Universidade de Brasília
Departamento de Estatística

Estudo da Evasão nos Curso de Ciência da Computação da Universidade de Brasília:
Uma aplicação a modelos de Análise de Sobrevivência

Mathews de Noronha Silveira Lisboa

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
2021

Mathews de Noronha Silveira Lisboa

Estudo da Evasão nos Curso de Ciência da Computação da Universidade de Brasília: Uma aplicação a modelos de Análise de Sobrevivência

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
2021

Resumo

Este trabalho teve como objetivo identificar fatores que levam alunos de graduação em Ciência da Computação do Instituo de Ciências Exatas da Universidade de Brasília a cometer evasão. O conjunto de dados é composto por 879 observações e 16 variáveis com as características dos alunos que se matricularam no curso durante o período de 2014/1 a 2019/2. Além disso, foi selecionado um segundo banco de dados a partir do primeiro que corresponde a 449 observações e 16 variáveis, coletando alunos que se matricularam no curso durante o período de 2014/1 a 2016/2, afim de avaliar os fatores que levam a evasão nos primeiros três semestres. Para avaliar o tempo de evasão dos alunos foi utilizado a metodologia de análise de sobrevivência, visto que é uma técnica que permite a inclusão de informação a respeito do tempo de falha (cometer evasão) e também de informações parciais no caso do tempo de censura (não cometer evasão). O modelo de regressão Log-Normal foi utilizado para identificar a influência de covariáveis no tempo de sobrevivência dos alunos para ambos os bancos. Embora a divisão de bancos tenha sido feita, a interpretação do efeito das covariáveis foi muito semelhante.

Palavras-chave: Evasão; Educação; Ensino superior; Análise de sobrevivência; Censura; Ciência da computação; Distribuição Log-Normal; Modelo de regressão

Abstract

This work aimed to identify factors that lead undergraduate students in Computer Science of the Instituo de Ciências Exatas of the Universidade de Brasília to commit evasion. The dataset is composed of 879 observations and 16 variables with the characteristics of students who enrolled in the course during the period from 2014/1 to 2019/2. In addition, a second database was selected from the first one that corresponds to 449 observations and 16 variables, collecting students who enrolled in the course during the period from 2014/1 to 2016/2, in order to evaluate the factors that lead to dropout in the first three semesters. The survival analysis methodology was used to evaluate the students' dropout time, since it is a technique that allows the inclusion of information about the failure time (committing dropout) and also partial information in the case of the censored time (not committing dropout). The Log-Normal regression model was used to identify the influence of covariates on student survival time for both banks. Although the division of banks was made, the interpretation of the effect of covariates was very similar.

Keywords: School dropout; Education; University education; Survival analysis; Censured; Computer science; Log-Normal distribution; Regression model

Lista de Figuras

*

Lista de Tabelas

*

Sumário

1 Introdução

A evasão escolar é um problema crônico no Brasil que atinge a educação do país em todos os níveis. O Brasil atualmente tem a terceira maior taxa de evasão escolar básica entre os 100 países com maior IDH (Índice de Desenvolvimento Humano) (FILHO; ARAÚJO, 2017). Por mais que seja um dos maiores desafios para a democratização da educação no Brasil, ainda há uma confusão do termo evasão escolar com abandono escolar, como visto em Santos e Albuquerque (2019). A evasão escolar é caracterizada quando o aluno que está matriculado em um tempo t, não está mais matriculado no tempo t+1. Enquanto o abandono é caracterizado por um aluno que não comparece mais no ambiente de ensino por um período de tempo t, mas ainda continua matriculado no t+1 (SANTOS; ALBUQUERQUE, 2019).

O ambiente de ensino superior brasileiro também sofre muito com o problema de evasão, sendo considerado um prejuízo social e perda de recursos de acordo com Lobo (2012). O Brasil tem um dos maiores índices de evasão escolar no ensino superior, considerando instituições públicas e privadas, chegando a casos de metade da turma evadir antes da conclusão do curso. Mesmo com o aumento de ingressantes jovens, entre 18 a 24 anos, no ensino superior, ainda é um índice baixo de ingressantes levando em conta os países emergentes como visto em Andrade (2012). Dito isso, nota-se que é fundamental entender as razões que levam a não conclusão de um ensino superior no Brasil, dado o custo social que gera uma grande parcela de jovens não formados.

Sendo assim, este trabalho tem como objetivo estudar a evasão nos cursos de graduação do Instituto de Ciências Exatas da UnB, especificamente o curso de Ciência da Computação. Uma forma de estudar a evasão no ensino superior é medir o tempo até a evasão, como visto em Junior, Silveira e Ostermann (2012). Para estudar o comportamento de dados dessa natureza, na estatística utiliza-se a metodologia de Análise de Sobrevivência. Essa metodologia define a variável resposta como o tempo até a ocorrência de um evento de interesse, também conhecido como tempo de falha. Porém, para casos que não se observa a ocorrência do evento de interesse, o tempo é considerado tempo de censura. As censuras são observações que, por diversas razões não chegaram a experimentar o evento de interesse, porém a Análise de Sobrevivência permite que essa informação parcial seja utilizada.

O tempo de falha, normalmente, é de natureza contínua. Porém, há ocasiões que devido a forma de registro dessa observação a variável pode tomar uma natureza discreta. No contexto de evasão, a forma de registro semestral faz com que o tempo de falha seja justamente dessa natureza discreta, visto que observações que teriam tempos de falha distintos acabam sendo englobados na mesma categoria, ou seja, mesmo semestre. Uma

forma usual de se trabalhar é supor que os dados são contínuos e utilizar de modelos contínuos para ajustar os dados. Essa técnica, porém, nem sempre apresenta resultados satisfatórios, como explica Nakano e Carrasco (2006). Outra metodologia utilizada quando os tempos apresentam-se de forma discreta é utilizar a abordagem de discretização de modelos contínuos, como proposto em Santos (2017). Neste trabalho serão avaliadas as duas metodologias.

É sabido que diversos fatores tais como: curso, turno do curso, idade entre outros fatores podem influenciar na ocorrência da evasão. Par avaliar quais fatores aumentam ou diminuem as chances de um aluno evadir, este trabalho também propõe construir um modelo de regressão. Todas as análises estatísticas, cálculos matemáticos e modelagens serão realizadas no *software* estatístico R (R Core Team, 2020), em sua versão 4.0.0 de 2020.

2 Revisão de Literatura

2.1 Análise de Sobrevivência

A análise e sobrevivência consiste em um conjunto de técnicas estatísticas com finalidade de estudar dados em que a variável resposta é o tempo até a ocorrência de um determinado fenômeno de interesse, chamado **tempo de falha**. Além disso, uma das características principais em análise de sobrevivência é a presença de informações incompletas ou parciais, chamadas de censuras. Mesmo que o indivíduo não tenha experimentado o evento de interesse, não se deve excluí-lo da análise, pois o mesmo contém informações sobre o tempo decorrido até o momento da censura e sua omissão será prejudicial para estimativas, podendo torná-las viciadas.

A seguir será apresentado diferentes tipos de censura:

2.1.1 Censura

A principal característica de dados de sobrevivência é a presença de censura, que são observações incompletas ou parciais que acontecem quando, por alguma razão, o acompanhamento do indivíduo em estudo foi interrompido. É também considerado censura casos em que o evento de interesse não acontece até o término do estudo, por exemplo, a pessoa não morreu devido ao câncer no período em que foi realizado o estudo. Mas, mesmo sendo incompletas, as observações censuradas trazem algumas informações dos indivíduos e, por isso, devem ser incluídas na análise estatística. Se retiradas, os resultados do estudo podem ser viciados.

Dessa forma, existe a necessidade da introdução de uma variável extra na análise, que indica se o valor do tempo de sobrevivência de um determinado indivíduo é, ou não, um tempo de falha. Essa variável, geralmente representada por δ, é conhecida como variável indicadora de censura e é expressa por:

$$\delta_i = \begin{cases} 1, & \text{se o i-ésimo indivíduo falhou.} \\ 0, & \text{se o i-ésimo indivíduo foi censurado.} \end{cases} \tag{2.1.1}$$

Dito isso, é importante ressaltar que a censura pode ser dividida em três subgrupos distintos de acordo com suas características:

- **Censura do Tipo I:** O estudo será terminado após um período de tempo fixo, e ao final dele, uma ou mais observações em estudo não falharam. Esse tempo deve ser determinado antes do início do estudo.

- **Censura do Tipo II:** O estudo termina após ocorrer o evento de interesse em um número fixo de indivíduos. O número de falhas deve ser determinado antes do início do estudo.

- **Censura aleatória:** São todos os casos em que as observações não experimentam o evento de interesse por motivos não controláveis.

2.1.2 Função de Probabilidade

Nessa subseção será apresentada a função de probabilidade, considerando que a variável T seja contínua ou discreta. A função de probabilidade pode ser denotada como $f(t)$ no caso contínuo ou $p(t)$ no caso discreto e pode ser interpretada como a probabilidade de um individuo experimentar o evento de interesse no intervalo de tempo $[t, t+t]$.

A função densidade de probabilidade, para uma variável aleatória T que seja contínua, é definida como o limite da probabilidade de um indivíduo experimentar o evento de interesse em um intervalo de tempo $[t; t+\Delta t)$ por unidade de Δt (comprimento do intervalo), ou simplesmente por unidade de tempo.

É expressa por :

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, \tag{2.1.2}$$

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

Se considerarmos uma variável aleatória X em um um espaço de probabilidade $(\Omega, \mathcal{F}, P)$ é uma função real definida no espaço Ω tal que $[X \leq x] \in \mathcal{F}, \forall x \in \mathbb{R}$ visto em Santos (2017).

Dessa forma, X uma variável aleatória contínua, a variável discreta obtida por $T = [X]$, sendo $[X]$ a parte inteira de X. Assim a distribuição probabilidade para T no caso discreto é definida por:

$$\begin{aligned} p(t) &= P(T = t) \\ &= P(t \leq X < t + 1) \\ &= P(X < t1) - P(X \leq t) \\ &= F_x(t + 1) - F_x(t), \end{aligned} \tag{2.1.3}$$

Em que $F_x(t)$ é a função de probabilidade acumulada para a variável T definida.

2.1.3 Função de Sobrevivência

A função de sobrevivência, denotada por $S(t)$, é definida para uma variável aleatória T contínua, como a probabilidade de um indivíduo sobreviver a um tempo t, ou seja, a probabilidade de um indivíduo não falhar até um certo tempo t. Ela é expressa por:

$$S(t) = P(T \geq t) = \int_t^{\infty} f(x)dx. \tag{2.1.4}$$

Deve-se considerar uma das principais função utilizadas em análise de sobrevivência e tem relação com a função de distribuição acumulada $F(t) = P(T \leq t)$, ou seja,

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t).$$

Sendo que $S(t)$ é uma função monótona decrescente e contínua. Considerando que a variável T é contínua e não negativa, no tempo zero tem-se $S(0) = 1$, o que não acontece quando a variável T é discreta, podendo ter ocorrido a falha no tempo zero e $S(0) < 1$, sendo uma das principais características de dados discretos.

Outro ponto a se considerar é a hipótese de que todos os indivíduos irão falhar durante o estudo; nesse caso a probabilidade de sobrevivência por um período de tempo muito grande é igual a 0, isto é, para $\lim_{t\to\infty} S(x) = 0$

Quando a variável T é discreta, ou seja, $t = 0, 1, 2...$, a função de sobrevivência discreta é representada como:

$$S(t) = P(T \geq t) = \sum_{k=t+1}^{\infty} P(T = k). \tag{2.1.5}$$

2.1.4 Função de Risco

A função de risco é definida como o limite da probabilidade de um indivíduo falhar no intervalo de tempo $[t; \Delta t)$. Representada por $h(t)$, é também chamada de taxa de falha. E dito isso, segundo Carvalho et al. (2011), é importante dizer que $h(t)$ se trata de uma taxa,e não uma probabilidade. Sendo assim assumindo que T é uma variável contínua e que este mesmo indivíduo sobreviveu até o tempo t, dividida pelo comprimento do intervalo e é representada por:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.1.6}$$

Além disso, $h(t)$ está correlacionada com $f(t)$ e $S(t)$ da seguinte forma:

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

Considerando o caso discreto,de acordo com Fernandes (2013), a função de risco é igual a zero, exceto nos pontos em que pode ocorrer uma falha. Em adição a isso, a função de risco é definida no intervalo $0 \leq h(t) \leq 1$ e pode ser expressa por:

$$h(t) = P(t = t | T \geq t) = \frac{P(T = t)}{P(\geq t)} = \frac{P(T = t)}{P(T > t) + P(T = t)} = \frac{p(t)}{S(t) + p(t)} \tag{2.1.7}$$

.

Para fazer suposição sobre qual o modelo que melhor representa os dados em estudo, $h(t)$ pode ser mais informativa do que $S(t)$. Tal forma que, diferentes funções de sobrevivência assumem formas semelhantes, enquanto que as funções de risco podem ser diferentes drasticamente.

Um outra função utilizada para representar o tempo de sobrevivência é a taxa de falha acumulada que é obtida a partir de uma função de risco e é representada por:

$$H(t) = \int_0^t h(\mu) d\mu. \tag{2.1.8}$$

A função acima fornece a taxa de falha do individuo e pode ser usada para obter $h(t)$ numa estimação não-paramétrica.Outra forma de obter-se a função de taxa de falha acumulada é pela relação com a função de sobrevivência e pode ser obtida pela seguinte forma: (COLOSIMO; GIOLO, 2006)

$$H(t) = -[\log(S(t))]. \tag{2.1.9}$$

Considerando que existem várias formas para as quais a função de risco da variável T pode assumir, é importante que já tenha um método definido para identificar o tipo de modelo mais apropriado para a variável.

Para as principais formas que a função de risco pode assumir, foi feito a seguinte metodologia para identificar:

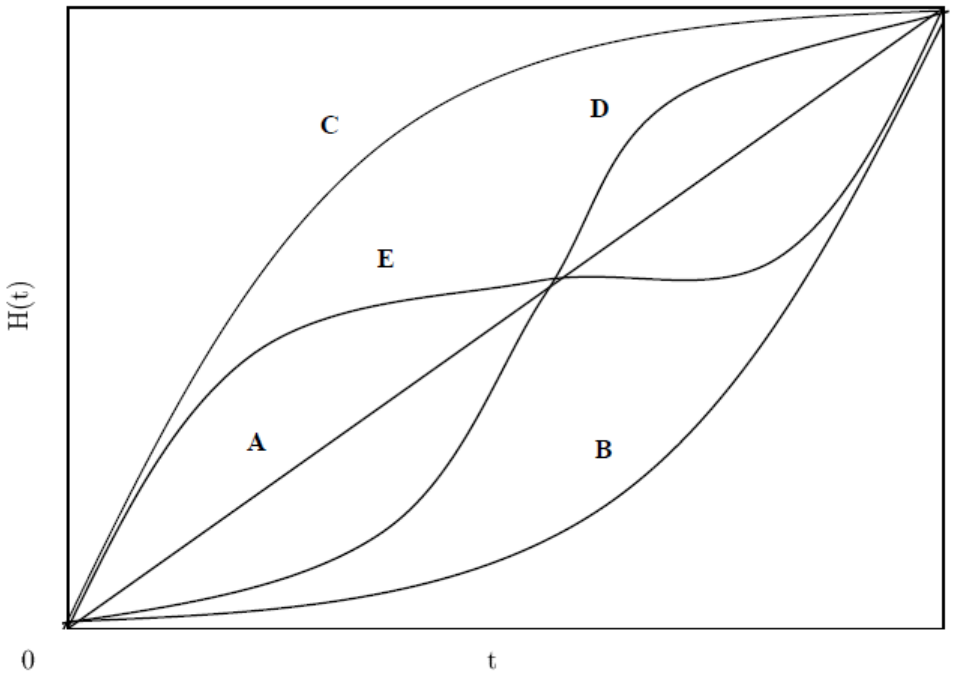


Figura 1: Diferentes formas da função de risco Santos (2017)

- Reta Diagonal(A): Função de risco constante
- Curva Convexa (B) ou Côncava (C) : Função de risco monotonicamente crescente ou decrescente
- Curva convexa e depois côncava (D): Função de risco unimodal
- Curva côncava e depois convexa (E): Função de risco em forma de U

2.1.5 Estimador de Kaplan-Meier

Nessa subseção será apresentado o conhecido estimador de Kaplan-Meier que é, sem dúvida, o mais utilizado em estudos de sobrevivência. O estimador de Kaplan-Meier por posto em (KAPLAN; MEIER, 1958) para estimar a função de sobrevivência. Também de acordo com o Kaplan e Meier (1958), foi mostrado que $\hat{S}(t)$ é um estimador de máxima verossimilhança não-paramétrico de S(t), também não é viciado e é fracamente consistente. E por essa razão este estimador vem sendo amplamente utilizado em estudos clínicos e de confiabilidade. Na ausência de censuras, o estimador de Kaplan-Meier é definido como:

$$\hat{S}(t) = \frac{\text{número de observações que não falharam até o tempo t}}{\text{número total de observações no estudo}}. \tag{2.1.10}$$

$\hat{S}(t)$ é uma função de escadas com degraus nos tempos observados de falha no tempo $1/n$, em que é n é o tamanho da amostra. Se existirem empates em um certo tempo t, o tamanho do degrau fica multiplicado pelo número de empates (COLOSIMO; GIOLO, 2006).

O estimador de Kaplan-Meier, foi construído considerado tantos intervalos de tempo quantos forem os números de falhas distintas. Os limites dos intervalos de tempo são os tempos de falha da amostra.

A expressão geral do estimador de Kaplan-Meier pode então ser apresentada após estas considerações preliminares. Considere que:

- $t_1 < t_2 < t_3 \cdots < t_k$ os k tempos distintos e ordenados de falha;
- d_j o número de falhas t_j, j = 1,·, k e
- n_j o número de indivíduos sob o risco t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

Sendo assim, o estimador de Kaplan-Meier pode ser expresso de forma genérica como:

$$\hat{S}(t) = \prod_{j:t_j<1} \left(1 - \frac{d_j}{n_j}\right). \tag{2.1.11}$$

2.2 Distribuição Log-Logística

A distribuição de Log-Logística proposta em Tadikamalla e Johnson (1982) vem ganhando espaço como alternativa às distribuições de Weibull e a log-normal, conforme é discutido em Colosimo e Giolo (2006). Considerando a análise de sobrevivência, devido ao comportamento de função de sobrevivência e também à sua densidade de probabilidade apresentar caudas pesadas, tem sido utilizado em áreas de finanças e atuária.

Seja T uma variável aleatória com distribuição Log-Logística, cuja a função de densidade de probabilidade, pode ser denotada for $f(t)$, essa é dada por:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}[1 + (t/\alpha)^\gamma]^{-2}, \tag{2.2.1}$$

sendo que $t > 0$, $\alpha > 0$ é o parâmetro de escala e $\gamma > 0$ é o parâmetro de forma. Além disso, a função de distribuição acumulada é denotada como $F(t)$ e deve ser expressa por:

$$F(t) = \left[1 + \left(\frac{t}{\alpha}\right)^\gamma)\right]^{-1}. \tag{2.2.2}$$

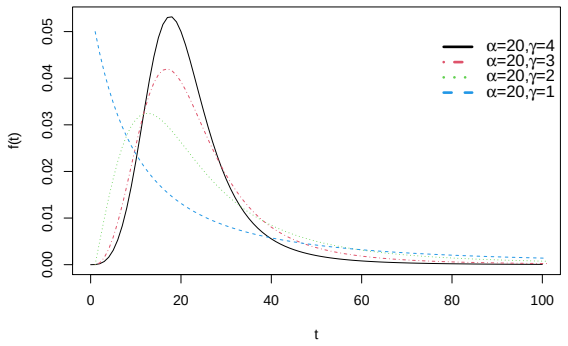
Uma das vantagens de utilizar a Log-Logística, de acordo com Santos (2017) é a simplicidade da expressão para as funções de sobrevivência e de risco. A função de sobrevivência é expressa por:

$$S(t) = \left[1 + \left(\frac{t}{\alpha}\right)^{\gamma})\right]^{-1} = \frac{1}{1 + (t/\alpha)^{\gamma}}, t > 0. \tag{2.2.3}$$

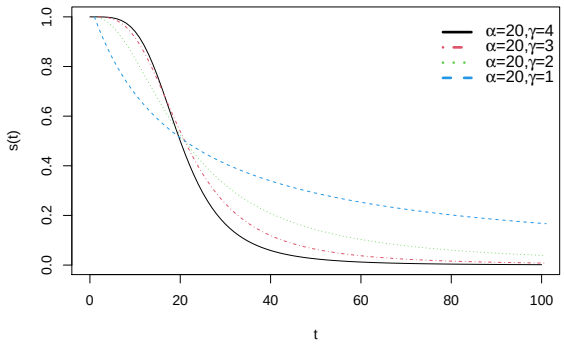
Enquanto a função de risco pode ser dada por:

$$h(t) = \frac{\gamma(t/\alpha)^{\gamma-1}}{\alpha[1 + (t/\alpha)^{\gamma}}. \tag{2.2.4}$$

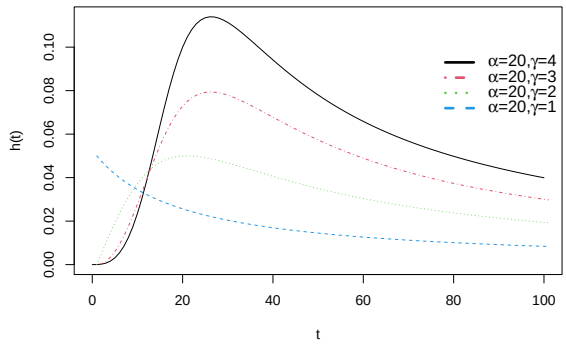
Na Figura 2 é apresentada as funções de probabilidade, sobrevivência e risco da distribuição Log-Logística. Para a construção dos gráficos o parâmetro α, de escala, foi mantido igual a 20 e o parâmetro γ, de forma, variou de 1 à 4.



(a) Função de Probabilidade



(b) Função de Sobrevivência



(c) Função de Risco

Figura 2: Função de probabilidade, sobrevivência e risco da distribuição Log-Logística para diferentes valores do parâmetro γ

A esperança e variância da distribuição Log-Logística podem ser dadas pelas

seguintes expressões, respectivamente: (TADIKAMALLA; JOHNSON, 1982) :

$$E(T) = \alpha \frac{\pi/\gamma}{sen(\pi/\gamma)}, \gamma > 1; \tag{2.2.5}$$

$$Var(T) = \alpha^2 \frac{2\pi/\gamma}{sen(2\pi/\gamma)} \left[\alpha \frac{\pi/\gamma}{sen(\pi/\gamma)}\right]^2, \gamma > 2.$$

Para encontrar o tempo mediano de sobrevivência, é preciso utilizar a função quantílica da distribuição Log-Logística que é dada por:

$$q_m(m, \alpha, \gamma) = \alpha \left[\frac{m}{(1-m)}\right]^{1/\gamma}, \tag{2.2.6}$$

sendo $0 \leq m \leq 1$.

2.3 Distribuição Log-Normal

A distribuição Log-Normal é muito utilizada para caracterizar tempos de vida de produtos e indivíduos. Isto inclui, fadiga de metal, semicondutores, diodos e isolação elétrica. Ela também é bastante utilizada para descrever situações clínicas, como o tempo de vida de pacientes com leucemia.

Como o nome anuncia, o logaritmo de uma variável com distribuição Log-Normal com parâmetros μ e σ possui uma distribuição normal com média μ e desvio-padrão σ. Esta relação significa que dados provenientes de uma distribuição Log-Normal podem ser analisados segundo uma distribuição normal, desde de que, evidentemente, se considere o logaritmo dos dados.

A função de densidade de uma variável aleatória T contínua com distribuição Log-Normal é dada por:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{\frac{-(\log(t) - \mu)^2}{2\sigma^2}\right\}, \tag{2.3.1}$$

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

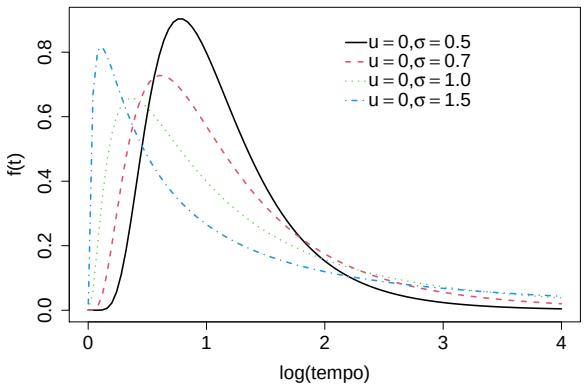
A sua função de sobrevivência não possui forma explícita, e ela é representada por:

$$S(t) = \Phi\left(\frac{-\log(t) + \mu}{\sigma}\right), \tag{2.3.2}$$

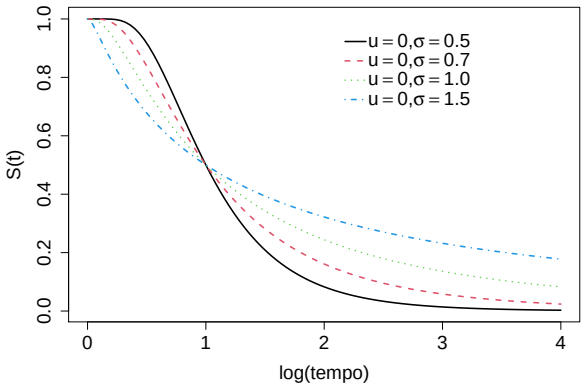
em que $\Phi(.)$ é a função de distribuição acumulada de uma normal padrão.

Assim como a sua função de sobrevivência, a taxa de falha de uma distribuição Log-Normal não possui forma analítica explícita, por isso ela é representada por:

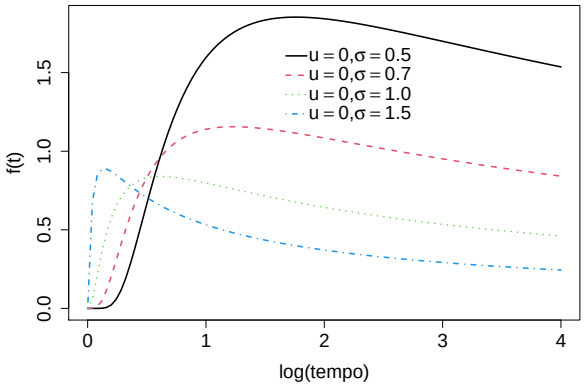
$$h(t) = \frac{f(t)}{S(t)}. \tag{2.3.3}$$



(a) Função de Probabilidade



(b) Função de Sobrevivência



(c) Função de Risco

Figura 3: Função de probabilidade, sobrevivência e risco da distribuição Log-Normal para diferentes valores do parâmetro σ

A esperança e variância da distribuição Log-Normal podem ser dadas pelas seguintes expressões, respectivamente: (COLOSIMO; GIOLO, 2006) :

$$E(T) = \exp\left(\mu + \sigma^2/2\right); \tag{2.3.4}$$

$$Var(T) = \exp(2\mu + \sigma^2).(\exp[\sigma^2] - 1).$$

2.4 Distribuição Log-Logística Discreta

Como foi citado na introdução, neste trabalho o tempo será considerado discreto.

Ao considerar a definição da função de probabilidade 2.1.3, função de sobrevivência 2.1.5 e função de risco 2.1.7 quando T é uma variável discreta e ao considerar as funções da distribuição Log-Logística definidas na Seção 2.2, para a distribuição Log-Logística discreta as funções de probabilidade, sobrevivência e risco, respectivamente, são definidas por:

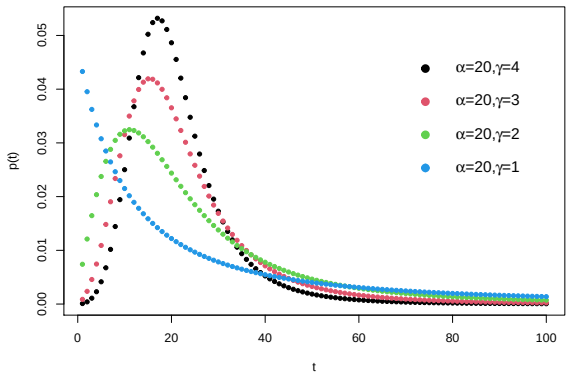
$$p(t,\alpha,\gamma) = \frac{1}{1+(t/\alpha)^{\gamma}} - \frac{1}{1+[(t+1)/\alpha]^{\gamma}}, t = 0,1,2,...; \tag{2.4.1}$$

$$S(t,\alpha,\gamma) = \frac{1}{1+[(t+1)/\alpha]^{\gamma}}, t = 0,1,2,...; \tag{2.4.2}$$

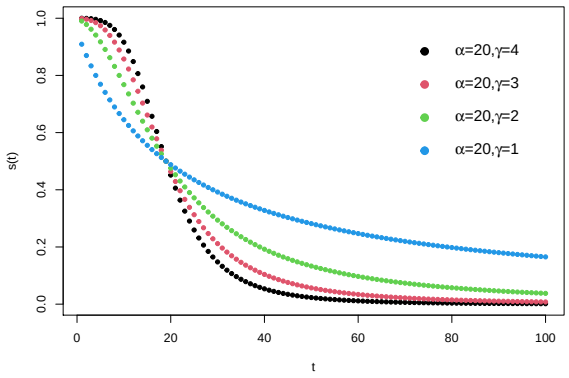
$$h(t,\alpha,\gamma) = 1 - \frac{1+(t/\alpha)^{\gamma}}{1+[(t+1)/\alpha]^{\gamma}}, t = 0,1,2,...; \tag{2.4.3}$$

sendo que $\alpha > 0$ é o parâmetro de escala e $\gamma > 0$ é o parâmetro de forma.

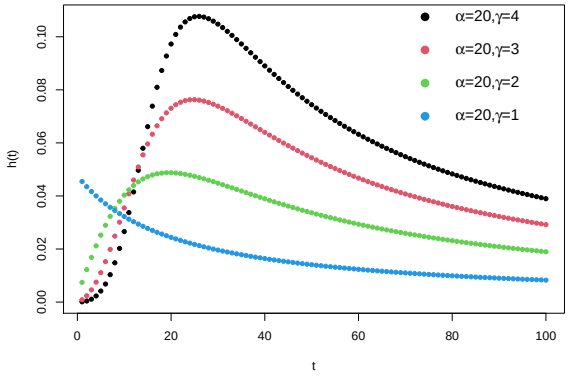
A Figura 4, mostra o gráfico das função de probabilidade, função de sobrevivência e função de risco para a distribuição Log-Logística discreta para diferentes valores do parâmetro de forma γ.



(a) Função de Probabilidade



(b) Função de Sobrevivência



(c) Função de Risco

Figura 4: Imagem para as 3 funções da distribuição Log-Logística discreta alterando o parâmetro de forma γ

Além das funções 2.4.1, 2.4.2 e 2.4.3, outra função que deve ser citada é a função acumulada denotada por $F(t)$. A função acumulada é bastante usada pois a partir dela pode-se encontrar a função quantil. Sendo que $S(t)$, uma função de sobrevivência da distribuição Log-Logística discreta, então $F(t)$ é encontrada utilizando a seguinte relação:

$$F(t) = P(T \leq t) = 1 - S(t) = 1 - \frac{1}{1 + [(t+1)/\alpha]^{\gamma}}, t = 0, 1, 2, ... \tag{2.4.4}$$

Para obter a função quantil de $T \sim LLD(\alpha, \gamma)$, sendo que $LLD(\alpha, \gamma)$ é uma distribuição Log-Logística com parâmetros α e γ, utiliza-se a função inversa da distribuição acumulada $F(t)$, sendo assim a função quantil de acordo com Santos (2017) é descrita como:

$$\begin{aligned} q_m(\alpha,\gamma) &= inf\{t : F(t) \geq m\} \\ &= inf\{t : S(t) \leq 1-m\} \\ &= inf\left\{t : \frac{1}{1+[(t+1)/\alpha]^{\gamma}}\right\} \\ &= inf\left\{t : \frac{1}{1-m} \leq 1+[(t+1)/\alpha]^{\gamma}\right\} \\ &= inf\left\{t : \left[\frac{1}{1-m}-1\right]^{1/\gamma} \leq \frac{t+1}{\alpha}\right\} \\ &= inf\left\{t : \alpha\left[\frac{m}{1-m}\right]^{1/\gamma} - 1 \leq t\right\} \end{aligned}$$

Considere que T é uma variável aleatória e que $T \sim LLD(\alpha,\gamma)$ então a expressão para a esperança de T e de variância pode ser obtida, respectivamente, por:

$$E(T) = \alpha^{\gamma}\left[\sum_{k=0}^{\infty}\left(\frac{k}{a^{\gamma}+k^{\gamma}}\right) - \sum_{k=0}^{\infty}\left(\frac{k^2}{a^{\gamma}+(k+1)^{\gamma}}\right)\right] \tag{2.4.5}$$

e

$$Var(T) = \alpha^{\gamma}\left[\sum_{k=0}^{\infty}\left(\frac{k^2}{a^{\gamma}+k^{\gamma}}\right) - \sum_{k=0}^{\infty}\left(\frac{k^2}{a^{\gamma}+(k+1)^{\gamma}}\right)\right] - E[(T)]^2. \tag{2.4.6}$$

De acordo com a simulação realizada em Santos (2017), observa-se que as equações 2.4.5 e 2.4.6 são satisfatórias para obter, respectivamente, a esperança e a variância de uma distribuição Log-Logística discreta para valores de $\gamma = 4$ e $\gamma = 5$. Isso ocorre pois esses valores de γ satisfaz a condição de convergência de $\gamma - 1 > 2$. Porém, para valores que não satisfazem essa condição as simulações em Santos (2017) não encontrou valores satisfatórios para as equações 2.4.5 e 2.4.6 quando se refere a calcular esperança e variância, respectivamente.

2.5 Seleção de modelos

É importante para que possa selecionar um modelo mais adequado. Seja comparando modelos de distribuições diferentes ou com variáveis explicativas diferentes que tenha-se um critério claro para avaliação.

2.5.1 Teste de razão de verossimilhança

Este teste é realizado com base na função de verossimilhança dos modelos que são suposto modelos encaixados. As hipóteses do teste são:

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

O teste é realizado a partir de dois ajustes. O modelo generalizado e obtenção do logaritmo de sua função de verossimilhança ($\log L(\hat{\theta}_G)$) e modelo de interesse e obtenção do logaritmo de sua função de verossimilhança ($\log L(\hat{\theta}_M)$). Com isso, a estatística do teste é:

$$TRV = -2 \log \left[\frac{\log L(\hat{\theta}_M)}{\log L(\hat{\theta}_G)} \right], \tag{2.5.1}$$

que, sob H_0, tem aproximadamente uma distribuição qui-quadrado com graus de liberdade igual a diferença do número de parâmetros dos modelos sendo comparados. É importante ressaltar que o teste só pode ser realizado para modelos encaixados.

2.5.2 Critério de informação

Para a comparação de modelos vindo de um mesmo conjunto de dados, é comum se utilizar critério de informação para escolher um modelo que se ajusta bem aos dados, mas que não seja super parametrizado. Por isso, utiliza-se medidas como AIC, BIC e AIC_C para a seleção do modelo probabilístico a ser utilizado. O critério de informação pode ser escrito como:

$$-2 \log(L(\hat{\theta})) + kp, \tag{2.5.2}$$

em que $\log(L(\hat{\theta}))$ é o logaritmo da verossimilhança do modelo, p é o número de parâmetros utilizado no modelo e k é uma medida para punir modelos com excesso de parâmetros, com:

- $k = 2$ para o Critério de Informação de Akaike (AIC);
- $k = \log(n)$ (em que n é o número de observações) para o Critério de Informação Bayesiano (BIC);
- $k = 2 + (2(p+1))/(n-p-1)$ (em que n é o número de observações) para o Critério de Akaike Corrigido (AIC_C);

Para a escolha do modelo, é escolhido aquele com o menor valor para o critério de informação utilizado (AIC, BIC, ou AIC_C). O critérios de informação diferentes do teste de razão de verossimilhança podem ser usados tanto para comparar modelos encaixados quanto modelos não encaixados, sendo assim uma escolha para comparação de modelos com diferentes distribuições.

2.6 Adequação do modelo

Segundo (COLOSIMO; GIOLO, 2006), é comum verificar a adequação do modelo ajustado por meio da inspeção de gráficos dos resíduos. Em dados censurados, como os resíduos não seguem a distribuição Normal e são assimétricos, é necessário utilizar resíduos especiais sendo que o mais utilizado é o resíduo de Cox-Snell.

2.6.1 Resíduo Cox-Snell

De acordo com Colosimo e Giolo (2006), o resíduo de Cox-Snell auxilia a examinar o ajuste global do modelo final e é definido por:

$$\hat{e}_i = \hat{H}(t_i|x_i), \tag{2.6.1}$$

em que $\hat{H}(\cdot)$ é a função de risco acumulada obtida do modelo ajustado e x é o vetor de covariáveis associados ao i-ésimo indivíduo.

Os resíduos $\hat{e}_i$, de acordo com Colosimo e Giolo (2006), vêm de uma população homogênea e devem seguir uma distribuição exponencial padrão. Logo, o gráfico de $\hat{e}_i$ versus $\hat{H}(\hat{e}_i)$ deve ser aproximadamente uma reta. Outras análises gráficas indicadas para este resíduo podem ser vistas em (COLOSIMO; GIOLO, 2006, p.87-88).

2.7 Inferência Estatística

2.7.1 Método de Máxima Verossimilhança

Ao considerar um modelo probabilístico para descrever a variável aleatória T, é fundamental estimar os parâmetros do modelo. Sendo T uma variável aleatória contínua com função de densidade igual a $f(t|\boldsymbol{\theta}$ e $t_1, t_2, ..., t_n$ uma amostra aleatória observada de tamanho n de T. Em que $\boldsymbol{\theta}$ é o vetor dos parâmetros, a função de verossimilhança para $\boldsymbol{\theta}$ é expressa por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i, \boldsymbol{\theta}), \tag{2.7.1}$$

no caso de $t_1, t_2, ..., t_n$ seja uma amostra de uma variável aleatória discreta T e $\boldsymbol{\theta}$ é o vetor dos parâmetros, a função de verossimilhança é dada por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(t_i, \boldsymbol{\theta}). \tag{2.7.2}$$

A partir da função acima é possível encontrar o valor de $\boldsymbol{\theta}$ que maximiza a probabilidade observada ocorrer, ou seja, o valor de $\boldsymbol{\theta}$ que maximiza a função L($\boldsymbol{\theta}$). A expressão (2.7.1) é utilizada quando todos os indivíduos experimentaram o evento de interesse, ou seja, todos eram suscetíveis.

No entanto, quando ocorre casos de indivíduos que não venham experimentar o evento de interesse, os dados precisarão ser separadas de tal forma que as observação são divididas em dois conjuntos. Um dos conjuntos com as r observações não censuradas,e outro, por consequência, com as $n - r$ observações censuradas. Dessa forma, a função de verossimilhança $L(\boldsymbol{\theta})$ é $f(t_i, \boldsymbol{\theta})$ para cada observação que o caso for o tempo de falha e, para cada observação censurada, a contribuição para é sua função de sobrevivência $S(t_i, \boldsymbol{\theta})$. Sendo assim, a função de verossimilhança é expressa por :

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{r} f(t_i, \boldsymbol{\theta}) \prod_{i=r+1}^{n} S(t_i, \boldsymbol{\theta}), \tag{2.7.3}$$

e isso é equivalente à :

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \left[f(t_i, \boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i, \boldsymbol{\theta})\right]^{\delta_i}, \tag{2.7.4}$$

em que δ_i é a variável indicadora de falha ou censura que foi apresentada na equação 2.1.1, $f(.)$ é a função de densidade de probabilidades e $S(.)$ a função de sobrevivência.

No caso de T ser uma variável aleatória discreta e que $\boldsymbol{\theta}$ o vetor de parâmetros, ao considerar $L(\boldsymbol{\theta})$ como a contribuição de $f(p_i, \boldsymbol{\theta})$ para cada observação que o caso for o tempo de falha e, para cada observação censurada, a contribuição para é sua função de sobrevivência $S(t_i, \boldsymbol{\theta})$, tem-se que a função de verossimilhança é expressa por:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{r} p(t_i, \boldsymbol{\theta}) \prod_{i=r+1}^{n} S(t_i, \boldsymbol{\theta}), \tag{2.7.5}$$

e isso é equivalente à :

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \left[p(t_i, \boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i, \boldsymbol{\theta})\right]^{\delta_i}, \tag{2.7.6}$$

em que δ_i é a variável indicadora de falha ou censura que foi apresentada na equação 2.1.1.

Por motivos de otimização dos cálculos, é comum trabalhar com o logaritmo da função de verossimilhança. Dessa forma, aplicando o log para a equação 2.7.4, tem-se:

$$\log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \delta_i log[f(t_i, \boldsymbol{\theta}) + (1 - \delta_i) log[S(t_i, \boldsymbol{\theta}] + C, \tag{2.7.7}$$

sendo que C é uma constante não independente.

Já se considerando que T é uma variável aleatória discreta o log da função de verossimilhança por:

$$\log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \delta_i log[p(t_i, \boldsymbol{\theta}) + (1 - \delta_i) log[S(t_i, \boldsymbol{\theta}] + C, \tag{2.7.8}$$

sendo que C é uma constante não independente.

Os estimadores de máxima verossimilhança (EMV) são os valores de $\boldsymbol{\theta}$ que maximizam $L(\boldsymbol{\theta})$ ou, de forma equivalente, o conjunto de $\boldsymbol{\theta}$ maximiza o logaritmo de $L(\boldsymbol{\theta})$. Eles são encontrados resolvendo-se o sistema de equações:

$$U(\boldsymbol{\theta}) = \frac{\partial \log(L(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}. \tag{2.7.9}$$

A solução deste sistema de equações para um determinado conjunto de dados deve ser obtida por meio de um método numérico, sendo que normalmente, é utilizado o método de Newton-Raphson e a utilização de um pacote computacional para realizar esse trabalho. Neste trabalho será utilizado o *software* R (R Core Team, 2020) para obter $\hat{\boldsymbol{\theta}}$ por meio da função *optim*.

2.7.2 Intervalo de confiança para os parâmetros

Após estimar os parâmetros é importante a construção de um intervalo de confiança para os mesmos. O intervalo de confiança é construído a partir da matriz de informação de Fisher definida como:

$$I_f(\boldsymbol{\theta}) = E\left[\left(\frac{\partial l(\mathbf{t}, \boldsymbol{\delta}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2\right] = -E\left[\frac{\partial^2 l(\mathbf{t}, \boldsymbol{\delta}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right] \tag{2.7.10}$$

e que $l(t, \boldsymbol{\delta}|\boldsymbol{\theta}) = log(L(\boldsymbol{\theta})$.

Então utiliza-se a distribuição assintótica do estimador de máxima verossimilhança. Para amostras grandes, sob condições regulares, essa propriedade estabelece que a distribuição de $\hat{\boldsymbol{\theta}}$ converge de forma assintótica para uma distribuição normal multivariada de média igual a $\boldsymbol{\theta}$ e a matriz de variância e covariância igual a $Var(\boldsymbol{\theta})$, isto é:

$$\hat{\boldsymbol{\theta}} \sim N_k\left(\boldsymbol{\theta}, Var(\hat{\boldsymbol{\theta}})\right);$$

em que k é a dimensão de $\boldsymbol{\theta}$. Além disso, tem-se:

$$Var(\hat{\boldsymbol{\theta}}) \simeq \left[I_f(\hat{\boldsymbol{\theta}})\right]^{-1};$$

em que $I_f(\hat{\boldsymbol{\theta}})$ é a informação de Fisher observada na amostra.

Desta forma, um intervalo aproximado de $(1-\alpha)$ de confiança para $\boldsymbol{\theta}$ é dado por:

$$\hat{\boldsymbol{\theta}} \pm Z_{1-\alpha/2}\sqrt{\hat{Var}(\hat{\boldsymbol{\theta}})}, \tag{2.7.11}$$

em que $Z_{1-alpha/2}$ é o quantil (1-α/2) de uma distribuição normal padrão.

Há casos que torna-se necessário estimar uma função dos parâmetros e uma das propriedades dos estimadores de máxima verossimilhança é a propriedade de invariância, ou seja, se $\hat{\boldsymbol{\theta}}$ é um EMV de $\boldsymbol{\theta}$ e seja $g(\hat{\boldsymbol{\theta}})$ uma função bijetora, então $g(\hat{\boldsymbol{\theta}})$ é um EMV de $g(\boldsymbol{\theta})$.

Sendo assim, para a construção de intervalo de confiança para $g(\boldsymbol{\theta})$, é preciso obter uma estimativa para o erro padrão de $g(\hat{\boldsymbol{\theta}})$ e isso é feito utilizando o método delta. Além disso, para grandes amostras, tem-se que:

$$g(\hat{\boldsymbol{\theta}}) \sim N_k(g(\boldsymbol{\theta}), Var(\hat{\boldsymbol{\theta}})[g'(\boldsymbol{\theta})]^2). \tag{2.7.12}$$

Assim, $g(\hat{\boldsymbol{\theta}})$ converge de forma assintótica para uma distribuição normal multivariada com média $g(\boldsymbol{\theta})$ e a matriz de variância e covariância $Var(\hat{\boldsymbol{\theta}})[g'(\boldsymbol{\theta})]^2$, em que k é dimensão de $g(\hat{\boldsymbol{\theta}})$ e $g'(\boldsymbol{\theta})$ é a derivada de primeira ordem de $g(\boldsymbol{\theta})$.

3 Metodologia

3.1 Banco de Dados Original

Para realizar esse estudo primeiramente foi feita uma solicitação a Universidade de Brasília,precisamente, a Secretaria de Tecnologia da Informação,para disponibilizar os dados necessários para avaliar a evasão dos estudantes do curso de graduação bacharelado em Ciência da Computação. Bem como, dados para conhecer quais fatores influenciam na evasão dos alunos.Dentro do banco de dados solicitado, pode ser listado as seguintes variáveis:

1. Identificação Aluno;
2. Identificação de Pessoa;
3. Cotas
4. Sexo;
5. Período de ingresso na UnB;
6. Período de ingresso no curso;
7. Período de saída do curso;
8. Forma de ingresso (Vestibular, PAS,...);
9. Sistemas de cotas (sim ou não);
10. Forma de saída do curso (Informação se concluiu o curso e/ou se mudou de curso ...);
11. Índice de rendimento acadêmico (IRA);
12. Total de crédito do curso;
13. Período que cursou disciplina;
14. Nome de disciplina;
15. Código de disciplina;
16. Menção na disciplina;
17. Créditos da disciplina cursada;

18. Créditos no período;
19. Curso;
20. Data de nascimento;
21. Escola (Pública ou Privada);
22. CEP;
23. Média do aluno no semestre;
24. Total de créditos cursados pelo aluno;
25. Mínimo de créditos para se formar;
26. Créditos aprovados no períodos;

A respeito das dimensões do banco de dados, originalmente, possui 348679 linhas e 26 colunas. É importante ressaltar que essas linhas não correspondem a um único aluno nesse momento, visto que o mesmo aluno pode ter cursado disciplinas em diferentes semestres, logo aparecendo em diversas vezes no banco. Dito isso, o número de Alunos diferentes no banco inicial é de 28396 com dados correspondente durante o período de entrada na UnB desde do segundo semestre de 1985 até o segundo semestre de 2019. Também é importante ressaltar que entre as observações há informações completas de um acompanhamento apenas de discentes que estivesse matriculados nos seguintes cursos: Ciência da Computação ,Computação (licenciatura) e Engenharia de Computação. Além disso, a variável "Cotas"é uma coluna com valores vazios. Dessa forma, será necessário uma limpeza no banco de dados para que seja possível as análises.

3.2 Limpeza do Banco

Como dito anteriormente é preciso limpar o banco e realizar alguns filtros para que obtenha-se um banco de dados apenas com informações completas.

Primeiramente foi realizado um descarte da coluna da variável "Cotas", dado que a coluna continua apenas registros vazios. Também foi realizado filtro a respeito de curso, visto que o objetivo do trabalho é estudar o curso de Ciência da Computação. Foi selecionado apenas observações as quais a variável "Curso"correspondesse o curso de interesse.

Em seguida, foi preciso realizar um filtro em relação ao tempo de ingresso no curso, pois com todas as observações e com mudanças no currículo da graduação em Ciência da Computação era necessário um recorte que fosse mais significativo ao que o

curso é efetivamente na atualidade. Dessa forma, foi escolhido o primeiro semestre de 2014 filtro de tempo para ingressantes no curso em Ciência da Computação,isto é foi selecionado apenas ingressantes que ingressam no primeiro semestre de 2014 ou após esse período. Esse período foi escolhido levando em consideração o tempo máximo de entrada de um novo aluno no banco, segundo semestre de 2019, sendo assim foi considerado que os alunos de 2014/1 teriam tempo suficiente para completar um ciclo de graduação visto que o curso tem como recomendado a realização em 9 semestres ou então 4 anos e 6 meses.

Além disso, outras variáveis também tiveram de ser descartadas por razões de valores incoerentes ou diversos erros de digitação, entre essas variáveis se encontram: Total de créditos cursados pelo aluno, créditos do período, créditos aprovados no semestre e média do aluno no semestre. Também foi descartada a variável de "Identificação de Pessoa", visto que era o mais adequado trabalhar com a "Identificação do Aluno".

Após todas as modificações o banco fica reduzido a 16551 linhas e 17 variáveis, sendo que essas linhas ainda não correspondem diretamente a quantidade de alunos distintas. Essa quantidade é equivalente a 879 alunos distintos no curso de graduação de Ciência da Computação entre o período de entrada no curso de 2014/1 e 2019/2.

3.3 Construção da Variável Tempo e Censura

Neste trabalho, a intenção de utilizar análise de sobrevivência tem como objetivo avaliar o tempo até a evasão de alunos de graduação de curso de bacharel em Ciência da Computação da Universidade de Brasília. Para a construção da variável de Tempo e Censura foi preciso utilizar as variáveis: Período de ingresso no curso, Período de saída do curso e Forma de saída do curso.

Primeiramente para a criação da variável censura, ou Status, que vai indicar se o aluno sofreu um tempo de falha ou censura. É uma variável dicotômica com dois valores indicando 0 para caso tenha sofrido censura, 1 para caso tenha tenha sofrido falha, como visto na subseção 2.1.1. Para essa construção primeiramente é preciso identificar todas as formas de saída do banco de dados. Será considerado a definição de evasão, quando por qualquer razão que seja, o aluno é desvinculado da matrícula do curso de graduação como visto em Santos e Albuquerque (2019).

A tabela abaixo mostra todas as 11 formas de saída do curso que ocorreram no banco, sendo que a forma de saída "Ativo"significa que o aluno em questão ainda está matriculado no curso de graduação. As únicas formas de saída que foram classificadas como censura foram "Ativo"e "Formatura", dado que em uma o aluno ou ainda está em curso na graduação enquanto que na outra não há mais possibilidade de cometer evasão dado que se formou.

Tabela 1: Tabela de formas de Saída

Formas de Saída do Curso	Status
Ativo	Censura
Formatura	Censura
Desligamento - Abandono	Falha
Desligamento-Força de Convênio	Falha
Desligamento por Força de Intercâmbio	Falha
Desligamento - não cumpriu condição	Falha
Desligamento Voluntário	Falha
Mudança de Curso	Falha
Novo Vestibular	Falha
Reprovar 3 vezes na mesma disciplina obrigatória	Falha
Transferência	Falha

Em seguida foi criado a variável tempo, para tanto a utilização das variáveis "Período de ingresso no curso"e "Período de saída no curso"sendo que o tempo será medido em semestres e será obtido pela subtração do período de saída pelo de entrada. Além disso, para os casos de censura que fosse do tipo "Ativo"foi utilizado o último período do aluno para "Período que cursou disciplina"como período de saída, ou seja, o último período que o aluno ativo cursou pelo menos alguma disciplina. Pois não há a informação de período de saída desse aluno, uma vez que ainda está ativo no curso. Para realizar essa subtração porém, foi necessário antes fazer uma mudança nos períodos e identificar os casos em que a entrada ou a saída ocorreu em um semestre de verão. A Universidade de Brasília não reconhece o verão como um semestre normal, logo, uma decisão de não trabalhar com o semestre de verão no tempo foi tomada. Todos os períodos de verão, seja de ingresso, saída ou que determina quando uma disciplina foi cursada durante o verão passaram para o primeiro semestre do respectivo ano. Dessa forma, um aluno que tivesse ingressado no curso no verão de 2015 passaria a ser tratado como se tivesse ingressado no primeiro semestre de 2015, como um exemplo.

Outro fator determinante a se contar sobre o tempo foi considerar ou não a ocorrência de falha e/ou censura no tempo igual a 0 (zero). Depois de muita elaboração e pesquisa em trabalhos anteriores, não encontrou-se nenhuma regra a respeito do aluno que ingressa e comete evasão no mesmo semestre. Nesse caso, há duas possibilidades, tratá-lo como evasão no tempo 0 ou então no tempo 1, dado que os dados são discretos. A decisão de tratar casos de evasão que ocorressem no mesmo semestre de ingresso como 1 foi tomada considerando que o aluno em questão não comete a evasão no momento exato 0 e sim tem uma experiência dentro da Universidade de Brasília. Sendo assim, a variável tempo foi criada com amplitude de 1 a 13 semestres.

3.4 Criação de Variáveis

Algumas variáveis foram criadas a partir das variáveis originais do banco, assim como algumas variáveis sofreram algumas alterações, principalmente, nos fatores que as compõem.

3.4.1 Taxa de reprovação

Taxa de Reprovação consiste em uma variável que pudesse levar em consideração os créditos que os alunos cursaram e a proporção desses créditos que eles reprovaram. Para essa construção, foi feito a seguinte definição utilizando as variáveis "Menção na disciplina"e "Créditos da disciplina". Foi considerado devidamente cursado a disciplina que obtivesse as menções: SS, MS, MM, M", II e SR. Sendo assim, para cada aluno foi somado o número de créditos que o aluno cursou com disciplinas nas quais obteve menções citadas acima, essa variável de apoio foi chamada de "Total de créditos cursados". Ainda era necessário identificar os créditos dos quais os alunos reprovaram, sendo assim foi considerado como menções de reprovação as seguintes menções: MI, II e SR. Utilizando essas últimas três menções em um processo similar ao de encontrar o Total de créditos cursados, encontra-se o Total de créditos reprovados pelo aluno. Dessa forma, pode-se dizer que para cada aluno i a fórmula para encontrar a Taxa de reprovação segue:

$$\text{Taxa de reprovação} = \frac{\text{Total de créditos reprovados}}{\text{Total de créditos cursados}} \tag{3.4.1}$$

Observe que a Taxa de reprovação é um valor numérico que vai de 0 a 1, visto que os créditos reprovados estão contido em créditos cursados.

3.4.2 Total de trancamentos

Total de trancamentos segue uma linha muito similar ao de taxa de reprovação, porém dessa vez não será utilizado os créditos. O interesse é contar quantas vezes um aluno i trancou alguma disciplina. Sendo assim, para a construção dessa variável foi considerado como menções de trancamento os categorias da variável Menção da disciplina que correspondessem a: TJ e TR. Para cada aluno foi feito a soma de quantas vezes essas duas menções pareciam e essa contagem foi denominada de Total de trancamentos.

3.4.3 Cursou Verão

Cursou Verão é uma variável binária que vai indicar se o aluno i cursou alguma disciplina em um semestre de verão ou não. Para isso foi avaliado a variável "Período cursou disciplina"e feito uma marcação da seguinte forma, 1 para se o aluno i cursou verão e 0 caso o aluno i não tenha cursado verão.

3.4.4 Idade em Anos

O banco originalmente não oferece informação a respeito da idade dos alunos, porém oferece a data de nascimento e foi a partir dessa variável que foi criada a variável de idade em anos, foi considerado para a idade dos alunos a idade que teriam na data da realização desse trabalho de acordo com a data de nascimento, ou seja, quantos anos completos até a data de 21/10/2021.

3.4.5 Forma de ingresso

Forma de Ingresso não é uma variável nova, porém é uma variável que sofreu alterações nos fatores que a compõem. Originalmente, para os 879 alunos diferentes as formas de ingresso estavam distribuídas conforme se observa na tabela abaixo:

Tabela 2: Formas de Ingresso UnB original

Forma de Ingresso	Frequência Absoluta
Acordo Cultural-PEC-G	1
Convênio-Int	1
Convênio - Andifes	1
Matrícula Cortesia	1
Enem UnB	5
Transferência Obrigatória	29
Transferência Facultativa	31
Portador Diplom Curso Superior	63
Sistema de Seleção Unificada (SISU)	177
Programa de Avaliação Seriada (PAS)	216
Vestibular	354

Fica evidente que existe uma concentração em três formas de ingresso principais: Vestibular, Programa de Avaliação Seriada(PAS) e Sistema de Seleção Unificada (SISU). Além desses três, a partir de 2019 a Universidade de Brasília passou a aceitar alunos

inscritos em seu próprio critério (Exame Nacional do Ensino Médio) e aparece no banco como "Enem UnB", levando em consideração que o critério de avaliação e número de vagas é o mesmo do SISU, faz sentido que sejam aglomerados na mesma classe.

Todas as demais formas de ingresso serão classificadas como "Outras formas de ingresso"visto que não possuem frequências significativas para justificar uma classe própria para cada, dessa forma, a nova tabela de frequência para os 879 alunos e suas respectivas formas de ingresso pode ser visualizada abaixo:

Tabela 3: Tabela Formas de Ingresso Alterada

Formas de Ingresso	Frequência Absoluta
Outras Formas de Ingresso	127
Sistema de Seleção Unificada(SISU e ENEM)	182
Programa de Avaliação Seriada (PAS)	216
Vestibular	354

3.5 Retirar informações duplicadas

Como citado anteriormente o banco de dados não correspondia a uma linha por aluno, visto que um mesmo aluno poderia aparecer diversas vezes no banco de dados. De fato, isso ocorre que o aluno i aparecia no banco de dados equivalente ao número de disciplinas n que tivesse cursado em cada semestre s. Esse formato de banco não é interessante para a realização de análises mais elaboradas, para isso, é de interesse que cada linha seja correspondente apenas a um único aluno. Para isso, foi necessário o descarte de diversas variáveis, tais como: Período cursou a disciplina, Créditos da disciplina, Código da disciplina, Nome da disciplina, Menção na disciplina, Mínimo de créditos para se formar e Total de crédito do curso. Uma vez que essas variáveis foram utilizadas para criar outras, sua informação foi resumida em valores únicos para cada aluno. Outras variáveis foram descartadas por motivos de que não é possível extrair mais informação delas para o modelo, após a criação de novas variáveis a partir das originais, entre elas se encontram: CEP, Data de Nascimento e Período de Ingresso na UnB.

Sendo assim, o banco final teve dimensão reduzida a 879 linhas e 16 colunas, sendo que agora cada linha corresponde apenas a um aluno distinto. As variáveis do banco a ser considerado a partir de agora são listada a baixo:

1. Identificação Aluno;
2. Tempo (semestres);
3. Status (Censura ou Falha);

4. Sexo;
5. Período de ingresso no curso;
6. Período de saída do curso;
7. Forma de ingresso;
8. Sistema de cotas (sim ou não);
9. Forma de saída do curso ;
10. Índice de rendimento acadêmico (IRA);
11. Curso;
12. Idade (anos);
13. Taxa de reprovação;
14. Total de Trancamentos;
15. Cursou verão (sim ou não);
16. Escola (Pública ou Privada);

3.6 Divisão em dois Bancos

Depois de todas as alterações no banco foi feito uma investigação a respeito do número grande de censuras que o banco tinha e descobriu-se que ocorria devido ao número alunos ativos. Pois o recorte temporal fez com que os alunos de 2014/1 tivessem tempo suficiente para se formar, mas não foi feito um filtro com as entradas em diante, logo o número de alunos ativos é muito alto.

Para contornar esse caso foi realizada uma separação em dois bancos. Um dos bancos que continuaria com todas as 879 observações com o período de ingresso no curso de 2014/1 até 2019/2. E um banco reduzido que tem período de entrada de 2014/1 até 2016/2, essa escolha de um novo banco é para que se possa construir um modelo apenas com alunos que já tenham cursado pelo menos 3 semestres do curso ou tenham evadido antes disso, assim passando pelas disciplinas reconhecidas como formação básica do curso.

A justificativa para realizar modelos para dois bancos é baseada no currículo do curso de Ciência da Computação e na análise descritiva que será exposta na seção 4.1. Foi avaliado que boa parte das evasões do curso ocorre até o terceiro semestre, além disso, o terceiro semestre é o que compreende grande parte do que o departamento de graduação

em Ciência da Computação chama de "formação básica"do curso, que determina disciplinas que são chave para que o aluno complete o curso de Ciência da Computação. Dessa forma, um banco de dados em que todos os alunos tiveram a oportunidade de passar pelo menos três semestres no curso é interessante, pois assim, pode-se tentar identificar fatores mais relevantes para a evasão nesse momento inicial do curso. Enquanto que o banco completo compreende todas as informações de alunos de 2014/1 a 2019/2, sendo assim, possui uma quantidade significativa de alunos mais recentes e também para servir de comparação se há uma diferença entre os modelos que avaliam o tempo de evasão de alunos em bancos que apresentam uma quantidade significativa de alunos que não passaram pela "formação básica", ou seja, se há grande diferença na interpretação dos coeficientes no banco completo ou reduzido.

A ideia de não escolher apenas uma forma do banco de dados é o interesse de avaliar todos os alunos, os que possuem menos informação a respeito, considerando os alunos ativos com pouco tempo no banco para observação, mas também avaliar um banco com alunos que já tenham passado por mais tempo de universidade e ainda se mantém ativos.

Para motivo de comparação o banco reduzido possui 449 observações e o mesmo número de variáveis explicativas.

3.7 Análise dos Dados

Primeiramente a análise de dados será realizada com as técnicas de análise descritiva: gráficos de *boxplot*, gráficos de colunas, histogramas e tabelas de frequência para as variáveis de forma individual. Então será feito uma análise descritiva clássica, com as variáveis explicativas e a variável "Status"a fim de encontrar alguma relação entre a variável explicativa e a variável que indica falha ou censura.

Depois é feito uma análise de correlação entre as variáveis explicativas. Essa investigação é de extrema importância quando se trata de decidir que tipo de variável será eleita para entrar no modelo, dado que, variáveis explicativas com grande correlação entre si não devem fazer parte do mesmo modelo.

Em seguida, será feita uma análise com gráficos e técnicas descritivas utilizando a metodologia de análise de sobrevivência. Como construir diferentes curvas de sobrevivências utilizando o estimador de Kaplan-Meier para estimar $S(t)$ (KAPLAN; MEIER, 1958), assim como também construção de curvas de função de risco acumulada e curvas de tempo total em teste, ou, curva (TTT). Como visto na subseção 2.1.4, a função de risco acumulada é de extrema importância para identificar o tipo de distribuição dos dados. É utilizado o estimador de Kaplan-Meier, pois esse é o estimador de máxima verossimilhança

não-paramétrico de $S(t)$ e não viciado e também fracamente consistente. Essa análise é importante para identificar a distribuição dos dados para o tempo de sobrevivência, dessa forma identificar o modelo mais adequado.

3.8 Modelagem

Para a modelagem propriamente dita, será considerado que a distribuição Log-Logística discreta como definida na subseção 2.4 é a distribuição da variável resposta, dado a investigação apresentada em Santos (2017), porém ressalta-se que a Log-Logística é apenas um ponto de partida. Então utiliza-se de uma extensão da distribuição escolhida para incluir os parâmetros de regressão e assim concluir com os objetivos do trabalho, ou seja, encontrar e analisar os fatores que levam alunos de graduação no curso de bacharel em Ciência da Computação a evasão.

No entanto, mesmo que a distribuição Log-Logística discreta seja um ponto de partida como a distribuição para a variável resposta, ainda será feito comparações com outras distribuições afim de afirmar que o modelo com a Log-Logística discreta seja realmente o que melhor ajusta os dados. Comparações utilizando as técnicas apresentadas em 2.5.

Outras distribuições candidatas para as comparações com a Log-Logística discreta são a Log-Logística contínua e também a Log-Normal contínua. Dessa forma utilizando os critérios discutidos em 2.5 pode-se escolher a distribuição que mais se ajusta aos dados.

No caso de uma distribuição contínua apresente resultados melhores que a distribuição Log-Logística discreta, então será considerado a metodologia mais usual que é supor modelos contínuos para tempos discreto e trabalhar com os mesmos caso os resultados forem satisfatórios.

3.9 Modelo de Regressão

De maneira geral nos estudos de sobrevivência pode-se verificar que algumas covariáveis influenciam de maneira significativa o tempo de sobrevivência do individuo. O uso dessas covariáveis em um modelo de regressão é uma maneira importante de representar a heterogeneidade em uma população.

Considerando que $\boldsymbol{x}^T = (1, x_1, ...x_p)$ seja um vetor de covariáveis dos indivíduos, utiliza-se então de uma função de ligação g(.) que conecta a variável resposta às variáveis explicativas. Para um conjunto de p covariáveis, o vetor de parâmetros θ que será estimado utilizando o vetor $\boldsymbol{x}$, passa a ser definido como:

$$\theta = g(\boldsymbol{x}^T\boldsymbol{\beta}) \tag{3.9.1}$$

em que $x^t = (1, x_1, ...x_p)$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ é o vetor de coeficientes de regressão.

Considere que T é uma variável aleatória com distribuição Log-Normal assim como definida em 2.3 por:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{\frac{-(\log(t) - \mu)^2}{2\sigma^2}\right\}, \tag{3.9.2}$$

Dessa forma, ao utilizar o parâmetro μ como $\mu = \boldsymbol{x}^T\boldsymbol{\beta}$, tem-se que a função de ligação para o caso da Log-Normal é a função de identidade $I(.)$. Sendo assim, o modelo de regressão Log-Normal é definido por :

$$f(t|x) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{\frac{-(\log(t) - \boldsymbol{x}^T\boldsymbol{\beta})^2}{2\sigma^2}\right\}, \tag{3.9.3}$$

A função de sobrevivência correspondente é dada por:

$$S(t) = \Phi\left(\frac{-\log(t) + \boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right), \tag{3.9.4}$$

e a função de risco corresponde a:

$$h(t) = \frac{f(t)}{S(t)} \tag{3.9.5}$$

Para estimar os parâmetros do modelo de regressão Log-Normal será utilizado o método de máxima verossimilhança que foi exposto em 2.7.1.

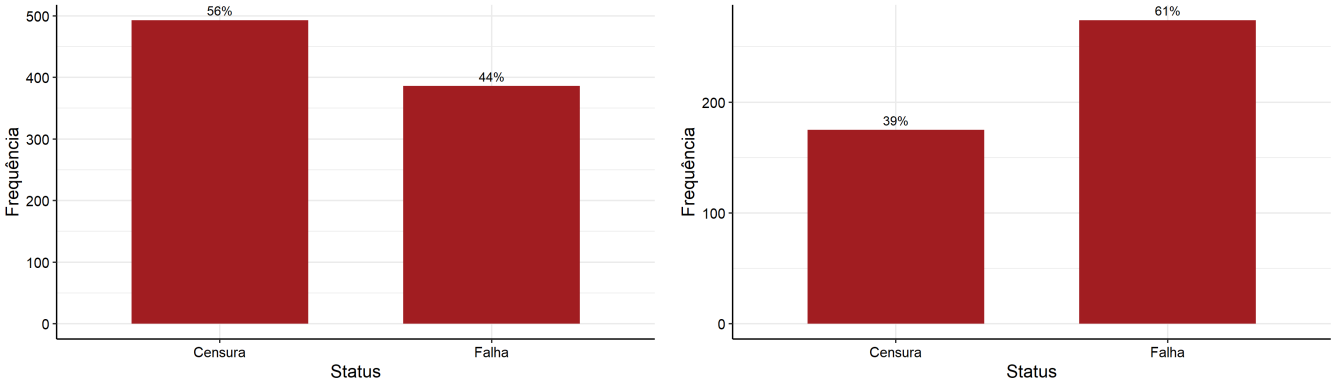
Para implementar os modelos, análises estatísticas e calcular todas as estimativas será utilizado o *software* estatístico R (R Core Team, 2020) em sua versão 4.0.0 .

4 Análise e Resultados

4.1 Análise Descritiva

4.1.1 Status

Status é a variável binária que indica se o aluno i teve um tempo de falha ou de censura, como mostrado na subseção 2.1.1. A sua construção foi devidamente evidenciada na subseção 3.3. Abaixo é apresentado os gráficos de barra para o banco completo e o banco reduzido:



(a) Banco completo (b) Banco Reduzido

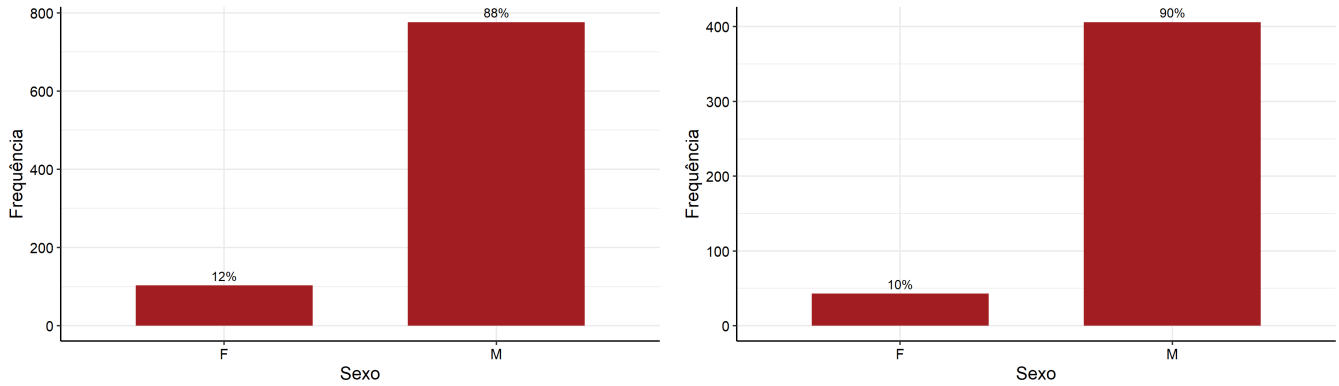
Figura 5: Gráficos de barras para Status para ambos os bancos

Apesar da escala dos gráficos serem muito diferentes, pois o banco reduzido tem 449 observações enquanto o banco completo tem 889, é interessante observar para os valores relativos no topo de cada barra. Enquanto para o banco completo observasse que 54% dos valores do banco são de tempo de censura enquanto 46% são de tempo de falha. Essa dinâmica muda drasticamente no banco reduzido, no qual 39% das observações tem tempo de censura e 61% apresenta tempo de falha.

Essa diferença era esperada e também faz parte da justificativa de separar um banco reduzido que tivessem alunos que tiveram a chance de cursar pelo menos três semestres.

4.1.2 Sexo

A variável Sexo é a respeito do sexo do aluno, possui apenas dois fatores sendo que F para designar sexo feminino e M para designar sexo masculino. Abaixo encontra-se o gráfico de barras para ambos os bancos para a variável sexo.



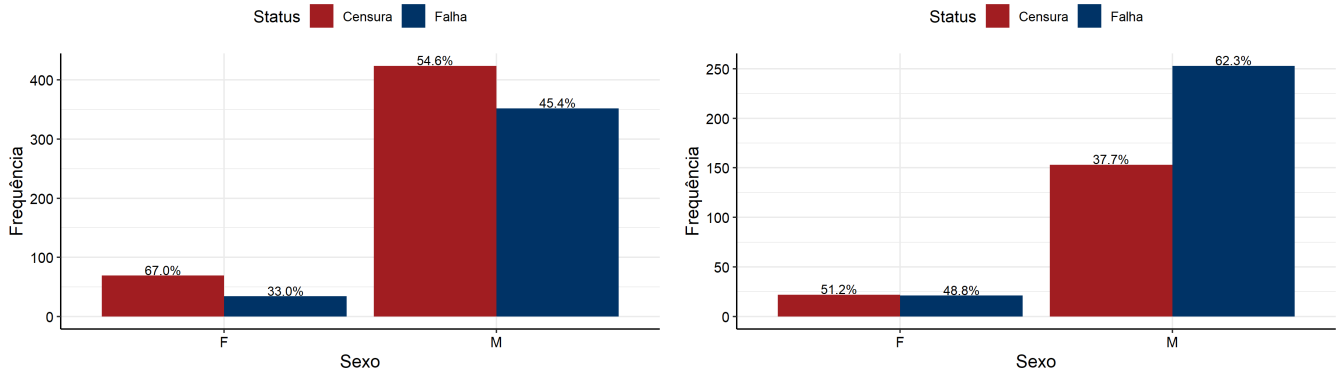
(a) Banco completo (b) Banco Reduzido

Figura 6: Gráficos de barras para Sexo para ambos os bancos

Não há grandes diferenças entre a distribuição de sexos dos dois bancos.Ou seja, essa distribuição de sexo predominante masculino não é um fator gerado apenas de matrículas mais antigas, como também não tem surtido sinal de reversão contando com matrículas mais novas.

De modo geral o curso de bacharel em Ciência da Computação na UnB é predominante masculino, com 12% de integrantes do sexo feminino para o banco completo e 10% para o banco reduzido.

Abaixo apresenta-se o gráfico de barras de sexo por Status, para verificar como os diferentes sexos estão distribuídos para o tempo de falha ou censura.

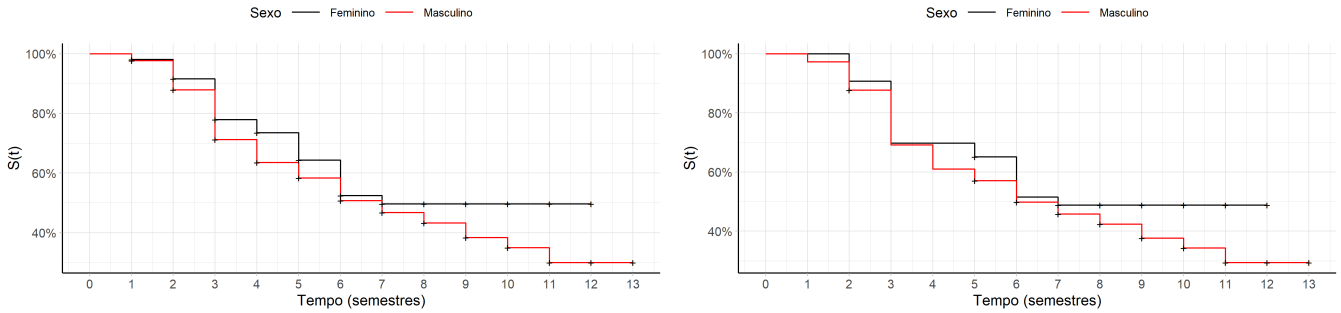


(a) Banco completo (b) Banco Reduzido

Figura 7: Gráficos de barras para Sexo vs Status para ambos os bancos

Como visto anteriormente em 4.1.1, a distribuição de Status muda de um banco para o outro. Desse modo, como mais de 88% das observações são do sexo masculino, é interessante observar que o sexo masculino reflete para ambos os bancos o que é encontrado na Figura 5. Porém para o sexo feminino, uma leve vantagem na frequência relativa para tempo de censura.

Abaixo encontra-se a curva de sobrevivência para a variável Sexo:



(a) Banco completo (b) Banco Reduzido

Figura 8: Gráficos de curvas de sobrevivência por Sexo para ambos os bancos

As duas curvas de sobrevivência por sexo são muito similares. De modo geral, parece que a probabilidade de sobrevivência para o sexo masculino é um pouco menor que para o sexo masculino, porém para verificar essa diferença foi feito um teste de *logRank*.

As hipóteses do teste são:

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

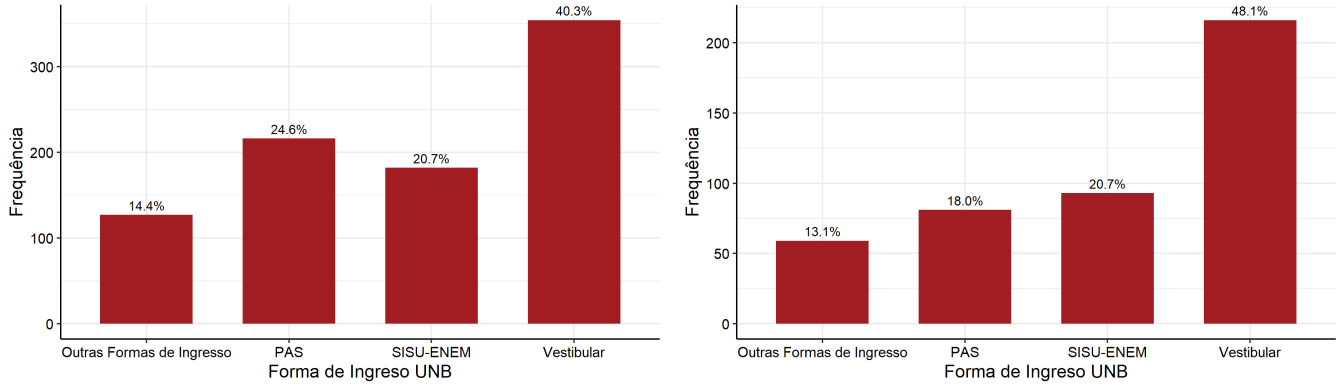
Tabela 4: Resultados do teste de *logRank* de Sexo

Banco de Dados	**Estatística do teste**	**Graus de liberdade**	**P-valor**
Banco Completo	$2,8$	1	$1 \cdot 10^{-1}$
Banco reduzido	$1,8$	1	$2 \cdot 10^{-1}$

Pode-se notar que de acordo com o teste *logRank*, considerando um nível de significância de 5%, que não há diferença entre os dois sexos quando se trata da curva de sobrevivência.

4.1.3 Forma de ingresso

A variável Forma de ingresso é uma variável categórica com 4 fatores, conforme a construção é evidenciada na subseção 3.4. Abaixo encontra-se o gráfico de barras para Forma de ingresso:

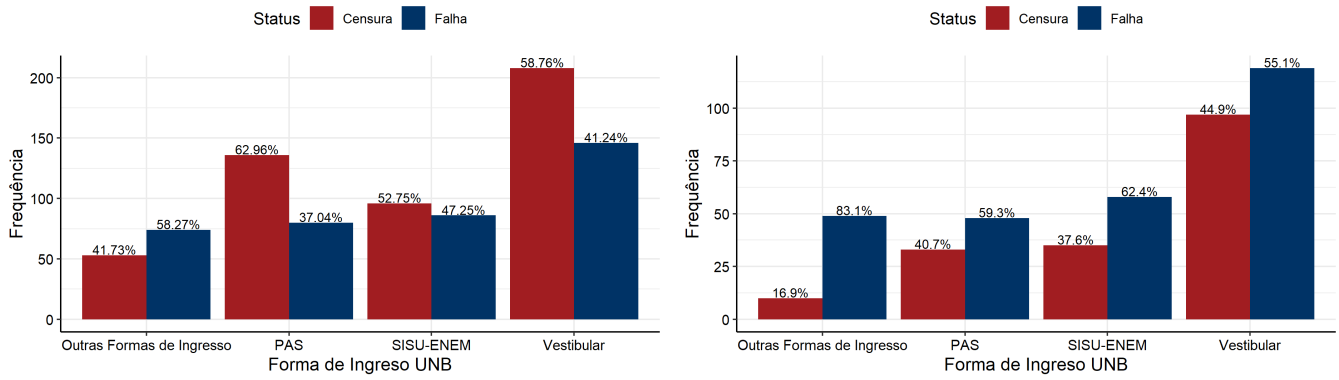


(a) Banco completo (b) Banco Reduzido

Figura 9: Gráficos de barras para Forma de ingresso para ambos os bancos

Nota-se que apesar da diferença de escala entre os gráficos, devido a quantidade de observações diferente entre o banco completo e o reduzido, porém a distribuição dos dados nos 4 fatore de Forma de Ingresso na UnB, como esperado a principal forma de ingresso na UnB é o vestibular.

É de interesse observar também a relação da variável com o Status, sendo assim foi realizado o gráfico de barras de Forma de ingresso por Status:

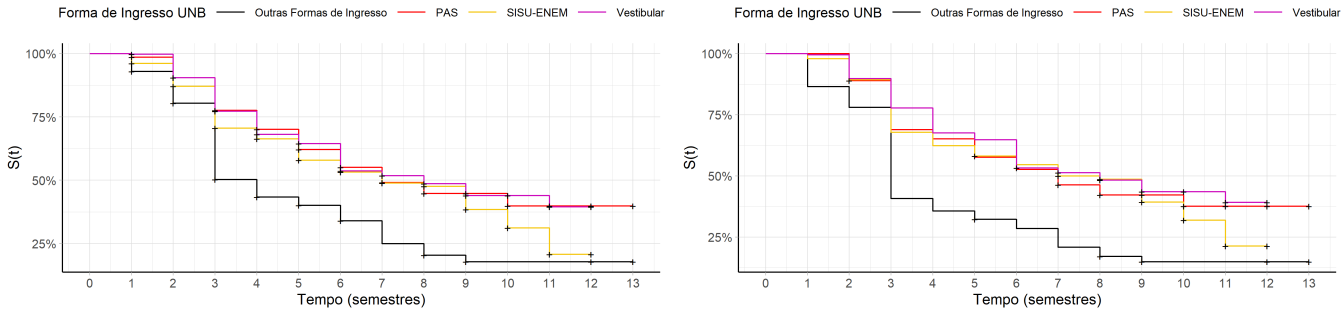


(a) Banco completo (b) Banco Reduzido

Figura 10: Gráficos de barras para Forma de ingresso vs Status para ambos os bancos

É interessante observar que na Figura 10a nota-se que as 3 principais formas de ingresso possuem tempo de censura com maior proporção do que de falha, sendo que apenas Outras formas de ingresso possui o oposto dado o comportamento diferente em Status em ambos os bancos. Nota-se que na Figura 10b todos passam a ter tempo de falha com maior proporção que o de censura, porém é evidente que no caso do Outras formas de ingresso é bem maior o tempo de falha.

A seguir as curvas de sobrevivência para Forma de ingresso:



(a) Banco completo (b) Banco Reduzido

Figura 11: Gráficos de curvas de sobrevivência por Formas de ingresso UnB para ambos os bancos

Para ambos os bancos nota-se um comportamento semelhante, enquanto Outras formas de ingresso parece ter uma curva de sobrevivência mais baixa que a curva das demais formas de ingresso. Também observa-se que enquanto PAS e Vestibular estão quase sobrepostas, SISU-ENEM está um pouco abaixo, indicando que essa talvez seja diferente das outras duas também.

Para certificar que há diferença entre pelo menos uma das curvas de sobrevivência foi feito o teste *logRank*.

As hipóteses do teste são:

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

Tabela 5: Resultados do teste de *logRank* de Forma de ingresso

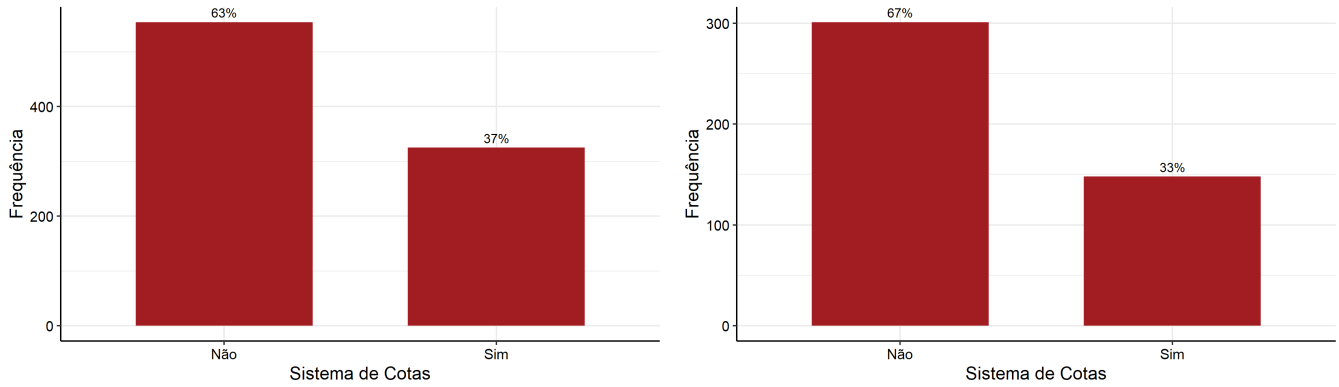
Banco de Dados	**Estatística do teste**	**Graus de liberdade**	**P-valor**
Banco Completo	$34,7$	3	$1 \cdot 10^{-7}$
Banco reduzido	$28,6$	3	$2 \cdot 10^{-6}$

Conclui-se com o resultado do teste, levando em consideração um nível de significância de 5% que há pelo menos uma das formas de ingresso com uma curva de sobrevivência diferente na variável para ambos os bancos.

4.1.4 Sistema de cotas

Sistema de cotas é uma variável binária na qual indica se o estudante fez uso de um dos sistemas de cotas previstos na Universidade de Brasília, sendo que a variável é divida em dois fatores: Sim ou Não. Ou seja, se Sim o aluno fez uso do sistema de cotas ou Não o aluno não fez uso do sistema de cotas.

Para avaliar como estão distribuídos os alunos de ambos bancos de dados para a variável Sistema de cotas foi feito o gráfico de barras abaixo:

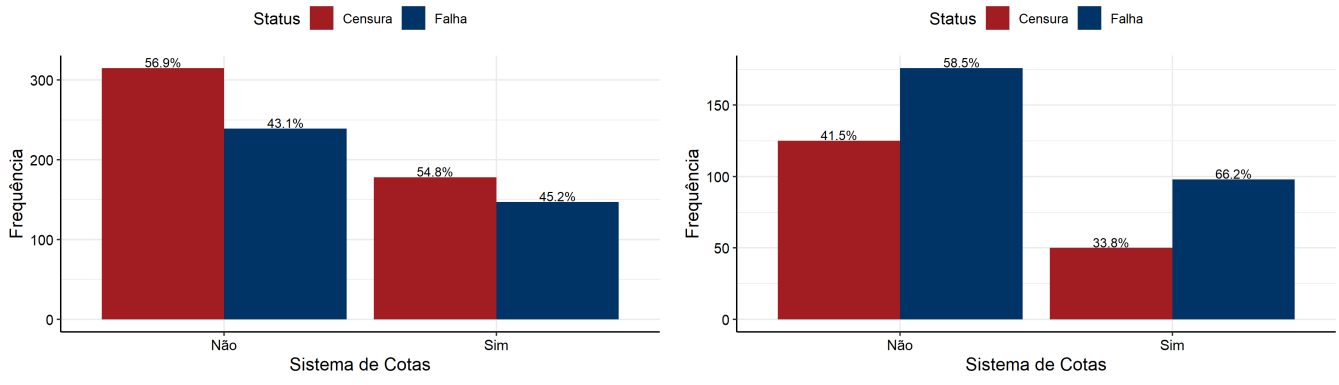


(a) Banco completo (b) Banco Reduzido

Figura 12: Gráficos de barras para Sistema de cotas para ambos os bancos

Novamente apesar da diferença entre a escala dos gráficos nota-se um comportamento muito semelhante em relação a proporção nas quais estão dividas as classes do banco , por volta de 63% para estudantes que não fizeram uso do sistema de cotas e 37% que utilizaram-se de algum sistema de cotas.

Para identificar como a variável interage com o Status foi feito o gráfico a seguir de barras de Sistema de cotas por Status:



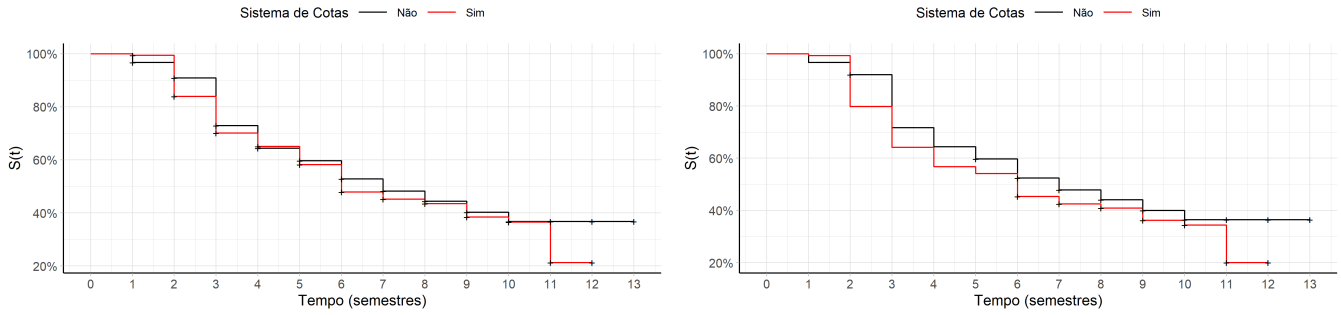
(a) Banco completo (b) Banco Reduzido

Figura 13: Gráficos de barras para Sistema de cotas vs Status para ambos os bancos

É interessante observar que enquanto tanto para alunos que usaram algum tipo de sistema cotas quanto para os que não usaram a distribuição quanto a Status no banco completo é semelhante em ambos os casos. Observa-se na Figura 13a os valores para tempo de censura ficam próximos a 55% enquanto os valores de tempo de falha ficam em 45%.

Ao observar a Figura 13b nota-se que para ambos os casos do uso do sistema de cotas o tempo de falha tem maior proporção do tempo de censura, mas não apresentam valores relativos tão próximos como na Figura 13a. Nota-se que para os alunos que não usaram sistema de cotas observa-se tempo de censura com $41,5\%$ e tempo de falha $58,5\%$ porém para os que usaram o sistema de cotas tem-se tempo de censura igual a $33,8\%$ e tempo de falha $66,2\%$.

Abaixo apresenta-se as curvas de sobrevivência para sistema de cotas para os dois bancos:



(a) Banco completo (b) Banco Reduzido

Figura 14: Gráficos de curvas de sobrevivência por Sistema de Cotas para ambos os bancos

Ao observar a 13 tem-se a impressão de que curva de sobrevivência de quem usa o sistema de cotas é abaixo da curva de sobrevivência dos que não usam, para afirmar algo com mais precisão foi feito o teste de *logRank*.

As hipóteses do teste são:

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

Tabela 6: Resultados do teste de *logRank* de Sistema de cotas

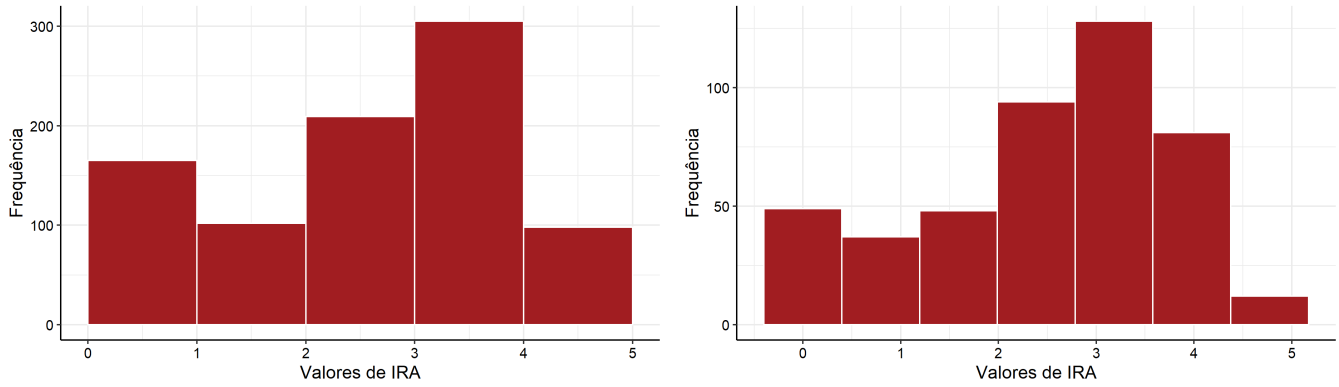
Banco de Dados	**Estatística do teste**	**Graus de liberdade**	**P-valor**
Banco Completo	$1,3$	1	$3 \cdot 10^{-1}$
Banco reduzido	$3,0$	1	$9 \cdot 10^{-2}$

Considerando o nível de significância de 5%, pode-se dizer que para nenhum dos dois bancos a curva de sobrevivência tem uma diferença significativa entre os alunos que usaram o sitema de cotas e aqueles que não

4.1.5 Índice de rendimento acadêmico (IRA)

Índice de rendimento acadêmico é uma variável numérica com amplitude de 0 a 5.É utilizada pela Universidade de Brasília para atribuir uma métrica ao rendimento de seus alunos, sendo que a nota 5 é considerada o melhor desempenho possível e 0 o pior.

Para verificar como da-se a distribuição do IRA para os bancos de dados foi feito o histograma abaixo:

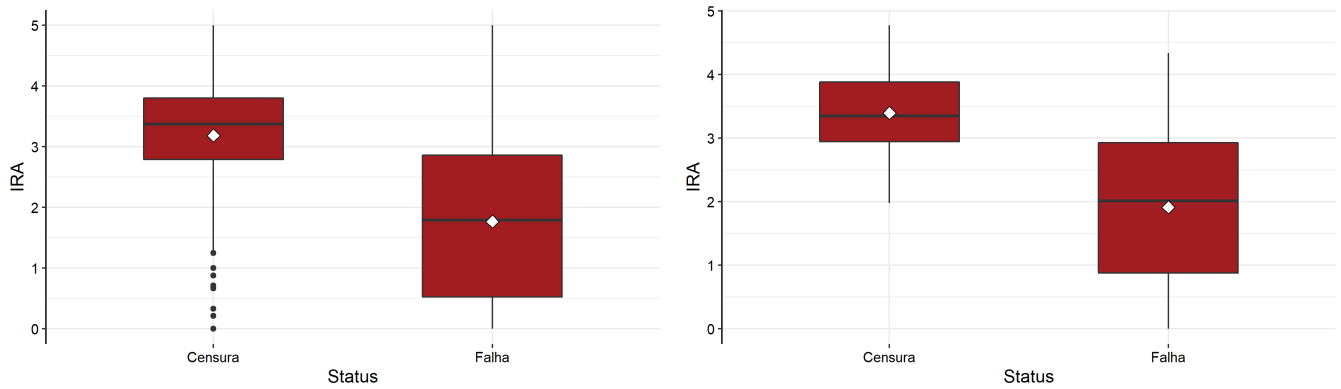


(a) Banco completo (b) Banco Reduzido

Figura 15: Gráficos de histograma para Índice de rendimento acadêmico para ambos os bancos

Ao olhar para a Figura 15 nota-se que há uma maior concentração de alunos com IRA entre os valores de 3 e 4, seguido pelos valores de 2 e 3. Como era de se esperar o IRA acima de 4 ou abaixo de 1 é reservado para poucos alunos. Porém algo que chama atenção é o fato de a faixa de IRA com valores de 0 ter uma frequência maior que o para 1 a 2. Isso ocorre para ambos os bancos.

Para verificar a relação de IRA com a variável Status foi feito *boxplots* de IRA por Status:



(a) Banco completo (b) Banco Reduzido

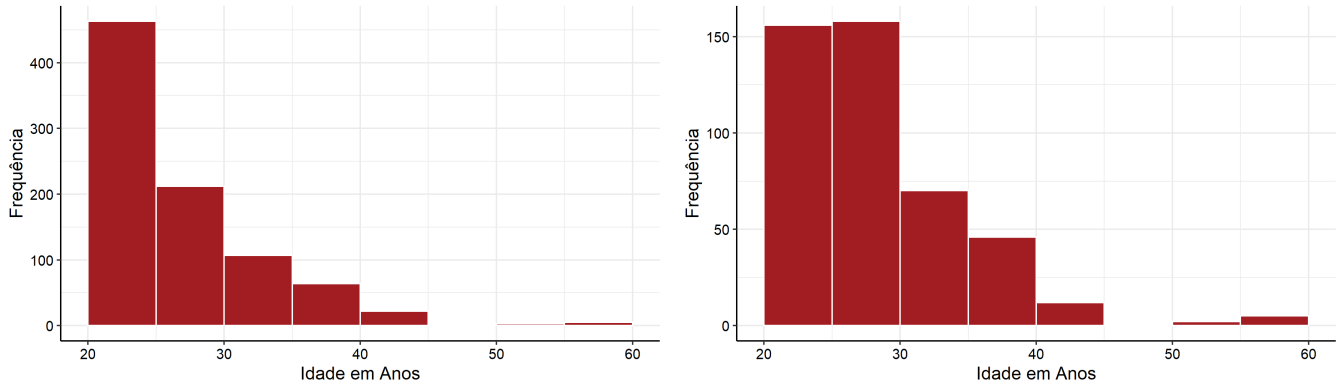
Figura 16: Gráficos *boxplot* para Índice de rendimento acadêmico vs Status para ambos os bancos

Ao observar a Figura 16 nota-se que ambos os bancos apresentam caixas muito diferentes para os casos de tempo de falha e tempo de censura. É nítido que para os valores de tempo de censura possuem, de modo geral, maiores valores para IRA.Também nota-se que a dispersão dos dados para tempo de falha é maior que para tempo de censura, ou seja, os alunos de tempo de censura são mais homogêneos entre si do que os alunos com tempo de falha. Porém, ao observar a Figura 16a tem-se que há valores discrepantes abaixo, ou seja, observações de alunos que tiveram tempo de censura porém com valores baixíssimos de IRA. Acontece também que o valor máximo da caixa de tempo de falha é justamente o valor máximo de IRA. Já para o caso da Figura 16b nota-se que não há mais os valores *outliers* para o caso de alunos tempo de censura e no caso dos alunos de tempo de falha o valor máximo não é mais igual a 5, encontra-se um pouco abaixo. Ou seja, ao observar alunos apenas com mais tempo de curso o IRA torna-se ainda mais determinante na diferenciação entre censura e falha.

4.1.6 Idade (anos)

A variável Idade foi construída conforme apresentado em 3.4.

Para verificar a distribuição de idade nos bancos de dados foi feito os seguintes gráficos:

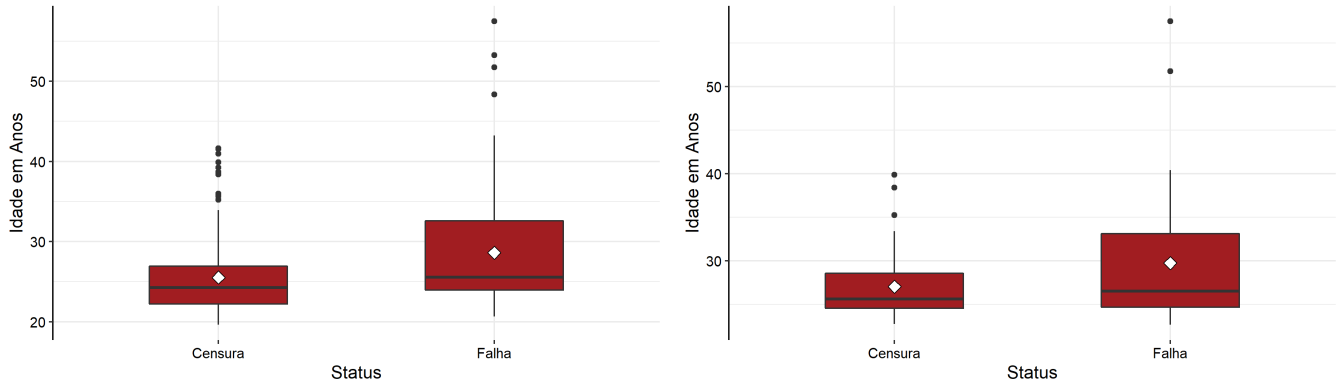


(a) Banco completo (b) Banco Reduzido

Figura 17: Gráficos de histograma para Idade para ambos os bancos

Considerando que o banco completo tem observações de período de entrada mais recentes, até 2019/2, enquanto o banco reduzido por sua vez tem período de entrada até 2016/2 espera-se encontrar uma distribuição de alunos mais velhos no banco reduzido, isso se confirma ao observar os histogramas na Figura 17.

Afim de apresentar a relação da variável Idade com Status foi feito os seguintes gráficos de *boxplot*:



(a) Banco completo (b) Banco Reduzido

Figura 18: Gráficos *boxplot* para Idade vs Status para ambos os bancos

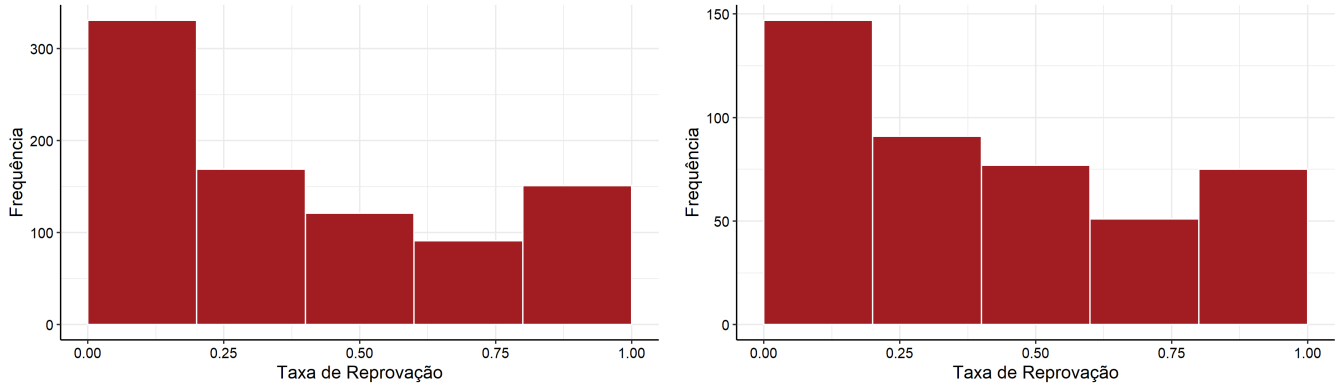
Observa-se que as caixas entre os alunos de tempo de censura e de falha são diferentes, porém não é tão acentuado. Ambas as caixas estão concentradas em valores

abaixo dos 30 anos, no caso de tempos de censura pelo menos 75% visto que o terceiro quartil está abaixo dessa idade, para o tempo de censura essa porcentagem cai pois o terceiro quartil se apresenta um pouco acima dos 30 anos. Novamente as observações com tempo de censura apresentam-se mais homogêneos entre si que os alunos com tempo de falha.

4.1.7 Taxa de reprovação

Taxa de reprovação é uma variável que foi construída, não sendo original do banco. O processo de construção da variável é detalhado na subseção 3.4. É uma variável que tem como intensão medir a proporção de créditos que o aluno efetivamente cursou até o fim a disciplina e obteve uma das menções que a Universidade de Brasília classifica como reprovação.

Para verificar a distribuição de taxa de reprovação foi realizado os gráficos abaixo:

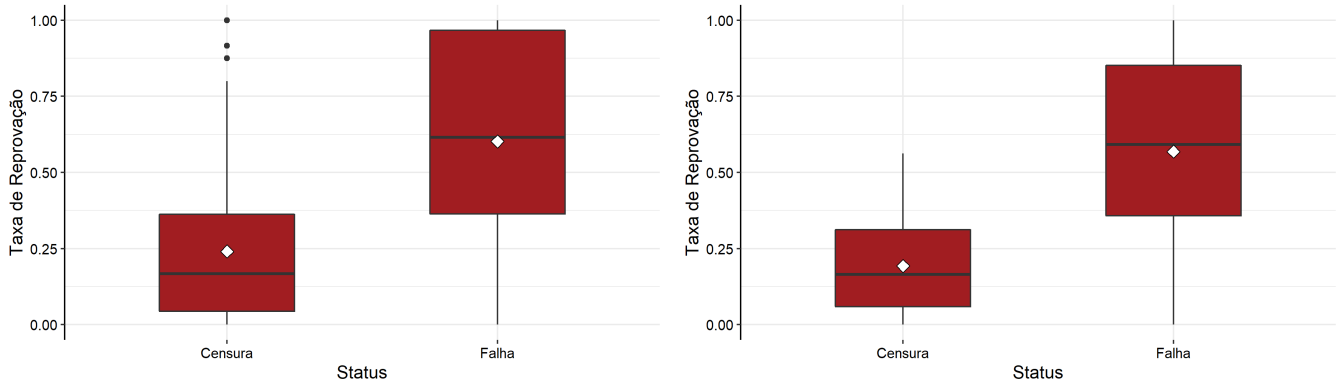


(a) Banco completo (b) Banco Reduzido

Figura 19: Gráficos de histograma para Taxa de reprovação para ambos os bancos

Nota-se na Figura 19 que a classe de valores de 0 a 0,2 é a que tem a maior concentração de alunos no caso de ambos os bancos. Chama a atenção também a quantidade alunos com taxa de reprovação entre 0,8 e 1, pois essa é a área mais crítica para a variável, uma vez que significa que o aluno em questão reprovou pelo menos 80% dos créditos que cursou até então.

Afim de investigar a relação entre taxa de reprovação e Status foi feito os gráficos a seguir:



(a) Banco completo (b) Banco Reduzido

Figura 20: Gráficos *boxplot* para Taxa de reprovação vs Status para ambos os bancos

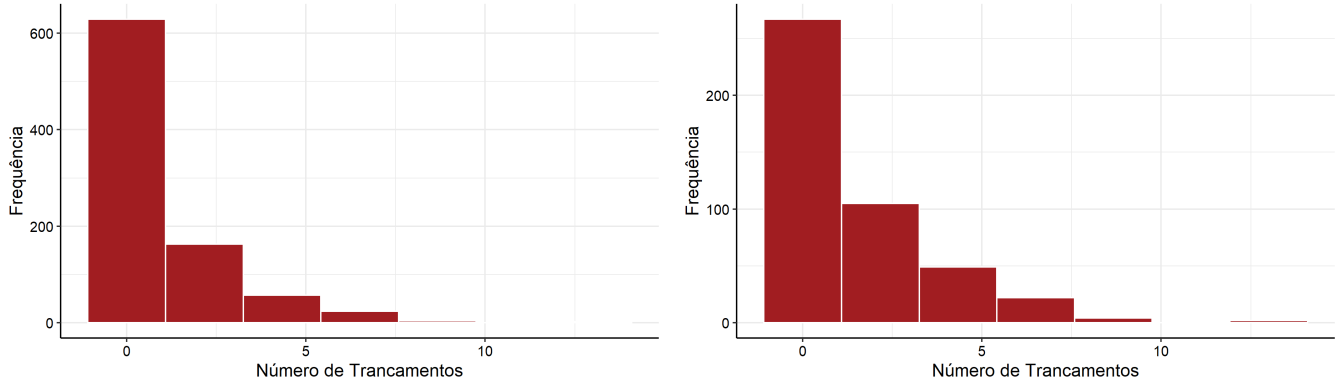
Ao observar a Figura 16 observa-se que ambos os bancos apresentam comportamento distinto para observações com tempo de falha e de censura. Começa pela dispersão das dados com tempo de falha, novamente, é muito maior do que para tempo de censura. Também há uma diferença no posicionamento das caixas, uma vez que o terceiro quartil do tempo de censura encontra-se no mesmo valor de taxa de reprovação do que o primeiro quartil do tempo de falha, ou seja, 75% das observações de tempo de censura encontra-se abaixo de $0,375$ de taxa de reprovação enquanto 75% das observações com tempo de falha estão acima desse valor, isso ocorre para ambos os bancos.

Entre as diferenças das caixas de um banco para o outro, enquanto que a dispersão dos dados é maior para as caixas de banco completo, destaque para os valores atípicos superiores em tempo de censura, no caso do banco reduzido essa dispersão é menor e não apresenta nenhum valor discrepante para nenhuma das caixas.

4.1.8 Total de trancamentos

Total de trancamentos é uma variável quantitativa que conta quantas disciplinas o aluno trancou no período que esteve matriculado no curso até o tempo observado no banco de dados. A construção em dessa variável está detalhado na subseção 3.4.

Abaixo apresentação o gráfico de histograma para o total de trancamentos para ambos os bancos:

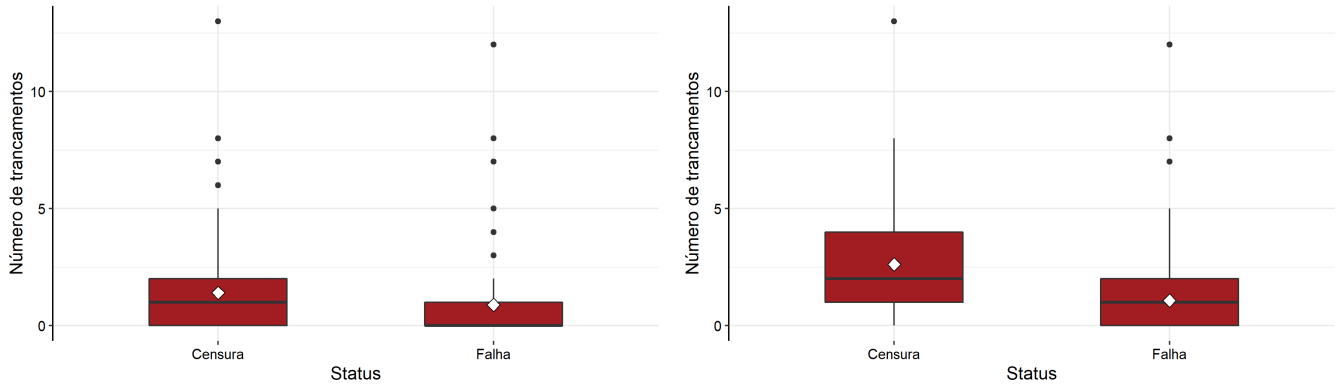


(a) Banco completo (b) Banco Reduzido

Figura 21: Gráficos de histograma para Taxa de reprovação para ambos os bancos

Observa-se que há uma grande concentração em valores baixos de trancamentos em ambos os bancos e o formato da distribuição para o total de trancamentos é muito semelhante. O gráfico então sugere uma concentração em valores próximos de zero para total de trancamentos, ou seja, os alunos não costumam trancar diversas disciplinas no curso.

Para visualização da relação entre o total de trancamentos e Status apresenta-se abaixo o gráfico de *boxplot* de total de trancamentos e Status.



(a) Banco completo (b) Banco Reduzido

Figura 22: Gráficos *boxplot* para Taxa de reprovação vs Status para ambos os bancos

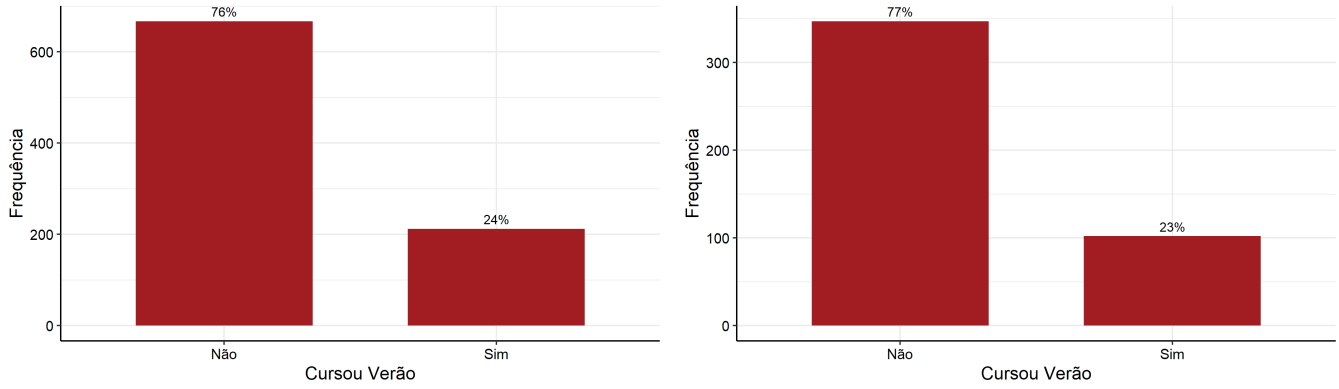
Ao observar a Figura 22 em primeiro momento não parece haver muita diferença entre o número de trancamentos das observações com tempo de falha e tempo de censura, principalmente, se observar a Figura 22a. De modo geral, tem-se que os alunos no banco de dados costumam ter menos de 5 trancamentos por curso visto que em ambos os bancos o terceiro quartil fica abaixo da marca de 5 e os valores atípicos que aparecem acima do valor limite das caixas.

Porém com mais atenção nota-se que a caixa de tempo de censura para ambos os bancos possui valores maiores para o total de trancamentos. Na Figura 22a nota-se que a mediana está próxima de zero para observações com tempo de falha. A diferença fica ainda mais evidente para o banco reduzido quando observa-se 22b, no qual a caixa de tempo de censura está inteira deslocada para cima, com a mediana superior ao terceiro quartil das observações com tempo de falha.

4.1.9 Cursou verão

A variável é a respeito da informação se o aluno cursou ou não verão, sendo assim uma variável categórica. A descrição detalhada da criação da variável cursou verão está explicitada na subseção 3.4.

Para identificar como está distribuído os dados a respeito da variável cursou verão para ambos os bancos foi feito o seguinte gráfico de barras:



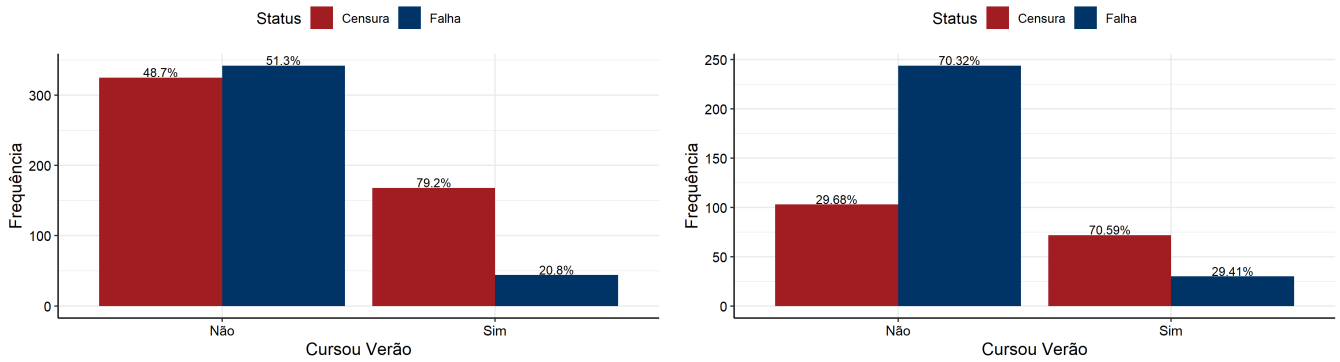
(a) Banco completo (b) Banco Reduzido

Figura 23: Gráficos de barras para Cursou Verão para ambos os bancos

Nota-se que apara ambos os bancos os dados estão distribuídos de forma semelhante na frequência relativa em relação a se cursou verão ou não. Observa-se na figura 23a para o banco completo 76% das observações não cursaram nenhuma disciplina no verão enquanto para o banco reduzido apresenta-se 77% de observações nessa situação.

Para analisar a relação entre se o aluno cursou verão ou não e a variável de Status

sobre o tempo de censura e de falha foi feito o gráfico de barras de cursou verão por Status:

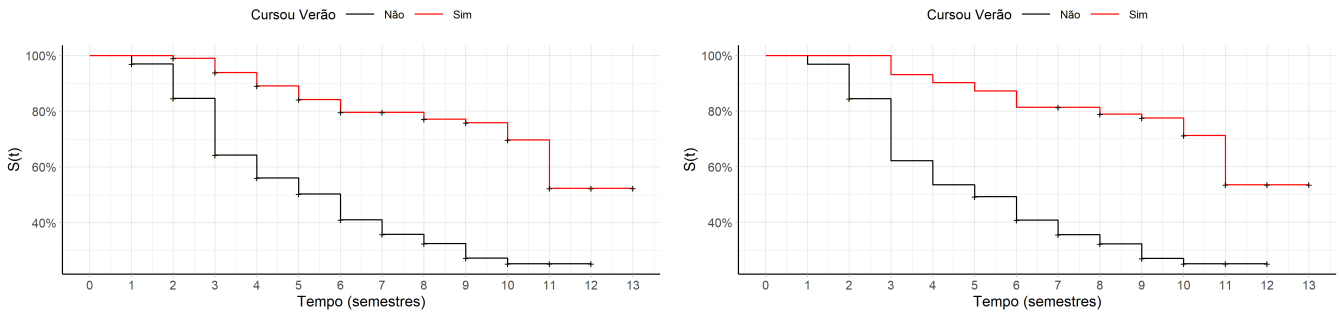


(a) Banco completo (b) Banco Reduzido

Figura 24: Gráficos de barras para Cursou Verão vs Status para ambos os bancos

Ao voltar a atenção a Figura 24a observa-se que para o banco completo há uma vantagem relativa para tempo de falha entre as observações que não cursaram verão, por outro lado, quando se olha para os alunos que cursaram verão uma maior proporção de tempo de censura. No caso do banco reduzido, que tem de forma relativa mais falhas que censuras se comparado com o banco completo essa diferença fica ainda mais clara. Na Figura 24b enquanto que para alunos que estão em não cursou disciplinas no verão apresentam mais de 70% de tempo de falha, o oposto ocorre para os que cursaram verão, mais de 70% de tempo de censura.

Para investigar melhor o efeito da variável cursou verão tem-se os gráficos de curva de sobrevivência para alunos de cursou verão ou não:



(a) Banco completo (b) Banco Reduzido

Figura 25: Gráficos de curvas de sobrevivência por Cursou verão para ambos os bancos

Ao observar as curvas de sobrevivência na Figura 25 nota-se que há uma distância entre as curvas de sobrevivência de cursou verão e no caso das observações em não cursou verão. No caso, a curva de sobrevivência para observações que não cursaram verão parece estar abaixo da curva para os que cursaram verão. Para identificar de fato que essa

diferença entre as curvas de sobrevivência são significativas foi feito o uso do teste *logRnak*.

As hipóteses do teste são:

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

Tabela 7: Resultados do teste de *logRank* Cursou Verão

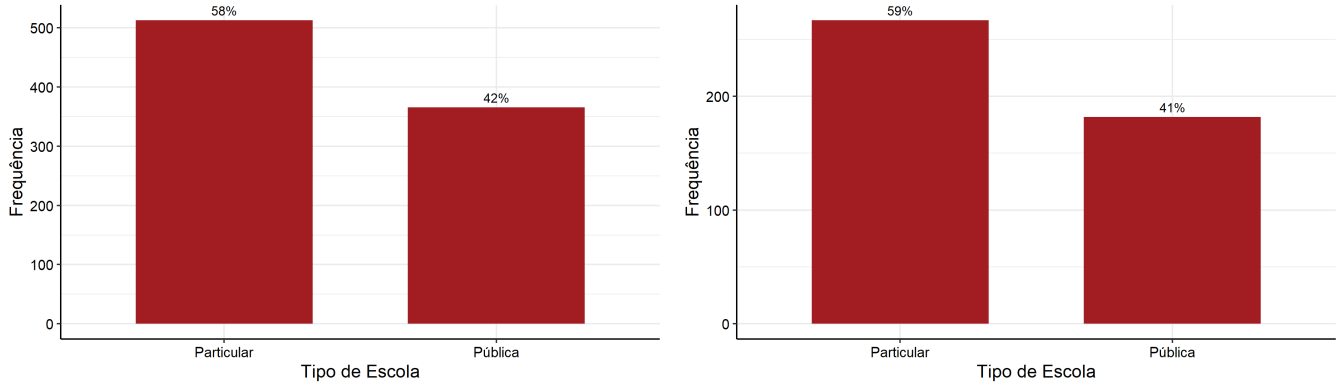
Banco de Dados	Estatística do teste	Graus de liberdade	P-valor
Banco Completo	$89,2$	1	$1 \cdot 10^{-16}$
Banco reduzido	$59,4$	1	$1 \cdot 10^{-14}$

Considerando os resultados do teste de hipótese considerando que o p-valor virtualmente igual a 0 para ambos os bancos, pode-se afirmar que as curvas de sobrevivência são diferentes para o caso de observações que cursaram verão e que não cursaram.

4.1.10 Escola (Pública ou Privada)

Essa variável indica se o aluno estudou em uma escola pública ou privada antes de sua vida acadêmica, fica a critério do aluno responder isso na entrada da Universidade sendo que é requisitado que o aluno considere o ensino médio como referência para responder.

Para melhor entendimento de como estão distribuídos os dados a respeito da escola que frequentaram no ensino médio foi feito o seguinte gráfico:



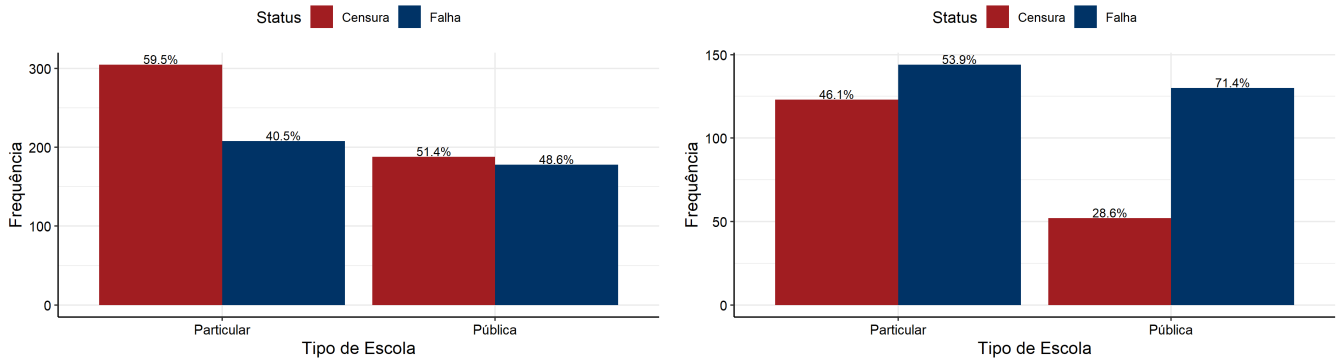
(a) Banco completo (b) Banco Reduzido

Figura 26: Gráficos de barras para Escola para ambos os bancos

Observa-se que a distribuição em relação ao tipo de escola é semelhante para os dois bancos, completo ou reduzido. Encontra-se 58% e 42% para escola particular e

pública respectivamente, no banco completo. Enquanto para o banco reduzido os valores encontrados são de 59% e 41% para particular e pública respectivamente.

Para entender a distribuição dos dados em relação a escola e também ao tempo de falha e censura foi feito a seguinte visualização:



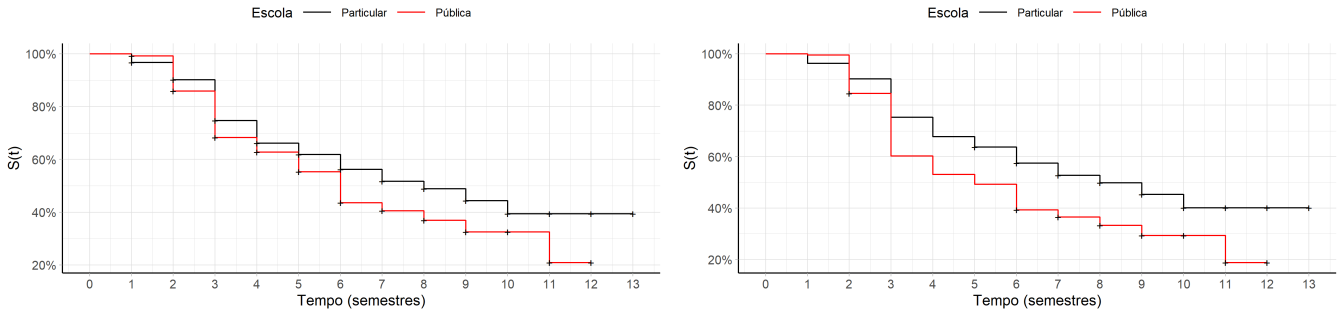
(a) Banco completo (b) Banco Reduzido

Figura 27: Gráficos de barras para Escola vs Status para ambos os bancos

É interessante notar na Figura 27a que enquanto os alunos que marcaram estudaram em escolas pública possuem uma proporção praticamente igual de tempo de falha e censura, para os alunos da escola particular o tempo de censura encontra-se em maior proporção clara do que o tempo de falha.

O que acontece na Figura 27b é que o para os alunos que marcaram escola pública fica evidente que são maioria em tempo de falha dentro de seu grupo, sendo que o valor para tempo de falha é de $71,4\%$ enquanto para tempo de censura é de $28,6$. Já para os alunos de escola particular o tempo de falha é de $53,9\%$ das observações enquanto o tempo de censura é de $48,1\%$.

Afim de verificar a curva de sobrevivência para os dois grupos de diferentes tipo de escola foi feito o gráfico abaixo:



(a) Banco completo (b) Banco Reduzido

Figura 28: Gráficos de curvas de sobrevivência por Escola para ambos os bancos

Ao observar a Figura 28 nota-se que parece que a curva de sobrevivência para observações de escola pública é menor que a curva para escola particular.

Para verificar a diferença apresentada no gráfico foi feito o teste *logRank*.

As hipóteses do teste são:

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

Tabela 8: Resultados do teste de *logRank* Escola

Banco de Dados	**Estatística do teste**	**Graus de liberdade**	**P-valor**
Banco Completo	$7,9$	1	$5 \cdot 10^{-3}$
Banco reduzido	$14,6$	1	$1 \cdot 10^{-4}$

Considerando um nível de significância de 5% pode-se afirmar que para ambos os bancos há uma diferença na curva de sobrevivência entre alunos de escola particular e privada.

4.2 Correlação entre variáveis

Nesta subseção será apresentado uma análise de correlação entre algumas variáveis, dado que não é ideal utilizar variáveis que sejam correlacionadas no mesmo modelo.

4.2.1 Índice de rendimento acadêmico e Taxa de reprovação

Primeiramente, serão consideradas duas variáveis que apresentam alguma forma de avaliar o desempenho do aluno. Enquanto o Índice de rendimento acadêmico é uma das variáveis originais do banco de dados, Taxa de reprovação foi criada conforme explicado na subseção 3.4.

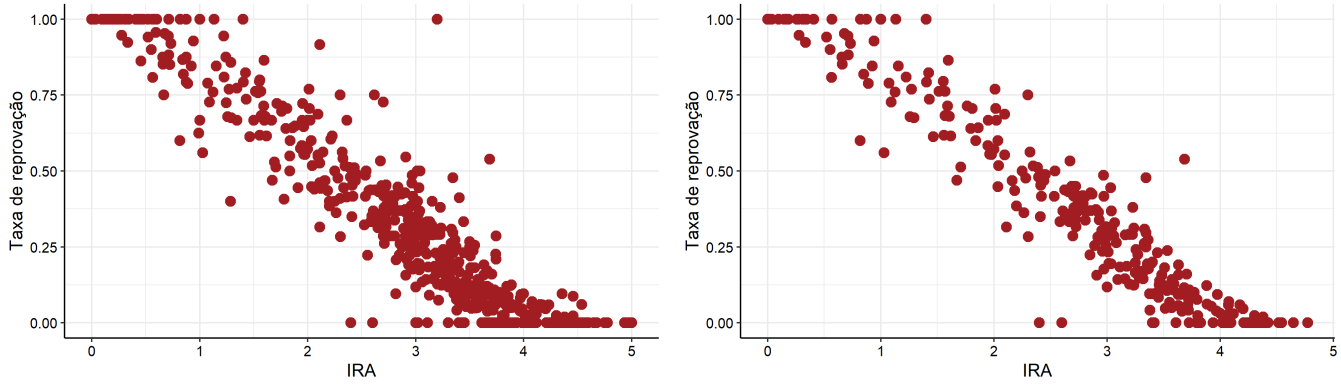
Considerando que são duas variáveis quantitativas contínuas, foi calculado a medida de correlação de Pearson para identificar o grau de correlação entre as variáveis.Esse processo foi feito para ambos os bancos.

Tabela 9: Correlação de Pearson entre IRA e Taxa de reprovação

Banco de Dados	$\boldsymbol{\rho}$
Banco Completo	$-0,9486$
Banco Reduzido	$-0,9536$

Pelos valores do coeficiente de correlação de Pearson (ρ) para os bancos, há evidências de correlação linear entre as variáveis IRA e Taxa de reprovação. Além disso, uma correlação forte e negativa, ou seja, inversamente proporcionais.

Também foi feito o gráfico de dispersão para as variáveis:



(a) Banco completo (b) Banco Reduzido

Figura 29: Gráfico de dispersão para IRA e Taxa de reprovação para ambos os bancos

Os gráficos da Figura 29 apresentam um padrão de reta decrescente na diagonal principal, acompanhando o resultado que foi obtido pela medida resumo do coeficiente de correlação. Confirmando que as variáveis estão mesmo correlacionadas linearmente.

4.2.2 Escola e Sistema de Cotas

No caso de Escola e Sistema de Cotas, ambas são variáveis categóricas com dois fatores. Há suspeita que estejam associadas dado que um dos sistemas de cotas possíveis tem como critério o estudo em escolas públicas durante o ensino médio. Dessa forma, para verificar essa suspeita foi feito o teste de independência χ^2 de Pearson na tabela de contingência de escola por sistema de cotas. Isso foi feito para ambos os bancos.

Tabela 10: Tabela de contingência Escola por Cursou verão banco completo

Escola	Cursou Verão	
	Não	Sim
Particular	475	38
Pública	79	287

Tabela 11: Tabela de contingência Escola por Cursou verão banco reduzido

Escola	Cursou Verão	
	Não	Sim
Particular	244	23
Pública	57	125

As hipóteses do teste são:

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

Tabela 12: Resultados do teste de independência χ^2 para Escola e Sistema de cotas

Banco de Dados	**Estatística do teste**	**Graus de liberdade**	**P-valor**
Banco Completo	$459,13$	1	$2 \cdot 10^{-16}$
Banco reduzido	$174,01$	1	$2 \cdot 10^{-16}$

Considerando esses resultados com o p-valor virtualmente igual a zero para ambos os bancos, pode-se afirmar que as varáveis Escola e Sistema de cotas não são independentes.

4.3 Modelo para o banco completo

Nesta subseção será apresentado o processo de seleção do modelo para o banco completo. Primeiramente foi feito a curva de sobrevivência utilizando o estimador de máxima verossimilhança Kaplan-Meier.

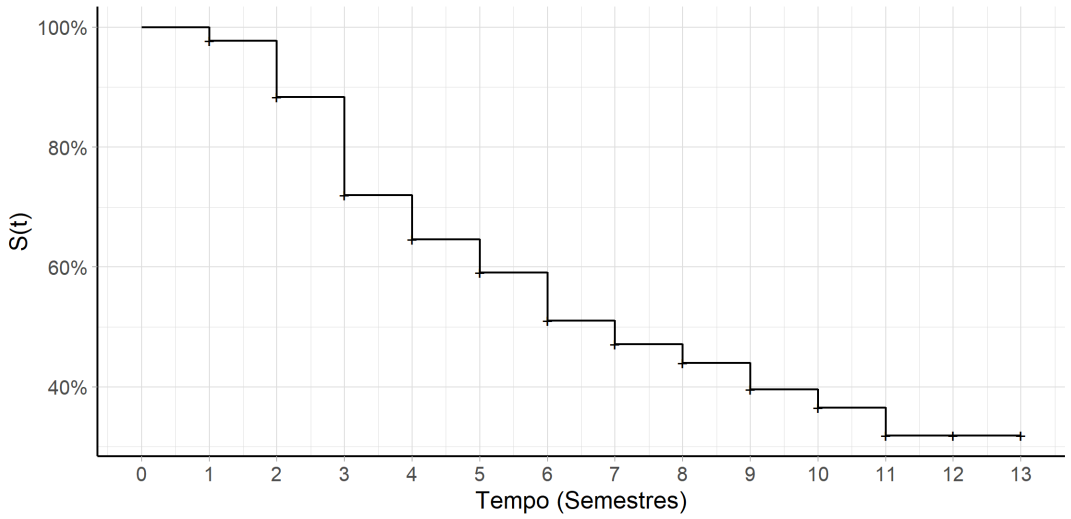


Figura 30: Curva de Sobrevivência por Kaplan-Meier para banco cheio

Considerando a Figura 30 observa-se que ao final do tempo a curva de sobrevivência não chega a zero, mas sim a 40% indicando a presença de valores censurados.

4.3.1 Selecionar distribuição

O primeiro passo é encontrar a distribuição de probabilidade mais adequada para ajustar os dados. Como dito anteriormente, a distribuição de ponto de partida para os dados será a Log-Logística discreta, visto que em primeiro momento tem-se a hipótese de que os dados possuem tempo discreto. Foi considerado como comparação, também o ajuste da distribuição Log-Logística para tempos contínuos. Foi feito a curva de sobrevivência para ambas as distribuições e a comparação com Kaplan-Meier.

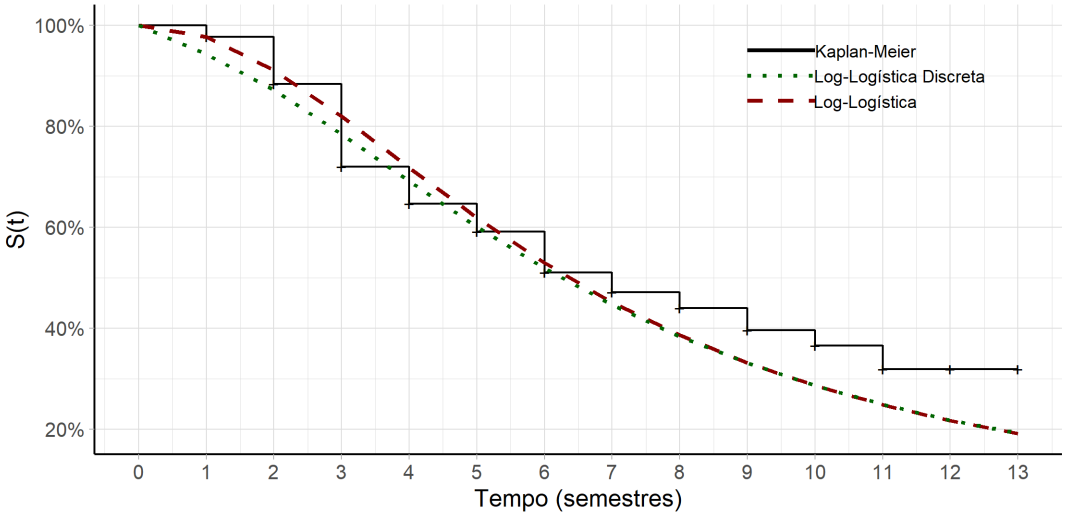


Figura 31: Comparação entre Log-Logística Discreta e Log-Logística contínua

Considerando a Figura 31 nota-se que a distribuição Log-Logística discreta quanto a Log-Logística contínua se ajustam aos dados de maneira muito parecida. Porém, há uma pequena vantagem para a Log-Logística contínua (em vermelho).

Nesse caso, a distribuição candidata para ajustar os dados seria contínua. Sendo assim, desse modo foi feito mais uma comparação entre a Log-Logística contínua e a Log-Normal.

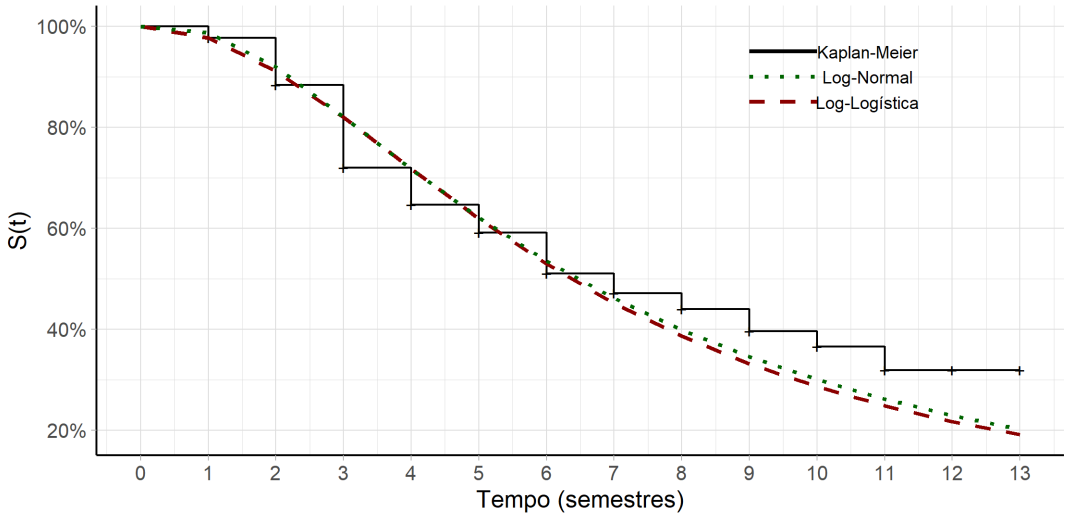


Figura 32: Comparação entre Log-Logística e Log-Normal

Observa-se no caso da Figura 32 que a distribuição Log-Normal apresenta um ajuste um pouco melhor que a Log-Logística. Embora, as distribuições estejam muito próximas no recurso gráfico.

Para comparação também foi feito o uso dos critérios de avaliação como foi explícito em 2.5.2.

Tabela 13: Medidas de informação para o banco completo

Distribuição	AIC	$\boldsymbol{AIC_C}$	BIC
Log-Logística Discreta	$2537,30$	$2537,32$	$2547,86$
Log-Logística Contínua	$2449,98$	$2449,99$	$2459,54$
Log-Normal	2423.44	$2423,46$	$2433,00$

Considerando os valores dos critérios, a Log-Normal obteve valores melhores em todos os critérios em relação as outras distribuições. A distribuição escolhida para modelar os dados então é a Log-Normal, dado que graficamente e pelos resultados dos critérios de informação é a que teve os melhores resultados.

4.3.2 Seleção de Variáveis

A seleção de variáveis é um processo crucial para a modelagem de dados, após escolher a distribuição de probabilidade que melhor se ajusta aos dados. As variáveis podem definir a precisão do modelo, assim como sua interpretação. A estratégia utilizada para construção do modelo foi o StepWise, no qual se começa com a variável mais significativa e a cada inclusão de variável se verifica uma possível exclusão de uma variável. A inclusão ou exclusão de variáveis do modelo foi feita a partir do resultado de testes de razão da verossimilhança apresentado na subseção 2.5.1 para modelos encaixados. Além disso, para modelos não encaixados, foi utilizado os critérios de informação apresentados na subseção 2.5.2. Foi considerado a distribuição Log-Normal para ajuste de todos os modelos a partir daqui.

Primeiramente, foi feito uma investigação do efeito e da significância das variáveis de maneira isolada, ou seja, modelos com apenas uma variável para cada uma das variáveis do banco. Sendo assim, pode-se obter o efeito que essa variável tem na probabilidade de sobrevivência de maneira isolada. Os resultados estão na Tabela 14:

Tabela 14: Coeficientes estimados, erro padrão, estatística do teste e p-valor dos modelos contendo apenas uma variável explicativas para o banco de dados completo

Parâmetro	Estimativa	Erro Padrão	Estatística do teste	P-valor
$\beta_{Sexo\,Masculino}$	$-0,15$	$0,11$	$-1,34$	$1,8 \cdot 10^{-1}$
$\beta_{Forma\,de\,Ingresso\,PAS}$	$0,51$	$0,10$	$4,84$	$1,3 \cdot 10^{-6}$
$\beta_{Forma\,de\,Ingresso\,SISU}$	$0,39$	$0,11$	$3,69$	$2,2 \cdot 10^{-4}$
$\beta_{Forma\,de\,Ingresso\,Vestibular}$	$0,55$	$0,10$	$5,77$	$7,9 \cdot 10^{-9}$
$\beta_{Sistema\,de\,Cotas\,Sim}$	$-0,07$	$0,07$	$-0,99$	$3,2 \cdot 10^{-1}$
β_{IRA}	$0,38$	$0,02$	$17,10$	$2,0 \cdot 10^{-16}$
β_{Idade}	$-0,03$	$5 \cdot 10^{-3}$	$-6,01$	$1,9 \cdot 10^{-9}$
$\beta_{Taxa\,de\,reprovacao}$	$-1,51$	$0,08$	$-17,42$	$2,0 \cdot 10^{-16}$
$\beta_{Total\,de\,trancamentos}$	$0,23$	$0,02$	$12,23$	$2,0 \cdot 10^{-16}$
$\beta_{Cursou\,verao\,sim}$	$0,88$	$0,08$	$10,47$	$2,0 \cdot 10^{-16}$
$\beta_{Escola\,publica}$	$-0,15$	$0,06$	$-2,33$	$2,0 \cdot 10^{-2}$

Destaca-se da Tabela 14 que apenas 2 variáveis deram como não significativas, considerando uma significância de 5%, ao observar o p-valor: Sexo e Sistema de cotas. Mesmo que sozinhas no modelo não sejam significativas, ainda serão testadas a cada rodada do método pra verificar se em conjunto com demais variáveis o efeito e significância são alterados.

Foi tomado a decisão de ter dois modelos para o banco de dados, visto que a variável IRA e Taxa de reprovação são correlacionadas e ambas apresentaram ter um

efeito muito significativo na curva de sobrevivência. Ao observar o β_{IRA} apresenta valor de $0,38$ e p-valor virtualmente igual a 0, enquanto para $\beta_{Taxa\,de\,reprovacao}$ apresenta valor de $-1,51$ p-valor aproximadamente igual a 0. Ou seja, a relação do IRA é positiva quanto maior o IRA maior a probabilidade de sobrevivência, enquanto a Taxa de reprovação é negativa, quanto maior a taxa de reprovação menor a probabilidade de sobrevivência.

A respeito da variável Forma de ingresso, nota-se que a referência para o modelo é o fator Outras forma de ingresso. Nesse caso, todas as demais formas apresentam uma probabilidade de sobrevivência maior, com destaque para Vestibular que apresenta o maior coeficiente positivo e o menor p-valor dos fatores, sendo respectivamente iguais a: $0,55$ e $7,9 \cdot 10^{-9}$.

A variável Idade apresenta um valor de coeficiente negativo, indicando que quanto maior a idade menor a probabilidade de sobrevivência. Porém, vale lembrar que esse valor está na segunda casa decimal.

Nota-se que a variável Total de trancamentos possui um efeito positivo, que é contra intuitivo. Pois assim, quanto mais trancamentos maior a probabilidade de sobrevivência. Isso pode ocorrer devido a concentração de valores próximos a zero para total de trancamentos, e pelo fato de vários alunos com tempo de censura apresentarem na Figura 22a pelo menos 1 trancamento.

Enquanto que para a variável Cursou verão o fator para referência para o modelo é Não, logo o fator Sim apresenta um efeito positivo em relação a probabilidade de sobrevivência. Isto é, se o estudante cursou verão, ele terá maior probabilidade de sobrevivência.

A variável Escola, possui dois fatores, sendo que o de referência para o modelo é Particular. Logo, o efeito de Escola quando igual a Pública dar negativo significa que tem uma probabilidade de sobrevivência menor se comparada com Privada.

Outra decisão que precisou ser tomada foi quanto ao uso da variável Escola ou da variável Sistema de cotas. Visto que as duas também apresentam uma associação, as duas não poderia entrar no mesmo modelo. Então foi utilizado os critérios de informação AIC, BIC e AIC_C para decidir qual das duas permaneceria no modelo.

4.3.3 Modelo Final incluindo IRA

A seleção começa incluindo a variável IRA, pois essa apresentou o maior coeficiente em módulo em conjunto como menor p-valor.E então fazendo teste para a inclusão ou exclusão de novas variáveis. Durante a construção do modelo se chega em dois modelos alternativo com as mesmas variáveis exceto que um inclui a variável Escola e outro, no lugar de Escola, inclui-se a variável Sistema de Ingresso. Utilizando os critérios de

informação AIC, BIC e AIC_C chega-se no seguinte resultado:

Tabela 15: Medidas de informação seleção de variáveis com IRA para o banco completo

Modelo	**AIC**	$\boldsymbol{AIC_C}$	**BIC**
$Modelo_{Sistema\,de\,Cotas}$	$1877,86$	$1877,88$	$1887,42$
$Modelo_{Escola}$	$1895,84$	$1895,85$	$1905,392$

Considerando os resultados da Tabela 15, o modelo selecionado como o modelo final para o banco completo incluindo IRA o seguinte modelo:

Tabela 16: Coeficientes estimados, erro padrão, estatística do teste e p-valor do modelo final iniciando-se com a variável IRA para o banco de dados completo

Parâmetro	**Estimativa**	**Erro Padrão**	**Estatística do teste**	**P-valor**
β_0	$0,68$	$0,07$	$9,87$	$2,0 \cdot 10^{-16}$
β_{IRA}	$0,21$	$0,02$	$9,67$	$2,0 \cdot 10^{-16}$
$\beta_{Sistema\,de\,Cotas\,Sim}$	$-0,39$	$0,09$	$-4,16$	$3,2 \cdot 10^{-5}$
$\beta_{Total\,de\,trancamentos}$	$0,17$	$0,02$	$10,96$	$2,0 \cdot 10^{-16}$
$\beta_{Cursou\,verao\,sim}$	$0,47$	$0,06$	$7,44$	$1,3 \cdot 10^{-13}$
$\beta_{Forma\,de\,Ingresso\,PAS}$	$0,33$	$0,08$	$4,03$	$5,5 \cdot 10^{-5}$
$\beta_{Forma\,de\,Ingresso\,SISU}$	$0,21$	$0,08$	$2,58$	$9,8 \cdot 10^{-3}$
$\beta_{Forma\,de\,Ingresso\,Vestibular}$	$0,35$	$0,07$	$4,92$	$8,8 \cdot 10^{-7}$
$\beta_{IRA:\,Sistema\,de\,cotas\,sim}$	$0,19$	$0,04$	$5,22$	$1,8 \cdot 10^{-7}$
$\log Scale$	$-0,63$	$0,03$	$-17,48$	$2,0 \cdot 10^{-16}$

Primeiramente é importante ressalta a interação entre a variável IRA e Sistema de cotas. Em um primeiro momento, enquanto se construía o modelo foi visto que quando IRA e Sistema de cotas entravam em conjunto no modelo o sinal do coeficiente de $\beta_{Sistema\,de\,cotas\,sim}$ mudava para positivo, ou seja, os alunos que fizeram uso do sistema de cotas passam a ter maior probabilidade de sobrevivência. Considerando essas variáveis importantes para o modelo, foi feito a interação de Sistema de cotas e IRA. Ao inserir a interação observa-se que Sistema de cotas volta a ter o coeficiente negativo, enquanto a interação possui sinal positivo, ambas significantes considerando um nível de significância de 5%.

Nota-se que β_{IRA} tem valor de $0,21$, desse modo pode-se afirmar que para valores maiores de IRA, o aluno possui maior probabilidade de sobrevivência, ou seja, de não cometer evasão.

Para $\beta_{Sistema\,de\,cotas\,sim}$ tem o valor de $-0,39$ nota-se que o coeficiente é negativo, logo os alunos que fizeram uso de sistema de cotas tem uma probabilidade de sobrevivência menor do que os alunos que não tiveram sistema de cotas.

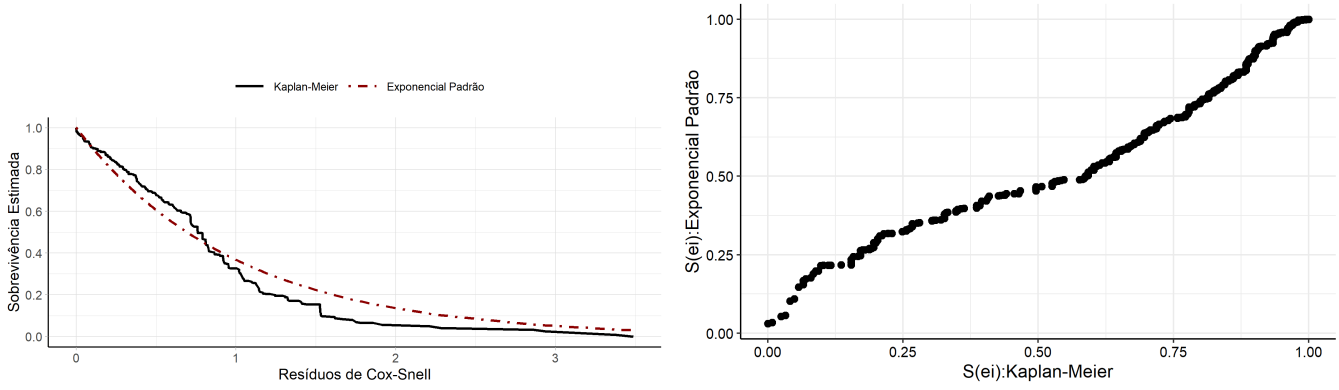
Considerando $\beta_{Total\,de\,trancamentos}$ observa-se um valor positivo, assim como no modelo de apenas uma variável, o que isso é contra intuitivo, quanto mais trancamentos maior a probabilidade de sobrevivência do aluno, ou seja, de não cometer evasão.

Ao observar o $\beta_{Cursou\,verao\,sim}$ com o valor de $0,47$, observa-se que os alunos que cursaram alguma disciplina no verão tem uma probabilidade de cometer evasão menor que alunos que nunca cursaram nenhuma disciplina no verão.

Considerando formas de ingresso, observa-se que a forma de ingresso que está de referência para o modelo como na Tabela 14 foi "Outras formas de ingresso"e que todas as demais possuem probabilidade de sobrevivência maior,ou seja, que "Outras formas de ingresso"possui maior probabilidade de cometer evasão que as demais. Destaque para $\beta_{Forma\,de\,ingresso\,Vestibular}$ que possui o maior coeficiente com valores de $0,35$.

Ao que se considerar a interação entre sistema de cotas e IRA no $\beta_{IRA:\,Sistema\,de\,cotas\,sim}$ com valor de $0,19$ apresentando um valor positivo, nota-se que em relação ao que se acompanha o IRA o sistema de cotas muda. Ou seja, para alunos com sistema de cotas e maiores valores de IRA apresenta maior probabilidade de sobrevivência estimada que alunos sem o sistema de cotas.

Para validar os resultados é necessário verificar a adequabilidade do modelo, que será realizado por meio da análise de resíduos. A Figura 33a apresenta o comportamento dos resíduos de Cox-Snell.



(a) Resíduo de Cox-snell (b) Estimada x Kaplan-Meier

Figura 33: Gráficos de resíduos Cox-Snell modelo incluindo IRA banco completo

Considerando a Figura 33a o modelo parece adequado visto que segue a distribuição exponencial padrão. É evidente que em momentos nos quais a curva de sobrevivência estimada se distância da exponencial padrão, porém ainda é próxima o suficiente para considerar que o modelo tem boa adequação global.

Além disso, ao observar a Figura 33b apresenta com algumas distorções, uma figura próxima de uma reta principalmente para valores maios altos.

4.3.4 Modelo Final incluindo Taxa de reprovação

Como dito anteriormente, também foi considerado um modelo para a variável de taxa de reprovação.

A seleção começa incluindo a variável Taxa de reprovação, pois essa apresentou o maior coeficiente em módulo em conjunto como menor p-valor como nota-se na Tabela 14.E então fazendo teste para a inclusão ou exclusão de novas variáveis.

Assim como no modelo que inclui a variável IRA, outra decisão importante foi considerando as variáveis escola e sistema de cotas. Visto que as duas variáveis entram juntas no mesmo modelo, foi tomada a decisão de acordo com os critérios de informação.

Tabela 17: Medidas de informação seleção de variáveis com Taxa de reprovação banco completo

Modelo	**AIC**	$\boldsymbol{AIC_C}$	**BIC**
$Modelo_{Sistema\,de\,Cotas}$	$1812,27$	$1812,28$	$1821,82$
$Modelo_{Escola}$	$1831,88$	$1831,89$	$1841,44$

Considerando os resultados da Tabela 17 o modelo final escolhido ficou com a variável sistema de cotas.

Os coeficientes do modelo final são mostrados na Tabela 18.

Tabela 18: Coeficientes estimados, erro padrão, estatística do teste e p-valor do modelo final iniciando-se com a variável Taxa de reprovação para o banco de dados completo

Parâmetro	**Estimativa**	**Erro Padrão**	**Estatística do teste**	**P-valor**
β_0	$1,60$	$0,07$	$20,35$	$2,0 \cdot 10^{-16}$
$\beta_{Taxa\,de\,reprovacao}$	$-0,92$	$0,08$	$-10,44$	$2,0 \cdot 10^{-16}$
$\beta_{Sistema\,de\,Cotas\,Sim}$	$0,45$	$0,10$	$4,48$	$7,6 \cdot 10^{-6}$
$\beta_{Total\,de\,trancamentos}$	$0,17$	$0,02$	$10,30$	$2,0 \cdot 10^{-16}$
$\beta_{Cursou\,verao\,sim}$	$0,46$	$0,07$	$7,21$	$5,8 \cdot 10^{-13}$
$\beta_{Forma\,de\,Ingresso\,PAS}$	$0,37$	$0,08$	$4,61$	$4,1 \cdot 10^{-6}$
$\beta_{Forma\,de\,Ingresso\,SISU}$	$0,21$	$0,08$	$2,53$	$1,1 \cdot 10^{-2}$
$\beta_{Forma\,de\,Ingresso\,Vestibular}$	$0,35$	$0,07$	$4,89$	$1,0 \cdot 10^{-6}$
$\beta_{Taxa:\,Sistema\,de\,cotas\,sim}$	$-0,74$	$0,15$	$-4,85$	$1,2 \cdot 10^{-6}$
$\log Scale$	$-0,64$	$0,04$	$-17,36$	$2,0 \cdot 10^{-16}$

Novamente foi feito uma interação utilizando a variável Sistema de cotas, porém dessa vez com Taxa de reprovação. Essa interação segue os mesmos preceitos que foram discutidos no modelo que inclui a variável IRA. Dito isso é notório a semelhança entre o modelo da Tabela 18 e o modelo apresentado na Tabela 16 em relação as variáveis

selecionadas e também aos coeficientes de cada um.

Há porém diferenças crucias, pois com a inclusão de Taxa de reprovação, temos um $\beta_{Taxa\,de\,reprovacao}$ com valor de $-0,92$ apresentando assim o valor negativo muito significativo. Ou seja, para valores maiores de taxa de reprovação maior é a probabilidade do aluno cometer evasão.

Outra diferença que merece a atenção é a em relação ao $\beta_{Sistema\,de\,Cotas\,Sim}$ pois observa-se o valor de $0,45$. Um valor de sinal positivo, ao contrário do efeito da variável quando observada isoladamente. Isso leva a uma interpretação de que os alunos com sistema de cotas igual a "Sim"tem maior probabilidade de sobrevivência que alunos que não utilizaram o sistema de cotas.

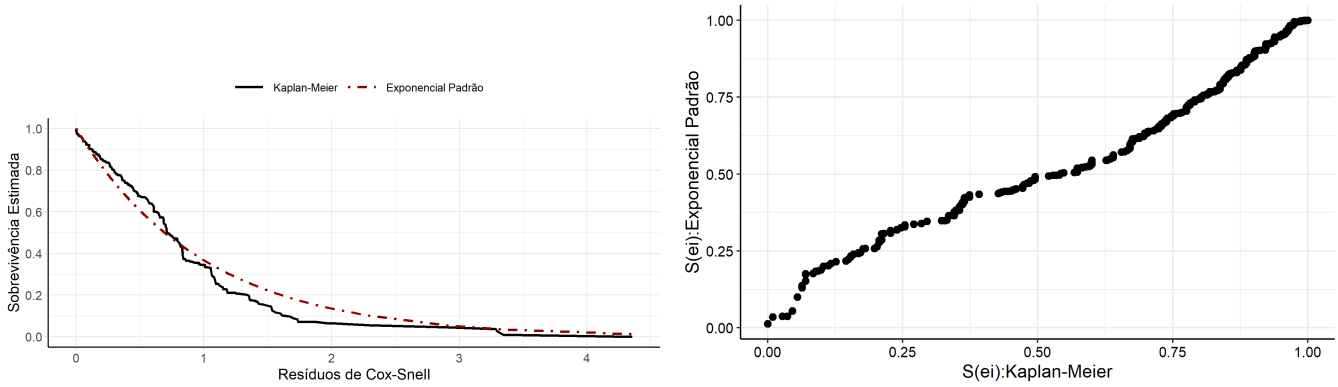
Para entender essa mudança é preciso observa a interação, que encontra-se em $\beta_{Taxa:Sistema\,de\,cotas\,sim}$ com valor de $-0,74$ negativo. Resumindo, para aqueles alunos que tem maior taxa de reprovação e sistema de cotas igual a sim possuem menor probabilidade de sobrevivência, com isso, uma explicação para o a alteração do sinal no $\beta_{Sistema\,de\,Cotas\,Sim}$ é que agora o coeficiente avalia os alunos com baixa taxa de reprovação com sistema de cotas igual a sim, sendo assim, alunos com baixa taxa de reprovação e de sistema de cotas tem maior probabilidade de sobrevivência que alunos com taxa de reprovação baixa e que não fazem uso do sistema de cotas.

Além disso, nota-se que $\beta_{Total\,de\,trancamentos}$ tem valor de $0,17$ positivo o que significa que quanto maior o valor em total de trancamentos maior é a probabilidade de sobrevivência, ou de não cometer evasão.

Também observa-se que $\beta_{Cursou\,verao\,sim}$ possui um valor de $0,46$, ou seja, aqueles que cursam verão possuem uma probabilidade de sobrevivência maior que alunos que não cursaram disciplinas no verão.

A respeito da Forma de Ingresso, ao observar para os diferentes β dessa variável nota-se que todas formas de ingresso possuem maior probabilidade de sobrevivência se comparadas com "Outras formas de ingresso".

Para realizar a validação dos resultado é necessário verificar a adequabilidade do modelo, que será feito por meio da análise de resíduos. A Figura 34a apresenta o a curva de sobrevivência dos resíduos de Cox-Snell.



(a) Resíduo de Cox-snell (b) Estimada x Kaplan-Meier

Figura 34: Gráficos de resíduos Cox-Snell modelo incluindo Taxa de reprovação banco completo

Considerando a Figura 34a o modelo parece adequado no ajuste global visto que segue, apesar das diferenças, uma forma semelhante a exponencial padrão.

Além disso, ao observar a Figura 34b apresenta com algumas distorções, uma figura próxima de uma reta principalmente para valores maios altos.

4.4 Modelo para o banco reduzido

Como dito na subseção 3.6, foi feito a divisão dos bancos em dois. Foi apresentado até então resultados da modelagem para o banco completo. A partir daqui será apresentado os resultados para o banco reduzido.

4.4.1 Seleção de distribuição

O primeiro passo é identificar a distribuição mais adequada para a modelagem. Dessa forma, foi considerado em primeiro momento que os dados poderiam ter distribuição Log-Logística discreta e foi feito uma comparação com a Log-Logística contínua.

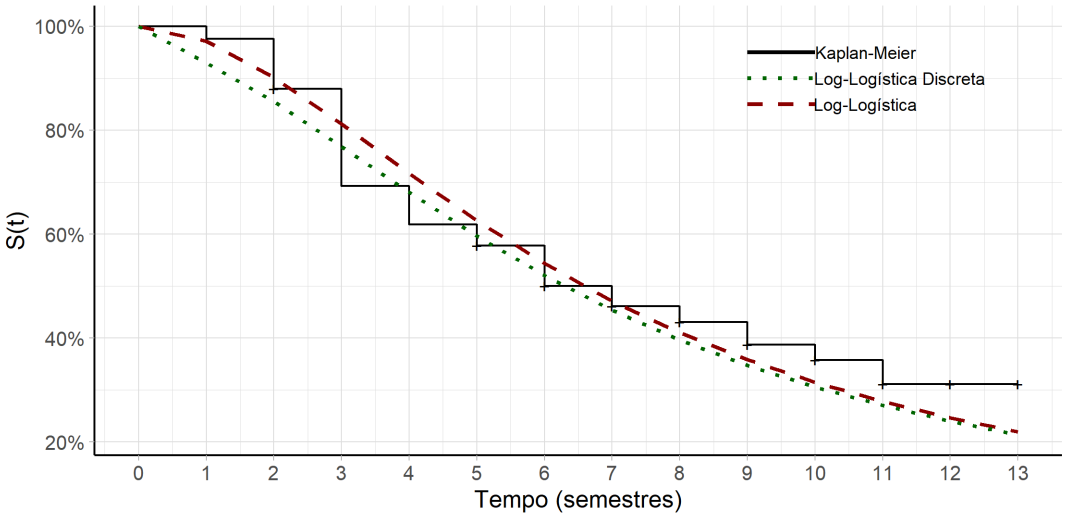


Figura 35: Comparação entre Log-Logística Discreta e Log-Logística contínua

Observa-se na Figura 35 uma leve vantagem para a distribuição Log-Logística contínua.

Considerando então que os dados podem ser analisados por uma distribuição contínua foi feito ainda a comparação com a distribuição Log-Normal.

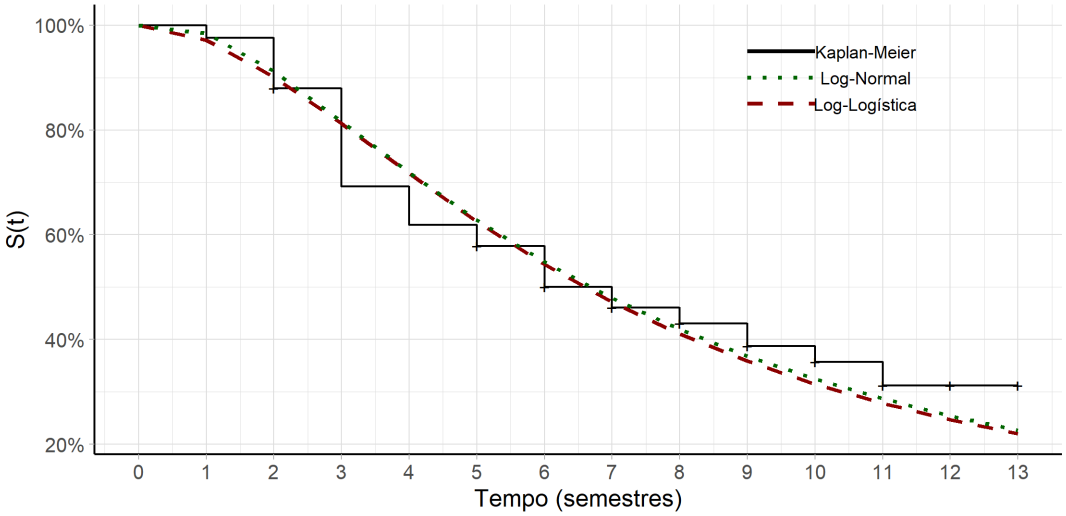


Figura 36: Comparação entre Log-Logística e Log-Normal

Na Figura 36 nota-se que há uma vantagem ínfima para a distribuição Log-Normal.

Visto que a análise gráfica não é o suficiente para determinar a distribuição foi feito também uma comparação utilizando os critérios apresentados na subseção 2.5.2, os resultados estão na tabela a seguir.

Tabela 19: Medidas de informação para o banco reduzido

Distribuição	AIC	$\boldsymbol{AIC_C}$	BIC
Log-Logística Discreta	$1786,93$	$1786,96$	$1795,14$
Log-Logística Contínua	$1746,82$	$1746,85$	$1755,04$
Log-Normal	$1729,93$	$1729,96$	$1738,15$

Considerando os resultados apresentados a distribuição de probabilidade selecionada para ajustar os dados foi a Log-Normal. A respeito da distribuição pode-se ler na subseção 2.3.Todos os modelos ajustados a seguir consideram a distribuição Log-Normal.

4.4.2 Seleção de Variáveis

A seleção de variáveis é uma etapa de extrema importância para a modelagem estatística. Selecionar as variáveis explicativas mais significativas e com maior poder de explicação tendem a melhorar o modelo.

Primeiramente foi feito modelos utilizando apenas uma variável de forma isolada para verificar a significância e o efeito das variáveis explicativas quando isoladamente em um modelo.

Tabela 20: Coeficientes estimados, erro padrão, estatística do teste e p-valor dos modelos contendo apenas uma variável explicativas para o banco de dados reduzido

Parâmetro	**Estimativa**	**Erro Padrão**	**Estatística do teste**	**P-valor**
$\beta_{Sexo\,Masculino}$	$-0,18$	$0,16$	$-1,16$	$2,5 \cdot 10^{-1}$
$\beta_{Forma\,de\,Ingresso\,PAS}$	$0,62$	$0,15$	$4,06$	$5,0 \cdot 10^{-5}$
$\beta_{Forma\,de\,Ingresso\,SISU}$	$0,56$	$0,13$	$3,83$	$1,3 \cdot 10^{-4}$
$\beta_{Forma\,de\,Ingresso\,Vestibular}$	$0,69$	$0,05$	$5,33$	$9,7 \cdot 10^{-8}$
$\beta_{Sistema\,de\,Cotas\,Sim}$	$-0,16$	$0,09$	$-1,68$	$9,0 \cdot 10^{-2}$
β_{IRA}	$0,42$	$0,03$	$14,15$	$2,0 \cdot 10^{-16}$
β_{Idade}	$-0,04$	$6 \cdot 10^{-3}$	$-6,07$	$1,3 \cdot 10^{-8}$
$\beta_{Taxa\,de\,reprovacao}$	$-1,68$	$0,11$	$-14,72$	$2,0 \cdot 10^{-16}$
$\beta_{Total\,de\,trancamentos}$	$0,24$	$0,02$	$10,22$	$2,0 \cdot 10^{-16}$
$\beta_{Cursou\,verao\,sim}$	$0,98$	$0,11$	$8,86$	$2,0 \cdot 10^{-16}$
$\beta_{Escola\,publica}$	$-0,30$	$0,08$	$-3,38$	$7,2 \cdot 10^{-4}$

Nota-se pelos resultados apresentados na Tabela 20 que considerando um nível de significância de 5% apenas duas variáveis foram não significativas: Sexo e Sistema de cotas. A respeito dessas duas variáveis vale destaque para sistema de cotas que apresenta coeficiente estimado negativo.

As demais variáveis que foram significativas isoladamente no modelo, ressalta-se a Escola. Pois essa variável por mais que tenha um isoladamente um coeficiente de $-0,30$ com p-valor menor que os 5% utilizado de nível de significância, quando em conjunto com outras variáveis principalmente as mais significativas como IRA e Taxa de reprovação a variável Escola deixa de ser significativa.

Sendo assim, durante a seleção de variáveis, a variável Escola acabou não entrando para os modelos. Porém, a variável Sistema de cotas que apesar de isoladamente não ser significativa, com p-valor maior que o nível de significância de referência, ao inserida em conjunto com as demais variáveis então Sistema de cotas torna-se significativa. É importante ressaltar a diferença entre essa interação e a que aconteceu com os modelos feitos com o banco completo. No caso anterior, a variável Escola também era significativa na seleção de variáveis com as demais, porém Sistema de cotas ajustava um modelo que era melhor se comparado com os critérios de informação. No caso que ocorre com o banco reduzido, o processo de seleção de variáveis exclui a variável Escola sempre que essa entra no modelo.

4.4.3 Modelo Final Incluindo IRA

Como visto anteriormente, foi escolhido realizar a modelagem considerando separadamente as variáveis IRA e Taxa de reprovação. Nessa seção será apresentado o modelo final que foi selecionado incluindo a variável IRA para o banco reduzido.

Tabela 21: Coeficientes estimados, erro padrão, estatística do teste e p-valor do modelo final iniciando-se com a variável IRA para o banco de dados reduzido

Parâmetro	**Estimativa**	**Erro Padrão**	**Estatística do teste**	**P-valor**
β_0	$1,15$	$0,23$	$5,00$	$5,5 \cdot 10^{-7}$
β_{IRA}	$0,22$	$0,03$	$7,45$	$9,7 \cdot 10^{-14}$
$\beta_{Sistema\,de\,Cotas\,Sim}$	$-0,55$	$0,13$	$-4,15$	$3,4 \cdot 10^{-5}$
$\beta_{Total\,de\,trancamentos}$	$0,16$	$0,02$	$9,40$	$2,0 \cdot 10^{-16}$
$\beta_{Cursou\,verao\,sim}$	$0,47$	$0,08$	$5,66$	$1,5 \cdot 10^{-8}$
$\beta_{Forma\,de\,Ingresso\,PAS}$	$0,26$	$0,12$	$2,24$	$2,5 \cdot 10^{-2}$
$\beta_{Forma\,de\,Ingresso\,SISU}$	$0,29$	$0,10$	$2,84$	$4,6 \cdot 10^{-3}$
$\beta_{Forma\,de\,Ingresso\,Vestibular}$	$0,35$	$0,09$	$3,69$	$2,2 \cdot 10^{-4}$
β_{Idade}	$-0,02$	$0,01$	$-2,88$	$3,9 \cdot 10^{-4}$
$\beta_{IRA:\,Sistema\,de\,cotas\,sim}$	$0,19$	$0,05$	$3,87$	$1,1 \cdot 10^{-4}$
$\log Scale$	$-0,61$	$0,04$	$-13,79$	$2,0 \cdot 10^{-16}$

Ao observar os resultados na Tabela 21 tem uma principal diferença se comparado com a Tabela 16 é a inclusão da variável Idade ao modelo.

A variável que foi a primeira a ser incluída no modelo foi a variável IRA, com o coeficiente β_{IRA} com valor igual a $0,22$. Ou seja, para alunos com valores maiores de IRA, possuem probabilidade de não cometer evasão maior que alunos com valores menores de IRA.

Outro ponto interessante é a interação entre IRA e Sistema de cotas com coeficiente igual a $0,19$. Ou seja, para os alunos de sistema de cotas igual a "Sim"o valor de IRA impacta mais positivamente que em relação ao impacto do valor de IRA para alunos que não fazem uso do sistema de cotas.

Ao observar o $\beta_{Sistema\,de\,cotas\,Sim}$ tem o valor de $-0,55$ um valor expressivo e negativo, ou seja, o sistema de cotas igual a "Sim"probabilidade de cometer evasão maior que os alunos com sistema de cotas igual a "Não".

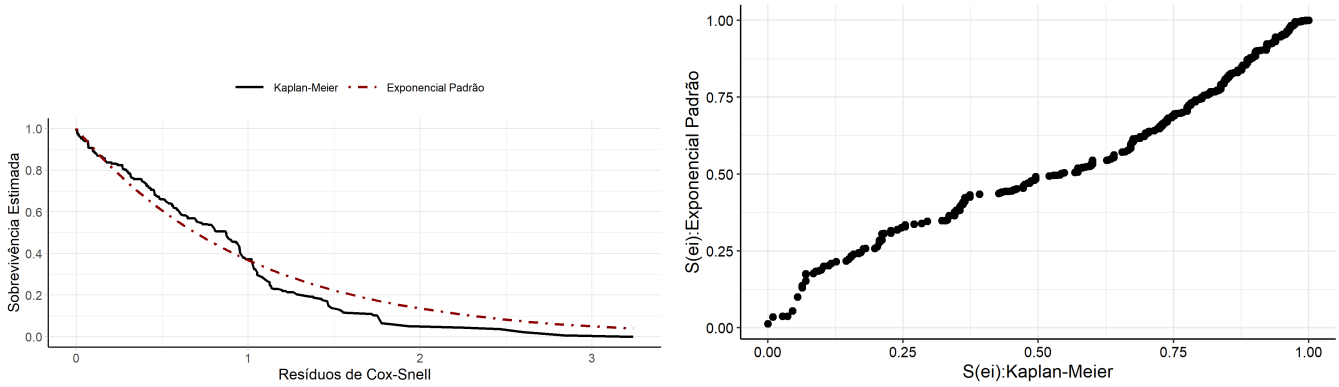
Considerando a variável Total de trancamentos tem um efeito positivo de $0,16$, ou seja, para maiores valores de total de trancamentos os alunos tem maior probabilidade de sobrevivência, isto é, maior probabilidade de não cometer evasão.

Nota-se que a variável Cursou verão, tem um coeficiente estimado de $0,47$. Dessa forma, os alunos que cursaram alguma disciplina em um semestre de verão tem maior probabilidade de sobrevivência do que alunos que nunca cursaram nenhuma disciplina em semestre de verão.

Voltando a atenção para Forma de ingresso, nota-se que todas as formas de ingresso no modelo tem um efeito positivo em relação a referência "Outras formas de ingresso". Mais precisamente pode-se dizer que em relação a "Outras formas de ingresso"tem-se que os alunos que ingressaram por meio de "PAS", "SISU"e "Vestibular"tem o coeficiente de $0,26$, $0,29$ e $0,35$ respectivamente, se comparado com a forma de ingresso de referência.

Ao voltar a atenção para a variável Idade, encontra-se um coeficiente de $-0,02$ sendo assim, alunos com maior idade tem menor probabilidade de sobrevivência, isto é, maior probabilidade de cometer evasão.

Para validar os resultados é importante verificar a adequação do modelo. A técnica mais usual para verificação de adequação de modelos é utilizar os resíduos de Cox-Snell.



(a) Resíduo de Cox-snell (b) Estimada x Kaplan-Meier

Figura 37: Gráficos de resíduos Cox-Snell modelo incluindo IRA banco reduzido

Considerando o resíduo de Cox-Snell parece que o modelo está bem ajustado globalmente, considerando que a Figura 37a a curva de sobrevivência estimada acompanha o formato da exponencial padrão com pequenos desvios. Também na Figura 37a apresenta-se em um formato parecido com de uma reta.

4.4.4 Modelo Final incluindo Taxa de reprovação

Ainda foi feito um modelo para o banco reduzido que inclui-se no modelo a variável Taxa de reprovação, uma vez que essa variável e IRA são fortemente correlacionadas e não entrariam juntas no modelo. Considerando Taxa de reprovação, os resultados para o modelo encontram-se na Tabela 22

Tabela 22: Coeficientes estimados, erro padrão, estatística do teste e p-valor do modelo final iniciando-se com a variável Taxa de reprovação para o banco de dados reduzido

Parâmetro	**Estimativa**	**Erro Padrão**	**Estatística do teste**	**P-valor**
β_0	$2,11$	$0,21$	$10,18$	$2,0 \cdot 10^{-16}$
$\beta_{Taxa\,de\,reprovacao}$	$-0,94$	$0,11$	$-7,87$	$3,4 \cdot 10^{-15}$
$\beta_{Sistema\,de\,Cotas\,Sim}$	$0,32$	$0,12$	$2,60$	$9,3 \cdot 10^{-3}$
$\beta_{Total\,de\,trancamentos}$	$0,15$	$0,02$	$8,74$	$2,0 \cdot 10^{-16}$
$\beta_{Cursou\,verao\,sim}$	$0,45$	$0,08$	$5,56$	$2,7 \cdot 10^{-8}$
$\beta_{Forma\,de\,Ingresso\,PAS}$	$0,31$	$0,11$	$2,66$	$7,8 \cdot 10^{-3}$
$\beta_{Forma\,de\,Ingresso\,SISU}$	$0,31$	$0,10$	$2,90$	$3,8 \cdot 10^{-3}$
$\beta_{Forma\,de\,Ingresso\,Vestibular}$	$0,35$	$0,09$	$3,46$	$5,4 \cdot 10^{-4}$
β_{Idade}	$-0,02$	$0,05$	$-2,75$	$5,9 \cdot 10^{-4}$
$\beta_{Taxa:\,Sistema\,de\,cotas\,sim}$	$-0,72$	$0,19$	$-3,65$	$2,6 \cdot 10^{-4}$
$\log Scale$	$-0,62$	$0,05$	$-13,74$	$2,0 \cdot 10^{-16}$

Considerando os resultados da Tabela 22 pode-se obter as seguintes interpretações.

Primeiramente a variável que mede a taxa de reprovação com coeficiente de $\beta_{Taxa\,de\,reprovacao}$ com valor de $-0,94$, sendo assim, para maiores valores de taxa de reprovação maior é a probabilidade estimada do aluno cometer evasão.

Em seguida, um ponto que chama a atenção é o fato de que sistema de cotas igual a "Sim"antes era negativo quando no modelo de uma única variável, porém agora muda de sinal. Dessa forma, sugere que o aluno que fez uso de sistema de cotas tem maior probabilidade de sobrevivência do que o aluno que não fez uso desse sistema. Porém, dado essa mudança de interpretação, deve-se observar que há uma interação no modelo entre taxa de reprovação e sistema de cotas que apresenta valor de coeficiente igual a $-0,72$ o que torna ambas as interpretações mais complexas. Pois uma das explicações para que o sinal de sistema cotas seja diferente no modelo em que se faz a interação é que a interação está com o efeito negativo em taxa de reprovação para o sistema de cotas igual a "Sim", ou seja, alunos com sistema de cota igual a sim que tenham maior taxa de reprovação tem probabilidade de sobrevivência menor que alunos com mesma taxa de reprovação porém de sistema de cota igual a "Não".

Em seguida passando para total de trancamentos que apresenta coeficiente de $0,15$, dessa forma, alunos com maior número de disciplinas trancadas durante o curso possuem maior probabilidade de não cometer evasão.

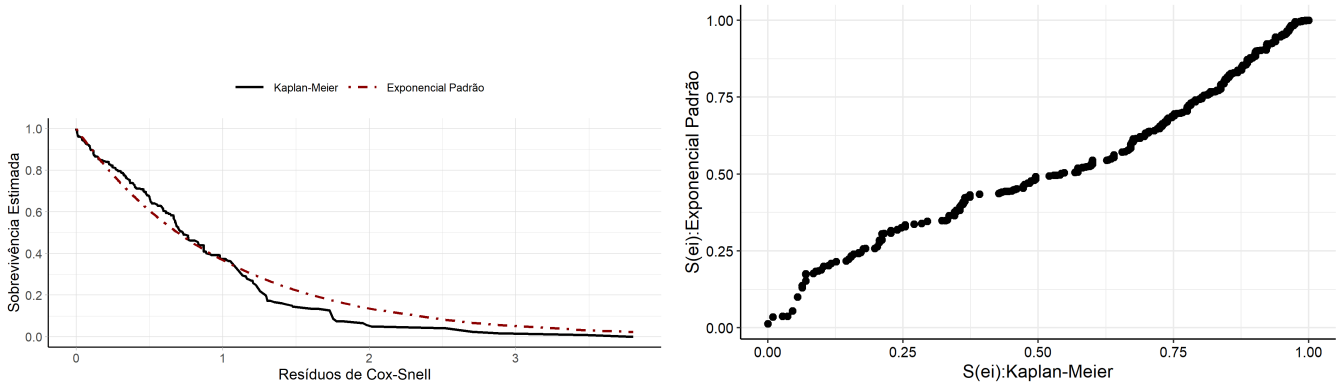
Quando se trata da variável que indica se o aluno cursou verão ou não, o modelo apresenta um coeficiente de $0,45$. Dessa forma, alunos que cursaram alguma disciplina no verão apresentam maior probabilidade de sobrevivência que alunos que jamais cursaram alguma disciplina no verão.

A respeito de forma de ingresso, o fator que é referência para o modelo é "Outras formas de ingresso"e observa-se que todos os demais fatores representados no modelo apresentam coeficientes positivos de $0,31$, $0,31$ e $0,35$ que representam os coeficientes para o sistema de ingresso "PAS", "SISU"e "Vestibular"respectivamente.

A idade é significativa no modelo com coeficiente de $-0,02$ o que representa que alunos de maior idade apresentam menor probabilidade de sobrevivência estimada, logo, maior probabilidade de cometer evasão.

É necessário verificar a adequação do modelo para que possa validar os resultados encontrados. Nesse sentido, utiliza-se o resíduos de Cox-Snell apresentados na subseção 2.6.1.

Na Figura 38 é apresentado os resíduos de Cox-Snell para o modelo apresentado na Tabela 22.



(a) Resíduo de Cox-snell modelo IRA banco reduzido (b) Estimada x Kaplan-Meier banco reduzido

Figura 38: Gráficos de resíduos Cox-Snell modelo incluindo Taxa de reprovação banco reduzido

Ao observa a Figura 38a tem-se que a curva dos resíduos acompanha a curva de uma exponencial padrão enquanto a Figura 38b observa uma linha que se assemelha a uma reta, logo, pode-se afirmar que o modelo está bem ajustado globalmente.

5 Conclusões e Considerações Finais

A motivação inicial deste trabalho surgiu de um problema real, a evasão escolar no curso de graduação de Ciência da Computação da Universidade de Brasília. Os dados foram medidos em semestres e apresentam três grupos: 1. Alunos que sofreram evasão, 2. Alunos que ainda estão ativos no curso e; 3. Alunos que se formaram. A partir desses dados, a variável resposta do estudo foi definida como o tempo, desde o período que o aluno entrou no curso de Ciência da Computação, até evadir do referido curso.

Neste contexto, o objetivo deste trabalho foi formular um modelo que considere os tempos de sobrevivência dos alunos do curso de Ciência da Computação. Para a análise dos dados foi utilizado o modelo de regressão Log-Normal como uma aproximação para tempos contínuos. Dado o contexto apresentado na subseção 3.6, o banco de dados foi dividido em dois: banco completo e banco reduzido, com observações com período de entrada até 2016/1. E como discutido nas subseções: 4.3 e 4.4; também foi considerado modelos que incia-se com duas variáveis distintas (IRA e Taxa de reprovação), que em razão da correlação entre elas, não poderiam ser incluídas no mesmo modelo. Nesse caso, foi feito então quatro modelos ao todo: 1. Modelo para o banco completo incluindo IRA, 2. Modelo para ao banco completo incluindo Taxa de reprovação, 3. Modelo para o banco reduzido incluindo IRA e; 4. Modelo para o banco reduzido incluindo Taxa de reprovação.

De acordo com as estimativas dos parâmetros, de maneira geral, verificou-se para modelos que incluíram a variável IRA que os alunos com IRA maior tem maior probabilidade de sobreviver à evasão. Além disso, para ambos os bancos, completo e reduzido, observou-se que um número de trancamentos maior também resulta em maior probabilidade de sobrevivência dos alunos. Outro resultado observado é que no caso da interação entre sistema de cotas e IRA, nota-se que alunos com sistema de cotas tem maior probabilidade de sobrevivência que os alunos sem o sistema de cotas ao considerar o aumento de IRA. Somado a isso, observou-se que alunos que cursaram uma disciplina no verão tem ,de modo geral, maior probabilidade de sobrevivência que alunos que não cursaram verão.

A respeito dos modelos que incluíram a variável taxa de reprovação, verificou-se, de modo geral, que alunos com maior taxa de reprovação possuem menor probabilidade de sobrevivência. Observou-se ainda que o coeficiente de sistema de cotas é positivo, o oposto ao que é apresentado para a variável sistema de cotas quando em um modelo isolado. Essa relação, entretanto, deve ser fruto da interação com taxa de reprovação e a interpretação mais correta é que para alunos com sistema de cotas e altos valores de taxa de reprovação observa-se menor probabilidade de sobrevivência. Além disso, as interpretações para as demais variáveis permaneceram as mesmas que para os modelos que incluíram IRA.

De maneira geral, os modelos de regressão Log-Normal tiveram bons ajustes para todos os modelos propostos e, principalmente, resultados coerentes. Como proposta para trabalhos futuros sugere-se:

- Ampliar a coleta de dados para pesquisar a possibilidade de outras covariáveis serem significativas para o tempo de sobrevivência dos alunos do curso de Ciência da Computação;
- Realizar estudo semelhante para outros cursos de graduação da Universidade de Brasília;
- Propor outros modelos de sobrevivência para análise dos dados.

Referências

ANDRADE, C. Y. de. Acesso ao ensino superior no brasil: equidade e desigualdade social. *Revista Ensino Superior Unicamp*, v. 6, p. 18–27, 2012. 10

CARVALHO, M. S. et al. *Análise de sobrevivência: teoria e aplicações em saúde*. [S.l.]: SciELO-Editora FIOCRUZ, 2011. 14

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. [S.l.]: Editora Blucher, 2006. 15, 17, 20, 25

FERNANDES, L. M. Inferencia bayesiana em modelos discretos com fracao de cura. 2013. 15

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. 10

JUNIOR, P. L.; SILVEIRA, F. L. d.; OSTERMANN, F. Análise de sobrevivência aplicada ao estudo do fluxo escolar nos cursos de graduação em física: um exemplo de uma universidade brasileira. *Revista brasileira de ensino de física*, SciELO Brasil, v. 34, n. 1, p. 1–10, 2012. 10

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. 16, 37

LOBO, M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. 2012. 10

NAKANO, E. Y.; CARRASCO, C. G. Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. *TEMA (São Carlos)*, v. 7, n. 1, p. 91–100, 2006. 11

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: ⟨https://www.R-project.org/⟩. 11, 27, 39

SANTOS, D. F. d. *Modelo de regressão log-logístico discreto com fração de cura para dados de sobrevivência*. Dissertação (Mestrado) — Universidade de Brasília, 2017. 3, 11, 13, 16, 18, 22, 23, 38

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. 10, 31

TADIKAMALLA, P. R.; JOHNSON, N. L. Systems of frequency curves generated by transformations of logistic variables. *Biometrika*, Oxford University Press, v. 69, n. 2, p. 461–465, 1982. 17, 19