Universidade de Brasília
Departamento de Estatística

Evasão no Curso de Bacharelado em Estatística da Universidade de Brasília: Uma Aplicação a modelos de Análise de Sobrevivência

Arthur de Oliveira Dias

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

Arthur de Oliveira Dias

Evasão no Curso de Bacharelado em Estatística da Universidade de Brasília: Uma Aplicação a modelos de Análise de Sobrevivência

Orientadora: Profª. Drª. Juliana Betini Fachini Gomes

Relatório parcial 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

Resumo

O Fenômeno da evasão de estudantes do ensino superior é maléfico ao país como um todo, à própria Universidade e também ao aluno evadido. A solução de tal problema passa pelo entendimento cada vez mais profundo dos fatores que influenciam na evasão escolar nessa população. O presente trabalho tem, portanto, o objetivo de colaborar no conhecimento mais profundo dessas causas nos cursos de bacharelado em Matemática e bacharelado em Estatística da Universidade de Brasília. A base de dados da Estatística possui 451 observações e 195 a base de dados da Matemática, que contém alunos ingressantes entre o primeiro semestre de 2011 e o segundo semestre de 2019 para a matemática e entre o primeiro semestre de 2014 e o segundo semestre de 2019 para a Estatística. A metodologia utilizada foi a análise de sobrevivência, que com o modelo de regressão log-normal proposto identificou a influência das variáveis sistema de cotas, forma de ingresso, cursou verão e taxa de reprovação, além de uma interação entre taxa de reprovação e sistema de cotas para a Estatística e a influencia de taxa de reprovação e cursou verão para a matemática. Portanto, faz-se relvantes estes resultados para subsidiar as discussões acerca da evasão e como combatê-la.

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

Abstract

The phenomenon of the evasion of higher education students is evil to the country as a whole, to the University itself and also to the evaded student. The solution to this problem involves an ever deeper understanding of the factors that influence school dropout in this population. The present work has, therefore, the objective of collaborating in the deeper knowledge of these causes in the bachelor's degree in Mathematics and bachelor's degree in Statistics from the University of Brasília. The Statistics database has 451 observations and 195 the Mathematics database, which contains students entering between the first semester of 2011 and the second semester of 2019 for mathematics and between the first semester of 2014 and the second semester of 2019 for Statistics. The methodology used was survival analysis, which with the proposed log-normal regression model identified the influence of the variables quota system, form of entry, summer course and failure rate, in addition to an interaction between failure rate and quota system for statistics and the influence of failure rate and summer course for mathematics. Therefore, these results are made to support discussions about evasion and how to combat it.

Keywords: Evasion; Fault; Censorship; Degree; Log-normal; Regression Model.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

A Educação é uma arte milenar e tem se mostrado essencial para sobrevivência da civilização e da cultura. Tal arte sofre de um problema importante, de difícil combate e que tem muitas causas: a evasão escolar (FILHO; ARAÚJO, 2017). Há diversos conceitos de evasão escolar, que é frequentemente confundida com o abandono escolar, conceitos que embora parecidos apresentam diferenças substanciais. O abandono escolar é identificado quando o aluno não conclui o período letivo em que ele está, ou seja, ele inicia o período letivo e antes do final do mesmo, deixa de frequentar o ambiente educacional. Já a evasão se faz presente na situação em que o aluno estava matriculado em t e não está mais em $t + 1$ (SANTOS; ALBUQUERQUE, 2019).

Esta interrupção no processo educacional no ensino superior é prejudicial por diversos fatores: no âmbito do serviço público de educação é claramente um desperdício de recursos como tempo e dinheiro, gerando impactos de ordem econômica e social, além do prejuízo acadêmico com a não formação dos estudantes como visto em Filho et al. (2007). Mas também para as instituições privadas a evasão representa a perda de receitas.

Conhecer o perfil desses estudantes que evadem mostra-se de exterma necessidade e importância, uma vez que dispondo dessas informações as instituições públicas e privadas podem estabelecer mecanismos e medidas que visem a maior eficiência de seus gastos, o sucesso do processo educacional do início ao fim, formando profissionais competentes e também o bem estar dos próprios estudantes com a realização de mais uma etapa em suas vidas ao invés de uma frustração de ter começado um curso e não ter conseguido terminar.

Compreender quantas e quais são as variáveis que podem influenciar a evasão fornece ao gestor da educação, e ao sistema educacional como um tudo, os insumos para montar seu estratagema para vencer esse mal que tanto afeta a vida dos estudantes e das instituições de ensino no Brasil. Há de se considerar também o grande fluxo de discentes entre cursos por perceberem que fizeram uma escolha errada. Tal fato enriquece ainda mais o valor do estudo da evasão e suas causas uma vez que fornece informações importantes até mesmo para o ensino básico. Já que não seria absurdo imaginar que uma má formação básica, um mau discernimento a respeito de aptidões e de vocação e uma falsa ideia do mundo da universidade pode contribuir para conturbar o processo de aprendizado no ensino superior.

Além do supracitado, a escassez de estudos sobre a evasão no Brasil, principalmente se comparado com outras partes do mundo motiva o presente trabalho que tem por finalidade contribuir no conhecimento do caso brasileiro da evasão e suas causas para iluminar um caminho de soluções.

1.1 Objetivo Geral

O presente trabalho tem como finalidade a investigação a respeito da evasão e de seus fatores contribuintes nos cursos de bacharelado em Estatística e bacharelado em Matemática da Universidade de Brasília (UnB). Para tal pretende-se fazer uso dos dados fornecidos pela Secretaria de Tecnologia da Informação da referida universidade.

1.2 Objetivos Específicos

1. Investigar e concluir a respeito de fatores que possivelmente influenciam na evasão nos curso de bacharelado em Estatística e bacharelado em Matemática da UnB, como sexo, idade, semestre, entre outros;

2. Aferir, uma vez identificados os fatores, o quanto eles influenciam na ocorrência da evasão no curso de bacharelado em Estatística da UnB;

3. Comparar tais fatores e aferições em relação a ambos os sexos;

2 Revisão de Literatura

2.1 Conceitos iniciais

A Análise Sobrevivência é uma área da Ciência Estatística que visa o estudo de dados relacionados ao tempo até a ocorrência de um evento de interesse (tempo de falha). Semelhante à análise de regressão, porém em análise de sobrevivência a informação da censura é considerada, ou seja, quando por algum motivo não se observa, durante o estudo o evento observado. Tal informação apesar de incompleta, tem valor na análise sobrevivência enquanto é desprezada por outras técnicas. Faz-se necessário, primeiramente, a definição de conceitos como tempo de falha, a unidade em que vai ser medido esse tempo, a falha, entre outros.

2.1.1 Tempo de Falha

O tempo de falha é composto substancialmente por três elementos:

(i) tempo inicial: é o tempo em que a observação se inicia, ou seja, queando começa o estudo. É de suma importância que as observações estajam sujeitas à mesma linha temporal;

(ii) escala: é a maneira como o tempo será medido, por exemplo, se é em horas, dias, meses, anos, semestres, etc;

(iii) evento de interesse: é o evento que - geralmente não se deseja de fato observar mas que - está sob estudo, como por exemplo a evasão, ou a falha de um equipamento, ou outro fato qualquer que seja bem especificado e identificável.

Definidos esses três elementos é possível construir a variável tempo de falha.

2.1.2 Censura

Como já mencionado, nem sempre há de se observar a falha em todas as observações. A essa não observação do evento de interesse se dá o nome "censura", uma vez que a informação é incompleta sobre aquele indivíduo observado, porém existe alguma informação. Existem diversos tipos de censura, uma vez que a não observação do evento pode ocorrer por diferentes motivos.

Censura à direita

Este tipo de censura acontece quando o tempo de falha está à direita do tempo final de observação do estudo. Tal tipo ainda pode ser subdivido em três casos:

(i) Censura Tipo I: Se faz presente quando o tempo final do estudo é fixado antes de iniciar o mesmo, e até esse tempo final não se observou a falha em um ou mais indivíduos observados;

(ii) Censura Tipo II: Ocorre quando o estudo é finalizado ao ter um número fixo de falhas, previamente definido, e ao final do estudo uma ou mais observações não falharam;

(iii) Censura aleatória: Quando por algum qualquer outro motivo, que não os supracitados, ao menos uma observação não falha.

Censura à esquerda

Acontece esse tipo de censura quando o tempo de falha está à esquerda do tempo inicial do estudo, ou seja, quando o estudo foi iniciado, já se tinha observado a falha.

Censura Intervalar Quando o acompanhamento dos indivíduos é períodico, não se sabe exatamente o tempo em que ocorreu a falha, apenas o intervalo $T \in (I, S)$ em que ele ocorreu. Os casos anteriores de censura são casos particulares desse. Basta tomar $I = 0$ (censura à esquerda) ou $S = 0$ (censura à direita).

Para a identificação das informações da variável reposta é utilizado para cada indivíduo a trinca (t_i, δ_i, x_i), em que, t_i é o tempo de falha ou de censura, x_i são as variáveis explicativas, se são consideradas, e δ_i é a variável que denota se aquela indivíduo falhou ou foi censurado, isto é,

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

2.2 Funções do tempo de Sobrevivência

Para o estudo da variável tempo T, são analisadas três funções, a saber: função densidade de probabilidade $f(t)$, função de Sobevivência $S(t)$ e função de risco $h(t)$.

2.2.1 Função densidade de probabilidade

A função densidade de probabilidade é definida como a probabilidade de um indivíduo falhar em um dado intervalo de tempo $[t, t + \Delta t)$ sobre o comprimento do intervalo(Δt). Para o caso contínuo, defin-se da seguinte maneira:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}, \tag{2.2.1}$$

sendo que $f(t)$ é uma função não-negativa e a área abaixo da curva de $f(t)$ soma 1.

2.2.2 Função de sobrevivência

A função de sobrevivência é a probabilidade de um indivíduo não falhar até um tempo t, isto é, a probabilidade de sobreviver até esse momento:

$$S(t) = P(T \geq t) = \int_t^{\infty} f(x)\, dx. \tag{2.2.2}$$

A partir desta função, tem-se a probabilidade de o indivíduo não sobreviver até o tempo t: $F(t) = 1 - S(t)$.

Temos como particularidades da Função de Sobrevivência $S(t)$:

$$\lim_{t \to 0} S(t) = 1 \text{ e} \tag{2.2.3}$$

$$\lim_{t \to \infty} S(t) = 0. \tag{2.2.4}$$

Quando acontece o acima descrito dizemos que $S(t)$ é uma função de sobrevivência própria. Pode ocorrer também de $S(t)$ ser uma função de sobrevivência imprópria, o que acontece quando

$$\lim_{t \to 0} S(t) = 1 \text{ e} \tag{2.2.5}$$

$$\lim_{t \to \infty} S(t) = p, \tag{2.2.6}$$

em que p é uma probabilidade.

2.2.3 Função de risco

A função de risco, ou taxa de falha, é o limite da probabilidade de um indivíduo falhar no intervalo $[t, \Delta t)$ (assume-se que ele sobreviveu até t) dividida pelo comprimento do intervalo Δt:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}. \tag{2.2.7}$$

A partir dessa função, obtém-se a função de risco acumulada, que é de grande importância na Análise de Sobrevivência:

$$H(t) = \int_o^t h(u)\, du. \tag{2.2.8}$$

A função de risco acumulada apresenta ótimas propriedades e é útil na avaliação de $h(t)$, além de auxiliar na escolha do modelo que melhor se ajusta aos dados.

2.2.4 Ralação entre $f(t)$, $S(t)$ e $h(t)$

As funções $f(t)$, $S(t)$ e $h(t)$ possuem relações matemáticas entre si que possibilitam a obtenção de uma pelas outras. São elas:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(logS(t)), \tag{2.2.9}$$

$$H(t) = -log(S(t)) \text{ e} \tag{2.2.10}$$

$$S(t) = exp{-}H(t) = exp\left\{\int_o^t h(u)\, du\right\}. \tag{2.2.11}$$

2.3 Técnicas não-paramétricas

2.3.1 Estimador de Kaplan-Meier

O estimador proposto por Kaplan e Meier (1958) é um estimador de máxima verossimilhança não-paramétrico da função de sobrevivência $S(t)$. Para grandes amostras o estimador é não viesado e quando não há censuras é definido por:

$$\hat{S}(t) = \frac{\text{número de observações que não falharam até o tempo t}}{\text{número total de observações no estudo}}. \tag{2.3.1}$$

Na prsença de censuras, o estimador fica definido da seguinte maneira:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right), \tag{2.3.2}$$

considerando:

(i) $t_1 < t_2 < ... < t_k$ sejam os k tempos distintos e ordenados de falha;

(ii) tantos intervalos de tempo quantos forem o número de falhas distintas.Os limites dos intervalos de tempo são os tempos de falha da amostra;

(iii) que d_j seja o número de falhas em t_j , $j = 1, ..., k$; e

(iv) que 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 .

2.3.2 Curva do tempo total em teste

A curva do tempo total em teste, conhecida como TTTplot foi proposta por Aarset (1987) e ajuda na identificação do modelo mais adequado para a variável T. A construção da curva pe obtida através de:

$$G(r/n) = \frac{[(\sum_{i=1}^{r} T_{i:n}) + (n-r)T_{r:n}]}{(\sum_{i=1}^{r} T_i)}, \tag{2.3.3}$$

$r = 1, ..., n$ e $T_i : n, i = 1, ..., n$ são as estatísiticas de ordem da amostra.

Esta curva pode ter diversos formatos e cada comportamento nos fornece uma informação diferente, como mostra a figura a seguir:

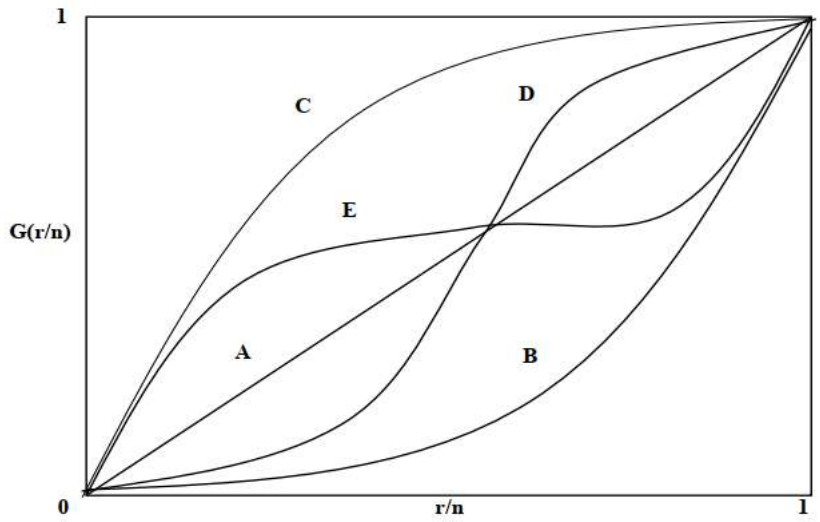


Figura 1: Ilustração de algumas formas da curva TTT

•Reta diagonal (A): A função de risco constante é apropriada.

• Curva convexa (B): A funçãode risco é monotonicamente decrescente.

• Curva côncava (C): A funçãoo 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.

2.3.3 Curva da função de risco acumulada $H(t)$

Nos casos em que há presença de censuras, a Função de risco acumulada é mais adequada. É possível ser obtida pelo estimador de Kaplan-Meier e sua interpretação é a inversa da curva TTT, isto é:

• 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.

2.4 Distribuições de probabilidade

2.4.1 Distribuição Weibull

A Distribuição Weibull (WEIBULL, 1954) é uma das principais distribuições utilizadas em modelos probabilísticos paramétricos. A função de risco dessa distribuição pode apresentar várias formas, dependendo do seu parâmetro γ:

$$\begin{cases} \gamma < 1, & \text{a função é decrescente} \\ \gamma > 1, & \text{a função é crescente} \\ \gamma = 1, & \text{a função é constante.} \end{cases}$$

• Densidade de probabilidade

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

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}, t \geq 0, \tag{2.4.1}$$

onde $\gamma \geq 0$, $\alpha \geq 0$ e α tem a mesma unidade de medida que t e γ não tem unidade.

• Função de Sobrevivência

A Função de Sobrevivência da distribuição Weibull é dada por:

$$S(t) = exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}, t \geq 0, \tag{2.4.2}$$

onde $\gamma \geq 0$, $\alpha \geq 0$.

• Função de risco

A Função de risco da distribuição Weibull é denotada por:

$$h(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1}, t \geq 0, \tag{2.4.3}$$

onde $\gamma \geq 0$, $\alpha \geq 0$.

A média e a varância são dadas por:

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

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

onde $\Gamma(\alpha) = (\alpha - 1)!$ para α inteiro.

Tomando $\gamma = 1$, temos a seguinte densidade:

$$f(t) = \frac{1}{\alpha} exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, t \geq 0, \tag{2.4.4}$$

que é a densidade da distribuição Exponencial, ou seja, a distribuição exponencial é um caso particular da distribuição Weibull: quando $\gamma = 1$.

2.4.2 Distribuição Log-logística

A Distribuição Log-Logística apresenta uma forma fechada para suas funções de sobrevivência e de risco. Seja T uma variável aleatória com distribuição Log-logística. T apresenta a seguintes funções:

• Densidade de probabilidade

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \left[1 + \left(\frac{t}{\alpha}^\gamma\right)\right]^{-2}, \tag{2.4.5}$$

onde $t, \alpha, \gamma > 0$. α é parâmetro de escala e γ de forma.

• Função de Sobrevivência

A Função de Sobrevivência da distribuição Log-logística é dada por:

$$S(t) = \frac{1}{1 + (t/\alpha)^\gamma}, t \geq 0, \tag{2.4.6}$$

• Função de risco

Sua função de risco é denotada por:

$$h(t) = \frac{\gamma\,(t/\alpha)^{\gamma-1}}{\alpha\,[1 + (t/\alpha)^\gamma]}, t \geq 0, \tag{2.4.7}$$

Se T tem distribuição Log-logística, $W = log(T)$ tem distribuição logística, com $\mu \in \Re$ e $\sigma > 0$, em que $\gamma = 1/\sigma$ e $\alpha = exp(\mu)$.

2.4.3 Distribuição Log-normal

Assim como as anteriormente apresentadas, a distribuição Log-normal é bastante utilizada no estudo de tempos de vida de indivíduos e produtos. Seja T uma variável aleatória com distribuição de probabilidade Log-normal, então temos:

• Densidade de probabilidade

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{log(t)-\mu}{\sigma}\right)^2\right\}, t \geq 0 \tag{2.4.8}$$

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

• Função de Sobrevivência e função de risco

As Função de Sobrevivência ede Risco da distribuição Log-normal não possuem forma analítica explícita e são dadas, respectivamente, por :

$$S(t) = \Phi\left(\frac{-log(t)+\mu}{\sigma}\right) \text{ e } h(t) = \frac{f(t}{S(t)}, \tag{2.4.9}$$

onde $\Phi(.)$ é a função de densidade acumulada da normal padrão.

2.5 Estimação dos Parâmetros

Os parâmetros da distribuição escolhida para modelar os dados precisam ser estimados através dos dados amostrais. O Método mais adequado é o de máxima verossimilhnaça, uma vez que mínimos quadrados, por exemplo, não incorpora dados censurados.

2.5.1 Método da Máxima Verossimilhança

O Método de máxima verossimilhança busca encontrar os valores dos parâmtros que tenham maior possibilidade de ter gerado os dados da amostra. Portanto, tomando uma amostra de tempor t_i tamanho n, isto é, $i = 1, ..., n$ com ausência de censura e $f(t)$ a função de probabilidade da população, a função de verossimilhança para um vetor $\boldsymbol{\theta}$ é dada por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i, \boldsymbol{\theta}). \tag{2.5.1}$$

Definida essa função basta encontrar um $\boldsymbol{\theta}$ que maximize $L(\boldsymbol{\theta})$. Quando há censuras, podemos dividir em dois grupos as observações, onde para as r observações que falharam, sua contribuição em $L(\boldsymbol{\theta})$ será $f(t_i, \boldsymbol{\theta})$, e para as $n - r$ censuradas sua contribuição em $L(\boldsymbol{\theta})$ será $S(t_i, \boldsymbol{\theta})$. Dessa maneira a função de verossimilhança é denotada por:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{r} f(t_i, \boldsymbol{\theta}) \prod_{i=r+1}^{n} S(t_i, \boldsymbol{\theta}), \tag{2.5.2}$$

que pode ser descrita também como

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [f(t_i, \boldsymbol{\theta})]^{\delta_i} [S(t_i, \boldsymbol{\theta})]^{1-\delta_i} = \prod_{i=1}^{n} \left[\frac{f(t_i, \boldsymbol{\theta})}{S(t_i, \boldsymbol{\theta})}\right]^{\delta_i} S(t_i, \boldsymbol{\theta}) = \prod_{i=1}^{n} [h(t_i, \boldsymbol{\theta})]^{\delta_i} S(t_i, \boldsymbol{\theta}). \tag{2.5.3}$$

Para facilitar o cálculo, é conveniente trabalhar com o logaritmo de $L(\boldsymbol{\theta})$. Dessa maneira, aplicando o *log* em $L(\boldsymbol{\theta})$, temos

$$l(\boldsymbol{\theta}) = log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \delta_i log[f(t_i, \boldsymbol{\theta})] + (1 - \delta_i) \sum_{i=1}^{n} \delta_i log[S(t_i, \boldsymbol{\theta})]. \tag{2.5.4}$$

Trabalhando com $l(\boldsymbol{\theta})$, que é mais tratável, basta derivar em relação à $\boldsymbol{\theta}$ e encontrar o valor de $\boldsymbol{\theta}$ que zera essa derivada para encontrar o ponto de máximo da função.

2.5.2 Intervalo de confiança

A estimação intervalar dos parâmetros é o passo seguinte à estimação pontual, que é possível graças às propriedades que valem para grandes amostras. A principal trata da variâcia de $\boldsymbol{\theta}$:

$$Var(\boldsymbol{\theta}) \approx -[E(I_F(\boldsymbol{\theta})]^{-1}. \tag{2.5.5}$$

Para o caso de $\boldsymbol{\theta}$ ser um escalar, a construção clássica de um intervalo de confiança com nivél $\gamma 100\%$ de confiança é dado por:

$$\theta \pm z_{\frac{(1+\gamma)}{2}} \sqrt{\hat{Var}(\hat{\theta})}. \tag{2.5.6}$$

Para o caso em que $\boldsymbol{\theta}$ é um vetor, utiliza-se a matriz de variância e covariância para se obter a estimativa para o erro pardão e constroe-se um intervalo para cada parâmetro.

2.6 Escolha do modelo probabilístico

2.6.1 Teste da Razão de Verossimilhança

As técnicas gráficas para a escolha do modelo probabilístico mais adequado, amplamentes utilizadas, tem intrinsecamente uma dose de subjetividade, que pode ser excluída utilizando um teste de hipóteses, objetivo, para a confirmação ou não do que se viu graficamente. As hipóteses a serem testadas são:

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

Para a realização do teste, ajusta-se dois modelos: (1) modelo generalizado e se obtém o valor do logaritmo da sua função de verossimilhança ($logL(\hat{\theta_G})$); (2) modelo de interesse e se obtém o valor do logaritmo de sua função de verossimilhança ($logL(\hat{\theta_M})$). A partir dessas quantidades, calcula-se a estatística do teste da razão de verossimilhança, que é dada por:

$$TRV = -2log\left[\frac{L(\hat{\theta_M})}{L(\hat{\theta_G})}\right], \tag{2.6.1}$$

que, sob H_0, tem distribuição qui-quadrados com o número de graus de liberdade igual à diferença do número de parâmetros entre os dois modelos.

2.7 Critérios de Informação

Outros métodos objetivos de seleção de modelos são os critérios de Akaike e Bayeiano, que são muito utilizados para selecionar o modelo mais adequado.

2.7.1 Critério de Akaike - AIC

Essa estimativa baseia-se no logaritmo da função de verossimilhança no ponto de máximo, somado a uma penalidade associada ao número de parâmtros, que pretende corrigir um viés advindo da comparação de modelos com diferente número de parâmetros. Tal estimativa é definida por:

$$AIC = -2logL(\hat{\theta}) + 2p, \tag{2.7.1}$$

onde p é o número de parâmetros do modelo. A recomendação é o uso desse critério quando $n/p \geq 40$. O modelo considerado mais adequado deve apresentar o menor valor do AIC dentre os modelos analisados.

2.7.2 Critério de Akaike Corrigido - AICc

Quando a condição $n/p \geq 40$ não é atendida, pode-se utiliza a correção:

$$AICc = AIC + \frac{2p(p+1)}{n-p-1}, \tag{2.7.2}$$

Escolhe-se também o modelo com menor valor do $AICc$ dentre os modelos analisados.

2.7.3 Critério de Informação Bayesiano - BIC

O BIC é uma opção para seleção de modelos. Este critério penalisa mais os modelos que contém maior número de parâmetros, tendendo asism, a selecionar modelos com menos parâmetros. A fórmula do BIC é dada por

$$BIC = -2logL(\hat{\theta}) + plog(n), \tag{2.7.3}$$

À semelhança dos outros critérios, seleciona-se o modelo com menor valor do BIC.

2.8 Adequação do modelo Ajustado

Após o ajuste do modelo e sua seleção, é necessário que se verifique a adequação desse ajuste. Essa parte é fundamental na análise dos dados e para tal são propostas diversas técninas de analise de resíduo, como os resíduos de Cox-Snell.

2.8.1 Resíduos de Cox-Snell

Os resíduos de Cox-Snell são definidos como

$$\hat{e_i} = \hat{H}(t_i|x_i), \tag{2.8.1}$$

sendo que $\hat{H}(.)$ é a função de risco acumulado obtida do modelo ajustado e x o vetor de covariáveis. Os resíduos de Cox-Snell são oriundos de uma população homogênea com distribuição exponencial padrão. Portanto, o gráfico de $\hat{e_i}$ versus $\hat{H}(\hat{e_i})$ deve ser aproximadamente uma reta.

3 Metodologia

3.1 Base de Dados

O Presente trabalho utiliza dados da Secretaria de Tecnologia da Informação da Universidade de Brasília.

Em 2020 aconteceu uma migração do sistema SIGRA para o sistema SIGAA na Universidade. Portanto foram disponibilizados 4 bancos de dados: 2 bancos referentes ao sistema SIGRA, dos cursos de Bacharelado em Estatística e em Matemática, e outros dois dos mesmos cursos, referentes ao sistema SIGAA. É importante notar que entre as bases dos dois sistemas para um mesmo curso não há mistura, ou seja, um mesmo aluno só pode estar em uma delas simultaneamente. Portanto, foram unidas as base de dados dos dois sistemas para os respectivos cursos e utilizados os dados de ambos sistemas, nos períodos definidos como tempo de observação para cada curso.

3.2 Variáveis

As 4 bases contém as mesmas 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;

A base do sistema SIGRA para Estatística contém 70814 observações enquanto que a do sistema SIGAA possui 22927 observações. Já para a Matemática, a base referente ao sistema SIGRA possui 43137 observações e 4422 a que se refere ao SIGAA.

3.3 Filtragem e junção das bases

Após unir as bases dos dois sistemas para cada curso, filtrou-se o período de ingresso dos alunos para que não excluir efeitos da diferença de currículos. Portanto no

curso da Estatística foram excluidos os alunos que entraram antes do 1º semestre de 2014 e para a Matemática foram excluídos os alunos que entraram antes do 1º semestre de 2011. Como os dados se referem à data de maio de 2022, além deste, foi feito um filtro para desconsiderar o acontecido na pandemia(2020-2022), ou seja, excluiu-se os alunos que ingressaram e/ou saíram depois do 2º semestre de 2019, além das disciplinas cursadas pelos alunos ativos no curso durante o período da pandemia. Inicialmente, cada observação represeta um par (aluno,disciplina). Após esses filtros, fez-se um procedimento de agrupamento para que cada observação representasse um aluno, fazendo com que a base da matemática ficasse com 195 obsrvações e a da Estatística com 451.

3.4 Criação de Variáveis

Considerou-se importante criar, a partir da base de dados, algumas variáveis relevantes para o estudo. São elas:

3.4.1 Total de Trancamentos

Variável que contabiliza quantas disciplinas foram trancadas pelo aluno durante o curso. Foi considerado trancamento a disciplina que continha na variável "Menção disciplina"os valores TR ou TJ, que significam trancamento e trancamento justificado, respectivamente.

3.4.2 Taxa de reprovação

Corresponde à razão de créditos das disciplinas reprovadas pelo aluno no curso em relação ao total de créditos das disciplinas cursadas. Assim como a Universidade, definiu-se como reprovação as menções SR,II e MI e como cursado as menções SR,II,MI,MM,MS e SS. Tais informações vieram da variável Menção disciplina.

3.4.3 Cursou verão

Variável qualitativa binária que assume o valor 1 caso o aluno tenha cursado alguma disciplina em algum verão durante o curso e 0 caso contrário.Essa informação foi verificada a partir da variável período cursou disciplina.

3.4.4 Censura

Essa variável é de suma importância pois é a definição se o aluno em questão falhou (evadiu) ou sobreviveu (não evadiu). Foi considerada a falha, conforme a tabela 1 a seguir:

Tabela 1: Formas de saída do curso e a definição de falha/censura

Forma de saída	Falha/censura
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

3.4.5 Tempo

Esta variável será medida em semestres e para construí-la utilizou-se das variáveis Período saída curso e Período ingresso como a seguir:

$$Tempo = (\text{período saída curso - período ingresso curso}) \times 2 + 1, \tag{3.4.1}$$

Para os alunos com forma de saída igual a ativo (censura) foi considerado o período de saída o último semestre que tinham cursado alguma disciplina (Período cursou disciplina). As variáveis período saída curso e período ingresso curso são dadas em anos, então faz-se necessária a multiplicação da diferença por 2 para a obtenção do tempo em semestres. À essa quantidade soma-se 1 pois considerou-se que uma ingresso na Universidade o aluno já está sujeito à evasão, ou seja, não existe tempo de falha/censura igual a zero, isto é, se o aluno ingressou no semestre 1/2016 e evadiu no 2/2016, seu tempo será 2, pois ele evadiu no segundo semestre cursado.

3.4.6 Idade

Variável que indica a idade do aluno ao ingressar no curso estudado, construída a partir da data de nascimento, presente no banco.

3.5 Reclassificação da Variável forma de ingresso no curso

Esta variável possui 11 categorias,contudo apenas 3 são expressivas, sendo as demais poucos casos. Portanto, optou-se por fazer uma reclassificação, colocando as categorias inexpressivas em "outras"formas de ingresso. Tem-se a nova classificação da variável: SISU, PAS, Vestibular e outras.

3.6 Análise dos dados

A análise se inicia com a descritiva dos dados, observando as covariáveis por meio de gráficos e tabelas de frequência. Para as variáveis qualitativas, os gráficos de barra serão amplamente utilizados e boxplot e histograma para as variáveis quantitativas. Além dessas técnicas comuns, serão utilizados recursos mais específicos da análise de sobrevivência que são os gráficos com as curvas de sobrevivência, a partir do indicador de Kaplan-Meier e a função de risco acumulada para encontrar indicativos de uma distribuição de probabilidade que modele bem os dados.

3.7 Modelagem

O primeiro passo para proceder à modelagem é a seleção da distribuição de probabilidade. Para tal, utiliza-se-á os métodos gráficos e critérios definidos em 3.7 Critérios de informação. Em seguida procede-se à seleção de veriáveis, seguindo Collett (1994).

3.8 Modelo de Regressão

Seja $x^T = (1, x_1, x_2, ...x_p)$ um vetor de covariáveis e g uma função de ligação. Dado um conjunto de p variáveis, o vetor de parâmetros θ é definido como:

$$\boldsymbol{\theta} = g(\boldsymbol{x}^T\boldsymbol{\beta}), \tag{3.8.1}$$

onde $\beta = (\beta_0, beta_1, ...beta_p)^T$ é o vetor dos coeficientes de regressão.

Seja T uma varável aleatória que segue uma das distribuições de probabilidade definidas na seção 3.4, um modelo de regressão pode ser definido ao considerar um dos parâmetros da densidade de probabilidade igual a $g(\boldsymbol{x}^T\boldsymbol{\beta})$.

Assim, ao considerar, por exemplo, a distribuição log-normal e tomando o parâmetro μ como $\mu = x^T\beta$, a função de ligação identidade $I(.)$, tem-se o modelo de regressão log-normal definido por:

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{log(t) - x^T\beta}{\sigma}\right)^2\right\}.t \geq 0 \tag{3.8.2}$$

A função de sobrevivência e a função de risco são dadas por:

$$S(t) = \Phi\left(\frac{-log(t) + x^T\beta}{\sigma}\right) \text{ e } h(t) = \frac{f(t}{S(t)}. \tag{3.8.3}$$

A estimação dos parâmetros do modelo de regressão log-normal seguirão método de máxima verossimilhança descrito na subseção 3.5.1. . Para a relização d todas as análises estatísticas pertinentes e calcular as estimativas para o modelo, será utilziado o software estatístico livre R.

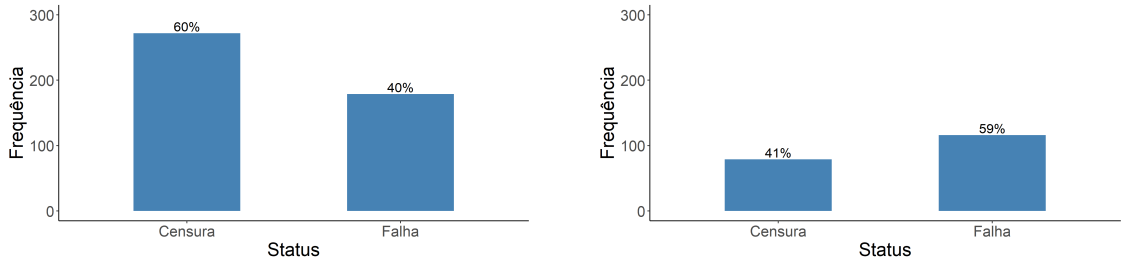
4 Resultados

4.1 Análise Descritiva

O primeiro passo em qualquer análise Estatística é conhecer os dados, e na análise de sobrevivência não é diferente. Com esse objetivo, pretende-se expor uma visão geral sobre as covariáveis mais suscetíveis a entrarem no modelo.

4.1.1 Censura

A variável Censura é a variável indicadora que fornece a informação se o tempo é de falha ou de censura.



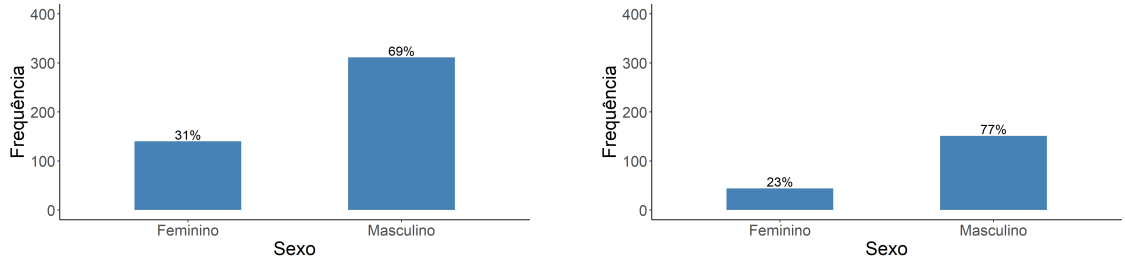
(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 2: Gráficos de barras para a variável Status

No curso de Estatística há uma maior presença de censuras, ou seja, que alunos que não evadiram, enquanto na Matemática essa situação se inverte e há uma proporção um pouco maior de falhas em relação às censuras.

4.1.2 Sexo

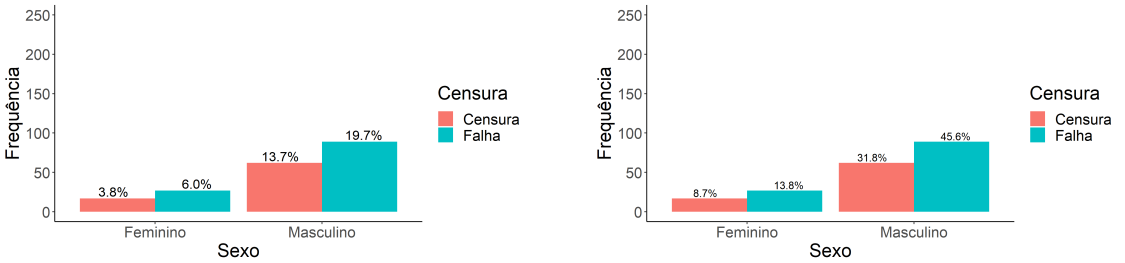
Esta é a variável que guarda a informação do sexo do aluno, podendo ser "Masculino"ou "Femino", declarado pelo mesmo.



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 3: Gráficos de barras para a variável Sexo

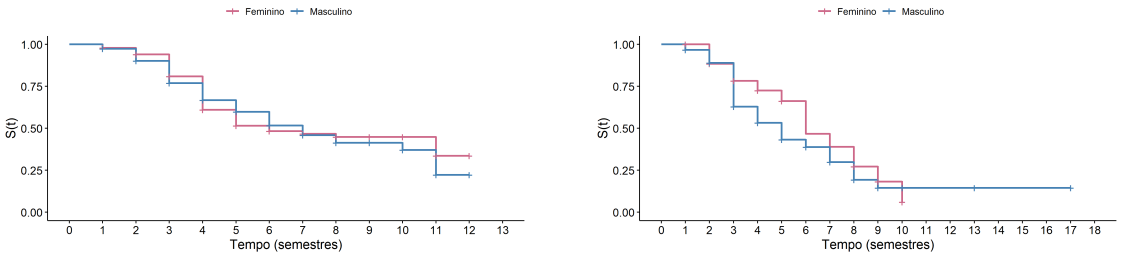
Ambos os cursos tem uma presença majoritária de homens em relação às mulheres, mostrando um perfil muito parecido em relação à esse aspecto, o que é natural já que os cursos são da mesma natureza.



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 4: Gráficos de barras para a variável Sexo vs Status

Os gráficos acima indicam comportamentos semelhantes entre os sexos em relação à porporção de falha e censura.



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

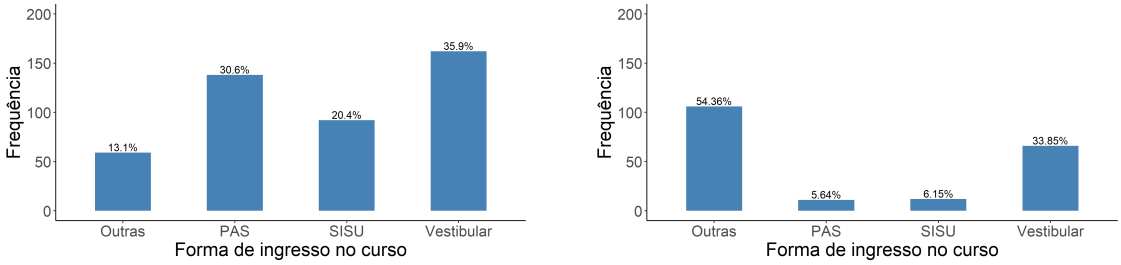
Figura 5: Curvas de sobrevivência para a variável Sexo

As curvas de sobrevivência dos sexos para Estatística parecem bem semelhantes, se cruzando em vários pontos e com um comportamento idêntico. Já para a Matemática a curva de sobrevivência das mulheres fica acima da dos homens em um certo período de

tempo (2 e 9 semestres), mostrando que aparentemente as mulheres tem uma probabilidade de sobrevivência um pouco maior do que a dos homens nesse período.

4.1.3 Forma de ingresso

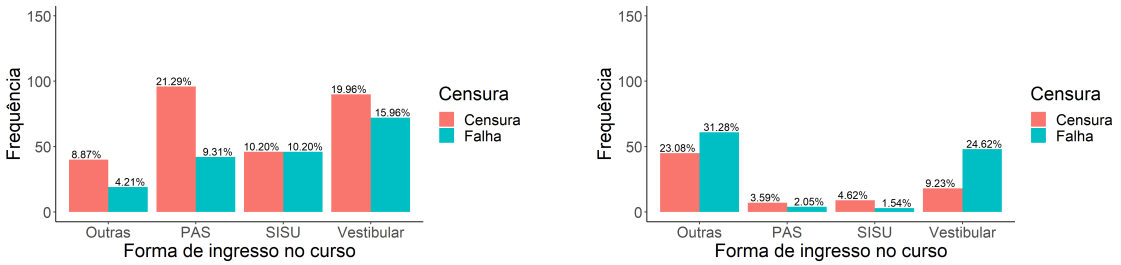
Variável que guarda a informação do tipo de ingresso no curso na UnB, podendo ser "PAS", "SISU", "Vestibular"ou outros.



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

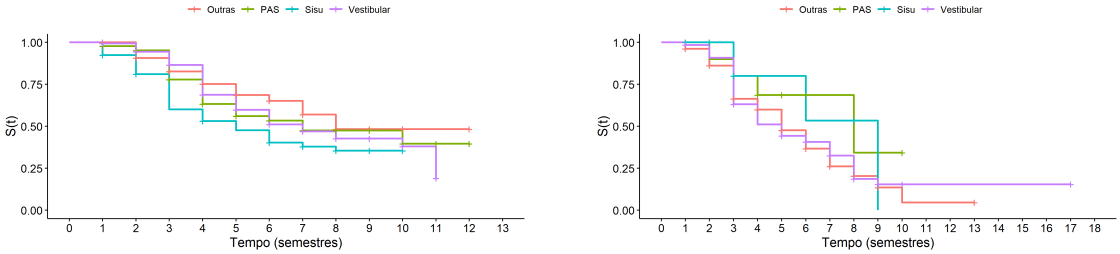
Figura 6: Gráficos de barras para a variável Forma de ingresso

No curso da Estatística, PAS e vestibular são as principais formas de ingresso enquanto que na matemática as "outras"formas junto com vestibular são as que tem o maior percentual. Isso caracteriza o pefil da maioria dos alunos que entram no curso de Bacharelado em Estatística que geralmente vêm de outros cursos ou ingressa como dupla habilitação, ou seja, não ingressando nem pelo "PAS"nem pelo "SISU"nem pelo "Vestibular".



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 7: Gráficos de barras para a variável Forma de ingresso vs Status



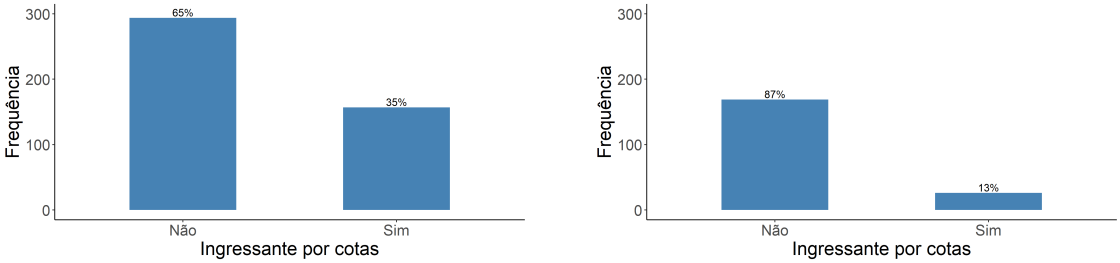
(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 8: Curvas de sobrevivência para a variável Forma de ingresso

As Figuras 7 e 8 nos mostram comportamentos diferentes entre as curvas e sobrevivênica. No curso de Estatística, a forma de ingresso "SISU"aparece quase sempre com uma probabilidade de sobrevivência menor do que "PAS", "Vestibular"e "Outras", que se comportam de maneira semelhantes. Já na Matemática, as curvas de "SISU"e "Outras"são semelhantes, e estas bem diferente das curvas de "PAS"e "Vestibular".

4.1.4 Sistema de cotas

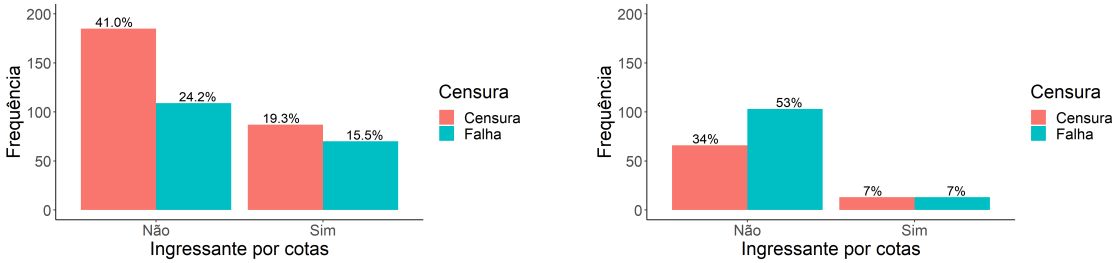
Variável binária que indica se o aluno entrou no curso pelo uso de cotas ou não, sejam quais forem.



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 9: Gráficos de barras para a variável Sistema de cotas

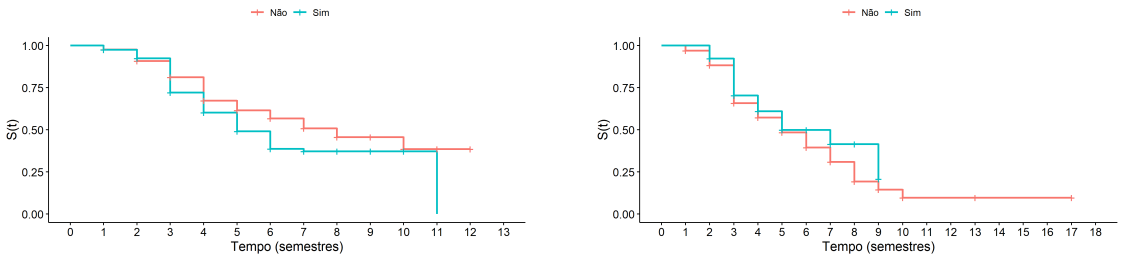
Em ambos os cursos, a proporção de de cotistas é inferior à proporçao de não cotistas, cerca de 40%.



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 10: Gráficos de barras para a variável Sistema de cotas vs Status

Observa-se comportamentos bem distintos entre os cotistas e não cotistas, com os alunos que não entraram através de cotas com uma proporção de falha bem menor em relação aos cotistas, no curso de Estatística. Na Matemática o comportamento já é bem diferente, tendo os não cotistas com proporção maior de falha do que os cotistas.



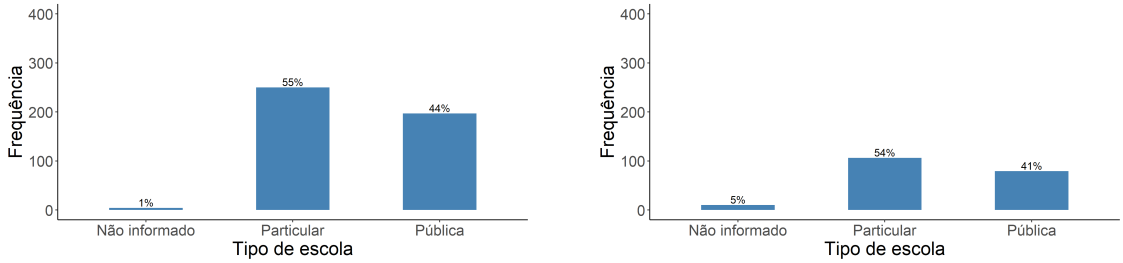
(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 11: Curvas de sobrevivência para a variável Sistema de cotas

As curvas de sobrevivência da Estatística se diferenciam em alguns pontos e de distânciam em outros, principalmente do meio pro final do curso. Já na Matemática há uma diferença importante no final do curso.

4.1.5 Escola

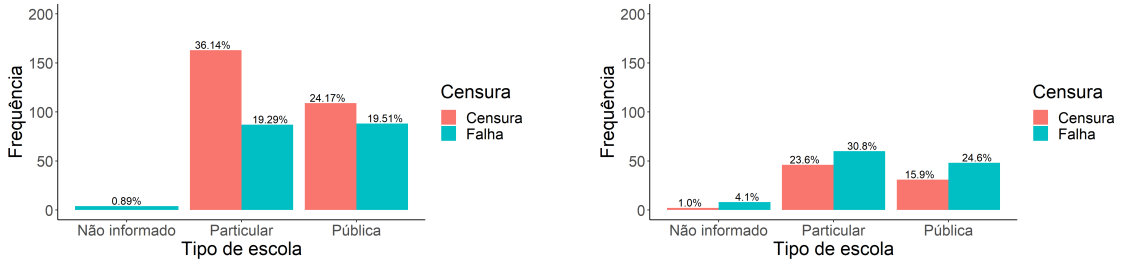
Informa de que tipo de escola é proveniente o aluno, seja "Pública"ou "Particular". Alguns não informaram tal dado.



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 12: Gráficos de barras para a variável Escola

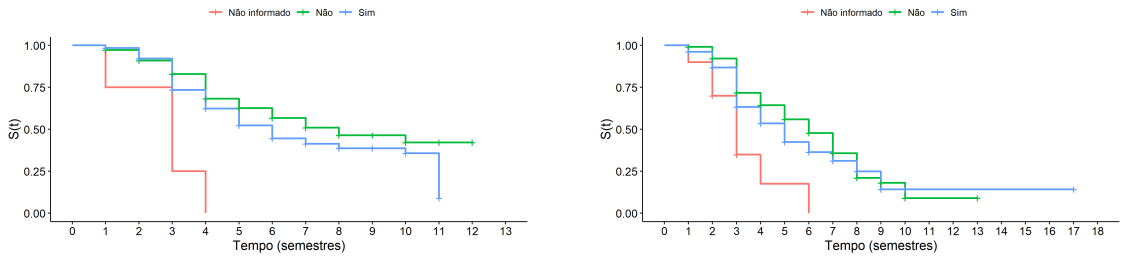
A maior parte dos alunos que cursam Estatística e Matemática advém de escolas particulares, porém essa diferença não é tão grande, cerca de 10 pontos percentuais.



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 13: Gráficos de barras para a variável Escola vs Status

Na Estatística a proporção de falha é bem menor entre os alunos oriundos de escolas particulares, já na matemática, o comportamento é bem semelhante entre alunos provenientes de escolas particulares e públicas.

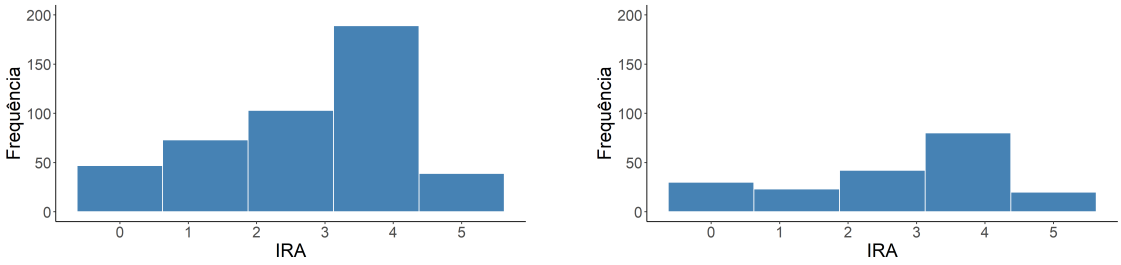


(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 14: Curvas de sobrevivência para a variável Escola

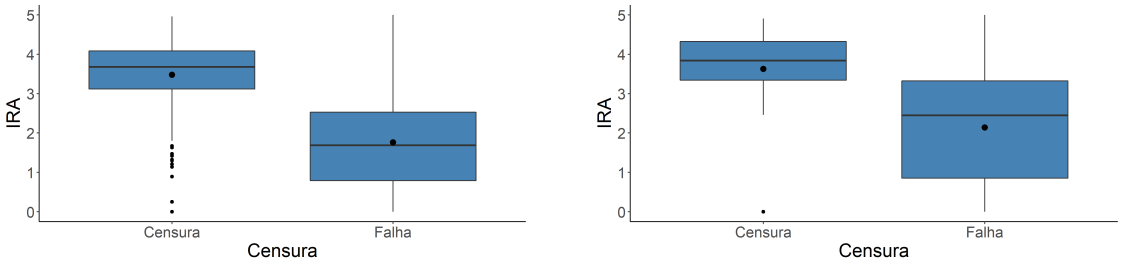
4.1.6 Índice de Rendimento Acadêmico (IRA)

O IRA é um índice que é calculado observando as menções nas disciplinas cursadas e a quantidade de trancamentos de disciplinas, onde quanto maior a menção, e quanto mais créditos maior peso tem aquela disciplina no referido índice, além de considerar se a disciplina é obrigatória ou optativa.



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 15: Histograma para a variável IRA



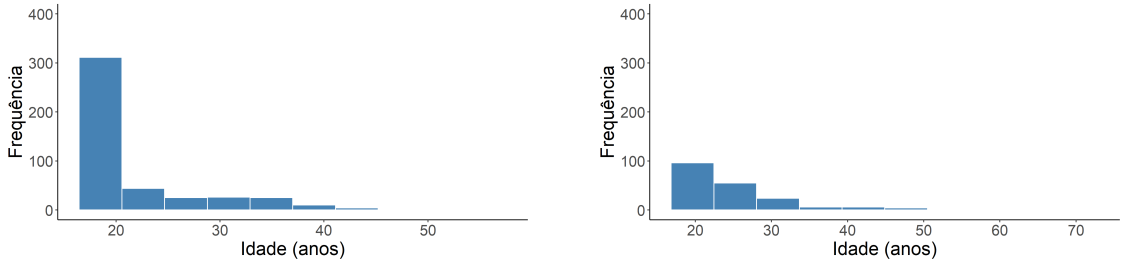
(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 16: Boxplot para a variável IRA vs Status

Os gráficos acima nos fornecem informações interessantes sobre o IRA dos alunos: (1) a distribuição do ira se concentra principalmente nos maiores valores; (2) IRAs de alunos que não evadiram são mais altos do que os IRAs de alunos que evadiram.

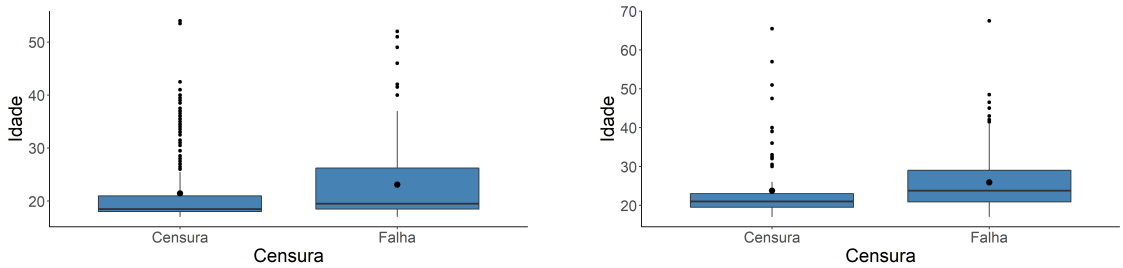
4.1.7 Idade

Idade (em anos) ao ingressar no curso calculada a partir da data de nascimento, presente na base e do período de ingresso.



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 17: Histograma para a variável Idade



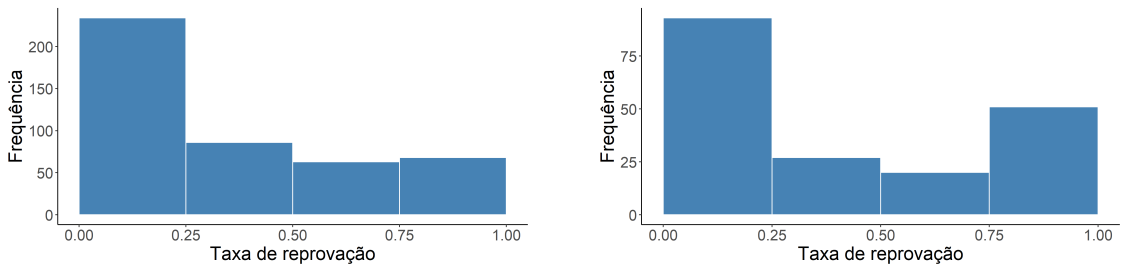
(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 18: Boxplot para a variável Idade vs Status

A concentração de alunos está entre idades mas baixas, até os 30 anos, porém na Matemática há uma presença maior de alunos com idades um pouco maiores que 20 anos.

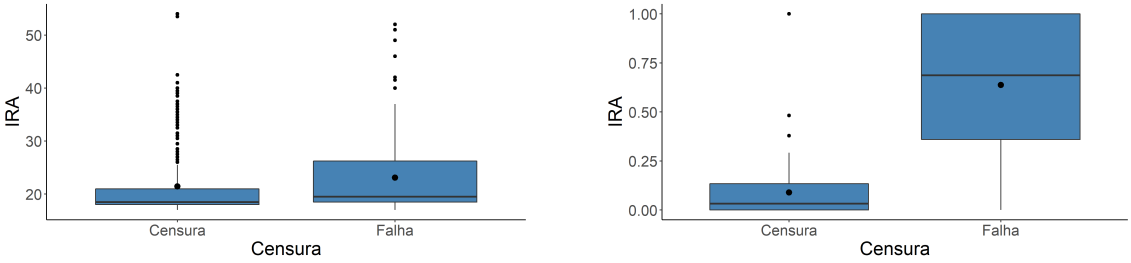
4.1.8 Taxa de reprovação

Taxa de reprovação foi uma variável criada levando em consideração as disciplinas e os créditos reprovados pelo aluno ao longo do curso.



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 19: Histograma para a variável Taxa de reprovação



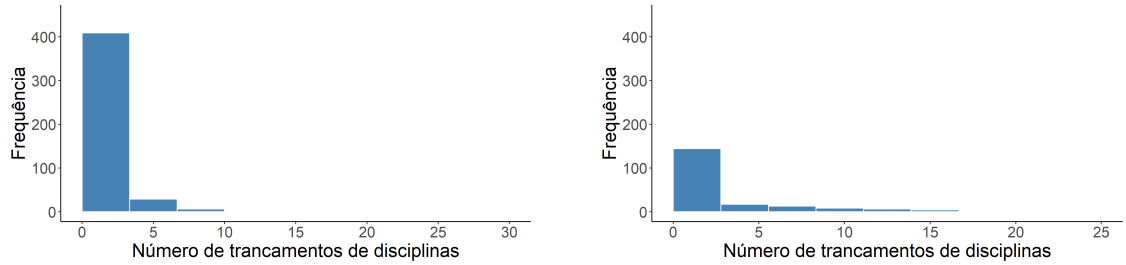
(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 20: Boxplot para a variável Taxa de reprovação vs Status

Os gráficos acima mostram uma grande concentração da taxa de reprovação entorno de 0, o que faz sentido tendo-se em conta a análise do IRA acima, e um comportamento muito díspare, na Matemática, entre os que falharam e os que são censura.

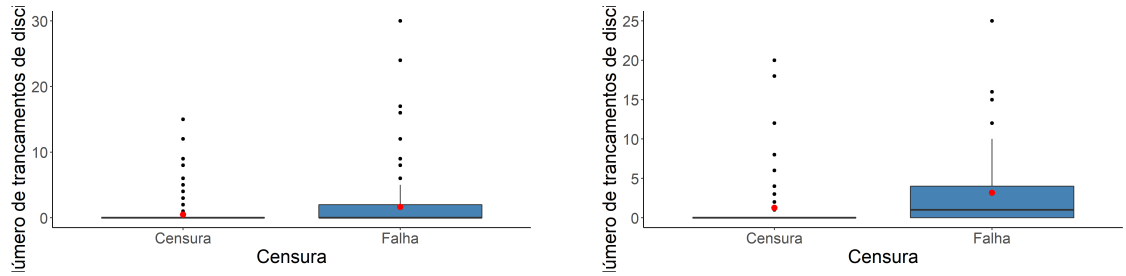
4.1.9 Total de trancamentos

Variável que denota a quantidade de disciplinas trancadas durante o curso ou período de observação.



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 21: Histograma para a variável Total de trancamentos

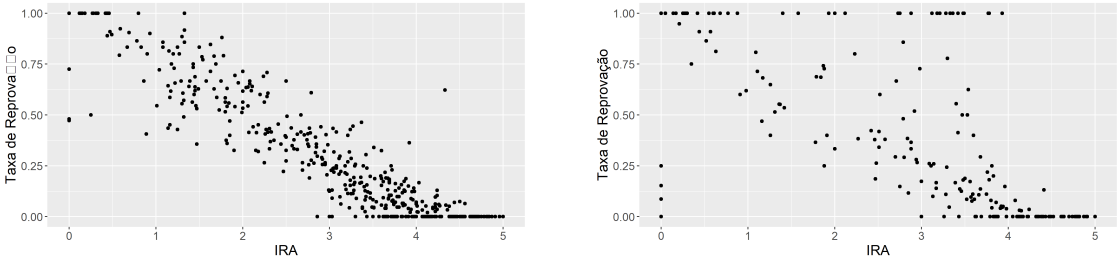


(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 22: boxplot para a variável Total de trancamentos vs Status

Há uma concentração muito grande entorno do 0 da variável total de trancamentos e uma variabilidade difente entre os que falharam e os que não falharam.

4.1.10 Correlação entre Taxa de reprovação e IRA



(a) Bacharelado em Estatística (b) Bacharelado em Matemática

Figura 23: Gráficos de barras para a variável Status

Como é de se esperar, IRA e taxa de reprovação tem uma correlação forte, negativa, ou seja, quanto maior o IRA menor a taxa de reprovação e vice-versa.

4.1.11 Associação entre Sistema de cotas e Escola

É Razoável a suposição de que essas variáveis não sejam independentes, em virtude do tipo de cota que beneficia alunos egressos de escola pública. Para testar tal suposição e confirmá-la ou não, o teste qui-quadrado de independência pode ser utilizado:

Tabela 2: Teste χ^2 da independência entre Sistema de cotas e Escola

Bacharelado	Estatística do teste	p-valor
Estatística	220.54	$< 2,2 \times 10^{-16}$
Matemática	16.748	$4,3 \times 10^{-5}$

Rejeita-se a hipótese nula de independência das variáveis em ambos os casos. Portanto, há evidências para afirmar que existe associação entre Sistema de Cotas e Escola.

4.2 Seleção da distribuição de probabilidade

O primeiro passo para modelar os dados é escolher a função de probabilidade que melhor se ajuste ao conjunto de dados. Para tal as figuras e tabelas a seguir podem fornecer informações importantes.

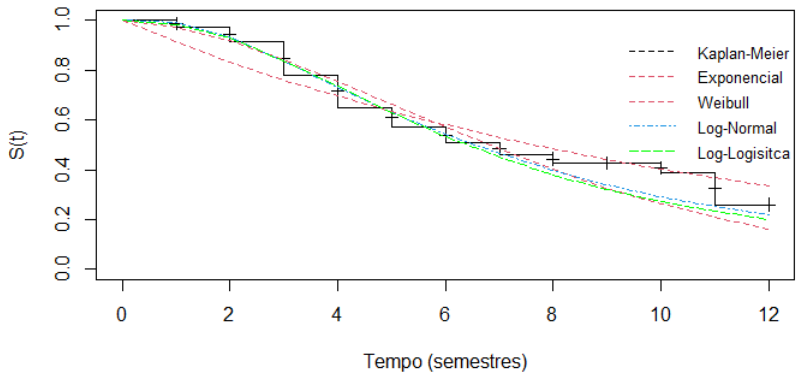


Figura 24: Distribuições de probabilidade - Estatística

Tabela 3: Critérios de informação para Bacharelado em Estatística

Distribuições	AIC	AICc	BIC
Exponencial	1218.99	1219.00	1223.10
Log-logística	1133.99	1134.02	1142.21
Log-normal	1125.81	1125.83	1134.03
Weibull	1154.28	1154.31	1162.50

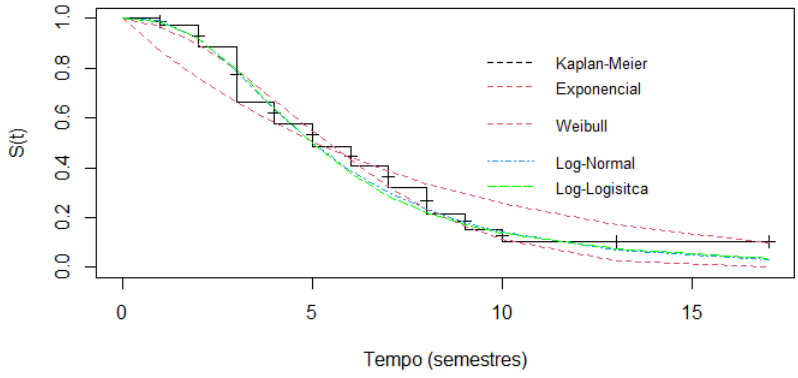


Figura 25: Distribuições de probabilidade - Matemática

Tabela 4: Critérios de informação para Bacharelado em Matemática

Distribuições	AIC	AICc	BIC
Exponencial	696.88	696.90	700.15
Log-logística	627.48	627.54	634.02
Log-normal	624.61	624.67	631.16
Weibull	639.19	639.25	645.74

O gráfico e a tabela acima nos ajudam a escolher uma distribuição adequada aos dados. Com exceção da distribuição Exponencial, todas estão próximas da funçaõ de sobrevivência do estimador de Kaplan-Meier. A tabela nos ajuda a tomar a decisão mais acertada, analisando os valores dos critérios de Akaike e Bayesiano.

Analisando assim, a distribuição log-normal obteve os menores valores para os três critérios, sendo então a distribuição escolhida.

4.3 Modelagem para o curso de Bacharelado em Estatística

Dentre as variáveis possíveis, foram selecionadas 9 potencialmente importantes na descrição do comportamento da variável resposta. Os passos de seleção das variáveis são:

Passo 1: Ajustar todos os modelos com as covariáveis isoladamente. Selecionar aquelas significativasao nível de 25% de significância;

Passo 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;

Passo 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;

Passo 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.

Passo 5 Por fim, verificar a possibilidade de incluir intera¸cões duas a duas entre as covariáveis. O modelo final ser´a composto pelas covariáveis remanescentes no passo 4 e os termos de interação significativos nesta etapa.

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

Tabela 5: Covariáveis individualmente para o curso de Estatística

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
$\beta_{\text{Sistema de cotas sim}}$	-0.1674	0.0950	-1.76	0.078
$\beta_{\text{Sexo masculino}}$	-0.0219	0.0990	-0.22	0.82
$\beta_{\text{Forma de ingresso Outras}}$	0.4522	0.1579	2.86	0.00419
$\beta_{\text{Forma de ingresso PAS}}$	0.2938	0.1264	2.32	0.02015
$\beta_{\text{Forma de ingresso Vestibular}}$	0.3829	0.1162	3.30	0.00098
$\beta_{\text{Escola pública}}$	-0.1583	0.0880	-1.80	0.072
$\beta_{\text{Cursou verão sim}}$	0.7493	0.1233	6.08	$1,2 \times 10^{-9}$
β_{IRA}	0.3529	0.0326	10.83	$< 2 \times 10^{-16}$
β_{Idade}	-0.01264	0.00656	-1.93	0.054
$\beta_{\text{Taxa de reprovação}}$	-1.4124	0.1270	-11.12	$< 2 \times 10^{-16}$
$\beta_{\text{Total de trancamentos}}$	0.00454	0.01442	0.31	0.75

Considerando no nível de 25% de significância, com a exceção de sexo e total de trancamentos, todas as variáveis são significantes. Como constatou-se na análise descritiva, há uma correlação entre taxa de reprovação e IRA e uma associação entre escola e sistema de cotas, portanto, para evitar o problema de multicolinearidade, vamos tomar 4 modelos iniciais diferentes, a saber:

Modelo 1: inclui IRA e sistema de cotas

Modelo 2: inclui IRA e escola

Modelo 3: inclui taxa de reprovação e escola

Modelo 4: inclui taxa de reprovação e sistema de cotas

Tais modelos estão representados na tabela abaixo:

Tabela 6: Covariáveis conjuntamente para o curso de Estatística

Modelo	Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
Modelo 1	$\beta_{\text{Sistema de cotas sim}}$	0.13377	0.08825	1.52	0.12955
	$\beta_{\text{Forma de ingresso Outras}}$	0.57212	0.15153	3.78	0.00016
	$\beta_{\text{Forma de ingresso PAS}}$	0.18503	0.11304	1.64	0.10167
	$\beta_{\text{Forma de ingresso Vestibular}}$	0.42251	0.10191	4.15	$3,4 \times 10^{-5}$
	$\beta_{\text{Cursou verão sim}}$	0.43084	0.11155	3.86	0.00011
	β_{IRA}	0.31157	0.03097	10.06	$< 2 \times 10^{-16}$
	β_{Idade}	-0.01281	0.00635	-2.02	0.04368
Modelo 2	$\beta_{\text{Escola Pública}}$	0.08558	0.08234	1.04	0.29868
	$\beta_{\text{Forma de ingresso Outras}}$	0.49134	0.14599	3.37	0.00076
	$\beta_{\text{Forma de ingresso PAS}}$	0.13370	0.11421	1.17	0.24174
	$\beta_{\text{Forma de ingresso Vestibular}}$	0.36907	0.10105	3.65	0.00026
	$\beta_{\text{Cursou verão sim}}$	0.42412	0.11146	3.81	0.00014
	β_{IRA}	0.31461	0.03144	10.01	$< 2 \times 10^{-16}$
	β_{Idade}	-0.01534	0.00657	-2.33	0.01961
Modelo 3	$\beta_{\text{Sistema de cotas sim}}$	0.19246	0.08973	2.14	0.0320
	$\beta_{\text{Forma de ingresso Outras}}$	0.47456	0.15082	3.15	0.0017
	$\beta_{\text{Forma de ingresso PAS}}$	0.15034	0.11321	1.33	0.1842
	$\beta_{\text{Forma de ingresso Vestibular}}$	0.40518	0.10216	3.97	$7,3 \times 10^{-16}$
	$\beta_{\text{Cursou verão sim}}$	0.43684	0.10954	3.99	$6,7 \times 10^{-16}$
	$\beta_{\text{Taxa de reprovação}}$	-1.31498	0.12578	-10.45	$< 2 \times 10^{-16}$
	β_{Idade}	-0.01208	0.00635	-1.90	0.0571
Modelo 4	$\beta_{\text{Escola Pública}}$	0.13651	0.08243	1.66	0.09770
	$\beta_{\text{Forma de ingresso Outras}}$	0.39994	0.14497	2.76	0.00580
	$\beta_{\text{Forma de ingresso PAS}}$	0.12040	0.11294	1.07	0.28642
	$\beta_{\text{Forma de ingresso Vestibular}}$	0.36778	0.09997	3.68	0.00023
	$\beta_{\text{Cursou verão sim}}$	0.43966	0.10990	4.00	$6,3 \times 10^{-5}$
	$\beta_{\text{Taxa de reprovação}}$	-1.27910	0.12269	-10.43	$< 2 \times 10^{-16}$
	β_{Idade}	-0.01554	0.00659	-2.36	0.01832

Considerando um nível de significância de 10%, todas as variáveis são significativas nos modelo 1 e 2, à exceção de cotas e escola, respectivamente, portando os modelos são iguais ao final dessa etapa. Para os modelos 3 e 4, todas as variáveis foram significativas, inicialmente. No processo de retirar as variávies uma a uma e verificar, por meio do p-valor, sua significância, Cotas e Escola foram de fato não significativas nos modelos 1 e 2, respectivamente, Idade foi não significativa no modelo 3 e Escola foi não significativa no modelo 4, a 10%. Logo, sai Escola e cota dos modelos 1 e 2, respectivamente, Idade sai do modelo 3 e Escola sai do modelo 4.

Retira-se agora as covariáveis restantes, uma a uma, e verifica-se a sua signi-

ficância através do TRV. Testaremos também a interação entre as variáveis IRA e taxa de reprovação com a variável sistema de cotas.

Tabela 7: Seleção de covariáveis significativas para a Estatística pelo TRV

Modelo	Hipótese nula	TRV	P-valor
Modelos 1 e 2	$\beta_{\text{Forma de ingresso}} = 0$	18.82266	$1,43 \times 10^{-5}$
	$\beta_{\text{Sistema de cotas}} = 0$	6.829218	0.008967851
	$\beta_{\text{Cursou verão}} = 0$	13.50134	0.000238393
	$\beta_{\text{Idade}} = 0$	2.904878	0.08831195
	$\beta_{\text{IRA}} = 0$	112.2734	0
	$\beta_{\text{IRA*Cotas}} = 0$	4.149667	0.0416427
Modelo 3	$\beta_{\text{Forma de ingresso}} = 0$	17.22993	$3,31 \times 10^{-5}$
	$\beta_{\text{Sistema de cotas}} = 0$	12.01052	0.0005290099
	$\beta_{\text{Cursou verão}} = 0$	14.73672	0.0001236153
	$\beta_{\text{Taxa de reprovação}} = 0$	127.2887	0
	$\beta_{\text{Taxa de reprovação*Sistema de cotas}} = 0$	6.233522	0.01253542
Modelo 4	$\beta_{\text{Forma de ingresso}} = 0$	12.59487	0.0003868068
	$\beta_{\text{Cursou verão}} = 0$	13.30662	0.0002644709
	$\beta_{\text{Idade}} = 0$	3.528982	0.0603048
	$\beta_{\text{Taxa de reprovação}} = 0$	109.1182	0

Temos portanto que todas as variáveis restantes em todos modelos são significativas, pois rejeitamos todas as hipóteses nulas a 10% de significância.

Chega-se então aos modelos candidatos.

4.3.1 Modelos candidatos

Temos 4 possíveis modelos, quais sejam:

Modelos 1 e 2: Forma de ingresso + Sistema de cotas + Cursou verão + Idade + IRA + IRA*Sistema de cotas

Modelo 3: Forma de ingresso + Sistema de cotas + Cursou verão + Taxa de reprovação + Taxa de reprovação*Sistema de cotas

Modelo 4: Forma de ingresso + Cursou verão + Idade + Taxa de reprovação

Uma das formas de rejeitar modelos inadequados é a análise gráfica dos resíduos de Cox-Snell. As figuras que se seguem apresentam os referidos resíduos, que devem seguir a distribuição exponencial padrão caso o modelo log-normal para o tempo de falha seja adequado para os dados.

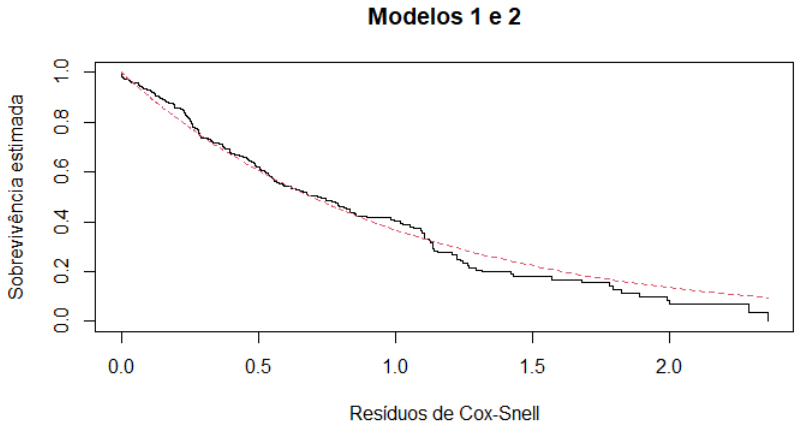


Figura 26: Resíduos de Cox-Snell para os modelos 1 e 2

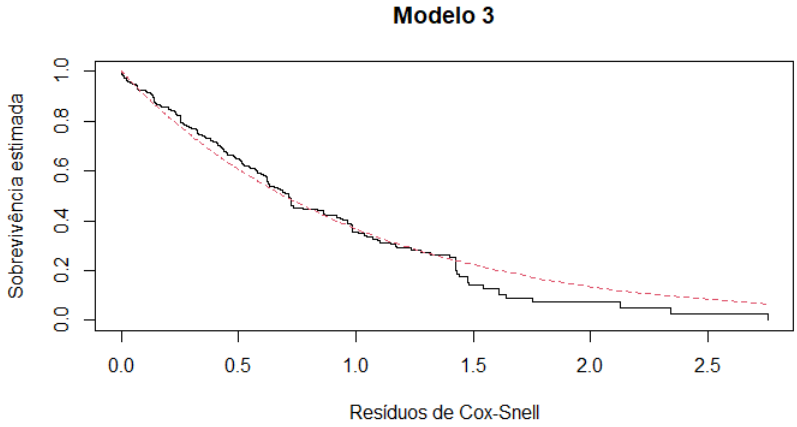


Figura 27: Resíduos de Cox-Snell para o modelo 3

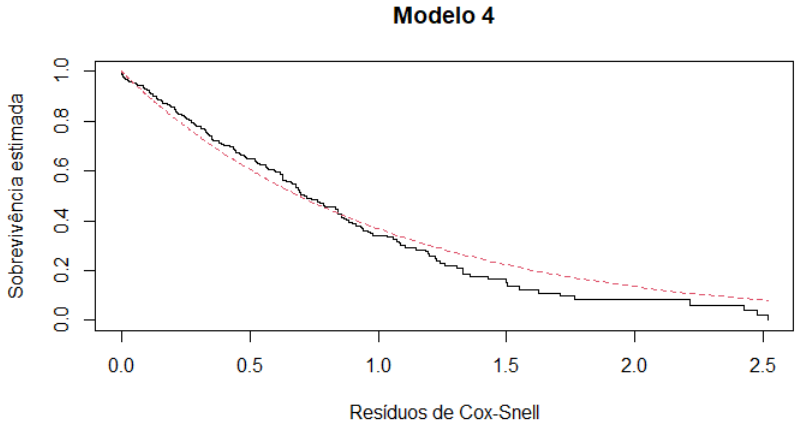


Figura 28: Resíduos de Cox-Snell para o modelo 4

As figuras acima nos permitem concluir que os 3 modelos são adequados para modelar nossos dados. Como critério objetivo para a escolha de um modelo específico, utilizar-se-á os critérios de Akaike, de Akaike corrigido e Bayesiano.

Tabela 8: Critérios de informação para os modelos candidados do curso de Bacharelado em Estatística

Modelos	AIC	AICc	BIC
Modelos 1 e 2	965.11	965.52	1002.11
Modelo 3	955.21	955.53	988.10
Modelo 4	961.69	961.94	990.47

Pelas medidas acima, o modelo 3 com Forma de ingresso, Sistema de cotas, Cursou verão, Taxa de reprovação e a interação entre Taxa de reprovação e Sistema de cotas, foi escolhido como modelo final para a Estatística.

Tabela 9: Estimativas para o modelo final da Estatística

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
$\beta_{\text{Intercepto}}$	1.8853	0.1055	17.88	$< 2 \times 10^{-16}$
$\beta_{\text{Sistema de cotas sim}}$	0.5828	0.1758	3.32	0.00091
$\beta_{\text{Forma de ingresso Outras}}$	0.4119	0.1408	2.92	0.00345
$\beta_{\text{Forma de ingresso PAS}}$	0.2330	0.1084	2.15	0.03166
$\beta_{\text{Forma de ingresso Vestibular}}$	0.4113	0.1009	4.08	$4,6 \times 10^{-5}$
$\beta_{\text{Cursou verão sim}}$	0.4167	0.1104	3.77	0.00016
$\beta_{\text{Taxa de reprovação}}$	-1.1425	0.1434	-7.97	$1,6 \times 10^{-16}$
$\beta_{\text{Taxa de reprovação*Sistema de Cotas}}$	-0.6546	0.2641	-2.48	0.01320
Log(Scale)	-0.4996	0.0542	-9.22	$< 2 \times 10^{-16}$

Os coeficientes estimados desse modelo possuem as seguintes interpretações:

1. Alunos que ingressaram na universidade por meio de cotas têm maior probabilidade de sobreviver do que alunos que não entraram por cotas;

2. Em relação às formas de ingresso, tendo o SISU como referência, todas as outras formas de ingresso tiveram efeito positivo, ou seja, todos os alunos que ingressaram por elas têm maior probabilidade de sobrevivência;

3. Os alunos que cursaram verão têm maior probabilidade de sobreviver dos que alunos que não cursaram verão;

4. O valor negativo do coeficiente da variável taxa de reprovação indica que alunos que têm uma taxa de reprovação menor possuem maior probabilidade de sobrevivência em relação aos alunos com taxa de reprovação maior;

5. A interação entre taxa de reprovação e sistema de cotas nos indica que os alunos com menor taxa de reprovação e que ingressaram por cotas têm maior probabilidade de

sobreviver, ou seja, de não evadir.

4.3.2 Disciplinas cursadas no Verão

A variável cursou verão se mostrou significativa no modelo escolhido, portanto é importante saber também quais disciplinas foram cursadas por esses alunos nos verões que foram observados. Para tal, tem-se abaixo a tabela que informa as disciplinas cursadas no verão por alunos observados pelo estudo e a quantidade de alunos desse universo.

Tabela 10: Disciplinas cursadas no verão na Estatística

Disciplinas	Quantidade de alunos
Administração de Sistemas de Informação	4
Algoritmos e Programação de Computadores	6
Cálculo 1 - Semipresencial	2
Cálculo 3	4
Cálculo Financeiro	4
Cálculo Numérico	7
Física 1	9
Física 1 Experimental	2
Formação Econômica do Brasil	3
Fundamentos Matemáticos da Física A	1
Inglês Instrumental 1	13
Inglês Instrumental 2	3
Introdução à Álgebra Linear	62
Introdução à Econometria	9
Introdução à Economia	11
Introdução à Probabilidade	4
Metodologia Científica Aplicada	1
Pesquisa Operacional	1

Vê-se uma grande variedade de disciplinas, optativas, obrigatórias, enfim. A que teve mais Matrículas foi Introdução à àlgebra linear, disciplina obrigatória, que é pré-requisito para outras disciplinas obrigatórias no curso da Estatística. Logo, o cursar essa disciplina no verão, não prejudicando o andamento do curso e as disciplinas que dependem dela, pode ajudar o aluno a não evadir, principalmente no começo do curso.

4.4 Modelagem para o curso de Bacharelado em Matemática

A tabela seguinte mostra os resultados para o passo 1 (modelo com apenas uma variável).

Tabela 11: Covariáveis individualmente para o curso de Matemática

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
$\beta_{\text{Sistema de cotas sim}}$	0.1348	0.1557	0.87	0.39
$\beta_{\text{Sexo masculino}}$	-0.1992	0.1233	-1.62	0.11
$\beta_{\text{Forma de ingresso Outras}}$	-0.4301	0.3081	-1.40	0.16
$\beta_{\text{Forma de ingresso PAS}}$	-0.1121	0.3845	-0.29	0.77
$\beta_{\text{Forma de ingresso Vestibular}}$	-0.3825	0.3115	-1.23	0.22
$\beta_{\text{Escola pública}}$	-0.0660	0.1026	-0.64	0.52
$\beta_{\text{Cursou verão sim}}$	0.4740	0.1546	3.07	0.0022
β_{IRA}	0.2303	0.0350	6.57	5×10^{-11}
β_{Idade}	0.00226	0.00625	0.36	0.72
$\beta_{\text{Taxa de reprovação}}$	-1.2356	0.1156	-10.7	$< 2 \times 10^{-16}$
$\beta_{\text{Total de trancamentos}}$	0.0057	0.0114	0.50	0.62

Considerando o nível de 25% de significância, as variáveis sistema de cotas, Escola, Idade e total de trancamentos não são significativas. Desse modo, ajusta-se o modelo sem as referidas variáveis. Como visto anteriormente, IRA e taxa de provação possuem correlação, portando não podem entrar em conjunto no mesmo modelo, o que nos dá duas possibilidades de modelos iniciais para o próximo passo:

Modelo 1: inclui o IRA;

Modelo 2: inclui a Taxa de reprovação.

Seguem as estimativas dos modelos ajustados:

Tabela 12: Covariáveis conjuntamente para o curso de Matemática

Modelo	Parâmetro	Estimativa	E.P.	Estatística do Teste	P-valor
Modelo 1	$\beta_{\text{Forma de ingresso Outras}}$	-0.4566	0.2864	-1.59	0.111
	$\beta_{\text{Forma de ingresso PAS}}$	-0.0851	0.3562	-0.24	0.811
	$\beta_{\text{Forma de ingresso Vestibular}}$	-0.0455	0.2919	-0.16	0.876
	$\beta_{\text{Cursou verão sim}}$	0.3321	0.1407	2.36	0.018
	β_{IRA}	0.2679	0.0371	7.22	$5,1 \times 10^{-13}$
	$\beta_{\text{Sexo masculino}}$	-0.0517	0.1103	-0.47	0.639
Modelo 2	$\beta_{\text{Forma de ingresso Outras}}$	-0.2642	0.2556	-1.03	0.30
	$\beta_{\text{Forma de ingresso PAS}}$	-0.1489	0.3183	-0.47	0.64
	$\beta_{\text{Forma de ingresso Vestibular}}$	-0.0752	0.2591	-0.29	0.77
	$\beta_{\text{Cursou verão sim}}$	0.0694	0.1287	0.54	0.59
	$\beta_{\text{Taxa de reprovação}}$	-1.2148	0.1192	-10.19	$< 2 \times 10^{-16}$
	β_{Sexo}	-0.0901	0.1019	-0.88	0.38

Para o modelo 1 apenas cursou verão e taxa de reprovação foi significativa a 10% e para o modelo 2 apenas taxa de reprovação. Foram, portanto, ajustados os modelos retirando as demais variáveis, uma a uma, e de fato foram não significativas.

Retirou-se as variáveis restantes em cada modelo, uma a uma, e realizou-se o TRV para verificar sua significância.

Tabela 13: Seleção de covariáveis significativas para a Estatística pelo TRV

Modelo	Hipótese nula	TRV	P-valor
Modelos 1	$\beta_{\text{Cursou verão}} = 0$	6.506812	0.0107462
	$\beta_{\text{IRA}} = 0$	40.74529	$1,7 \times 10^{-10}$
Modelo 2	$\beta_{\text{Taxa de reprovação}} = 0$	50.16054	$1,4 \times 10^{-12}$

Tem-se que o TRV rejeita a hipótese nula em todos os casos, ou seja, as covariáveis restantes são de fato significantes.

Tentou-se também, a partir desse modelo, incluir alguma das variáveis que não entraram nos passos anteriores.

Tabela 14: Tentativa de inclusão pelo TRV das variáveis que saíram nos passos anteriores

Modelo	Hipótese nula	TRV	P-valor
Modelos 1	$\beta_{\text{Forma de ingresso}} = 0$	-16.33249	1
	$\beta_{\text{Sexo}} = 0$	-2.071452	1
Modelo 2	$\beta_{\text{Forma de ingresso}} = 0$	-116.7477	1
	$\beta_{\text{Sexo}} = 0$	-112.2899	1
	$\beta_{\text{Cursou verão}} = 0$	-112.2899	1

A 10% de significância o TRV não rejeitou a hipótese nula em nenhum caso.

4.4.1 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: Cursou verão + IRA

Modelo 2: Taxa de reprovação

A 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.

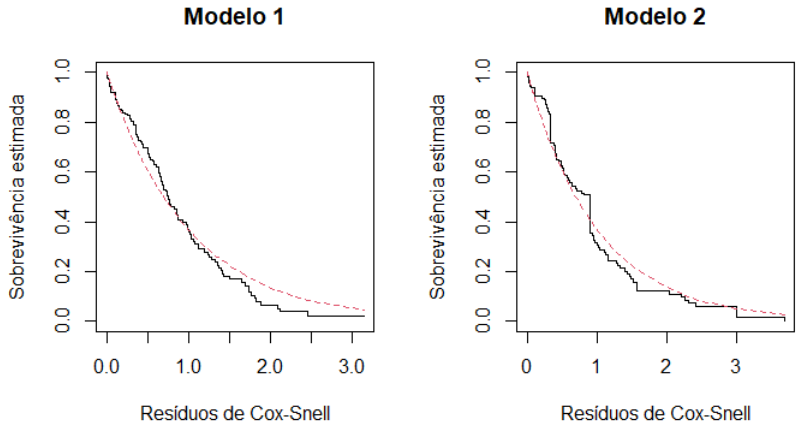


Figura 29: Resíduos de Cox-Snell

A partir da figura acima 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, AICc e BIC serão utilizados para este propósito.

Tabela 15: Critérios de informação para os modelos candidatos do curso de Bacharelado em Matemática

Modelos	AIC	AICc	BIC
Modelo 1	562.62	562.75	572.38
Modelo 2	503.21	503.27	509.72

Apesar do modelo 2 apresentar valores menores para os critérios, e como a diferença não é muito grande (Cerca e 10%), optou-se pelo modelo 1 tendo em vista a informação que ele contém a mais que o modelo preterido e o fato de o modelo 1 possuir apenas uma variável.

Tabela 16: Estimativas para o modelo final da Matemática

Parâmetro	Estimativa	Erro Padrão	Estatística do Teste	P-valor
$\beta_{\text{Intercepto}}$	0.9812	0.1004	9.77	$< 2 \times 10^{-16}$
$\beta_{\text{Cursou verão sim}}$	0.3678	0.1488	2.47	0.013
β_{IRA}	0.2248	0.0359	6.27	3.6×10^{-10}
Log(Scale)	-0.5410	0.0664	-8.14	$3,9 \times 10^{-16}$

Com isso, os coeficientes estimados possuem as seguintes interpretações:

1. Alunos que cursaram verão têm maior chance de sobreviver;
2. Alunos com maior IRA tem maior probabilidade de sobrevivência.

4.4.2 Disciplinas cursadas no Verão

A variável cursou verão se mostrou significativa no modelo escolhido, portanto é importante saber também quais disciplinas foram cursadas por esses alunos nos verões que foram observados. Para tal, tem-se abaixo a tabela que informa as disciplinas cursadas no verão por alunos observados pelo estudo e a quantidade de alunos desse universo.

Disciplinas	Quantidade de alunos
Álgebra Linear 2	9
Algoritmos e Programação de Computadores	1
Cálculo Financeiro	4
Contabilidade Geral 1	3
Didática Fundamental	2
Física 1	1
Física 2	8
Física 3	5
Fundamentos de Desenvolvimento e Aprendizagem	3
Fundamentos Matemáticos da Física A	11
Introdução à Administração	1
Introdução à Álbegra Linear	4
Introdução à Economia	1
Introdução à Filosofia	1
Metodologia de Observação	2
Organização da Educação Brasileira	2
Probabilidade e Estatística	4
Psicologia da Sexualidade	2
Seminário de Tópicos em Matemática aplicada	3
Termodinâmica	8

Observa-se aqui também uma variedade de disciplinas e de tipos de disciplinas cursadas pelos alunos. Destaca-se com mais matrículas Fundamentos Matemáticos da Física A, Álgebra Linear 2 e Termodinâmica que certamente ajudam os alunso que as cursaram a deixar o semestre com menos disciplinas, possibilitando maior aprovação nas demais disciplinas, colaborando com a manutenção do curso.

5 Conslusão

O Presente trabalho teve por objetivo avaliar a relação entre o tempo até um aluno evadir dos cursos de Bacharelado em Estatística ou Matemática da Universidade de Brasília e diversos fatores que possivelmente influenciariam nesse tempo.

A análise descritiva inicial foi de suma importância para fornecer uma direção a seguir quais alternativas tentar. Os indícios que forneceu a análise descritiva foram confirmados pela análise de sobrevivência como a influência de IRA, cursou verão, entre outras variáveis.

Para a Estatística, chegamos a conclusão que ser ou não cotista, a forma de ingresso na universidade, a quantidade de créditos reprovados, entre outros fatores fazem diferença na probabilidade de sobrevivência do aluno, ou seja, tais fatores influenciam na chance de um determinado aluno evadir do curso. A interação entre sistema cotas e taxa de reprovação resultou um coeficiente negativo, isto é, os alunos que são cotistas e tem uma taxa de reprovação baixa, tem maior chance de sobreviver dos que não atendem à essas condições.

Já para a Matemática, encontrou-se menos fatores: taxa de reprovação e cursou verão, fato que pode ser decorrência do menor número de dados e outras peculiaridades do curso de Bacharelado em Matemática, como as formas de ingresso serem em sua maioria diferente de "PAS", "SISU"e "Vestibular", tendo uma presença grande de "dupla habilitação"e "mudança de curso".

Os métodos para escolha do modelo adequado levam em conta a parcimônia de cada modelo, principalmente para o caso da Estatística, onde foi escolhido o modelo com menor valor dos critérios de informação. Para a Matemática, as peculiaridades do curso aliadas à menor quantidade de observações na nossa base em relação à Estatística nos levam a escolher o modelo com valores para os critérios de informação maior, dando também uma peso para a quantidade de variáveis e levando em consideração o objetivo do estudo, que é investigar a influência de fatores sobre a evasão escolar.

Sugere-se para trabalhos futuros a inclusão de variáveis de cunho mais pessoal, como local onde mora, se mora sozinho ou não, meios de transporte para verificar a influencia nesse acontecimento da evasão.

6 Referências

FILHO, R. B. S.; ARAÚJO, R. M. de L. Evasão e abandono escolar na educação básica no brasil: fatores, causas e possíveis Consequências. Educação por escrito, v. 8, n. 1, p.35 - 48, 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.

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.

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

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

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