

Estudo sobre ajuste de modelos de probabilidade a dados de sobrevivência

Study on fitting probability models to survival data

Estudio sobre ajuste de modelos de probabilidad en datos de supervivencia

Recebido: 23/03/2022 | Revisado: 30/03/2022 | Aceito: 03/04/2022 | Publicado: 10/04/2022

Daniel Leonardo Ramírez Orozco

ORCID: <https://orcid.org/0000-0003-0563-2907>
Universidade Federal Rural de Pernambuco, Brasil
E-mail: orozco.dlr@gmail.com

Resumo

Neste artigo, utiliza-se modelos de distribuição de probabilidade conhecidos na literatura. O objetivo deste estudo é procurar alguns modelos de probabilidade que melhor se ajustem a dois conjuntos de dados específicos na área da análise de sobrevivência. O primeiro faz referência à resistência de rolamentos de esferas e o segundo, ao período de falhas sucessivas do sistema de ar condicionado de uma frota de aviões Air Boeing. A estimação dos parâmetros é realizada utilizando o método de máxima verossimilhança. Uma aplicação a dois conjuntos de dados está dada para ilustrar a veracidade dos ajustes. As distribuições que mostraram ótimos resultados foram a Exponencial Exponencializada, Exponencial Burr XII, Gdus Exponencializada e Dagum, para a primeira análise. Para a segunda análise, a Weibull Exponencializada, Kumaraswamy Weibull, Kumaraswamy Burr XII e Dagum. Utiliza-se o Akaike Information Criterion, Bayesian Information Criterion, Anderson Darling e Cramér von Mises, como medidas de comparação. A análise será feita com ajuda do pacote (AdequacyModel) no software R.

Palavras-chave: Bondade de ajuste; Distribuição de probabilidade; Estimação de parâmetros.

Abstract

In this article, probability distribution models known in the literature are used. The aim of the present study is to search some probability models with better fit in two specific data sets in survival analysis. The first one refers to the resistance of ball bearings, and the second to the period of successive failures of the air conditioning system of a fleet of Air Boeing aircraft. The estimation of the parameters is performed using the maximum likelihood method. An application to two data sets is given to illustrate the veracity of the fits. The distributions that showed great results were Exponentialized Exponential, Exponential Burr XII, Gdus Exponentialized and Dagum, for the first analysis. For the second analysis, the Exponentialized Weibull, Kumaraswamy Weibull, Kumaraswamy Burr XII and Dagum. The Akaike Information Criterion, Bayesian Information Criterion, Anderson Darling and Cramér von Mises are used as comparison measures. The analysis will be done with the help of the package (AdequacyModel) in the R software.

Keywords: Goodness of fit; Parameter estimation; Probability distribution.

Resumen

En este artículo se utilizan modelos de distribución de probabilidad conocidos en la literatura. El objetivo de este estudio es buscar algunos modelos de probabilidad que se ajusten mejor a dos conjuntos de datos específicos en el área de análisis de supervivencia. El primero se refiere a la resistencia de los rodamientos de esferas y el segundo al período de fallas sucesivas del sistema de aire acondicionado de una flota de aviones Air Boeing. La estimación de los parámetros se realiza mediante el método de máxima verosimilitud. Se da una aplicación a dos conjuntos de datos para ilustrar la veracidad de los ajustes. Las distribuciones que mostraron grandes resultados fueron Exponencial Exponencializada, Exponencial Burr XII, Gdus Exponencializada y Dagum, para el primer análisis. Para el segundo análisis, Weibull Exponencializada, Kumaraswamy Weibull, Kumaraswamy Burr XII y Dagum. El criterio de información de Akaike, el criterio de información bayesiano, Anderson Darling y Cramér von Mises se utilizan como medidas de comparación. El análisis se realiza con la ayuda del paquete (AdequacyModel) en el software R.

Palabras clave: Bondad de ajuste; Estimación de parámetros; Distribución de probabilidad.

1. Introdução

Existem diferentes modelos de distribuição existentes na literatura usados para analisar dados de tempo de vida. Neste trabalho, trata-se uma ampla variedade de distribuições em vista de uma melhor escolha ao momento de comparar as medidas dos ajustes.

Neste artigo, analisa-se as distribuições Weibull (W) (Weibull, 1951), Exponencial Exponencializada (ExpExp) (Gupta et al.,1998), Weibull Exponencializada (WE) (Pal et al., 2006), Kumaraswamy Weibull (KwW) (Cordeiro et al.,2010), Beta Weibull (BW) (Lee et al., 2007), Burr XII (Burr, 1942), Burr XII Exponencializada (ExpBurrXII) (Afify & Mead, 2017), Kumaraswamy Burr XII (KwBXII) (Paranaíba et al., 2013), Beta Burr XII (BBXII) (Paranaíba et al., 2011), Gumbel Exponencializada (ExpGum) (Nadarajah, 2006), Nadarajah Haghghi (NH) (Nadarajah & Haghghi, 2011), Gdus Exponencializada (GdusExp) (Maurya et al., 2017), Dagum (Dagum, 1977), Kumaraswamy Log Logistic (KmLL) (Santana et al., 2012) e Gamma Exponencializada (ExpGam) (Nadarajah & Gupta, 2007) nos conjuntos de dados.

Especificamente, se ilustra a primeira parte do estudo com dados que fazem referência a resistência de rolamentos de esferas. O segundo conjunto de dados, está relacionado ao período de falhas sucessivas do sistema de ar condicionado de uma frota de aviões. Na primeira análise, utiliza-se os modelos ExpExp, ExpBurrXII, GdusExp e Dagum, que deram melhores resultados para o primeiro conjunto de dados. Já para a segunda análise os modelos WE, KwW, KwBXII e Dagum mostraram resultados satisfatórios para os ajustes.

O objetivo principal deste artigo é ajustar algumas distribuições a conjuntos de dados específicos. Portanto, na Seção 2 ilustra-se duas tabelas que contêm os modelos utilizados e que mostraram os melhores resultados obtidos. Também o método de máxima verossimilhança é apresentado. A Seção 3 mostra a análise dos dados. Finalmente as conclusões.

2. Metodologia

2.1 Distribuições

Nas Tabelas 1 e 2 apresenta-se os modelos que melhor se ajustaram aos conjuntos de dados analisados neste estudo. Mostra-se a função de distribuição acumulada (fda), função de densidade de probabilidade (fdp), os parâmetros e o suporte de cada distribuição.

Tabela 1. Modelos utilizados nos ajustes dos Dados 1.

Modelo	fda	fdp	Parâmetros	Suporte
ExpExp	$F(x; \alpha, \beta) = (1 - e^{-\beta x})^\alpha$	$f(x; \alpha, \beta) = \alpha \beta e^{-\beta x} (1 - e^{-\beta x})^{\alpha-1}$	$\alpha, \beta > 0$	$x > 0$
ExpBurrXII	$F(x; b, \alpha, \beta) = \left(1 - (1 + x^b)^{-\alpha}\right)^\beta$	$f(x; b, \alpha, \beta) = b \alpha \beta x^{b-1} (1 + x^b)^{-(\alpha+1)} \left(1 - (1 + x^b)^{-\alpha}\right)^{\beta-1}$	$b, \alpha, \beta > 0$	$x > 0$
GdusEXP	$F(x; \alpha, \beta) = \frac{e^{(1-e^{-\beta x})^\alpha} - 1}{e - 1}$	$f(x; \alpha, \beta) = \frac{\alpha \beta e^{-\beta x} (1 - e^{-\beta x})^{\alpha-1} e^{(1-e^{-\beta x})^\alpha}}{e - 1}$	$\alpha, \beta > 0$	$x > 0$
Dagum	$F(x; b, \alpha, \beta) = (1 + \alpha x^{-\beta})^{-b}$	$f(x; b, \alpha, \beta) = b \alpha \beta x^{-\beta-1} (1 + \alpha x^{-\beta})^{-b-1}$	$b, \alpha, \beta > 0$	$x > 0$

Fonte: Autoria própria.

Tabela 2. Modelos utilizados nos ajustes dos Dados 2.

Modelo	fda	fdp	Parâmetros	Suporte
WE	$F(x; b, \alpha, \beta) = \left(1 - e^{-(x/\beta)^\alpha}\right)^b$	$f(x; b, \alpha, \beta) = b \alpha / \beta^\alpha x^{\alpha-1} e^{-(x/\beta)^\alpha} \left(1 - e^{-(x/\beta)^\alpha}\right)^{b-1}$	$b, \alpha, \beta > 0$	$x > 0$
KwW	$F(x; a, b, \alpha, \beta) = 1 - \left(1 - \left(1 - e^{-(x/\beta)^\alpha}\right)^a\right)^b$	$f(x; a, \alpha, \beta) = a \alpha / \beta^\alpha x^{\alpha-1} e^{-(x/\beta)^\alpha} \left(1 - e^{-(x/\beta)^\alpha}\right)^{a-1}$	$a, b, \alpha, \beta > 0$	$x > 0$
KWBXII	$F(x; a, b, \alpha, \beta) = 1 - \left(1 - \left(1 - (1 + x^a)^{-b}\right)^\alpha\right)^\beta$	$f(x; a, b, \alpha, \beta) = \alpha \beta a b \left(1 - \left(1 - (1 + x^a)^{-b}\right)^\alpha\right)^{\beta-1} \times \left(1 - (1 + x^a)^{-b}\right)^{\alpha-1} x^{a-1} / (1 + x^a)^{b+1}$	$a, b, \alpha, \beta > 0$	$x > 0$
Dagum	$F(x; b, \alpha, \beta) = (1 + \alpha x^{-\beta})^{-b}$	$f(x; b, \alpha, \beta) = b \alpha \beta x^{-\beta-1} (1 + \alpha x^{-\beta})^{-b-1}$	$b, \alpha, \beta > 0$	$x > 0$

Fonte: Autoria própria.

2.2 Estimação de parâmetros pelo método de máxima verossimilhança

O método de máxima verossimilhança é um dos mais utilizados para encontrar os estimadores. Seja uma amostra aleatória X_1, X_2, \dots, X_n independente identicamente distribuída de tamanho n , de uma população com função de densidade $f(x; \theta)$. A função de verossimilhança está dada a seguir

$$L(\theta; \mathbf{x}) = L(\theta_1, \theta_2, \dots, \theta_k; x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta_1, \theta_2, \dots, \theta_k).$$

Espera-se que para cada ponto amostral \mathbf{x} , $\hat{\theta}(\mathbf{x})$ seja um valor paramétrico no qual $L(\theta; \mathbf{x})$ atinge seu máximo como uma função de θ , com \mathbf{x} fixo (Casella & Berger, 2010).

Se a função de verossimilhança é diferenciável em θ_i os possíveis candidatos para os estimadores de máxima verossimilhança são os valores de $(\theta_1, \theta_2, \dots, \theta_k)$ que solucionam

$$\frac{\partial}{\partial \theta_i} L(\theta; \mathbf{x}) = 0, \quad i = 1, 2, \dots, k.$$

3. Resultados e Discussão

Ajustar distribuições a dois conjuntos de dados é o principal objetivo deste trabalho, portanto, escolher o melhor modelo ajustado, baseando-se em critérios e estatísticas de bondade de ajuste é o que tenta se mostrar no que segue. De forma análoga, Ximenes et al. (2020), fizeram uma análise com ajustes, mas com relação à precipitação no estado de Pernambuco-Brasil. Já, Borges et al. (2021), estudaram distribuições de probabilidade procurando o melhor ajuste em dados climáticos no estado de Manaus-Brasil.

A seguir, analisa-se o conjunto de dados (endurance of deep groove ball bearings), disponíveis em Lawless (2011), que fazem referência a resistência de rolamentos de esferas. Kundu e Gupta (2008), também analisaram esses dados. A seguir os dados apresentados são o número de milhões de revolução antes da falha para cada um dos 23 rolamentos das esferas no teste de vida e serão analisados com ajuda do pacote (AdequacyModel) no software R na versão 3.4.2. Algumas estatísticas descritivas deste conjunto de dados são apresentadas na Tabela 3.

Dados 1: 17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.80, 51.84, 51.96, 54.12, 55.56, 67.80, 68.64, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40.

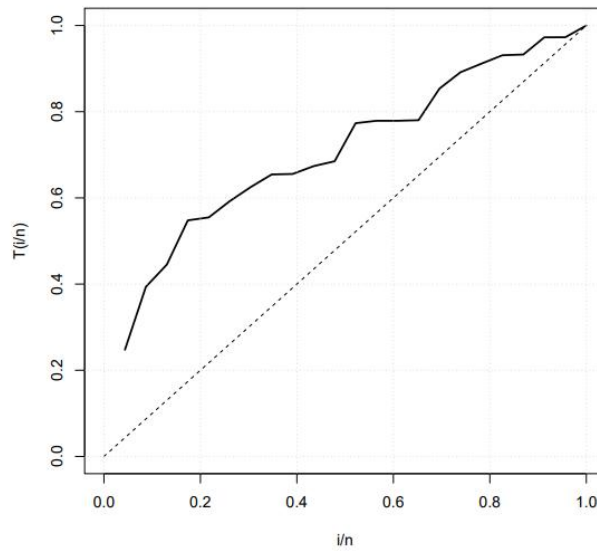
Tabela 3. Estatísticas descritivas dos dados resistência de rolamentos de esferas.

Tamanho amostral	Mínimo	1 Quartil	Mediana	Media	3 Quartil	Máximo
23	17.88	47.20	67.80	72.24	95.88	173.40

Fonte: Autoria própria

Intenta-se identificar o modelo mais apropriado para o conjunto de dados com ajuda do gráfico do tempo total em teste (TTT) proposto por Aarset (1987). O TTT para este caso, na Figura 1, mostra uma curva côncava, isto indica que distribuições com taxa de falha crescentes podem ser apropriadas ou serão boas candidatas para modelar o conjunto de dados.

Figura 1. Curva TTT para o conjunto de dados resistência de rolamentos de esferas.

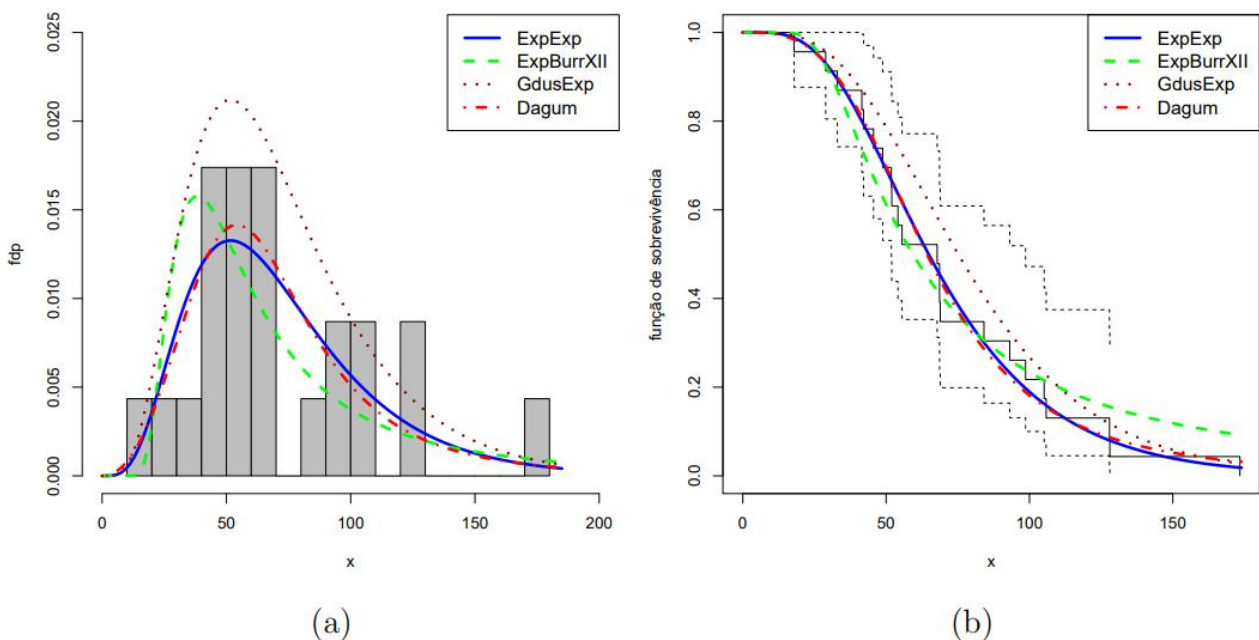


Fonte: Autoria própria

Os melhores resultados dos ajustes dos modelos mencionados para o conjunto resistência de rolamentos de esferas, foram das distribuições ExpExp, ExpBurrXII, GdusExp e Dagum. As densidades estimadas para esses modelos ajustados são apresentados na Figura 2 junto com as curvas da função de sobrevivência estimada para esses modelos versus a função de sobrevivência empírica pelo método de Kaplan-Meier, observa-se de modo geral, que a distribuição GdusExp aparentemente apresenta melhor ajuste comparada com os outros modelos.

Para fazer uma avaliação da qualidade dos ajustes, serão apresentados na Tabela 4 as estimativas dos parâmetros, os erros das estimativas entre parêntesis e se fornecem as estatísticas de bondade de ajuste Anderson-Darling (A^*) e Cramér-von Mises (W^*), também os critérios Akaike Information Criterion (AIC) e Bayesian Information Criterion (BIC) como medidas comparativas.

Figura 2. Densidades ajustadas e funções de sobrevivência ao conjunto de dados resistência de rolamentos de esferas.



Fonte: Autoria própria

Tabela 4. Modelos para os dados resistência de rolamentos de esferas.

Modelo	a	b	α	β	AIC	BIC	W*	A*
ExpExp	-	-	5.4109 (2.1056)	30.5993 (6.0492)	229.9546	232.2256	0.0320	0.1882
ExpBurrXII	-	12.4053 (1.2379)	14.5253 (22.7151)	0.1263 (0.1973)	237.5889	240.9954	0.0755	0.5567
GdusExp	-	-	0.0330 (0.0059)	5.4894 (1.9386)	208.9642	211.2352	0.0372	0.2092
Dagum	-	0.9065 (0.7254)	3.4675 (1.1711)	66.8484 (24.7660)	232.7184	236.1249	0.0340	0.2034

Fonte: Autoria própria

Os valores do AIC e BIC da ExpBurrXII permitem descartar esse modelo por possuir os valores mais altos. Observa-se que o modelo da distribuição ExpExp obtém os menores valores de ajuste de W* e A*, mas os erros padrão da GdusExp são menores comparados com os outros modelos. Porém, os valores de W* e A* da GdusExp mesmo não sendo os menores, variam pouco em comparação a ExpExp e Dagum com três parâmetros. Assim, o GdusExp pode ser escolhido como melhor modelo para analisar os dados resistência de rolamentos de esferas.

No que segue do trabalho, apresenta-se o segundo conjunto de dados tomado de Proschan (1963), relacionado ao período de falhas sucessivas do sistema de ar condicionado de uma frota de aviões Air Boeing 720. Esses dados também foram estudados por Dahiya e Gurland (1972), entre outros. Algumas estatísticas descritivas são apresentadas na Tabela 5.

Dados 2: 194, 413, 90, 74, 55, 23, 97, 50, 359, 50, 130, 487, 102, 15, 14, 10,57, 320, 261, 51, 44, 9 , 254, 493, 18, 209, 41, 58, 60, 48, 56, 87, 11, 102, 12, 5, 100, 14,29, 37, 186, 29, 104, 7, 4, 72, 270, 283, 7, 57, 33, 100, 61, 502, 220, 120, 141, 22, 603, 35,98, 54, 181, 65, 49, 12, 239, 14, 18, 39, 3, 12, 5, 32, 9, 14, 70, 47, 62, 142, 3, 104, 85, 67,169, 24, 21, 246, 47, 68, 15, 2, 91, 59, 447, 56, 29, 176, 225, 77, 197, 438, 43, 134, 184,20, 386, 182, 71, 80, 188, 230, 152, 36, 79, 59, 33, 246, 1, 79, 3, 27, 201, 84, 27, 21, 16,88, 130, 14, 118, 44, 15, 42, 106, 46, 230, 59, 153, 104, 20, 206, 5, 66, 34, 29, 26, 35, 5,82, 5, 61, 31, 118, 326, 12, 54, 36, 34, 18, 25, 120, 31, 22, 18, 156, 11, 216, 139, 67, 310,3, 46, 210, 57, 76, 14, 111, 97, 62, 26, 71, 39, 30, 7, 44, 11, 63, 23, 22, 23, 14, 18, 13, 34,62, 11, 191, 14, 16, 18, 130, 90, 163, 208, 1, 24, 70, 16, 101, 52, 208, 95.

Tabela 5. Estatísticas descritivas dos dados falhas do sistema de ar condicionado de aviões.

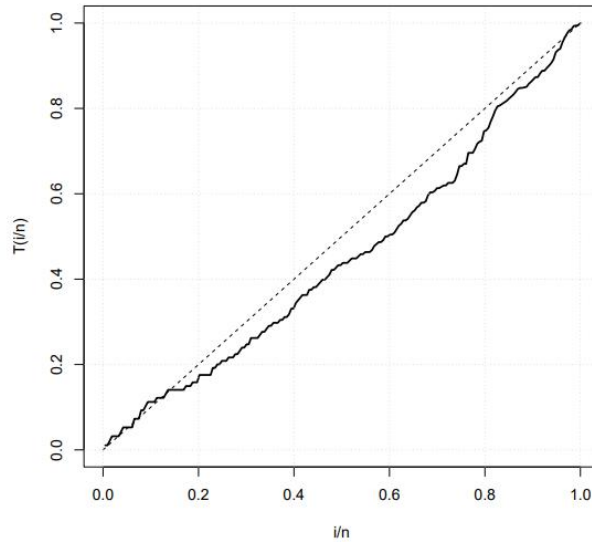
Tamanho amostral	Mínimo	1 Quartil	Mediana	Media	3 Quartil	Máximo
213	1.00	22.00	57.00	93.14	118.00	603.00

Fonte: Autoria própria

A Figura 3 mostra que a curva TTT para os dados falhas do sistema de ar condicionado de aviões é convexa, o que indica que modelos com taxa de falha decrescente poderão ser utilizados para o ajuste.

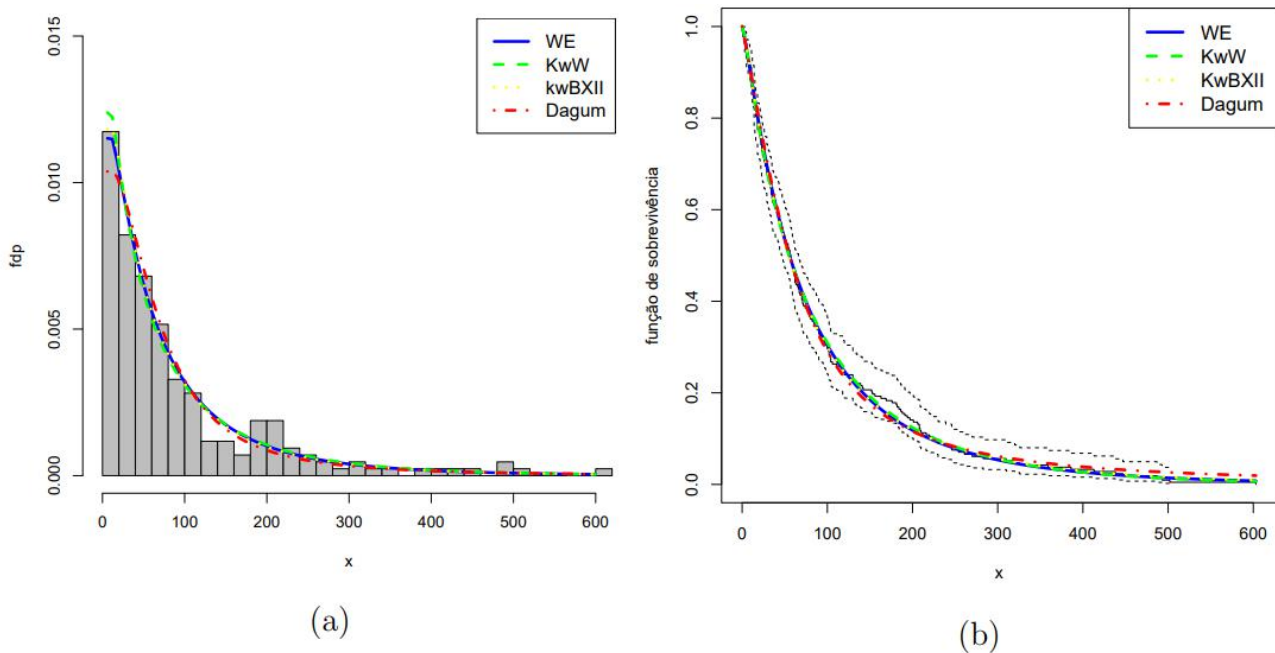
A Figura 4 mostra algumas das densidades estimadas para os diferentes modelos ajustados, junto com as curvas da função de sobrevivência para esses modelos aos dados falhas do sistema de ar condicionado de aviões. Pode ser observado que as distribuições WE, KwW, KwBXII e Dagum apresentam bons ajustes.

Figura 3. Curva TTT para o conjunto de dados falhas do sistema de ar condicionado de aviões.



Fonte: Autoria própria

Figura 4. Densidades ajustadas e funções de sobrevivência ao conjunto de dados falhas do sistema de ar condicionado de aviões



Fonte: Autoria própria.

A Tabela 6 mostra as estimativas dos parâmetros com seus respectivos erros padrão, pode se observar que o KwBXII tem erros bastante grandes, entretanto, a Dagum e KwBXII possuem os valores AIC e BIC maiores que os outros modelos. Assim, verificando o valor das estatísticas W^* e A^* , o KwW com quatro parâmetros mostra o menor valor se comparado com WE que tem três parâmetros. Isto indica que KwW fornece o melhor ajuste para esse conjunto de dados se comparado aos outros modelos probabilísticos.

Tabela 6. Modelos para os dados falhas do sistema de ar condicionado de aviões.

Modelo	a	b	α	β	AIC	BIC	W*	A*
WE	-	2.4651	0.5916	31.2325	2355.2340	2365.3179	0.0373	0.2723
	-	(1.1135)	(0.1255)	(19.8454)				
KwW	2.5265	0.1631	0.7563	6.2682	2355.9609	2369.4061	0.0294	0.2188
	(0.0532)	(0.0150)	(0.0183)	(0.0309)				
KwBXII	75.5108	85.5661	0.1054	3.0057	2357.3027	2370.7479	0.0314	0.2447
	(318.3185)	(111.4633)	(0.1679)	(5.5994)				
Dagum	-	0.6032	1.7559	85.8780	2363.4352	2373.5191	0.0777	0.5787
	-	(0.1575)	(0.2379)	(21.0942)				

Fonte: Autoria própria

4. Conclusões

Foram tomadas duas bases de dados de sobrevivência, apresenta-se as estatísticas descritivas e gráficos dos ajustes aos dados. Se estimam os parâmetros mediante o método de máxima verossimilhança com seus respectivos erros das estimativas. Adicionalmente, mediante o Akaike Information Criterion e Bayesian Information Criterion, e também junto com estatísticas de bondade de ajuste como Anderson-Darling e Cramér-von Mises permitem comparar os ajustes.

Sugere-se como trabalhos futuros utilizar outros testes de bondade de ajuste, além de utilizar outros métodos de estimação dos parâmetros como também realizar estudos de simulação com modelos de probabilidade utilizados em outras áreas da estatística.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Brasil – código de financiamento 001.

Referências

- Aarset, M. V. (1987). How to identify a bathtub hazard rate. *IEEE Transactions on Reliability*, 36(1):106–108.
- Afify, A. Z. & Mead, M. (2017). On five-parameter burr xii distribution: properties and applications. *South African Statistical Journal*, 51(1):67–80.
- Borges, Y. M., da Silva, B. G., de Melo, B. A. R., & da Silva, R. R. (2021). Evaluation of probability distributions in the analysis of minimum temperature series in Manaus-AM. *Research, Society and Development*, 10(3): e46210313616–e46210313616.
- Burr, I. W. (1942). Cumulative frequency functions. *The Annals of mathematical statistics*, 13(2):215–232.
- Casella, G. & Berger, R. L. (2010). Inferência estatística. *Cengage Learning*.
- Cordeiro, G. M., Ortega, E. M., & Nadarajah, S. (2010). The kumaraswamy weibull distribution with application to failure data. *Journal of the Franklin Institute*, 347(8):1399–1429.
- Dagum, C. (1977). A new model of personal income distribution, *Ieconomie appliquée*.
- Dahiya, R. C. & Gurland, J. (1972). Goodness of fit tests for the gamma and exponential distributions. *Technometrics*, 14(3):791–801.
- De Santana, T. V. F., Ortega, E. M., Cordeiro, G. M., & Silva, G. O. (2012). The kumaraswamy-log-logistic distribution. *Journal of Statistical Theory and Applications*, 11(3):265–291.
- Gupta, R. C., Gupta, P. L., & Gupta, R. D. (1998). Modeling failure time data by lehman alternatives. *Communications in Statistics-Theory and methods*, 27(4):887–904.
- Kundu, D. & Gupta, R. D. (2008). Generalized exponential distribution: Bayesian estimations. *Computational Statistics & Data Analysis*, 52(4):1873–1883.
- Lawless, J. F. (2011). Statistical models and methods for lifetime data, volume 362. *John Wiley & Sons*.

- Lee, C., Famoye, F., & Olumolade, O. (2007). Beta-weibull distribution: some properties and applications to censored data. *Journal of modern applied statistical methods*, 6(1):17.
- Maurya, S., Kaushik, A., Singh, S., & Singh, U. (2017). A new class of distribution having decreasing, increasing, and bathtub-shaped failure rate. *Communications in Statistics-Theory and Methods*, 46(20):10359–10372.
- Nadarajah, S. (2006). The exponentiated gumbel distribution with climate application. *Environmetrics: The official journal of the International Environmetrics Society*, 17(1):13–23.
- Nadarajah, S. & Gupta, A. K. (2007). The exponentiated gamma distribution with application to drought data. *Calcutta Statistical Association Bulletin*, 59(1-2):29–54.
- Nadarajah, S. & Haghghi, F. (2011). An extension of the exponential distribution. *Statistics*, 45(6):543–558.
- Pal, M., Ali, M. M., & Woo, J. (2006). Exponentiated weibull distribution. *Statistica*, 66(2):139–147.
- Paranaíba, P. F., Ortega, E. M., Cordeiro, G. M., & Pascoa, M. A. d. (2013). The kumaraswamy burr xii distribution: theory and practice. *Journal of Statistical Computation and Simulation*, 83(11):2117–2143.
- Paranaíba, P. F., Ortega, E. M., Cordeiro, G. M., & Pescim, R. R. (2011). The beta burr xii distribution with application to lifetime data. *Computational statistics & data analysis*, 55(2):1118–1136.
- Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*, 5(3):375–383.
- Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of applied mechanics*.
- Ximenes, P. d. S. M. P., da Silva, A. S. A., Ashkar, F., & Stosic, T. (2020). Ajuste de distribuições de probabilidade a precipitação mensal no estado de pernambuco-brasil. *Research, Society and Development*, 9(11):e4869119894–e4869119894.