ESTIMAÇÃO DA PROBABILIDADE DE CORRETA SELEÇÃO ENTRE AS DISTRIBUIÇÕES INVERSA-GAUSSIANA E LOG-NORMAL VIA RAZÃO DAS VEROSSIMILHANÇAS E MÉTODOS BASEADOS EM DISTÂNCIAS

Danielle PERALTA[1]

Josmar MAZUCHELI[1]

▪ RESUMO: As distribuições Inversa-Gaussiana e Log-Normal são duas distribuições bastante utilizadas na análise de dados de sobrevivência. Em muitas situações práticas estas competem entre si na modelagem de um mesmo conjunto de dados. Neste artigo busca-se discriminá-las utilizando o teste da razão das verossimilhanças e oito outros baseados em estatísticas usadas para avaliar a qualidade do ajuste. Esses métodos basicamente calculam a distância entre a função de distribuição teórica e empírica. Utilizando simulações Monte Carlo e considerando-se vários cenários estimou-se a probabilidade de correta seleção de todos os métodos. O estudo de simulação mostrou que em alguns cenários os métodos avaliados apresentam uma baixa probabilidade de correta seleção, mesmo para tamanhos de amostra razoavelmente grandes. Para fins ilustrativos seis conjuntos de dados reais, retirados da literatura, foram analisados.

▪ PALAVRAS-CHAVE: Distribuição Inversa-Gaussiana; distribuição Log-Normal; razão das verossimilhanças; função de distribuição empírica; probabilidade de correta seleção.

1 Introdução

As distribuições Inversa-Gaussiana e Log-Normal são duas dentre uma série de outras distribuições usadas para a análise de dados de sobrevivência, principalmente

[1]Universidade Estadual de Maringá – UEM, Departamento de Estatística, CEP: 87020-900, Maringá, PR, Brasil. E-mail: *danielleperalta@outlook.com*; *jmazucheli@gmail.com*

em situações em que o risco de morte ou falha, $h(t)$, é unimodal. Esta característica as tornam concorrentes naturais para a modelagem de um mesmo conjunto de dados. Embora elas apresentem risco com comportamento semelhante, tem-se uma diferença importante no seu valor para $t \to \infty$. Na distribuição Log-Normal $\lim_{t\to\infty} h(t) = 0$ enquanto que para a distribuição Inversa-Gaussiana $\lim_{t\to\infty} h(t) = 0,5\lambda\mu^{-2}$ (SESHADRI, 1999; KARADENIZ et al., 2012). O comportamento similar das densidades e por outro lado distinto das duas funções de risco, para $t \to \infty$, tem implicações na escolha de uma dentre as duas distribuições para a modelagem de um mesmo conjunto de dados. As informações intrínsecas aos dados ou oriundas da curva empírica do risco (HESS; SERACHITOPOL; BROWN, 1999) podem ser úteis para a seleção entre a Inversa-Gaussiana e a Log-Normal. Naturalmente isso é válido também para outras distribuições sob julgamento. É de se esperar que quando o risco empírico $h(t) \to 0$ para $t \to \infty$ que a distribuição Log-Normal seja mais apropriada. Por outro lado quando observa-se um decaimento assintótico para algum valor diferente de zero é de se esperar que a distribuição Inversa-Gaussiana seja mais adequada. A distribuição Inversa-Gaussiana tem se mostrado mais precisa que a distribuição Log-Normal em situações com alta variabilidade (KARMESHU; AGRAWAL, 2007) e caudas pesadas (MONTROLL; SHLESINGER, 1983; CHHIKARA; FOLKS, 1977; MARCUS, 1976).

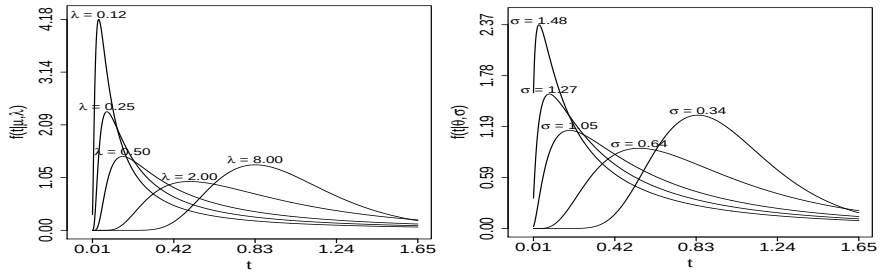
O comportamento similar das densidades e por outro lado distinto das funções de risco ocorre também quando as distribuições candidatas são a Weibull exponenciada (MUDHOLKAR; SRIVASTAVA, 1993) e a *Odd* Weibull (COORAY, 2006). Ambas apresentam, entre outras formas, densidade unimodal e risco em forma de banheira, entretanto o que se chama de período de vida útil é bem mais longo na distribuição *Odd* Weibull.

A escolha entre duas ou mais distribuições de probabilidade para o ajuste de um conjunto de observações é um problema clássico na literatura estatística. Cox (1961, 1962) foi um dos primeiros a abordar este problema tendo desenvolvido um procedimento para a discriminação entre duas distribuições. Desde então, diversos trabalhos cujo interesse recai em discriminar uma entre duas ou mais distribuições têm sido publicados.

Em análise de sobrevivência este problema tem recebido bastante atenção graças a infinidade de novas distribuições que vêm sendo propostas nos últimos anos. Trabalhos mais recentes como o de Gupta e Kundu (2003) considera o problema de discriminação entre as distribuições Weibull e Exponencial Generalizada; Kundu, Gupta e Manglick (2005) considera a discriminação entre a Log-Normal e Exponencial Generalizada; Kundu e Manglick (2005) trabalha com as distribuições Log-Normal e Gama enqualto que Kundu e Raqab (2007) considera as distribuições Log-Normal e Rayleigh. Em Dey e Kundu (2010) e Dey e Kundu (2012) são consideradas, respectivamente, as distribuições Log-Normal e Log-Logística, e Weibull e Log-Normal. Todos esses trabalhos utilizam como critérios de discriminação a razão das verossimilhanças e a distância de Kolmogorov-Smirnov.

Neste trabalho considera-se o problema de discriminação entre as distribuições Inversa-Gaussiana e Log-Normal. A motivação para a adoção destas distribuições

relaciona-se ao fato de que em muitas situações o comportamento de suas funções de densidade são praticamente indistinguíveis (ver Figura 1). Como critérios de discriminação considera-se o baseado na razão das verossimilhanças e oito outros baseados em estatísticas usadas para avaliar a qualidade do ajuste. Algumas propriedades das distribuições Inversa-Gaussiana e Log-Normal são apresentadas, respectivamente, nas Seções 2 e 3. Os procedimentos de discriminação são apresentados na Seção 4. A Seção 5 apresenta os resultados do estudo de simulação usado na estimação da probabilidade de se selecionar o modelo julgado como correto. Na Seção 6 os procedimentos de discriminação são aplicados em 6 conjuntos de dados reais da literatura. Algumas conclusões, na Seção 7, finaliza o artigo.



(a) Distribuição Inversa-Gaussiana (b) Distribuição Log-Normal

Figura 1 - Função densidade de probabilidade das distribuições Inversa-Gaussiana e Log-Normal.

2 A Distribuição Inversa-Gaussiana

A função densidade de probabilidade de uma variável aleatória T, não negativa, com distribuição Inversa-Gaussiana é dada por:

$$f(t \mid \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi t^3}} \exp\left[-\frac{\lambda\,(t-\mu)^2}{2\mu^2 t}\right] \tag{1}$$

em que $\mu > 0$ é a média da distribuição e $\lambda > 0$ é o parâmetro de forma. A função de densidade é unimodal com máximo em $T_{\max} = \mu\sqrt{1 + \left(\frac{3\mu}{2\lambda}\right)^2} - \frac{3\mu^2}{2\lambda}$. Verifica-se facilmente que a distribuição Inversa-Gaussiana pertence a família exponencial com $E(T) = \mu$ e $Var(T) = \mu^3\lambda^{-1}$.

De (1) tem-se, respectivamente, as funções de distribuição e de risco, dadas por:

$$F(t \mid \mu, \lambda) = \Phi\left[\sqrt{\frac{\lambda}{t}}\left(\frac{t}{\mu} - 1\right)\right] + \exp\left(\frac{2\lambda}{\mu}\right)\Phi\left[-\sqrt{\frac{\lambda}{t}}\left(\frac{t}{\mu} + 1\right)\right], \tag{2}$$

$$h(t \mid \mu, \lambda) = \frac{\sqrt{\frac{\lambda}{2\pi t^3}} \exp\left[-\frac{\lambda(t-\mu)^2}{2\mu^2 t}\right]}{\Phi\left[-\sqrt{\frac{\lambda}{t}}\left(\frac{t}{\mu}-1\right)\right] - \exp\left(\frac{2\lambda}{\mu}\right)\Phi\left[-\sqrt{\frac{\lambda}{t}}\left(\frac{t}{\mu}+1\right)\right]} \tag{3}$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada de uma variável aleatória com distribuição normal padrão.

Chhikara e Folks (1977) mostraram que a função de risco $h(t \mid \mu, \lambda)$ é unimodal, ou seja, cresce até um valor máximo e depois decresce assintoticamente para $0,5\lambda\mu^{-2}$. Essa característica faz da distribuição Inversa-Gaussiana uma boa candidata para a modelagem de tempos de vida quando existe uma alta probabilidade de falhas iniciais.

Em contraste com outras distribuições usadas na análise de dados de sobrevivência, os estimadores de máxima verossimilhança de μ e λ, denotados respectivamente por $\widehat{\mu}$ e $\widehat{\lambda}$ podem ser obtidos analiticamente.

3 A Distribuição Log-Normal

A função densidade de probabilidade de uma variável aleatória não negativa T com distribuição Log-Normal pode ser escrita na forma:

$$f(t \mid \theta, \sigma) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left[-\frac{1}{2}\left(\frac{\log t - \theta}{\sigma}\right)^2\right] \tag{4}$$

em que $\theta \in \mathbb{R}$ e $\sigma > 0$, são respectivamente, os parâmetros escala e forma. A função densidade é unimodal com máximo em $T_{\max} = \exp\left(\theta - \sigma^2\right)$. Verifica-se facilmente que $E(T) = \exp(\theta + 0,5\sigma^2)$ e $Var(T) = \exp(2\theta + \sigma^2)\left[\exp(\sigma^2) - 1\right]$.

De (4) tem-se, respectivamente, as funções de distribuição e de risco, dadas por:

$$F(t \mid \theta, \sigma) = \frac{1}{2}\left[1 + \Phi\left(\frac{\log t - \theta}{\sqrt{2}\sigma}\right)\right], \tag{5}$$

$$h(t \mid \theta, \sigma) = \frac{\frac{1}{\sqrt{2\pi}t\theta} \exp\left[-\frac{1}{2}\left(\frac{\log t - \theta}{\sigma}\right)^2\right]}{1 - \frac{1}{2}\left[1 + \Phi\left(\frac{\log t - \theta}{\sqrt{2}\sigma}\right)\right]}. \tag{6}$$

Da Expressão (6) verifica-se que o risco é unimodal e $\lim_{t\to\infty} h(t \mid \theta, \sigma) = 0$. Da mesma forma que na distribuição Inversa-Gaussiana, os estimadores de máxima verossimilhança de θ e σ podem ser obtidos analiticamente.

4 Procedimentos de discriminação

Esta seção apresenta os procedimentos de discriminação baseado na razão das verossimilhanças, assim como outros baseados em estatísticas usadas para avaliar

a qualidade do ajuste. Nota-se que os modelos considerados possuem o mesmo número de parâmetros. Isso faz com que a discriminação se desenvolva de forma adequada, pois caso contrário, o modelo com o maior número de parâmetros teria uma vantagem natural sobre o outro (MARSHALL; MEZA; OLKIN, 2001).

4.1 Método da Razão das Verossimilhanças

Suponha que $T_1, \ldots, T_n$ são variáveis aleatórias $i.i.d$ com distribuição Inversa-Gaussiana (Log-Normal). Sejam $l_{IG}(\widehat{\mu}, \widehat{\lambda})$ e $l_{LN}(\widehat{\theta}, \widehat{\sigma})$, respectivamente, as funções log-verossimilhanças maximizadas das distribuições Inversa-Gaussiana e Log-Normal. A estatística da razão das verossimilhanças, LR, é definida pela diferença entre os logaritmos das funções de verossimilhanças, escrita na forma:

$$LR = l_{IG}\left(\widehat{\mu}, \widehat{\lambda}\right) - l_{LN}\left(\widehat{\theta}, \widehat{\sigma}\right). \tag{7}$$

Como regra de decisão, discrimina-se em favor da distribuição Inversa-Gaussiana se $LR > 0$, caso contrário, escolhe-se a distribuição Log-Normal. A probabilidade de correta seleção, PCS, naturalmente dependerá da distribuição base. Se os dados forem originalmente provenientes de uma $IG(\mu, \lambda)$, a probabilidade de correta seleção pode ser escrita na forma:

$$PCS = P\left(LR > 0 \mid \text{Inversa-Gaussiana}\right). \tag{8}$$

No entanto, para o cálculo de (8) é necessário que se conheça a distribuição exata da estatística LR (DEY; KUNDU, 2010). Quando os modelos sob julgamento são encaixados pode-se usar como critério de seleção, dentre outras, as estatísticas da razão das verossimilhanças, Wald ou escore (LEHMANN; CASELLA, 1998; COX; HINKLEY, 1974). Por outro lado, quando isso não ocorre tem-se como alternativa utilizar a distribuição assintótica da estatística LR. Este é um problema clássico de teste de hipótese e muitas soluções foram propostas na literatura. Cox (1961, 1962) propôs um teste baseado na razão das verossimilhanças modificadas. A fim de sanar dificuldades numéricas na estatística proposta por Cox, **??**) propôs uma aproximação bastante utilizada nos recentes artigos que tratam do problema de discriminação entre distribuições não encaixadas. Sob certas condições de regularidade, Cox (1961, 1962) e **??**) mostraram que, sob a distribuição verdadeira, a estatística proposta e modificada, respectivamente por eles, é assintoticamente normal com média zero e variância v^2. Este é um resultado bastante interessante entretanto envolve a resolução de integrais que não apresentam soluções analíticas explícitas. Existem ainda dificuldades na obtenção da estimativa de v^2.

Ao leitor interessado recomenda-se as excelentes leituras de Pesaran e Ulloa (2008), Pesaran e Weeks (2001) ou Pesaran (1984) e as propostas discutidas em Vuong (1989) e Clarke (2003).

Neste artigo, como alternativa a distribuição assintótica de LR, considera-se a estimação da PCS via simulações Monte Carlo. O procedimento consiste em gerar B amostras da Inversa-Gaussiana (Log-Normal) e contar o número de vezes em que $LR > 0$ (< 0). Naturalmente, $\widehat{PCS} = B^{-1}\#\left(LR > 0\right)$.

4.2 Métodos baseados em distâncias

A distância entre a função de distribuição teórica e a função de distribuição empírica é uma grandeza natural usada para avaliar a qualidade do ajuste (THAS, 2010). Os procedimentos de discriminação baseados em distâncias, basicamente, avaliam as distâncias entre a função de distribuição acumulada e a função de distribuição empírica (LUCEÑO, 2006). Estas estatísticas são usadas em geral para a avaliação da qualidade do ajuste (D'AGOSTINO; STEPHENS, 1986) apresentadas a seguir.

Kolmogorov-Smirnov (KS)

A estatística de Kolmogorov-Smirnov é definida por:

$$D_n = \frac{1}{2n} + \max_{1\leq i\leq n}\left|z_i - \frac{i-0,5}{n}\right|. \tag{9}$$

Cramér-von-Mises (CvM)

Que tem como estatística a quantidade:

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left(z_i - \frac{i-0,5}{n}\right)^2. \tag{10}$$

Anderson-Darling (AD)

A estatística de Anderson-Darling é escrita na forma:

$$A_n^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left(2i-1\right)\left[\log z_i + \log\left(1 - z_{n+1-i}\right)\right]. \tag{11}$$

Variações do teste de Anderson-Darling

Algumas variações na estatística de Anderson-Darling são dadas pelas expressões abaixo:

Right-tail AD (ADR)

$$R_n^2 = \frac{n}{2} - 2\sum_{i=1}^{n} z_i - \frac{1}{n}\sum_{i=1}^{n} (2i-1)\log\left(1 - z_{n+1-i}\right). \tag{12}$$

Left-tail AD (ADL)

$$L_n^2 = -\frac{3n}{2} + 2\sum_{i=1}^{n} z_i - \frac{1}{n}\sum_{i=1}^{n} (2i-1)\log z_i. \tag{13}$$

Right-tail AD de Segundo Grau (AD2R)

$$r_n^2 = 2\sum_{i=1}^{n} \log\left(1 - z_i\right) + \frac{1}{n}\sum_{i=1}^{n} \frac{2i-1}{1 - z_{n+1-i}}. \tag{14}$$

Left-tail AD de Segundo Grau (AD2L)

$$l_n^2 = 2\sum_{i=1}^{n} \log z_i + \frac{1}{n}\sum_{i=1}^{n} \frac{2i-1}{z_i}. \tag{15}$$

AD de Segundo Grau (AD2)

$$a_n^2 = 2\sum_{i=1}^{n} \left[\log z_i + \log\left(1 - z_i\right)\right] + \frac{1}{n}\sum_{i=1}^{n} \left(\frac{2i-1}{z_i} + \frac{2i-1}{1 - z_{n+1-i}}\right). \tag{16}$$

em que $t_{(i)}$ são as as estatísticas de ordem da amostra, n o tamanho da amostra, $z_i = F\left(t_{(i)} \mid \widehat{\theta}\right)$ a função de distribuição acumulada localmente na estimativa de máxima verossimilhança de θ e $\frac{i-0,5}{n}$ a função de distribuição empírica.

Da mesma maneira que no teste da razão das verossimilhanças escolhe-se o modelo com o menor valor da estatística (LUCEÑO, 2006). O número de vezes que o modelo é escolhido dividido pelo número de simulações estima a probabilidade de correta seleção.

É importante mencionar que as estatísticas, dadas pelas Expressões de $(9-16)$ também podem ser usadas como métodos de estimação como alternativa ao método da máxima verossimilhança (LUCEÑO, 2006).

5 Estudo de simulação

Nesta seção serão apresentados alguns experimentos numéricos para avaliar o desempenho dos métodos de discriminação discutidos na seção anterior. Precisamente, o que se avalia é o comportamento das probabilidades de correta

seleção (PCS) para diferentes tamanhos de amostra e diferentes valores dos parâmetros.

No estudo de simulação foram geradas $B = 500.000$ amostras de tamanhos $n = 20$, 50, 80, 110, 140, 170 e 200. Os valores dos parâmetros foram tomados de modo que, independente da distribuição base, tem-se a mesma média, mesma variância e mesmo coeficiente de variação (ver Tabela 1).

Tabela 1 - Valores dos parâmetros usados no estudo de simulação

$E(X)$	1	1	1	1	1	1	1
$Var(X)$	$0,125$	$0,25$	$0,50$	1	2	4	8
CV	$0,35$	$0,50$	$0,71$	1	$1,41$	2	$2,83$
Parâmetro	Distribuição Inversa-Gaussiana						
μ	1	1	1	1	1	1	1
λ	8	4	2	1	$0,5$	$0,25$	$0,125$
Parâmetro	Distribuição Log-Normal						
θ	$-0,06$	$-0,11$	$-0,20$	$-0,35$	$-0,55$	$-0,80$	$-1,10$
σ	$0,34$	$0,47$	$0,64$	$0,83$	$1,05$	$1,27$	$1,48$

As Figuras 2 e 3 apresentam as $PCS's$ para cada cenário e método de discriminação. É importante mencionar que os valores das estatísticas baseados na qualidade do ajuste foram calculados localmente nos estimadores de máxima verossimilhança.

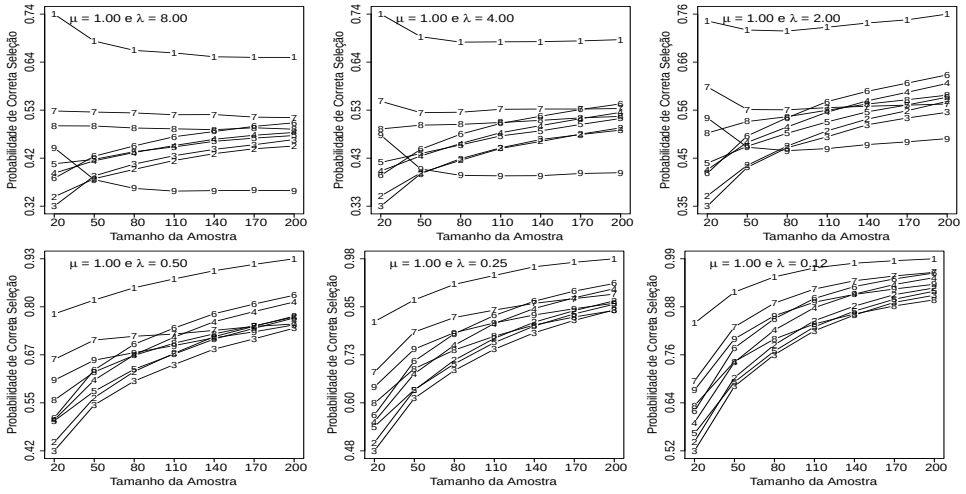


Figura 2 - Estimativa da probabilidade de correta seleção quando os dados são provenientes da distribuição Log-Normal. (1: LR, 2: CvM, 3: KS, 4: AD, 5: ADR, 6: ADL, 7: AD2R, 8: AD2L e 9: AD2).

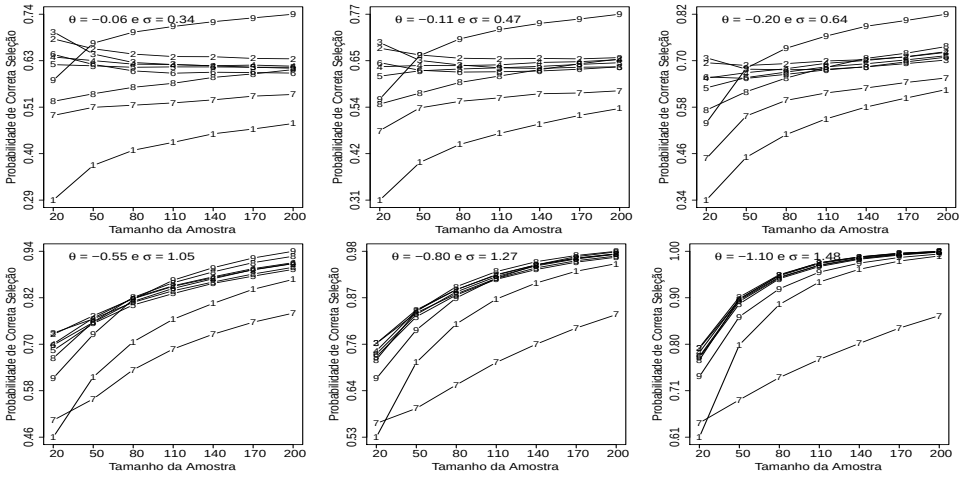


Figura 3 - Estimativa da probabilidade de correta seleção quando os dados são provenientes da distribuição Log-Normal. (1: LR, 2: CvM, 3: KS, 4: AD, 5: ADR, 6: ADL, 7: AD2R, 8: AD2L e 9: AD2).

Como era de se esperar, estas figuras evidenciam, independente da distribuição base, que a medida que o tamanho da amostra aumenta a probabilidade de correta seleção também aumenta. Apesar de um perceptível aumento, a velocidade com que a PCS cresce depende do cenário e da distribuição base.

Nota-se na Figura 2, para valores do parâmetro de forma da distribuição Inversa-Gaussiana λ maiores do que 1, que há um baixo poder discriminatório, as $PCS's$ se mantêm praticamente constante para todos os critérios de discriminação. A medida que λ aumenta, menos influência o tamanho da amostra tem sobre a PCS. Ao passo que para valores de $\lambda < 1$, o que acarreta no aumento de falhas iniciais e da variabilidade dos dados, a probabilidade de correta seleção também aumenta. Em alta variabilidade o tamanho da amostra mínimo necessário para se ter uma probabilidade de correta seleção acima de 50% é de $n \geq 50$. O critério da razão das verossimilhanças mostra-se superior aos outros critérios, mesmo em baixa variabilidade e amostras pequenas.

Na Figura 3 tem-se as probabilidades de correta seleção quando a distribuição base é a Log-Normal. Observa-se que, assim como a distribuição Inversa-Gaussiana, quanto menor o valor do parâmetro de forma σ, e consequentemente menor variabilidade dos dados, menor é a influência do tamanho da amostra na PCS. Nota-se que a PCS se mantêm praticamente constante para a grande maioria dos critérios de discriminação. A medida que σ aumenta, consequentemente aumenta a variabilidade dos dados e também a probabilidade de correta seleção. Em alta variabilidade, o tamanho mínimo da amostra necessário para uma $PCS > 55\%$, é de $n \geq 50$. O critério Anderson-Darling de segundo grau mostra-se superior aos outros critérios mesmo em baixa variabilidade e amostras pequenas.

Desse modo tem-se que o critério da razão das verossimilhanças é mais eficiente quando a distribuição base é a Inversa-Gaussiana, mesmo para amostras pequenas. Os outros critérios de discriminação baseados nas estatísticas que avaliam a qualidade ao ajuste (com excessão do critério $AD2R$) apresentam comportamentos similares e são mais eficientes quando a distribuição base é a Log-Normal, para todos tamanhos de amostras. Para a distribuição Log-Normal o critério da razão das verossimilhanças mostra-se ineficiente para $n < 50$ e de acordo com os valores de σ.

6 Análise de dados reais

Nesta seção as distribuições Log-Normal e Inversa-Gaussiana são ajustadas a seis conjuntos de dados reais retirados da literatura. O objetivo é aplicar os procedimentos de discriminação, descritos anteriormente, para a seleção do modelo mais apropriado. É importante mencionar que os 6 conjuntos de dados foram analisados na literatura segundo o ajuste da distribuição Log-Normal.

As estimativas da máxima verossimilhança e os valores das estatísticas baseados em distâncias foram calculadas, respectivamente, no procedimento *SAS/NLMIXED* (SAS, 2010) e a partir das funções disponíveis na biblioteca *fitdistrplus*, (DELIGNETTE-MULLER et al., 2014), do ambiente *R* (R Development Core Team, 2011).

Dados 1: O primeiro conjunto de dados ($n = 30$), extraído de Kumagai et al. (1989), indicam as concentrações médias diárias do tolueno no ar (medidas em $TWAs/8h$) durante um período de trinta dias em uma determinada fábrica. Os valores observados foram:

13,6 11,1 66,9 4,9 5,8 6,9 4,9 2,1 3,2 8,3 10,5 1,1 17,4 0,9 50,9 7,8
2,6 21,9 3,1 2,9 6,2 58,3 20,4 22,4 16,6 8,7 5,2 58,6 1,9 57,4

Dados 2: Uma amostra com as concentrações de clorobenzeno no ar (dado em $TWAs/15min$) de uma fábrica foi apresentada em Kumagai e Matsunaga (1995). Os 31 valores observados são referentes às concentrações médias calculadas a cada quinze minutos em um turno de oito horas. As médias foram:

13,7 10,2 9,9 4,3 5,6 45,6 42,0 14,1 3,8 9,3 10,6 91,3 2,2 3,8 6,0 17,8
131,8 31,0 4,2 2,6 27,6 1,7 7,0 2,1 1,5 7,5 2,5 2,4 51,9 12,9 12,3

Dados 3: Em Gupta, Kannan e Raychaudhuri (1997) analisou-se os tempos de sobrevivência (em dias) de cobaias que receberam injeções de bacilos causadores da tuberculose. Os tempos até a morte das 72 cobaias foram:

12 15 22 24 24 32 32 33 34 38 38 43 44 48 52 53 54 54 55 56 57 58
58 59 60 60 60 60 61 62 63 65 65 67 68 70 70 72 73 75 76 76 81 83
84 85 87 91 95 96 98 99 109 110 121 127 129 131 143 146 146 175
175 211 233 258 258 263 297 341 341 376

Dados 4: Chhikara e Folks (1977) analisou 46 observações referentes a tempos de reparo (em horas) de um transceptor aéreo:

0,2 0,3 0,5 0,5 0,5 0,5 0,6 0,6 0,7 0,7 0,7 0,8 0,8 1,0 1,0 1,0 1,0 1,1
1,3 1,5 1,5 1,5 1,5 2,0 2,0 2,2 2,5 2,7 3,0 3,0 3,3 3,3 4,0 4,0 4,5 4,7
5,0 5,4 5,4 7,0 7,5 8,8 9,0 10,3 22,0 24,5

Dados 5: Um estudo com tempos (em dias) de conservação de um produto alimentício foi apresentada em Gacula e Kubala (1975). Os tempos de conservação das 26 amostras analisadas desse produto foram:

24 24 26 26 32 32 33 33 33 35 41 42 43 47 48 48 48 50 52 54 55 57
57 57 57 61

Dados 6: Os dados apresentados em Chen (2006) foram coletados em um processo de fabricação de um laminado plástico (supostamente sob controle estatístico) cuja resistência deve ultrapassar alguns quilos mínimos por polegada quadrada (psi). Os dados abaixo são referentes a resistências de laminado plástico (em psi) de 49 amostras:

21,87 23,8 24,83 25,8 29,95 30,26 31,23 31,29 31,86 32,48 33,38 33,73
33,88 33,93 34,03 34,5 34,9 35,57 35,66 39,44 41,76 41,96 42,21 42,66
43,27 43,41 44,06 45,32 47,39 47,98 48,81 50,76 51,54 54,67 54,92
55,33 57,24 59,3 60,41 60,89 61,63 68,93 71,96 72,65 73,51 76,15
78,48 81,37 99,43

As estimativas de máxima verossimilhança para os parâmetros das distribuições Inversa-Gaussiana e Log-Normal considerando os conjuntos de dados descritos acima, encontram-se nas Tabelas 2 e 3. Estas tabelas apresentam também os valores da função de log-verossimilhança calculadas localmente nas estimativas de máxima verossimilhança e serão usadas para a discriminação do modelo mais apropriado. São apresentadas também as estimativas das médias e desvios padrão.

Na Tabela 4 tem-se os valores das estatísticas calculadas para os nove métodos de discriminação. No método da razão das verossimilhanças a regra de decisão é favorável a distribuição Inversa-Gaussiana se $LR > 0$. Nos outros métodos, a regra de decisão é favorável ao modelo que apresenta o menor valor da estatística. Com base nos valores das estatísticas os conjuntos de dados 1 e 3 a distribuição Log-Normal é escolhida como a melhor opção. Para os outros conjuntos de dados, a distribuição Inversa Gaussiana foi a melhor.

A Figura 4 mostra a curva da distribuição empírica acumulada contra função de distribuição ajustadas pelas distribuições Inversa-Gaussiana e Log-Normal. Observa-se que a sobreposição das curvas de sobrevivência teóricas às curvas empíricas são praticamente indistinguíveis.

Tabela 2 - Estimativas de máxima verossimilhança (erros padrão) e valores da função log-verossimilhança para a distribuição Inversa-Gaussiana

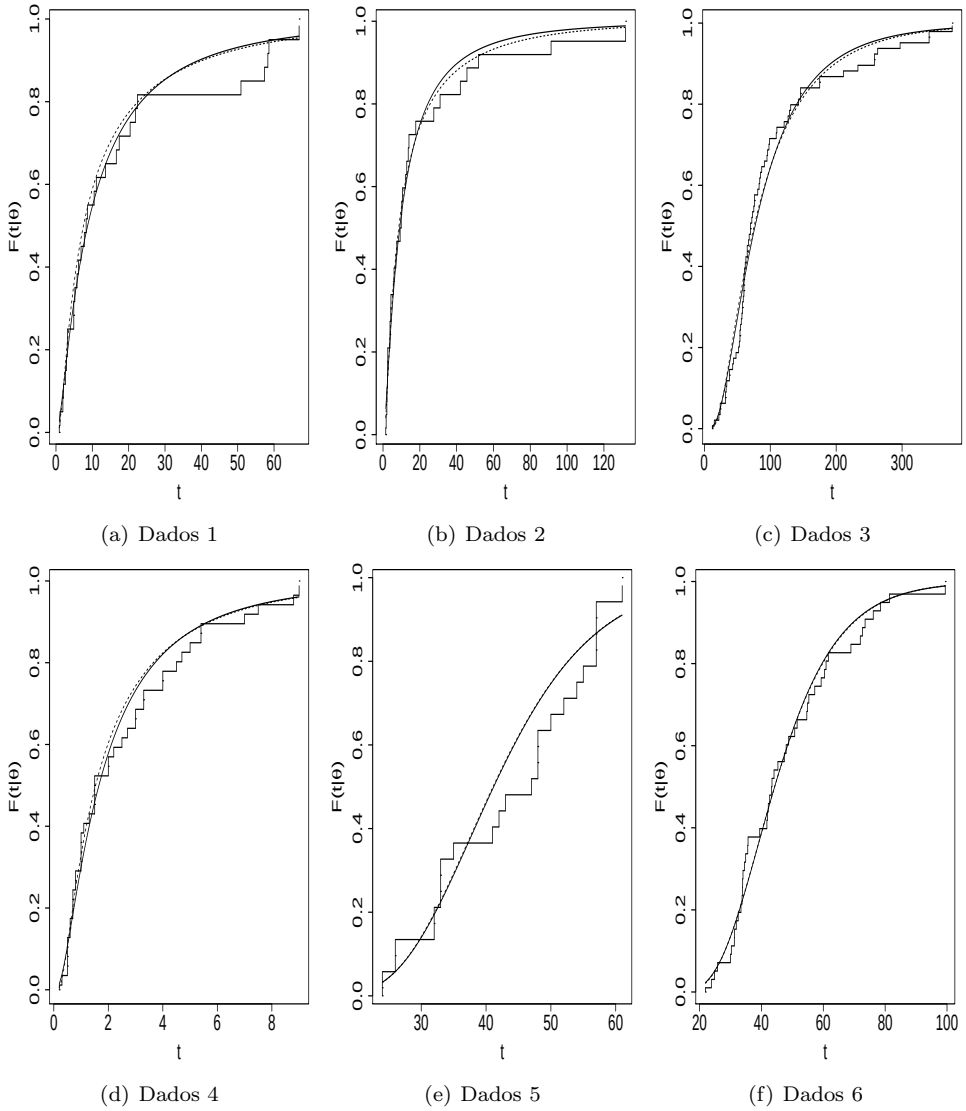
Dados	$\widehat{\mu}$	$\widehat{\lambda}$	$\widehat{E(X)} = \widehat{\mu}$	$\widehat{S(X)}$	$-2 \times l\left(\widehat{\mu}, \widehat{\lambda} \mid t\right)$
1	16,7500 (4,9228)	6,4641 (1,6690)	16,7500 (4,9228)	26,9631 (12,3857)	223,9328
2	19,0062 (5,5336)	7,2326 (1,8371)	19,0062 (5,5336)	30,8102 (14,0127)	231,6046
3	99,8194 (9,5639)	151,02 (25,1704)	99,8194 (9,5639)	81,1525 (13,4819)	781,4446
4	3,6065 (0,7841)	1,6589 (0,3459)	3,6065 (0,7841)	5,3177 (1,8206)	198,1186
5	42,8846 (2,5028)	484,25 (134,31)	42,8846 (2,5028)	12,7619 (2,0929)	203,0396
6	47,1508 (2,4434)	358,34 (72,3962)	47,1508 (2,4434)	17,1035 (2,1800)	408,0334

Tabela 3 - Estimativas de máxima verossimilhança (erros padrão) e valores das função log-verossimilhança para a distribuição Log-Normal

Dados	$\widehat{\theta}$	$\widehat{\sigma}$	$\widehat{E(X)}$	$\widehat{S(X)}$	$-2 \times l\left(\widehat{\theta}, \widehat{\sigma} \mid t\right)$
1	2,1643 (0,2148)	1,1765 (0,1519)	17,3977 (4,8611)	30,0905 (14,1178)	224,7448
2	2,2039 (0,2107)	1,1733 (0,1490)	18,0337 (4,9378)	31,0336 (14,2699)	234,5268
3	4,3443 (0,08372)	0,7104 (0,05920)	99,1514 (9,2895)	80,3328 (13,6721)	780,6704
4	0,6584 (0,1625)	1,1018 (0,1149)	3,5444 (0,7299)	5,4527 (1,8920)	200,0326
5	3,7180 (0,05735)	0,2924 (0,04055)	42,9801 (2,5170)	12,8420 (2,1410)	203,1834
6	3,7903 (0,05040)	0,3528 (0,03564)	47,1123 (2,4474)	17,1531 (2,2327)	408,4086

Tabela 4 - Distâncias entre os métodos de discriminação

Dados	Dist.	LR	CvM	KS	AD	ADR	ADL	AD2R	AD2L	AD2
1	IG	-0,4060	0,033	0,095	0,303	0,202	0,101	3,029	1,829	4,858
	LN		0,029	0,099	0,282	0,207	0,075	3,257	1,118	4,376
2	IG	-1,4611	0,033	0,088	0,234	0,095	0,139	1,130	1,871	3,001
	LN		0,049	0,093	0,365	0,164	0,202	2,112	2,633	4,745
3	IG	0,3871	0,161	0,111	0,882	0,405	0,477	3,422	8,182	11,605
	LN		0,150	0,097	0,833	0,436	0,397	4,457	2,750	7,208
4	IG	-0,9570	0,062	0,092	0,395	0,234	0,162	2,570	2,461	5,031
	LN		0,071	0,094	0,436	0,227	0,209	2,452	1,629	4,081
5	IG	-0,0719	0,148	0,176	0,912	0,519	0,393	3,934	3,188	7,122
	LN		0,147	0,173	0,916	0,507	0,409	3,889	3,443	7,332
6	IG	-0,1876	0,063	0,116	0,378	0,160	0,218	1,510	1,960	3,470
	LN		0,068	0,119	0,412	0,181	0,231	1,658	2,037	3,696



(a) Dados 1 (b) Dados 2 (c) Dados 3

(d) Dados 4 (e) Dados 5 (f) Dados 6

Figura 4 - Função distribuição empírica e distribuições teóricas ajustadas (linha tracejada: distribuição Inversa-Gaussiana, linha contínua: distribuição Log-Normal e escada: função distribuição empírica).

7 Considerações finais

Neste artigo, considerou-se o problema da discriminação entre as funções de distribuição de probabilidade Log-Normal e Inversa-Gaussiana. Foram utilizados o método da razão das verossimilhanças e vários métodos baseados em distâncias. Os desempenhos dos métodos de discriminação (Seção 5) foram verificados via simulações Monte Carlo. Precisamente, o que se avaliou foram os comportamentos das probabilidades de correta seleção (PCS), considerando diferentes cenários, ou seja, diferentes tamanhos de amostra e diferentes valores dos parâmetros de forma. Observou-se que a medida que o tamanho da amostra aumenta a probabilidade de correta seleção também aumenta, no entanto a velocidade desse aumento depende do cenário e da distribuição base. O método baseado na razão das verossimilhanças, para os cenários considerados, apresentou uma melhor performace, principalmente quando a distribuição base era a Inversa-Gaussiana.

Foram analisados também seis conjuntos de dados reais para fins ilustrativos. Apesar do ajuste dos modelos serem próximos (Figura 6) a distribuição Inversa-Gaussiana foi mais vezes escolhida como a melhor opção. Observou-se que em determinados casos, a escolha da melhor distribuição são dadas de forma clara, no entanto para alguns conjuntos de dados essa escolha não é tão simples.

Agradecimentos

Os autores agradecem aos pareceristas pelos valiosos comentários e sugestões dadas ao artigo e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

PERALTA, D.; MAZUCHELI, J. Estimation of the probability of correct selection between distributions Inverse-Gaussian and Log-Normal via ratio of likelihoods and methods based on distances. *Rev. Bras. Biom.,* São Paulo, v.32, n.4, p.553-569, 2014.

▪ *ABSTRACT: The Inverse-Gaussian and Log-Normal distribution are two distributions widely used in the analysis of survival data. In many practical situations they compete among themselves in the modeling of the same data set. This paper seeks discriminate between the two distributions using the likelihood ratio test and eight other based on statistics used to evaluate the goodness of fit. These methods basically calculates the distance between the theoretical distribution function and empirical distribution function. Using Monte Carlo simulations and considering various scenarios, the probability of correct selection was estimated for all methods. The simulation study showed that in some scenarios the methods evaluated have a low probability of correct selection, even for reasonably large sample sizes. For illustrative purposes six real data sets, taken from the literature, were analyzed.*

▪ *KEYWORDS: Inverse-Gaussian distribution; Log-Normal distribution; likelihood ratio test; empirical distribution function; probability of correct selection*

Referências

CHEN, C. Tests of fit for the three-parameter lognormal distribution. *Computational Statistics & Data Analysis*, v. 50, n. 6, p. 1418–1440, 2006.

CHHIKARA, R. S.; FOLKS, J. L. The inverse gaussian distribution as a lifetime model. *Technometrics*, v. 19, n. 4, p. 461–468, 1977.

CLARKE, K. Nonparametric model discrimination in international relations. *Journal of Conflict Resolution*, v. 47, n. 1, 2003.

COORAY, K. Generalization of the Weibull distribution: the odd Weibull family. *Statistical Modelling. An International Journal*, v. 6, n. 3, p. 265–277, 2006.

COX, D. R. Tests of separate families of hypotheses. In: PROC. 4th BERKELY SYMPOS. MATH. STATIST. AND PROB., Vol. I. California: Univ. California Press, 1961. p. 105–123.

COX, D. R. Further results on tests of separate families of hypotheses.*Journal of the Royal Statistical Society. Series B. Methodological*, v. 24, p. 406–424, 1962.

COX, D. R.; HINKLEY, D. V. *Theoretical statistics*. [S.l.]: Chapman and Hall, 1974.

D'AGOSTINO, R. B.; STEPHENS, M. A. (Ed.). *Goodness-of-fit techniques*. [S.l.]: Marcel Dekker, Inc., 1986. (Statistics, Textbooks and Monographs, v. 68).

DELIGNETTE-MULLER, M. L.; DUTANG, C. *fitdistrplus: help to fit of a parametric distribution to non-censored or censored data*. [S.l.], 2014. R package version 1.0-2.

DEY, A. K.; KUNDU, D. Discriminating between the log-normal and log-logistic distributions. *Communications in Statistics. Theory and Methods*, v. 39, n. 1-2, p. 280–292, 2010.

DEY, A. K.; KUNDU, D. Discriminating between the Weibull and log-normal distributions for type-II censored data. *Statistics. A Journal of Theoretical and Applied Statistics*, v. 46, n. 2, p. 197–214, 2012.

GACULA, M.; KUBALA, J. Statistical models for shelf life failures.*Food Sci.*, v. 40, p. 404–409, 1975.

GUPTA, R. C.; KANNAN, N.; RAYCHAUDHURI, A. Analysis of lognormal survival data.*Mathematical Biosciences*, v. 139, n. 2, p. 103 – 115, 1997.

GUPTA, R. D.; KUNDU, D. Discriminating between Weibull and generalized exponential distributions.*Computational Statistics & Data Analysis*, v. 43, n. 2, p. 179–196, 2003.

HESS, K. R.; SERACHITOPOL, D. M.; BROWN, B. W. Hazard function estimators: a simulation study.*Statistics in Medicine*, v. 18, n. 22, p. 3075–3088, 1999.

KARADENIZ, P. G. et al. Comparison of inverse Gaussian distribution with survival analysis in advanced chronic heart failure patients.*International Journal of Cardiology*, v. 155, n. 3, p. 508 – 509, 2012.

KARMESHU, V. K.; AGRAWAL, R. On efficacy of Rayleigh-inverse gaussian distribution over K-distribution for wireless fading channels. *Wireless Communications and Mobile Computing*, v. 7, n. 1, p. 1–7, 2007.

KUMAGAI, S.; MATSUNAGA, I. Changes in the distribution of short-term exposure concentration with different averaging times. *American Industrial Hygiene Association Journal*, v. 56, n. 1, p. 24–31, 1995.

KUMAGAI, S. et al. Assessment of occupational exposures to industrial hazardous sustances. *Japanese Journal of Industrial Health*, v. 31, p. 216–226, 1989.

KUNDU, D.; GUPTA, R. D.; MANGLICK, A. Discriminating between the log-normal and generalized exponential distributions. *Journal of Statistical Planning and Inference*, v. 127, n. 1-2, p. 213–227, 2005.

KUNDU, D.; MANGLICK, A. Discriminating between the log-normal and gamma distributions. *Journal of Applied Statistical Science*, v. 14, n. 1-2, p. 175–187, 2005.

KUNDU, D.; RAQAB, M. Z. Discriminating between the generalized Rayleigh and log-normal distribution. *Statistics. A Journal of Theoretical and Applied Statistics*, v. 41, n. 6, p. 505–515, 2007.

LEHMANN, E. L.; CASELLA, G. *Theory of point estimation*. Second. ed. New York: Springer-Verlag, 1998. (Springer Texts in Statistics).

LUCEÑO, A. Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators. *Computational Statistics & Data Analysis*, v. 51, n. 2, p. 904–917, 2006.

MARCUS, A. H. Power sum distributions: An easier approach using the wald distribution. *Journal of the American Statistical Association*, v. 71, n. 353, p. 237–238, 1976.

MARSHALL, A. W.; MEZA, J. C.; OLKIN, I. Can data recognize its parent distribution? *Journal of Computational and Graphical Statistics*, v. 10, n. 3, p. 555–580, 2001.

MONTROLL, E. W.; SHLESINGER, M. F. Maximum entropy formalism, fractals, scaling phenomena, and $1/f$ noise: a tale of tails. *J. Statist. Phys.*, v. 32, n. 2, p. 209–230, 1983.

MUDHOLKAR, G. S.; SRIVASTAVA, D. K. Exponentiated weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability*, v. 42, n. 2, p. 299–302, 1993.

PESARAN, M. H. Asymptotic power comparisons of tests of separate parametric families by Bahadur's approach. *Biometrika*, v. 71, n. 2, p. 245–252, 1984.

PESARAN, M. H.; ULLOA, M. R. D. non-nested hypotheses. In: DURLAUF, S. N.; BLUME, L. E. (Ed.). *The New Palgrave Dictionary of Economics*. Basingstoke: Palgrave Macmillan, 2008.

PESARAN, M. H.; WEEKS, M. Nonnested hypothesis testing: an overview. In: *A companion to theoretical econometrics*. Massachusetts: Blackwell, Malden, 2001, (Blackwell Companions Contemp. Econ.). p. 279–309.

R Development Core Team. *R: A Language and Environment for Statistical Computing.*Vienna, Austria, 2011.

SAS. *The NLMIXED Procedure, SAS/STAT® User's Guide, Version 9.22*. Cary, NC: SAS Institute Inc.: [s.n.], 2010. p. 4967–5062

SESHADRI, V. *The inverse Gaussian distribution*. New York: Springer-Verlag, 1999. (Lecture Notes in Statistics, v. 137).

THAS, O. *Comparing distributions*. New York: Springer, 2010. (Springer Series in Statistics).

VUONG, Q. H. Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica*, v. 57, n. 2, p. 307–333, 1989.

Recebido em 10.09.2014.

Aprovado após revisão em 05.12.2014.