Universidade Federal da Bahia

Instituto de Matemática e Estatística

Colegiado de Estatística

JÉSSICA FAGUNDES GÓES

Análise do tempo de permanência em cursos da UFBA: uma aplicação de modelagem com tempos discretos

Salvador

2022

JÉSSICA FAGUNDES GÓES

Análise do tempo de permanência em cursos da UFBA: uma aplicação de modelagem com tempos discretos

Trabalho de Conclusão de Curso apresentado ao Curso de Graduaçãoem Estatística, Colegiado de Estatística, Instituto de Matemática e Estatística, Universidade Federal da Bahia, como requisito parcial à aprovação na disciplina de Trabalho de Conclusão de Curso II.

Orientadora: Profa. Dra. Edleide de Brito

Salvador

2022

AGRADECIMENTOS

A Deus e aos espíritos de luz que me guiam nessa caminhada aqui na Terra. Aos meus pais, Rita e Fernando, por terem me dado a vida. Pelo incentivo, carinho, cuidado e por toda a base que me deram. Sem eles não seria possível chegar até aqui. À minha irmã, Patrícia, por estar sempre comigo, por todas as sábias palavras, por todo companheirismo, por me apoiar tanto e não largar a minha mão. Aos meus avós maternos *(in memorian)* Maria e Osório, pela história de luta e de vida que viveram juntos, e que me inspiram. Sei que eles me guiam aonde quer que estejam. Aos meus avós paternos Nize e Armando, por cuidarem tão bem de mim. A todos os familiares presentes na minha vida, pelas palavras de incentivo e por apostarem no meu potencial.

Ao meu noivo, João Victor, pelos 8 anos em que estamos juntos, inclusive minha etapa nessa graduação. Por todo o seu apoio nos incontáveis dias sentadas na minha mesinha de estudos e sua companhia sempre me motivando. Por ter me dado uma segunda família, "A Grande Família", a qual vibram pelas minhas conquistas. Aos meus amigos e colegas da faculdade, sem os quais essa caminhada certamente seria mais difícil: Marcos, Natália, Guilherme, Michelle, Leonardo, Renan, Rodrigo *(in memorian)*, Kézia, Caio e Rafael. Meu muito obrigada por partilharem comigo os seus conhecimentos.

Às minhas irmãs do coração: Lorena, Criscia e Carol. Por todo carinho e amor, por tudo que representam em minha vida, por serem mulheres fortes e guerreiras pelas quais eu tanto me espelho. Não podia deixar de agradecer à toda comunidade UFBA, técnicos administrativos, pessoal da limpeza e funcionários envolvidos na manutenção e preservação dessa maravilhosa Universidade. Em especial, a todos os professores do Instituto de Matemática e Estatística, principalmente aos professores e professoras do Departamento de Estatística, por lutarem incansavelmente pela educação de ensino superior público de qualidade, e por terem me ensinado tanto. Meu respeito e admiração eternos.

Por fim, não posso deixar de agradecer e enaltecer o papel da minha orientadora Edleide de Brito. Profissional ímpar, ser humano de um coração grandioso. Sem ela este trabalho não seria possível. Acompanhou-me durante 4 disciplinas na graduação, me ensinou tantas coisas as quais não consigo enumerar, mas a principal delas: ter alguém em que me espelhar não só como futura profissional em estatística, mas por todo seu empenho e comprometimento em dar o seu melhor sempre. Sua didática de ensino e presteza com a arte da docência, sua personalidade muito parecida com a minha, alinhada em muitos pensamentos, que me fez escolhê-la como a minha orientadora. Orientador é aquele que direciona, dirige, conduz e guia. É aquele que te mostra por onde seguir, te dando uma luz. Minha gratidão e encatamento por esse papel que desempenha com excelência.

"A estatística é a arte de torturar os números até que eles confessem. E eles sempre confessam."

Abraham Laredo Sicsú

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABI	Área Básica de Ingresso
Andifes	Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior
CAE	Conselho Acadêmico de Ensino
CARE	Coordenação de Atendimento e de Registros Estudantis
CENSUP	Censo da Educação Superior
CONSEPE	Conselho de Ensino, Pesquisa e Extensão
CONSUNI	Conselho Universitário
CFE	Conselho Federal de Educação
DCE	Diretório Central dos Estudantes
DEED	Diretoria de Estatísticas Educacionais do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
EAD	Ensino à Distância
ENEM	Exame Nacional do Ensino Médio
FIA	Funções de Incidência Acumuladas
IES	Instituições de Ensino Superior
Ifes	Instituições Federais de Ensino Superior
IGC	Índice Geral de Cursos
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LDB	Lei de Diretrizes e Base da Educação Nacional
MEC	Ministério da Educação
OCDE	Organização para Cooperação e Desenvolvimento Econômico
PDE	Plano de Desenvolvimento da Educação
PDI	Plano de Desenvolvimento Institucional

PIBIC	Programa Institucional de Bolsas de Iniciação Científica
PIBIEX	Programa Institucional de Bolsas de Iniciação à Extensão Universitária
PNE	Plano Nacional da Educação
PROEXT	Pró-Reitoria de Extensão
PROPCI	Pró-Reitoria de Pesquisa, Criação e Inovação
REGPG	Regulamento de Ensino de Graduação e Pós-Graduação *stricto sensu*
REUNI	Programa de Apoio a Planos de Restruturação e Expansão das Universidades Federais
SESU	Secretaria de Educação Superior
SIAC	Sistema Acadêmico
SiSU	Sistema de Seleção Unificada
UFBA	Universidade Federal da Bahia
UNE	União Nacional dos Estudantes
UNESCO	Organização das Nações Unidas para a Educação, a Ciência e a Cultura

SUMÁRIO

1 INTRODUÇÃO

O ensino superior é, em qualquer sociedade moderna, um dos motores do desenvolvimento econômico, sendo, igualmente, o instrumento principal de transmissão da experiência cultural e científica acumulada pela humanidade (FERREIRA, 2009). Como agente de transformação social, a diplomação em nível superior pode trazer inúmeras oportunidades de crescimento profissional, intelectual e social para os seres humanos. Segundo consta na *Declaração Mundial sobre Educação Superior no Século XXI: Visão e Ação, de 1998*, elaborada pela Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO), a educação superior é entendida como uma prova de viabilidade, no decorrer dos séculos, de transformar e induzir mudanças na sociedade (UNESCO, 1998). Embora sua relevância seja de vasto conhecimento, o acompanhamento dos desafios e circunstâncias encontradas nesse âmbito ainda necessitam de um maior enfoque.

De maneira geral, independente da cultura, a jornada enfrentada por estudantes na busca da obtenção de uma graduação permeia múltiplos aspectos que definem sua permanência no ensino superior. Esses aspectos podem ser vistos desde a identificação de uma vocação/escolha profissional ainda no ensino médio, até fatores de cunho econômico, sociais, psicológicos, acadêmicos e oportunidades do mercado de trabalho. Seja de qual natureza for o motivo associado a permanência dos estudantes em uma graduação, é cada vez mais importante a criação de pautas que discutam sobre esses desafios vivenciados na rede de ensino superior. Vale ressaltar que para as Instituições de Ensino Superior (IES) brasileiras públicas e privadas os desafios são comuns, embora possam também ter motivos associados distintos ou com magnitudes diferentes.

A rede pública de ensino superior participa com 24,2% do total de IES brasileiras (BRASIL, 2019a). Dentre as categorias públicas, a rede federal foi a que apresentou maior número de matrículas no período de 2009 a 2019, registrando um aumento de 59,2%. No entanto, entre os anos de 2018 e 2019, o sistema público de ensino registrou uma queda de 3,7% no número de ingressantes. Além disso, nesse mesmo período, o número de concluintes apontou uma diminuição de 3,1% (BRASIL, 2019a). Outras estatísticas trazem ainda números preocupantes, 18,4% foi a taxa de evasão registrada em cursos presenciais das instituições públicas no Brasil em 2019, conforme aponta o Mapa do Ensino Superior do Brasil, elaborado com base nos dados divulgados do Censo da Educação Superior (SEMESP, 2021). Essa taxa de evasão é calculada através da razão entre o somatório do número de matrículas trancadas, desvinculações do curso e falecimentos, pelo número total de alunos da instituição (SEMESP, 2016). Ainda em 2019, 21% das pessoas de 25 a 34 anos tinham diploma de ensino superior no Brasil em comparação com 45%, em média, nos países da Organização para a Cooperação e Desenvolvimento Econômico

(OCDE) (OECD, 2020).

Qualquer que seja a mudança vivenciada, é evidente a necessidade em acompanhar e verificar a trajetória dos estudantes que ingressam em uma graduação, a fim de proporcionar o sucesso dos mesmos quanto ao objetivo de formação, bem como monitorar a manutenção dos investimentos que são necessários na preservação do estudante nesse sistema. O cenário da educação superior brasileira passou por diversas transformações ao longo do tempo, e, em 2020, vivenciou impactos e desafios com a propagação da pandemia do novo coronavírus (COVID-19) que assolou o mundo, e, portanto, esse cenário tende a passar por desafios maiores nos próximos anos.

O acompanhamento dos estudantes em uma graduação é de extrema importância, e, baseado nesse contexto, o presente trabalho busca verificar o tempo de permanência dos estudantes dos cursos de graduação de Estatística e Matemática da Universidade Federal da Bahia (UFBA), bem como os fatores que estão associados e que impedem o estudante de se formar, traçando possíveis perfis entre aqueles que foram concluintes ou não no ano de 2019. Há também o propósito de escolher um conceito de estudante que está retido no curso e classificar discentes evadidos, uma vez não foram encontrados referências da UFBA definindo estes conceitos.

Os resultados obtidos são embasados na análise de microdados do Censo da Educação Superior (CENSUP) de 2019. O CENSUP é realizado anualmente pelo Instituto Nacional de Pesquisas Educacionais Anísio Teixeira (INEP), órgão vinculado diretamente ao Ministério da Educação (MEC). Pretende-se então contribuir e agregar às informações já existentes sobre esse aspecto do tempo de permanência dos estudantes, no âmbito da UFBA, incorporando, através da modelagem de eventos competitivos em Análise de Sobrevivência, elementos que auxiliem na tomada de decisão, melhorias e debates acerca do tema.

1.1 O problema de pesquisa

Uma vez introduzidos os aspectos da seção anterior, é necessário também enfatizar as questões relacionadas ao problema desta pesquisa.

Segundo consta no mais atual Plano de Desenvolvimento Institucional (PDI) da UFBA (2018-2022), a universidade possui, dentre outros, o seguinte objetivo:

> "Objetivo estratégico 1 e diretriz estratégica 3: Assegurar maior eficiência no uso de recursos institucionais para o processo de formação dos alunos dos cursos de Graduação e Pós Graduação, com a ampliação de oportunidades e vagas, buscando aprofundar a integração entre diversos níveis de ensino, tendo como ação acompanhar e avaliar os currículos, bem como a taxa de retenção e a de evasão dos cursos, para estabelecer estratégias de permanência e sucesso na Instituição. (...)". UFBA (2017a).

Contudo, ao procurar documentos oficiais que tragam essas taxas, bem como características do corpo estudantil inserido nesses perfis de evasão e retenção, temos acesso apenas ao anuário produzido pela UFBA denominado *UFBA em Números*, os quais não abordam essas estatísticas. Esse mesmo documento apresenta, dentre outras informações, alguns registros relacionados ao número de matrículas e concluintes nos cursos, conforme sintetizados na Tabela 1. Todas as tabelas e gráficos aqui presentes são de autoria própria.

Tabela 1 – Número de estudantes matriculados e concluintes por ano - UFBA - 2015 a 2019.

Ano Base	Matrículas	Concluintes
2015	33.804	2.999
2016	35.211	3.635
2017	37.985	3.625
2018	39.795	3.407
2019	40.727	3.288

Fonte: UFBA (2016a), UFBA (2017b), UFBA (2018), UFBA (2019a), UFBA (2020a).

O conceito de permanência está diretamente relacionado com os conceitos de evasão, retenção e diplomação. Na UFBA, no entanto, existe uma lacuna referente ao verdadeiro significado de evasão. No *Regulamento de Ensino de Graduação e Pós-Graduação stricto sensu (REGPG), da Universidade Federal da Bahia*, artigo 75 da Subseção III, apresenta-se a seguinte definição:

> "O estudante da graduação poderá ter a sua matrícula cancelada caso: não conclua o curso no prazo máximo fixado para a integralização do respectivo currículo; não conclua a nova modalidade/habilitação/opção no prazo definido pelo Colegiado do curso, quando se tratar de reingresso."(UFBA, 2015).

E ainda, no inciso 1, define-se que:

> "O estudante será notificado pela Coordenação de Atendimento e de Registros Estudantis (CARE) se, ao atingir cinquenta por cento (50%) do tempo máximo previsto para integralização, não tiver cumprido, pelo menos, cinquenta por cento (50%) da carga horária total do curso." (UFBA, 2015).

No tocante ao termo jubilamento, segundo consta em DCE-UFBA (2014) sobre o novo regulamento instituído, o mesmo sucedeu-se em: extinção de todas as modalidades de jubilamento na graduação, como, por exemplo, reprovação em todas as disciplinas em dois semestres seguidos ou reprovações seguidas em uma mesma disciplina, exceto no caso de estudantes que excedem o tempo máximo de curso, ou, no caso de estudantes reingressos, que excedam o tempo determinado pelo colegiado.

Todavia, conforme é trazido por Cruz (2019), ainda existe um número considerável de estudantes que mesmo sem matricular-se em pelo menos alguma disciplina por mais de dois semestres consecutivos, são considerados pelo sistema como estudantes com *status* de ativos. A exemplo, menciona-se os dados encontrados por Lima, Coutinho e Santos (2015), em um estudo com uma amostra de estudantes ingressantes no curso de Psicologia da UFBA, após terem passado pelo Bacharelado Interdisciplinar, nenhum dos estudantes que foram caracterizados como evadidos havia solicitado desistência do curso de maneira oficial.

Sob esse olhar, pretende-se, através da utilização dos dados públicos disponíveis do CENSUP, analisar o tempo de permanência de estudantes dos cursos de Estatística e Matemática da UFBA, escolhendo também um conceito para estudantes retidos e evadidos, de modo a traçar perfis entre os discentes e contribuir para o melhor mapeamento dessa questão na comunidade.

1.2 Justificativa

Alguns trabalhos já existentes e que fazem menção ao contexto de evasão e permanência de estudantes, no panorama da UFBA, são ainda escassos. A saber: Santos (2013), Andrade (2014), Lima, Coutinho e Santos (2015), Santos (2017), Cruz (2019), Campos (2020) e Pinheiro (2021).

A dissertação de Santos (2013) traz a ótica de estudantes cotistas e não cotistas da universidade, sua vivência acadêmica e a intenção de evadir. A autora faz uso do procedimento estatístico de Análise de Variância (ANOVA) para comparar os grupos de cotistas e não cotistas; a dissertação de Andrade (2014) aborda o contexto de evasão referente aos Bacharelados Interdisciplinares (BI) da UFBA; o artigo de Lima, Coutinho e Santos (2015) discorre sobre um estudo com discentes do curso de Psicologia da UFBA e que evadiram; o trabalho de conclusão de curso de Santos (2017) trata a questão da evasão no curso de Educação Física. Cruz (2019) aborda sobre a evasão de discentes de cursos na UFBA utilizando uma amostra de 369 estudantes considerados como evadidos (que não estavam inscritos em nenhuma disciplina por mais de dois semestres consecutivos) no período de 2014 e 2019 e utiliza a metodologia de classes latentes em sua pesquisa; Campos (2020) disserta sobre as causas de evasão e retenção no curso de Biblioteconomia e Documentação da instituição; e, por fim, Pinheiro (2021) com a tese de doutorado que aplica a metodologia de análise de sobrevivência, com modelos de longa duração, no estudo da evasão nos cursos de engenharia da UFBA.

A maioria dessas produções tem caráter mais descritivo e exploratório. Além disso, os estudos aqui citados investigam um número menor de estudantes ou sob uma ótica restrita de um ou mais cursos, através de questionários aplicados ou informações obtidas através do Sistema Acadêmico (SIAC) da UFBA. Todos extremamente relevantes para a

pauta de discussão sobre um dos elementos-chaves de uma universidade: o estudante, e com ele, a sua permanência na IES. O presente trabalho se mostra pioneiro no âmbito da UFBA, comparado aos citados anteriormente, pois fará uso de dados públicos fornecidos pelo mais completo instrumento de coleta de informações da educação superior brasileira, o CENSUP.

Uma razão a ser levada em consideração é o uso da Análise de Sobrevivência como recurso de investigação do tempo de permanência dos estudantes nos cursos de graduação. No Brasil ainda são limitados os estudos sobre os fenômenos da retenção, evasão e tempo até formatura dos estudantes que se apropriam dessa metodologia para avaliar a associação existente entre esses fenômenos e eventuais fatores catalisadores do mesmo, especialmente na identificação de perfis e padrões existentes nessas circunstâncias. Grande parte dos estudos descrevem as problemáticas sob uma perspectiva de técnicas exploratórias e descritivas. No entanto, nas pesquisas relacionadas às questões sobre a evasão em IES, muitas vezes se faz de interesse a verificação do tempo em que o estudante permaneceu no curso até se formar, se o mesmo ficou retido por muito tempo ou se evadiu, por exemplo. Neste trabalho é apresentada uma proposta de como classificar os conceitos de evasão e retenção na UFBA.

Considerando que no final do estudo o aluno pode não experimentar o evento de interesse, não diplomando-se e ficando retido por mais tempo do que o previsto ou ainda ser considerado evadido, temos o caso de presença de dados censurados ou ditos incompletos. O final do estudo aqui figura-se como o ano referência do censo (2019). Então, cada estudante é acompanhado da data de ingresso na UFBA até o ano de 2019. Esse contexto será melhor abordado na Seção 3.4. Na literatura, o procedimento indicado para lidar com dados dessa natureza é a metodologia estatística de Análise de Sobrevivência, que consegue abarcar observações com essas características, incorporando-as em todas as etapas das análises.

Além disso, é necessário frizar que o tempo de permanência do estudante na IES, especialmente nas Instituições Federais de Ensino (Ifes), depende de investimentos e gastos públicos para a manutenção desses estudantes na rede de ensino superior. Sobre esse contexto, Amaral e Pinto (2010) mencionam os debates relacionados a quanto se gasta para formar um estudante de graduação, o que é comumente chamado de custo-aluno.

Os dados do CENSUP serão usados no estudo do tempo de permanência de estudantes dos cursos de Estatística e Matemática da UFBA Com a aplicação da metodologia de Análise de Sobrevivência, e visando disponibilização em sítio *web* dos resultados obtidos através da aplicação *web* `R Shiny`, este trabalho tende à agregar mais conhecimento sobre esse tema tão importante. Sendo assim, no combate à evasão e retenção de estudantes na UFBA, a identificação dos fatores relacionadas a essas situações contribuem para traçarmos padrões, tendências e perfis de discentes.

O estudo aqui realizado se mostra pertinente, uma vez que pode auxiliar na tomada de decisão que diz respeito a essas questões na universidade, além de melhorar o monitoramento do percurso acadêmico dos estudantes e contribuir para as reflexões sobre a redução do tempo de permanência dos discentes no ensino superior. A escolha dos cursos de Estatística e Matemática foram motivados por estarem na realidade mais próxima vivenciada no decorrer do trabalho (pertencem ao mesmo instituto, o Instituto de Matemática e Estatística da UFBA), podendo ser facilmente estendida para análises futuras com outros cursos da instituição.

1.3 Objetivos

Discorridas as justificativas para realização deste estudo, e todo contexto em que a UFBA se insere nessas circunstâncias, este trabalho tem como objetivo geral analisar o tempo de permanência dos estudantes nos cursos de Estatística e Matemática da UFBA através da utilização dos microdados do CENSUP do ano de 2019, aplicando modelagem para tempos discretos em eventos competitivos.

Como objetivos específicos, pretende-se:

i) Escolher um conceito de aluno retido e evadido, a fim de verificar estudantes que ainda permanecem no sistema de ensino mas que possam ser considerados retidos e também aqueles que podem ser classificados como evadidos.

ii) Verificar os fatores associados e que impedem os estudantes desses cursos de se formarem, traçando possíveis perfis de estudantes concluintes e não concluintes, aplicando a metodologia de eventos competitivos com tempos discretos.

iii) Construir aplicativo *web* através do `R Shiny`, com todos os resultados obtidos nas análises realizadas, e disponibilizar publicamente em endereço na *internet*, visando contribuir com os estudos acerca do tema junto à comunidade UFBA.

Para direcionar o alcance do objetivo geral já definido, a estrutura do trabalho está organizada como descrito a seguir. Na Seção 2 é apresentada a revisão de literatura sobre o ensino superior no país, definição de alguns conceitos relacionados com o tema, explanação de como o CENSUP é aplicado e sua importância para o Brasil, além de uma breve abordagem sobre a UFBA. Na Seção 3, são apresentadas as metodologias necessárias para a manipulação dos microdados do CENSUP bem como a modelagem de tempos discretos em eventos competitivos. Na Seção 4, são dispostos os resultados descritivos e a modelagem com riscos competitivos realizada. A Seção 5 traz as considerações finais e sugestões para trabalhos futuros. O Apêndice A pode ser consultado para verificar as variáveis que foram utilizadas neste trabalho. No Apêndice B disponibiliza-se os códigos dos modelos e rotinas utilizadas no *software* `R`. No Apêndice C é disponibilizado os resultados preliminares obtidos para todos os cursos de graduação da UFBA.

2 EDUCAÇÃO SUPERIOR NO BRASIL: BREVE HISTÓRICO E ATUAL CONJUNTURA

Para verificar as razões relacionadas com o tempo de permanência do estudante no ensino superior superior, faz-se necessário entender alguns conceitos e referências sobre a história da educação superior, em especial no quadro em que estamos inseridos: o Brasil.

> "A educação, direito de todos e dever do Estado e da família, será promovida e incentivada com a colaboração da sociedade, visando ao pleno desenvolvimento da pessoa, seu preparo para o exercício da cidadania e sua qualificação para o trabalho". (BRASIL, 1988).

Conforme rege a Constituição Brasileira de 1988, artigo 205, a educação tem papel fundamental na vida em sociedade, potencializando o desenvolvimento cognitivo, físico e emocional dos indivíduos, capacitando-o para tornar-se um cidadão integrado ao convívio social.

No Brasil, a educação vivenciou inúmeras transformações ao longo do tempo até chegarmos aos dias atuais. De 1808, quando foram criadas as primeiras escolas de nível superior, até 1934, a referência da educação superior foi voltada para a formação de profissões liberais tradicionais, com a presença de cursos de direito, medicina e engenharias (SAMPAIO, 1991).

O ano de 1808 marca o início da fundação de instituições superiores em território nacional com a chegada da família real portuguesa ao país. Nesse ano, foram criadas a atual Escola de Medicina da Universidade Federal da Bahia e da Universidade Federal do Rio de Janeiro, bem como a Academia da Guarda Marinha, também com sede no Rio de Janeiro. Com a Proclamação da República em 1889 e a transição da monarquia para o novo sistema assumido pelo governo, a segunda Constituição do Brasil e primeira Constituição do sistema republicano de 1891 demarca a criação dos estabelecimentos do ensino superior pela iniciativa privada (MARTINS, 2002).

No ano de 1938 houve a criação da União Nacional dos Estudantes (UNE), movimento estudantil que preconizava uma reforma universitária por meio de debates, com a finalidade de combater a natureza ultrapassada e elitista das instituições universitárias do país na época (ZOCCOLI, 2009). Já a reforma universitária ocorrida em 1968 simboliza marcantes transformações no ensino superior brasileiro: 1- instituiu o departamento como unidade mínima de ensino, 2 – criou os institutos básicos, 3 – organizou o currículo em ciclos básico e profissionalizante, 4 – alterou o exame vestibular, 5 – aboliu a cátedra (cadeira a qual o professor ocupava em sala de aula, em um plano superior ao dos ouvintes e detiam amplos poderes (FÁVERO, 2000)), 6 – tornou as decisões mais democráticas, 7

– institucionalizou a pesquisa e 8 – centralizou decisões em órgão federais (MARTINS, 2002).

No entanto, cabe lembrar que o ano de 1968 esteve inserido no período da Ditadura Militar Brasileira (1964-1985), e mesmo a educação sendo pauta relevante na época, impulsionada principalmente pela Reforma Universitária ocorrida nesse ano, o Governo enfrentava dificuldades em relação a insuficiência de capital monetário (COSTA; BARBOSA; GOTO, 2011). A opção encontrada nessa época foi a criação de medidas que favoreceram a expansão do ensino superior privado, de fins lucrativos, que proporcionou um regaste financeiro ao país através da presença de empreendedores nesse setor. Com isso, o número de instituições privadas cresceu rapidamente.

Complementado esses anos-chave da história da educação superior brasileira, outros pontos vivenciados na sua expansão podem ser citados, com igual relevância:

- A Lei Nº 9.394 de 1996 denominada Lei de Diretrizes e Base da Educação (LDB) (BRASIL, 1996), definiu mudanças significativas para a educação, em particular para o ensino superior público e privado. No ensino superior público, a LDB permitiu autonomias, tais como: a possibilidade em desenvolver e aplicar seu próprio orçamento; reavaliarem operações de crédito; receberem doações, heranças, legados e obter auxílio financeiro de associações público-privadas, tornando legítima a busca pelas mais diferentes origens de financiamento (COSTA; BARBOSA; GOTO, 2011).

- Estabelecido através da Lei Nº 10.172, o Plano Nacional da Educação (PNE) de 2001 foi sancionado em 9 de janeiro de 2001. O PNE instituiu premissas básicas como a educação como um direito de todos, educação como instrumento de combate à pobreza e inclusão social e a educação como fator de desenvolvimento econômico e social do país (BRASIL, 2001).

- No ano de 2007 houve a implantação do Plano de Desenvolvimento da Educação (PDE), através do Decreto Nº 6.094 (BRASIL, 2007). Para a competência de nível superior, o PDE trouxe como princípios a expansão da oferta de vagas com garantia de qualidade, promoção da inclusão social pela educação, ordenação territorial, desenvolvimento econômico e social. Fazendo da educação superior uma formadora de recursos humanos altamente preparados, além de peça primodial na produção científico-tecnológica, elemento-chave da integração e da formação da nação (BRASIL, 2008). A implantação do PDE em 2007 proporcionou a criação e ampliação de programas públicos para a educação superior no Brasil. Destaca-se aqui o REUNI - Restruturação e Expansão das Universidades Federais no ensino público, criado pelo MEC. O programa REUNI teve duração de cinco anos, compreendendo o período de 2007 a 2011. Os dados do Relatório de Primeiro Ano do REUNI, em 2008, mostram

que houve um aumento de 14.826 novas oportunidades de ingresso em universidade públicas federais de 2007 para 2008 (MEC, 2009).

- No período de 2012-2021, foram criadas as cotas para a rede pública nas universidades federais através da publicação, em agosto de 2012, da Lei Nº 12.711. A Lei instituiu que no mínimo 50% das vagas das Ifes fossem reservadas para estudantes que tivessem cursado integralmente ensino médio em escolas públicas, das quais seriam subdivididas e 50% deveriam ser reservadas aos estudantes oriundos de famílias com renda igual ou inferior a 1,5 salário-mínimo (um salário-mínimo e meio) *per capita* (BRASIL, 2012).

- Ainda no período de 2012-2021, também ocorreram cortes no orçamento devido ao ajuste fiscal de 2015 e que abalaram os recursos destinados à educação (em todos os níveis), tendo efeito direto no funcionamento das Ifes, com o atraso de pagamentos de trabalhadores terceirizados, bolsas de pesquisa e assistência estudantil (MOREIRA; MOREIRA; SOARES, 2018). De 2014 a 2018 o investimento em educação superior no Brasil caiu em 15% (MAZIEIRO, 2019).

A atual conjuntura da educação superior pública brasileira continua sendo de redução orçamentária. A Proposta de Legislação Orçamentária Anual (PLOA) de 2021 foi aprovada com redução de 18,16% no orçamento das 69 universidades em todo o país. Esta redução afetou diretamente o cumprimento do pagamento de despesas básicas como contas de água, luz, contratos com empresas de segurança, bolsas de pesquisa, alimentação e apoio a alunos carentes (OLIVEIRA, 2021).

No que tange a permanência do estudante na universidade, a atual situação da educação superior pública brasileira mostra uma realidade alarmante e que pode também interferir no sucesso na obtenção do grau acadêmico superior. As atuais restrições orçamentárias tem impacto direto nos repasses destinado à bolsas de pesquisa, de apoio permanência e de extensão, e contas básicas para a rotina em uma Ifes, como água e luz.

O ensino superior público de qualidade gera retorno à sociedade através da produção científica de conhecimento, formação de recursos humanos, desenvolvimento tecnológico e prestação de inúmeros serviços à população. Mas, a atual desvalorização desse ensino, em função dos cortes orçamentários, coloca em risco todo esse retorno benéfico que é produzido pelas IES públicas.

Para essa discussão, no entanto, cabe um estudo mais focado. Por ora, o tempo de permanência dos estudantes, no âmbito UFBA, será discutido sob o ponto de vista dos dados disponibilizados pelo CENSUP.

2.1 Conceitos básicos: evasão, retenção e diplomação no ensino superior brasileiro

Para vislumbrar o contexto a qual está inserido este trabalho, é necessário conhecer alguns conceitos que estão diretamente relacionados com a questão do tempo de permanência dos discentes na graduação. O tempo de permanência, ou até mesmo retenção do discente na IES está intimamente relacionado com as concepções de diplomação e evasão. Isso porque, o estudante que permanece na rede de ensino superior pode vir a evadir/desistir do curso ou então diplomar-se.

O olhar acerca das situações de evasão, retenção e diplomação no ensino superior público brasileiro ganhou pauta relevante com a publicação, em outubro de 1996, do relatório apresentado pela comissão especial de estudos sobre a evasão nas universidades públicas brasileiras à Andifes/Associação Brasileira de Reitores das Universidades Estaduais e Municipais (ABRUEM)/Secretaria de Educação Superior (SESU)/MEC. O documento pioneiro, intitulado por "Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas" reuniu dados referentes ao desempenho das universidades públicas brasileiras levando em consideração os índices de diplomação, retenção e evasão dos estudantes de seus cursos de graduação (MEC, 1996).

Ainda nesse documento, o mesmo carrega as seguintes definições sobre os termos evasão, retenção e diplomação:

> "Diplomado: Aluno que concluiu o curso de graduação dentro do prazo máximo de integralização curricular, fixado pelo Conselho Federal de Educação (CFE), contado a partir do ano/período-base de ingresso.
>
> Retido: Aluno que, apesar de esgotado o prazo máximo de integralização curricular fixado pelo CFE, ainda não concluiu o curso, mantendo-se, entretanto, matriculado na universidade.
>
> Evadido: Aluno que deixou o curso sem concluí-lo." (MEC, 1996).

Na narrativa, o MEC (1996) traz, dentre outras informações, que alguns cursos listados apresentaram, na época, uma taxa baixa de evasão. Nesse caso, o que esperava-se era que a taxa de diplomação fosse elevada. No entanto, esses mesmos cursos apresentaram tanto uma taxa de diplomação baixa como uma taxa de retenção alta. Para essa situação, o texto menciona que deveriam existir possíveis contrariedades do Controle Acadêmico dos alunos ou não cumprimento das normas legais de jubilamento.

O termo jubilamento é entendido como o desligamento ou afastamento do aluno da IES por ter ultrapassado o prazo máximo estabelecido para a conclusão do curso (RODRIGUES, 2007). No entanto, temos que a LDB em 1996 revogou a Lei Nº 5.540/1968 que deliberava anteriormente sobre essa questão, como tal:

> "A Lei n.º 9.394/1996 – LDB, em seu artigo 92, revogou expressamente a Lei n.º 5.540/1968. Nesse sentido, no plano das normas gerais do Direito

> Educacional brasileiro, não há mais qualquer base legal para desligar estudantes, no âmbito da educação superior, tendo por base o argumento de que ultrapassaram o prazo máximo para a conclusão dos cursos aos quais estariam vinculados" (MEC, 2018).

Isto é, a Lei n.º 9.394 de 1996 revogou essa obrigatoriedade do desligamento segundo esse critério de jubilamento. Ressalta-se, no entanto, que o jubilamento ainda é um tema controverso, uma vez que ainda é utilizado em algumas universidades, como por exemplo, na Universidade Federal do Maranhão (PINTO, 2020).

Na literatura, são numerosos os trabalhos que falam sobre os fenômenos da retenção e principalmente evasão no ensino superior. Como exemplos, temos os de Filho et al. (2007), Freitas (2009), Lobo (2012) e Cunha, Nascimento e Durso (2016). A busca das causas relacionadas a esses acontecimentos tem sido alvo de muitos estudos, uma vez que afetam diretamente o sistema educacional como um todo. Contudo, grande parte desses estudos enfatizam a necessidade de aprofundamento das análises sobre o tema.

Em um de seus trechos, Filho et al. (2007) trazem que, no primeiro ano de curso, a taxa de evasão chega a ser 2 a 3 vezes maior do que nos anos seguintes, o que se torna uma questão muito examinada, visto ter influência direta nos índices de evasão e diplomação. Outro ponto trazido nesse artigo faz menção à evasão decorrente de dificuldades financeiras enfrentadas pelos estudantes, o que impede que os mesmos levem adiante os seus estudos. Esse motivo é mencionado tanto por parte das instituições de ensino como também por parte dos alunos. No entanto, Filho et al. (2007) frisam que a real causa não se resume apenas a isso, pois dizem respeito também às:

> "questões de ordem acadêmica, as expectativas do aluno em relação à sua formação e a própria integração do estudante com a instituição, constituem, na maioria das vezes, os principais fatores que acabam por desestimular o estudante a priorizar o investimento de tempo ou financeiro, para conclusão do curso. Ou seja, ele acha que o custo benefício do "sacrifício" para obter um diploma superior na carreira escolhida não vale mais a pena" (FILHO et al., 2007).

O artigo também conclui que, na esfera pública, os recursos direcionados à manutenção do aluno na rede de ensino superior pública são investimentos sem nenhum retorno caso o aluno evada do curso. Cunha, Nascimento e Durso (2016) corroboram com essa afirmação, trazendo o fato de que os investimentos, tanto em IES particulares quanto em públicas, que são destinados à formação do indivíduo acabam sendo perdidos sem possibilidade de ressarcimento uma vez que a vaga ocupada por um estudante que desiste do curso não poderá ser aproveitada por outro discente.

Lobo (2012) destaca que, com o grande número de trabalhos existentes sobre o fenômeno da evasão, torna-se custoso padronizar tudo referente a esse tema. Desse modo,

Lobo (2012) define que a evasão é um fenômeno multifatorado e pode manifestar-se em três tipos de configurações: evasão do curso, evasão do sistema e evasão da IES.

As múltiplas razões que podem levar o estudante a não permanecer no seu curso recaem em um ponto único: prevenir que estes evadam do sistema, garantindo a sua permanência, seja por motivo financeiro seja outro motivo que necessite de um acompanhamento próximo por parte dos Departamentos, Colegiados e órgãos ligados a essa questão.

Dentro dessa perspectiva, Freitas (2009) abrange como foco não somente a questão da evasão, mas sim as ações que podem ser praticadas com o intuito de encorajar os estudantes a persistir na vida escolar e permanecer no sistema de ensino com sucesso. Esse contexto exprime relevância, principalmente no âmbito do ensino superior, uma vez que os estudantes estão sendo preparados para o mercado de trabalho.

No que concerne a compreensão sobre o quesito da retenção, abordada como a condição em que o aluno necessita um tempo maior para a conclusão do que o mínimo previsto na matriz curricular do curso (PEREIRA et al., 2015), Carvalho (2019) discorre que a evasão, além de ser uma possibilidade de decisão repentina, pode ser ocasionada devido a um grande número de retenções em um curso, daí a importância em estudar também a retenção. No seu entendimento, a atenção com a retenção diz respeito a esta ser um fenômeno que sempre ocorre anteriormente à evasão. Ainda nesse sentido, Carvalho (2019) salienta ser imprescindível conhecer as características e perfis que predominam na retenção de um determinado curso de uma IES, buscando formentar nos gestores dos cursos a introdução de ações que mirem à queda do número de retenções.

Campello e Lins (2008) aborda esse mesmo pensamento, discorrendo sobre os impactos negativos trazidos pela retenção, haja vista que acaba não permitindo que profissionais de nível superior passem a atuar nas suas áreas do conhecimento no tempo inicialmente previsto, além da possibilidade desses alunos em algum momento evadirem do sistema.

Sob o ponto de vista de Garcia, Lara e Antunes (2020), o acadêmico obtém sucesso quando a universidade cumpre o seu papel de fornecer profissionais qualificados para a sociedade e para o mercado de trabalho. Casos opostos a esses configuram-se como um insucesso, uma vez que trazem como exemplo os fenômenos da evasão e retenção no ensino superior, que ocasionam a redução drástica do número de concluintes em cursos superiores.

Diante de todas as abordagens que são trazidas sobre o tema de evasão e permanência de estudantes no ensino superior, ainda é necessário enfatizar a importância da gestão dessas informações no ambiente acadêmico. Afinal, a boa conduta em gerir os dados a respeito dessas questões traz a possibilidade de crescimento e melhor monitoramento

das instituições (HOFFMANN; NUNES; MULLER, 2019). Os autores Hoffmann, Nunes e Muller (2019) discorrem sobre a seriedade e dificuldade da gestão do conhecimento das organizações a respeito das causas e medidas de combate à evasão em IES. Ainda, chamam à atenção sobre a quantidade de informações que são coletadas, examinadas e assimiladas, mas que geralmente ficam restritas às coordenações de cursos ou publicações sem uma determinada abrangência.

Hoffmann, Nunes e Muller (2019) enfatizam a necessidade da utilização de dados padronizados. Desse modo, segundo os autores, é possível realizar comparações com estudos de outros locais, uma vez que diferentes instituições realizam estudos sob diferentes fontes de dados, o que resulta em informações que podem não ser necessariamente equivalentes. Sendo assim, os autores concluem a necessidade de gestão do conhecimento das instituições acadêmicas sustentadas na análise de dados do CENSUP, de abrangência nacional e que é padronizado para todas as IES do país.

Nessa linha de raciocínio, as produções existentes que são fundamentadas na investigação dos microdados do CENSUP, e associadas com os fenômenos da evasão e retenção de alunos nas IES, ainda são tímidas. A exemplo, algumas produções são mencionadas.

Apoiada nos dados do CENSUP, Saccaro (2016) investiga os fatores associados a evasão em cursos de ciências naturais e engenharias (de instituições privadas e públicas brasileiras) através da metodologia de análise de sobrevivência. Nos seus achados, a autora enfatiza sobre a investigação do fenômeno nessas áreas, visto que, segundo Salerno et al. (2013), esses setores estão diretamente ligados com a evolução tecnológica e aumento de produtividade para a economia brasileira. Santos e Giraffa (2013) também utilizam dados públicos ao analisar a ocorrência da evasão em cursos de modalidade EAD de IES públicas e privadas no Brasil.

Marques (2020) recorre aos microdados do censo com o objetivo de mapear a volta ao ensino superior dos estudantes que foram considerados evadidos, em instituições públicas e privadas, analisando o intervalo de tempo de 2009 a 2017. Cândido (2019), por sua vez, realiza a análise da base de dados do censo no período de 2009 a 2018, objetivando a construção e cálculo de indicadores que quantifiquem as taxas de evasão universitária, tanto para IES de natureza pública quanto privada.

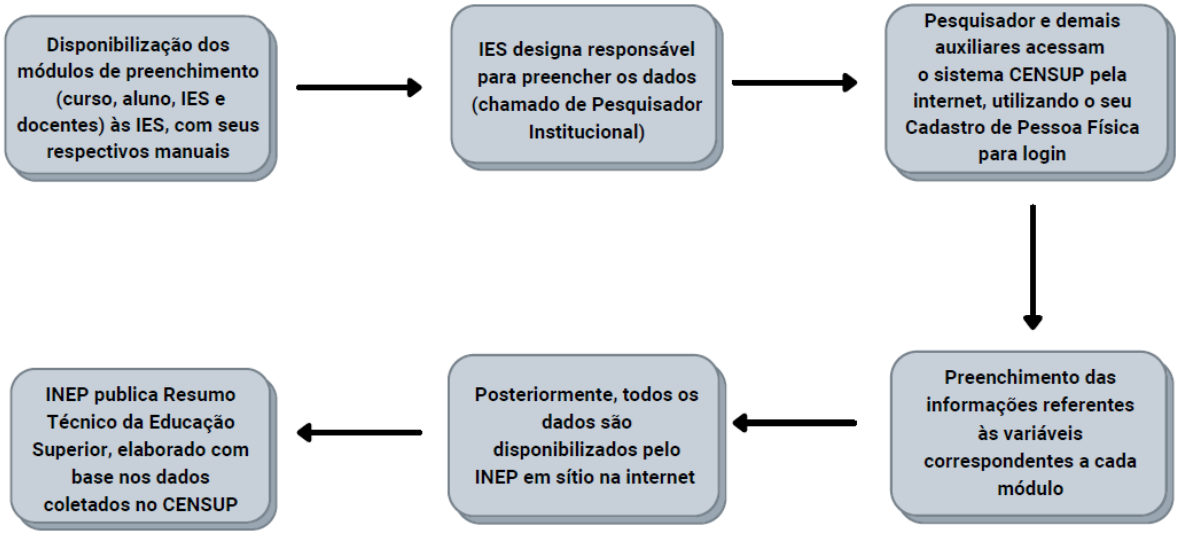
Por fim, para o ângulo de instituições públicas e principalmente federais, é apresentado, como exemplo, os estudos de Carvalho (2017), Ferreira (2018) e Souza e Freitas (2021). Em Carvalho (2017) e Souza e Freitas (2021) o objeto de estudo é relacionado a Ifes, buscando observar os perfis de alunos que evadem. Ferreira (2018) traz no seu estudo a proposta de caracterização da assistência estudantil, no contexto de prevenir a evasão e contribuir para a permanência do estudante na instituição. A abordagem de Ferreira (2018) é feita no contexto da Universidade do Estado da Bahia (UNEB) utilizando microdados do CENSUP dos anos de 2014 e 2015.

Desse modo, faz-se necessário conhecer as razões que levam os estudantes a permanecerem no ensino superior por um prazo maior que o previsto, e também verificar as razões associadas à evasão. E mais, é visível a vantagem em utilizar dados públicos de extensão nacional e padronizados, a fim de verificarmos panoramas relacionadas a educação superior brasileira, de maneira a trazer as devidas contribuições e diversas abordagens acerca dessa temática. Conforme visto, embora numerosas as produções a respeito desses temas (e ainda tímidas a utilização embasada em dados do CENSUP), as mesmas carecem de estudos mais aprofundados na tentativa de traçar perfis de estudantes nessas situações, impulsionar ações que vislumbrem a não evasão e fometem a diplomação dos discentes, bem como na conscientização referente ao custo-aluno envolvido nesse processo.

2.2 O Censo da Educação Superior (CENSUP)

O INEP, órgão vinculado ao MEC, é a entidade responsável pela coleta e divulgação dos dados do CENSUP. A coleta de dados é realizada anualmente e tem como referência as diretrizes deliberadas no Decreto Nº 6.425 de 4 de abril de 2008. O CENSUP tem como objetivo reunir informações sobre as IES brasileiras de natureza pública e privada, cursos ofertados de maneira presencial e à distância e informações acerca dos alunos e docentes (BRASIL, 2008). Amparado pela LDB (BRASIL, 1996), que tornou a União encarregada de coletar, analisar e disseminar informações sobre a educação, a obrigatoriedade de prestar informações ao INEP através do CENSUP teve seu início em 2008, conforme consta na série histórica do INEP (2019a). Os primeiros dados disponibilizados referente a realização do censo datam de 2009. Para a coleta dos dados, o INEP realiza o procedimento listado na Figura 1.

Figura 1 – Procedimento de coleta dos dados do CENSUP.



Fonte: INEP (2019b).

A Diretoria de Estatísticas Educacionais (DEED), unidade interna do INEP, é responsável pela realização do CENSUP e pela publicação do Resumo Técnico da Educação Superior, que é o produto final do estudo. Esse documento sintetiza os principais achados da pesquisa, em que:

> "os resultados apresentados são fundamentais para traçar o atual panorama da educação no País, sendo uma fonte de consulta para dirigentes de instituições de ensino, gestores de políticas educacionais, órgãos governamentais, pesquisadores e demais interessados na educação brasileira, de modo a subsidiar análises, pesquisas, planejamentos e processos de tomada de decisão" (BRASIL, 2019b).

A importância das informações obtidas através do CENSUP também pode ser verificada com a criação de indicadores de fluxo da educação superior. Segundo MEC (2020), esses indicadores servem de base para diferentes análises, bem como para medida da eficiência de cada tipo de formação, podendo ser combinados com outros indicadores ou insumos, auxiliando na criação de novos parâmetros de controle de eficiência do curso, além de qualificar a oferta e a demanda dessas graduações. Também são capazes de subsidiar discussões acerca da eficácia do sistema de ensino superior, principalmente quanto à capacidade deste para formar pessoas.

Portanto, podemos perceber a notoriedade em realizar-se as análises dos dados fornecidos pelo CENSUP, haja vista que o mesmo pode trazer resultados relevantes sobre o atual cenário educacional brasileiro, não somente em nível nacional como também em níveis regionais, estaduais e municipais.

2.3 A Universidade Federal da Bahia

A UFBA é uma das mais antigas instituições de ensino superior do estado, somando a sua história mais de 200 anos. Sua origem remonta a chegada da família real portuguesa ao Brasil, no ano de 1808. A Escola de Cirurgia da Bahia, fundada em 18 de fevereiro de 1808 com sede na cidade de Salvador marca o nascimento da universidade. Posteriormente, a Escola transformou-se na atual Faculdade de Medicina da Bahia e a sua fundação simboliza o ano de criação da UFBA, conforme aprovado por unanimidade pelo Conselho Universitário (CONSUNI) em 2009 (TOUTAIN; SILVA, 2010).

Formalmente foi estabelecida como universidade no ano de 1946 atráves do Decreto-Lei Nº 9.155 de 8 de abril de 1946 (BRASIL, 1946), tendo instalado-se oficialmente em 2 de julho do mesmo ano, atual data de aniversário da instituição. Inicialmente a UFBA teve o seu funcionamento com as seguintes unidades: Faculdade de Medicina da Bahia (e suas escolas anexas Odontologia e Farmácia), Academia de Belas Artes (1877), Faculdade de Direito da Bahia (1891), Escola Politécnica da Bahia (1896), Faculdade de Filosofia da Bahia (1941) e Faculdade de Ciências Econômicas (1905), tendo como seu primeiro

reitor, o médico Edgard Santos (UFBA, 2006). A partir de então, a universidade cresceu e se modernizou com a criação de inúmeros outros cursos.

Para a Bahia, o primeiro reitor da UFBA trouxe com os seus 15 anos de mandato (1946-1961) a construção do Hospital Universitário, criação do Centro de Estudos Afro-Orientais e dos *campi* do Canela, Federação e Ondina, tendo o Estado também alcançado com o seu mandato o prestígio para a Dança, Teatro e Música, cursos universitários pioneiros do gênero no país (UFBA, 2021a).

No ano de criação da UFBA, em 1946, registravam-se 617 inscritos em seus processos seletivos via vestibular, para um total de 17 cursos (UFBA, 2006). Esse número contrasta-se de forma significativa com o atual quantitativo possuído pela instituição: 106 cursos distribuídos por 3 locais em todo o estado (Salvador, Vitória da Conquista e Camaçari) e 40.727 matriculados (UFBA, 2020a). Vistos de uma série história, até o ano de 2016, a UFBA havia graduado aproximadamente 105 mil alunos, titulado cerca de 3 mil doutores e 12 mil mestres (UFBA, 2016b), o que demonstra a grandeza e importância dessa instituição no retorno de cidadãos qualificados para a sociedade brasileira.

Na universidade, os cursos de graduaçao estão distribuídos em 5 grandes áreas: Área I - Ciências Físicas, Matemática e Tecnologia; Área II - Ciências Biológicas e Profissões da Saúde; Área III - Filosofia e Ciências Humanas; Área IV - Letras e Área V - Artes (UFBA, 2021b). Os cursos escolhidos para o objetivo deste trabalho foram os de Estatística e Matemática, que fazem parte do grupo de cursos da Área I.

Não é difícil verificar a relevância da UFBA enquanto IES pública e de qualidade. Em agosto de 2021, a divulgação da *QS Latin America University Rankings* para o ano 2022, da consultoria britânica Quacquarelli Symonds (QS) trouxe a UFBA como 70ª melhor universidade da América Latina. Para o Brasil, a universidade ocupa a 19ª posição dentre as melhores, e, em nível regional, é eleita a 3ª melhor universidade da região Nordeste (QS, 2021).

Em abril de 2021, e mesmo enfrentando nos últimos anos um panorama danoso de redução orçamentária, a UFBA alcançou 3,84 pontos no Índice Geral de Cursos (IGC). O IGC é o indicador do MEC responsável em avaliar a qualidade dos cursos de graduação e de pós-graduação das instituições de educação superior brasileiras. O resultado alcançado colocou a UFBA como conceito 4 de classificação do IGC, em uma escala que vai no máximo até 5, revelando o empenho da instituição em manter um padrão de qualidade mesmo diante de desafios (SANGIOVANNI, 2021a). E ainda, em julho de 2021, a universidade conseguiu subir duas posições no *The World University Ranking*, sendo considerada a 26ª melhor universidade da América Latina, como também a 16ª melhor do Brasil e 1ª do Nordeste (SANGIOVANNI, 2021b).

Além de toda a produção acadêmica, a UFBA conta com uma vasta rede de pres-

tação de serviços públicos para toda a população. Em momentos de encontro e elo entre universidade e sociedade, a universidade disponibiliza serviços de atenção à saúde bucal, assistência psicossocial, complexo hospitalar, maternidade, hospital veterinário, diversos laboratórios, bibliotecas, museus e cursos livres aos cidadãos. Ações e serviços estes que promovem a saúde pública, educação e cultura na vida em sociedade (UFBA, 2020b).

A magnitude e extensão de tudo e todos que compõem a UFBA, bem como o papel que desempenha enquanto agente de transformação social como IES são inumeráveis. Por essa razão e por muitos outros fatores, se torna tão necessário verificar os quantitativos e perfis dos fenômenos da retenção/permanência, evasão e diplomação na universidade, especialmente por ser essa de natureza federal.

Uma vez deliberado como meta do PDI, é necessário reiterar que o acompanhamento das taxas de evasão e retenção na universidade é uma dificuldade enfrentada pelas IES em todo o território nacional. Desse modo, as ações e o empenho voltados para a manutenção do estudante no ensino superior, evitando que o mesmo venha a evadir bem como consiga o sucesso na sua diplomação, é urgente e indispensável, especialmente na instituição a qual este trabalho delimita-se.

2.3.1 Política de reserva de vagas

Ao pontuar o histórico da UFBA é necessário, também, apresentar alguns fatores cruciais e alicerces para o seu funcionamento, e que norteiam a expansão da acessibilidade ao seu ensino. Entre esses fatores, é apresentado a política de reserva de vagas na instituição. No presente estudo, a variável de reserva de vagas, indicando se o estudante entrou por cotas ou não, também foi utilizada nas análises, conforme será descrito na Seção 4.

Para alunos cotistas, por exemplo, Santos (2013) apresenta em seus resultados, com uma amostra de 3.844 estudantes cotistas e não costistas matriculados na UFBA, a valorização desse grupo em relação ao ensino superior. Santos (2013) ainda pontua que várias pesquisas mostram que a determinação de ingressar em uma universidade pública de prestígio e qualidade, e a responsabilidade, com isso, na trajetória escolar, constituem elementos comuns entre estudantes que ingressam por reservas de vagas em universidades públicas. Santos (2013) ainda disserta trazendo que esse comportamento é reflexo de uma orientação voltada a uma boa formação acadêmica, advinda de um projeto familiar no qual é extremamente valorizado a conquista em passar em um curso superior, sendo essas razões colocadas em primeiro plano.

Na UFBA, as ações afirmativas para ingresso à instituição tiveram suas discussões iniciadas no ano de 2002. Com o programa Universidade Nova, iniciou-se na comunidade os debates relacionados ao tema. O programa estabeleceu, entre outras ações, a reserva de vagas para alunos pretos e pardos oriundos de escola pública. A política de ações afirmativas da universidade foi aprovada em julho de 2004, através da resolução Nº 01/04

elaborada pelo Conselho de Ensino, Pesquisa e Extensão (CONSEPE) (PEIXOTO et al., 2016). E, no final do ano de 2004, o edital do vestibular de 2005 foi publicado, sendo o primeiro do Brasil a trazer como opção a reserva de vagas para pretos e pardos provenientes de escola pública (UFBA, 2019b).

A reserva de vagas, ou também conhecida como sistema de cotas, é a ação com maior enfoque do programa de ações afirmativas da UFBA. Também é a geradora de maior polêmica, visto que modificou a prática história de privilégios de grupos mais favorecidos em benefício daqueles com menos condições de ingresso no ensino superior. A UFBA foi a terceira universidade pública federal do país a implantar um sistema de cotas para entrada de estudantes de grupos minoritários. No processo seletivo de 2005, a universidade reservou 43% das vagas de todos os seus cursos de graduação para alunos de escola pública sendo 85% destinadas a estudantes autodeclarados negros (pretos e pardos) (SANTO, 2013).

O sistema de reserva de vagas da universidade sofreu modificação em 12 de novembro de 2012. Com a Resolução Nº 03/2012, o CONSEPE aprovou as novas diretrizes para o processo seletivo do ano de 2013 (SANTO, 2013). Em conformidade com a Lei Nº 12.711, sancionada para determinar que a reserva de vagas fosse de 50% para estudantes de escolas públicas, para todas as universidades públicas federais, foi também estabelecido critérios adicionais de etnia e de renda familiar (BRASIL, 2012). Além dessas mudanças, a primeira etapa do processo seletivo de 2013 foi substituída pela pontuação obtida do candidato com o Exame Nacional do Ensino Médio (ENEM).

A implantação da forma única de ingresso via Sistema de Seleção Unificada (SiSU) ocorreu a partir de 2014, e desde então os candidatos participam do processo seletivo da UFBA utilizando a nota alcançada com o ENEM. Nos editais de ingresso de 2015 até 2019 via SiSU para a UFBA, as modalidades de vagas eram as seguintes:

- Candidatos de Escola Pública/Pretos/Pardos/Indígenas com renda menor ou igual a 1,5 salário mínimo;
- Candidatos de Escola Pública/renda menor ou igual a 1,5 salário mínimo;
- Candidatos de Escola Pública/Pretos/Pardos/Indígenas;
- Candidatos de Escola Pública;
- Candidatos de Ampla Concorrência, que não concorrem através das demais reservas de vagas e usam apenas a nota do ENEM como critério de classificação.

No período de 2015 a 2019, essas categorias eram comuns para todos os editais. A exceção é verificada no edital de 2015, que ainda trazia uma modalidade de vagas somente para candidatos autodeclarados indígenas e também uma modalidade para candidatos índios aldeados ou moradores das comunidades remanescentes dos quilombos; e os editais

de 2018 e 2019 que incorporaram a modalidade de vagas contemplando candidatos deficientes e que se encaixavam nas outras modalidades (escola pública, renda e etnia - pretos, pardos e indígenas) (UFBA, 2021c).

A implantação da Lei de Cotas teve reflexo direto na expansão do acesso ao ensino superior nas instituições federais brasileiras. De acordo com o estudo realizado por Senkevics e Mello (2019) baseados em dados do CENSUP e do ENEM, a participação de jovens de 18 a 24 anos nas Ifes brasileiras, autodeclarados pretos, pardos ou indígenas e com renda igual ou inferior a 1,5 salário mínimos aumentou de 33,9% para 42,7% de 2012 a 2016. Para os estudantes de escola pública, esse percentual teve um aumento de 15% no mesmo período, segundo os autores.

Na UFBA também é notória a mudança substancial do perfil de estudantes após a Lei de Cotas. Conforme divulgado na *V Pesquisa do Perfil Socioeconômico dos Estudantes das Universidades Federais*, promovida pela Andifes, junto com o Fórum Nacional de Pró-Reitores de Assistência Estudantil (Fonaprace), a amostra de 5.774 estudantes de graduação da UFBA de um total de 38.674 alunos no ano de 2018 revelou que três em cada quatro estudantes são considerados negros. Além disso, o Índice de Inclusão Racial (IIR) trazido pela pesquisa informa que a universidade apresenta um IIR de 1,02, o que simboliza que o perfil racial dos estudantes da instituição aproxima-se bastante daquele encontrado na população do estado, de acordo com os dados levantados. Isto é, quanto mais próximo de 1 é o valor do IIR, mais a proporção de negros na universidade reflete a proporção dessa população no estado (ANDIFES, 2019).

2.3.2 Atividades extracurriculares

As atividades extracurriculares citadas nesse tópico se limitam a quatro específicas: pesquisa, extensão, monitoria e estágio. Isso porque, os dados trazidos pelo CENSUP abordam estritamente a realização desses tipos de atividades pelos discentes.

Para complementar a integralização curricular dos cursos de graduação da universidade, os estudantes precisam, além da carga horária de disciplinas obrigatórias e optativas, cumprirem carga horária em atividades complementares. A quantidade e a forma como cada tipo de atividade realizada pelo estudante será aproveitada fica a critério de cada colegiado de curso, que possui geralmente resoluções específicas para esse fim.

Como exemplo para a Pesquisa, podemos citar a atividade realizada pelo estudante na graduação em projetos de iniciações científicas, que são executadas mediante o Programa Institucional de Bolsas de Iniciação Científica (PIBIC). O PIBIC propõe-se a formação de estudantes de graduação da UFBA em pesquisa científica. Atualmente, a gestão da Pesquisa Universitária na universidade é de competência da Pró-Reitoria de Pesquisa, Criação e Inovação (PROPCI). À PROPCI, cabe, portanto, as funções de fomentar, coordenar, supervisionar, avaliar e controlar as políticas, os programas e os

projetos de pesquisa, criação e inovação da instituição, conforme descrito no Regimento da Reitoria (UFBA, 2013).

Na UFBA, o PIBIC é financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela Fundação de Amparo à Pesquisa no Estado da Bahia (FAPESB) e pela própria universidade (UFBA, 2021e). O programa geralmente é ofertado com bolsa, tendo em alguns casos a opção voluntária, a depender do projeto. As atividades desempenhadas pelo aluno nessa modalidade são conciliadas com o turno desempenhado para as atividades acadêmicas.

Nas atividades de Extensão, como exemplo, é possível citar aquelas que dizem respeito às modalidades de projeto, cursos, eventos, entre outros. Na UFBA, essas atividades são promovidas por docentes, servidores técnico-administrativos e por instâncias da universidade cuja coordenação dessas ações são de incumbência de docentes e servidores técnicos-administrativos da instituição (UFBA, 2014). Dentre as atividades fornecidas, há o Programa de Bolsas de Iniciação à Extensão Universitária (PIBIEX), na qual o aluno desenvolve um projeto de extensão de forma remunerada, com valor decidido de acordo com disponibilidade orçamentária da Pró-Reitoria de Extensão Universitária (PROEXT) (UFBA, 2014). Essas atividades possuem quantidade de horas estabelecidas de acordo com cada ação realizada, e permitem ao aluno uma formação profissional com responsabilidade social.

Outro tipo de atividade extracurricular que pode ser desempenhada pelos estudantes de graduação são as monitorias. As atividades de monitoria no âmbito dos cursos de graduação da UFBA foram regulamentadas na Resolução Nº 05/2021 pelo CAE. Por essa tarefa, o CAE discorre que seus objetivos estão abarcados na contribuição da melhoria da qualidade do processo de ensino - aprendizagem - avaliação, uma vez que os projetos em monitoria envolvem alunos dos cursos de graduação na execução de atividades que estão vinculadas a componentes curriculares do curso (UFBA, 2021d).

Por fim, a UFBA também dispõe de realização da estágio por parte dos discentes. O estágio pode ser de natureza obrigatória, participando da matriz curricular do curso do estudante, ou não obrigatória, quando os estudantes não possuem o estágio como um componente da matriz curricular da sua graduação. Em consonância com a Lei Nº 11.788, de 25 de setembro de 2008, que dispõe sobre os estágios obrigatórios e não obrigatórios no país (BRASIL, 2008), cada colegiado de curso da UFBA define suas próprias diretrizes em resoluções acerca dos critérios para realização do estágio.

No caso dos dados do CENSUP, que trazem a variável de atividade extracurricular, para a modalidade de estágio estão sendo considerados apenas àqueles estágios de natureza não-obrigatória (não participam das matrizes curriculares dos cursos de graduação).

3 METODOLOGIA

Para a tratativa e abordagem dos microdados do CENSUP, são empregados os seguintes métodos estatísticos: análise descritiva e exploratória para verificação quantitativa das variáveis envolvidas neste estudo, e análise de sobrevivência, para verificar o tempo de permanência dos estudantes na UFBA. Nessa seção são dispostos os tópicos a respeito da natureza dos dados, variáveis utilizadas e ênfase nas técnicas metodológicas empregadas.

3.1 Descrição dos dados e ferramentas utilizadas

Os microdados do CENSUP podem ser obtidos em sítio na *internet* para *download*. O arquivo é constituído de uma pasta zipada, contendo planilhas em formato `csv`. Para cada módulo distinto abordado no censo (aluno, curso, instituição e docentes) existe uma planilha preenchida referente às informações da competência de 2019. Há também uma planilha com um dicionário de todas as variáveis de cada módulo (INEP, 2019c). Para o objetivo deste trabalho, apenas as planilhas de aluno e curso foram utilizadas. Cabe ressaltar que em fevereiro de 2022, no decorrer deste estudo, houve uma atualização no *website* do INEP, em que a planilha contendo informações referente ao aluno foi retirada do ar. Desse modo, estes dados não encontram-se mais disponíveis.

Sobre a base contendo informações dos alunos, as variáveis dizem respeito ao sexo, idade, cor/raça, código de identificação do aluno, nacionalidade, ano de ingresso no curso, tipo de ingresso, dentre outras. Para a planilha de dados sobre os cursos são encontradas referências sobre o nome do curso, turno, carga horária, prazo mínimo de integralização, etc. No Apêndice A é disponibilizado a tabela com a listagem completa das variáveis que foram utilizadas neste trabalho bem como as variáveis que foram criadas para prosseguir com as análises estatísticas, todas com as suas devidas definições.

A base de dados do curso diz respeito a uma planilha razoavelmente simples de tamanho 9.659 KB, facilmente manuseada através do programa `Microsoft Excel (2019)`. Já a planilha de alunos possuem informações referentes aos estudantes de todas as IES do Brasil, públicas e privadas e, portanto, o seu tamanho se resume a um arquivo de quase 3 GB, sendo difícil a pré-visualização através do programa `Microsoft Excel (2019)` em uma máquina intermediária de 8 GB de memória ram e processador Intel CoreI5 (configuração do computador utilizado para as análises deste trabalho).

A ferramenta aplicada para todas as análises a serem apresentadas aqui foi o `R`. O `R` é um *software* livre e gratuito, amplamente utilizado para análise de dados utilizando técnicas estatísticas e possui linguagem de programação própria. A versão utilizada foi a

4.1.1. Os pacotes disponíveis no `R` que foram usados para o manuseio, tratativa da base de dados do CENSUP e criação de gráficos foram especificamente o `dplyr`, `sparklyr`, `lubridate`, `ggplot2` e `survminer`. Na modelagem de eventos competitivos utilizou-se os pacotes `survival`, `cmprsk`, `mstate` e `discSurv`. Para a construção do aplicativo *web Shiny* foram utilizados os pacotes `shiny`, `shinydashboard`, `shinythemes`, `shinyWidgets`, `dashboardthemes`, `plotly` e `rsconnect`.

3.2 Pré-processamento e limpeza dos microdados do CENSUP

Devido ao tamanho da base de dados contendo informações dos alunos ser muito grande e dificultar uma pré-visualização dentro do ambiente `R`, se fez necessário a utilização do pacote `sparklyr`. O pacote possibilitou realizar a filtragem de interesse: alunos da IES UFBA, uma vez que o banco total continha dados referentes a todas as IES do país.

O `sparklyr` é um pacote do `R` que permite uma interface entre o *software* estatístico e o *Spark*. O *Apache Spark* é um *framework* de código fonte aberto que tem o objetivo de processar grandes conjuntos de dados de forma paralela e distribuída (SPARK, 2021). O `sparklyr` faz o papel de "ponte" para que seja possível acessar os dados na fonte em que ele encontra-se armazenado, permitindo assim o manuseio dos dados no ambiente `R`.

Ao carregar o pacote `sparklyr` no *software*, é necessário que se faça a utilização em conjunto do pacote `dplyr`, uma vez que o *Spark* comunica-se com o ambiente *R* através das funções disponíveis nesse outro pacote. O *Spark* dentro do `R`, facilitado pela ferramenta do `sparklyr`, é requerido apenas para a filtragem do código da IES de interesse (código 578 para a UFBA). Após a realização do filtro, não foi mais necessário seguir com a conexão do *Spark* na interface `R`, visto que os dados após esse processamento já conseguiam ser visualizados dentro do *software* e também manuseados para as limpezas e análises que foram feitas posteriormente. A filtragem para obtenção do banco contendo as informações dos estudantes da UFBA retornou um total de 44.212 observações.

Referente a limpeza da base de dados, após o pré-processamento citado anteriormente, realizou-se algumas filtragens. Foram retiradas observações com quantitativo ínfimo de respostas (para as categóricas) e neste trabalho considerou-se: apenas alunos ingressantes a partir de 2009; ingressos via Vestibular, ENEM e Vagas Remanescentes (eliminadas as demais opções existentes); estudantes de escola pública e privada (eliminadas as "não respostas"); estudantes cursando, formado, desvinculado, transferido para outro curso UFBA ou com matrícula trancada (falecidos retirados).

Uma das principais características de interesse neste estudo é a situação do estudante no CENSUP 2019, que é definida da seguinte maneira:

- Cursando: aluno que está matriculado em alguma disciplina e que não concluiu a totalidade da carga horária exigida para a conclusão do curso;

- Matrícula trancada: aluno que está com a matrícula trancada na IES;
- Desvinculado do curso: aluno que, não possui vínculo com o curso em decorrência de evasão, abandono, desligamento ou transferência para outra IES;
- Transferido para outro curso da mesma IES: aluno que foi transferido para outro curso de graduação da mesma IES;
- Formado: aluno que concluiu a totalidade dos créditos acadêmicos exigidos para titulação no curso durante o ano de referência da coleta. Para o CENSUP, não é obrigatório que o aluno tenha realizado a colação de grau e/ou participado do Exame Nacional de Desempenho de Estudantes (ENADE).

Optou-se por agrupar as categorias desvinculado do curso e transferido para outro curso da UFBA como uma nova categoria denominada evadido. Essa foi uma escolha particular feita, visto que a UFBA não possui o conceito bem definido. Para classificar estudantes retidos, utilizou-se o seguinte critério: caso o estudante possuísse tempo (em semestres) superior ao tempo mínimo de integralização do curso + 50% desse tempo mínimo (considerado tempo máximo para conclusão), e não tivesse ao menos integralizado 50% das horas totais prevista na matriz curricular, seria considerado retido no curso. Desse modo, uma nova variável foi criada englobando as novas situações de vínculo: formado, cursando, evadido, retido e matrícula trancada.

Uma problemática encontrada nos dados tem relação com a variável de identificação única do aluno. A análise dessa variável apontou que códigos de identificação, a princípio únicos, se repetiam 2 e até 3 vezes na base geral. Ao examinar-se, foi constatado que essas diferenças decorriam de estudantes que faziam troca de curso no decorrer do ano (o ano do censo de 2019). No caso específico de repetição desse códigos por 3 vezes na base, os estudantes se enquadravam em cursos de Área Básica de Ingresso (ABI).

Cursos dessa natureza (ABI) possuem tanto o grau em bacharelado quanto em licenciatura. Os estudantes ingressam nos cursos de sua preferência com a denominação de ABI. No decorrer da graduação, após a conclusão de um conjunto básico de disciplinas, ou próximo a se formar, o estudante pode decidir para qual dos graus fará a migração (bacharelado ou licenciatura) (INEP, 2019d). Essa situação ocorre em um dos cursos que escolheu-se trabalhar, o de Matemática, visto que o mesmo possui as duas modalidades, bacharelado e licenciatura.

Foram encontrados, na base total UFBA, 1231 códigos de estudantes que se repetiam 2 vezes e 82 códigos que se repetiam 3 vezes. Para fazer a escolha de qual tipo de situação, dentre as repetições do aluno, que seria deixada na base, usou-se um critério de hierarquia: preferência em retirada da situação cursando, em seguida matrícula trancada, desvinculado do curso/transferido para outro curso da UFBA e só em último caso retirar

o aluno formado. Essa hierarquia foi baseada no alto percentual de censura encontrada em toda a base UFBA (92,20%, que englobam todos os casos opostos a estudantes formados, como tratou-se em um primeiro momento, mas que foi modificado para a análise em eventos competitivos). A definição de censura será trazida com maior ênfase e detalhamento na Seção 3.4.

Como o evento de interesse (formado) compunha apenas 7,80% do total de observações na base UFBA, priorizou-se em deixar estudantes, dentre os repetidos, com essa situação visto que é o nosso evento de interesse. A escolha da retirada, como primeira opção, dos discentes dentre as repetições que estivessem com situação cursando, também foi apoiada com base nesse quantitativo de estudantes representar mais de 50% do conjunto de dados com as informações de toda a UFBA. Logo, impactaria em um menor percentual de perdas das demais categorias.

Uma outra variável necessária a criação para dar suporte às análises em sobrevivência foi referente ao semestre de conclusão do aluno no censo de 2019 (1º ou 2º semestre do respectivo ano). Para universidades federais, essa informação não é disponibilizada pelo CENSUP, conforme consta no Manual de Preenchimento do Módulo Aluno (INEP, 2019e). Assim sendo, foi preciso uma estratégia para preencher essa variável, uma vez que a variável de interesse que também seria criada: tempos, em semestres, dos estudantes na UFBA, dependia também do conhecimento da data de conclusão (1º ou 2º semestre de 2019) para os discentes que se formaram no ano de 2019.

A solução encontrada e mais viável neste momento foi atribuir, para os estudantes concluintes em 2019, a data de formatura como 2º semestre caso o discente possuísse o ingresso em 1º semestre (variável de ano de ingresso do estudante ao curso, já trazida pelo censo), e data de formatura como 1º semestre de 2019 caso o estudante tivesse ingressado no 2º semestre na UFBA. Admitindo um discente semestralizado, em tese, esse é o comportamento corriqueiro visto informalmente na instituição. Ressalva-se, no entanto, a limitação em atribuir-se "arbitrariamente" esse critério, pela possibilidade de introdução de algum tipo de viés nas análises. Por hora, a solução encontrada para tal foi proceder dessa maneira.

Para verificar o tempo, em semestres, que os estudantes tinham até o ano de 2019, incluindo todas as possíveis situações dos alunos, criou-se uma nova variável contendo os tempos respectivos que cada um possuía na instituição. A variável foi obtida mediante operação de diferença entre a data de ingresso no curso (que é também informada com referência ao semestre, exemplo: 2015.1) e data do censo de 2019 (2º semestre de 2019 para os alunos com situação diferente de concluinte), considerando para isso uma conta em semestres completos. Sendo assim, a variável resposta de interesse (tempo de permanência) é do tipo quantitativa e discreta. Para os estudantes formados, conforme já mencionado no parágrafo anterior, essa conta foi feita com a diferença entre a data de ingresso no

curso e o semestre de conclusão, dado pela variável de semestre de conclusão, que já havia sido preenchida com as respostas de interesse.

A carga horária mínima necessária para o estudante se formar no curso é uma variável referenciada tanto no módulo aluno, quanto no módulo curso. Ao incoporar essa variável de carga horária mínima (informada na planilha de curso) para o banco original contendo as informações do aluno, percebeu-se que para muitos cursos haviam inconsistências e divergências de informações. Essas divergências foram possíveis de serem identificadas ao realizar a comparação dessa variável com a variável de quantidade de carga horária total do curso (que já é trazida no banco dos alunos, sendo, portanto, variáveis que retratam a mesma característica).

Em consulta ao manual de preenchimento do CENSUP, verificou-se que, enquanto a variável de quantidade de carga horária total (contida na base de dados dos alunos) é preenchida manualmente pelos encarregados da IES, a variável de carga horária mínima (contida na base de dados dos cursos) é declarado pelo sistema E-MEC, não sendo possível a sua modificação por parte das IES (INEP, 2019e). Uma das incosistências, inclusive, foi observada para alunos ingressantes do próprio ano de 2019, en que as cargas horárias de todos os cursos (informada pela variável de carga horária na planilha de alunos) estavam menores do que as cargas horárias informadas pelo E-MEC na planilha de cursos.

Consequentemente, pela variável de carga horária (na planilha de alunos) tratar-se de um heteropreenchimento, optou-se pela utilização da variável de carga horária presente na planilha de cursos do CENSUP. No entanto, foi percebido uma outra contradição. Para a variável de carga horária preenchida pelo responsável da instituição, haviam diferenças de acordo com a variável de ano de ingresso do aluno. Alguns cursos verificados continham duas e até três cargas horárias distintas de matriz curricular.

Como a checagem curso a curso para correção de todas as incoerências dessa natureza seria algo trabalhoso e que demandaria mais tempo, e, haja vista que a variável atualizada do E-MEC (na planilha de curso) informaria apenas a atual carga horária do curso em 2019, estaríamos admitindo que os alunos ingressantes de anos anteriores também teriam a mesma carga horária, o que incorporaria mais um viés para o estudo. Isto posto, outras fontes complementares aos microdados do CENSUP precisaram ser buscadas.

Construiu-se uma nova planilha a partir dos microdados de 2009 a 2018, contendo as informações dos códigos dos cursos e respectivas cargas horárias para cada ano, bem como os prazos de integralização mínima para cada graduação. Sendo assim, foi possível incorporar uma nova coluna, fazendo referência a essa "nova" carga horária, contendo essas informações atualizadas. A nova coluna foi trazida ao banco original de alunos através das chaves de identificação dos códigos dos cursos e ano de ingresso. Nesse processo foi vivenciada uma outra adversidade. Feita a incorporação da nova variável de carga horária,

constatou-se que, para algumas linhas, os respectivos valores retornaram dados faltantes (denominados *missings*). Foi então verificada a quais códigos de cursos isso acontecia e observou-se que ao decorrer dos anos, muitos cursos sofreram modificação na sua cifragem, passando a carregar outro tipo de código equivalente. Então, outra planilha foi elaborada com esses códigos faltantes para criação da variável de carga horária.

Como critério de escolha para esse quesito, optou-se em identificar o código correspondente no ano imediatamente anterior ou posterior comparando caso a caso para essa decisão e, verificando se havia alguma mudança substancial entre os anos. Em seguida incorporou-se a carga horária dos anos escolhidos como a oficial para cada situação em que faltava a informação. Esses novos códigos de cursos e cargas horárias foram inseridas na planilha contendo todos os códigos dos cursos, dos anos de 2009 a 2019 e as respectivas cargas horárias. Isso feito, os dados dessa nova planilha foram trazidos para o banco original para prosseguir com as análises. Nessa mesma planilha a informação dos prazos de integralização mínima dos cursos também foi extraída dos censos dos anos de 2009 a 2018.

Observou-se que nos anos 2009 e 2010 o prazo de integralização dos cursos eram informados em semestres. Em 2011, passou a ser declarado em anos ou fração de anos. Desse modo, para adaptar ao nosso objetivo, o prazo de integralização de 2011 a 2019 foi convertido em semestres. Importante ressaltar que o prazo de integralização, para alguns cursos de modalidade integral e noturno, possuíam valores distintos. Nesse caso, a chave para incorporar a variável de integralização em semestres foi composta pelo código do curso, respectivo turno do mesmo e o ano correspondente.

As duas planilhas que foram construídas com base em fonte de dados complementares ao utilizado neste trabalho, referente a carga horária e integralização dos cursos, bem como para os dados faltantes dos códigos de cursos específicos, encontram-se disponível no repositório pessoal do `GitHub`, podendo ser consultado através do link <https://github.com/jessicafagundesg>.

Uma outra variável construída para auxiliar no objetivo das análises foi definida como a porcentagem da carga horária integralizada pelo estudante até o ano de 2019, a mesma foi calculada através do percentual equivalente da variável de carga horária total (quantidade de horas integralizadas do aluno no curso) em relação a variável que já havia sido criada, a carga horária nova.

Após todas as limpezas e tratativas efetuadas, a base com dados da UFBA resultou em um total de 30.693 informações de estudantes. No projeto inicial deste trabalho foram apresentados resultados para todos os estudantes da UFBA. Esses resultados estão disponibilizados no aplicativo *web* `R` *Shiny* <https://jessica-fagundesg.shinyapps.io/tempodepermanenciaufba/>. Nesta etapa final, no entanto, os resultados referentes a modelagem com tempos discretos e eventos competitivos restringiram-se aos estudantes

dos cursos de Estatística e Matemática. As análises trazidas aqui dizem respeito a um quantitativo de 434 discentes desses cursos.

3.3 Análise descritiva e exploratória

Uma vez realizadas as abordagens iniciais de pré-processamento dos dados e limpeza dos mesmos, a etapa de análise descritiva e exploratória foi iniciada. Nessa metodologia, empregam-se técnicas estatísticas com o objetivo de conhecer o comportamento das covariáveis trazidas nos microdados do CENSUP. As técnicas são baseadas em análises quantitativas e qualitativas, porcentagens, criação de tabelas e gráficos que demonstrem algum comportamento das variáveis envolvidas no estudo.

Para as covariáveis referente a características dos estudantes (sexo, idade, ter entrada por reserva de vaga/tipo de escola que concluiu o ensino medio e se realiza ou não atividade extracurricular) empregou-se as técnicas descritivas já conhecidas da estatística, e para a variável resposta de interesse (tempo de permanência dos estudantes na UFBA) foi utilizado a abordagem de eventos competitivos, em que estima-se o risco de subdistribuição (será abordado na Seção 3.4.3.2). Neste caso, as outras situações de vínculo do estudante ao curso (retido, evadido e matrícula trancada) são tratados como eventos que estão competindo entre si e que impedem o estudante de se formar (evento de interesse). Os estudantes classificados como "cursando" são tratados como as censuras, conceito que será abordado na Seção 3.4

Essa etapa descritiva e exploratória tem papel primordial em qualquer pesquisa, seja ela de natureza qualitativa, quantitativa, básica, aplicada, exploratória, descritiva, explicativa ou até mesmo inferencial (quando se têm o objetivo de realizar inferências sobre uma população a partir de uma amostra de dados provenientes da mesma). Um conjunto de dados bem conhecido e explorado possibilita que o andamento de etapas mais elaboradas e técnicas estatísticas mais avanças sejam feitas da maneira mais correta possível. Além disso, possíveis resultados observados na etapa da análise descritiva refletem nas investigações que serão feitas posteriormente.

Os pacotes do *software* `R` mais utilizados nessa etapa foram o `dplyr`, para manipulação e verificação dos quantitativos de interesse, e o `ggplot2` para a criação dos gráficos.

3.4 Análise de Sobrevivência

Como uma das áreas da Estatística que mais cresceu nas últimas décadas do século passado, a Análise de Sobrevivência tem como variável resposta o tempo até a ocorrência de um determinado evento (COLOSIMO; GIOLO, 2006). Com grande aplicação, principalmente na área de Medicina, o uso desse procedimento cresceu de 11% em 1979 para

32% em 1989, segundo Bailar-III e Mosteller (1992). Esse tempo é, geralmente, denominado como tempo de falha e pode representar o tempo até a morte de um paciente após descobrir uma determinada doença, ou mesmo até a cura e recidiva, por exemplo.

A aplicação de Análise de Sobrevivência também é vista na área de Engenharia, em que a mesma recebe o nome de Análise de Confiabilidade. Nesses casos, o interesse quase sempre é verificar o tempo de vida de componentes, peças e sistemas utilizados nesse campo. Além disso, sua aplicação também pode ser vista nas áreas de Sociologia, Biologia, Finanças, ou em qualquer outra circunstância que se faça necessário analisar o tempo até a ocorrência de um determinado acontecimento.

Neste trabalho a Análise de Sobrevivência será empregada na área da Educação, investigando o tempo de permanência dos estudantes de Estatística e Matemática da UFBA. Esse método permite a identificação de fatores que diminuem ou aumentam a probabilidade de sobrevivência do indivíduo ao evento. No presente trabalho, o evento principal de interesse é a diplomação dos estudantes, conforme discorrido nos objetivos. Ou seja, visamos identificar possíveis fatores que impedem os estudantes de diplomar-se no tempo inicialmente previsto, através da análise do tempo de permanência desses discentes na UFBA.

O uso dessa metodologia como aliada na investigação dos fenômenos envolvendo o tempo de permanência/retenção ou até mesmo evasão de alunos no ensino superior no país ainda é pouco vista. Algumas produções já mencionas como Saccaro (2016), por exemplo, utiliza o ajuste de modelos paramétricos e não paramétricos para verificar os fatores associados ao fenômeno da evasão nas IES públicas e privadas do país. Saccaro (2016) se apropria dos dados do CENSUP de 2009 a 2014 no seu projeto. Já a tese de Pinheiro (2021) utiliza outra abordagem estatística dentro da Análise de Sobrevivência como ferramenta de investigação da evasão em cursos de Engenharia da UFBA: o modelo de longa duração, que considera indivíduos suscetíveis ou não ao evento de interesse.

Outras produções que podem ser citadas são Echeveste (1997) e Martins e Rocha (2011). Echeveste (1997) investiga a evasão de discentes no curso de Estatística da Universidade Federal do Rio Grande do Sul ajustando modelo de regressão semiparamétrico de Cox. Martins e Rocha (2011) também trazem em seu estudo a investigação da evasão e do tempo de permanência de alunos do curso de Estatística da Universidade Federal do Paraná. Os autores fazem uso da análise não paramétrica com o estimador de Kaplan Meier, Modelos de Riscos Proporcionais ou Modelo de Cox e Modelo de Chances Proporcionais ou Modelo Logístico.

Podemos perceber que as possibilidades são diversas nesse âmbito de Análise de Sobrevivência. Neste trabalho, aplicou-se outra abordagem, através do uso de eventos competitivos considerando tempos discretos (os tempos em semestres até formatura/conclusão dos estudantes). Neste caso, a abordagem considera que as diversas situações de

vínculo do aluno (matrícula trancada, evadido, retido) são eventos que estão competindo entre si para a não diplomação do discente no prazo inicialmente previsto.

O emprego da Análise de Sobrevivência se mostra bastante adequado uma vez que, nas investigações desses fenômenos de evasão e retenção, é evidenciado que muitos estudantes podem não vivenciar o evento final de interesse (como a evasão, por exemplo, ou ainda não formar-se ao final do estudo, ficando retido), o que os caracteriza como um dado incompleto dentro dessa metodologia, reforçando a sua adequabilidade. Como a principal característica de dados de sobrevivência é a presença de censuras, que será definido a seguir, essa metodologia é capaz de incorporar a informação carregada por esse dado incompleto às análises realizadas. Outros métodos como Modelos de Regressão e Análise de Variância, por exemplo, seriam mais indicados caso houvessem dados completos, sem a presença dessa característica (COLOSIMO; GIOLO, 2006).

3.4.1 Conceitos iniciais

Os conjuntos de dados de sobrevivência são geralmente caracterizados pela presença de tempos de falha e tempos de censura. Ao admitirmos que a variável resposta nessa abordagem é o tempo até a ocorrência de um dado evento, caso o indivíduo venha a experienciar esse acontecimento, o tempo em que isso ocorre é caracterizado como tempo de falha desse indivíduo. Caso contrário, se ao final do estudo isso não ocorre então registra-se este tempo como um tempo de censura. Esses dois componentes constituem a variável resposta. Nos estudos envolvendo essa metodologia, muitas vezes também há o interesse em investigar possíveis razões que influenciam nesses tempos, o que é feito através de covariáveis incorporadas no estudo de interesse (COLOSIMO; GIOLO, 2006).

O tempo de falha é composto por três elementos: o tempo inicial, a escala em que é medido e o evento de interesse (falha). Todos os elementos devem estar definidos de forma clara e precisa para o bom andamento da investigação que se tenha interesse em realizar. Além disso, os indivíduos deverão ser comparáveis na origem desse estudo, exceto pelas diferenças evidenciadas através das covariáveis também presentes. Aqui, considerou-se como tempo inicial o 1º semestre/ingresso do estudante da UFBA, e efetuou-se as devidas manipulações para o cálculo do tempo de permanência do estudante até o ano do censo de 2019.

Referente a escala de medida, a mesma é representada pelo tempo real, tempo de calendário ou "de relógio". Pode ser também caracterizada por outras possibilidades (comuns na análise de confiabiliade, como por exemplo número de ciclos em um aparelho/sistema, quilometragem de veículos, etc), embora as primeiras alternativas listadas sejam as mais comuns. Neste trabalho, a escala de medida é simbolizada pelo tempo, em semestres, que o estudante possui na graduação da UFBA até o ano de 2019.

Os eventos de interesse, em grande parte dos estudos, especialmente na área mé-

dica, são representados por acontecimentos indesejáveis, como por exemplo o óbito de um paciente e a recidiva de uma doença, chamadas portanto de uma falha. No entanto, há situações que o evento pode não representar algo desfavorável. Contextualizando para o principal evento de interesse do presente trabalho, temos que o tempo medido é o tempo de permanência dos estudantes na UFBA até se formarem. Nesse caso, deixamos claro que o principal evento de interesse é desejado e positivo.

O evento de interesse também pode ocorrer devido a uma única causa ou a mais de uma. Em situações de mais de uma causa, dizemos que as causas de falha competem entre si, sendo denominadas de Riscos Competitivos (COLOSIMO; GIOLO, 2006). Essa é a metodologia que foi utilizada no presente trabalho, conforme já mencionado.

Uma outra definição a ser abordada aqui faz referência a presença das chamadas censuras na metodologia de Análise de Sobrevivência. Mesmo em estudos de longos acompanhamentos nesse tipo de análise, os indivíduos podem não vir a falhar ao término dos mesmos. Temos então a presença de informações incompletas, ou seja, os indivíduos não falharam até o momento final do acompanhamento, podendo vir a falhar em um tempo superior, a qual não se tem informação. Portanto, são dados os quais não se tem conhecimento do evento, sendo necessariamente tratados como referências incompletas. Sendo assim, todo esclarecimento que se tem referente aos indivíduos com tempos de censuras são de que o tempo até a ocorrência do evento, para cada um deles, será superior ao tempo que foi registrado até o último acompanhamento (COLOSIMO; GIOLO, 2006).

Em contrapartida, salienta-se que, embora os resultados sejam censurados, todos os achados advindos de um estudo com natureza de dados do tipo de sobrevivência devem ser incorporados nas análises estatísticas. Isso porque, mesmo sendo incompletas, as observações censuradas fornecem informações sobre o tempo de vida de pacientes, em caso de estudos clínicos, por exemplo (COLOSIMO; GIOLO, 2006). Analogamente ao estudo desenvolvido aqui com o tempo de permanência dos estudantes, ainda que não formados até 2019, os seus tempos de censura fornecem a informação de que o tempo em que esses concluirão o curso é posterior ao ano de 2019, ou seja, poderá ocorrer em algum momento após esse ponto do tempo.

As razões usuais para a ocorrência de uma censura pode ter relação não somente com o fato do indivíduo não ter "falhado" até o término do estudo, como também o seu acompanhamento ser perdido por alguma razão (mudança de cidade, o indivíduo é retirado do estudo ou vem à óbito por alguma outra causa). Aqui, considerou-se como censuras os casos de estudantes com situação de vínculo cursando (que ainda não alcançaram o tempo máximo previsto na matriz curricular). Ou seja, não experimentaram nenhum dos eventos até aquele ponto no tempo. Os demais tipo de situação de vínculo (evadido, retido e matrícula trancada) são tratados como os eventos competitivos, sendo o evento formar o nosso evento de interesse.

A censura pode apresentar três mecanismos distintos: censura à direta, à esquerda ou intervalar. A censura à direita é a situação frequentemente encontrada envolvendo estudos com dados de sobrevivência, e ocorre quando o tempo de ocorrência do evento de interesse está à direita do tempo registrado (poderá ocorrer depois do tempo final registrado na pesquisa). A censura à esquerda ocorre quando o evento de interesse já sucedeu quando o indivíduo foi observado, como por exemplo em uma investigação da idade em que crianças começaram a ler. No início do estudo podem ser observadas crianças que já sabiam ler (evento de interesse). Por fim, a censura do tipo intervalar ocorre quando o evento de interesse acontece em um determinado intervalo entre os tempos de medição/acompanhamento do indivíduo (COLOSIMO; GIOLO, 2006). Novamente contextualizando com a temática deste trabalho, a censura verificada aqui é do tipo à direita, uma vez que os estudantes poderão vivenciar o evento de interesse (formatura) após o tempo de referência de 2019. Isto é, o tempo de ocorrência da diplomação dos discentes está à direita do tempo final registrado.

Os dados de sobrevivência para um determinado indivíduo i $(i = 1, ..., n)$ são representados em geral pela dupla $(\tilde{T}_i, \Delta_i)$. Tem-se que T_i é a variável contendo os tempos de falha, C_i é a variável contendo os tempos de censura e $\tilde{T}_i = \min(T_i, C_i)$ é uma variável contendo os tempos de falha e de censura dos indivíduos. Δ_i representa a variável indicadora de falha ou censura. Em outras palavras, temos que:

$$\Delta_i = \begin{cases} 1, & \text{se } \tilde{T}_i \text{ é um tempo de falha} \\ 0, & \text{se } \tilde{T}_i \text{ é um tempo censurado.} \end{cases}$$

Neste trabalho, construiu-se uma variável para abranger os tempos de falha e de censura conjuntamente, e utilizou-se uma variável indicadora de estudante concluinte (aquele que já integralizou todas as horas previstas na matriz curricular) como sendo a variável indicadora de censura (informa se o aluno é ou não concluinte em 2019). Os tempos de falhas são referentes aos demais tipos de situação de vínculo do estudante à UFBA (formado, evadido, retido ou matrícula trancada).

Na presença de covariáveis medidas nos i indivíduos, como por exemplo sexo, idade, etc, tem-se a representação desses como $\boldsymbol{Z}_i$, e então os dados de sobrevivência são representados por $(\tilde{T}_i, \Delta_i, \boldsymbol{Z}_i)$. A variável aleatória não-negativa, T, usualmente contínua (mas que neste trabalho é classificada como discreta, medindo o tempo, em semestres), que representa o tempo de falha é especificada através da sua função de sobrevivência ou função taxa de falha (risco). Essas funções são usadas de forma extensiva em Análise de Sobrevivência (COLOSIMO; GIOLO, 2006).

A função de sobrevivência é uma das principais funções probabilísticas utilizadas para descrever trabalhos com dados nessa estrutura. É definida como sendo a probabili-

dade de um indivíduo não vir a falhar até um determinado instante de tempo t. Isto é, a probabilidade de o indivíduo sobreviver ao tempo t. Ou seja, seria a probabilidade da variável aleatória T ser maior ou igual a um instante no tempo t (COLOSIMO; GIOLO, 2006). Por se tratar de um sentido ambíguo quando comparado a abordagem dada neste trabalho, essa função de sobrevivência expressaria então a probabilidade de o indivíduo permanecer no ensino superior (não vir a formar) em um determinado tempo t.

Cabe também mencionar que a presença de censuras acarreta problemas para as técnicas convencionais de análise descritiva envolvendo média, desvio-padrão, técnicas gráficas como histograma, *boxplot*, etc. Nesse caso, o principal componente de análise descritiva envolvendo dados desse tipo diz respeito à função de sobrevivência. Basicamente o procedimento consiste em encontrar uma estimativa para a função de sobrevivência e só então estimar as estatísticas de interesse como tempo médio, tempo mediano e alguns percentis. As técnicas estatísticas especializadas nesse contexto precisam acomodar as informações contidas nas censuras, que indicam que o tempo até a falha (evento de interesse) é maior do que aquele que foi registrado (COLOSIMO; GIOLO, 2006). A técnica não-paramétrica mais utilizada e tida como padrão nessas análises é o conhecido estimador de Kaplan-Meier, que será definido a seguir.

3.4.2 Estimador não paramétrico de Kaplan-Meier

Embora a metodologia utilizada para a abordagem dos tempos de permanência dos estudantes tenha sido os modelos com tempos discretos em riscos competitivos, faz-se necessário a definição do estimador de Kaplan-Meier, uma vez que os conceitos englobados no tópico de eventos competitivos utilizam esse estimador como parte de algumas notações. Menciona-se também que na primeira etapa deste trabalho (com o Trabalho de Conclusão de Curso I), analisou-se, preliminarmente, os dados dos estudantes de todos os cursos da UFBA através das curvas de Kaplan-Meier. Estes resultados inicias, bem como as respectivas análises descritivas estão disponíveis para consulta através do aplicativo *web* `R` *Shiny*: <https://jessica-fagundesg.shinyapps.io/tempodepermanenciaufba/>.

O estimador de Kaplan-Meier foi proposto por Kaplan e Meier (1958) para estimar a função de sobrevivência. Também chamado de estimador limite-produto, ele é uma adaptação da função de sobrevivência empírica, que, na ausência de censuras, é definida como:

$$\widehat{S}_{(t)} = \frac{\text{nº de observações que não falharam até o tempo } t}{\text{nº total de observações no estudo}}.$$

$\widehat{S}_{(t)}$ é uma função "escada" com degraus nos tempos observados de falha de tamanho $1/n$, em que n é tamanho da amostra. Caso existam empates (mais de uma falha simultânea) em um determinado instante de tempo t, o tamanho do degrau fica multipli-

cado pelo número de empates. Além disso, o estimador ao ser construído considera tantos intervalos de tempo quantos forem o número de falhas distintas existentes no conjunto de dados de sobrevivência. Os limites dos intervalos de tempos são dados pelos tempos de falha observados na amostra (COLOSIMO; GIOLO, 2006).

Dito isto, consideremos que existem n indivíduos sob teste e $r \leq n$ falhas distintas e ordenadas nos tempos $t_1 < t_2 < \cdots < t_r$, em que temos:

- d_j é o número de falhas em t_j, $j = 1, ..., r$, e
- n_j é o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o momento imediatamente anterior a t_j.

Nessas condições, o estimador de Kaplan-Meier é definido por:

$$\widehat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right). \tag{3.1}$$

As principais propriedades do estimador de Kaplan-Meier se resumem a: não ser viciado para grandes amostras, ser fracamente consistente, possuir distribuição assintótica normal além de ser estimador de máxima verossimilhança de $S(t)$ (COLOSIMO; GIOLO, 2006). Todas essas propriedades são desejáveis para um bom estimador. A consistência e normalidade assintótica de $\widehat{S}_{(t)}$ foram provadas, sob certas condições de regularidade, por Breslow e Crowley (1974) e Meier (1975). No seu artigo original, Kaplan e Meier (1958) mostram que $\widehat{S}_{(t)}$ é o estimador de máxima verossimilhança de $S(t)$.

Quando o maior tempo observado na amostra for um tempo de falha, a curva de $\widehat{S}(t)$ cairá para zero, o que é geralmente visto. Caso o maior tempo observado na amostra seja um tempo de censura, então a curva $\widehat{S}(t)$ não decairá para zero.

Através do estimador não paramétrico de Kaplan-Meier é possível a obtenção de estatísticas de interesse como o tempo médio e tempo mediano de vida dos indivíduos em observação, além de ser viável a obtenção de percentis de interesse. A utilização direta da curva fornecida com o estimador nos informa a probabilidade estimada de sobrevivência por um determinado tempo.

Uma estimativa para o tempo médio de vida do indíviduo é obtida calculando-se a área sob a curva da estimativa de Kaplan-Meier, tal que:

$$t_m = E[T] = \int_0^\infty S(t)t.$$

Como essa curva é uma função escada, essa integral é simplesmente a soma de áreas de retângulos, ou seja:

$$\widehat{t}_m = t_1 + \sum_{j=1}^{k-1} \widehat{S}(t_j)(t_{j+1} - t_j),$$

em que $t_1 < \cdots < t_k$ são os k tempos distintos e ordenados de falha (COLOSIMO; GIOLO, 2006).

Contudo, quando o maior tempo observado na amostra for um tempo de censura, a curva de Kaplan-Meier não atinge o valor zero, o que implica que o valor do tempo médio de vida fica subestimado, e a estimativa deve ser interpretada com bastante cuidado ou preferencialmente ser evitada. Uma possibilidade para esses casos é fazer uso do tempo mediano de vida ao invés do tempo médio de vida, visto que ambas são medidas de tendência central e representam um valor típico da distribuição do tempo de vida da população sob estudo (COLOSIMO; GIOLO, 2006).

O tempo de sobrevivência mediano é a medida-sumário mais citada e interpretada nos artigos científicos que trabalham com Análise de Sobrevivência. É uma medida mais usada do que a média, pois como a distribuição do tempo de sobrevivência é pouco simétrica, tendo um comportamento mais assimétrico, as medidas robustas são mais indicadas. Por definição, o tempo mediano é o tempo depois do qual 50% dos indivíduos estão vivos, ou seja, o tempo no qual $\widehat{S}(t) = 0,5$. Aqui, definiríamos que seria o tempo depois do qual 50% dos indivíduos ainda estão cursando (não diplomaram-se). Em outras palavras, o tempo de sobrevivência mediano é definido como o menor tempo para o qual o valor estimador de $S(t)$ é menor ou igual a 50%, tal que:

$$\widehat{t}_{mediano} = min\{t_j | \widehat{S}(t_j) \leq 0,5\},$$

em que j é o índice dos tempos observados de ocorrência do evento (CARVALHO et al., 2019).

Como a curva de Kaplan-Meier é uma função escada, as estimativas mais adequadas são obtidas através de interpolação linear. Essa forma de estimar esses valores é equivalente a conectar por retas as estimativas de Kaplan-Meier ao invés de se utilizar $\widehat{S}(t)$ na forma de escada. Gera também uma melhor representação da distribuição dos tempos contínuos de falha (COLOSIMO et al., 2002). Outros percentis da distribuição do tempo de vida dos indivíduos podem ser obtidos de maneira análoga.

3.4.3 Eventos competitivos

Na modelagem do tempo de sobrevivência pode surgir um problema clássico em que é observado a presença de eventos ditos competitivos. Nesses casos, somente é possível observar o tempo até a ocorrência do primeiro (e único) evento, que impede que

outros eventos aconteçam. Logo, é como se esses eventos estivessem competindo entre si, justificando então o nome atribuído (CARVALHO et al., 2019).

Na presença de riscos competitivos, cada indivíduo está simultaneamente em risco para k eventos, e a ocorrência de um desses eventos elimina a chance que qualquer outro ocorra. A exemplo, supondo um indivíduo que, ao entrar no estudo, encontra-se no estado vivo, isso é, ele está em risco para k tipos de eventos alternativos. Sendo $\lambda_A(t)$ o risco do indivíduo sofrer o evento A, e $\lambda_B(t)$ o risco de sofrer o evento B, a passagem do indivíduo do estado inicial (vivo) para o estado A (óbito pela causa A, por exemplo) é definitiva. Ou seja, os demais estados/causas de morte não poderão ocorrer, visto que o evento A já ocorreu (CARVALHO et al., 2019).

No contexto do estudo atual, temos que o estudante classificado como "evadido", por exemplo, não poderá se formar (evento de interesse). Analogamente, a situação ocorre para os outros eventos que estão competindo entre si e impedindo, naquele momento analisado, a diplomação/formatura do estudante no curso (como por exemplo, o discente ficar retido ou trancar sua matrícula).

Em estudos de análise de sobrevivência em que não há presença de riscos competitivos, mas com censuras, a maneira mais imediata e fácil de descrever os dados é por meio do gráfico da função de sobrevivência $S(t)$, calculada pelo estimador de Kaplan-Meier através da expressão (3.1). No entanto, quando há presença de riscos competitivos, a função de distribuição acumulada $F(t)$ estará subestimada. A função de distribuição acumulada, no contexto aqui inserido, é igual a $1 - S(t)$. E, na presença de riscos competitivos, essa função não será mais uma função de distribuição de probabilidade acumulada usual (CARVALHO et al., 2019).

Para visualizar essa problemática acerca da $F(t)$, temos que, na prática, quando calcula-se a $F(t)$ para todos os eventos competitivos juntos e depois comparamos esse valor com o somatório das funções de distribuição acumulada de cada evento $F_k(t)$, essa soma será maior que o valor $F(t)$ global. Inclusive, $\sum_k F_k(t)$ pode ultrapassar o valor 1, o que não condiz com a definição de uma função de distribuição acumulada. Ou seja, a estimativa das funções de sobrevivência para cada evento em separado estará viesado (CARVALHO et al., 2019).

Para contornar esse problema, foi proposto a utilização das chamadas funções de incidência acumuladas (FIA), definidas na seção 3.4.3.1.

3.4.3.1 Funções de incidência acumuladas (FIA)

As funções de incidência acumuladas ou FIA, também chamadas de subdistribuições, consideram não somente o desfecho estudado, mas a totalidade dos desfechos existentes (CARVALHO et al., 2019). Ao calcularmos a FIA para os dados deste traba-

lho, estaremos levando em consideração não somente o evento de interesse formar, mas também os eventos evadir, ficar retido e trancar a matrícula.

O procedimento para estimar corretamente o risco de formar (nesse caso enfatizamos que o risco para esse evento é algo desejável), é assumir que os eventos evadir, ficar retido e trancar matrícula são eventos competitivos, e vice-versa. Ou seja, a cada momento definimos um evento de interesse, e os demais admitimos como eventos competitivos. Nessa abordagem, teremos que a soma das subdistribuições será exatamente a $F(t)$ global. No entanto, as subdistribuições também não se caracterizam como funções de distribuição acumuladas, pois, na presença de eventos competitivos, elas nunca alcançam o valor 1. (CARVALHO et al., 2019).

Sejam $0 < t_1 < t_2 < \cdots < t_m$ os tempos ordenados da ocorrência de qualquer um dos eventos e das censuras, e $d_k(t_j)$ o número de indivíduos (neste trabalho, os estudantes) que experimentaram o evento k no tempo t_j. O total de ocorrência por todos os eventos no tempo t_j é dado por:

$$d(t_j) = \sum_{k=1}^{K} d_k(t_j).$$

Seja ainda $R(t_j)$ definido como o número de estudantes ainda sob risco (de sofrer qualquer um dos eventos ou ser censurado) no tempo t_j. Lembrando do então estimador de Kaplan-Meier apresentado na expressão (3.1), pode-se definir uma probabilidade geral de sobrevivência a toda e a qualquer causa como sendo:

$$\widehat{S}(t) = \prod_{j:t_j \leq t} \left(\frac{R(t_j) - d(t_j)}{R(t_j)} \right).$$

Temos então que a FIA pode ser calculada como a soma, em todos os tempos t_j, das probabilidades de se observar o evento k no tempo t_j entre estudantes que não experimentaram qualquer dos possíveis eventos. A probabilidade de estar livre de qualquer evento antes do tempo t_j é dada por $\widehat{S}(t_{j-1})$. E assim, a probabilidade acumulada de se sofrer o evento k e estar livre de qualquer evento em t, será dada por:

$$\widehat{F_k}(t) = \sum_{\forall j, t_j \leq t} \widehat{\lambda}_k(t_j) \widehat{S}(t_{j-1}), \tag{3.2}$$

em que $\lambda_k(t_j)$ é o risco da causa específica para o evento k no tempo t_j (CARVALHO et al., 2019).

Seja $\epsilon = 1, ..., k$ a causa da falha (evento), esse risco por causa específica é dado por:

$$\lambda_k(t_j) = \lim_{\Delta_t \to 0} \frac{P(t \leq T < t + \Delta_t, \epsilon = k | T \geq t)}{\Delta_t} \tag{3.3}$$

e representa a taxa de ocorrência do evento k ($\epsilon = k$) no tempo t, dado a sobrevivência até o tempo t, na presença de todas as possíveis causas de falha. Ou seja, de todos os eventos competitivos.

Como os indivíduos ainda sob risco não sofreram nenhum evento até o tempo t_{j-1}, o risco pode então ser calculado por $d_k(t_j)/R(t_j)$. Sendo assim, substituindo essa razão na expressão (3.2), podemos reescrevê-la como:

$$\widehat{F_k}(t) = \sum_{\forall j, t_j \leq t} \frac{d_k(t_j)}{R(t_j)} \widehat{S}(t_{j-1}),$$

em que $\widehat{S}(t)$ é o estimador de Kaplan-Meier calculado como em (3.1), obtido considerando todos os eventos como sendo do mesmo tipo (PINTILIE, 2006). Logo, a FIA é uma proporção entre o número de estudantes que sofreram o evento k e todos aqueles que não sofreram evento algum até aquele momento (CARVALHO et al., 2019).

Modelos de regressão baseados na proposta de subdistribuições foram desenvolvidos a fim de estimar adequadamente o efeito das covariáveis quando os eventos são competitivos. Outras abordagens utilizadas para modelar eventos competitivos dizem respeito a abordagem da sobrevivência livre de eventos e risco específico por causa, que não serão tratados aqui. O procedimento de modelagem da subdistribuição do risco possui a vantagem de não supor (ao contrário das abordagens citadas anteriormente), que os eventos competitivos sejam independentes (CARVALHO et al., 2019).

3.4.3.2 Subdistribuição dos riscos para tempos contínuos

Fine e Gray (1999) introduziram uma forma de estimar o efeito de covariáveis através da modelagem direta das FIA's, sem supor independência entre os tempos dos eventos. A proposta dos autores consideram tempos contínuos, mas serviram de base para a adaptação que foi realizada por Berger et al. (2018) para tempos discretos, conforme será verificada na Seção 3.4.3.3. Essa adaptação foi a utilizada para modelagem realizada neste trabalho, no entanto, realiza-se também o ajuste usual contínuo proposto por Fine e Gray (1999) a fim de comparações das estimativas obtidas.

Em um modelo de riscos competitivos, os dados observados para cada estudante, por exemplo, serão representados por uma tripla de variáveis aleatórias (T, C, Δ): T representa os tempos de falha, C os tempos de censura, e Δ é uma variável indicadora de censura. Seja $\epsilon = 1, ..., k$ as causas de falha, assumidas como observáveis. Em nossos dados, $\epsilon = 1, 2, 3, 4$, em que 1 é o nosso evento de interesse (formar), 2 é o evento evadir, 3 ficar retido e 4 trancar a matrícula.

Criou-se uma variável auxiliar contendo a codificação respectiva 0, 1, 2, 3 e 4 para identificação das censuras (0) e dos eventos competitivos (1, ...,4). Essa foi a variável

utilizada para realizar os devidos ajustes ao modelo final, e, portanto, esta variável inclui tantos os valores quantos forem os eventos, incluindo a censura.

Fine e Gray (1999) propuseram uma alternativa ao modelo clássico de riscos proporcionais de Cox, que já era utilizado na época e baseava-se nas funções de risco específico por causa como em (3.3), e também sob a formulação de riscos proporcionais. Essa formulação, feita no modelo de Cox, assume que as covariáveis tem um efeito multiplicativo na função de risco, e portanto a razão entre o risco de ocorrência do evento para dois indivíduos s e l, dado as suas covariáveis $\boldsymbol{Z}$, é constante ao longo do tempo. O modelo de riscos proporcionais de Cox ajusta a seguinte função de risco $\lambda(t|\boldsymbol{Z})$:

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)\exp(Z_1\beta_1 + Z_2\beta_2 + \cdots + Z_p\beta_p) = \lambda_0(t)\exp(\boldsymbol{Z}^T\boldsymbol{\beta}),$$

em que $\lambda_0(t)$ é uma função não negativa denominada risco basal e $\boldsymbol{\beta}$ o vetor de parâmetros que se deseja estimar. E ainda:

$$\frac{\lambda_s(t, \boldsymbol{Z}_s)}{\lambda_l(t, \boldsymbol{Z}_l)} = \frac{\exp(\boldsymbol{Z}_s\boldsymbol{\beta})}{\exp(\boldsymbol{Z}_l\boldsymbol{\beta})},$$

que é constante no tempo (CARVALHO et al., 2019).

Isso significa (com o exemplo do contexto aqui inserido) dizer que o risco de formar, para o estudante do sexo feminino, é sempre o mesmo ao longo do tempo. Algumas estudantes vão formar rapidamente e outras mais tarde, porém sempre na mesma proporção, estimada pela exponencial do coeficiente ($\exp(\beta)$), daí o nome modelo de riscos proporcionais. Na prática, nem sempre essa suposição será atendida, conforme veremos na Seção 4.2.2, através da análise de pressupostos do modelo.

Essa metodologia proposta por Cox, no entanto, não permitia estimar adequadamente o efeito de covariáveis quando os eventos são competitivos. Desse modo, Fine e Gray (1999) introduziram o conceito de um novo modelo semiparamétrico de riscos proporcionais para a subdistribuição.

Seja T os tempos de falha, C os tempos de censura, $\epsilon \in (1, ..., k)$ a causa da falha (assumida como observável) e $\mathbf{Z}$ o vetor de covariáveis independentes do tempo, de dimensão $p \times 1$. Para os dados usuais com censura à direita, observamos $X = min(T, C)$, $\Delta = I(T \leq C)$ e $\mathbf{Z}$, em que Δ é uma função indicadora de censura. Supomos que $\{X_i, \Delta_i, \Delta_i\epsilon_i, \mathbf{Z}_i\}$ são independentes e identicamente distribuídos para $i = 1, ..., n$. Com interesse de modelar a FIA para a falha da causa k, condicional às covariáveis, Fine e Gray (1999) definiram que a mesma é dada por:

$$F_k(t, \mathbf{Z}) = P(T \leq t, \epsilon = k|\mathbf{Z}),$$

e a função de risco de subdistribuição:

$$\lambda_k(t, \mathbf{Z}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P\{t \leq T \leq t + \Delta t, \epsilon = k | T \geq t \cup (T \leq t \cap \epsilon \neq k, \mathbf{Z})\}$$

$$= \left\{ \frac{\frac{dF_k(t,\mathbf{Z})}{dt}}{1 - F_k(t, \mathbf{Z})} \right\} = -\frac{d}{dt} \log\{1 - F_k(t, \mathbf{Z})\}.$$

Fine e Gray (1999) também impuseram a suposição de riscos proporcionais sob os riscos de subdistribuição:

$$\lambda_k(t, \mathbf{Z}) = \lambda_{k,0}(t) \exp\{\mathbf{Z}^T(t)\boldsymbol{\beta}\}, \tag{3.4}$$

em que $\lambda_{k,0}(t)$ é uma função completamente não especificada e não negativa em t, chamada de função de risco basal, $\mathbf{Z}$ é o vetor de covariáveis, e $\boldsymbol{\beta}$ é o vetor de parâmetros que desejamos estimar.

O vetor de parâmetros $\boldsymbol{\beta}$ é estimado a partir de uma verossimilhança parcial, similar ao que é feito no modelo de riscos proporcionais de Cox, porém difere no conceito do grupo que está sob risco $R(t_j)$ para cada evento. O procedimento consiste em eliminar a função de risco basal em (3.4), considerando-se apenas, a cada tempo t, as informações dos indivíduos que estão sob o risco para cada evento. É uma formulação semelhante ao modelo não paramétrico de Kaplan-Meier, com a vantagem de poder estimar também o efeito das covariáveis, isto é, o efeito de fatores de risco no tempo de sobrevivência (CARVALHO et al., 2019). Aqui, seria estimar o efeito de fatores de risco no tempo dos estudantes na UFBA até formar, por exemplo.

Para entender a formulação da verossimilhança parcial é necessário primeiro entender em como cada indivíduo que está sob risco de sofrer determinado evento k irá contribuir para a função de verossimilhança. Chamemos então de evento de interesse aquele para o qual estamos estimando o modelo, e de evento competitivo os demais. O conjunto de risco $R^*(t_j)$ agora inclui além de todos os indivíduos que não sofreram qualquer evento em t, todos aqueles que falharam por outros eventos competitivos em t, e que a censura também permanecerá neste conjunto. No entanto, não faz sentido que esses indivíduos contribuam para $R^*(t_j)$ da mesma forma que os demais. Por isso, é atribuído aos indivíduos que sofreram um evento competitivo um peso que muda a cada tempo t_j no qual ocorre o evento de interesse (CARVALHO et al., 2019).

Seguindo esse raciocínio, temos então que os estudantes que sofreram o evento competitivo antes da ocorrência de qualquer evento de interesse nos pontos $t_j > t_i$ permanecem no grupo de risco, mas não contribuem de maneira integral para a função de

verossimilhança. Neste caso, cada discente receberá um peso sempre que ocorre um evento de interesse. Esse peso, denominado por $w_l(t_j)$, de cada estudante l a cada momento t_j no qual ocorre um evento de interesse é definido como:

$$w_l(t_j) = \begin{cases} 1, & \text{se } l \text{ não tiver sofrido o evento de interesse ou censura;} \\ \frac{\widehat{G}_{km}(t_j)}{\widehat{G}_{km}(t_i)}, & \text{se } l \text{ sofreu evento competitivo em } t_i < t_j; \\ 0, & \text{quando o indivíduo é retirado do estudo.} \end{cases} \tag{3.5}$$

A definição de peso igual a zero significa que o indivíduo l será retirado da base de dados a partir do momento em que ele sofre o evento de interesse ou é censurado. Essa definição de ponderação descrescente baseou-se na curva observada de sobrevivência das censuras, em que considera-se censura tudo aquilo que não for evento. Então, no cálculo de $G(.)$ censura vira evento e evento (seja ele qual for), vira censura. A função de sobrevivência $\widehat{G}_{km}(.)$ é estimada de forma não paramétrica via Kaplan-Meier como em (3.1) (CARVALHO et al., 2019). Em resumo, a permanência dos estudantes que sofreram o evento competitivo no grupo de risco é tanto menor quanto mais distante do evento de interesse, seguindo a própria estimativa que decai com o tempo dada pela estimativa de Kaplan-Meier das censuras.

Agora, visto que cada estudante recebe um peso conforme descrito em (3.5), consideremos que a contribuição individual de cada um deles para o tempo de sobrevivência t_i é denominada de verossimilhança individual $\mathcal{L}_i$. A $\mathcal{L}_i$ será dada pela razão entre o risco do estudante i experimentar um evento k no tempo t_j, e a soma dos riscos de ocorrência do evento para todos os discentes que estão sob o risco no mesmo período, considerando que seus pesos mudam conforme em (3.5). Isto é:

$$\mathcal{L}_i = \frac{\lambda_k(t, \mathbf{Z}_i)}{\sum\limits_{l \in R(t_j)} w_l(t_j)\lambda_k(t, \mathbf{Z}_l)}. \tag{3.6}$$

Substituindo a expressão (3.4) em (3.6), e sabendo-se que para a verossimilhança em questão apenas é considerada a informação dos estudantes sob risco (desconsiderando a expressão do risco basal), temos então que:

$$\mathcal{L}_i = \frac{\exp(\mathbf{Z}_i^T \boldsymbol{\beta})}{\sum\limits_{l \in R(t_j)} w_l(t_j) \exp(\mathbf{Z}_l^T \boldsymbol{\beta})}. \tag{3.7}$$

A verossimilhança parcial $\mathcal{L}(\boldsymbol{\beta})$ é o produto das verossimilhanças individuais expressas em (3.7), e, portanto, será dada, como Pintilie (2006), por:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{\exp(\mathbf{Z}_i^T \boldsymbol{\beta})}{\sum\limits_{l \in R(t_j)} w_l(t_j) \exp(\mathbf{Z}_l^T \boldsymbol{\beta})}$$

.

No `R`, os pacote disponíveis para o ajuste do modelo de riscos competitivos de Fine e Gray (1999) são os `mstate` e `cmprsk`. Através da função `crprep` da biblioteca `mstate` é possível calcular os pesos como descrito na equação (3.5). Além disso, essa função calcula os pesos considerando cada possível evento como de interesse, sendo os demais considerados competitivos. Portanto, a nova base de dados fica com um grande número de linhas, em que cada estudante terá tantas linhas quantas as mudanças de peso. Após os dados estarem preparados da forma adequada, estima-se os coeficientes dos modelos utilizando as subdistribuições do risco com a função `coxph` disponível na biblioteca `survival`. Maiores detalhes podem ser consultados no Apêndice B, com os modelos e pacotes utilizados.

Em síntese, para cada evento competitivo teremos um modelo com as estimativas dos parâmetros associados às covariáveis. O objetivo deste trabalho consiste na verificação do tempo de permanência do estudante na UFBA. Além do modelo para o evento formar, também é trazido o modelo para os demais eventos competitivos (evadir, ficar retido e trancar a matrícula).

3.4.3.3 Subdistribuição dos riscos para tempos discretos

O modelo amplamente utilizado de Fine e Gray (1999) para a análise de tempos contínuos de eventos não se aplica quando os tempos desses eventos são medidos em uma escala de tempo discreta, que pode ser um cenário provável de ocorrer, como figura-se a análise deste trabalho. Aqui, temos que o tempo medido é em semestres, e, portanto, temos uma escala discreta e não contínua. Visando adaptar a abordagem de Fine e Gray (1999) para essas situações, Berger et al. (2018) propuseram uma técnica para modelagem da subdistribuição dos riscos com tempos discretos com dados censurados à direita. Conforme definem os autores, o método proposto também resulta em estimadores consistentes e assintoticamente normais dos parâmetros do modelo, e é baseado em uma estimativa de máxima verossimilhança ponderada utilizando uma regressão binária (BERGER et al., 2018).

Seja T_i o tempo do evento e C_i o tempo de censura do indivíduo i, $i = 1, ..., n$. Ambos T_i e C_i são assumidos como sendo variáveis aleatórias independentes tomando valores discretos em $\{1, 2, ..., q\}$, em que q é um número natural. Para dados censurados à direita, o período de tempo durante o qual o indivíduo está sob observação é denotado por $\tilde{T}_i = \min(T_i, C_i)$, isto é, $\tilde{T}_i$ corresponde ao tempo do evento se $T_i \leq C_i$ e tempo de censura caso contrário. Também, tem-se que a variável aleatória $\Delta_i := I(T_i \leq C_i)$ indica

se $\tilde{T}_i$ é censurado à direita ($\Delta_i = 0$) ou não ($\Delta_i = 1$) (BERGER et al., 2018).

Supõe-se que existem k eventos competindo e que o tipo de evento do i-ésimo indivíduo (estudante) em T_i é denotado por $\epsilon_i \in \{1, ..., k\}$. Para um conjunto de variáveis preditoras constante no tempo $\boldsymbol{Z}_i = (Z_{i1}, ..., Z_{ip})^T$, o objetivo é estimar a função de incidência acumulada para um determinado evento k. Supondo o evento de interesse do tipo 1, temos que essa FIA será dada por:

$$F_1(t|\boldsymbol{Z}_i) = P(T_i \leq t, \epsilon_i = 1|\boldsymbol{Z}_i). \tag{3.8}$$

Tanto no estudo publicado por Fine e Gray (1999) quanto em Berger et al. (2018), ambos trataram as suas definições com foco na modelagem da ocorrência de um evento do tipo 1 ($\epsilon_i = 1$). O presente trabalho tem enfoque no evento de interesse formar (aqui, chamamos também de evento 1). De acordo com Fine e Gray (1999), é possível vincular a função de incidência acumulada como em (3.8) a um tempo de subdistribuição ϑ_i, que mede o tempo até a ocorrência do evento do tipo k. Na presença de riscos competitivos, ϑ_i precisa levar em conta a possível ocorrência de um evento diferente de 1, por exemplo.

No caso de Fine e Gray (1999), com a definição do evento de interesse do tipo 1 em suas notações matemáticas, a suposição básica feita pelos autores é de que o evento do tipo 1 nunca será o primeiro a ser observado, uma vez que outro evento competitivo tenha ocorrido. Isso implica que não há tempo de evento finito para a ocorrência de um evento tipo 1 se $\epsilon_i \neq 1$. Desse modo, o tempo de subdistribuição discreta para o evento do tipo 1 é dada por:

$$\vartheta_i = \begin{cases} T_i, & \text{se } \epsilon_i = 1; \\ \infty, & \text{se } \epsilon_i \neq 1. \end{cases} \tag{3.9}$$

Análogo ao que foi proposto por Fine e Gray (1999), Berger et al. (2018) definiram a função de risco de subdistribuição discreta para o evento de interesse do tipo 1 como sendo:

$$\lambda_1(t|\boldsymbol{Z}_i) = P(T_i = t, \epsilon_i = 1|(T_i \geq t) \cup (T_i \leq t-1, \epsilon_i \neq 1, \boldsymbol{Z}_i) = P(\vartheta_i = t|\vartheta_i \geq t, \boldsymbol{Z}_i), \tag{3.10}$$

com $t = 1, ..., q$. A expressão em (3.10) é então a função de risco discreta no tempo do evento ϑ_i, que foi definido na equação (3.9).

Agora, para modelar o risco de subdistribuição discreta para um evento do tipo 1,

consideremos a classe de modelos de regressão:

$$\lambda_1(t|\boldsymbol{Z}_i) = h(\gamma_{0t} + \boldsymbol{Z}_i^T\boldsymbol{\gamma}), \tag{3.11}$$

em que h(.) é uma função de distribuição crescente estritamente monótona; o preditor linear $\eta_{it} = \gamma_{0t} + \boldsymbol{Z}_i^T\boldsymbol{\gamma}$ contém os interceptos dependentes do valor real γ_{0t}, $t = 1, ..., q-1$ (referidos como os coeficientes basais) e $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)$ um vetor contendo os coeficentes de regressão independentes de t (BERGER et al., 2018).

Para estimar os coeficientes basais e os coeficientes de regressão do modelo (3.11), Berger et al. (2018) propõem uma metodologia de estimação de máxima verossimilhança que é baseada em uma ponderação discreta. Esse método dos autores está enraizado na modelagem clássica de riscos discretos sem a presença de riscos competitivos, proposta anteriormente por Tutz e Schmid (2016). Para entender a estimação feita na presença de riscos competitivos, vamos definir brevemente o procedimento utilizado para estimar os coeficientes de (3.11) quando não há a presença de eventos competitivos, proposta por Tutz e Schmid (2016).

Na situação mais simples em que o indivíduo experimentou um evento do tipo 1 ou um evento de censura, o risco de subdistribuição discreta se reduz a função de risco $\lambda_1(t|\boldsymbol{Z}_i) = P(T_i = t|T_i \geq t, \boldsymbol{Z_i})$. O tempo de subdistribuição é igual a $\vartheta_i = T_i$ neste caso, e a verossimilhança por indivíduo é dada por:

$$L_i = \lambda_1(\tilde{T}_i|\boldsymbol{Z}_i)^{\Delta_i}(1 - \lambda_1(\tilde{T}_i|\boldsymbol{Z}_i))^{1-\Delta_i} \prod_{t=1}^{\tilde{T}_i-1}(1 - \lambda_1(t|\boldsymbol{Z}_i)) \tag{3.12}$$

e Tutz e Schmid (2016) trazem maiores detalhes sobre a expressão (3.12).

As estimativas dos parâmetros do modelo em (3.12) podem ser obtidas especificando $\lambda_1(t|\boldsymbol{Z}_i) = h(\gamma_{0t} + \boldsymbol{Z}_i^T\boldsymbol{\gamma})$ e maximizando o logaritmo da verossimilhança:

$$\ell(\gamma_{01}, ..., \gamma_{0,q-1}, \boldsymbol{\gamma}^T) = \sum_{i=1}^{n} \log(L_i(\gamma_{01}, ..., \gamma_{0,q-1}, \boldsymbol{\gamma}^T))$$

$$= \sum_{i=1}^{n} \log\left[h(\gamma_{0\tilde{T}_i} + \boldsymbol{Z}_i^T\boldsymbol{\gamma})^{\Delta_i}(1 - h(\gamma_{0\tilde{T}_i} + \boldsymbol{Z}_i^T\boldsymbol{\gamma}))^{1-\Delta_i} \prod_{t=1}^{\tilde{T}_i-1}(1 - h(\gamma_{0t} + \boldsymbol{Z}_i^T\boldsymbol{\gamma}))\right]. \tag{3.13}$$

A estimativa dos parâmetros da equação (3.13) consegue ser simplificada pelo fato de que L_i é equivalente a verossimilhança de um modelo de resposta binária com valores $y_{it} \in \{0, 1\}$, $t = 1, ..., q-1$. Os últimos valores indicam se o indivíduo i experimentou o evento do tipo 1 no tempo t ($y_{it} = 1$) ou não ($y_{it} = 0$). Além disso, y_{it} só é definido se

$t \leq \tilde{T}_i$, isto é, enquanto o indivíduo i estiver sob o risco (TUTZ; SCHMID, 2016). Com isso, obtém-se

$$L_i = \prod_{t=1}^{\tilde{T}_i} \lambda_1(t|\boldsymbol{Z}_i)^{y_{it}}(1 - \lambda_1(t|\boldsymbol{Z}_i))^{1-y_{it}}, \tag{3.14}$$

em que $(y_{i1}, ..., y_{i\tilde{T}_i}) = (0, ..., 0, 1)$ se $\Delta_i = 1$ e $(y_{i1}, ..., y_{i\tilde{T}_i}) = (0, ..., 0, 0)$ se $\Delta_i = 0$.

A verossimilhança binomial em (3.14) implica que modelos de regressão binária podem ser usados via *software* padrão de análise para ajustar o modelo de risco discreto expresso em (3.11) (BERGER et al., 2018). Além disso, uma expressão alternativa para L_i em (3.14) (e que será útil ao modelar o risco de subdistribuição discreta na presença de riscos competitivos, que será definida posteriormente nessa subseção) pode ser escrita como:

$$L_i = \prod_{t=1}^{q-1} \left\{\lambda_1(t|\boldsymbol{Z}_i)^{y_{it}}(1 - \lambda_1(t|\boldsymbol{Z}_i))^{1-y_{it}}\right\}^{W_{it}}, \tag{3.15}$$

em que $W_{it} = I(t \leq \tilde{T}_i)$, $i = 1, ..., n$, $t = 1, ..., q-1$ é o conjunto de pesos que indica se o indivíduo i está em risco no tempo t ou não (BERGER et al., 2018).

Os valores binários correspondentes a y_{it} usados na equação (3.15) são definidos por:

$$(y_{i1}, ..., y_{i,\tilde{T}_i}, ..., y_{i,q-1}) = \begin{cases} (0, ..., 0, 1, 0, ..., 0), & \text{se } \Delta_i = 1; \\ (0, ..., 0, 0, 0, ..., 0), & \text{se } \Delta_i = 0. \end{cases} \tag{3.16}$$

E, com essa representação, teremos que:

$$\ell = \sum_{i=1}^{n}\sum_{t=1}^{q-1} W_{it}\left\{y_{it}\log(\lambda_1(t|\boldsymbol{Z}_i)) + (1 - y_{it})\log(1 - \lambda_1(t|\boldsymbol{Z}_i))\right\} \tag{3.17}$$

é o logaritimo da verossimilhança total do modelo de risco de subdistribuição discreta sem a presença de riscos competitivos (BERGER et al., 2018). Portanto, estando definido o conceito da estimação da subdistribuição dos riscos para tempos discretos na ausência de eventos competitivos, é necessário definir o processo de estimação na presença destes eventos.

Para modelar o risco de subdistribuição na presença de riscos competitivos do tipo $2, ..., k$ (com λ_1 agora denotando o risco de subdistribuição para o evento do tipo 1), o logaritmo da função de verossimilhança será especificado de maneira semelhante a (3.17). No entanto, os pesos W_{it} precisam ser redefinidos adequadamente para esse novo contexto. Para isso, considera-se o conjunto sob risco no tempo t dado por $r(t)$, definido

como o conjunto de estudantes que não experimentaram um evento do tipo 1 nem um evento de censura anterior ao tempo t. Se o i-ésimo estudante é conhecido por estar nesse conjunto, a ideia é definir $W_{it} = 1$, como antes. Portanto, define-se $W_{it} = 0$ se o discente não for membro de $r(t)$. Quando um estudante, por exemplo, experimenta um evento competitivo antes de t, $r(t)$ não é totalmente conhecido, o que torna-se um problema. Como $\vartheta_i = \infty$ para esses casos, os discentes ainda continuam sob risco além de $\tilde{T}_i$ até que eles eventualmente experimentem o evento de censura (BERGER et al., 2018).

Consequentemente, como os tempos de censura C_i não são observados se $C_i > \tilde{T}_i$, não se pode determinar se um indivíduo com $\epsilon_i > 1$ ainda faz parte do risco estabelecido em $t > \tilde{T}_i$. Então, de acordo com a abordagem de tempo contínuo de Fine e Gray (1999), os autores Berger et al. (2018) propõem estimar a probabilidade de cada indivíduo fazer parte do conjunto de risco $r(t)$, e definir os pesos em (3.17) igual às probabilidades estimadas.

Mais especificamente, Berger et al. (2018) propõem os pesos da seguinte maneira:

i) Para indivíduos não censurados que vivenciam um evento do tipo 1 ($\Delta_i\epsilon_i = 1$), definem que $W_{it} := I(t \leq \tilde{T}_i)$. Essa definição implica que:

$$(W_{i1}, W_{i2}, ..., W_{i\tilde{T}_i}, W_{i,\tilde{T}_{i+1}}, ..., W_{i,(q-1)}) = (1, 1, ..., 1, 0, ..., 0),$$

contabilizando o fato de que os indivíduos deixam de estar em risco após seus respectivos eventos do tipo 1 terem sido observados.

ii) Para indivíduos que experimentam primeiro o evento de censura ($\Delta_i\epsilon_i = 0$), também propõem definir $W_{it} := I(t \leq \tilde{T}_i)$. Essa definição leva em conta o fato de que os indivíduos deixam de estar em risco após $\tilde{T}_i$.

iii) Para os indivíduos sem censura que experimentam primeiro um evento competitivo ($\Delta_i\epsilon_i > 1$), os autores propõe definir $W_{it} := 1$ se $t \leq \tilde{T}_i$, levando em conta o fato de que os indivíduos estão em risco pelo menos até $\tilde{T}_i$. Para $t > \tilde{T}_i$, estima-se a probabilidade de pertencer ao conjunto de risco $r(t)$ por:

$$W_{it} := \frac{\widehat{G}(t-1)}{\widehat{G}(\tilde{T}_i - 1)}, \tilde{T}_i < t \leq q-1,$$

em que $\widehat{G}(t)$ é uma estimativa da função de sobrevivência de censura $G(t) = P(C_i > t)$.

Temos que, ao combinar (i) a (iii), os pesos W_{it} podem ser expressos na forma fechada:

$$W_{it} = I(C_i \geq \min(T_i, t)) \frac{\widehat{G}(t-1)}{\widehat{G}\left(\min(T_i, C_i, t) - 1\right)} \Big(I(t \leq T_i) + I(T_i \leq t-1, \epsilon_i \neq 1) \Big)$$

$$= \frac{\widehat{G}(t-1)}{\widehat{G}(\min(\tilde{T}_i, t)-1)}\Big(I(t \leq \tilde{T}_i) + I(\tilde{T}_i \leq t-1, \Delta_i\epsilon_i > 1)\Big).$$

Assim como em (3.16), os valores binários y_{it} também são definidos por Berger et al. (2018):

$$(y_{i1}, ..., y_{i,\tilde{T}_i}, ..., y_{i,q-1}) = \begin{cases} (0, ..., 0, 1, 0, ..., 0), & \text{se } \Delta_i\epsilon_i = 1; \\ (0, ..., 0, 0, 0, ..., 0), & \text{se } \Delta_i\epsilon_i \neq 1. \end{cases} \tag{3.18}$$

Dado essas premissas e definições, Berger et al. (2018) comprovam o principal resultado do seu trabalho, o Teorema 1, o qual diz que a solução para o problema de otimização:

$$\underset{\boldsymbol{\gamma}_0, \boldsymbol{\gamma}}{\arg\max}\, \ell(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}) =$$

$$\underset{\boldsymbol{\gamma}_0, \boldsymbol{\gamma}}{\arg\max} \left\{ \sum_{i=1}^{n} \sum_{t=1}^{q-1} W_{it} \{ y_{it}\ \log(h(\gamma_{0t} + \boldsymbol{Z}_i^T\boldsymbol{\gamma})) + (1 - y_{it})\ \log(1 - h(\gamma_{0t} + \boldsymbol{Z}_i^T\boldsymbol{\gamma})) \} \right\} \tag{3.19}$$

define um estimador consistente e assintoticamente normal ($n \to \infty$) dos parâmetros $\boldsymbol{\gamma}_0$ e $\boldsymbol{\gamma}$ do modelo de risco de subdistribuição em (3.11). Os autores demonstram o Teorema 1, e concluem que esse Teorema implica que os parâmetros do modelo do risco de subdistribuição discreta em (3.11) pode ser estimado de forma consistente ajustando um modelo de regressão binária ponderado com valores de resultado y_{it} e pesos W_{it}.

Computacionalmente, o procedimento consiste em, primeiramente configurar uma matriz de dados, que será composta por um conjunto de matrizes de dados aumentadas definidas separadamente para cada indivíduo (estudante) (BERGER et al., 2018). Mais especificamente, para estudantes não censurados que experimentaram o evento do tipo 1 ($\Delta_i\epsilon_i = 1$) a matriz de dados aumentada e o vetor de pesos são definidos, respectivamente, por:

$$\begin{pmatrix} \boldsymbol{y}_i & \boldsymbol{t} & & \boldsymbol{Z}_i & \\ 0 & 1 & z_{i1} & \dots & z_{ip} \\ 0 & 2 & z_{i1} & \dots & z_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \tilde{T}_i & z_{i1} & \dots & z_{ip} \\ 0 & \tilde{T}_{i+1} & z_{i1} & \dots & z_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & k-1 & z_{i1} & \dots & z_{ip} \end{pmatrix} \quad \mathbf{e} \quad \begin{pmatrix} \boldsymbol{W}_i \\ 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{3.20}$$

em que $\boldsymbol{t}$ é uma variável preditora adicional que se refere aos coeficientes basais.

A matriz de dados aumentada e o vetor de pesos para indivíduos censurados ($\Delta_i\epsilon_i = 0$) são definidas da mesma forma que em (3.20), exceto que $y_i := (0, ..., 0)^T$ nesses casos (BERGER et al., 2018).

Para indivíduos sem censura que experimentam um evento competitivo primeiro ($\Delta_i\epsilon_i > 1$), define-se:

$$\begin{pmatrix} \boldsymbol{y}_i & \boldsymbol{t} & & \boldsymbol{Z}_i & \\ 0 & 1 & z_{i1} & \dots & z_{ip} \\ 0 & 2 & z_{i1} & \dots & z_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \tilde{T}_i & z_{i1} & \dots & z_{ip} \\ 0 & \tilde{T}_{i+1} & z_{i1} & \dots & z_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & k-1 & z_{i1} & \dots & z_{ip} \end{pmatrix} \quad \textbf{e} \quad \begin{pmatrix} \boldsymbol{W}_i \\ 1 \\ 1 \\ \vdots \\ 1 \\ \frac{\widehat{G}(\tilde{T}_i)}{\widehat{G}(\tilde{T}_i - 1)} \\ \vdots \\ \frac{\widehat{G}(k-2)}{\widehat{G}(\tilde{T}_i - 1)} \end{pmatrix}.$$

A matriz de dados aumentada completa é obtida concatenando-se as matrizes de dados aumentadas individuais. A matriz resultante $n(k-1) \times (p+2)$ e o vetor de pesos de comprimento $n(k-1)$ são posteriormente passados para a função de ajuste de uma regressão binária, a fim de resolver o problema de otimização em (3.19).

No `R`, para gerar a matriz de dados aumentada, aplica-se a função disponível chamada `dataLongSubDist()` do pacote `discSurv`. Em seguida, as estimativas dos parâmetros podem ser obtidas usando a função `glm`. Esse processo é feito separadamente para cada evento que se tenha interesse, e, portanto, serão 4 modelos ajustados, para 4 matrizes de dados do tipo `dataLongSubDist()`, com referência ao evento de interesse no momento do ajuste. Detalhes com o modelo ajustado e pacotes utilizados podem ser consultados no Apêndice B.

Ressalta-se que a estrutura do modelo proposto por Berger et al. (2018), e aplicado na análise deste trabalho, também permite que a modelagem seja realizada mesmo quando os riscos de subdistribuição não são proporcionais (como ocorreu nos resultados deste trabalho, conforme será visto na Seção 4).

3.4.3.4 Verificação dos pressupostos do modelo

Nessa seção, descreve-se alguns métodos utilizados para verificar o ajuste do modelo. Os aspectos a serem investigados serão a proporcionalidade global dos riscos, a proporcionalidade de cada variável considerada no modelo e a presença de pontos aberrantes.

Para isso, serão utilizados: teste de proporcionalidade global, resíduo de Schoenfeld, resíduo Martingale e resíduo *Deviance*.

Os resíduos apresentados aqui são os mesmos utilizados na abordagem dos modelos de Cox, com a ressalva de que o grupo sob risco está contribuindo de forma ponderada como expresso em (3.5). Além disso, foram analisados para o ajuste feito com o modelo de Fine e Gray (1999), para o desfecho formar, considerando os tempos, em semestres (escala discreta).

Para responder a pergunta se o risco relativo de um estudante experimentar um evento competitivo, dado uma covariável, é sempre o mesmo durante todo o tempo de observação, utiliza-se os resíduos de Schoenfeld, definidos para cada estudante i e covariável p. Caso o risco não seja o mesmo durante todo o tempo observado, o efeito da covariável é tempo-dependente, e, portanto, a pressuposição de riscos proporcionais não estará satisfeita. Esses resíduos são definidos por:

$$r_{ip} = \delta_i(z_{ip} - a_{ip}),$$

em que δ_i é o indicador de ocorrência do evento de interesse no discente i, e por isso quando ocorre censura (estudante que está com *status* cursando), o resíduo é nulo. Define-se a_{ip} como uma média ponderada dos valores das covariáveis dos estudantes em risco no tempo t_i:

$$a_{ip} = \frac{\sum\limits_{j \in R(t_i)} z_{jp} \exp(\boldsymbol{z}_j \widehat{\boldsymbol{\beta}})}{\sum\limits_{j \in R(t_i)} \exp(\boldsymbol{z}_j \widehat{\boldsymbol{\beta}})},$$

em que $R(t_i)$ é o conjunto de discentes em risco no tempo t_i e z_{jp} representa o valor da covariável p do estudante j pertecente ao grupo de risco (CARVALHO et al., 2019). A única diferença é que internamente, o cálculo desse resíduo estará considerando o modelo ajustado pela função `coxph` após a preparação dos dados com os devidos pesos, como foi descrito na Seção 3.4.3.2.

Cabe ressaltar que haverá tantos vetores de resíduos quanto covariáveis ajustadas no modelo, e que esses são definidos somente nos tempos t_j em que ocorreu um evento. Além disso, é possível também definir resíduos padronizados de Schoenfeld. Consideremos que o vetor de coeficientes $\boldsymbol{\beta}$ varia com o tempo t. Sendo assim, o vetor $\boldsymbol{\beta}$ pode ser dividido em uma parte considerada uma média constante e a outra parte uma função $\boldsymbol{U}(t)$ que representa valores que variam no tempo. Desse modo, o resíduo padronizado de Schoenfeld em t_i é dado por:

$$\boldsymbol{r_i^*} = [\widehat{V}(\boldsymbol{r_i})]^{-1} \boldsymbol{r_i},$$

em que $\widehat{V}(\boldsymbol{r_i})$ é a matriz de covariância estimada de vetor de resíduos de Schoenfeld (CARVALHO et al., 2019). Ainda, demonstra-se que o valor esperado desse resíduo padronizado é aproximadamente igual à parte de $\boldsymbol{\beta}$ que varia no tempo - a função $\boldsymbol{U}(t)$. Portanto, o gráfico dos resíduos padronizados de Schoenfeld contra os tempos de sobrevivência permite verificar se estes estão distribuídos igualmente ao longo do tempo, ou se existe uma forma sugestiva de não proporcionalidade. Nesse caso, se a suposição de riscos proporcionais for satisfeita, não deverá existir tendência sistemática no gráfico de r^*_{ik} contra o tempo de sobrevivência (CARVALHO et al., 2019).

Além da análise gráfica, pode-se testar a existência de correlação linear entre o tempo de sobrevivência e o resíduo Schoenfeld. Sob a hipótese nula, de correlação igual a zero, teremos que a distribbuição do teste é uma qui-quadrado com um grau de liberdade. Portanto, se a hipótese nula não é rejeitada, não se rejeita a premissa de proporcionalidade dos riscos. O teste realizado para cada covariável é baseado em uma regressão dada da seguinte maneira:

$$\beta_p(t) = \beta_p + \theta_p U_p(t),$$

em que θ_k é o parâmetro de variação no tempo, e a hipótese nula é de que $\theta_k = 0$ (CARVALHO et al., 2019).

Outro resíduo analisado diz respeito aos resíduos Martingale, que possibilitam a identificação de pontos aberrantes (estudantes que demoram muito tempo para sofrer o evento ou que sofrem o evento muito rapidamente, dadas as covariáveis). Denominados de M_i, os resíduos martingale são baseados no processo de contagem individual e são definidos por:

$$M_i = N_i - E_i,$$

tal que N_i é igual ao número de eventos observados no intervalo $[0, \infty]$ e E_i é o número de eventos esperados sob o modelo ajustado no intervalo $[0, \infty]$. Estima-se o resíduo martingale como:

$$\widehat{M}_i = \delta_i - \widehat{\Lambda}_0(t_i) \exp(\boldsymbol{x}_i \widehat{\boldsymbol{\beta}}) = \delta_i - r_{C_i},$$

em que $r_{C_i} = \Lambda_0(t_i) \exp(\boldsymbol{x}_i \widehat{\boldsymbol{\beta}})$ é chamado de resíduo de Cox-Snell (CARVALHO et al., 2019), e $\Lambda_0 = \int_0^t \lambda_0(u) d(u)$ é a função de risco acumulado.

Algumas propriedades dos resíduos martingale (algumas semelhantes aos resíduos dos modelos de regressão linear) são:

- o valor esperado de M_i é 0, quando avaliado no valor verdadeiro (e desconhecido) do vetor de parâmetros $\boldsymbol{\beta}$;
- os resíduos M_i não são simetricamente distribuídos em torno de 0, variando de $(-\infty, 1]$ e quando o tempo de sobrevivência é censurado o resíduo é negativo;
- o somatório dos resíduos observados baseados no valor estimado de $\boldsymbol{\beta}$ é igual a 0; e
- os resíduos M_i calculados usando o verdadeiro vetor de parâmetros $\boldsymbol{\beta}$ são não correlacionados, mas as estimativas $\widehat{M}_i$ são negativamente correlacionadas, ainda que fracamente (CARVALHO et al., 2019).

Por fim, assim como os resíduos martingale, os resíduos *deviance* permitem detectar pontos aberrantes (*outliers*). Denominado de D_i, esses resíduos são definidos por:

$$D_i = sinal(\widehat{M}_i)\sqrt{-2 \times (l_{i(modelo)} - l_{i(saturado)})},$$

sendo o sinal de $\widehat{M}_i$ dado pelo sinal do resíduo martingale e $l_{i(modelo)} - l_{i(saturado)}$ os valores do logaritmo da função de verossimilhança para cada observação i do modelo em questão e do modelo saturado. Esses resíduos são distribuído simetricamente em torno do zero, portanto sendo de interpretação gráfica mais simples. Além disso, pontos aberrantes podem ser identificados quando não estão dentro do intervalo $[-2, 2]$ dos resíduos *deviance*.

A rotina com todo o *script* utilizado nesse tópico pode ser consultada no Apêndice B.

4 RESULTADOS

Nesta seção apresenta-se os principais resultados obtidos com as análises dos microdados do CENSUP 2019 para os alunos dos cursos de Estatística e Matemática da UFBA. Visando alcançar um dos objetivos específicos deste trabalho, todos os achados deste trabalho encontram-se disponibilizados em endereço *web* através do aplicativo R *Shiny*: <https://jessica-fagundesg.shinyapps.io/tempodepermanenciaimeufba/>. Estes resultados preliminares também encontram-se disponíveis para consulta no Apêndice C. Construiu-se também um segundo aplicativo *web Shiny* reunindo todos os resultados encontrados no projeto deste trabalho. Estes resultados podem ser consultados em: <https://jessica-fagundesg.shinyapps.io/tempodepermanenciaufba/>.

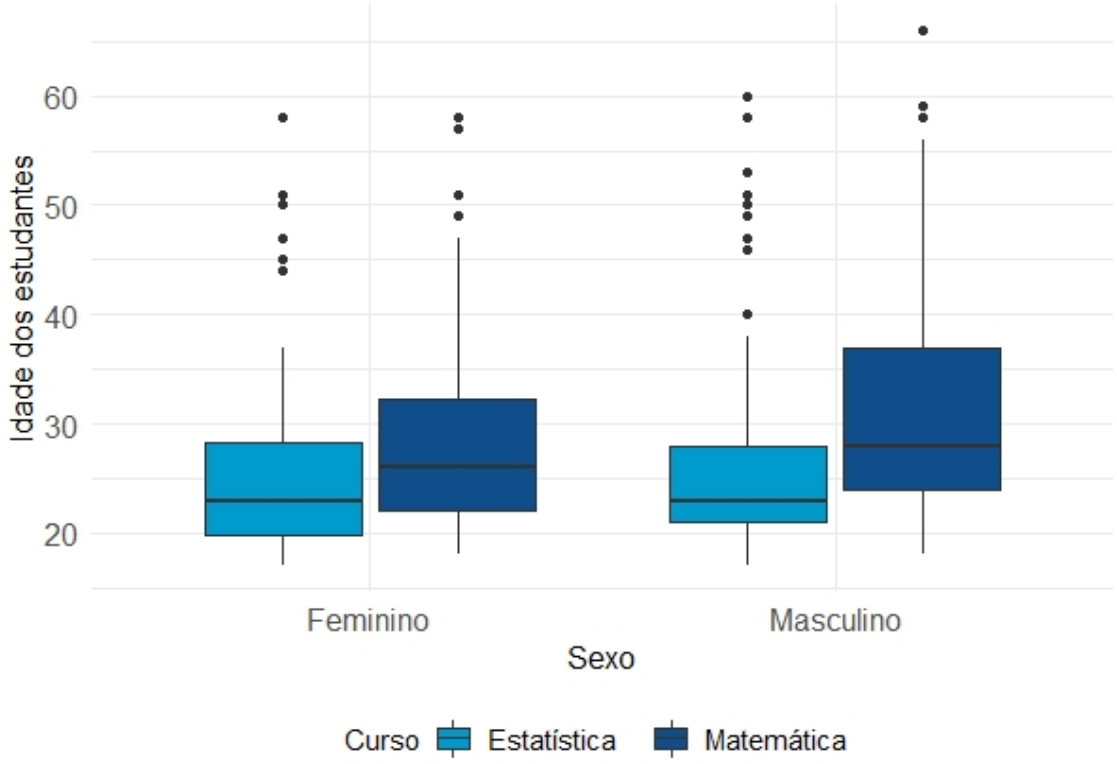
4.1 Análises descritivas

A base de dados final corresponde a informações sobre 434 estudantes dos cursos de Estatística e Matemática da UFBA. Desses estudantes, 72% eram do sexo masculino, e 53% do total de discentes cursavam Matemática. Referente a idade (no ano do censo em 2019) segundo o sexo e o curso, podemos verificar na Figura 2 que a idade mediana dos estudantes do sexo masculino, em ambos os cursos, é maior do que para os estudantes do sexo feminino.

Além disso, a variabilidade das idades entre os sexos parece ser similar, destacando-se os estudantes do sexo masculino no curso de matemática, com uma maior variabilidade entre as idades. Podemos verificar também que, tanto para o sexo quanto para o curso, há presença de pontos atípicos, com estudantes mais velhos em todos os *boxplots* apresentados na Figura 2. A maior idade registrada foi de um estudante de 66 anos do sexo masculino e do curso de matemática, e de 60 anos do sexo masculino e do curso de estatística.

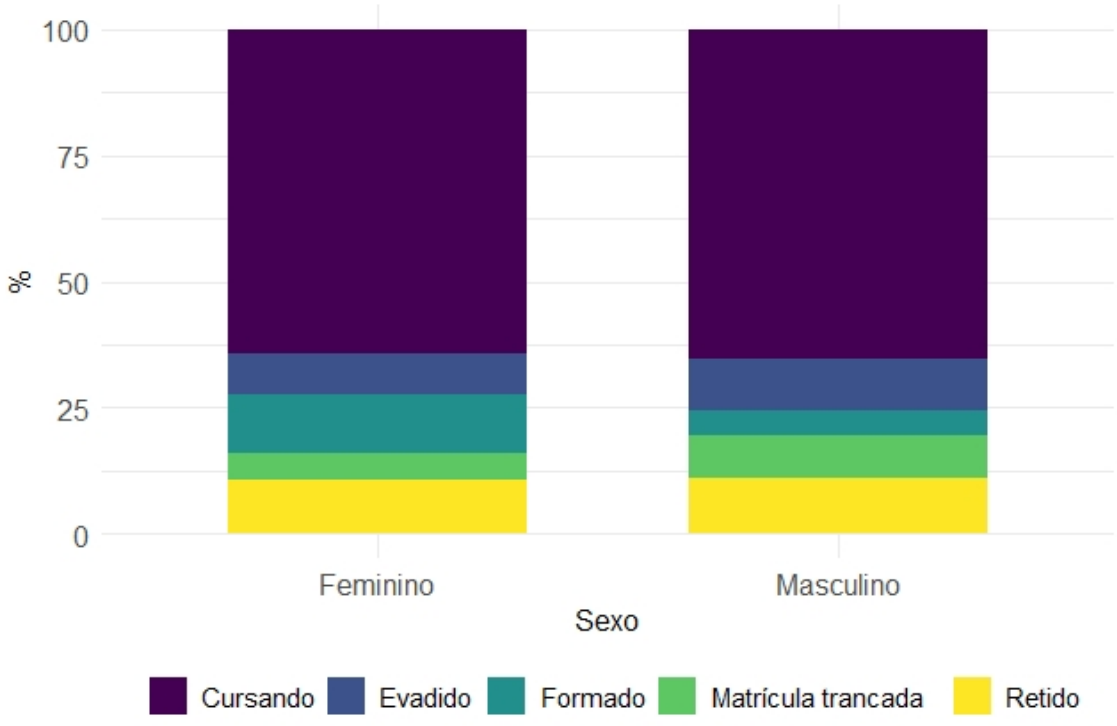
Referente ao tipo de situação de vínculo com a UFBA, tivemos que 65% estavam cursando (tratados como censuras); 11% dos discentes foram classificados como retidos (segundo a definição proposta neste trabalho na Seção 3.2); 10% dos alunos foram definidos como evadidos; 7% haviam formado no ano de 2019 (evento de interesse), e o restante com situação de vínculo de matrícula trancada. O comportamento da situação de vínculo em relação ao sexo dos discentes pode ser consultada na Figura 3. Destaca-se um maior percentual de estudantes do sexo feminino formados naquele ano, e de matrículas trancadas por estudantes do sexo masculino.

Figura 2 – Idade dos estudantes segundo o sexo e o curso - CENSUP 2019 - UFBA.



Fonte: INEP (2019c).

Figura 3 – Situação de vínculo dos estudantes segundo o sexo - CENSUP 2019 - UFBA.



Fonte: INEP (2019c).

Sobre a variável de reserva de vaga e tipo de escola de conclusão de ensino médio, foi necessária a criação de uma variável auxiliar resultante da combinação dessas duas variáveis originais. Constatou-se uma alta associação entre estas variáveis (possivelmente pela existência de reserva de vagas destinadas à estudantes oriundos de escola pública), o que poderia impactar nas estimativas dos coeficientes do modelo ajustado. Desse modo, e para não retirar essas covariáveis das análises, criou-se uma nova variável com os quantitativos por categoria apresentados na Tabela 2. Destaca-se que 50% do total de estudantes não ingressaram na UFBA por reserva de vagas e eram de escola privada, e 43% eram ingressantes por reserva de vagas e advindos de escola pública.

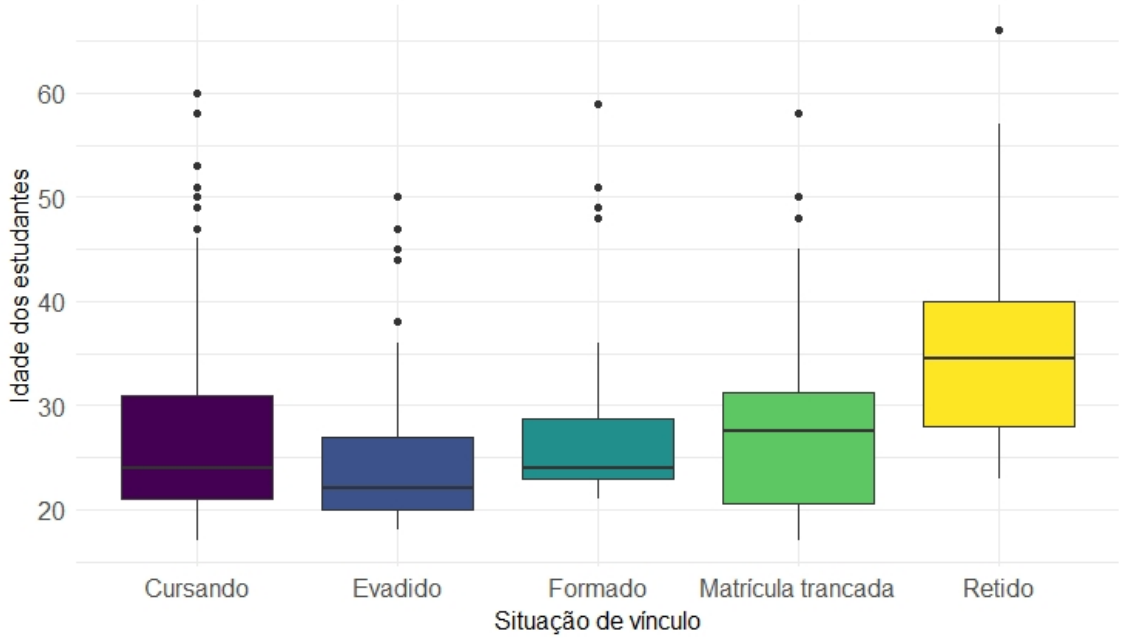
Tabela 2 – Quantidade de estudantes dos cursos de Estatística e Matemática da UFBA segundo reserva de vaga/tipo de escola de conclusão ensino médio - CENSUP 2019.

Reserva/Escola	Quantidade/%
Sem reserva/escola privada	217 (50%)
Com reserva/escola pública	188 (43%)
Sem reserva/escola pública	24 (6%)
Com reserva/escola privada	5 (1%)

Fonte: INEP (2019c).

Também podemos visualizar o comportamento da situação de vínculo dos discentes conforme a idade dos mesmos. Para isso, é trazido o *boxplot* da Figura 4, em que observa-se que estudantes classificados como evadidos possuíam idade um pouco menor, no geral, comparado aos estudantes com os demais tipo de vínculos à UFBA.

Figura 4 – Situação de vínculo dos estudantes segundo a idade - CENSUP 2019 - UFBA.



Fonte: INEP (2019c).

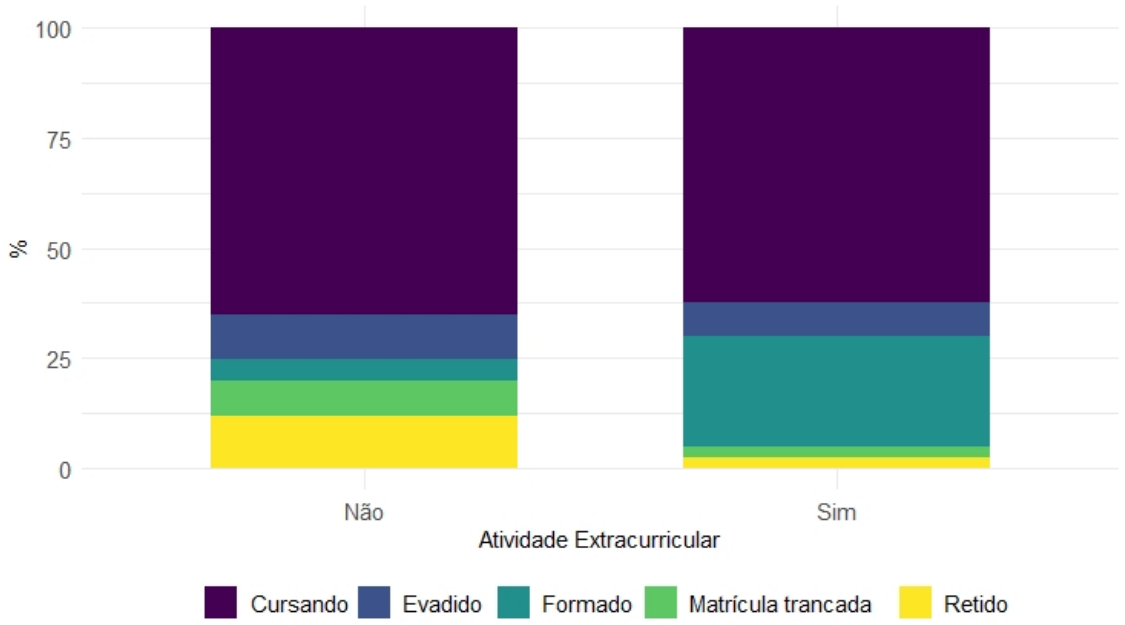
Destaca-se ainda que, entre os estudantes evadidos, o tempo registrado (em semestres), do ingresso dos mesmos até serem classificados como situação de evadido no curso, foi de 2 semestres para 33% do total de discentes nessa categoria. Esse alto percentual corrobora com resultados já encontrados na literatura, como em (FILHO et al., 2007), que diz que no primeiro ano de curso, a taxa de evasão observada chega a ser 2 a 3 vezes maior do que nos anos seguintes.

Verifica-se também na Figura 4 que a idade mediana no grupo de estudantes classificados como retidos mostra-se superior aos demais casos de vínculos. Em média, mesmo com a existência de *outliers*, estudantes mais velhos encontram-se com tempo de curso maior e com integralização de horas mínima abaixo do que o previsto para a sua respectiva matriz curricular.

A outra variável investigada e trazida no ajuste dos modelos refere-se a realização de algum tipo de atividade extracurricular por parte dos estudantes. Constatou-se que, entre os discentes de Estatística e Matemática, apenas 9% realizavam algum tipo de atividade extracurricular como monitoria, pesquisa, extensão ou estágio não obrigatório.

Em relação a situação de vínculo dos discentes segundo a realização de algum tipo de atividade extracurricular, observa-se que, de acordo com a Figura 5, dentre os formandos naquele ano, o maior percentual realizou algum tipo de atividade extracurricular. Além disso, o maior percentual de retenção e trancamento de matrícula concentrou-se entre aqueles que não faziam nenhum tipo de atividade.

Figura 5 – Situação de vínculo dos estudantes segundo a realização de atividade extracurricular - CENSUP 2019 - UFBA.

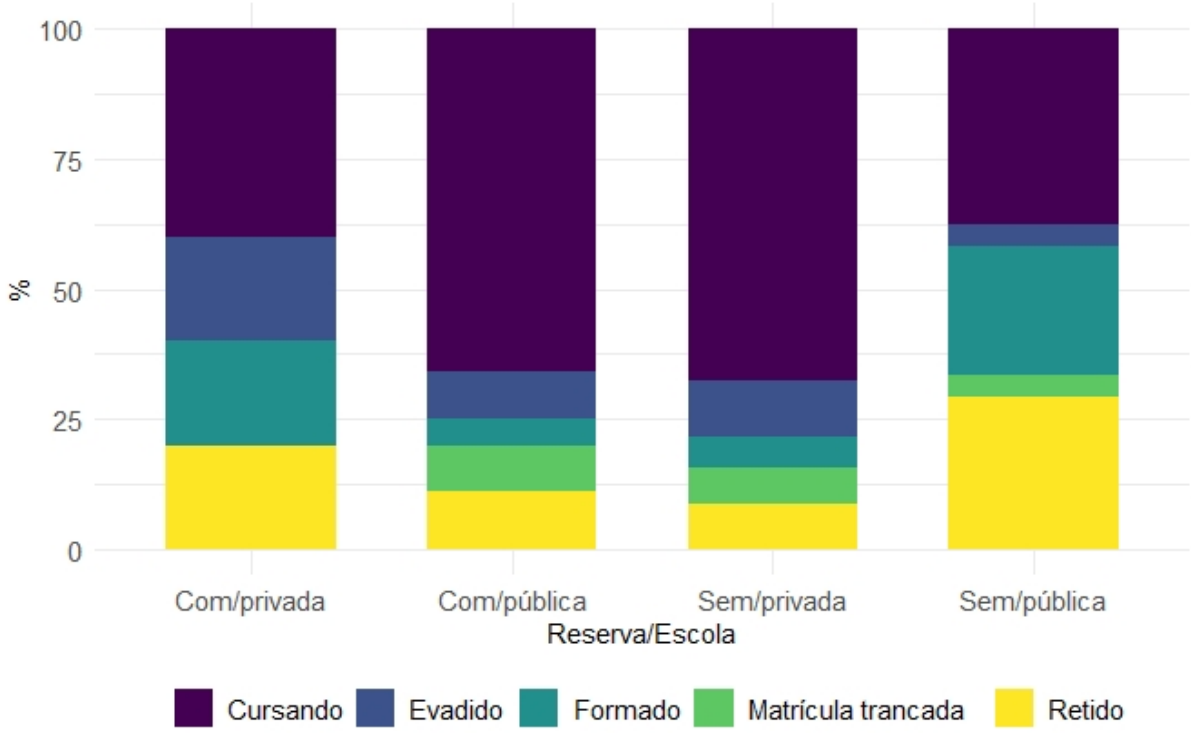


Fonte: INEP (2019c).

Outra análise trazida diz respeito a situação de vínculo segundo a variável de reserva de vagas/tipo de escola de conclusão do ensino médio. Analisando a Figura 6, é possível verificar que houve um maior percentual de estudantes formados que não entraram por reserva de vaga e eram oriundos de escola pública. Entre os estudantes oriundos de escola privada que eram cotistas não houve trancamento de matrícula. No entanto, enfatizando os quantitativos já evidenciados na Tabela 2, os estudantes de escola pública e cotistas, em comparação aos de escola privada e não cotistas, tiveram um comportamento semelhante quanto à situação de vínculo a UFBA.

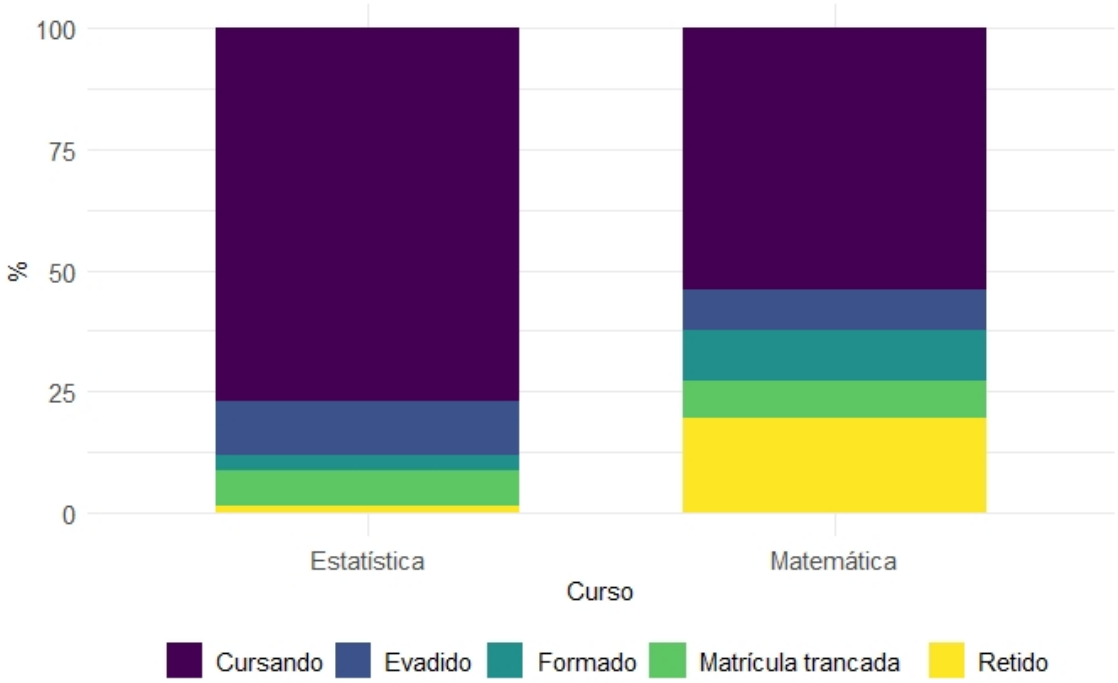
Por fim, na Figura 7, é possível verificar o comportamento da situação de vínculo em relação a cada curso aqui estudado. Destaca-se o maior percentual de formados sendo do curso de Matemática, e alto percentual de estudantes classificados como retidos também neste curso. Além disso, verifica-se maior percentual de evadidos no curso de Estatística.

Figura 6 – Situação de vínculo dos estudantes segundo reserva de vagas/escola de conclusão do ensino médio - CENSUP 2019 - UFBA.



Fonte: INEP (2019c).

Figura 7 – Situação de vínculo dos estudantes segundo o curso - CENSUP 2019 - UFBA.



Fonte: INEP (2019c).

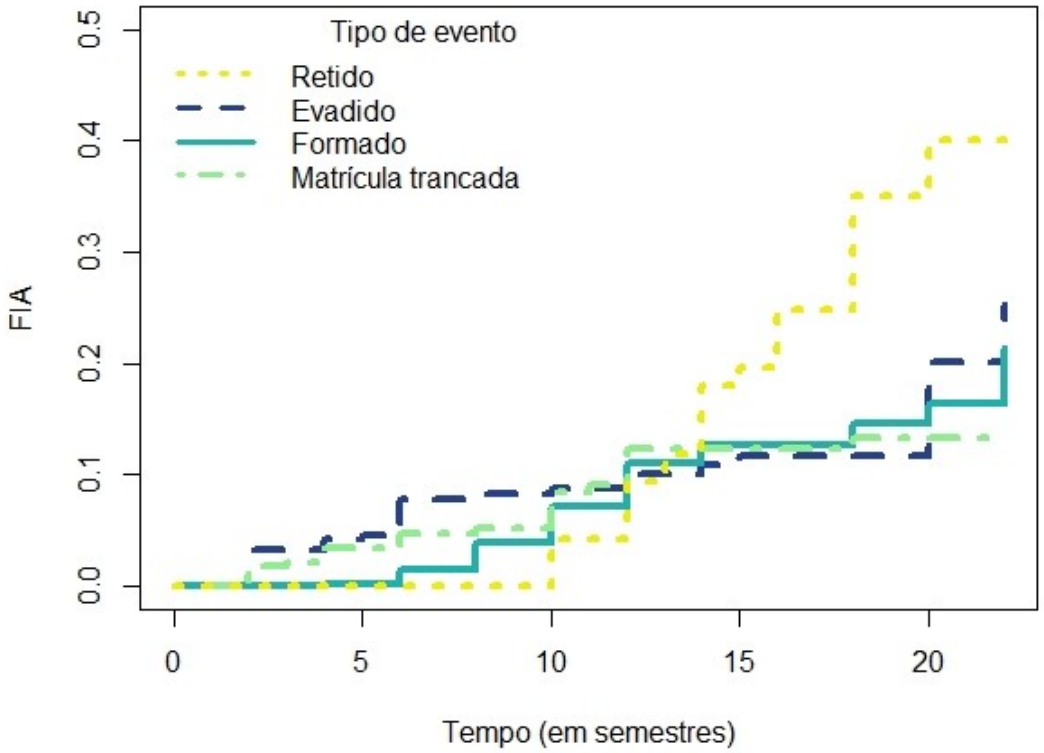
4.2 Modelagem com riscos competitivos

Antes da realização do processo de modelagem da subdistribuição dos riscos, analisou-se os gráficos das FIA's para as covariáveis ajustadas no modelo. Registrou-se que o maior tempo de permanência, até a formatura, foi de 22 semestres para 2 estudantes do curso de Matemática, e de 12 semestres para 2 estudantes do curso de Estatística.

Inicialmente, verificou-se o gráfico da FIA para todos os tipos de eventos competitivos. A Figura 8 indica as curvas de risco para os tipos de desfechos dos estudantes, sem levar em conta nenhuma covariável. A figura sugere que o risco do estudante ficar retido no curso é maior do que para os demais desfechos, próximo ao 14º semestre, e é maior de evadir nos primeiros semestres.

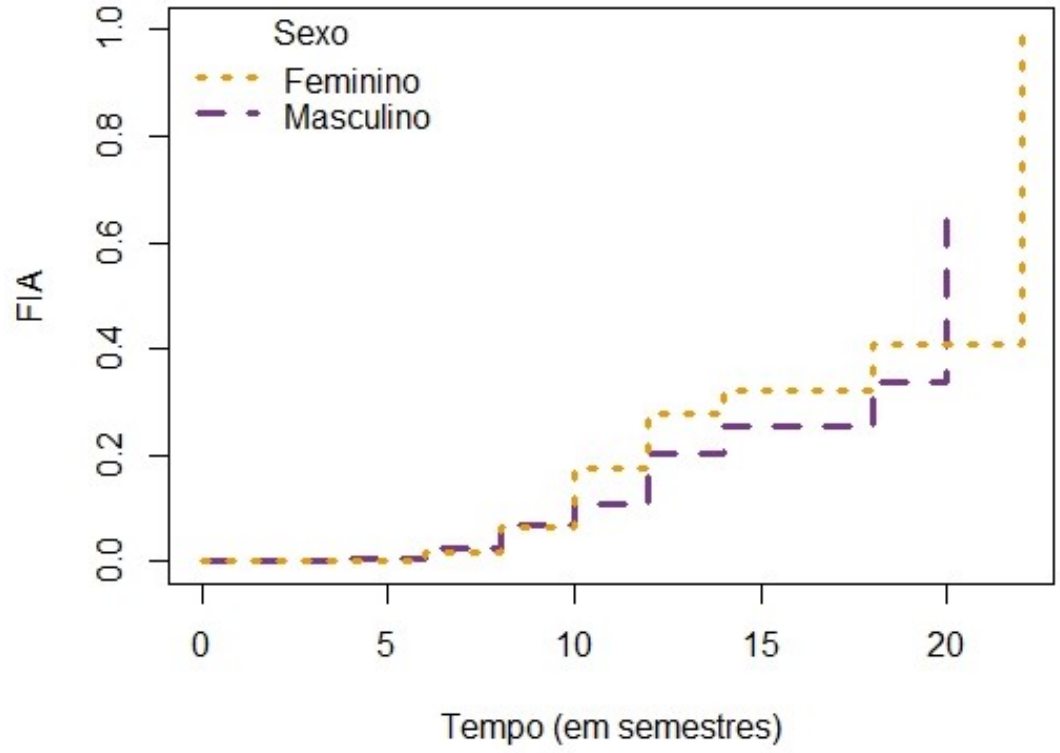
Referente ao evento de interesse deste trabalho, verificou-se a diferença na associação entre as variáveis categóricas e as curvas das FIA's. Podemos observar que há diferença em relação ao sexo dos estudantes para o tempo até a formatura, conforme Figura 9, especialmente a partir do 10º semestre. Ou seja, essa figura sugere que estudantes do sexo feminino possuem risco maior de se formarem (o que é bom), a partir do 10º semestre, do que estudantes do sexo masculino, segundo os dados.

Figura 8 – FIA's para todos os eventos.



Fonte: INEP (2019c).

Figura 9 – FIA's para o evento formar, dado o sexo do estudante.

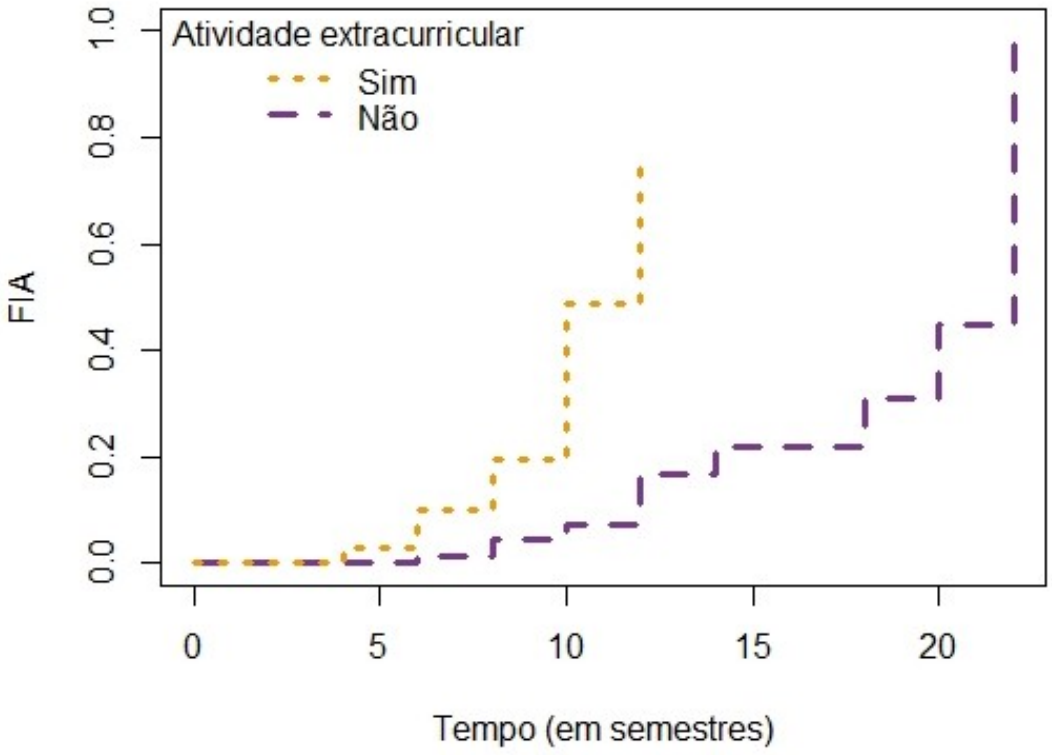


Fonte: INEP (2019c).

A mesma análise foi realizada para a variável de atividade extracurricular entre os estudantes. De acordo com a Figura 10, nota-se que estudantes que realizam algum tipo de atividade extracurricular possuem um risco maior para o evento formar, principalmente a partir do 4º semestre, em comparação aos discentes que não realizam nenhum tipo de atividade.

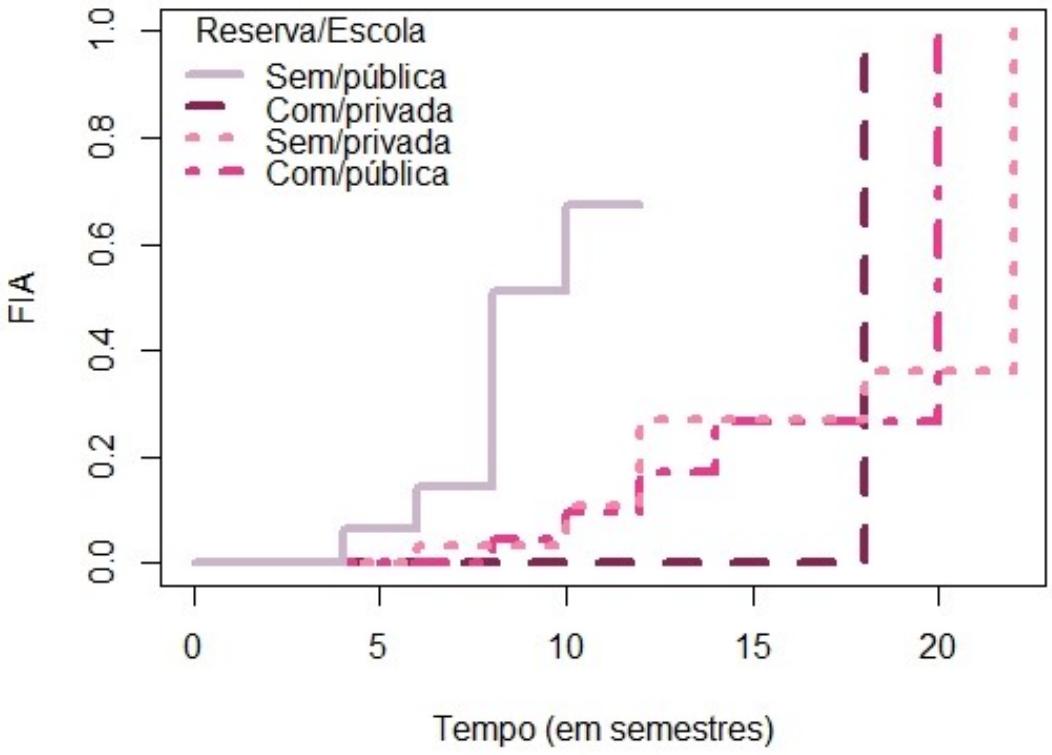
Em relação a ter ingressado na UFBA por reserva de vaga e a respectiva escola de conclusão de ensino médio, as curvas das FIA's para essa covariável são apresentadas na Figura 11. A figura sugere que estudantes não cotistas mas oriundos de escola pública possuem um risco maior em relação a formatura, do que os estudantes classificados nas demais categorias. Novamente, o alto desbalanceamento entre as categorias nessa covariável, conforme verificado na Tabela 2, exige cautela nas interpretações.

Figura 10 – FIA's para evento formar, dado a realização de atividade extracurricular pelo estudante.



Fonte: INEP (2019c).

Figura 11 – FIA's para evento formar, dado reserva de vagas/tipo de escola de conclusão do ensino médio do estudante.



Fonte: INEP (2019c).

4.2.1 Modelo com tempos contínuos *vs modelo com tempos discretos*

Nesta seção, serão apresentados os resultados que foram obtidos através da modelagem da subdistribuição dos riscos para o tempo de permanência dos estudantes de Estatística e Matemática da UFBA. Utilizou-se um nível de significância $\alpha = 0,05$.

Os modelos finais (para cada desfecho, aplicando a metodologia de Fine e Gray (1999) para tempos contínuos e de Berger et al. (2018) para tempos discretos) apresentados aqui consideram as seguintes variáveis: sexo, idade (em anos) em 2019, atividade extracurricular e se o aluno possui é cotista ou não, segundo o tipo de escola de conclusão do ensino médio.

Para o evento de interesse deste trabalho, a Tabela 3 traz os valores da exponencial dos coeficientes para cada covariável (que representam o risco de formar de cada indivíduo), além dos seus respectivos erros-padrão e p-valor indicando a significância estatística dessas covariáveis em relação ao risco do estudante formar. Observou-se que as estimativas obtidas com o modelo para tempos discretos foram bastante próximas do modelo de tempos contínuos proposto por Fine e Gray (1999). Esse resultado era esperado, conforme estudos de simulação realizados por Berger et al. (2018), que concluíram que em média, os coeficientes estimados pelo modelo de tempo discreto foram muito próximos dos coe-

ficientes obtidos através do modelo de tempo contínuo. No modelo discreto (binário), o tempo é estimado incorporando-o na estimação como covariável. Faz-se necessário este ajuste, pois o objetivo inicial é modelar o tempo t até o evento de interesse.

Tabela 3 – Estimativas, erros-padrão e p-valor obtidos com o modelo de tempo contínuo e tempo discreto para o evento formar.

Parâmetros	Tempo contínuo	Tempo discreto	Erro-Padrão*	p-valor*
Intercepto*	-	0,001	0,89	**<0,01**
Tempo*	-	1,16	0,03	**<0,01**
Sexo (Feminino)	1,91	2,22	0,37	**0,03**
Idade	0,98	0,97	0,02	0,35
Ativ. Extra. (Sim)	9,19	10,30	0,46	**<0,01**
Reserva/Escola (Sem reserva/pública)	4,60	4,38	0,52	**<0,01**
Reserva/Escola (Com reserva/privada)	2,21	2,15	1,06	0,47
Reserva/Escola (Com reserva/ pública)	0,79	0,82	0,43	0,65

*Valores referentes ao ajuste do modelo com tempos discretos - via função `glm`.

As mesmas covariáveis que foram estatisticamente significantes para o modelo de tempos discretos ajustados via `glm` (valores de p sinalizados em negrito), também foram significativas para o modelo de tempos contínuos. Ou seja, as variáveis de sexo, atividade extracurricular e categoria de ser não cotista e de escola pública foram importantes para explicar o risco do estudante de se formar, e consequentemente o seu tempo de permanência na UFBA. A variável idade (embora não tenha sido significativa estatisticamente) foi deixada no modelo pelo fato de ser uma variável vista e considerada na maioria dos estudos da área de educação. A última covariável, no entanto, deve ser interpretada com cautela, haja vista o desbalanceamento já trazido na Tabela 2 em comparação com a categoria de referência (sem reserva/privada). Estudos com maiores quantitativos de formandos, em outros cursos, por exemplo, podem analisar melhor essse cenário.

Destaca-se que, mantendo fixa as demais covariáveis, ser estudante do sexo feminino aumenta em 2,22 vezes o risco do estudante vir a se formar mais rapidamente, comparado aos estudantes do sexo masculino, para o cenário aqui analisado. Um dado interessante a ser citado envolvendo o contexto do sexo na diplomação, refere-se ao estudo de 2019 publicado pela OCDE, que diz que discentes do sexo feminino tem 34% mais chances de se formar no ensino superior do que os do sexo masculino (OECD, 2019).

Do mesmo modo, discentes que realizam algum tipo de atividade extracurricular possuem risco 10,30 vezes maior de formarem do que aqueles que não realizam nenhum tipo de atividade. Na linha de raciocínio de Astin (1984), quanto mais envolvido com a instituição, maior a probabilidade de permanência até a formatura do estudante. O propósito das atividades extracurriculares caminham no mesmo sentido da teoria de envolvimento defendido por Astin (1984), uma vez que essas atividades permitem aos estudantes uma vivência com situações ligadas à universidade e ao curso, permitindo a melhoria da qualidade do processo de ensino-aprendizagem. De acordo com o estudo de Kuh (1993), as

experiências nas atividades extracurriculares compõem a trajetória acadêmica dos estudantes, sendo responsável por 70% da formação dos discentes, auxiliando no processo de permanência e afiliação à cultura universitária.

Importante ressaltar que o modelo utilizado no processo de estimação permite apenas inferir os fatores associados ao tempo de permanência dos estudantes, de acordo com os dados do CENSUP, e para o contexto restrito dos cursos de Estatística e Matemática aqui estudados. Esses resultados não permitem concluir existência de relações de causalidade entre as variáveis e o risco do desfecho estudado. Ressalta-se que para um estudo mais detalhado e profundo, inúmeras covariáveis como por exemplo informações de renda familiar, notas dos estudantes, necessidade de trabalhar além de estudar, fatores psicológicos, etc deveriam ser consideradas. Quesitos esses que englobam muitos aspectos que também podem estar relacionados ao tempo de permanência de discentes na instituição, mas que não estão presentes nos dados do CENSUP.

Conforme dito anteriormente, os modelos para os demais eventos também serão trazidos aqui, mesmo não sendo o objetivo principal deste trabalho, apenas como um complemento ao trabalho que foi realizado. Nesse sentido sugere-se trabalhos futuros com maior enfoque também nos outros eventos, como será enfatizado na Seção 5.

A Tabela 4 sintetiza as estimativas obtidas para o evento evadir. Destaca-se que para este evento, a modelagem destes dados aponta somente a variável idade como fator estatisticamente significativo para o risco de evadir. E, neste caso, o aumento de 1 ano na idade do estudante, tem efeito protetitivo em relação ao risco de evadir. Esse resultado corrobora com a Figura 4 trazida nas análises descritivas, em que estudantes mais novos concentraram-se mais na categoria de evadidos.

Tabela 4 – Estimativas, erros-padrão e p-valor obtidos com o modelo de tempo contínuo e tempo discreto para o evento evadir.

Parâmetros	Tempo contínuo	Tempo discreto	Erro-padrão*	*p*-valor*
Intercepto*	-	0,16	0,73	**0,01**
Tempo*	-	1,05	0,02	0,07
Sexo (Feminino)	0,65	0,62	0,36	0,19
Idade	0,91	0,90	0,02	**<0,01**
Ativ. Extra. (Sim)	0,50	0,48	0,61	0,23
Reserva/Escola (Sem reserva/pública)	0,36	0,36	1,03	0,32
Reserva/Escola (Com reserva/privada)	1,20	1,03	1,03	0,97
Reserva/Escola (Com reserva/ pública)	0,93	0,94	0,32	0,85

*Valores referentes ao ajuste do modelo com tempos discretos - via função `glm`.

Nas Tabelas 5 e 6 são trazidas as estimativas verificadas para os eventos retenção e trancamento de matrícula, respectivamente. É possível observar na Tabela 5 que apenas a variável idade também teve significância estatística, em que o risco de ficar retido aumenta em 1,02 vezes para cada aumento de 1 ano na idade do discente. Esse comportamento

também foi possível identificar previamente por meio da Figura 4 com os *boxplots* das idades segundo a situação de vínculo. Neste gráfico identificamos que os estudantes que configuravam-se nessa categoria eram, em geral, mais velhos. Ou seja, quanto mais tempo se passa, e consequentemente os estudantes ficam mais velhos, maior é o risco que o mesmo fique retido na UFBA.

Tabela 5 – Estimativas, erros-padrão e p-valor obtidos com o modelo de tempo contínuo e tempo discreto para o evento ficar retido.

Parâmetros	Tempo contínuo	Tempo discreto	Erro-padrão*	*p*-valor*
Intercepto*	-	0,0003	0,64	**<0,01**
Tempo*	-	1,25	0,02	**<0,01**
Sexo (Feminino)	0,83	0,84	0,33	0,61
Idade	1,02	1,02	0,01	**0,03**
Ativ. Extra. (Sim)	0,29	0,27	1,03	0,21
Reserva/Escola (Sem reserva/pública)	2,36	2,82	0,46	0,02
Reserva/Escola (Com reserva/privada)	0,67	0,83	1,04	0,86
Reserva/Escola (Com reserva/ pública)	1,15	1,26	0.32	0,46

*Valores referentes ajuste do modelo com tempos discretos - via função `glm`.

Finalmente, temos a Tabela 6 com as estimativas obtidas para o evento de matrícula trancada. O ajuste deste modelo não trouxe significância estatística em nenhuma das covariáveis consideradas. Destaca-se o alto erro-padrão e a estimativa quase nula associados ao parâmetro estimado para os estudantes com reserva de vagas e oriundos de escola privada. Vale lembrar que essa categoria possui um baixo quantitativo (comparado com as demais categorias dessa variável), conforme trazido na Tabela 2. Ainda nesse contexto, observou-se que, para o evento de matrícula trancada segundo a variável de reserva de vaga/escola de conclusão do ensino médio, dentre os estudantes que trancaram a matrícula, nenhum era cotista de escola privada (observado na Figura 6). Ainda assim, o modelo foi trazido como um complemento do estudo, podendo servir para direcionar aplicações em projetos futuros.

Tabela 6 – Estimativas, erros-padrão e *p*-valor obtidos com o modelo de tempo contínuo e tempo discreto para o evento matrícula trancada.

Parâmetros	Tempo contínuo	Tempo discreto	Erro-padrão*	*p*-valor*
Intercepto*	-	0,017	0,64	**<0,001**
Tempo*	-	0,99	0,03	0,89
Sexo (Feminino)	0,52	0,53	0,45	0,16
Idade	0,98	0,98	0,02	0,37
Ativ. Extra. (Sim)	0,27	0,26	1,02	0,19
Reserva/Escola (Sem reserva/pública)	0,42	0,43	1,04	0,42
Reserva/Escola (Com reserva/privada)	<0,00000001	<0,000001	677,66	0,98
Reserva/Escola (Com reserva/ pública)	1,17	1,19	0,36	0,98

*Valores referentes ao ajuste do modelo com tempos discretos - via função `glm`.

4.2.2 Análise dos pressupostos do modelo

Para analisar o ajuste do modelo cujas estimativas foram apresentadas na Tabela 3, verificou-se o resíduo de Schoenfeld para cada covariável, os resíduos Martingale e *Deviance*.

Os resíduos de Schoenfeld para as variáveis ajustadas bem como os p-valores do teste de correlação linear com o tempo (verificação da proporcionalidade dos riscos) são apresentados na Figura 12. Ao nível de significância adotado, podemos perceber que para a covariável de idade é rejeitada a hipótese de que a variação no tempo seja igual a zero. Isto é, o efeito da idade parece ter uma tendência crescente ao longo do tempo. Na prática, equivale a dizer que o risco do estudante formar, relacionado a sua idade, aumenta à medida que o tempo passa, o que é esperado.

O teste de proporcionalidade global do modelo também rejeita a hipótese nula de proporcionalidade de todas as covariáveis (p-valor = 0,002608), como é possível verificar na Figura 12. Desse modo, o pressuposto principal do modelo trazido com a abordagem de Fine e Gray (1999) através do ajuste ao modelo de Cox (com ponderação dos indivíduos sob o risco), é violado. No entanto, o modelo final tratado aqui para tempos discretos permite a modelagem de riscos de subdstribuições não proporcionais, conforme trazido por (BERGER et al., 2018). Conclui-se que o modelo proposto foi satisfatório ao objetivo definido no trabalho.

Outro aspecto analisado diz respeito a verificação de pontos aberrantes presentes no modelo através dos resíduos martingale e *deviance*. Na Figura 13 podemos verificar dois pontos no limite do intervalo para os resíduos martingale, e para os resíduos *deviance* identificou-se mais cinco pontos diferentes fugindo do limite de intervalo em 2. Esses sete pontos identificam observações mal ajustados pelo modelo. Neste caso, os discentes identificados correspondem a estudantes que formaram com tempos menores (6 e 8 semestres) e também com tempos maiores, de 12 e 20 semestres. No entanto, optou-se por não retirar essas observações devido ao baixo quantitativo de estudantes que experimentaram o evento final de interesse (formar), totalizando 30 discentes, o que poderia ocasionar perda de informação para este contexto.

A análise dos resíduos obtidos, no entanto, se mostra razoável, e concluiu-se que o modelo contribuiu para identificar os fatores associados ao tempo de permanência dos estudantes de Estatística e Matemática da UFBA. Salienta-se o baixo quantitativo de formandos na base de dados final, conforme já mencionado, o que limita uma visão mais ampla desses cursos. E, mesmo com a violação do pressuposto de proporcionalidade global dos riscos, considera-se o modelo ajustado como potencial contribuinte na identificação do perfil de diplomação desses discentes.

Figura 12 – Resíduos padronizados de Schoenfeld e teste de proporcionalidade dos riscos.

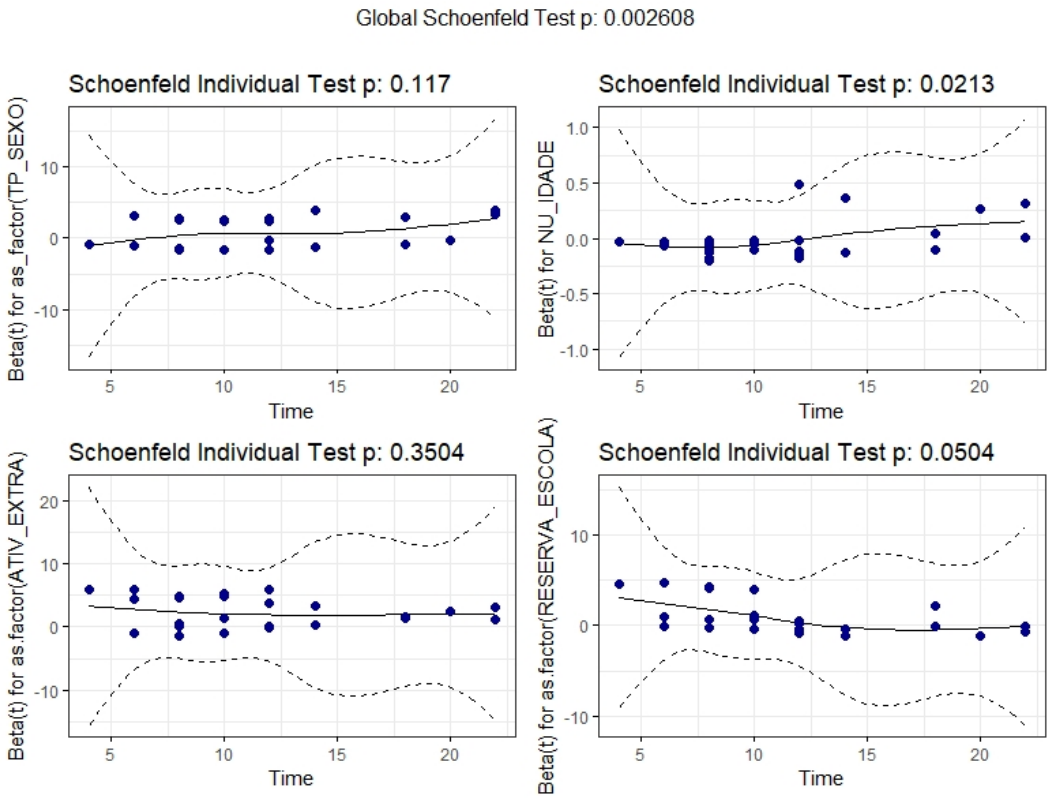
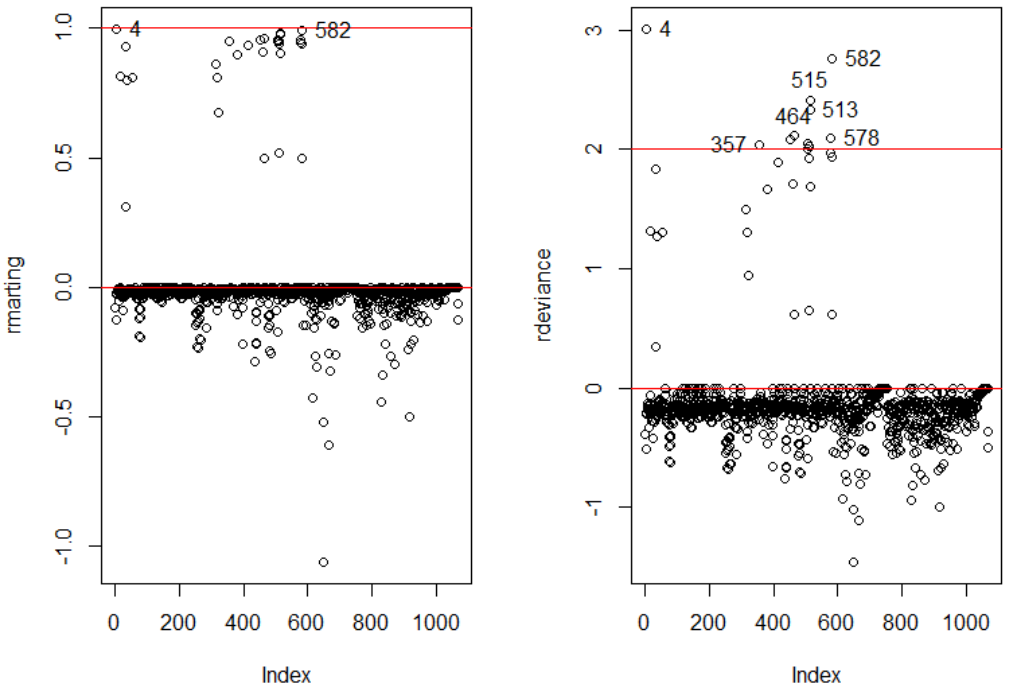


Figura 13 – Resíduos martingale e *deviance*.



5 CONSIDERAÇÕES FINAIS

Nesta seção são trazidas as conclusões obtidas com os resultados deste trabalho, bem como as sugestões para realização de pesquisas futuras que abordem a temática e procedimento tratados aqui.

5.1 Conclusões

Este trabalho trouxe a análise dos dados do CENSUP, maior fonte de dados públicos sobre o ensino superior no Brasil, como pilar na identificação de fatores associados ao tempo de permanência de estudantes dos cursos de Estatística e Matemática da UFBA. A proposta consistiu em incorporar um estudo baseado em dados de fornecimento público, o que até então não verificou-se nos trabalhos feitos sobre o tema na instituição. Sendo assim, foi possível agregar informações importantes sobre esse contexto na ótica desses cursos.

Através da aplicação da metodologia de eventos competitivos e utilizando covariáveis disponibilizadas pelo CENSUP, o trabalho conseguiu identificar fatores associados e que estão relacionados com o tempo de permanência desses estudantes na UFBA. Por exemplo, verificou-se que, mantendo fixa as demais covariáveis, ser estudante do sexo feminino, comparado aos estudantes do sexo masculino, aumenta 2,22 vezes o risco de formar em menos tempo (recorda-se que para esse evento, o risco é interpretado como algo bom). Além disso, ser estudante que realiza algum tipo de atividade extracurricular aumenta em 10,30 vezes o risco de diplomar-se.

Este trabalho também trouxe uma sugestão de conceito para classificar estudantes evadidos e retidos, visto que não foram encontradas referências acerca desses conceitos na UFBA. O conceito aqui proposto pode direcionar estudos posteriores cujo interesse seja investigar os temas de evasão ou retenção. Além disso, a análise feita introduziu materiais complementares para agregar na qualidade das informações trazidas com os dados. Esses materiais complementares, construídos em formato de planilhas, podem ser consultados no repositório virtual do `GitHub`: <https://github.com/jessicafagundesg>. Nesse ponto, o trabalho feito serve até mesmo como um tutorial de análise dos dados do CENSUP, podendo ser expandido para outros estudos em IES.

Por fim, foi trazida a consolidação de todas as análises feitas neste trabalho através do aplicativo *web* `R` *Shiny*, disponibilizado em *link* na *internet*: <https://jessica-fagundesg.shinyapps.io/tempodepermanenciaimeufba/>. Também foi desenvolvido um aplicativo *web* `R` *Shiny* com os resultados obtidos no projeto deste trabalho: <https://jessica-fagundesg.shinyapps.io/tempodepermanenciaufba>. Sendo assim, os respectivos aplicativos podem

ser consultados por toda a comunidade UFBA e pelos leitores que tenham interesse.

As limitações verificadas na realização deste trabalho dizem respeito a: baixa quantidade de formandos na base de dados final dos cursos escolhidos; não utilização da variável de cor/raça por alto percentual de não declarantes (ponto que merece atenção especial por parte da instituição junto ao MEC, visto ser variável importante nesse contexto); não utilização da variável de apoio social por quantitativo ínfimo. E ainda, não foi utilizada a variável de tipo de ingresso nas análises por essa não ter agregado informações importantes para o presente contexto.

Também observou-se que, mesmo não sendo o objetivo deste trabalho, o ajuste do modelo para os outros desfechos não demonstrou significância estatística para as covariáveis em questão (exceto idade para evadido e retido). No entanto, a aplicação abre possibilidade para que esses outros eventos possam ser melhor explorados em circunstância diferentes, como em outros cursos da UFBA. Outra limitação que vale ser pontuada refere-se a ausência de informações socioeconômicas (tais como renda familiar, nível de escolaridade dos pais), aspectos psicológicos, notas/rendimentos na instituição, necessidade de trabalho conciliado com a graduação, etc. Características essas que não são abordadas pelo CENSUP mas que poderiam agregar ainda mais às análises.

Conclui-se que os objetivos estabelecidos neste trabalho foram alcançados e espera-se que as informações trazidas possam contribuir com as tomadas de decisão e estudos da instituição para com os seus discentes.

5.2 Sugestões para pesquisas futuras

Como sugestão para trabalhos futuros, recomenda-se a aplicação da metodologia aos demais cursos e/ou áreas de conhecimento da UFBA, com o intuito de verificar se, por exemplo, os eventos competitivos tratados aqui teriam significância estatística para as covariáveis abordadas em outros diferentes contextos.

Recomenda-se também a examinação das covariáveis de apoio social e tipo ingresso no ajuste do modelo para os demais cursos, a fim de verificar se em outros casos é possível estabelecer um risco ou fator protetivo das covariáveis, dado os diferentes desfechos. Com essa recomendação, seria possível traçar perfis de discentes para os desfechos também em outros cursos da universidade, podendo contribuir para as discussõs acerca da temática de evasão, retenção, diplomação e trancamento de matrícula na UFBA.

Outra sugestão diz respeito a variável idade. Durante a realização deste trabalho, verificou-se que a recategorização da idade em faixas etárias não modificou a significância estatística dessa variável no ajuste do modelo para o desfecho formar. No entanto, uma análise que pode ser feita é se essa nova variável tem efeito sob o tempo de permanência até formar, evadir, etc, para outros cursos e áreas da UFBA. Em cursos com maiores

quantidades de formandos por ano talvez seja possível mapear melhor esse e outros fatores.

Finalmente, sugere-se que outros modelos sejam utilizados na análise destes dados, tal como, o modelo Weibull discreto. Neste caso, a variável de interesse (tempo) seria modelada diretamente pelas covariáveis de interesse. E os resultados encontrados poderiam ser comparados com os achados deste trabalho.

REFERÊNCIAS

AMARAL, N. C.; PINTO, J. M. de R. O financiamento das ies brasileiras em 2005: recursos públicos, privados e custo dos alunos. *Série Estudos - Periódico do Programa de Pós-Graduação em Educação da UCDB*, Campo Grande, MS, n. 30, p. 61, 2010.

ANDIFES. V pesquisa nacional de perfil socioeconômico e cultural dos (as) graduandos (as) das ifes - 2018. *Relatório Executivo*, Uberlândia, MG, 2019.

ANDRADE, J. B. A evasão nos bacharelados interdisciplinares da ufba: Um estudo de caso. *Repositório Institucional UFBA*, Salvador, BA, 2014.

ASTIN, A. W. Student involvement: A developmental theory for higher education. *Journal of College Student Development*, California, USA, 1984.

BAILAR-III, J. C.; MOSTELLER, F. Medical uses of statistics. *NEJM Books*, Boston, EUA, 1992.

BERGER, M. et al. Subdistribution hazard models for competing risks in discrete time. *Biostatistics*, Bonn, Germany, 2018. Disponível em: <https://academic.oup.com/biostatistics/article/21/3/449/5168329?login=false>. Acesso em: 02 abr. 2022.

BRASIL. Decreto-lei nº 9.155, de 8 de abril de 1946. cria a universidade da bahia e dá outras providências. *Diário Oficial da República Federativa do Brasil*, Brasília, DF, 1946.

BRASIL. Constituição (1988). *Constituição da República Federativa do Brasil de 1988, Capítulo 3: Da Educação, Da Cultura e do Desporto, Seção 1: Da Educação, Art. 205*, Brasília, DF, Centro Gráfico, p. 1, 1988.

BRASIL. Lei nº 9.394, de 20 de dezembro de 1996: Estabelece as diretrizes e bases da educação nacional. *Diário Oficial da República Federativa do Brasil*, Brasília, DF, p. 27833, 1996.

BRASIL. Lei nº 10.172, de 09 de janeiro de 2001: Aprova o plano nacional de educação e dá outras providências. *Diário Oficial da República Federativa do Brasil*, Brasília, DF, p. 1, 2001.

BRASIL. Decreto nº 6.094, de 24 de abril de 2007: Dispõe sobre a implementação do plano de metas compromisso todos pela educação, pela união federal, em regime de colaboração com municípios, distrito federal e estados, e a participação das famílias e da comunidade, mediante programas e ações de assistência técnica e financeira, visando a mobilização social pela melhoria da qualidade da educação básica. *Diário Oficial da República Federativa do Brasil*, Brasília, DF, p. 5, 2007.

BRASIL. *O Plano de Desenvolvimento da Educação: razões, princípios e programas*, Brasília, DF, p. 15, 2008.

BRASIL. Lei nº 11.788, de 25 de setembro de 2008: Dispõe sobre o estágio de estudantes; altera a redação do art. 428 da consolidação das leis do trabalho – clt, aprovada pelo decreto-lei no 5.452, de 1o de maio de 1943, e a lei no 9.394, de 20 de dezembro de 1996; revoga as leis nos 6.494, de 7 de dezembro de 1977, e 8.859, de 23 de março de 1994, o

parágrafo único do art. 82 da lei no 9.394, de 20 de dezembro de 1996, e o art. 6o da medida provisória no 2.164-41, de 24 de agosto de 2001; e dá outras providências. *Diário Oficial da República Federativa do Brasil*, Brasília, DF, p. 3, 2008.

BRASIL. Lei nº 12.711 de 29 de agosto de 2012: Dispõe sobre o ingresso nas universidades federais e nas instituições federais de ensino técnico de nível médio e dá outras providências. *Diário Oficial da República Federativa do Brasil*, Brasília, DF, p. 1, 2012.

BRASIL. *Notas Estatísticas do Censo da Educação Superior de 2019*, p. 16–17 e 13–14, 2019. Disponível em: <ttps://download.inep.gov.br/educacao_superior/censo_superior/documentos/2020/Notas_Estatisticas_Censo_da_Educacao_Superior_2019.pdf>. Acesso em: 30 set. 2021.

BRASIL. *Resumo Técnico do Censo da Educação Superior 2019*, p. 11, 2019. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/resumo_tecnico_censo_da_educacao_superior_2019.pdf>. Acesso em: 15 set. 2021.

BRESLOW, N. E.; CROWLEY, J. A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics*, v. 2, n. 3, p. 437–453, 1974.

CAMPELLO, A. de V. C.; LINS, L. N. Metodologia de análise e tratamento da evasão e retenção em cursos de graduação de instituições federais de ensino superior. *XXVIII Encontro Nacional de Engenharia de Produção*, p. 2, 2008.

CAMPOS, C. F. Causas da evasão e retenção no curso de biblioteconomia e documentação da universidade federal da bahia. *Repositório Institucional UFBA*, Salvador, BA, 2020.

CÂNDIDO, L. G. *Evasão universitária na graduação presencial brasileira: um panorama a partir do Censo da Educação Superior (2009-2018)*, Niterói, RJ, 2019.

CARVALHO, A. C. M. de. *Estudo sobre a retenção de estudantes do curso de ciências contábeis em uma IES pública*, p. 16, 2019.

CARVALHO, A. P. de. *Fatores institucionais associados à evasão na educação superior*, Goiânia, GO, 2017.

CARVALHO, M. S. et al. Análise de sobrevivência: Teoria e aplicações em saúde. *Editora Fiocruz, 1ª reimpressão*, São Paulo, SP, 2019.

COLOSIMO, E. A. et al. Empirical comparisons between kaplan-meier and nelson-aalen survival function estimators. *Journal Of Statistical Computation and Simulation*, v. 72, p. 299–308, 2002.

COLOSIMO, E. A.; GIOLO, S. R. Análise de sobrevivência aplicada. *Editora Edgard Blücher, 1ª Edição*, São Paulo, SP, 2006.

COSTA, D. de M.; BARBOSA, F. V.; GOTO, M. M. M. O novo fenômeno da expansão da educação superior no brasil. *Revista Reuna*, Belo Horizonte, MG, v. 16, n. 1, p. 15–29, 2011.

CRUZ, U. A. da. Evasão de discentes: Um estudo na universidade federal da bahia. *Repositório Institucional UFBA*, Salvador - BA, 2019.

CUNHA, J. V. A. da; NASCIMENTO, E. M.; DURSO, S. de O. Razões e influências para a evasão universitária: um estudo com estudantes ingressantes nos cursos de ciências contábeis de instituições públicas da região sudeste. *Advances in Scientific and Applied Accounting*, v. 9, n. 2, p. 146, 2016.

DCE-UFBA. *Aprovado Novo Regulamento de Ensino da UFBA*, 2014. Disponível em: <https://ufbadce.wordpress.com/2014/12/18/aprovado-novo-regulamento-de-ensino-da-ufba-2/>. Acesso em: 11 de out. 2021.

ECHEVESTE, S. S. Análise de sobrevivência: Um estudo na área educacional. *UFRGS - Sistema de Bibliotecas: Biblioteca Setorial de Matemática*, Porto Alegre, RS, 1997.

FÁVERO, M. de Lourdes de A. Da cátedra universitária ao departamento: subsídios para discussão. *23ª Reunião Anual da ANPed*, 2000.

FERREIRA, A. M. S. dos A. *Caracterização da assistência estudantil na Universidade do Estado da Bahia na perspectiva do Censo da Educação Superior*, Salvador, BA, 2018.

FERREIRA, M. Determinantes do rendimento académico no ensino superior. *Revista Internacional d´Humanitats 15*, CEMOrOc-Feusp, Univ. Autónoma de Barcelona, 2009.

FILHO, R. L. L. e S. et al. A evasão no ensino superior. *Cadernos de Pesquisa*, v. 37, n. 132, p. 643, 2007.

FINE, J. P.; GRAY, R. J. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, Alexandria, USA, 1999.

FREITAS, K. S. de. Alguns estudos sobre evasão e persistência de estudantes. *Revista Científica EccoS*, v. 11, n. 1, p. 247–264, 2009.

GARCIA, L. M. L. da S.; LARA, D. F.; ANTUNES, F. Análise da retenção no ensino superior: um estudo de caso em um curso de sistemas de informação. *Revista da Faculdade de Educação - Universidade do Estado do Mato Grosso*, p. 17, 2020.

HOFFMANN, I. L.; NUNES, R. C.; MULLER, F. M. As informações do censo da educação superior na implementação da gestão do conhecimento organizacional sobre evasão. *Gestão e Produção*, São Carlos, SP, v. 26, n. 2, 2019.

INEP. *Censo da Educação Superior: Histórico*, 2019. Disponível em: <http://portal.inep.gov.br/censo-da-educacao-superior/historico>. Acesso em: 15 out. 2021.

INEP. *Preenchimento: Censo da Educação Superior 2019*, 2019. Disponível em: <http://portal.inep.gov.br/preenchimento>. Acesso em: 02 out. 2021.

INEP. *Censo da Educação Superior: Microdados*, 2019. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>. Acesso em: 15 ago. 2021.

INEP. *Manual de Preenchimento do Censo da Educação Superior - Módulo Curso*, Brasília, DF, 2019. Disponível em: <https://download.inep.gov.br/educacao_superior/censo_superior/questionarios_e_manuais/2019/Modulo_Curso.pdf>. Acesso em: 02 nov. 2021.

INEP. *Manual de Preenchimento do Censo da Educação Superior - Módulo Aluno*, Brasília, DF, 2019. Disponível em: <https://download.inep.gov.br/educacao_superior/censo_superior/questionarios_e_manuais/2019/Modulo_Aluno.pdf>. Acesso em: 02 nov. 2021.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53, n. 282, p. 457–481, 1958.

KUH, G. D. In their own words: What students learn outside the classroom. *American Educational Research Journal*, p. 277–304, 1993.

LIMA, M.; COUTINHO, D.; SANTOS, V. Trajetórias interrompidas no curso de psicologia em relação ao bacharelado interdisciplinar na ufba. *Revista CAMINE: Caminhos da Educação*, v. 7, n. 2, p. 30–51, 2015. Disponível em: <https://periodicos.franca.unesp.br/index.php/caminhos/article/view/1364>. Acesso em: 12 de out. 2021.

LOBO, M. B. de C. M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. *Associação Brasileira de Mantenedoras de Ensino Superior, Cadernos ABMES 25*, p. 9–58, 2012.

MARQUES, F. T. A volta aos estudos dos alunos evadidos do ensino superior brasileiro. *Revista Cadernos de Pesquisa*, São Paulo, SP, v. 50, n. 178, p. 1061–1077, 2020.

MARTINS, A. C. P. *Ensino Superior no Brasil: da descoberta aos dias atuais*, Acta Cirúrgica Brasileira - (Suplemento 3), v. 17, 2002.

MARTINS, G. O.; ROCHA, S. H. Evasão e tempo de permanência no curso de estatística da universidade federal do paraná: Um estudo sobre os alunos que ingressaram no período de 1991 a 2011. *UFRGS - Sistema de Bibliotecas: Biblioteca Setorial de Matemática*, Curitiba, PR, 2011.

MAZIEIRO, G. Em 4 anos, brasil reduz investimento em educação em 56%; cortes continuam. *UOL - Educação*, 2019. Disponível em: <https://educacao.uol.com.br/noticias/2019/05/02/em-4-anos-brasil-reduz-investimento-em-educacao-em-56.htm>. Acesso em: 16 de out. 2021.

MEC. *Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas*, 1996. Disponível em: <http://www.dominiopublico.gov.br/download/texto/me001613.pdf>. Acesso em: 01 de out. 2021.

MEC. Programa de apoio a planos de reestruturação e expansão das universidades federais. *Reuni 2008 – Relatório de Primeiro Ano*, 2009.

MEC. *Jubilamento: 1) O aluno pode ser jubilado?*, 2018. Disponível em: <http://portal.mec.gov.br/component/tags/tag/perguntas-frequentes?start=20>. Acesso em: 04 de out. 2021.

MEC. *Indicadores de Fluxo da Educação Superior*, 2020. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-de-fluxo-da-educacao-superior>. Acesso em: 02 out. 2021.

MEIER, P. Estimation of a distribution function from incomplete observations. *Journal Of Applied Probability Trust*, Sheffield, England, 1975.

MOREIRA, L. K. R.; MOREIRA, L. R.; SOARES, M. G. Educação superior no brasil: discussões e reflexões. *Educação por Escrito*, v. 9, n. 1, p. 134–150, 2018.

OECD. *Education at a Glance 2019: OECD Indicators*, Paris, France, 2019.

OECD. *Education at a Glance: Country Notes - Brazil*, p. 3, 2020.

OLIVEIRA, E. Cortes no orçamento de universidades federais poderão afetar mais de 70 mil pesquisas. *G1-Educação*, 2021. Disponível em: <https://g1.globo.com/educacao/noticia/2021/05/31/cortes-no-orcamento-de-universidades-federais-podera-impactar-em-mais-de-70-mil-/pesquisas-relacionadas-a-pandemia.ghtml>. Acesso em: 16 de out. 2021.

PEIXOTO, A. de L. A. et al. Cotas e desempenho acadêmico na ufba: um estudo a partir dos coeficientes de rendimento. *Avaliação*, Campinas, Sorocaba, SP, v. 21, n. 2, p. 569–591, 2016.

PEREIRA, A. S. et al. Fatores relevantes no processo de permanência prolongada de discentes nos cursos de graduação presencial: um estudo na universidade federal do espírito. *Ensaio: Avaliação e Políticas Públicas em Educação*, 2015. Disponível em: <https://www.redalyc.org/articulo.oa?id=399542696010>. Acesso em: 05 mai. 2022.

PINHEIRO, S. M. C. Uma abordagem dos modelos de longa duração para análise de sobrevivência da evasão de estudantes em cursos de engenharia: Epufba como um estudo de caso. *Repositório Institucional UFBA*, Salvador, BA, 2021.

PINTILIE, M. Competing risks: A practical perspective. *Wiley*, New York, USA, 2006.

PINTO, D. Ufma anuncia desligamento de 525 estudantes nos cursos de graduação. *G1 - Maranhão: Rede Mirante*, 2020. Disponível em: <https://g1.globo.com/ma/maranhao/noticia/2020/02/04/ufma-anuncia-jubilamento-de-525-estudantes-nos-cursos-de-graduacao.ghtml>. Acesso em: 05 de out. 2021.

QS. *QS Latin America University Rankings 2022*, United Kingdom, 2021. Disponível em: <https://www.topuniversities.com/university-rankings/latin-american-university-rankings/2022>. Acesso em: 29 out. 2021.

RODRIGUES, H. W. Jubilamento ainda existe? *Revista Gestão Universitária*, 2007. Disponível em: <http://gestaouniversitaria.com.br/artigos/jubilamento-ainda-existe>. Acesso em: 04 de out. 2021.

SACCARO, A. *Ampliação do Ensino Superior Brasileiro: um estudo sobre as causas da evasão e o impacto da bolsa permanência do PNAES*, Porto Alegre, RS, 2016.

SALERNO, M. S. et al. *Uma proposta de sistematização do debate sobre falta de engenheiros no Brasil*, Brasília, DF, 2013.

SAMPAIO, H. *Evolução do ensino superior brasileiro, 1808-1990*, NUPES: Núcleo de Pesquisas sobre Ensino Superior da Universidade de São Paulo, 1991.

SANGIOVANNI, R. Mesmo em cenário adverso, ufba melhora e atinge sua maior pontuação no Índice geral de cursos. *EdgarDigital*, Salvador, BA, 2021. Disponível em: <https://www.edgardigital.ufba.br/?p=20099>. Acesso em: 29 out. 2021.

SANGIOVANNI, R. Ranking the: Ufba sobe duas posições e é a 26ª da américa latina. *EdgarDigital*, Salvador, BA, 2021. Disponível em: <https://www.edgardigital.ufba.br/?p=20970>. Acesso em: 29 out. 2021.

SANTO, A. C. do E. A trajetória acadêmica e o perfil dos estudantes da universidade federal da bahia, nos cursos de alta demanda, pós-sistema de cotas. *Repositório Institucional UFBA*, Salvador, BA, 2013.

SANTOS, P. K. dos; GIRAFFA, L. M. M. Evasão na educação superior: Um estudo sobre o censo da educação superior no brasil. *Congresos CLABES III*, Palacio de Minería, Universidad Nacional Autónoma de México, México, 2013.

SANTOS, P. V. S. Adaptação à universidade dos estudantes cotistas e não cotistas: Relação entre vivência acadêmica e intenção de evasão. *Repositório Institucional UFBA*, Salvador, BA, 2013.

SANTOS, V. B. dos. Evasão dos estudantes do curso de licenciatura em educação física da uba: Período entre 2011 e 2016. *Repositório Institucional UFBA*, Salvador, BA, 2017.

SEMESP. *Mapa do Ensino Superior no Brasil*, p. 14, 2016. Disponível em: <https://www.convergenciacom.net/pdf/mapa_ensino_superior_2016.pdf>. Acesso em: 05 jul. 2022.

SEMESP. *Mapa do Ensino Superior no Brasil*, v. 11, p. 29, 2021. Disponível em: <https://www.semesp.org.br/wp-content/uploads/2021/06/Mapa-do-Ensino-Superior-Completo.pdf>. Acesso em: 01 out. 2021.

SENKEVICS, A. S.; MELLO, U. M. O perfil discente das universidades federais mudou pós-lei de cotas? *Cadernos de Pesquisa*, São Paulo, SP, v. 49, n. 172, p. 184–208, 2019.

SOUZA, E. de; FREITAS, L. F. Um estudo sobre a evasão nos cursos de graduação dos institutos federais. *Revista Brasileira da Educação Profissional e Tecnológica*, v. 1, n. 20, 2021.

SPARK, A. *SparkR (R on Spark) - Spark 3.2.0 Documentation*, 2021. Disponível em: <https://spark.apache.org/docs/latest/sparkr.html#overview>. Acesso em: 30 out. 2021.

TOUTAIN, L. M. B. B.; SILVA, R. R. G. da. Universidade federal da bahia: Do século xix ao século xxi. *EDUFBA*, Edição Memorial, Salvador, BA, v. 1, p. 16, 2010.

TUTZ, G.; SCHMID, M. Modeling discrete time-to-event data. *Springer*, New York, USA, 2016.

UFBA. *UFBA em Números: Especial 60 anos*, Salvador - BA, p. 5–7, 2006.

UFBA. *Regimento Interno da Reitoria*, Salvador, BA, 2013.

UFBA. *Manual de Extensão Universitária da UFBA*, Salvador, BA, 2014.

UFBA. *Regulamento de Ensino de Graduação e Pós-Graduação stricto sensu (REGPG), da Universidade Federal da Bahia.*, p. 22–23, 2015. Disponível em: <https://www.ufba.br/sites/portal.ufba.br/files/Resolucao_n_012015_REGPG_atualizado_01-04-2015%29.pdf>. Acesso em: 10 de out. 2021.

UFBA. *UFBA em Números, Ano base 2015*, Salvador - BA, p. 3, 2016.

UFBA. *UFBA em números: Retrospectiva - Especial 70 anos*, Salvador - BA, 2016.

UFBA. *Plano de Desenvolvimento Institucional 2018 - 2022, Universidade Federal da Bahia*, Salvador - BA, p. 81, 2017.

UFBA. *UFBA em Números, Ano base 2016*, Salvador - BA, p. 3, 2017.

UFBA. *UFBA em Números, Ano base 2017*, Salvador - BA, p. 3, 2018.

UFBA. *UFBA em Números, Ano base 2018*, Salvador - BA, p. 3, 2019.

UFBA. Ações afirmativas, 15 anos: das cotas ao sucesso, profissionais contam suas histórias. *Portal UFBA*, Salvador, BA, 2019. Disponível em: <https://portal.ufba.br/ufba_em_pauta/acoes-afirmativas-15-anos-das-cotas-ao-sucesso-profissionais-contam-suas-historias>. Acesso em: 30 out. 2021.

UFBA. *UFBA em Números, Ano base 2019*, Salvador - BA, p. 3, 2020.

UFBA. Até 20 de novembro, ufba recebe informações para atualização de sua carta de serviços. *EdgarDigital*, Salvador, BA, 2020. Disponível em: <https://www.edgardigital.ufba.br/?p=18773>. Acesso em: 29 out. 2021.

UFBA. *Histórico: Universidade Federal da Bahia – A primeira do Brasil*, 2021. Disponível em: <https://www.ufba.br/historico>. Acesso em: 20 out. 2021.

UFBA. *SUPAC - Supertindência de Administração Acadêmica - Graduação 2021.2*, Salvador, BA, 2021. Disponível em: <https://supac.ufba.br/guia-matricula-graduacao>. Acesso em: 29 out. 2021.

UFBA. *Ingresso UFBA: Coordenação de Seleção e Orientação (CSOR) - SiSU UFBA*, 2021. Disponível em: <https://ingresso.ufba.br/>. Acesso em: 29 out. 2021.

UFBA. *Resolução Nº 05/2021: Regulamenta as atividades de monitoria no âmbito dos cursos de graduação, na UFBA e revoga as Resoluções n. 06/2012, 07/2017, 02/2018 e 11/2019.*, Salvador, BA, 2021.

UFBA. Apresentação. *Pró-Reitoria de Pesquisa, Criação e Inovação - Sistema de Gerenciamento de Bolsas de Iniciação - SISBIC*, Salvador, BA, 2021. Disponível em: <https://sisbic.ufba.br/sisbic/Welcome.do##>. Acesso em: 29 out. 2021.

UNESCO. *Declaração Mundial sobre Educação Superior no Século XXI: Visão e Ação - 1998*, Conferência Mundial sobre Educação Superior - UNESCO - Paris, 1998. Disponível em: <http://www.direitoshumanos.usp.br/index.php/Direito-a-Educacao/declaracao-mundial-sobre-educacao-superior-no-seculo-xxi-visao-e-acao.html>. Acesso em: 30 set. 2021.

ZOCCOLI, M. M. de S. Educação superior brasileira: Política e legislação. *Editora Ibpex*, Curitiba, PR, v. 3, 2009.

APÊNDICE A – VARIÁVEIS DO CENSO DA EDUCAÇÃO SUPERIOR UTILIZADAS E VARIÁVEIS AUXILIARES CRIADAS

Tabela 7 – Variáveis utilizadas e variáveis auxiliares criadas - CENSUP 2019.

VARIÁVEIS	DESCRIÇÃO
NU_ANO_CENSO	Ano de aplicação do censo (2019, neste trabalho).
CO_IES	Código de identificação da IES (578 - UFBA).
CO_CURSO	Código de identificação do curso gerado pelo E-MEC.
TP_GRAU_ACADEMICO	Tipo do grau acadêmico conferido ao diplomado (Bacharelado e Licenciatura).
ID_ALUNO	Código de identificação para o aluno da educação superior.
CO_ALUNO_CURSO	Código de identificação para o vínculo do aluno ao curso.
TP_SEXO	Informa o sexo do aluno (Feminino e Masculino).
NU_ANO_NASCIMENTO	Ano de nascimento do aluno.
NU_MES_NASCIMENTO	Mês de nascimento do aluno.
NU_DIA_NASCIMENTO	Dia de nascimento do aluno.
NU_IDADE	Idade que o aluno completa no ano de referência do Censo.
TP_SITUACAO	Tipo de situação de vínculo do aluno no curso (Cursando, Matrícula trancada, Desvinculado do curso, Transferido para outro curso da mesma IES, Formado e Falecido).
QT_CARGA_HORARIA_INTEG	Total de carga horária aproveitada pelo aluno.
QT_CARGA_HORARIA_TOTAL	Total de carga horária do curso - informado pela IES.
DT_INGRESSO_CURSO	Data de ingresso do aluno no curso correspondente ao 1º semestre (01/01/2015) e ao 2º semestre (01/07/2015).

IN_RESERVA_VAGAS	Informa se o aluno participa de programa de reserva de vagas.
IN_ATIVIDADE_EXTRACURRICULAR	Informa se o aluno participa de algum tipo de atividade extracurricular (estágio não obrigatório, extensão, monitoria e pesquisa).
TP_ESCOLA_CONCLUSAO_ENS_MEDIO	Tipo de escola que o aluno concluiu ensino médio (pública ou privada).
IN_CONCLUINTE	Informa se o aluno é concluinte em 2019.
NU_ANO_INGRESSO	Ano de ingresso do aluno no curso.
NU_CARGA_HORARIA	Carga horária mínima do curso - sistema E-MEC.
TP_SEMESTRE_CONCLUSAO	Semestre de conclusão do curso (Para alunos formados em 2019).
TEMPOS	Variável criada descrevendo os tempos de falha e tempos de censura.
NU_CARGA_HORARIA_NOVA	Variável criada com base em anos dos censos anteriores.
TP_SITUACAO_NEW	Variável criada, em que acrescentou-se a situação de evadido (desvinculado do curso e transferido para outro curso da UFBA), retido e matrícula trancada.
PCT_INTEG	Variável criada, representando a porcentagem de horas já integralizada pelo estudante no curso.
INTEG_SEMEST	Variável criada com base em anos dos censos anteriores.
EVENTOS_COD	Variável criada contendo informações sobre os eventos competitivos e a censura (0, 1, 2, 3 e 4).
RESERVA_ESCOLA	Variável criada cruzando as variáveis reserva de vagas e tipo de escola.

Fonte: INEP (2019c). Adaptado.

APÊNDICE B – MODELOS E PACOTES UTILIZADOS

1: Modelo de Fine e Gray (1999) para tempos contínuos.

```
# Obtendo FIA's

compete<-DADOS
fia<-survfit(Surv(TEMPOS,event=status_event>0)~1,
etype=TP_SITUACAO_NEW, data=compete)

plot(fia, ylab="FIA", xlab="Semestres", fun="event",
lty=c(2,1,4,3), ylim=c(0,.5))

legend("topleft", c("Retido", "Evadido", "Formado",
"Matricula trancada"), lty = c(3,2,1,4),
title ="Tipo de desfecho", bty="n")


# Definindo os pesos

compete.peso <-
  mstate::crprep(Tstop = "TEMPOS",
                 status = "status_event",
                 trans = 1:4,
                 keep = c("TP_SEXO","NU_IDADE","RESERVA_ESCOLA"
,"IN_ATIVIDADE_EXTRACURRICULAR"),
                 data = compete)

# Ajustando o modelo para tempos continuos

m1.cox <- coxph(Surv(Tstart, Tstop, status==1) ~ as_factor(TP_SEXO)
+ NU_IDADE  + IN_ATIVIDADE_EXTRACURRICULAR + RESERVA_ESCOLA,
            data = compete.peso,
            weights = weight.cens,
            subset = failcode == 1) #evento 1: formar
# similar para os demais eventos, trocando failcode e status
exp(coef(m1.cox)) #estimativas dos riscos
```

```
# Grafico do residuo de schoenfeld + teste de proporcionalidade global

resid.sch<-cox.zph(m1.cox, transform="identity")
plot(resid.sch)
cox.zph(m1.cox) #teste

#alternativo via ggplot

teste_ph<-cox.zph(m1.cox, transform = "identity")

ggcoxzph(teste_ph, resid = TRUE, ggtheme = theme_bw(),
point.size =2, point.col = "darkblue")


# residuos martingale e deviance

rmarting<-resid(m1.cox, type="martingale");plot(rmarting); abline(h=0)

rdeviance<-resid(m1.cox, type="deviance");plot(rdeviance); abline(h=0);
```

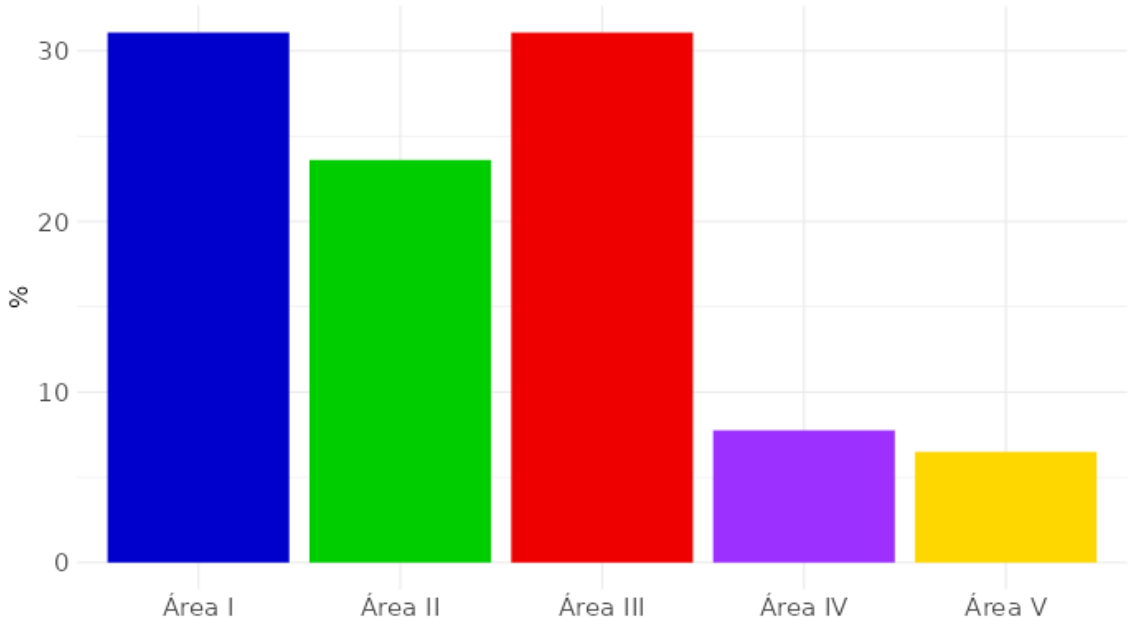
2: Modelo de Berger et al. (2018) para tempos discretos.

```
# Criando dummies dos eventos competivos

DADOS <- DADOS %>%
  mutate(censor1 = case_when(
    status_event  == 0 ~ 1, #cursando-censura
    TRUE ~ 0
  ))

DADOS <- DADOS %>%
  mutate(censor2 = case_when(
    status_event  == 1 ~ 1, #formado
    TRUE ~ 0
  ))

DADOS <- DADOS %>%
  mutate(censor3 = case_when(
    status_event  == 2 ~ 1,  #evadido
    TRUE ~ 0
  ))
```

```
DADOS <- DADOS %>%
  mutate(censor4 = case_when(
    status_event  == 3 ~ 1,  #retido
    TRUE ~ 0
  ))

DADOS <- DADOS %>%
  mutate(censor5 = case_when(
    status_event  == 4 ~ 1,  #matricula trancada
    TRUE ~ 0
  ))

# Construindo matriz esparsa

DadosDurLong1 <- discSurv::dataLongSubDist(dataShort=DADOS,
                          timeColumn="TEMPOS",
                         eventColumns=c( "censor2",
                         "censor3",
                         "censor4", "censor5"),
                        eventFocus="censor2") #formado

# Ajustando o modelo para tempos discretos

m1.glm <- glm(y ~ timeInt +  as_factor(TP_SEXO) +
NU_IDADE + RESERVA_ESCOLA + IN_ATIVIDADE_EXTRACURRICULAR ,
           family = binomial(link = "logit"),
           weights = DadosDurLong1$subDistWeights,
           control = list(maxit = 100),
           data=DadosDurLong1); summary(m1.glm)
exp(coef(m1.glm)) #estimativas do risco
```

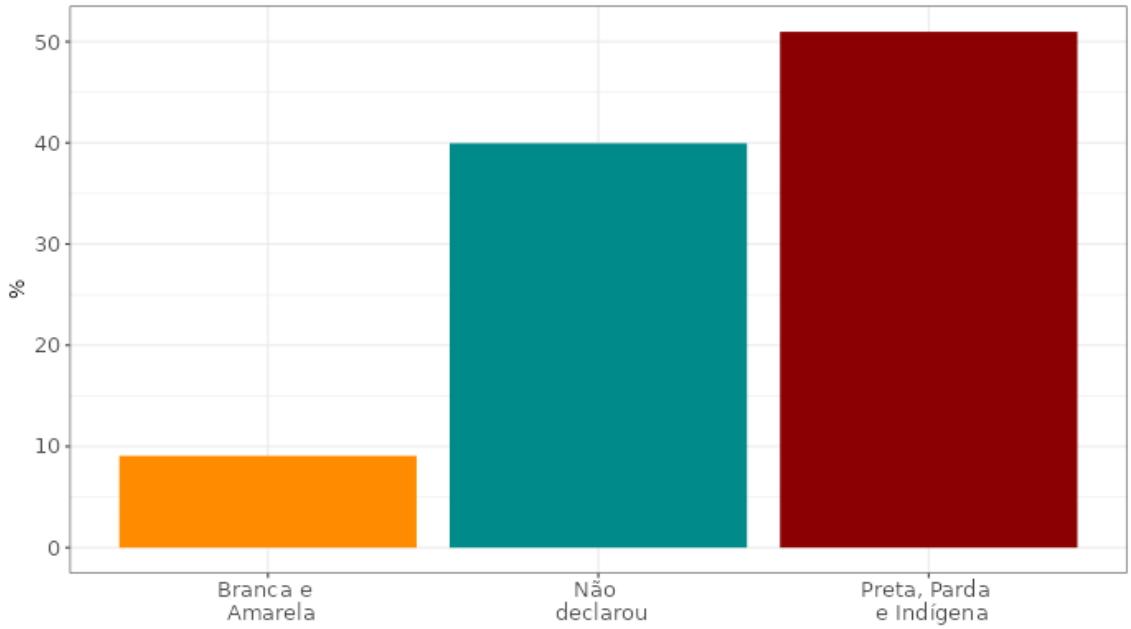
APÊNDICE C – RESULTADOS PRELIMINARES PARA TODOS OS CURSOS DE GRADUAÇÃO DA UFBA

Figura 14 – Porcentagem de estudantes da UFBA por área do curso a qual pertencem - CENSUP 2019.



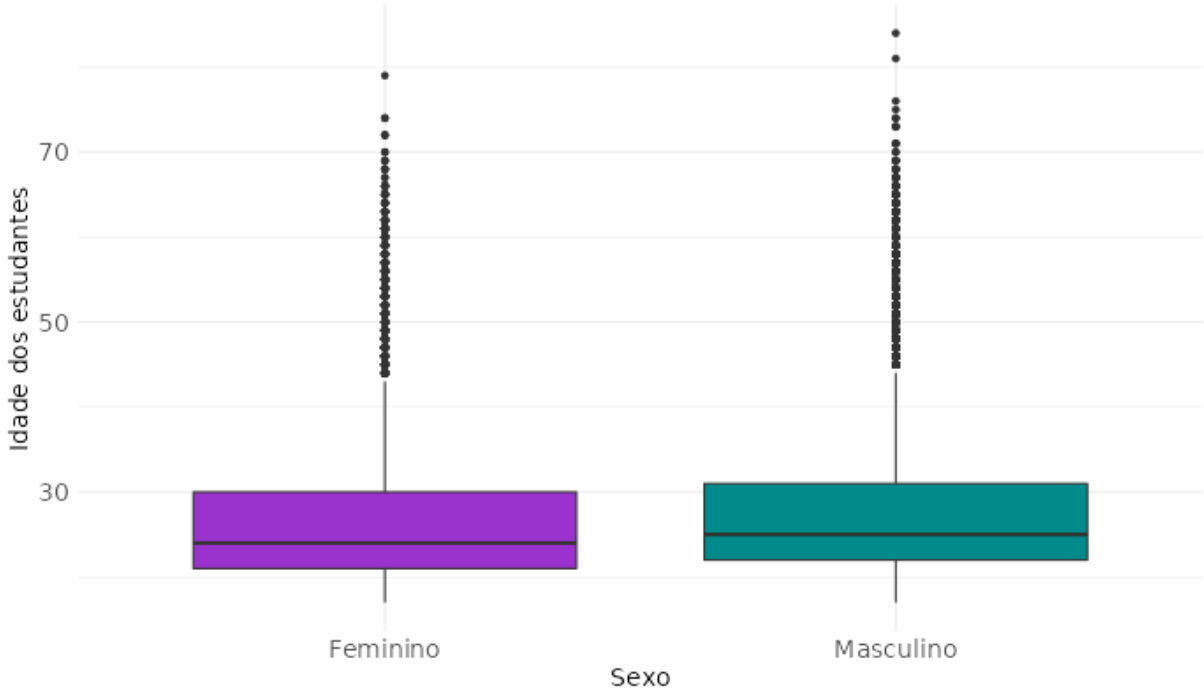
Fonte: INEP (2019c).

Figura 15 – Porcentagem de estudantes da UFBA por cor/raça - CENSUP 2019.



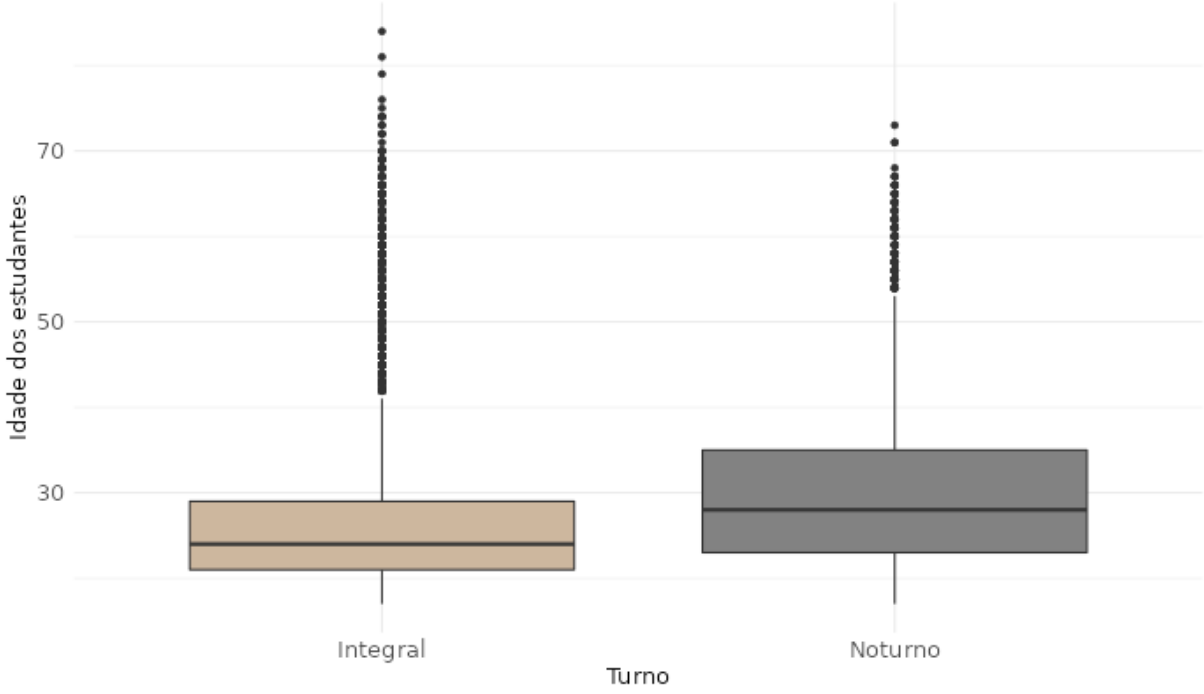
Fonte: INEP (2019c).

Figura 16 – Idade dos estudantes da UFBA segundo o sexo - CENSUP 2019.



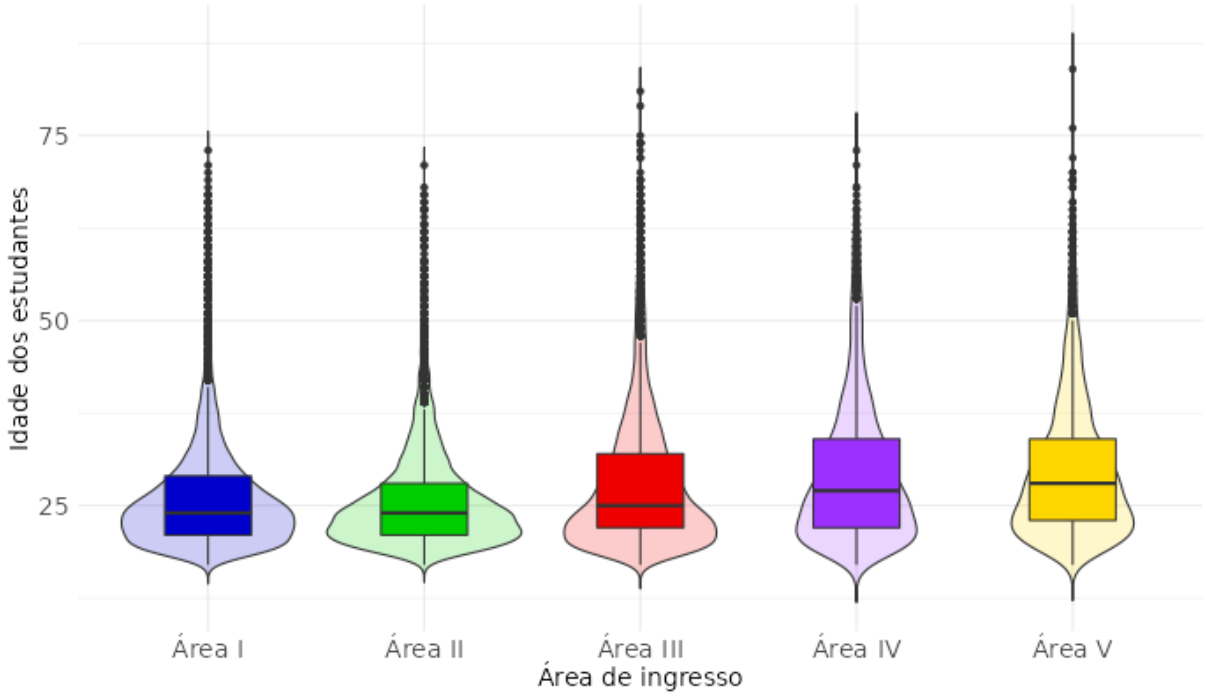
Fonte: INEP (2019c).

Figura 17 – Idade dos estudantes da UFBA segundo o turno do curso - CENSUP 2019.



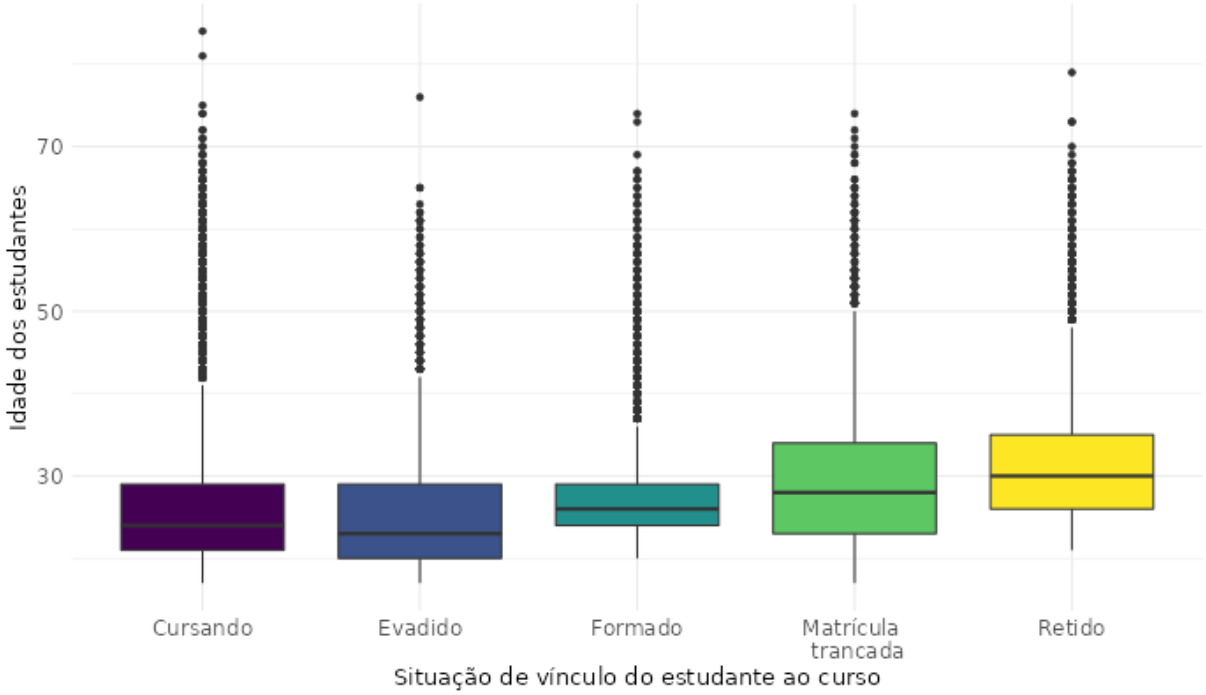
Fonte: INEP (2019c).

Figura 18 – Idade dos estudantes da UFBA segundo a área do curso - CENSUP 2019.



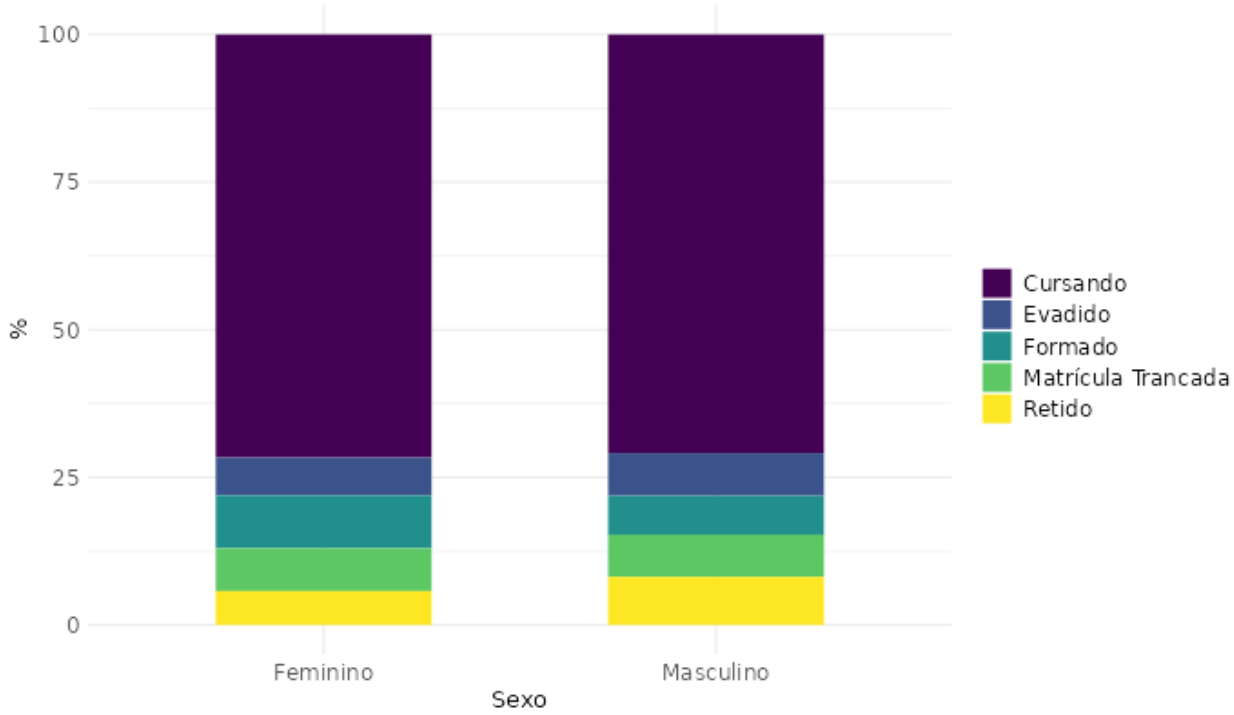
Fonte: INEP (2019c).

Figura 19 – Idade dos estudantes da UFBA segundo a situação de vínculo - CENSUP 2019.



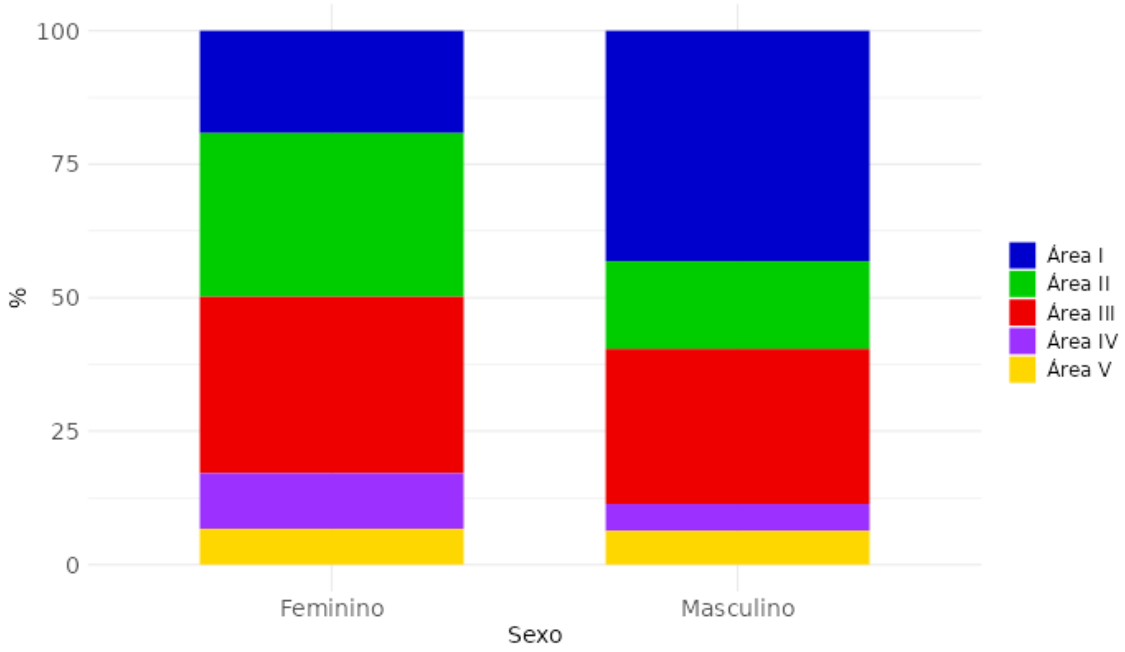
Fonte: INEP (2019c).

Figura 20 – Situação de vínculo dos estudantes da UFBA segundo o sexo - CENSUP 2019.



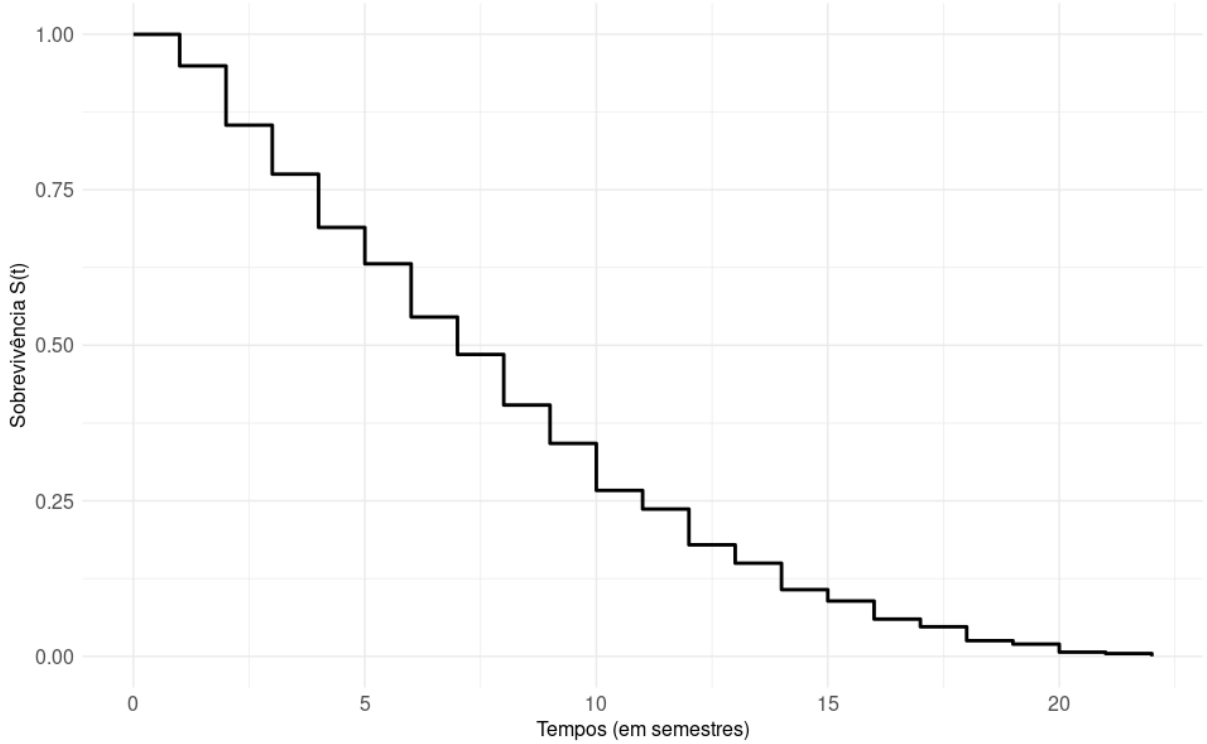
Fonte: INEP (2019c).

Figura 21 – Sexo dos estudantes da UFBA e área do curso - CENSUP 2019.



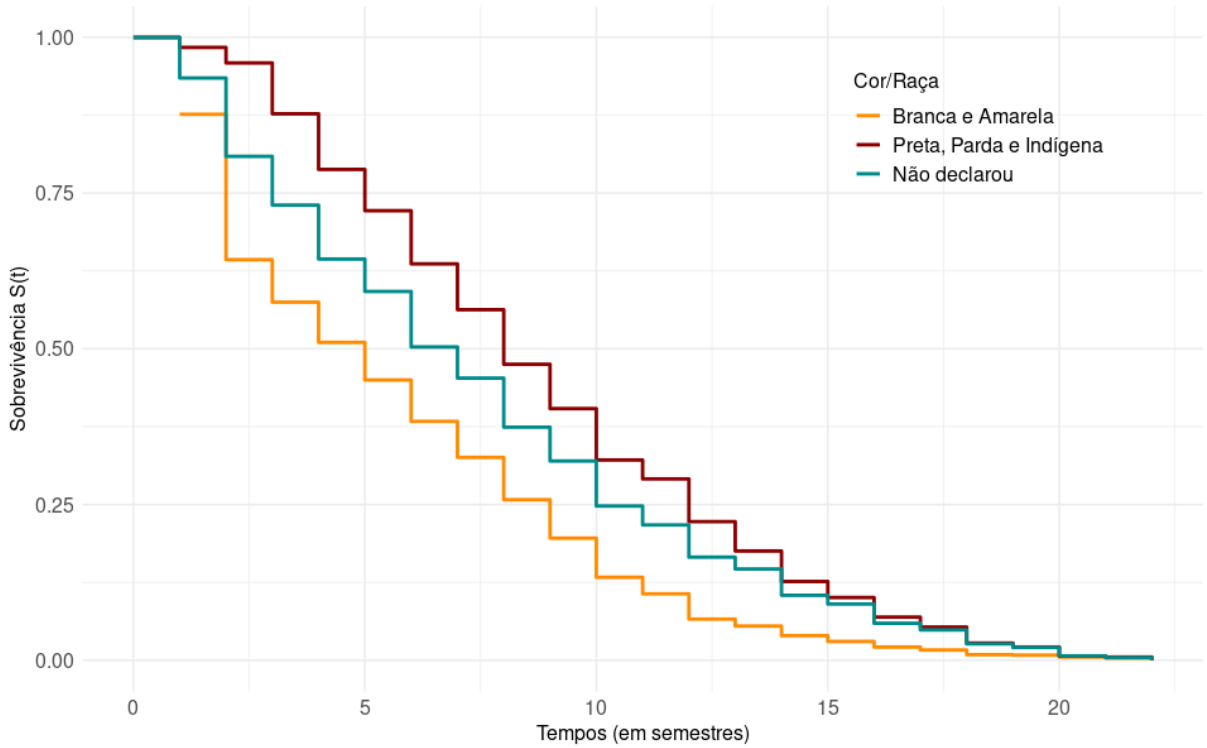
Fonte: INEP (2019c).

Figura 22 – Curva de Kaplan-Meier para o tempo de permanência (até formatura) dos estudantes da UFBA - CENSUP 2019.



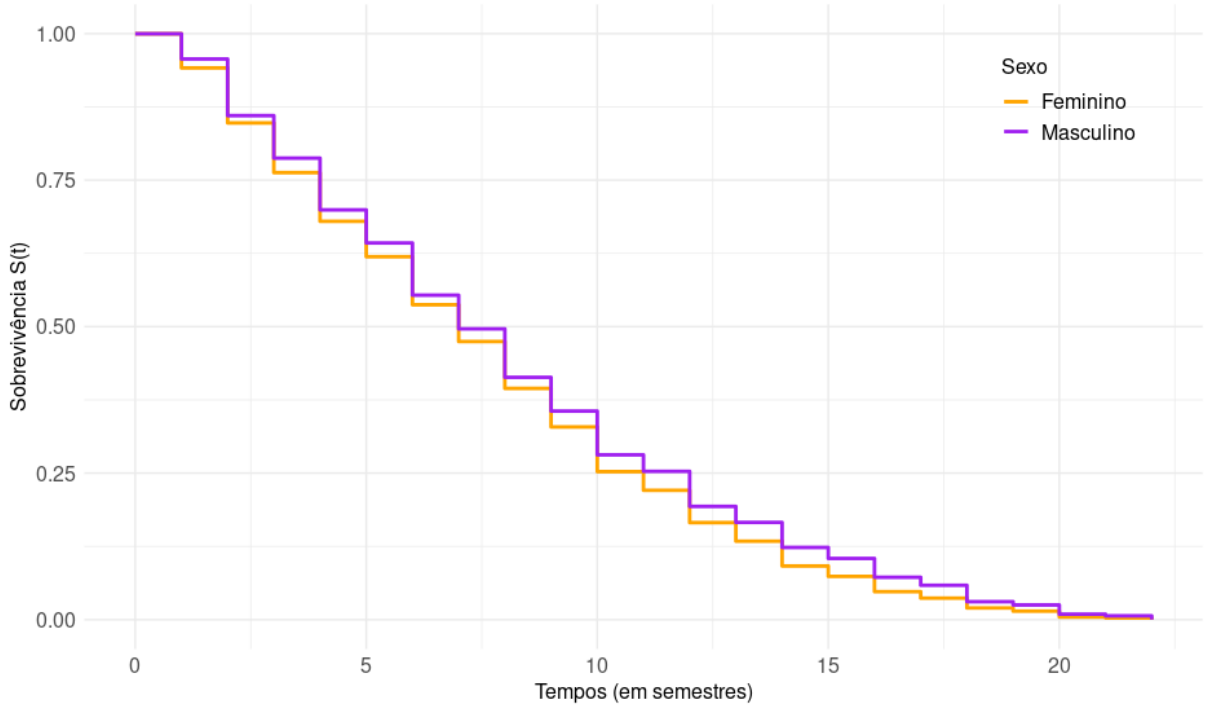
Fonte: INEP (2019c).

Figura 23 – Curva de Kaplan-Meier para o tempo de permanência (até formatura) dos estudantes da UFBA segundo a Cor/Raça - CENSUP 2019.



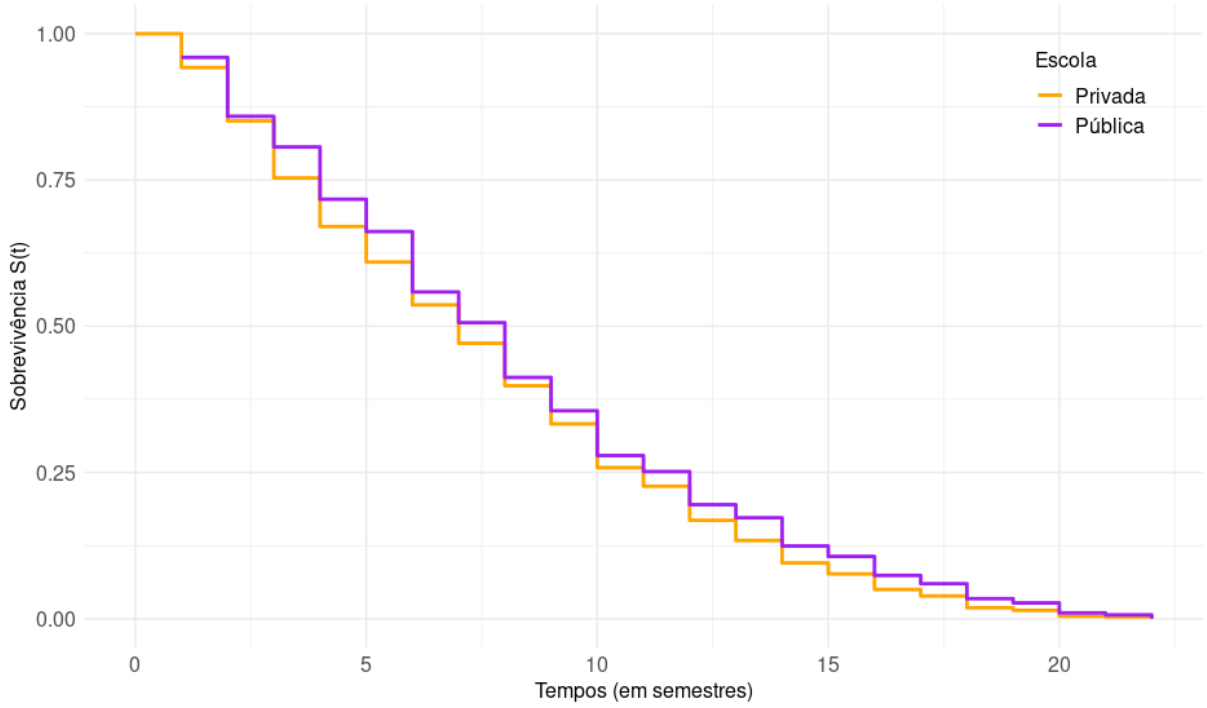
Fonte: INEP (2019c).

Figura 24 – Curva de Kaplan-Meier para o tempo de permanência (até formatura) dos estudantes da UFBA segundo o sexo - CENSUP 2019.



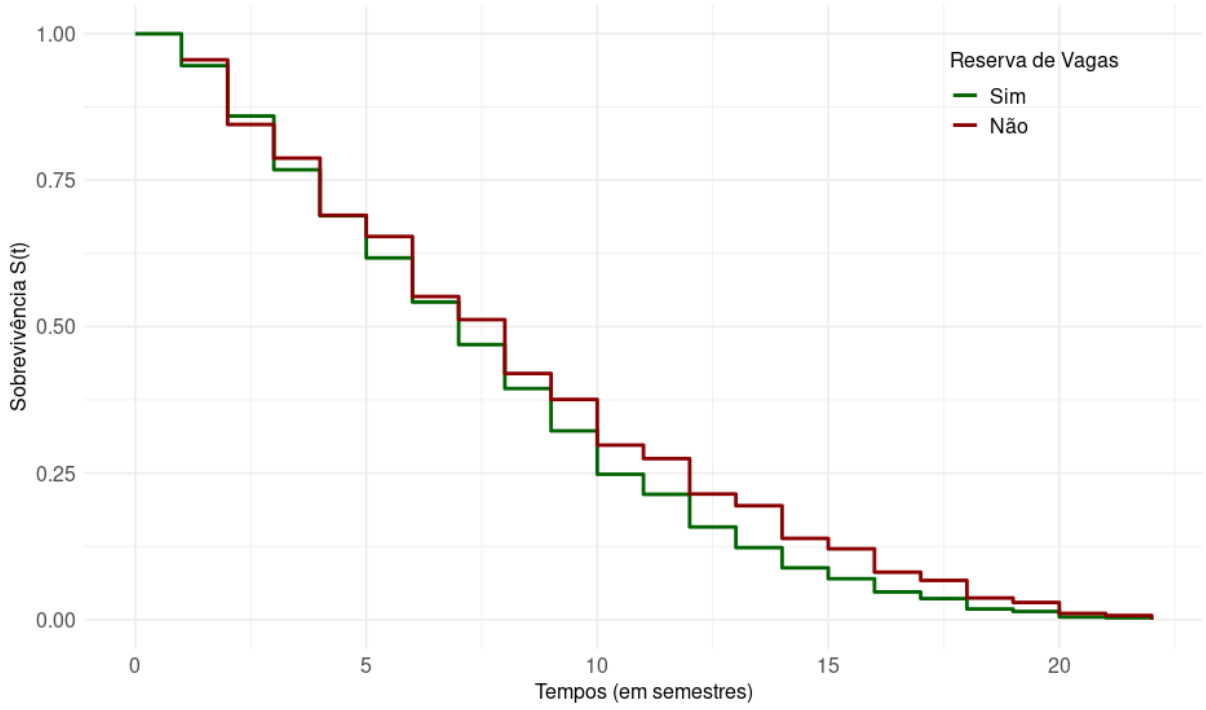
Fonte: INEP (2019c).

Figura 25 – Curva de Kaplan-Meier para o tempo de permanência (até formatura) dos estudantes da UFBA segundo escola de conclusão do Ensino Médio - CENSUP 2019.



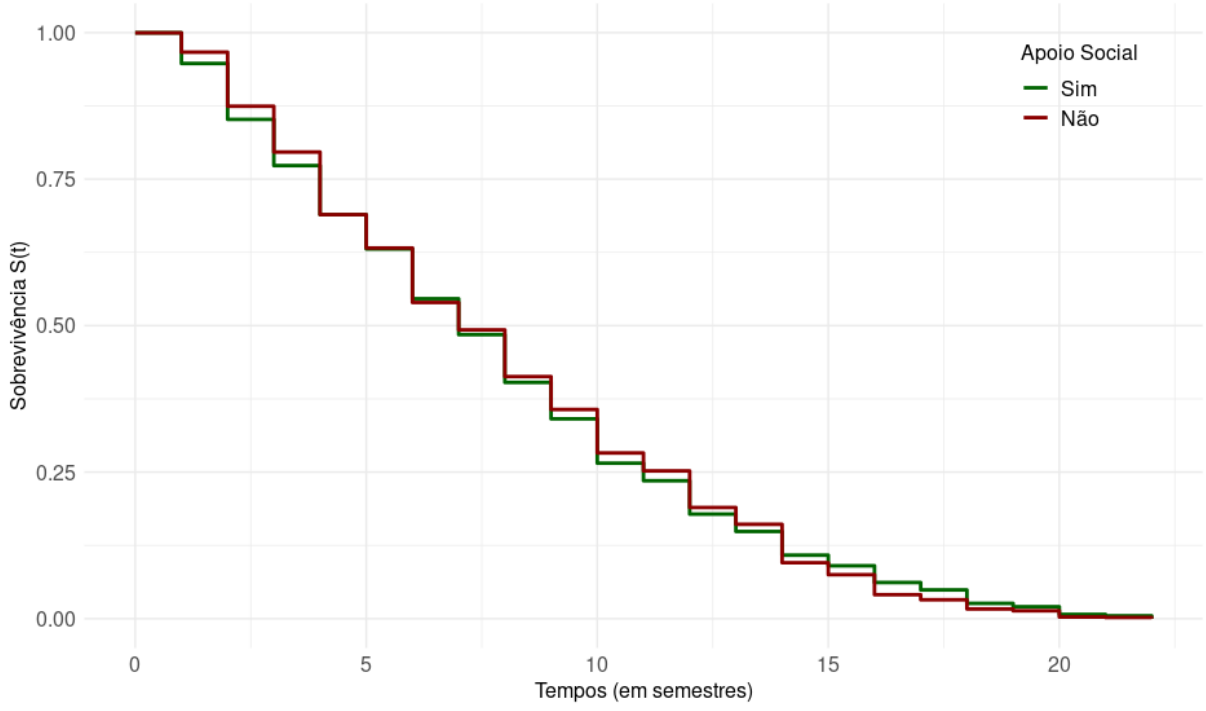
Fonte: INEP (2019c).

Figura 26 – Curva de Kaplan-Meier para o tempo de permanência (até formatura) dos estudantes da UFBA segundo ser cotista ou não - CENSUP 2019.



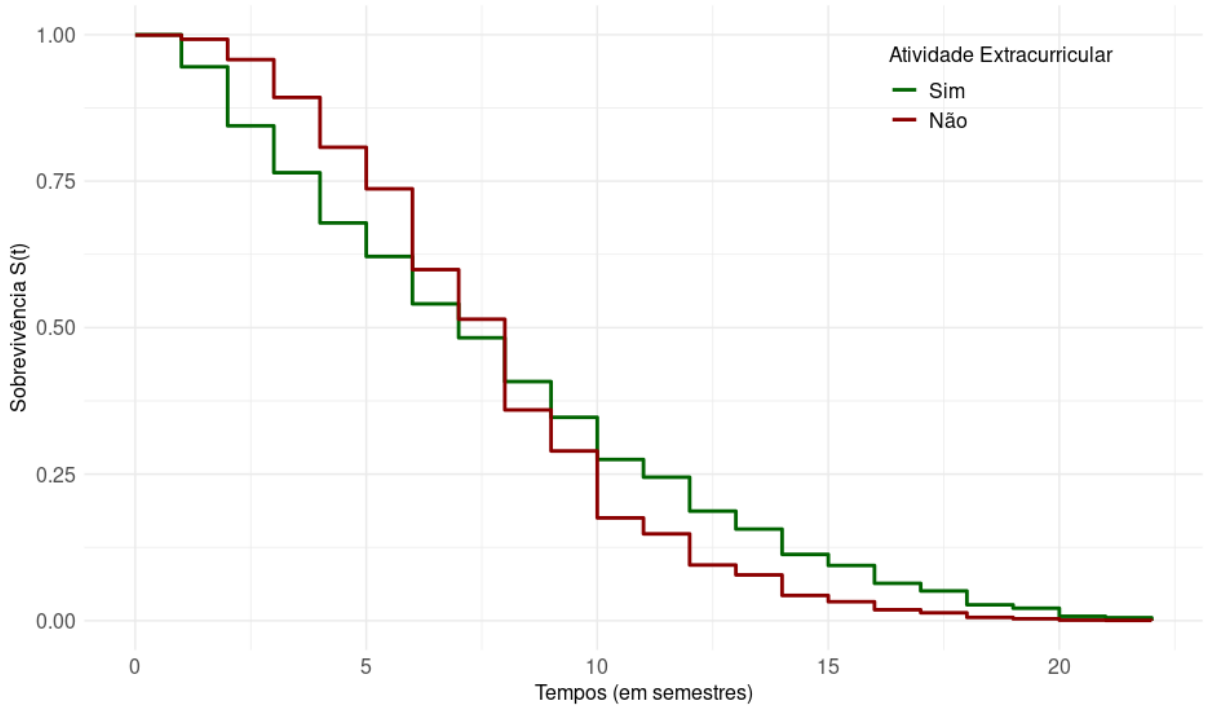
Fonte: INEP (2019c).

Figura 27 – Curva de Kaplan-Meier para o tempo de permanência (até formatura) dos estudantes da UFBA segundo possuir ou não apoio social - CENSUP 2019.



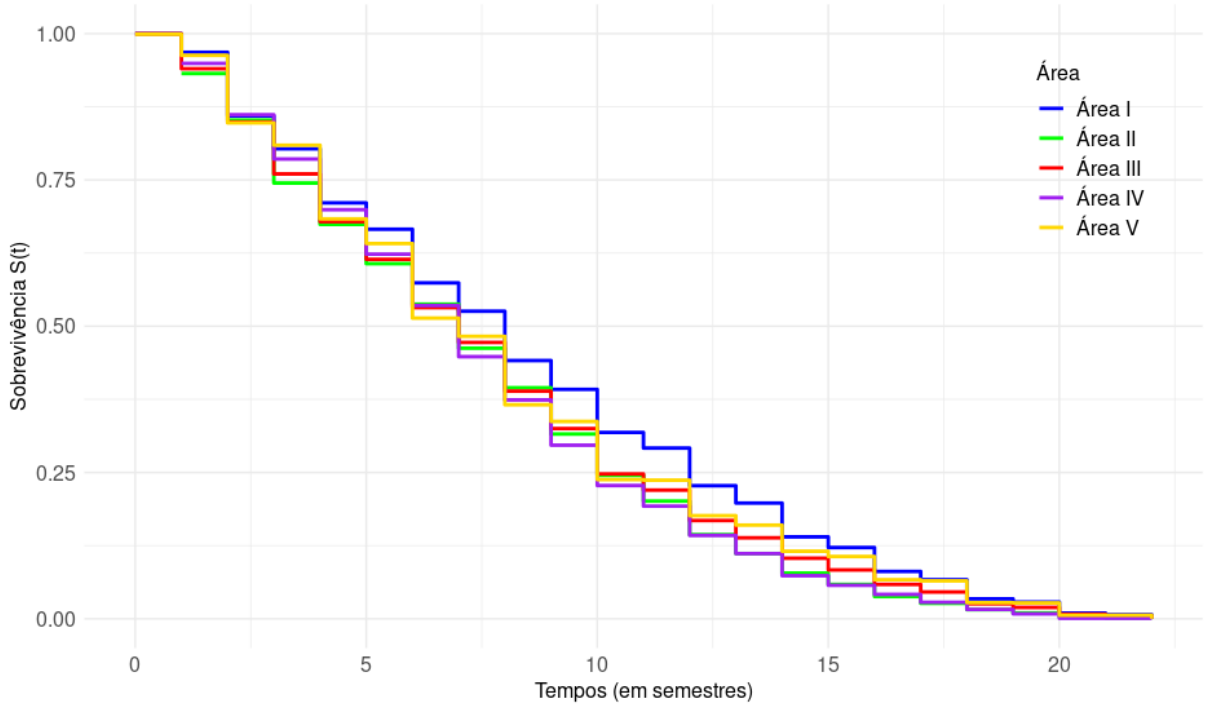
Fonte: INEP (2019c).

Figura 28 – Curva de Kaplan-Meier para o tempo de permanência (até formatura) dos estudantes da UFBA segundo realizar ou não alguma atividade extracurricular - CENSUP 2019.



Fonte: INEP (2019c).

Figura 29 – Curva de Kaplan-Meier para o tempo de permanência (até formatura) dos estudantes da UFBA segundo a área do curso - CENSUP 2019.



Fonte: INEP (2019c).