Acompanhamento e análise de processos em Shiny

Aluno: Fernando Cesar Moreira Valle
Orientador: Prof. Dr Eduardo Monteiro de Castro Gomes



Universidade de Brasília Instituto de Ciências Exatas - IE Departamento de Estatística - EST

22 de dezembro de 2020

Sumário

- Introdução e Justificativa
- Objetivos
 - Objetivo Geral
 - Objetivos Específicos
- Metodologia
 - Função de sobrevivência
 - Taxa de falha
 - Função taxa de falha acumulada
 - Tempo médio de vida
 - Vida média residual
 - Análise de modelos paramétricos
 - Análise de Adequação ao modelo ajustado
- Análise dos dados
 - Análise exploratória
 - Dashboard
 - Referências



Sumário

- 1 Introdução e Justificativa
- Objetivos
 - Objetivo Geral
 - Objetivos Específicos
- Metodologia
 - Função de sobrevivência
 - Taxa de falha
 - Função taxa de falha acumulada
 - Tempo médio de vida
 - Vida média residual
 - Análise de modelos paramétricos
 - Análise de Adequação ao modelo ajustado
- Análise dos dados
 - Análise exploratória
 - Dashboard
 - Referências



Introdução

- Querendo sempre melhorar a qualidade do serviço prestado ao público, a 5ª Vara da Justiça Federal buscou incentivar maneiras que otimizassem a visualização no tempo de processamento das ações judiciais, mas conjuntamente considerando os desafios a serem contemplados para se lidar diariamente com grandes volumes de demandas nos diversos setores internos.
- Preocupada com os prazos limites para resolução dos pleitos, a instituição buscou realizar parcerias com a UnB para criação de um conjunto de visualizações em Dashboard por meio do software R que fossem capazes de informar os servidores e advogados responsáveis sobre as análises documentais existente dos processos conjuntamente com suas análises descritivas e acompanhamento processual.

Justificativa

- Com o propósito de informar a instituição sobre o tempo sobressalente ou faltante das atividades processuais, delimitou-se um conjunto de procedimentos visuais e analíticos que possuem como intuito evitar a quebra dos prazos limites designados as ações.
- Para usufruir de modelos em análise de sobrevivência no banco de dados, criou-se uma variável indicadora de censura (processos que não tiveram seu status de processamento encerrado até o dia 02/12/2019) com a intenção de gerar um conjunto de gráficos que evidenciem a probabilidade de conclusão dos pleitos.

Sumário

- Introdução e Justificativa
- Objetivos
 - Objetivo Geral
 - Objetivos Específicos
- Metodologia
 - Função de sobrevivência
 - Taxa de falha
 - Função taxa de falha acumulada
 - Tempo médio de vida
 - Vida média residual
 - Análise de modelos paramétricos
 - Análise de Adequação ao modelo ajustado
- Análise dos dados
 - Análise exploratória
 - Dashboard
 - Referências



Objetivos

Objetivo Geral

Criar um sistema de Dashboard, por meio do software R, que contenha gráficos e tabelas que sejam capazes de atualizar e informar servidores e advogados sobre os prazos remanescentes para elaboração e estruturação dos processos descritos como de interesse. Dessa maneira, com auxilio do pacote *Shiny* e *ShinyDashboard* criou-se uma página web referente a instituição da 5ª vara da justiça federal com um layout bootstrap de fácil compreensão e interatividade para o usuário final.

Objetivos

Objetivos Específicos

- Realizar um conjunto de visualizações simples e diretas que sejam capazes de informar o usuário a cerca dos tempo restante ou tardios para conclusão dos processos;
- Gerar gráficos que mostrem o desenvolvimento dos pleitos dentro da instituição de maneira interativa e eficiente;
- Realizar um estudo descritivo e analítico na área de análise de sobrevivência utilizando o banco de dados da Justiça Federal;

Objetivos

Objetivos Específicos

- Estruturar um breve manual de uso afim de inteirar novos usuários sobre as características descritivas de cada processo analisado, conjuntamente evidenciando os prazos delimitados para as conclusões dos pleitos;
- Evidenciar medidas de como prosseguir com a utilização do software e dos códigos estruturados, disponibilizando documentação e concedendo arquivos via Github.

Sumário

- 1 Introdução e Justificativa
- Objetivos
 - Objetivo Geral
 - Objetivos Específicos
- Metodologia
 - Função de sobrevivência
 - Taxa de falha
 - Função taxa de falha acumulada
 - Tempo médio de vida
 - Vida média residual
 - Análise de modelos paramétricos
 - Análise de Adequação ao modelo ajustado
- Análise dos dados
 - Análise exploratória
 - Dashboard
 - Referências

- Os dados obtidos por meio de parceria com a 5ª vara da justiça federal correspondem a 55 (cinquenta e cinco) diferentes tipos de classes processuais com diferentes tempos de circulação entre etapas, sendo estas administradas nos quatro setores internos ao órgão (Secretaria, Gabinete, Central de mandados e Requerido);
- Inicialmente os dados necessitaram de limpeza e manipulação para posteriormente realizar-se as delimitações de censuras em relação aos pleitos que não obtiveram seu status de processamento encerrado até o dia 02/12/2019;

Função de sobrevivência

- Esta é uma das principais funções probabilísticas usadas para descrever estudos na área de análise de sobrevivência, e possui notação descrita por: S(t);
- Pode ser definido como a probabilidade de um objeto em estudo não falhar até um determinado período de tempo $t_{(j)}$, ou seja, a probabilidade da observação analisada não falhar até o tempo $t_{(j)}$;

$$S(t) = P[T > t] = \int_t^\infty f(t) dx, \ t \ge 0$$

- A função de Sobrevivência é uma função não-crescente, absolutamente contínua, tal que $\lim_{t\to 0} S(t)=1$ e $\lim_{t\to \infty} S(t)=0$.
- Uma das técnicas amplamente utilizadas na estimação de $S_{(t)}$ é o Estimador de Kaplan-Meier, sendo este um método não paramétrico de estimação;

Estimador de Kaplan-Meier

 O estimador não paramétrico de Kaplan-Meier proposto por Kaplan e Meier em 1958 se define como a medida mais popular para estimação da Função de Sobrevivência, sendo esta dada por:

$$\hat{S}_{\mathcal{KM}}(t) = \prod_{j:t_(j) \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

• Devido ao fato de $\hat{S}_{KM}(t)$ possuir uma variabilidade associada a estimação, utilizou-se do intervalo de confiança Log descrito abaixo:

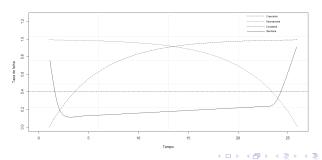
$$\left[e^{-Z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{KM(t)})}}\hat{S}_{KM}(t);e^{Z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{KM(t)})}}\hat{S}_{KM}(t)\right]$$

Taxa de falha

• É o risco instatâneo que o indivíduo tem de experimentar o evento de interesse em um determinado tempo t. No caso de uma variável aleatória contínua, esta função é definida como sendo a razão do limite da probabilidade condicional de um indivíduo experimentar o evento de interesse no intervalo de tempo $[t, t + \Delta t]$ dado que o mesmo não tenha experimentado o evento de interesse antes de t, sobre o intervalo de tempo Δt .

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}, \ t \ge 0$$

• A taxa de falha de um objeto pode vir a assumir quatro formatos diferentes entre si, sendo estas: Função crescente (indicando um aumento na taxa de falha ao longo do tempo), Função constante (evidencia que a taxa não se altera com o passar do tempo), Função banheira (representa uma taxa de falha que é alta inicialmente, baixa no centro e alta novamente no final da vida) e Função decrescente (demonstra uma redução da taxa de falha à medida que o tempo passa).



Função taxa de falha acumulada

• É uma função matemática útil na estimação não-paramétrica e na seleção de modelos mais apropriados para ajustar um determinado conjunto de dados. A função H(t) fornece o risco acumulado do indivíduo no tempo t. Caso esta seja definida como uma variável aleatória contínua, é descrita por:

$$H(t) = \int_0^t \lambda(u) du, \ t \ge 0$$

• Uma das técnicas amplamente utilizadas na estimação de H(t) é o estimador de Kaplan-Meier, sendo este um método não paramétricos de estimação.

Estimador de Kaplan-Meier

• Considerando-se que a função de sobrevivência de uma variável contínua pode vir a ser expressa em relação a função taxa de falha acumulada por meio da equação: $S(t) = exp\{-H(t)\}$. Considera-se que o estimador $\hat{H}_{KM}(t)$ para a função de risco acumulado pode ser obtida por meio de:

$$\hat{H}_{\mathcal{K}\mathcal{M}}(t) = log\left[\hat{S}_{km}(t)
ight]$$

Com seu respectivo intervalo de confiança Log:

$$\left[\hat{H}_{KM}(t) - Z_{1-(\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{KM}(t))}; \hat{H}_{KM}(t) + Z_{1-(\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{KM}(t))}\right]$$

Tempo médio de vida

• É a representação da área gerada abaixo do gráfico da Função de sobrevivência, ou seja, o tempo médio que o objeto sobrevive sem presenciar a falha. (James, 2015)

$$E(T^r) = r \int_0^\infty t^{r-1} S(t) dt$$
, para todo $r \ge 1$

Vida média residual

• É a representação do tempo médio de vida restante para os elementos de interesse no estudo. A vida média residual no tempo t é a área sob a curva de sobrevivência à direita do ponto t, dividido pelo valor da Função de sobrevivência neste ponto.

$$V(t)=rac{1}{S(t)}\int_{t}^{\infty}S(u)\,du,\;$$
 para todo $t\geq0$

Análise de modelos paramétricos

- Nos modelos paramétricos a função de sobrevivência e taxa de falhas dependem de um vetor de parâmetros θ que pode ser estimado via máxima verossimilhança.
- E esperado que as covariáveis possuam influencia no comportamento dos tempos de falhas. Sendo que estas podem explicar parte da heterogeneidade desses tempos.
- No estudo serão utilizadas as variáveis Classe processual e Etapa processual como variáveis preditoras.
- No modelo paramétrico, $\beta = (\beta 0, \beta 1, ..., \beta p)$ é o vetor de parâmetros desconhecidos associado às covariáveis.

Análise de Adequação ao modelo ajustado

Resíduos de Cox-Snell

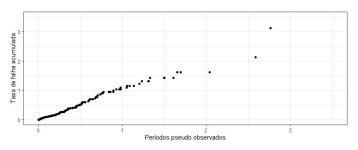
- Diferente da análise de resíduos efetuada em regressão linear, em análise de sobrevivência, não pode-se simplemente analisar os gráficos de resíduos devido a presença de censuras.
- Para analisar a qualidade do ajuste efetuado ao modelo, em 1968, foi criado os resíduos de Cox-Snell.

$$e_i = \hat{H}_0(t_i) exp\Big(\sum_{k=1}^p x_{ip}\hat{\beta}_k\Big)$$

• Onde H_0 (função de risco base acumulada) é uma função escada com saltos nos distintos tempos de falhas.

Resíduos de Cox-Snell

• Os coeficientes β medem o efeito das covariáveis sobre a taxa de falha sendo que, uma covariável pode acelerar, ou desacelerar, a função de risco.



(b) Resíduos de Cox-Snell

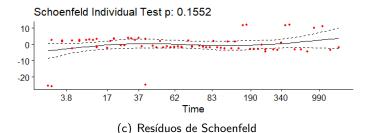
Resíduos de Schoenfeld

- O resíduo de Schoenfeld é a diferença entre os valores observados de covariáveis em um elemento contido no tempo (ti) e os valores esperados desde elemento também no período de tempo (ti) dado o grupo de risco $R_{(ti)}$.
- Um vetor de resíduos de Schoenfeld é obtido em cada tempo observado de falha. Assim, se o elemento *i* é verificado falhar, o correspondente resíduo é obtido por meio de:

$$r_i = x_i - \frac{\sum_{j \in R_{(ti)}} x_j e^{\hat{\beta}_{xj}}}{\sum_{j \in R_{(ti)}} e^{\hat{\beta}_{xj}}}$$

Resíduos de Schoenfeld

• Considerando-se o plot de resíduos padronizados de Schoenfeld contra o tempo é verificável a ocorrência ou não de proporcionalidade, ou seja, se as suposições de riscos proporcionais forem satisfeitas não deverá existir nenhuma tendência no gráfico (Ho: p=0).



Sumário

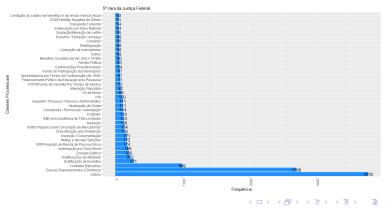
- Introdução e Justificativa
- Objetivos
 - Objetivo Geral
 - Objetivos Específicos
- Metodologia
 - Função de sobrevivência
 - Taxa de falha
 - Função taxa de falha acumulada
 - Tempo médio de vida
 - Vida média residual
 - Análise de modelos paramétricos
 - Análise de Adequação ao modelo ajustado
- Análise dos dados
 - Análise exploratória
 - Dashboard
 - Referências



Análise exploratória

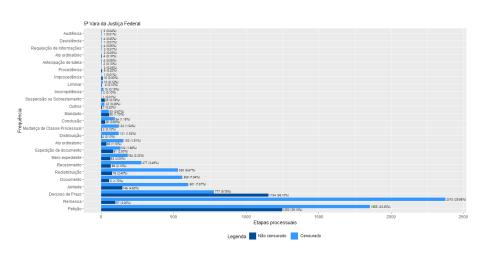
- O banco de dados consta inicialmente com 220.363 (duzentos e vinte mil trezentos e sessenta e três) observações separadas em duas classificações distintas, sendo estas Processual e Pje.
- Foram selecionados apenas os processos com classificações Pje que não incluíssem no mesmo dia duplicadas em suas etapas processuais, contabilizando-se 65.539 (sessenta e cinco mil quinhentos e trinta e nove) termos plausíveis de uso para averiguação das medidas descritivas em interesse.

 Inicialmente plotou-se as frequências das classes processuais para saber como funcionaram as demandas da instituição entre os anos de 2014 a 2019.

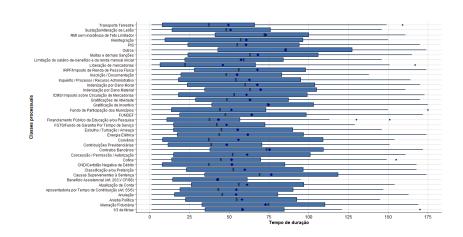


- Observações com contagens inferiores a 50 unidades sofreram ajustes de nomenclatura para "Outros" afim de reduzir a poluição gráfica e otimizar a compreensão dos resultados propostos.
- Dentre as classes processuais que obtiveram maior demanda durante os seis anos de análise, destacam-se: Causas supervenientes à sentença (2.708 24.3%), Contratos bancários (999 8.96%) e Gratificação de incentivo (271 2.43%).

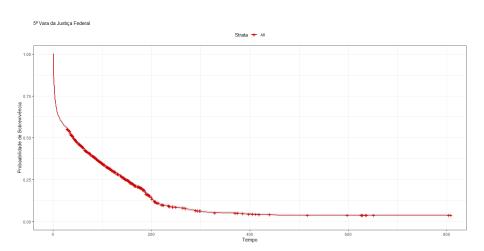
- Em relação a frequência de dados censurados e não censurados existentes em cada uma das etapas, foi realizado um estudo com o objetivo de evidenciar o número de vezes que estes estágios não foram concluídos ou foram concluídos dentre as diversas classes processuais analisadas.
- Assim como ocorrido anteriormente, observações com contagens inferiores a 50 unidades sofreram ajustes de nomenclatura para "Outros" afim de reduzir a poluição gráfica e otimizar a compreensão dos resultados propostos.



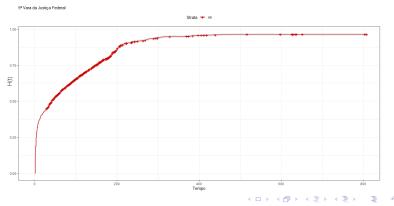
- Pode-se dizer que em relação as etapas não censuradas, "Decurso de prazo" (1154 observações 36.10%) e "Petição" (1250 termos 39.10%) sozinhas equivalem a 2.404 (duas mil quatrocentas e quatro) frequências de estágios com status finalizados até o dia de coleta das informações.
- Já em relação à frequências de etapas censuradas, destaca-se uma maior ocorrência das etapas em "Petição" e "Remessa" com respectivamente 1.855 e 2.374 casos distintos. Dessa maneira, vale ressaltar que sozinhas essas duas atividades foram responsáveis por 4.229 observações dentre os 7.890 casos, representando cerca de 53.59% do valor total dos registros observados no conjunto de dados.



- A função de sobrevivência serve para comparar os tempos de falha.
 No exemplo abaixo, pode-se comparar o comportamento do tempo até a conclusão do processo de acordo com outras variáveis de interesse, como classe processual e etapa processual
- Observa-se que a função se inicia em um determinado momento no tempo, com 100% dos processos ainda com status em aberto, nos permitindo calcular qual a percentagem desses processos permanecem em aberto em relação a outros momentos ao longo do tempo, ou seja, serve para evidenciar o percentual de chance dos pleitos serem concluídos em cada instante de tempo.



 Já a função da taxa de falha acumulada mostra a proporção de processos no todo que inicialmente encontram-se com status em aberto ou que não sofreram ainda o evento de interesse.



 Realizou-se o cálculo do melhor ajuste à distribuição sem a presença das covariáveis para o modelo paramétrico de falhas aceleradas (AFT)

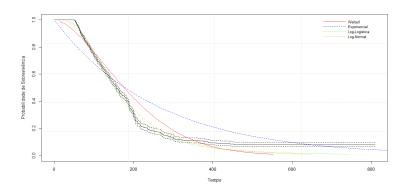


Tabela: Qualidade de ajuste das distribuições em relação ao AIC

	Weibull	Exponential	Log-logistic	Log-normal
AIC	26846.49	27752.56	25996.1	25957.6

 Considerando o resultado obtido na tabela acima com base no AIC (Critério de Informação de Akaike), seleciona-se a distribuição Log-normal como aquela que melhor de adequa aos dados propostos.

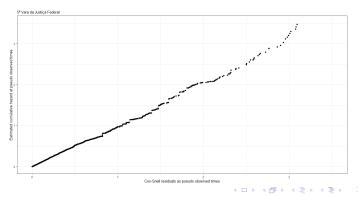
 Valendo-se novamente do critério de seleção AIC e alocando a etapa processual de **Redistribuição** como intercepto (covariáveis qualitativas), gerou-se os resultados observados na tabela abaixo. Indicando assim a escolha do modelo 3 como aquele que na presença das covariáveis assumiu um melhor resultado de adequação.

Tabela: Qualidade de ajuste dos modelos em relação ao AIC

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
AIC	25546.12	25872.4	25427.17	26291.22

 No Modelo 3 estão inseridas as covariáveis referentes as categorias de classes processuais, etapas processuais e as interações entre as covariáveis.;

 Após a seleção do melhor ajuste conjuntamente com sua distribuição, gera-se o interesse de qualificar os resíduos para averiguar a qualidade do modelo proposto. Dessa maneira, utilizou-se o gráfico de resíduos de Cox-Snell para realização dessa atividade.



 Utilizando-se da distribuição Log-normal em relação ao modelo 3, gerou-se por meio do software R as estimativas dos parâmetros e seus respectivos intervalos de confiança conjuntamente com o tempo médio de vida.

Tabela: Estimativas dos parâmetros de μ e σ

	Valor do parâmetro	Limite Inferior	Limite Superior
μ	5.050421	5.024123	5.076719
σ	0.6102394	0.5893945	0.6310844

Tabela: Tempo médio de vida para o modelo ajustado

	Tempo médio de vida	Limite Inferior	Limite Superior
E[T]	188.0329	182.6538	193.4121

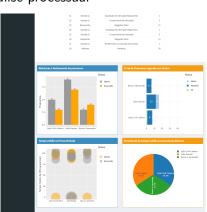
Dashboard

Tela referente a identificação de usuário



Tela referente a análise processual





Tela referente ao manual de uso

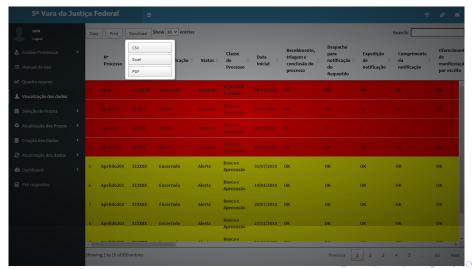




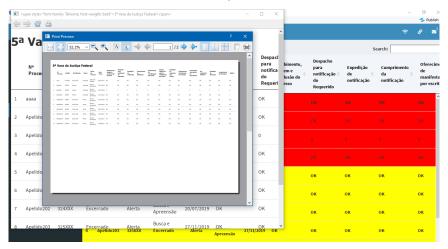
Tela referente ao quadro Resumo



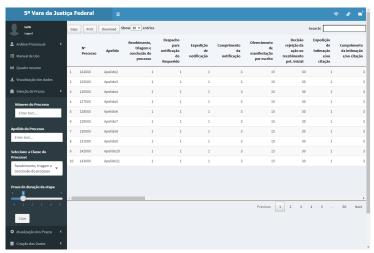
Tela referente a Visualização de dados



Tela referente a Impressão dos dados



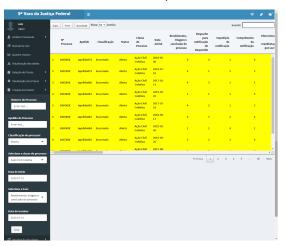
Tela referente a Seleção de prazos



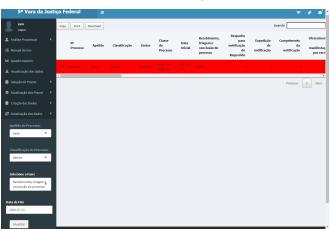
Tela referente a Atualização dos prazos



Tela referente a Criação de dados



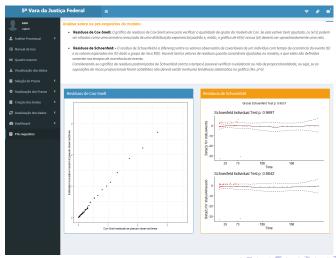
Tela referente a Atualização de dados



Tela referente ao Dashboard



Tela referente aos Pré-requisitos do modelo



Sumário

- 1 Introdução e Justificativa
- Objetivos
 - Objetivo Geral
 - Objetivos Específicos
- Metodologia
 - Função de sobrevivência
 - Taxa de falha
 - Função taxa de falha acumulada
 - Tempo médio de vida
 - Vida média residual
 - Análise de modelos paramétricos
 - Análise de Adequação ao modelo ajustado
- Análise dos dados
 - Análise exploratória
 - Dashboard
 - Referências



Colonismo, E.A.; Giolo, S.R. Análise de Sobrevivência Aplicada. São Paulo: Edgard Blucher, ano 2006.

J.F. Lawless.

Estatistical Models and Methods for Lifetime Data. John Wiley Sons, New York, and 1982.

Poder Judiciário Justiça Federal.

Corregedoria-Regional da Justiça Federal da 2ª Região. Manual de Rotinas e Procedimentos Internos. Brasília.

DF, ano 2009.

Cox, D.R.

Regression models and life tables.

Journal of Royal statistical society. Series V.39, P.1-38, 1972.

Cox, D.R.

Partial likehood..

Biometrika, V.62, N.2, P.269-276, MAR. 1975.

🍆 Giolo, S. R

Modelos de análise de sobrevivência para experimentos dose-resposta...

Campinas: Disseratação de Mestrado, 1994.

Giolo, S. R

Variáveis latentes em análise de sobrevivência e curvas de crescimento..

Piracicaba: Tese de Doutorado, 2003.

Nougaard, P.

Analysis of multivariate survival data..

New York: Springer Verlag, 2000.

🍆 Kaplan, E. L.; Meier, P.

Non-parametric estimation from incomplete observations.. Jour-non-parametric statistical association, V.53,P.547-481, 1958.

Giolo, S. R. Modelos de Riscos Proporcionais.

Paraná: Universidade Federal do Paraná, 2018.

Colonismo, E. A.; Giolo, S. R. Análise de Sobrevivência Aplicada. Edgar Blucher, 2006.

Rupert G. Miller; JR. Rupert. Survival Analysis. Wiley-Interscience, 1998.

Marilia ET AL. Análise de sobrevivência: Teoria e aplicações em saúde. Fiocruz, 2019.

A. Kaufmann; D. Grouchko; R. Cruon

Mathematical models for study of the reability of systems.

University of southern California, Volume 124, 1997.