Revista Brasileira de Ensino de Física, v. 34, n. 1, 1403 (2012)
www.sbfisica.org.br

Análise de sobrevivência aplicada ao estudo do fluxo escolar nos cursos de graduação em física: um exemplo de uma universidade brasileira

(Survival analysis applied to student flow in undergraduate Physics courses: an example from a Brazilian university)

Paulo Lima Junior[1], Fernando Lang da Silveira e Fernanda Ostermann

Instituto de Física, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil
Recebido em 4/5/2011; Aceito em 12/10/2011; Publicado em 2/3/2011

A análise de sobrevivência é um método estatístico muito utilizado nas ciências da saúde, possuindo aplicações em outras áreas do conhecimento. Seu objeto de estudo é o tempo entre eventos, geralmente chamado "tempo de vida". Neste trabalho, investiga-se a possibilidade de aplicação da análise de sobrevivência ao estudo do tempo de permanência de estudantes diplomados, evadidos e desligados de um curso de graduação em física. Como resultado, percebe-se que este método é adequado, permitindo descrever quando a evasão e a diplomação ocorrem e quais fatores estão relacionados à permanência prolongada no curso. O objetivo deste trabalho é o de fazer uma breve introdução ao tema, exemplificando sua aplicação ao estudo do fluxo escolar de alunos de física de uma universidade brasileira. Acreditamos que este trabalho representa uma contribuição metodológica relevante para pesquisadores interessados no tema da evasão e da retenção nos cursos de formação científica na medida em que a análise de sobrevivência constitui um instrumento eficiente de descrição e compreensão do fluxo escolar.
Palavras-chave: análise de sobrevivência, fluxo escolar, evasão, retenção.

Survival analysis is a statistical method widely used in health sciences with applications in other fields. Its object of study is the time between events, usually called "survival time". This paper investigates the possibility of applying survival analysis to study students' retention, dropout and degree in an undergraduate Physics course. As a result, it is possible to notice that survival analysis is a suitable method, enabling one to describe when dropouts and degrees are more likely to occur and what factors are related to retention in this course. The goal of this work is to make a brief introduction to survival analysis, illustrating its application to the study of students flow from a Physics course in a Brazilian university. We believe this paper is an actual methodological contribution to researchers interested in the topic of dropout and retention in science education since survival analysis is an efficient tool for describing and understanding student flow.
Keywords: survival analysis, student flow, dropout, retention.

1. Introdução

A análise de sobrevivência é um conjunto de métodos estatísticos que remontam a meados do século XX, mas foram desenvolvidos e se popularizaram majoritariamente em torno da década de 80 [1]. Apesar de estarem historicamente relacionados à pesquisa em epidemiologia e em clínica médica, os métodos da análise de sobrevivência são adequados à abordagem de diversas questões importantes em engenharia, sociologia, psicologia e educação [2].

O objeto de estudo da análise de sobrevivência é o tempo entre eventos, por exemplo: o tempo do diagnóstico à morte de um paciente, o tempo da remissão à recidiva de uma doença, o tempo da venda de um automóvel até seu primeiro defeito mecânico, o tempo da soltura de um preso à sua re-incidência no crime, o tempo do ingresso em um curso de graduação ao desligamento, evasão ou diplomação. Do ponto de vista estatístico, todas essas situações podem ser abordadas com as mesmas ferramentas.

O fluxo escolar é um conceito amplo que compreende diversas características das trajetórias estudantis dentro das instituições de ensino e está entre os objetos mais tradicionais da pesquisa educacional. Seus aspectos mais destacados são a evasão – que consiste da desistência do curso pelo discente – e a retenção – que consiste da permanência prolongada no curso. Desses, a evasão tem recebido maior destaque recentemente tanto nas políticas federais brasileiras para a educação superior [3] quanto na pesquisa em educação

[1]E-mail: paulolima@ufrgs.br.

científica [4–6]. Tanto a evasão quanto a retenção podem representar prejuízos para o estudante e para a instituição de ensino – que é pressionada pelos seus financiadores a produzir mais egressos em períodos cada vez menores. Com efeito, a questão do fluxo escolar ocupa um papel destacado no debate sobre a qualidade do ensino superior, em geral, e da educação científica superior, em particular.

Na literatura, são encontrados diversos fatores relacionados ao fluxo escolar de graduação, por exemplo [5, 7, 8]: desempenho no vestibular, sucesso acadêmico, orientação vocacional prévia, adequação do trabalho ao estudo, relações de gênero. Como cada instituição possui suas particularidades, é importante que as instituições de educação científica superior elaborem seus próprios indicativos, buscando caracterizar quais dos fatores apontados pela literatura são mais relevantes em seu contexto e elaborando políticas institucionais eficazes de combate à evasão e à retenção.

A análise de sobrevivência permite estimar probabilidades relacionadas ao fluxo escolar e testar sua dependência com diversos fatores. Por se tratar de uma técnica estatística, a análise de sobrevivência impõe poucos limites ao pesquisador que a utiliza, deixando-o relativamente livre para escolher sua orientação teórica e as variáveis potencialmente relevantes no contexto da sua pesquisa. Neste trabalho, utilizaremos dados longitudinais retirados do registro acadêmico dos estudantes de física da Universidade Federal do Rio Grande do Sul (UFRGS), uma universidade federal brasileira, com o objetivo de ilustrar o uso da análise de sobrevivência no contexto da pesquisa sobre o fluxo escolar de graduação. Assim, nesta pesquisa, perguntamos: De que maneira a análise de sobrevivência pode contribuir para a descrição e melhor compreensão do fluxo escolar de graduação?

2. Tempos de vida e observações censuradas

O ponto de partida de qualquer análise de sobrevivência é um conjunto de medidas de tempo entre eventos, chamadas tempo de vida ou tempo de falha. Considere, por exemplo, um estudo clínico da sobrevida de pacientes com alguma doença grave. Nesse caso, é possível tomar por tempo de vida o intervalo que se estende do diagnóstico da doença à morte do paciente. Por outro lado, na análise do fluxo escolar de graduação, pode ser considerado que o tempo de vida se estende do ingresso no curso à conclusão do registro do estudante – por desligamento, evasão ou diplomação. Em qualquer caso, é preciso especificar claramente os eventos que definem os intervalos de tempo.

Uma das características mais importantes dos dados de sobrevivência é a presença intervalos de tempo incompletos, que não foram concluídos pelo evento de interesse. Tais intervalos são chamados observações censuradas, carregam informações importantes e não devem ser descartados.

Usualmente, estudos longitudinais para coleta de dados de sobrevivência precisam ser encerrados antes que todos os pacientes experimentem o evento terminal. Por exemplo, no acompanhamento de um grupo de pacientes com doença grave, pode ser necessário encerrar a pesquisa antes que todos tenham falecido. Nesse caso, os pacientes que ainda estão vivos nos fornecem uma informação mais limitada, pois é possível saber somente que seus tempos de vida verdadeiros T_{VERD} são superiores aos tempos de vida observados T_{OBS}.

$$T_{VERD} > T_{OBS}. \tag{1}$$

De maneira análoga, ao consultar o registro discente da universidade, são encontradas informações sobre estudantes graduados, evadidos ou desligados ao mesmo tempo em que são identificados estudantes com matrícula ativa ou trancada – que ainda podem se tornar diplomados, evadidos ou desligados. Assim, a informação que temos sobre esses estudantes é semelhante àquela que possuímos sobre os pacientes que ainda estão vivos ao final do estudo: nos dois casos, o evento terminal ainda não ocorreu e o tempo de vida verdadeiro é superior ao tempo de vida observado.

Na literatura, encontra-se uma pequena variedade de mecanismos capazes de tratar as censuras nas observações de tempo de vida [9]. Desses, os casos mais comuns são: (1) censura tipo I, que ocorre nos estudos que começam com tempo pré-estabelecido para serem concluídos; (2) censura tipo II, em que a decisão por concluir o estudo está condicionada à ocorrência de uma quantidade pré-estabelecida de observações não censuradas.

À primeira vista, o procedimento mais correto seria descartar os casos censurados e manter no corpus somente os indivíduos para os quais o evento terminal foi realmente observado. No entanto, além de reduzir o poder dos testes estatísticos devido à perda de muitos graus de liberdade, a exclusão das observações censuradas introduz viés nas funções de sobrevivência [10]. Enfim, uma das maiores contribuições da análise de sobrevivência é permitir que dados censurados sejam levados em consideração.

3. Modelando o fluxo escolar

Assim como todas as instituições de ensino superior, a UFRGS mantém bancos de dados com várias informações sobre seus estudantes. O conjunto dessas informações, chamadas registros acadêmicos, é fundamental tanto para a avaliação da qualidade do ensino oferecido pela universidade quanto para a discussão de políticas institucionais de combate à evasão.

Os dados utilizados neste trabalho para ilustrar o uso da análise de sobrevivência foram fornecidos pelo

Departamento de Controle e Registro Acadêmico (DECORDI/ UFRGS) e consistem de informações sobre os estudantes que ingressaram por meio de vestibular no curso de física diurno (bacharelado e licenciatura) no período de 1995 a 2009. Como o primeiro ingresso para a licenciatura noturna ocorreu no ano 2000, optou-se por concentrar a análise sobre os estudantes das habilitações diurnas.

Sob essas condições, foram localizados 1447 registros de estudantes. As informações obtidas com respeito a esses estudantes são: (1) Tempo de permanência no curso; (2) Situação do registro (ativo, afastado, diplomado, transferido, evadido e outros); (3) Habilitação (bacharelado ou licenciatura); (4) Pontuação obtida no concurso vestibular à universidade (um escore composto pelos escores ponderados em nove provas, denominado "argumento de concorrência"); (5) Sexo.

Um dos primeiros passos de qualquer análise estatística do fluxo escolar é propor um modelo inicial que, reduzindo o detalhamento com que se descrevem as trajetórias dos estudantes, põe em destaque seus aspectos mais importantes. Há na literatura uma pequena variedade de modelos utilizados em análises do fluxo escolar [11]. O modelo proposto para este trabalho pode ser visualizado no diagrama da Fig. 1.

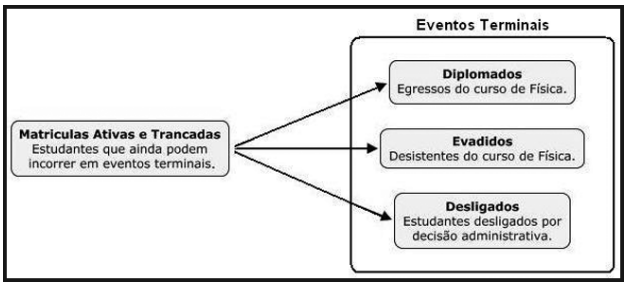

Figura 1 - Modelo proposto para descrever o fluxo escolar de graduação.

Estudantes com matrícula ativa ou trancada são aqueles que, no futuro, incorrerão em algum dos eventos terminais. Portanto, seus tempos de vida serão tratados como observações censuradas. Estudantes diplomados são os egressos do curso. Estudantes evadidos são todos aqueles que desistiram do curso. Nesta categoria, estão os alunos que mudaram de curso por meio de vestibular ou transferência interna além daqueles que simplesmente abandonaram a vaga. Enfim, estudantes são ditos desligados quanto perdem sua vaga por decisão administrativa. Esse desligamento ocorre somente em casos extremos e encontra-se regulamentado na legislação da universidade [12].

4. O fluxo escolar em tabelas de contingência

Após a realização de testes de consistência das informações, elas foram apropriadas em planilhas de dados do pacote estatístico SPSS – versão 16.0. Porém, antes de proceder com a análise de sobrevivência, foram elaboradas tabelas de contingência (ver Tabela 1), visando descrever a distribuição dos estudantes de física nas categorias do modelo proposto para o fluxo escolar. Com o objetivo de situar o fluxo escolar do curso de física na área de matemática, ciências e engenharia, os percentuais de estudantes evadidos e diplomados no período de 1995 a 2009 em diversos cursos de graduação dessa área são apresentados na Tabela 2.

Nessas tabelas, é possível perceber que a quantidade de estudantes desligados é pouco expressiva se comparada com os diplomados e com os evadidos. Por essa razão, a análise se concentrará sobre os eventos "diplomação" e "evasão". Nas tabelas de contingência, a ocorrência de diplomados no curso de física é baixa (16,5% dos registros). Descontando os casos de matrícula ativa ou trancada, percebe-se que o curso de física da UFRGS tem formado, em média, 24,1% dos seus ingressantes enquanto 72,7% evadem.

Abaixo das tabelas foram registrados coeficientes de contingência (representado pela letra C). Tais coeficientes são uma medida de associação ou de relação entre as variáveis em questão [13]. Quando ele é nulo, significa que não há relação entre as variáveis; quanto mais próximo da unidade, tanto mais intensa é a relação entre essas variáveis.

Tabela 1 - Tabelas de contingência situação *vs.* habilitação e situação *vs.* sexo.

		Habilitação[1]		Sexo[2]	
		Bacharelado	Licenciatura	Masculino	Feminino
Situação do Registro	Matricula Ativa ou Trancada	31,4%	32,2%	30,7%	35,0%
	Diplomado	16,6%	16,4%	16,3%	17,2%
	Evadido	50,0%	48,9%	50,4%	47,2%
	Desligado	2,0%	2,5%	2,5%	0,7%
Total de Registros por coluna (100%)		1093	354	1144	303

[1].- $C_1 = 0{,}018$; $p = 0{,}922$. [2].- $C_2 = 0{,}064$; $p = 0{,}115$.

Tabela 2 - Tabela de contingência curso *vs.* situação do registro.

		Situação do fegistro			Nº de registros por curso (100%)
		Diplomado	Evadido	Desligado	
Curso	Física	24,1%	72,7%	3,1%	990
	Biologia	53,3%	45,6%	1,0%	1343
	Engenharia Civil	58,2%	37,8%	3,9%	1358
	Engenharia de Materiais	51,7%	45,1%	3,1%	288
	Engenharia de Minas	21,6%	73,4%	5,0%	436
	Engenharia Elétrica	50,4%	47,2%	2,4%	922
	Engenharia Mecânica	45,5%	50,6%	3,9%	1068
	Engenharia Metalúrgica	34,1%	59,6%	6,3%	495
	Engenharia Química	55,5%	41,6%	2,9%	652
	Matemática	25,9%	69,9%	4,2%	1363
	Química	21,9%	74,5%	3,5%	734

[1].- Dados referentes aos cursos da UFRGS no período de 1995 a 2009. [2].- C = 0,282; p < 0,001.

Da Tabela 1, é possível perceber que, embora existam mais matrículas no bacharelado (1093 contra 354 da licenciatura, portanto 76% do total), os estudantes tendem a evadir e se diplomar na mesma proporção nas duas habilitações. Com efeito, o coeficiente de contingência C_1 confirma que a situação do registro e a habilitação não estão relacionadas, ou seja, a distribuição dos percentuais de alunos nas diversas situações de registro é semelhante no bacharelado e na licenciatura.

Ainda nessa tabela, é possível perceber que há mais homens ingressando no curso de física (1144 registros acadêmicos do sexo masculino contra 303 do sexo feminino, ou seja, 79% são homens). Entretanto, o sexo não está relacionado com a situação do registro acadêmico conforme atesta o valor quase nulo do coeficiente de contingência C_2.

Da Tabela 2, é possível perceber que a variável situação está relacionada ao curso, pois os percentuais de evadidos e de diplomados variam através dos diferentes cursos. Isto também fica evidenciado no fato de o coeficiente de contingência resultar em 0,28 (estatisticamente significativo em nível $p < 0{,}001$). A proporção de evadidos é semelhante nos cursos de física, matemática e química. Por outro lado, nos cursos de engenharia, predomina um percentual de diplomação significativamente mais elevado. A saber, nas estatísticas nacionais, a física aparece ao lado da matemática entre os cursos superiores de maior evasão anual [8].

A representação da associação entre as variáveis curso e situação pode ser obtida por meio de uma análise de correspondência [14]. Neste procedimento, as categorias das variáveis são representadas por pontos em um gráfico. Quanto mais próximos estão os pontos, mais similares são as categorias entre si. Desta forma, a representação gráfica possibilita compreender a relação entre as duas variáveis. O resultado da análise de correspondência encontra-se no gráfico apresentado na Fig 2.

Conforme se observa na Fig. 2, os pontos que representam os cursos de matemática, química e física estão mais próximos da categoria "evadido", identificando assim uma maior incidência percentual (de cerca de 70%) desta situação. Os pontos correspondentes aos cursos de engenharia de um modo geral (exceto engenharia de minas e engenharia metalúrgica) e Biologia situam-se próximos à categoria "diplomado", identificando a maior incidência de diplomação (cerca de 50%) nesses cursos. Finalmente o ponto que representa "desligado" está muito afastado da maioria dos cursos, indicando a baixa incidência dessa categoria.

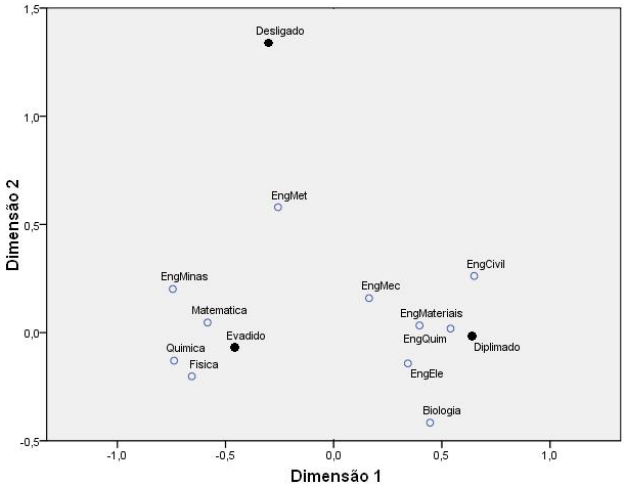

Figura 2 - Mapeamento bidimensional resultante da análise de correspondência do curso *vs.* situação do registro.

Como foi possível perceber, as tabelas de contingência, aliadas à análise de correspondência, permitem analisar as frequências relativas ou percentuais de ocorrência dos eventos do fluxo escolar em cada categoria de indivíduos (homens e mulheres, licenciatura e bacharelado). Esses procedimentos também foram utilizados por Arruda e cols. [4] para melhor compreender as relações entre as diversas situações dos alunos através dos cursos. Entretanto, é importante destacar que essa técnica de análise não incorpora uma informação relevante sobre o fluxo escolar: sua evolução temporal.

A análise de sobrevivência leva em consideração o tempo até a ocorrência do evento de interesse permitindo adicionalmente compreender as influências de outras variáveis (nominais e/ou intervalares) para essa evolução temporal. Assim, munidos dessa ferramenta, será possível avaliar, por exemplo, se os alunos que obtém diploma de licenciatura ficam retidos por mais

tempo no curso se comparados com os estudantes do bacharelado. Também será possível avaliar se a pontuação no concurso vestibular tem valor preditivo sobre a situação do aluno ao longo do tempo. Nas próximas seções apresentaremos alguns aspectos da análise de sobrevivência, exemplificando com o caso do fluxo escolar nos cursos de física da UFRGS.

5. Generalidades sobre a análise de sobrevivência

5.1. A descrição funcional do tempo de vida

Em análise de sobrevivência existem diversas maneiras de especificar como o tempo de vida se encontra distribuído em uma população de medidas. Chama-se função sobrevivência $S(t)$ a probabilidade de que um indivíduo possua tempo de vida maior que t. Ou seja, a probabilidade de que, decorrido um tempo t, esse indivíduo tenha sobrevivido ao evento terminal. De maneira semelhante, a distribuição do tempo de vida $F(t)$ é a probabilidade de que o evento terminal ocorra até o tempo t. Ou seja, a probabilidade de que esse indivíduo tenha experimentado algum evento terminal até o tempo t.

$$S(t) \equiv \Pr(T > t). \tag{2}$$

$$F(t) \equiv \Pr(T \leq t). \tag{3}$$

Em amostras numerosas, a função de sobrevivência S(t) pode ser pensada como a fração de sobreviventes em função do tempo. Da mesma maneira, a distribuição do tempo de vida $F(t)$ pode ser interpretada como a fração de indivíduos atingidos pelo evento terminal em função do tempo. Por construção, a distribuição do tempo de vida $F(t)$ deve ser uma função crescente com $F(t) + S(t) = 1$ em qualquer tempo t.

A densidade de evento $f(t)$ expressa a probabilidade de que um indivíduo qualquer da população seja atingido pelo evento terminal no intervalo de tempo $t \leq T < t + \Delta t$. A taxa de risco $h(t)$ expressa a probabilidade de que um indivíduo seja atingido pelo evento terminal com $t \leq T < t + \Delta t$ no universo daqueles que sobreviveram ao tempo t.

$$f(t)\Delta t \equiv \Pr(t \leq T < t + \Delta t) = F(t + \Delta t) - F(t) \rightarrow f(t) = \frac{dF}{dt}. \tag{4}$$

$$h(t)\Delta t \equiv \frac{\Pr(t \leq T < t + \Delta t)}{\Pr(T \geq t)} = \frac{f(t)}{S(t)}\Delta t \rightarrow h(t) = -\frac{d}{dt}\ln S(t). \tag{5}$$

Ainda que de maneiras diferentes, todas as funções acima são capazes de informar sobre o tempo de vida em um conjunto de medidas. Entretanto, de acordo com os objetivos do pesquisador e a técnica de análise utilizada, uma função pode se tornar preferível às outras.

5.2. Técnicas da análise de sobrevivência

De uma maneira geral, a análise de sobrevivência visa construir estimadores das funções usadas para especificar o tempo de vida, testando a dependência dessas funções com outras variáveis. As técnicas da análise de sobrevivência podem ser organizadas em três categorias [10]:

a) Técnicas não-paramétricas. Não fazem quase nenhuma restrição sobre a distribuição dos tempos de vida na população que gerou a amostra. Esses modelos são os mais flexíveis, porém menos poderosos. São também mais limitados porque não permitem testar o efeito de muitas variáveis ao mesmo tempo. A principal técnica não-paramétrica da análise de sobrevivência é o estimador de Kaplan-Meier.

b) Técnicas semi-paramétricas. Também são chamadas regressão de Cox. A principal vantagem dessa abordagem com relação aos modelos não-paramétricos é a possibilidade de testar a significância estatística de diversas co-variáveis (variáveis explicativas nominais e/ou interevalares) sobre a distribuição de tempos de vida.

c) Técnicas paramétricas. Também são chamadas modelos de tempo de vida acelerado. Elas definem com rigidez a forma funcional das distribuições de tempo de vida na população. Devido à sua falta de flexibilidade, essas técnicas são utilizadas com menos frequência.

Nas seções a seguir, os principais aspectos dos modelos não-paramétricos e semi-paramétricos serão discutidos de maneira sucinta, tomando como exemplo o caso do fluxo escolar de graduação do curso de física da UFRGS. Para uma compreensão mais detalhada das técnicas da análise de sobrevivência, recomenda-se a bibliografia deste trabalho.

6. Técnicas não-paramétricas da análise de sobrevivência

6.1. Definindo o estimador de Kaplan-Meier

O estimador de Kaplan-Meier (EKM), também chamado estimador limite-produto, é a técnica não-paramétrica mais utilizada em análise de sobrevivência. Intuitivamente, o EKM pode ser compreendido se levarmos em consideração que, para sobreviver a M intervalos de tempo, um indivíduo precisa ter sobrevivido a cada intervalo de tempo anterior [10]. Por exemplo, se um estudante sobreviveu ao evento "evasão" por 5 períodos letivos, isso significa ter sobrevivido ao evento

no primeiro período e no segundo período e no terceiro..., até o quinto período. Por essa razão, é possível construir um estimador a partir do produto das probabilidades de sobreviver a cada intervalo de tempo.

Suponha que uma amostra compreenda N medidas de tempo de vida e que, dentre essas medidas, haja somente k observações distintas e não censuradas. Considere que essas medidas distintas de tempo estejam dispostas em ordem crescente $t_1 < t_2 < t_3 < ... < t_k$. Sejam d_i o número de eventos terminais ocorridos em t_i e n_i o número de indivíduos que poderiam ser atingidos pelo evento terminal no tempo t_i. Assim, é possível definir formalmente o EKM por [10]

$$\hat{S}_{KM}(t) = \prod_{t_i=0}^{t} \left(1 - \frac{d_i}{n_i}\right). \tag{6}$$

O EKM pode ser calculado facilmente na maioria dos pacotes estatísticos comerciais que incorporam a análise de sobrevivência. Entretanto, conhecer sua definição formal é importante para compreender algumas propriedades desse estimador.

6.2. O EKM na descrição do fluxo escolar

Após a inserção dos registros acadêmicos no pacote estatístico SPSS, o EKM da distribuição do tempo de vida $F(t)$ foi calculado para os eventos "evasão" e "diplomação" nos cursos de física da UFRGS separadamente, usando a seguinte relação

$$\hat{F}_{KM}(t) = 1 - \hat{S}_{KM}(t). \tag{7}$$

O resultado desse procedimento pode ser visualizado nos gráficos representados em Figs. 3 e 4. Por convenção, os gráficos do EKM são representados com degraus para indicar os instantes do tempo em que ocorrem eventos terminais, e sinais (+) para indicar as observações censuradas. Como o tempo no registro acadêmico é contado em períodos de um semestre, os degraus em Figs. 3 e 4 aparecem somente nos múltiplos inteiros de meio ano.

A partir da Fig. 3, é possível perceber que a ocorrência de diplomações cresce após 3,5 anos de permanência no curso. A proporção de graduados por semestre (densidade de evento) é máxima no intervalo de 3,5 a 4,5 anos. Após esse intervalo, a probabilidade de que o estudante obtenha o grau é reduzida, tornando-se praticamente nula para os estudantes com 8 anos de permanência ou mais.

Na Fig. 4, nota-se que a evasão ocorre desde o primeiro semestre, atingindo uma saturação pouco antes de 10 anos, ou seja, após esse tempo, o risco de evadir é nulo. Embora esse risco seja relativamente mais intenso nos primeiros semestres, é possível observar que, após 4 anos de permanência, aproximadamente metade da evasão ainda está por ocorrer. Dessa maneira, fica evidente que grande parte dos estudantes de física decide abandonar o curso somente após vários anos de retenção.

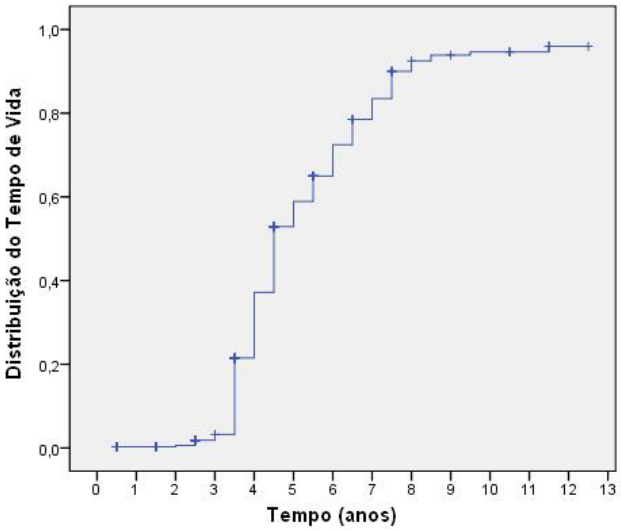

Figura 3 - EKM da distribuição do tempo de vida para o evento "diplomação" em função do tempo.

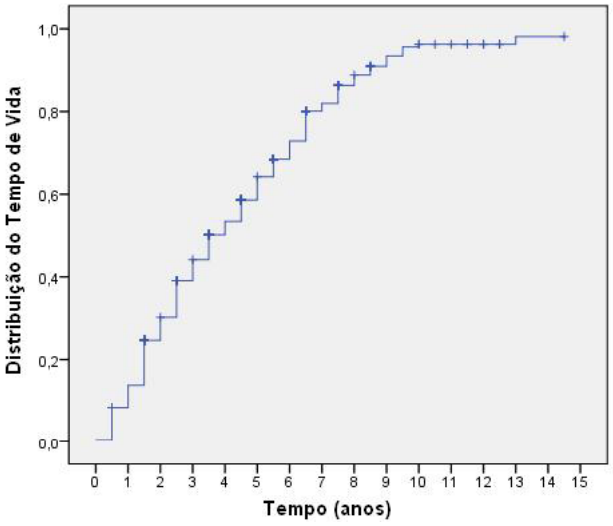

Figura 4 - EKM da distribuição do tempo de vida para o evento "evasão" em função do tempo.

6.3. Testes estatísticos não-paramétricos

Além servir à descrição do fluxo escolar, a abordagem não-paramétrica à análise de sobrevivência, utilizando o EKM, permite realizar testes de significância estatística para comparar categorias – por exemplo, homens e mulheres, licenciatura e bacharelado – em termos das características que a distribuição do tempo de vida $F(t)$ assume em cada grupo.

Os principais testes de significância estatística disponíveis nos pacotes comerciais são [9]: (1) Logrank, também chamado teste generalizado de Savage; (2) Breslow, também chamado teste generalizado de Wilcoxon; e (3) Tarone-Ware.

De maneira geral, esses testes consistem em calcular uma estatística ponderada que, sob a hipótese nula (hipótese de que os grupos de medidas são amostras originadas da mesma população), tem distribuição conhecida. Entretanto os testes possuem regras de atribuição de peso para melhor discriminar as curvas em determinadas etapas de sua evolução temporal. No teste Logrank, os pesos são atribuídos de maneira a enfatizar diferenças ao final das distribuições de tempo de vida. O teste de Breslow enfatiza diferenças no início da distribuição do tempo de vida. O teste de Tarone-Ware, por sua vez, foi desenvolvido com o objetivo de discriminar distribuições em fases intermediárias do tempo de vida.

Os gráficos representados nas Figs. 5 e 6 apresentam o EKM da distribuição do tempo de vida para os eventos "diplomação" e "evasão" nas categorias homens e mulheres. As Tabelas 3 e 4 apresentam o resultado dos testes de significância estatística em cada caso.

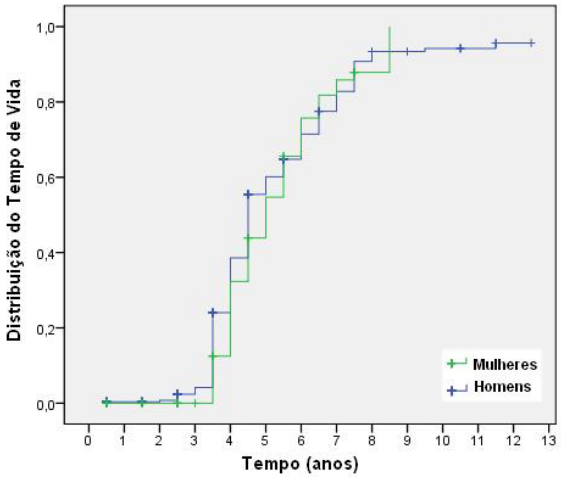

Figura 5 - Comparação das distribuições do tempo de vida em função do tempo para homens e mulheres. Evento terminal: "diplomação".

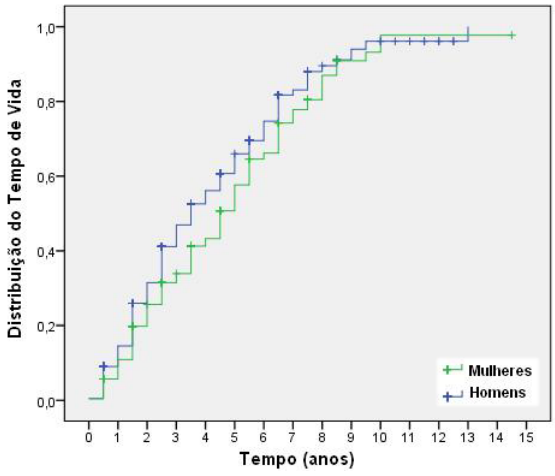

Figura 6 - Comparação das distribuições do tempo de vida em função do tempo para homens e mulheres. Evento terminal: "evasão".

Tabela 3 - Resultado da comparação das distribuições do tempo de vida. Evento terminal: "diplomação". Categorias: homens e mulheres.

	χ^2	Graus de liberdade	Significância
Logrank	0,235	1	0,628
Breslow	1,825	1	0,177
Tarone-Ware	0,997	1	0,318

Tabela 4 - Resultado da comparação das distribuições do tempo de vida. Evento terminal: "evasão". Categorias: homens e mulheres.

	χ^2	Graus de liberdade	Significância
Logrank	5,994	1	0,014
Breslow	7,397	1	0,007
Tarone-Ware	7,467	1	0,006

Na Fig. 5, as curvas estão sobrepostas e os testes estatísticos da Tabela 3 resultaram não-significativos. Assim, entre os homens e as mulheres que se graduam em física, os tempos de vida estão distribuídos aproximadamente da mesma maneira. Em contrapartida, a Tabela 4 aponta para uma diferença estatisticamente significativa entre as curvas apresentadas na Fig. 6 (ao nível de significância estatística $p < 0{,}05$). Nele, a curva da distribuição do tempo de vida para os homens está mais à esquerda, indicando que a decisão pela evasão é, em média, mais demorada para as mulheres.

A mesma análise foi realizada para as categorias licenciatura e bacharelado, porém, seus gráficos e tabelas serão omitidos. Dessa análise foi possível perceber que, entre os egressos do curso de física, os licenciados ficam retidos por mais tempo que os bacharéis (com $p_{LOGRANK} < 0,01$). Entre os evadidos, os licenciados também ficam retidos por mais tempo (com $p_{LOGRANK} < 0,01$).

Apesar de os exemplos apresentados acima envolverem a comparação entre somente duas categorias de variáveis nominais, os testes não-paramétricos podem ser facilmente estendidos para mais de duas categorias. Por outro lado, quando for necessário analisar o efeito de duas ou mais variáveis (nominais e/ou intervalares) ao mesmo tempo, é preciso recorrer aos modelos semi-paramétricos. Na próxima seção discutiremos uma técnica que permite tratar situações multivariadas: a regressão de Cox.

7. Técnicas semi-paramétricas da análise de sobrevivência

7.1. A regressão de Cox

Em várias situações de pesquisa, a única maneira de estudar a relação entre o tempo de vida e diversas variáveis explicativas (nominais e/ou intervalares) é construir um modelo paramétrico ou semi-paramétrico que incorpore os efeitos dessas variáveis. A regressão de Cox é um modelo de risco proporcional definido pela seguinte expressão geral [9]

$$h(t, \mathbf{x}, \mathbf{B}) = h_0(t) \cdot r(\mathbf{x}, \mathbf{B}), \quad (8)$$

com

$$r(\mathbf{x}, \mathbf{B}) = \exp\{b_1 x_1 + b_2 x_2 + \ldots + b_w x_w\} = e^{\mathbf{B}\cdot\mathbf{x}}. \quad (9)$$

Na expressão acima, $h_0(t)$ é chamada taxa de base de risco, $\mathbf{x}$ é o vetor das variáveis explicativas e $\mathbf{B}$ é o vetor dos parâmetros ajustáveis. Como é possível perceber, enquanto o EKM estima diretamente a função sobrevivência $S(t)$ e a distribuição do tempo de vida $F(t)$, a Regressão de Cox está ligada mais diretamente à taxa de risco $h(t)$.

Os modelos de Cox são ditos de risco proporcional porque as razões entre as taxas de risco para indivíduos com diferentes valores ou em diferentes categorias das variáveis explicativas $\mathbf{x}$ são assumidas como independentes do tempo. Considere, por exemplo, que a função $h(t, \mathbf{x}, \mathbf{B})$ descreva a taxa de risco de evasão no curso de física em função do tempo e do sexo do estudante. Considere também que a variável sexo tenha sido codificada com o escore 0 para a categoria "homem" e 1 para a categoria "mulher" (usual codificação para uma variável binomial). Se o resultado da regressão de Cox é $B_{SEXO} = -0{,}16$, decorre que

$$\frac{h(t, \mathbf{x} = 1, \mathbf{B})}{h(t, \mathbf{x} = 0, \mathbf{B})} = \frac{\exp\{B_{SEXO} \cdot 1\}}{\exp\{B_{SEXO} \cdot 0\}} = e^{-0{,}16} \cong 0,85. \quad (10)$$

Mas o resultado acima é a razão entre a taxa de risco das mulheres pela taxa de risco dos homens. Como ela resulta ser 0,85, significa que as mulheres possuem uma taxa de risco de evasão igual a 85% da taxa dos homens, ficando retidas no curso por mais tempo antes de evadir.

Considere agora que, ao introduzir uma variável intervalar no modelo – por exemplo, a pontuação obtida no concurso vestibular à universidade – deseja-se comparar as taxas de risco de dois estudantes que apresentem uma diferença de 10 pontos nessa variável. Se, o resultado da regressão de Cox é $\mathrm{B}_{VEST} = 0{,}07$, decorre que

$$\frac{h(t, \mathbf{x} = y + 10, \mathbf{B})}{h(t, \mathbf{x} = y, \mathbf{B})} = \frac{\exp\{B_{VEST} \cdot y + B_{VEST} \cdot 10\}}{\exp\{B_{VEST} \cdot y\}} = e^{0{,}07 \cdot 10} \cong 2,01. \quad (11)$$

Dessa maneira, a taxa de risco para cada estudante é aproximadamente o dobro da taxa de risco para outro estudante idêntico com 10 pontos a menos na pontuação do concurso vestibular. Por conseguinte, os estudantes com menos pontos ficam retidos no curso por mais tempo. Enfim, como é possível perceber, a interpretação dos parâmetros B depende da maneira como as variáveis foram quantificadas ou codificadas; toda a informação necessária para avaliar o efeito das variáveis explicativas sobre o tempo de vida em uma regressão de Cox encontra-se nos parâmetros ajustáveis da equação de regressão.

7.2. Obtendo estimativas dos parâmetros ajustáveis

O procedimento de regressão usado para ajustar os parâmetros do modelo de Cox é chamado método da máxima verossimilhança parcial [9] e pode ser realizado facilmente em diversos pacotes estatísticos comerciais. Como o auxílio do pacote SPSS foi realizada a regressão de Cox para os eventos "evasão" e "diplomação" separadamente, usando a pontuação no concurso vestibular como variável explicativa. Os resultados estão apresentados na Tabela 5.

Para realizar a regressão, a pontuação no vestibular foi transformada em um escore padronizado do tipo z [15] no qual a unidade de medida é o desvio padrão da pontuação no concurso vestibular. Na Tabela 5, Wald (que é igual a $[\mathrm{B}_{VEST}/\text{Desvio padrão}]^2$) é a estatística que permite obter o nível de significância do parâmetro B_{VEST} conforme a penúltima coluna dessa tabela. Como é possível perceber, o tempo de permanência no curso entre os estudantes que colam grau está relacionado à pontuação obtida no vestibular (ao nível de significância $p < 0{,}01$). Entre os estudantes que desistem do curso de física, a pontuação no vestibular não é relevante para determinar o tempo de permanência.

Adicionalmente a taxa de risco de diplomação para um estudante situado um desvio padrão (da pontuação no concurso vestibular) acima de outro estudante é cerca de 1,4 vezes maior conforme a última coluna da Tabela 5.

Portanto, entre os estudantes que colam grau em física, os ingressantes com maior pontuação tendem a ficar retidos por menos tempo, pois seu risco de diplomação é maior. É importante lembrar que esse resultado não implica que os estudantes com maior nota tendam a se diplomar mais frequentemente, mas que eles tendem a se diplomar mais cedo.

8. Considerações finais

Na análise, destacou-se que o curso de física da UFRGS apresenta uma alta proporção de estudantes evadidos (72,8%) tal como ocorre, em média, em todos os cursos de física no Brasil [8]. Usando tabelas de contingência e as estatísticas usuais em tais tabelas, percebeu-se que, comparados nas categorias sexo e habilitação, todos os estudantes são igualmente propensos incorrer em evasão e diplomação.

Tabela 5 - Resultado da regressão de Cox para os eventos terminais. Variável explicativa: escore padronizado z no concurso vestibular à universidade.

Evento terminal	B_{VEST}	Desvio padrão de B_{VEST}	Wald	Graus de liberdade	Nível de significância p	e^B
Diplomação	0,336	0,058	33,843	1	0,000	1,399
Evasão	0,045	0,041	1,202	1	0,273	1,046

Ao usar o EKM como ferramenta de descrição do fluxo escolar, foi possível perceber que, no intervalo de 3,5 a 4,5 anos, há maior densidade do evento diplomação. Após esse intervalo, a probabilidade de que o estudante obtenha o grau é reduzida, tornando-se praticamente nula para os estudantes com mais de 8 anos de permanência. Em contrapartida, a evasão já acontece desde o primeiro semestre, evoluindo de maneira suave ao longo do tempo. Somente após 10 anos, o risco de evasão torna-se quase nulo. Dessa maneira, foi possível perceber que a decisão por abandonar o curso de física não é tomada de maneira apressada, ocorrendo após vários anos de retenção.

Aplicando testes de significância estatística sobre o EKM, percebeu-se que, apesar de homens e mulheres serem igualmente propensos a evadir, as mulheres levam mais tempo até que ocorra a evasão. De maneira semelhante, os estudantes da licenciatura ficam retidos por mais tempo tanto entre os diplomados quanto entre os evadidos.

Usando a regressão de Cox, que permite inserir uma ou mais variáveis categóricas e/ou intervalares no modelo, percebeu-se que os ingressantes no curso de física com maior pontuação tendem a ficar retidos por menos tempo até a diplomação. Porém, isso não quer dizer que esses estudantes colem grau mais frequentemente, mas que eles tendem a se diplomar mais cedo.

Apesar de ser importante retomar os resultados acima nestas considerações finais, acreditamos que a maior contribuição do presente trabalho é de natureza *metodológica*. Nele, chamamos a atenção da comunidade às potencialidades da análise de sobrevivência – ferramenta estatística pouco adotada na pesquisa em educação científica – no estudo do fluxo escolar. O que de fato se deseja concluir das análises apresentadas neste artigo é a efetividade da análise de sobrevivência na detecção de variações sutis nas distribuições de tempo de vida e sua relevância na análise do fluxo escolar. Dessa maneira, é preciso reconhecer a necessidade de elaborar novos modelos de sobrevivência a partir de outras variáveis explicativas (escolhidas sob alguma perspectiva teórica definida) e aplicá-los a outros contextos educacionais, para que se avance mais significativamente com respeito à compreensão do fluxo escolar propriamente dito.

Embora a evasão tenha recebido maior atenção na caracterização do fluxo escolar da educação científica superior [4–6], a retenção também pode representar prejuízos para o estudante e para a instituição de ensino. Com efeito, aplicada ao estudo do fluxo escolar, a análise de sobrevivência permite descrever quando a evasão e a diplomação ocorrem e quais fatores estão relacionados à permanência prolongada no curso.

Acreditamos que, com este trabalho, contribui-se para que pesquisadores interessados no tema da evasão e da retenção nos cursos de formação científica se alertem para as potencialidades da análise de sobrevivência na descrição e compreensão do fluxo escolar dos cursos de graduação em ciências.

Agradecimentos

Agradecemos à diretora Denise Coutinho por ter fornecido acesso ao registro discente da UFRGS para os propósitos desta pesquisa (protegendo a identidade dos alunos) e ao estudante Jorge Luís Alves da Silva pelo auxílio técnico oferecido aos autores deste trabalho na preparação da planilha de dados antes de ela ser inserida no software de análise propriamente dito.

Referências

[1] J.F. Soares e E.A. Colosimo, *Métodos Estatísticos na Pesquisa Clínica* (USP, Ribeirão Preto, 1995).

[2] R.A. Maller and X. Zhou, *Survival Analysis with Long Term Survivors* (John Wiley, London, 1996).

[3] Brasil, Decreto n° 6.069, de 24 de abril de 2007: Institui o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais - REUNI. Diário Oficial da União, 79, 7 (2007).

[4] S.M. Arruda, M.A. Carvalho, M.M. Passos e F.L. Silveira, Cad. Bras. Ens. Fís. **23**, 3 (2006).

[5] M.M. Braga, M.C. Peixoto, L.F. Diniz e T.F. Bogutchi, Avaliação **7**, 1 (2002).

[6] O. Portilho, *Um Estudo da Evasão no Curso de Graduação em Física na UnB* (UnB, Brasília, 2009).

[7] W.B. Andriola, Ensaio: Aval. Pol. Públ. Educ. **11**, 40 (1993).

[8] R.L.L. Silva Filho, P.R. Montejunas, O. Hipólito e M.B.C.M.L. Lobo, Cad. Pesq. **37**, 132, (2007).

[9] D.W. Hosmer and S. Lemeshow, *Applied Survival Analysis: Regression Modeling of Time to Event Data* (John Wiley & Sons , New York, 1999).

[10] E.A. Colosimo e S.R. Giolo, *Análise de Sobrevivência Aplicada* (Edgard Blucher, São Paulo, 2006).

[11] M.E. Gonçalvez, M.E. *Análise de Sobrevivência e Modelos Hierárquicos Logísticos Longitudinais: Uma Aplicação à Análise da Trajetória Escolar* (Faculdade de Ciências Econômicas, Belo Horizonte, 2008).

[12] UFRGS, Resolução do Conselho de Ensino, Pesquisa e Extensão n° 38, de 6 de dezembro de 1995, *Sobre as Normas para Jubilamento e Recusa de Matrícula*. Disponível em `http://www.ufrgs.br/cepe/legislacao`.

[13] S. Siegel *Estatística Paramétrica e Não-Paramétrica* (McGraw-Hill , São Paulo, 1975).

[14] J.F. Hair Junior, R.E. Anderson, R.L. Tatham e W.C. Black, *Análise Multivariada de Dados* (Bookman, Porto Alegre, 2005).

[15] J.C. Nunally, *Psychometric Theory* (McGraw-Hill, New York, 1978).