INTER SCIENTIA

Vol. 6 • Nº 2 • Ano 2018

MODELOS DE SOBREVIVÊNCIA APLICADOS À EVASÃO DOS ALUNOS DE ESTATÍSTICA DA UFPB

Survival Models Applied to the Evasion of Statistics Students of the UFPB

Recebido em: 01/10/2018.
Aceito em: 08/11/2018.

Alisson de Oliveira Silva[1]
Antonio Guedes Gondim Filho[2]
Camila Ribeiro da Silva[3]
Danilo Rangel Arruda Leite[4]
Luana Cecília Meireles da Silva[5]
Wanessa Weridiana da Luz Freitas[6]

RESUMO

A educação é considerada um dos temas mais relevantes da atualidade, tendo grande impacto no desenvolvimento de vários segmentos de uma região. No ensino superior brasileiro, tem-se observado um importante crescimento do número de Instituições de Ensino Superior, matrículas, cursos e docentes. Juntamente com esse contínuo crescimento, verifica-se um aumento considerável da evasão nos cursos superiores. No curso de estatística da UFPB, em particular, nota-se também uma grande quantidade de alunos evadidos, tornando-se fundamental a identificação do perfil desses alunos, bem como a determinação de possíveis fatores que influenciam a sua desistência.

1 Mestre em Estatística pela Universidade Federal de Pernambuco (UFPE). Analista Administrativo da Empresa Brasileira de Serviços Hospitalares - EBSERH/HULW/UFPB. E-mail: alissonhulw@gmail.com

2 Mestre em Biometria e Estatística Aplicada pela Universidade Federal Rural de Pernambuco (UFRPE). Analista Administrativo da Empresa Brasileira de Serviços Hospitalares - EBSERH/HC/UFPE. E-mail: antonio.ebserh@gmail.com

3 Mestre em Modelos de Decisão e Saúde pela Universidade Federal da Paraíba (UFPB). Estaticista da Secretária Estadual de Saúde da Paraíba (SES-PB). E-mail: camilaribeiroufpb@hotmail.com

4 Doutorando em Modelos de Decisão e Saúde pela Universidade Federal da Paraíba. Analista de Tecnologia da Informação da Empresa Brasileira de Serviços Hospitalares – EBSERH/HULW/UFPB. E-mail: danilorangel@buscapb.com.br

5 Doutoranda em Estatística pela Universidade Federal de Pernambuco (UFPE). E-mail: ceciliameireles2006@hotmail.com

6 Doutoranda em Ciência da Computação pela Universidade Federal de Pernambuco (UFPE). E-mail: wanyweridiana@hotmail.com

Para isso, foi utilizada estatística descritiva e modelos de regressão em análise de sobrevivência.

Palavras-chave: Educação. Evasão. Modelo de regressão em sobrevivência.

ABSTRACT

Education is considered to be one of the most relevant themes of our today, having a great impact on the development of several segments of a region. In Brazilian higher education, there has been an important increase in the number of Higher Education Institutions, enrollments, courses and teachers. Along with this continued growth, there is a considerable increase in dropout in higher education. In the UFPB's statistics undergraduate course, in particular, a large number of students are also evaded, making it fundamental to identify the profile of these students, as well as determining possible factors that influence their withdrawal. To this end, descriptive statistics and regression models were used in survival analysis.

Keywords: Education. Evasion. Regression model in survival.

INTRODUÇÂO

A educação é considerada um dos temas mais relevantes da atualidade, tendo grande impacto no desenvolvimento de vários segmentos de uma região. Apesar disso, muitos países, principalmente em desenvolvimento, apresentam condições precárias nos sistemas de educação no que concerne a infraestrutura, investimentos em profissionais especializados, tecnologias, etc. Diante dessa realidade e de vários outros fatores, grande parte dos alunos tende a abandonar a escola, sendo atualmente uma das grandes preocupações da educação no Brasil.

No ensino superior brasileiro, têm-se observado um importante crescimento no número de Instituições de Ensino Superior (IES), matrículas, cursos e docentes. Juntamente com esse contínuo crescimento, é possível verificar um aumento da evasão. A evasão é um fenômeno social complexo, definido como a interrupção no ciclo de estudos (GAIOSO, 2005). É atualmente, um problema que vem preocupando as instituições de ensino em geral, sejam públicas ou particulares, pois a saída de alunos implica em graves consequências sociais, acadêmicas e econômicas.

Silva Filho (2007) evidencia que no período compreendido entre 2000 e 2005, no conjunto formado por todas as IES do Brasil, a evasão média foi de 22% e atingiu 12% nas públicas e 26% nas particulares. Além disso, revelou que são poucas as instituições que possuem um programa institucional regular de combate à evasão.

Para demonstrar as dimensões do problema da evasão das IES no país, cabe mencionar que a quantidade de matrículas, em 2008, foi de 5.080.056 alunos. Considerando a média apresentada por Silva Filho (2007) de 22%, cerca de 1.117.612 alunos saíram do sistema de ensino superior no referido ano. Zago (2006) apresenta outro fator a ser considerado: somente 9% dos jovens entre 18 e 24 anos frequentam o ensino superior, um dos índices mais baixos da América Latina. Segundo a autora, alguns estudos indicam que 25% dos potenciais alunos são carentes e não têm condições de ingressar no ensino superior, ainda que este seja gratuito. Silva Filho e Hipólito (2000) apontam que somente 8% da população adulta tem formação superior, enquanto outros países apresentam um percentual maior: Coréia, 32%; Espanha, 28%; Rússia, 55% e Chile, 13%, na década de 1990, o que comprova as grandes disparidades encontradas no ensino superior no Brasil quando comparado com outros países.

Determinados cursos superiores, em particular, apresentam altas taxas de evasão, principalmente da área de ciências exatas, como as engenharias de um modo geral, matemática, física, etc. No curso de bacharelado em estatística, presente em 27 IES do país, entre públicas e privadas, é possível verificar também altos índices de evasão, decorrentes de diversos fatores. Para o curso de estatística da UFPB, criado no ano 2000, nota-se também uma grande quantidade de alunos evadidos, tornando-se de suma importância a identificação do perfil desses alunos, bem como a determinação de possíveis fatores que influenciam na desistência dos alunos do bacharelado desta instituição, de modo a auxiliar na tomada de decisões. Desse modo, o presente trabalho tem por objetivo, identificar o perfil dos alunos evadidos do curso de estatística da Universidade Federal da Paraíba, e analisar o tempo até a evasão destes alunos. Para isso, serão utilizadas técnicas descritivas para análise do perfil, e modelos de regressão em sobrevivência para análise do tempo até a evasão dos alunos.

MATERIAIS

O banco de dados utilizado para análise do perfil e do tempo até a evasão dos alunos do bacharelado em estatística foi cedido pelo Núcleo de Tecnologia da Informação (NTI) da Universidade Federal da Paraíba, que possui uma extensa base de dados com informações sobre os alunos ingressantes, tais como sexo, procedência, ano do ingresso, ano da evasão, forma de ingresso, etc. Inicialmente foi realizado um filtro dos dados disponibilizados para contemplar apenas as informações dos alunos de estatística. Algumas inconsistências na base de dados foram retiradas, obtendo-se uma amostra de 132 alunos. A variável de interesse (falha) foi definida como sendo o tempo até a desistência, enquanto a censura foi considerada como sendo os alunos diplomados. Foi verificado um total de 14 observações censuradas, o que representa um percentual de aproximadamente 11%.

MODELOS EM ANÁLISE DE SOBREVIVÊNCIA

Para cumprir com o objetivo do presente trabalho, foi utilizada inicialmente estatística descritiva para verificar o perfil dos alunos evadidos do curso de estatística da UFPB, além dos métodos de análise de sobrevivência para analisar o tempo até a evasão dos alunos, uma vez que os dados possuem informações censuradas.

Para verificar possíveis relações entre a variável resposta e as covariáveis medidas em cada indivíduo, torna-se necessário utilizar técnicas especializadas que permitam levar em conta essas covariáveis. A técnica estatística que permite relacionar variáveis, mais especificamente uma variável dita dependente, com um conjunto de variáveis denominadas de independentes é a Análise de Regressão. Diversos modelos de regressão foram desenvolvidos na literatura como modelo linear (clássico), modelos lineares generalizados, modelos aditivos generalizados, modelos mistos, etc. Todos esses modelos supõem para a variável resposta uma distribuição de probabilidade de forma que os parâmetros do modelo sejam estimados. Porém, na presença de variável resposta estritamente positiva, assimétrica e possivelmente censurada, os modelos de regressão tradicionais são inapropriados, de forma que as informações das censuras não são levadas em consideração no processo de estimação dos parâmetros.

No entanto, uma forma de ultrapassar esses problemas referentes à estimação e formulação do modelo é supondo-se uma distribuição assimétrica para a resposta e utilizar o método de máxima verossimilhança para estimar os parâmetros desconhecidos do modelo. Formalmente temos:

$$T = \exp\{X\beta\}\exp\{\sigma v\} \tag{1}$$

onde T representa o tempo até a ocorrência do evento de interesse, uma matriz contendo as covariáveis, um vetor de parâmetros desconhecidos a serem estimados e parâmetros de escala. Um modelo definido desta forma, denomina-se modelo de tempo de vida acelerado. Diversas distribuições de probabilidade podem ser assumidas para a variável resposta tais como: Exponencial, Weibull, Log-Normal, Gama e Log-logística. Para estimação dos parâmetros, utiliza-se o método de máxima verossimilhança por permitir levar em consideração informações relacionadas à censura, o que não é permitido nos demais métodos de estimação (momentos, mínimos quadrados, mínimos quadrados generalizados). Matematicamente, o método consiste em maximizar a seguinte quantidade:

$$L(\beta) = \prod_{i=1}^{n}[f(y_i, \beta|x_i)]^{\delta_i}[S(y_i, \beta|x_i)]^{\delta_i} \tag{2}$$

em que . O primeiro termo do produtório é referente às falhas, enquanto que o segundo termo introduz informações relacionadas às censuras presente nos dados. Para obtenção dos estimadores de máxima verossimilhança, é necessário substituir as funções de densidade e sobrevivência na expressão (2).

Uma avaliação da adequação do modelo ajustado é parte fundamental da análise dos dados. No modelo de regressão linear usual, uma análise gráfica dos resíduos é usada para esta finalidade. Nos modelos de regressão apresentados aqui, a definição de resíduos não é tão clara e, desse modo, diversos resíduos têm sido propostos na literatura para acessar o ajuste do modelo, como os apresentados em (KLEIN e MOESCHBERGER, 1997), (LAWLESS, 1982) e (THERNEAU e GRAMBSCH, 2000).

Técnicas gráficas, que fazem uso dos diferentes resíduos propostos são, em particular, bastante utilizadas para examinar diferentes aspectos do modelo. Um desses aspectos é o de avaliar, por meio dos resíduos, a distribuição dos erros. Estas técnicas, no entanto, como bem observado por Klein e Moeschberger (1997), devem ser utilizadas como um meio de rejeitar modelos claramente inapropriados e não para "provar" que um particular modelo paramétrico está correto, mesmo porque, em muitas aplicações, dois ou mais modelos paramétricos podem fornecer ajustes razoáveis bem como estimativas similares das quantidades de interesse.

No presente trabalho, para verificação da qualidade do ajuste foram utilizados os resíduos propostos por Cox-Snell (1968), e os resíduos padronizados. Esses resíduos são definidos respectivamente por:

$$\hat{e}_i = \widehat{\Lambda}(t_i|x_i) \tag{3}$$

onde é a função de risco acumulada.

$$\hat{v}_i = \frac{(y_i - x_i\tilde{\beta})}{\hat{\sigma}} \tag{4}$$

RESULTADOS E DISCUSSÕES

Para verificar o perfil dos alunos evadidos do curso de bacharelado em estatística da UFPB, algumas estatísticas descritivas foram calculadas para algumas covariáveis, e que são sumarizadas na Tabela I. Pode-se observar que a maioria dos alunos evadidos é do sexo masculino, representando 69,5%, já que a demanda de alunos que buscam o bacharelado em estatística da UFPB em sua maioria é do sexo masculino. Quanto à procedência, verificamos que 56,1% são da capital, enquanto que 33,6% são de outras cidades do Estado. Das formas de ingresso no curso, verifica-se ainda uma predominância do vestibular tradicional através do Processo Seletivo Seriado (PSS),

com um percentual de 72,9%. Para as demais formas de ingresso como transferência voluntária, ainda são pouco expressivas. Para a variável ensino básico, verificamos que não há grandes diferenças dos percentuais entre alunos que cursaram o ensino fundamental e médio em escolas públicas ou privadas, evidenciando que o curso não é opção apenas para aqueles possivelmente mais carentes. A idade média dos alunos desistentes do curso foi de 28 anos, com um desvio padrão de aproximadamente 8 anos, sendo os extremos respectivamente, 18 e 62 anos.

Para analisar o tempo até a desistência dos alunos do curso de estatística, faz-se necessário o uso de técnicas da análise de sobrevivência já que essas lidam com informações censuradas, incluindo-as na análise, de forma a reduzir possíveis vieses na análise estatística. Inicialmente, a função de sobrevivência para a variável resposta, foi estimada usando o estimador não paramétrico de Kaplan-Meier (KAPLAN e MEIER, 1958).

Tabela 1: Distribuição de frequências das variáveis sexo, procedência, forma de ingresso e tipo de ensino básico de alunos evadidos do curso de estatística da UFPB.

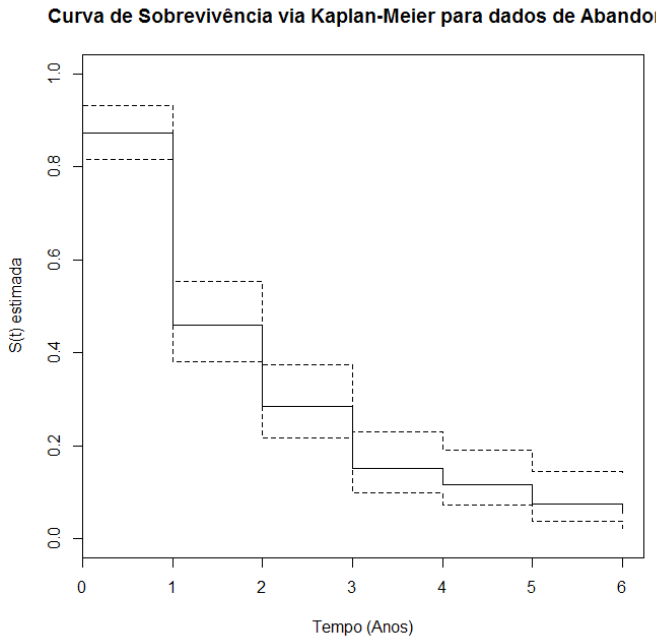
Variável	n	%
Sexo		
Masculino	92	69,5
Feminino	40	30,5
Procedência		
João Pessoa	74	56,1
Outras cidades da PB	44	33,6
Outros estados	14	10,3
Formas de ingresso		
PSS	96	72,9
Outras formas	36	27,1
Ensino básico		
Pública	62	46,7
Privada	58	43,9
Pública/Privada	12	9,3

Tabela 2: Medidas descritivas da variável idade de alunos evadidos do curso de estatística da UFPB.

Estatística	Estimativa
Média	28,03
Mediana	25,5
Mínimo	18
Máximo	62
Desvio padrão	7,82
Variância	61,17

Através dessas estimativas é possível verificar que a chance dos alunos que ingressam no bacharelado desistirem após o primeiro ano de curso é 45,85%, o que indica que mais de 50% dos alunos não conseguem sequer concluir o primeiro ano de curso, período referente às disciplinas básicas, onde estão inseridos os cálculos, que são responsáveis por grande parte das desistências. Em contrapartida, a probabilidade de um aluno desistir após o terceiro ano é de 15,07%, período relacionado às disciplinas especificas do curso de estatística. Ou seja, a partir do terceiro ano de curso a chance de desistência do aluno é bastante pequena.

Figura 1: Curva de Sobrevivência para o tempo até a evasão dos alunos de estatística.



No entanto, a análise feita até o momento, não permite identificar qualquer fator associado à evasão dos alunos. Para verificar efetivamente a relação entre o tempo até a evasão dos alunos e possíveis fatores, torna-se necessária a utilização de um modelo de regressão. Como em geral o tempo até a ocorrência do evento de interesse é assimétrica e existe a presença de observações censuradas, os métodos convencionais de modelagem tornam-se inadequados. Dessa forma, serão utilizados os modelos de regressão para dados de sobrevivência.

Um passo fundamental para iniciar a modelagem é selecionar de forma adequada as variáveis preditoras. A abordagem utilizada, deriva do método de Collet, cujos passos encontram-se na tabela a seguir para o modelo com distribuição Log-normal. Vale mencionar que esta abordagem também foi utilizada para selecionar as covariáveis para os modelos Weibull e Exponencial, para fins de comparação entre os três modelos.

Tabela 3: Seleção de variáveis explicativas para o modelo com distribuição log-normal para o tempo até a evasão de alunos do curso de estatística da UFPB.

Modelo	-2log(L)	Estatística	Valor P
Nulo	291,2214	-	-
Sexo(S)	290,7530	0,4684	0,4940
Naturalidade(N)	290,7878	0,4336	0,5100
Forma de Ingresso(F)	290,0834	1,1380	0,2860
Formação Básica(FB)	290,2744	0,9470	0,3300
Cor(C)	290,0532	1,1682	0,2800
Idade(I)	291,0662	0,1552	0,6940
N + F + C	274,0388	-	-
F + C	278,5254	4,4866	0,0320
N + C	277,4962	3,4574	0.065
N + F	278,4218	4,3830	0,0360
N + F + C	274,0388	-	-
N + F + C + S	274,0372	0,0016	0,9680
N + F + C + FB	274,0258	0,0130	0,9090
N + F + C + idade	271,4638	2,5750	0,1090
N + F + C	274,0388	-	-
N + F + C + N*F	272,1688	1,8700	0,1710
N + F + C + N*C	272,7180	1,3190	0,2520
N + F + C + F*C	274,0386	0,0002	0,9880

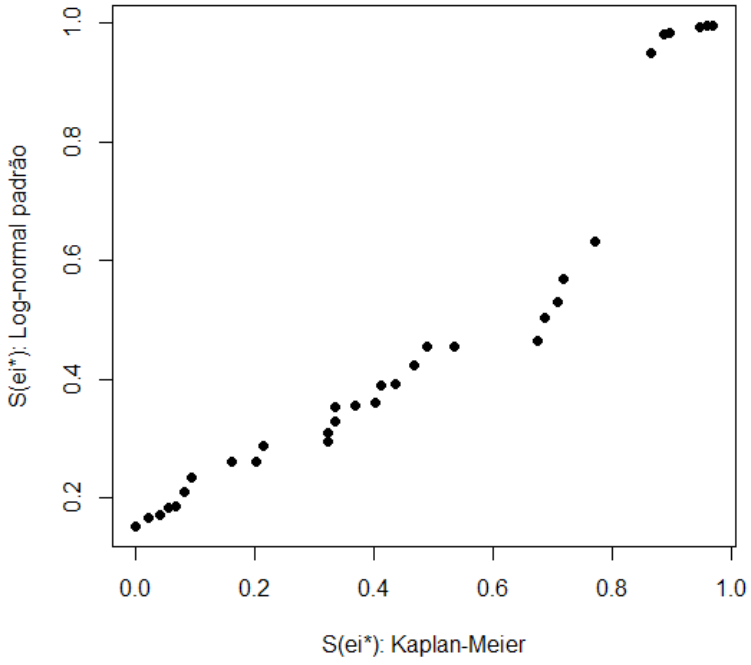
Verificou-se que os modelos com distribuição Weibull, e Exponencial, não se ajustaram adequadamente, já que nenhuma das covariáveis foi significativa ao nível de 10% no primeiro passo. Já para o modelo Log-normal, as variáveis naturalidade ($p < 0,10$) e cor ($p < 0,10$), foram significativas. Além dessas optou-se por incluir a variável forma de ingresso, já que o p-valor não se encontra tão discrepante em relação ao nível escolhido. No passo 2, as variáveis foram ajustadas conjuntamente, e posteriormente verificada a importância de cada uma das variáveis selecionadas através do teste da razão de verossimilhanças. Neste passo todas as variáveis foram significativas. O terceiro passo não se fez necessário já que todas as variáveis no passo 2 foram significativas. No passo quatro, foram incluídas as variáveis excluídas no passo um, juntamente com aquelas selecionadas no passo 2 para verificar efetivamente a não significância destas variáveis. Verificou-se que estas não são estatisticamente significativas. O passo cinco também não se fez necessário. Por fim, verificaram-se possíveis termos de interação entre as variáveis selecionadas. Nesse passo, nota-se através do p-valor das últimas três linhas da tabela anterior, que estes termos não foram significativos. Assim, temos que o modelo selecionado foi aquele com distribuição Log-normal e com as variáveis explicativas: naturalidade, forma de ingresso e cor, cujas estimativas de máxima verossimilhança encontram-se na tabela a seguir.

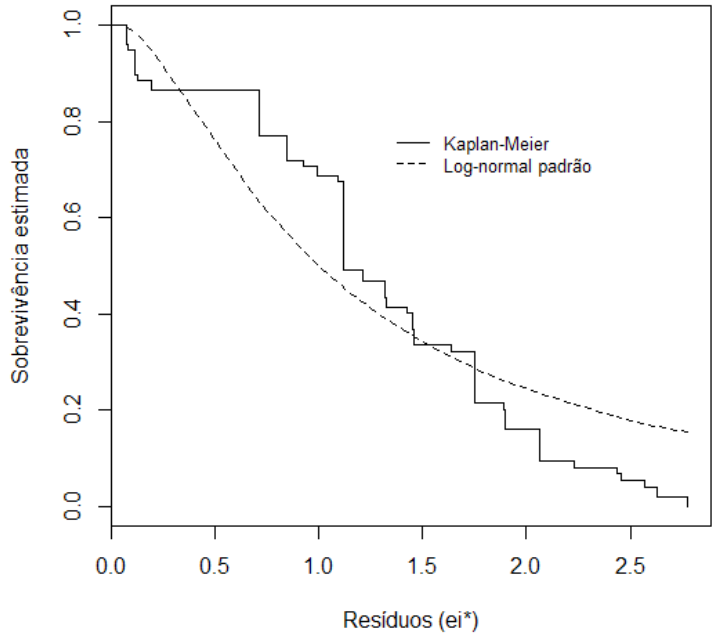
Tabela 4: Estimativos dos parâmetros do modelo log-normal para o tempo até a evasão de alunos do curso de estatística da UFPB.

Variáveis	Estimativa	Erro-padrão	p-valor
Intercepto	-2,7	1,2274	0,0280
Naturalidade	1,9	0,8868	0,0320
Forma de Ingresso	2,22	1,1843	0,0610
Raça	-1,89	0,8938	0,0350
Log(escala)	1,44	0,0769	0,0000

Para interpretação dos parâmetros estimados, torna-se fundamental antes verificar a adequação do modelo. Para isto, foi realizada uma análise dos resíduos padronizados do modelo ajustado, para verificação da adequação da distribuição aos dados. Para que o modelo log-normal seja adequado, é necessário que os resíduos sejam uma amostra censurada da distribuição Log-normal padrão, e, portanto, o gráfico das sobrevivências estimadas por Kaplan-Meier e pelo modelo ajustado deveriam estar dispersos em torno de uma reta, ou equivalentemente, a curvas de sobrevivências estimadas devem estar próximas. Através dos gráficos verificamos que o modelo se ajustou satisfatoriamente aos dados, apesar de haver um afastamento para os maiores tempos de sobrevivência. Para os resíduos de Cox-Snell, o bom ajuste também foi observado.

Figura 2: Gráficos dos Resíduos Padronizados para o Modelo Log-normal





Com o modelo validado, é de interesse interpretar os parâmetros estimados, de forma a verificar o padrão de associação entre as covariáveis e a resposta, ou seja, se estas aceleram ou desaceleram o tempo de sobrevivência. Porém, para interpretação destes parâmetros, foi inicialmente feita uma transformação exponencial destas estimativas, resultando na razão dos tempos medianos de sobrevivência (Hosmer e Lemeshow). Os resultados encontram-se na TABELA V.

Tabela 5: Razão dos tempos medianos de sobrevivência para o tempo até a evasão de alunos do curso de estatística da UFPB.

Variáveis	Razão dos tempos medianos
Naturalidade	6,69
Forma de Ingresso	9,21
Raça	0,15

Através destes valores, verifica-se para a variável naturalidade, que alunos que são naturais de João Pessoa possuem um tempo mediano até a evasão aproximadamente 7 vezes maior do que aqueles que são naturais de outras localidades. Para a variável Forma de Ingresso, nota-se que aqueles que foram submetidos a forma tradicional de ingresso (PSS) possuem um tempo mediano até a desistência 9 vezes maior que aqueles que ingressaram de outra forma, como por exemplo PSTV e graduados. Por fim para a variável Cor, o tempo mediano de sobrevivência daqueles que se declaram como branco é 85% menor em relação aqueles que se declaram como outra cor.

CONCLUSÕES

Através da análise do perfil dos alunos evadidos do curso de estatística da Universidade Federal da Paraíba, foi possível verificar para a variável sexo, que mais de 50% das evasões são do sexo masculino. Além disso, 56,1% dos alunos residem em João Pessoa onde está localizado o campus em que é oferecido o curso, e 33,6% das demais cidades do estado. Quanto à forma de ingresso, 72,9% desses alunos ingressa a partir processo seletivo seriado. Em relação à formação básica dos alunos não houve diferenças significativas nos percentuais de alunos que cursaram o ensino básico em escola pública ou privada.

Para verificar fatores relacionados ao tempo até a evasão, foram ajustados modelos de regressão de sobrevivência. Dentre os modelos ajustados, verificou-se através da análise residual, que o modelo Log-normal configurou-se como mais adequado para os dados. Através das estimativas dos parâmetros, mais especificamente da razão dos tempos medianos de sobrevivência que alunos naturais de João Pessoa, que ingressaram através do PSS e se declaram como ter outra cor possuem um tempo mediano de sobrevivência maior que aqueles que são provenientes de outras localidades, que ingressam por PSTV e outros e que se declararam como brancos.

Dessa forma, nota-se a utilidade dos modelos de sobrevivência para o estudo da evasão, por permitir identificar mais efetivamente fatores relacionados à desistência dos alunos do curso de estatística da UFPB.

REFERÊNCIAS

BRESLOW, N. E.; CROWLEY, J. A Large Sample Study of the Life Table and Product Limit Estimates under Random Censorship. **Annals of Statistics**, n. 2, p. 437-453, 1974.

COX, D. R.; SNELL, E. J. A General Definition of Residuals. **Journal of the Royal Statistical Society B**, n. 30, p. 248-275, 1968.

GAIOSO, N. P. L. **O fenômeno da evasão escolar na educação superior no Brasil**. 2005. 75 f. Dissertação (Mestrado em Educação) - Programa de Pós-Graduação em Educação da Universidade Católica de Brasília, Brasília, 2005.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **J. Amer. Statist. Assoc**, v. 53, n. 282, p. 457-481, 1958.

KLEIN, J. P.; MOESCHBERGER, M. L. **Survival Analysis: Thechniques for Censored and Truncated Data**. New York: Springer-Verlag, 1997.

LAWLESS, J. F. **Statistical Models and Methods for Lifetime Data**. New York: John Wiley ans Sons, 1982.

SILVA FILHO, R. L. L. et al. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, São Paulo, v. 37, n. 132, p. 641-659, 2007.

SILVA FILHO, R. L. L.; HIPOÓLITO, O. Financiamento e expansão do ensino superior. Disponível em: <http://www.jornaldaciencia.org.br/Detalhe.jsp?id=62770>. Acesso em: 23 abr. 2018.

THERNEAU, T. M.; GRAMBSCH, P. M. **Modeling Survival Data: Extending the Cox Models**. New York: Springer-Verlag, 2000.

ZAGO, N. Do acesso a permanência no ensino superior: percursos de estudantes universitários de camadas populares. **Revista Brasileira de Educação**, Rio de Janeiro, v.11, n.32, p.226-237, 2006.