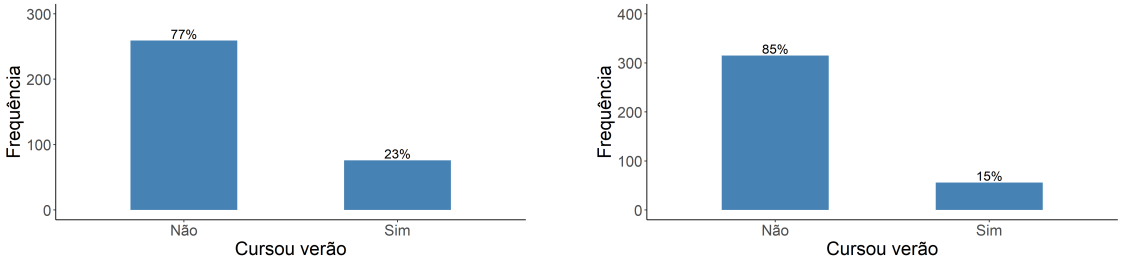
(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 23: Gráficos *boxplot* para as variáveis Total de trancamentos vs Status

Analisando agora a distribuição do total de trancamentos segundo a variável Status, chama a atenção o tempo de censura em ambas as licenciaturas, onde o gráfico *boxplot* é apenas um traço, ou seja, 1º quartil, mediana e 3º quartil são todos iguais a zero. Já no caso dos tempos de falha, o 1º quartil e a mediana coincidem, e a dispersão entre a mediana e o 3º quartil é maior.

5.1.9 Cursou Verão

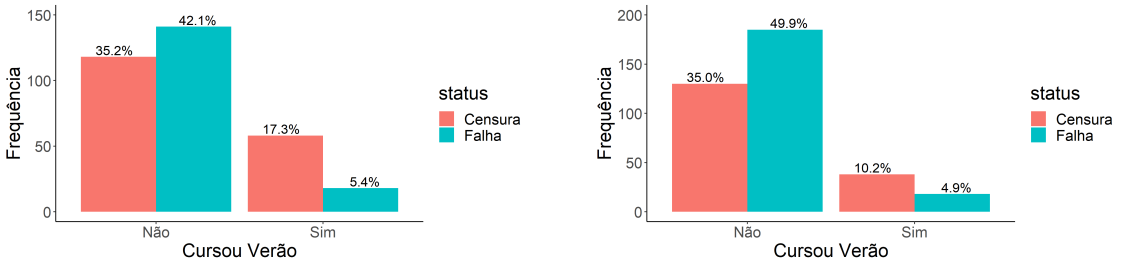
Esta é uma variável que classifica em "Sim"ou "Não"caso o aluno tenha cursado alguma disciplina dentre aquelas ofertadas no período do verão pela universidade.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 24: Gráficos de barras para a variável Cursou verão

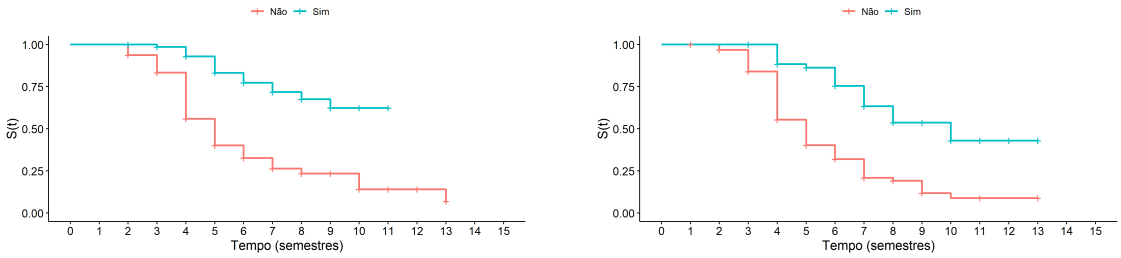
A diferença entre quem cursou verão e quem não cursou é considerável nas duas licenciaturas, como apresentado na Figura 24. Especialmente na noturna, onde 85% dos alunos não cursaram nenhuma disciplina no verão, contra 15% que cursaram ao menos uma disciplina nesse período.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 25: Gráficos de barras para as variáveis Cursou verão vs Status

Entre os tempos de falha, o comportamento da variável Cursou verão é o mesmo em ambos os bancos. O percentual de falhas é superior entre os que não cursaram verão, enquanto entre os que cursaram, a proporção de censuras é superior ao de falhas, com destaque para a licenciatura diurna, onde essa diferença chama mais a atenção.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 26: Gráficos das curvas de sobrevivência para a variável Cursou verão

As curvas de sobrevivência acima demonstram uma clara diferença entre quem cursou e quem não cursou verão. A probabilidade de sobrevivência dos que cursaram é muito superior aos que não cursaram.

No entanto, essa diferença deve ser significativa para afirmarmos que essa variável é determinante na probabilidade de sobreviver. Para isso, mais uma vez o teste de *logRank* se apresenta a fim de fornecer uma resposta mais precisa. As hipóteses do teste são as mesmas apresentadas na equação (5.1.1).

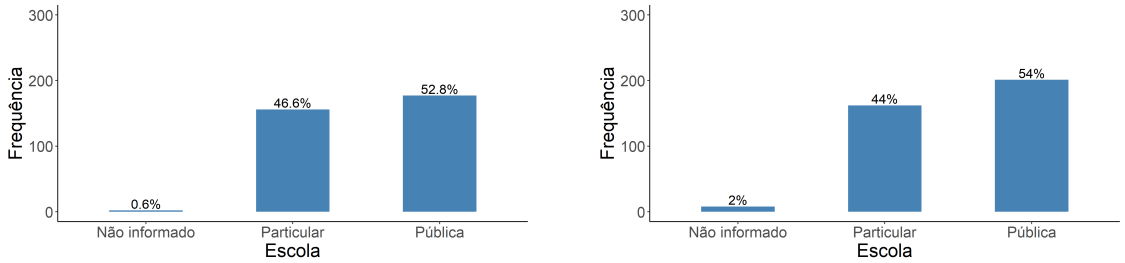
Tabela 6: Teste de *logRank* para a variável Cursou verão

Licenciatura	Estatística do teste	g.l.	p-valor
Diurna	41,1	1	1×10^{-10}
Noturna	33,2	1	9×10^{-9}

O teste rejeita a hipótese nula em ambas as licenciaturas, ou seja, temos evidências para dizer que as curvas de sobrevivência entre quem cursou e quem não cursou verão são diferentes.

5.1.10 Escola

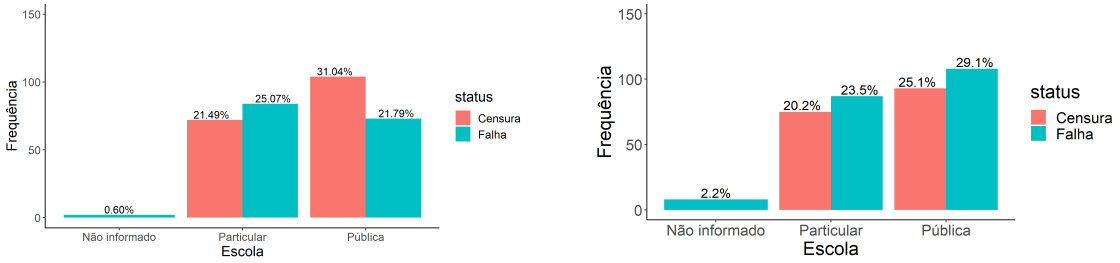
A variável Escola indica se o aluno cursou o ensino médio em escola pública ou particular. Importante lembrar que existem alguns casos em que não se sabe em que tipo de escola o aluno estudou, recebendo portanto a classificação "Não informado".



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 27: Gráficos de barras para a variável Escola

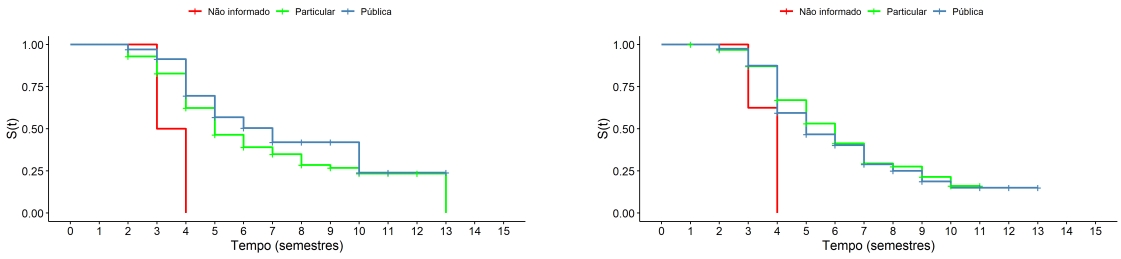
Pelos gráficos acima, a quantidade de alunos que não informaram a escola em que estudaram é irrisória no caso da licenciatura diurna. Na noturna, essa proporção é de 2%. Em geral, nos dois bancos a maioria dos alunos é proveniente de escolas públicas, representando cerca de 53% no diurno e 54% no noturno, valores bastante próximos.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 28: Gráficos de barras para as variáveis Escola vs Status

A Figura 28 mostra que, entre os que não informaram, todos falharam. No diurno, a proporção de alunos que falharam é maior para aqueles que estudaram em escola particular, enquanto os que estudaram em escola pública o percentual de censura é maior. No noturno a proporção de falhas é maior, tanto para provenientes de escola pública quanto de privada.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 29: Gráficos das curvas de sobrevivência para a variável Escola

No caso das curvas de sobrevivência, não parecem existir diferenças significativas entre a probabilidade de sobrevivência dos provenientes de escola pública ou particular na licenciatura noturna, já que as curvas em muitos momentos quase que se sobrepõem . Já na diurna, essa diferença é mais perceptível. Nesse sentido, como os casos dos alunos que não informaram a escola que estudaram são muito atípicos e não correspondem a uma parcela significativa dos dados, esses indivíduos precisarão ser removidos a fim de não influenciar em testes de hipóteses e no processo de modelagem, caso essa variável entre no modelo. No caso do teste de *logRank* abaixo, esses dados foram removidos, uma vez que o interesse aqui é avaliar se há diferença nas probabilidades de sobrevivência de quem estudou em escola pública ou particular. As hipóteses do teste são as mesmas da equação (5.1.1).

Tabela 7: Teste de *logRank* para a variável Escola

Licenciatura	Estatística do teste	g.l.	p-valor
Diurna	4,51	1	0,03
Noturna	0,37	1	0,50

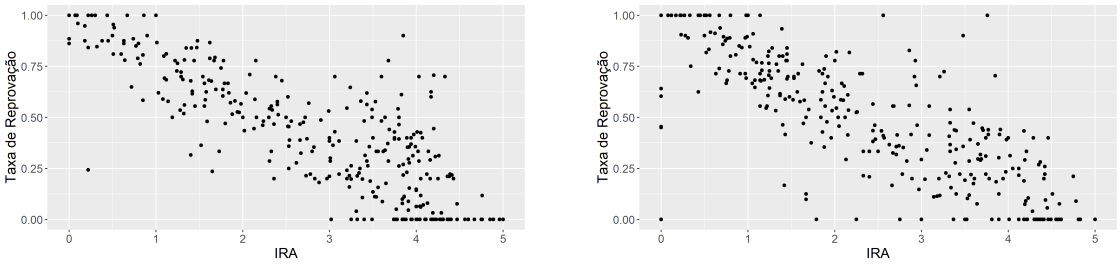
O teste acima rejeita a hipótese nula para a licenciatura diurna, e não rejeita para licenciatura noturna, corroborando a suspeita inicial vista na Figura 29.

5.2 Correlação entre as Variáveis

Algumas variáveis possuem íntima relação uma com a outra, podendo ocasionar em forte associação, no caso das variáveis qualitativas, ou forte correlação, no caso das variáveis quantitativas. Esta seção se dedica a examinar com mais profundidade essas associações/correlações, com a finalidade de evitar problemas de multicolinearidade, que ocorre quando o modelo inclui variáveis preditoras correlacionadas.

5.2.1 Taxa de Reprovação e IRA

O Índice de Rendimento Acadêmico de um aluno é calculado com base em diversos fatores. Entre eles, as menções obtidas nas disciplinas, que impactam diretamente no desempenho do aluno. Assim, quanto mais disciplinas reprovadas, espera-se que o IRA diminua consideravelmente, e vice-versa. Dessa maneira, a taxa de reprovação é intimamente ligado ao IRA, uma vez que considera em seu cálculo a quantidade de disciplinas reprovadas. Além disso, ambas as medidas são formas de mensurar o desempenho acadêmico do estudante.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 30: Gráficos de dispersão entre as variáveis IRA e Taxa de reprovação

A Figura 30 esclarece essa relação. Nota-se uma forte correlação linear nega-

tiva entre IRA e taxa de reprovação, ou seja, à medida que o IRA aumenta, a taxa de reprovação diminui.

Tabela 8: Coeficiente de Correlação de Pearson entre as variáveis IRA e Taxa de reprovação

Licenciatura	ρ
Diurna	-0,8497
Noturna	-0,8033

Tal fato se confirma por meio do coeficiente de correlação de Pearson, muito utilizado para medir a correlação entre variáveis quantitativas contínuas. O coeficiente apontou uma correlação de cerca de $-0,85$ para a licenciatura diurna, e de $-0,80$ para a noturna, corroborando a análise gráfica apresentada na Figura 30.

5.2.2 Sistema de Cotas e Escola

Há uma desconfiança de que essas duas variáveis sejam associadas devido ao fato de que um dos critérios para ingressar como cotista seja justamente ter estudado em escola pública no ensino médio. Desse modo, como são ambas variáveis qualitativas, um teste χ^2 de independência pode ser útil para medir o grau de associação.

Tabela 9: Tabela de contingência entre Escola e Sistema de cotas para a Licenciatura Diurna

	Escola	
Cotas	Particular	Pública
Não	139	49
Sim	17	128

Tabela 10: Tabela de contingência entre Escola e Sistema de cotas para a Licenciatura Noturna

	Escola	
Cotas	Particular	Pública
Não	145	84
Sim	17	117

Acima são apresentadas as tabelas de contingência para ambas as variáveis. O número de alunos cotistas e que estudaram em escola pública é muito maior em relação aos que estudaram em escola particular.

Para o teste χ de independência, as hipóteses são as seguintes:

$$\begin{cases} H_0: & \text{As variáveis são independentes} \\ H_1: & \text{As variáveis não são independentes} \end{cases}$$

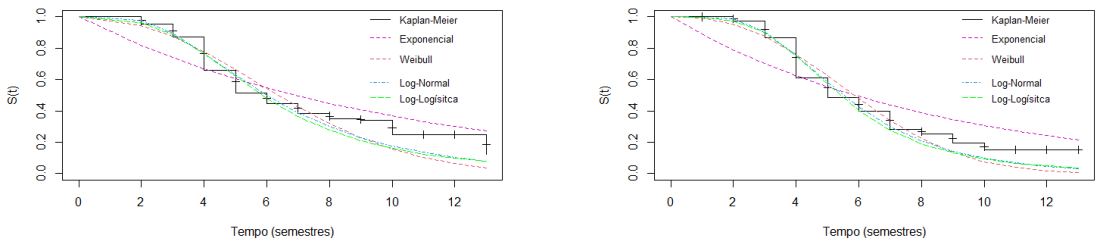
Tabela 11: Teste χ^2 de independência entre as variáveis Escola e Sistema de cotas

Licenciatura	Estatística do teste	g.l.	p-valor
Diurna	124,75	1	1×10^{-16}
Noturna	85,66	1	9×10^{-16}

O teste rejeita a hipótese de independência para ambas as licenciaturas. Logo, temos evidências para acreditar que Sistema de cotas e Escola são variáveis associadas.

5.3 Seleção da Distribuição de Probabilidade

O passo fundamental antes de qualquer processo de modelagem é justamente encontrar a distribuição de probabilidade que melhor se ajuste ao banco de dados em questão. Logo, a Figura 31 apresenta as curvas de sobrevivência de Kaplan-Meier para as duas licenciaturas, juntamente com o ajuste das distribuições mais utilizadas em dados de sobrevivência com intuito de compará-las.



(a) Licenciatura Diurna (b) Licenciatura Noturna

Figura 31: Comparação entre as principais distribuições

A distribuição Exponencial é prontamente descartada em ambos os bancos, visto que apresentou o pior desempenho em termos de ajuste aos dados. Por outro lado, tanto a log-normal quanto a log-logística obtiveram desempenho melhor se comparadas à Weibull, especialmente nas caudas das curvas. Dessa maneira, pela análise gráfica, a Exponencial e a Weibull são logo desconsideradas, restando uma dúvida em relação à log-normal e à log-logística. Em casos assim, critérios de parcimônia como as medidas de informação de Akaike e Bayesiano costumam ser utilizados para a escolha entre distribuições ou modelos candidatos. Contudo, vale ressaltar que, segundo Klein e Moeschberger (2003), esses

critérios não provam que um modelo particular é o correto, eles apenas tentam rejeitar modelos que são claramente inapropriados.

Tabela 12: Critérios de informação para a licenciatura diurna

Distribuição	AIC	AIC_c	BIC
Exponencial	1051,42	1051,43	1055,23
Weibull	936,18	936,21	943,80
Log-logística	905,05	905,08	912,68
Log-normal	897,21	897,25	904,84

Tabela 13: Critérios de informação para a licenciatura noturna

Distribuição	AIC	AIC_c	BIC
Exponencial	1276,04	1276,05	1279,95
Weibull	1082,90	1082,94	1090,73
Log-logística	1036,43	1036,46	1044,26
Log-normal	1029,69	1029,72	1037,52

Em termos das medidas AIC, AIC_c e BIC, a log-normal foi a distribuição com os menores valores, obtendo assim os melhores resultados tanto na licenciatura diurna quanto na noturna.

Portanto, levando em conta a análise gráfica e as medidas de informação listadas nas Tabelas 12 e 13, a log-normal será a distribuição utilizada em ambas as licenciaturas no processo de modelagem.

5.4 Modelagem para a Licenciatura Diurna

Nesta subseção será apresentado todo o processo de modelagem dos dados da licenciatura diurna em matemática.

5.4.1 Seleção de Variáveis

Nove covariáveis foram selecionadas do banco de dados como potencialmente relevantes a serem incluídas no modelo para descrever o comportamento da variável resposta.

Em vez de fazer uso de métodos de seleção automáticos, que muitas vezes não levam em consideração covariáveis significativas do ponto de vista educacional, a abordagem para o processo de seleção de variáveis se dará de forma diferente. O processo se baseia no método derivado da proposta de Collett (1994) e descrito por Colosimo e Giolo (2006), segundo o qual o estatístico, juntamente com o pesquisador da respectiva

área, adotem uma postura pro-ativa no processo de seleção. Nessa perspectiva, o processo apresenta os seguintes passos:

1. Ajustar todos os modelos com as covariáveis isoladamente. Selecionar aquelas significativas ao nível de 10% de significância.

2. Ajustar conjuntamente um modelo com todas as covariáveis significativas no passo 1. Em seguida, excluir do modelo as covariáveis que, conjuntamente, não são significativas, uma de cada vez, a fim de constatar a significância no modelo.

3. Retirar as covariáveis que restaram no passo 2, uma a uma, a fim de verificar se alguma delas pode ser retirada. Nesta etapa, o Teste da Razão de Verossimilhanças é recomendado para confirmar se o modelo com a variável é viável.

4. Com as variáveis restantes do passo 3, incluir as variáveis não significativas no passo inicial e verificar a possibilidade de inclusão de alguma delas. Novamente o uso do TRV é recomendado.

5. Por fim, verificar a possibilidade de incluir interações duas a duas entre as covariáveis. O modelo final será composto pelas covariáveis remanescentes no passo 4 e os temos de interação significativos nesta etapa.

Ao utilizar esse procedimento de seleção, Colosimo e Giolo (2006) aconselham evitar ser muito rigoroso ao testar cada nível de significância, sugerindo também um nível de significância próximo de 10%.

Passo 1

A Tabela 14 apresenta os resultados para o passo 1 (modelos com apenas uma variável).

Tabela 14: Passo 1: seleção de covariáveis significativas para a licenciatura diurna

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
$\beta_{\text{Forma de ingresso PAS}}$	0,33	0,09	3,57	$3{,}6 \times 10^{-4}$
$\beta_{\text{Forma de ingresso Vestibular}}$	0,19	0,09	2,13	$3{,}3 \times 10^{-2}$
$\beta_{\text{Outras formas de ingresso}}$	0,41	0,14	2,82	$4{,}8 \times 10^{-3}$
$\beta_{\text{Escola pública}}$	0,15	0,071	2,09	$3{,}6 \times 10^{-2}$
$\beta_{\text{Sexo Masculino}}$	-0,21	0,076	-2,79	$5{,}2 \times 10^{-3}$
$\beta_{\text{Sistema de cotas sim}}$	0,17	0,07	2,43	$1{,}5 \times 10^{-2}$
β_{IRA}	0,19	0,02	7,51	$5{,}8 \times 10^{-14}$
β_{Idade}	-0,01	0,006	-3,02	$2{,}6 \times 10^{-3}$
$\beta_{\text{Taxa de reprovação}}$	-0,74	0,11	-6,77	$1{,}3 \times 10^{-11}$
$\beta_{\text{Total de trancamentos}}$	0,007	0,012	0,62	0,54
$\beta_{\text{Cursou verão sim}}$	0,64	0,09	7,11	$1{,}1 \times 10^{-12}$

Usando um nível de 25% de significância, todas as variáveis são significativas, exceto total de trancamentos. Assim, o próximo passo é ajustar o modelo sem essa covariável.

Porém, como constatado na subseção 5.2, quatro covariáveis são correlacionadas: IRA é correlacionada com Taxa de reprovação e Escola é correlacionada com Sistema de cotas. Com isso, elas não poderão entrar conjuntamente no modelo. Temos portanto quatro possibilidades de modelos para o passo 2:

- Modelo 1: inclui IRA e Escola.
- Modelo 2: inclui IRA e Sistema de cotas.
- Modelo 3: inclui Taxa de reprovação e Escola.
- Modelo 4: inclui Taxa de reprovação e Sistema de cotas.

Passo 2

Para o passo 2, foram ajustados os quatro modelos encontrados no passo 1. Na Tabela 15 abaixo se encontram as estimativas.

Tabela 15: Passo 2: seleção de covariáveis significativas para a licenciatura diurna

Modelo	Parâmetro	Estimativa	EP	Estatística Z	P-valor
Modelo 1	β_{IRA}	0,15995	0,02533	6,31	$<$ 0,0001
	$\beta_{\text{Escola pública}}$	0,22288	0,06479	3,44	0.00058
	$\beta_{\text{Sexo Masculino}}$	-0,10011	0,06948	-1,44	0,14964
	$\beta_{\text{Outras formas de ingresso}}$	0,35717	0,14329	2,49	0,01268
	$\beta_{\text{Forma de ingresso PAS}}$	0,25613	0,08656	2,96	0,00309
	$\beta_{\text{Forma de ingresso Vestibular}}$	0,22688	0,08274	2,74	0,00610
	β_{Idade}	-0,00174	0,00671	-0,26	0,79557
	$\beta_{\text{Cursou verão sim}}$	0,49356	0,08806	5,60	$<$ 0,0001
Modelo 2	β_{IRA}	0,15708	0,02534	6,20	$<$ 0,0001
	$\beta_{\text{Sistema cotas sim}}$	0,20668	0,06886	3,00	0,00269
	$\beta_{\text{Sexo Masculino}}$	-0,07239	0,07074	-1,02	0,30619
	$\beta_{\text{Outras formas de ingresso}}$	0,41140	0,14631	2,81	0,00493
	$\beta_{\text{Forma de ingresso PAS}}$	0,29161	0,08778	3,32	0,00089
	$\beta_{\text{Forma de ingresso Vestibular}}$	0,25946	0,08402	3,09	0,00202
	β_{Idade}	0,00158	0,00671	0,24	0,81383
	$\beta_{\text{Cursou verão sim}}$	0,47860	0,08821	5,43	$<$ 0,0001
Modelo 3	$\beta_{\text{Taxa de reprovação}}$	-0,63439	0,10858	-5,84	5,1e-09
	$\beta_{\text{Escola pública}}$	0,23811	0,06449	3,69	0,00022
	$\beta_{\text{Sexo Masculino}}$	-0,12778	0,06825	-1,87	0,06118
	$\beta_{\text{Outras formas de ingresso}}$	0,33142	0,14377	2,31	0,02115
	$\beta_{\text{Forma de ingresso PAS}}$	0,25529	0,08507	3,00	0,00269
	$\beta_{\text{Forma de ingresso Vestibular}}$	0,19897	0,08138	2,44	0,01449
	β_{Idade}	-0,00347	0,00669	-0,52	0,60373
	$\beta_{\text{Cursou verão sim}}$	0,49267	0,08662	5,69	$<$ 0,0001
Modelo 4	$\beta_{\text{Taxa de reprovação}}$	-0,0625	0,109	-5,75	$<$ 0,0001
	$\beta_{\text{Sistema cotas sim}}$	0,230	0,0688	3,35	0,00081
	$\beta_{\text{Sexo Masculino}}$	-0,0960	0,0696	-1,38	0,16766
	$\beta_{\text{Outras formas de ingresso}}$	0,397	0,147	2,70	0,00687
	$\beta_{\text{Forma de ingresso PAS}}$	0,294	0,0856	3,40	0,00067
	$\beta_{\text{Forma de ingresso Vestibular}}$	0.236	0,0827	2,86	0,00429
	β_{Idade}	$<$ 0,0001	0,00671	-0,01	0,98939
	$\beta_{\text{Cursou verão sim}}$	0,477	0,0869	5,48	$<$ 0,0001

A 25% de significância, as variáveis sexo e idade não foram significativas no modelo 2. Nos demais modelos, somente a idade não foi significativa. Assim, foram

ajustados os respectivos modelos retirando essas variáveis, uma a uma, e de fato elas não foram significativas. Logo, idade e sexo saem do modelo 2, enquanto apenas a idade sai dos demais modelos.

Passo 3

Agora, retira-se as covariáveis que sobraram do passo 2, uma a uma, para verificar se algumas delas pode sair do modelo por meio do TRV.

Tabela 16: Passo 3: seleção de covariáveis significativas para a licenciatura diurna pelo TRV

Modelo	Hipótese nula (H_0)	TRV	P-valor
Modelo 1	$\beta_{\text{IRA}} = 0$	49,1019	$< 0{,}0001$
	$\beta_{\text{Escola}} = 0$	12,2830	0,0004
	$\beta_{\text{Sexo}} = 0$	2,16750	0,1409
	$\beta_{\text{Forma de ingresso}} = 0$	14,7772	0,0001
	$\beta_{\text{Cursou verão}} = 0$	33,2747	$< 0{,}0001$
Modelo 2	$\beta_{\text{IRA}} = 0$	45,8536	$< 0{,}0001$
	$\beta_{\text{Sistema cotas}} = 0$	0,5904	0,0011
	$\beta_{\text{Forma de ingresso}} = 0$	18,1098	$< 0{,}0001$
	$\beta_{\text{Cursou verão}} = 0$	30,5114	$< 0{,}0001$
Modelo 3	$\beta_{\text{Taxa de reprovação}} = 0$	39,8161	$< 0{,}0001$
	$\beta_{\text{Escola}} = 0$	13,6404	0,0002
	$\beta_{\text{Sexo}} = 0$	3,7782	0,0519
	$\beta_{\text{Forma de ingresso}} = 0$	18,4502	$< 0{,}0001$
	$\beta_{\text{Cursou verão}} = 0$	34,2623	$< 0{,}0001$
Modelo 4	$\beta_{\text{Taxa de reprovação}} = 0$	38,0796	$< 0{,}0001$
	$\beta_{\text{Sistema cotas}} = 0$	13,7866	0,0002
	$\beta_{\text{Sexo}} = 0$	1,9171	0,1661
	$\beta_{\text{Forma de ingresso}} = 0$	18,3307	$< 0{,}0001$
	$\beta_{\text{Cursou verão}} = 0$	31,5023	$< 0{,}0001$

Tem-se portanto que, a um nível de 10% de significância, o TRV não rejeita a hipótese nula para a variável sexo nos modelos 1 e 4, ou seja, essa variável sai de ambos os modelos. Nos demais casos, as covariáveis que sobreviveram ao passo 2 são de fato significativas e seguem todas nos respectivos modelos.

Passo 4

Nesta etapa ajusta-se os modelos do passo 3 agora com a única covariável que não entrou no modelo no passo 1: Total de trancamentos, fazendo uso do TRV mais uma vez para verificar se essa covariável realmente ficará de fora.

Tabela 17: Passo 4: seleção de covariáveis significativas para a licenciatura diurna pelo TRV

Modelo	Hipótese nula (H_0)	TRV	P-valor
Modelo 1	$\beta_{\text{Total de trancamentos}} = 0$	0,7561	0,3845
Modelo 2	$\beta_{\text{Total de trancamentos}} = 0$	1,0355	0,3088
Modelo 3	$\beta_{\text{Total de trancamentos}} = 0$	0,0009	0,9751
Modelo 4	$\beta_{\text{Total de trancamentos}} = 0$	0,0286	0,8655

O TRV não rejeita a hipótese nula nos quatro modelos listados na Tabela 17, ou seja, total de trancamentos segue fora dos modelos candidatos.

5.4.2 Modelos candidatos

Chegou-se portanto em quatro modelos candidatos que potencialmente modelam a variável resposta tempo de falha. São eles:

- Modelo 1: IRA + Escola + Forma de ingresso + Cursou verao.
- Modelo 2: IRA + Sistema cotas + Forma de ingresso + Cursou verao.
- Modelo 3: Taxa de reprovação + Escola + Forma de ingresso + Cursou verão + Sexo.
- Modelo 4: Taxa de reprovação + Sistema cotas + Forma de ingresso + Cursou verão.

Como forma de rejeitar modelos claramente inapropriados, uma análise de resíduos é bastante conveniente nesse caso. A Figura 32 abaixo apresenta os resíduos de Cox-Snell para os quatro modelos candidatos. Caso o modelo log-normal para a variável tempo de falha esteja bem ajustado aos dados, os resíduos devem seguir uma distribuição exponencial padrão.

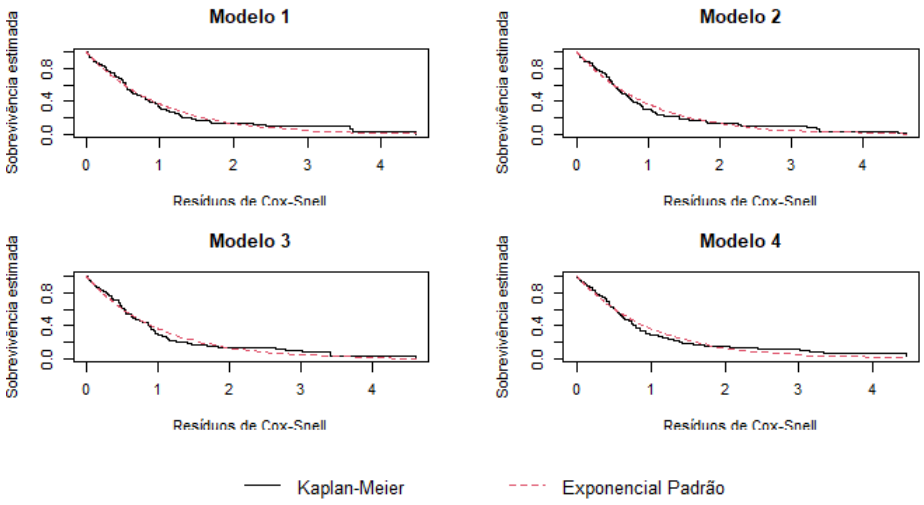


Figura 32: Resíduos Cox-Snell para os 4 modelos candidatos da licenciatura diurna.

A partir da Figura 32 pode-se considerar que o modelo de regressão log-normal se ajustou bem aos dados nos 4 modelos candidatos.

Como forma de escolher um modelo entre os quatro, critérios de parcimônia como as medidas AIC, AIC_c e BIC serão utilizados para este propósito.

Tabela 18: Critérios de informação para a escolha dentre os modelos candidatos para a licenciatura diurna

Modelo	AIC	AIC_c	BIC
Modelo 1	776,3723	776,7169	803,0293
Modelo 2	778,0649	778,4095	804,7219
Modelo 3	781,7884	782,1330	808,4454
Modelo 4	785,8388	786,1835	812,4958

Pelas medidas listadas na Tabela 18, o modelo 1, com IRA, Escola, Forma de ingresso e Cursou verão foi escolhido como modelo final para a licenciatura diurna. Nenhum termo de interação foi significativo.

5.4.3 Adequação do Modelo

Os métodos gráficos para validar o ajuste do modelo incluem a análise dos resíduos de Cox-Snell, já apresentado na Figura 32 para os modelos candidatos, bem como a distribuição dos resíduos $\hat{e}_i$ versus $\hat{H}(\hat{e}_i)$. Para que o modelo log-normal seja adequado, os resíduos Cox-Snell devem seguir uma distribuição exponencial padrão Pela Figura 33, o modelo log-normal parece estar bem ajustado aos dados.

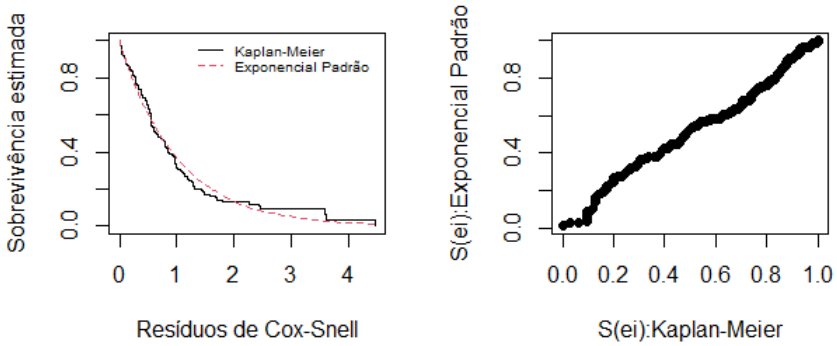


Figura 33: Curvas de sobrevivência estimadas (à esquerda) e resíduos de Cox-Snell estimados por Kaplan-Meier e pelo modelo exponencial padrão (à direita) para a licenciatura diurna

5.4.4 Modelo Final e Interpretação dos Coeficientes

Após todo o processo de seleção de variáveis e adequação pela análise de resíduos, pode-se partir para a interpretação dos coeficientes estimados. O modelo final possui as estimativas listadas na Tabela 19.

Tabela 19: Estimativas para o modelo final da licenciatura diurna

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
Intercepto	0,9583	0,0929	10,31	$< 0{,}0001$
β_{IRA}	0,1657	0,0244	6,79	$< 0{,}0001$
$\beta_{\text{Escola pública}}$	0,2271	0,0652	3,48	0,00049
$\beta_{\text{Outras formas de ingresso}}$	0,3606	0,1346	2,68	0,00740
$\beta_{\text{Forma de ingresso PAS}}$	0,2844	0,0827	3,44	0,00059
$\beta_{\text{Forma de ingresso Vestibular}}$	0,2364	0,0831	2,84	0,00445
$\beta_{\text{Cursou verão sim}}$	0,5010	0,0887	5,65	$< 0{,}0001$
$Log(scale)$	-0,7772	0,0565	-13,75	$< 0{,}0001$

Com isso, os coeficientes estimados possuem as seguintes interpretações:

1. Alunos com maiores valores de IRA possuem probabilidade maior de sobrevivência do que alunos com o IRA menor.

2. Alunos que estudaram em escola pública no ensino médio possuem maior probabilidade de sobreviver se comparado com alunos que estudaram em escola particular.

3. Para a variável Forma de ingresso, o modelo tomou como referência o Sisu. Nesse aspecto, todas as outras formas de ingresso possuem efeito positivo no modelo em

relação à referência, isto é, quem ingressou pelo PAS, vestibular ou outras formas de ingresso possui maior probabilidade de sobrevivência em relação a quem ingressou pelo Sisu.

4. O coeficiente positivo da variável cursou verão indica que os alunos que cursaram alguma disciplina no verão possuem maior probabilidade de não cometer evasão em relação a quem não cursou nenhuma disciplina no verão.

5.5 Modelagem para a Licenciatura Noturna

Nesta subseção será apresentado todo o processo de modelagem dos dados da licenciatura noturna em matemática. O procedimento será muito similar ao já feito na subseção 5.4 para a licenciatura diurna, uma vez que os dois bancos possuem as mesmas variáveis.

5.5.1 Seleção de Variáveis

Nove covariáveis foram selecionadas do banco de dados como potencialmente relevantes a serem incluídas no modelo para descrever o comportamento da variável resposta. O processo de seleção será o mesmo já descrito na subseção 5.4.1.

Passo 1

A Tabela 20 abaixo apresenta os resultados para o passo 1 (modelos com apenas uma variável).

Tabela 20: Passo 1: seleção de covariáveis significativas para a licenciatura noturna

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
$\beta_{\text{Outras formas de ingresso}}$	0,0475	0,0932	0,51	0,611
$\beta_{\text{Forma de ingresso PAS}}$	0,1536	0,0824	1,86	0,062
$\beta_{\text{Forma de ingresso Vestibular}}$	0,0468	0,0746	0,63	0,531
$\beta_{\text{Escola pública}}$	-0,0326	0,0572	-0,57	0,57
$\beta_{\text{Sexo Masculino}}$	0,0036	0,0721	0,05	0,96
$\beta_{\text{Sistema de cotas sim}}$	-0,0489	0,0587	-0,83	0,41
β_{IRA}	0,2198	0,0192	11,4	$<$ 0,0001
β_{Idade}	-0,00584	0,00317	-1,84	0,065
$\beta_{\text{Taxa de reprovação}}$	-0,7724	0,0811	-9,52	$<$ 0,0001
$\beta_{\text{Total de trancamentos}}$	-0,00570	0,00829	-0,69	0,49
$\beta_{\text{Cursou verão sim}}$	0,5168	0,0805	6,42	$<$ 0,0001

A 25% de significância, as variáveis Forma de ingresso, IRA, Idade, Taxa de reprovação e Cursou verão foram significativas. Assim, o próximo passo é ajustar o modelo sem as covariáveis não significativas (Escola, Sexo, Sistema de Cotas e Total de trancamentos).

Porém, como constatado na subseção 5.2, as covariáveis IRA e Taxa de reprovação são correlacionadas. Com isso, elas não poderão entrar conjuntamente no modelo. Temos portanto duas possibilidades de modelos para o passo 2:

- Modelo 1: inclui o IRA.
- Modelo 3: inclui a Taxa de reprovação.

Passo 2

Para o passo 2, foram ajustados os dois modelos encontrados no passo 1. Na Tabela 21 abaixo se encontram as estimativas.

Tabela 21: Passo 2: seleção de covariáveis significativas para a licenciatura noturna

Modelo	Parâmetro	Estimativa	EP	Estatística Z	P-valor
Modelo 1	β_{IRA}	0,20282	0,01953	10,38	$< 0{,}0001$
	β_{Idade}	0,0034	0,00320	1,06	0,29
	$\beta_{\text{Outras formas de ingresso}}$	0,05952	0,08441	0,71	0,48
	$\beta_{\text{Forma de ingresso PAS}}$	0,07558	0,07308	1,03	0,30
	$\beta_{\text{Forma de ingresso Vestibular}}$	0,07709	0,06213	1,24	0,21
	$\beta_{\text{Cursou verão sim}}$	0,31765	0,07634	4,16	$< 0{,}0001$
Modelo 2	$\beta_{\text{Taxa de reprovação}}$	-0,703324	0,078670	-8,94	$< 0{,}0001$
	β_{Idade}	0,000815	0,003234	0,25	0,80
	$\beta_{\text{Outras formas de ingresso}}$	-0,001320	0,085024	-0,02	0,99
	$\beta_{\text{Forma de ingresso PAS}}$	0,108980	0,074410	1,46	0,14
	$\beta_{\text{Forma de ingresso Vestibular}}$	0,081097	0,063715	1,27	0,20
	$\beta_{\text{Cursou verão sim}}$	0,388920	0,075636	5,14	$< 0{,}0001$

Em ambos os modelos as variáveis idade e forma de ingresso não são significativas. Assim, foram ajustados os modelos retirando essas variáveis, uma a uma, e de fato elas não foram significativas.

Passo 3

Agora, retira-se as covariáveis que sobraram do passo 2, uma a uma, para verificar se algumas delas pode sair do modelo por meio do TRV.

Tabela 22: Passo 3: seleção de covariáveis significativas para a licenciatura noturna pelo TRV

Modelo	Hipótese nula (H_0)	TRV	P-valor
Modelo 1	$\beta_{\text{IRA}} = 0$	111,2719	0
	$\beta_{\text{Cursou verão}} = 0$	18,0434	$< 0{,}0001$
Modelo 2	$\beta_{\text{Taxa de reprovação}} = 0$	77,4036	0
	$\beta_{\text{Cursou verão}} = 0$	30,1917	$< 0{,}0001$

Tem-se portanto que o TRV rejeita fotemente a hipótese nula em todos os casos, isto é, as covariáveis que sobreviveram ao passo 2 são de fato significativas e seguem todas nos respectivos modelos.

Passo 4

Nesta etapa ajusta-se os modelos do passo 3 agora com as covariáveis que não entraram no modelo no passo 1: Sexo, Sistema de Cotas, Escola e Total de trancamentos, fazendo uso do TRV mais uma vez para verificar se essas covariáveis realmente ficarão de fora.

Tabela 23: Passo 4: seleção de covariáveis significativas para a licenciatura noturna pelo TRV

Modelo	Hipótese nula (H_0)	TRV	P-valor
Modelo 1	$\beta_{\text{Sexo}} = 0$	0,1765	0,6743
	$\beta_{\text{Sistema cotas}} = 0$	0,0715	0,7891
	$\beta_{\text{Escola}} = 0$	4,2532	0,0391
	$\beta_{\text{Total de trancamentos}} = 0$	0,0345	0,8526
Modelo 2	$\beta_{\text{Sexo}} = 0$	0,0575	0,8104
	$\beta_{\text{Sistema cotas}} = 0$	0,0396	0,8421
	$\beta_{\text{Escola}} = 0$	1,4505	0,2284
	$\beta_{\text{Total de trancamentos}} = 0$	4,5340	0,0332

A 10% de significância, o TRV rejeita a hipótese nula para a variável Escola no modelo 1, e para a variável Total de trancamentos no modelo 2, ou seja, essas duas variáveis agora entram nos respectivos modelos.

5.5.2 Modelos candidatos

Chegou-se portanto em dois modelos candidatos que potencialmente modelam a variável resposta tempo de falha no caso da licenciatura noturna. São eles:

- Modelo 1: IRA + Escola + Cursou verao.
- Modelo 2: Taxa de reprovação + Total de trancamentos + Cursou verão.

A Figura 34 abaixo apresenta os resíduos de Cox-Snell para os dois modelos candidatos. Caso o modelo log-normal para a variável tempo de falha esteja bem ajustado aos dados, os resíduos devem seguir uma distribuição exponencial padrão.

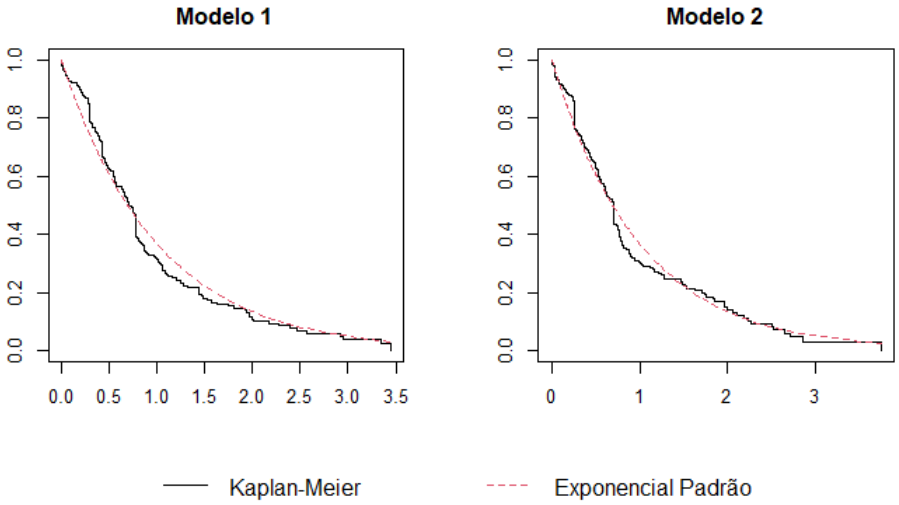


Figura 34: Resíduos Cox-Snell para os 4 modelos candidatos da licenciatura noturna.

A partir da Figura 34 pode-se considerar que o modelo de regressão log-normal se ajustou bem aos dados nos 2 modelos candidatos.

Como forma de escolher um modelo entre os dois, critérios de parcimônia como as medidas AIC, AIC_c e BIC serão utilizados para este propósito.

Tabela 24: Critérios de informação para a escolha dentre os modelos candidatos para a licenciatura noturna

Modelo	AIC	AIC_c	BIC
Modelo 1	844,9260	845,0377	860,5036
Modelo 2	878,5135	878,6252	894,0911

Pelas medidas listadas na Tabela 24, o modelo 1, com IRA, Escola e Cursou verão foi escolhido como modelo final para a licenciatura noturna. Nenhum termo de interação foi significativo.

5.5.3 Adequação do Modelo

Os métodos gráficos para validar o ajuste do modelo incluem a análise dos resíduos de Cox-Snell, bem como a distribuição dos resíduos $\hat{e}_i$ versus $\hat{H}(\hat{e}_i)$. Para que o modelo log-normal seja adequado, os resíduos Cox-Snell devem seguir uma distribuição exponencial padrão Pela Figura 35, o modelo log-normal parece estar bem ajustado aos dados.

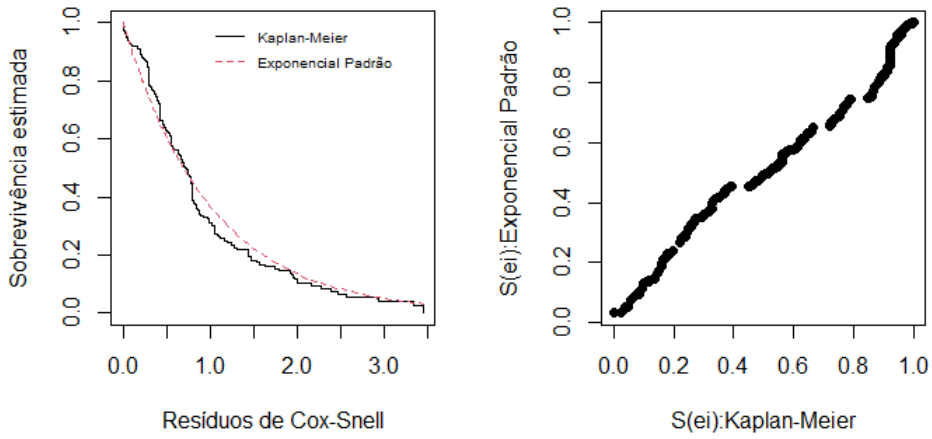


Figura 35: Curvas de sobrevivência estimadas (à esquerda) e resíduos de Cox-Snell estimados por Kaplan-Meier e pelo modelo exponencial padrão (à direita) para a licenciatura noturna

5.5.4 Modelo Final e Interpretação dos Coeficientes

Após todo o processo de seleção de variáveis e adequação pela análise de resíduos, pode-se partir para a interpretação dos coeficientes estimados. O modelo final possui as estimativas listadas na Tabela 25.

Tabela 25: Estimativas para o modelo final da licenciatura noturna

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
Intercepto	1,2415	0,0511	24,30	$<$ 0,0001
β_{IRA}	0,2071	0,0198	10,46	$<$ 0,0001
$\beta_{\text{Escola pública}}$	0,1035	0,0502	2,06	0,039
$\beta_{\text{Cursou verão sim}}$	0,3072	0,0756	4,06	$<$ 0,0001
$Log(scale)$	-0,9912	0,0505	-19,64	$<$ 0,0001

Com isso, os coeficientes estimados possuem as seguintes interpretações:

1. Alunos com maiores valores de IRA possuem probabilidade maior de sobrevivência do que alunos com o IRA menor.

2. Alunos que estudaram em escola pública no ensino médio possuem maior probabilidade de sobreviver se comparado com alunos que estudaram em escola particular.

3. Os alunos que cursaram alguma disciplina no verão possuem maior probabilidade de não cometer evasão em relação a quem não cursou nenhuma disciplina no verão.

6 Conclusão

Este trabalho teve como objetivo principal avaliar a relação entre o tempo que o aluno leva para cometer a evasão e outras nove covariáveis de interesse, no âmbito dos cursos de licenciatura diurna e noturna em Matemática da Universidade de Brasília.

A análise descritiva foi bastante esclarecedora no sentido de fornecer indicativos a respeito das covariáveis candidatas a entrar no modelo. O teste de *logRank* rejeitou a hipótese de igualdade das curvas de sobrevivência da variável Forma de ingresso apenas para o caso da licenciatura diurna. De fato, o modelo final para o diurno incluiu essa variável, além do IRA, Escola e Cursou verão. Já para o noturno, o modelo final incluiu todas as covariáveis do modelo do diurno, exceto justamente a Forma de ingresso.

Para a licenciatura diurna, a interpretação das estimativas dos parâmetros indicou que os alunos com maior IRA têm maior probabilidade de sobreviver em relação a alunos com menor IRA. Além disso, alunos que fizeram o ensino médio em escola pública tem probabilidade maior de sobreviver se comparados a alunos que estudaram em escola particular. A variável Forma de ingresso também entrou no modelo. Nesse sentido, quem ingressou pelo PAS, vestibular ou outras formas de ingresso possui maior probabilidade de sobrevivência em relação a quem ingressou pelo Sisu. Outro resultado que chamou a atenção foi o fato da variável Cursou verão ter sido muito significativa no modelo final, indicando que alunos que cursaram alguma disciplina no verão têm menor probabilidade de evadir do que alunos que não cursaram nenhuma disciplina nesse período.

Já no caso da licenciatura noturna, os resultados foram muito semelhantes aos encontrados na licenciatura diurna. Com a exceção da covariável Forma de ingresso, as demais covariáveis que entraram no modelo do diurno também entraram no noturno.

O método de seleção de variáveis proposto permitiu chegar em modelos escolhidos por critérios de parcimônia, uma vez que não existem modelos verdadeiros e absolutos, e sim aproximados da realidade. Quanto à qualidade do ajuste, ambos os modelos de regressão log-normal propostos tiveram bons resultados. A análise dos resíduos evidenciou o bom ajuste e a adequabilidade dos modelos.

Sugere-se, para trabalhos futuros: a inclusão de outras covariáveis que certamente acrescentariam ao modelo proposto, como por exemplo o local de residência e a distância percorrida pelo aluno até a universidade, ou até mesmo o tempo de deslocamento, se o aluno participa de algum projeto de extensão, se o aluno participa do Programa de Educação Tutorial em Matemática (PETMAT) e a realização de estudos semelhantes no âmbito dos demais cursos da Universidade de Brasília.

Referências

AARSET, M. V. How to identify a bathtub hazard rate. **IEEE Transactions on Reliability**, IEEE, v. 36, n. 1, p. 106–108, 1987.

ANDERSON, D.; BURNHAM, K. Model selection and multi-model inference. **Second. NY: Springer-Verlag**, v. 63, n. 2020, p. 10, 2004.

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

COLLETT, D. Modelling survival data. In: **Modelling survival data in medical research**. [S.l.]: Springer, 1994. p. 53–106.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. [S.l.]: Editora Blucher, 2006.

COOPER, H.; LINDSAY, J. J.; NYE, B.; GREATHOUSE, S. Relationships among attitudes about homework, amount of homework assigned and completed, and student achievement. **Journal of educational psychology**, American Psychological Association, v. 90, n. 1, p. 70, 1998.

COX, D.; OAKES, D. **Analysis of Survival Data**. [S.l.]: CRC Press, 1984. v. 21.

COX, D. R.; HINKLEY, D. V. **Theoretical statistics**. [S.l.], 1974.

FELDER, R. M.; FORREST, K. D.; BAKER-WARD, L.; DIETZ, E. J.; MOHR, P. H. A longitudinal study of engineering student performance and retention: I. success and failure in the introductory course. **Journal of Engineering Education**, Wiley Online Library, v. 82, n. 1, p. 15–21, 1993.

FILHO, R. L. L. S.; MOTEJUNAS, P. R.; HIPÓLITO, O.; LOBO, M. B. d. C. M. A evasão no ensino superior brasileiro. **Cadernos de pesquisa**, SciELO Brasil, v. 37, p. 641–659, 2007.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.

KLEIN, J. P.; MOESCHBERGER, M. L. **Survival analysis: techniques for censored and truncated data**. [S.l.]: Springer, 2003. v. 1230.

LEE, E. T.; WANG, J. **Statistical methods for survival data analysis**. [S.l.]: John Wiley & Sons, 2003. v. 476.

SACCARO, A.; FRANÇA, M. T. A.; JACINTO, P. d. A. Fatores associados à evasão no ensino superior brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de ciência, matemática e computação e de engenharia, produção e construção em instituições públicas e privadas. **Estudos Econômicos (São Paulo)**, SciELO Brasil, v. 49, p. 337–373, 2019.

SANTOS, D. F. d. Modelo de regressão log-logístico discreto com fração de cura para dados de sobrevivência. 2017.

SANTOS, R. dos; ALBUQUERQUE, A. E. M. Análise das taxas de abandono nos anos finais do ensino fundamental e do ensino médio a partir das características das escolas. **Cadernos de Estudos e Pesquisas em Políticas Educacionais**, v. 2, p. 34–34, 2019.

STINEBRICKNER, R.; STINEBRICKNER, T. Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model. **Journal of Labor Economics**, University of Chicago Press Chicago, IL, v. 32, n. 3, p. 601–644, 2014.

WEIBULL, W. **The propagation of fatigue cracks in light-alloy plates**. [S.l.]: Svenska Aeroplan Aktiebolaget, 1954.