

Análise do Tempo de Desmame de Suínos Utilizando a Distribuição Gama Generalizada

Fellipe Tavares Nogueira Cruz¹, Marcelino Alves Rosa de Pascoa², Carlos Alberto de Sousa Gondim³

Introdução

A distribuição gama generalizada foi introduzida por Stacy (1962) é de grande destaque em análise de sobrevivência por possuir como casos particulares outras distribuições muito usadas. A distribuição gama generalizada (GG) é caracterizada por três parâmetros, α , τ e k , todos positivos, sua função de densidade é dada por:

$$g(t) = \frac{\tau}{\Gamma(k)\alpha^{\tau k}} t^{\tau k-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\tau}\right\}, t > 0.$$

Em que $\Gamma(k)$ é a função gama definida por:

$$\Gamma(k) = \int_0^{\infty} t^{k-1} e^{-t} dt.$$

Se T é uma variável aleatória positiva e tem distribuição de probabilidade GG com parâmetros α , τ e k , então denota-se $T \sim GG(\alpha, \tau, k)$. As funções de distribuição acumulada $G(t)$, de sobrevivência $S(t)$ e de taxa de falha $h(t)$ são escritas, respectivamente por:

$$G(t) = P[T \leq t] = \frac{1}{\Gamma(k)} \int_0^{\left(\frac{t}{\alpha}\right)^{\tau}} u^{k-1} e^{-u} du = \left\{ \gamma_1 \left[k, \left(\frac{t}{\alpha} \right)^{\tau} \right] \right\},$$

$$S(t) = 1 - G(t) = 1 - \left\{ \gamma_1 \left[k, \left(\frac{t}{\alpha} \right)^{\tau} \right] \right\} e$$

$$h(t) = \frac{g(t)}{S(t)} = \frac{t^{\tau k-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\tau}\right\}}{\int_0^{\infty} x^{\tau k-1} \exp\left\{-\left(\frac{x}{\alpha}\right)^{\tau}\right\} dx}$$

¹UFMT. email: felliperambo@live.com

²UFMT. email: marcelino.pascoa@gmail.com

³UFMT. email: gondim1.cag@gmail.com

Sendo $\gamma(k, x) = \int_0^x u^{k-1} e^{-u} du$ é a função gama incompleta e $\gamma_1(k, x) = \frac{\gamma(k, x)}{\Gamma(k)} = \frac{1}{\Gamma(k)} \int_0^x w^{k-1} e^{-w} dw$ é a razão da função gama incompleta.

Objetivo

O objetivo desse trabalho foi estimar os parâmetros da distribuição GG, através do método de máxima verossimilhança, utilizando dados de desmame de suínos e comparar o seu melhor ajuste com o estimador de Kaplan-Meier.

Metodologia

Os dados foram disponibilizados por uma granja localizada em Campo Verde, Mato Grosso (MT). O período de coleta dessa amostra foi Setembro de 2018, utilizando-se, para a análise 50 porcas. Foi considerado como variável resposta o período do nascimento à desmama, contudo, nesse estudo não houve censura.

Uma forma empírica de determinar o comportamento da função risco se dá por meio da construção do gráfico do tempo total em teste (curva TTT), proposto por Aarset (1987). A curva TTT é obtida construindo um gráfico através da seguinte expressão:

$$G\left(\frac{r}{n}\right) = \frac{\sum_{i=1}^r T_{i:n} + (n-r)T_{r:n}}{\sum_{i=1}^n T_{i:n}} \quad \text{por} \quad \frac{r}{n},$$

em que n é o tamanho da amostra, $r = 1, \dots, n$ e $T_{i:n}$, $i = 1, \dots, n$ são estatísticas de ordem da amostra.

A estimação dos parâmetros foi feita pelo método de máxima verossimilhança. Para que fosse possível realizar inferências fundamentadas no modelo, foi necessário obter a função de verossimilhança, logo considere uma amostra $X = (x_1, \dots, x_n)$ de n observações independentes, e seja F o conjunto do logaritmo dos tempos de vida e C o conjunto do logaritmo dos tempos de censura. Assim, o logaritmo da função de verossimilhança para o vetor de parâmetros $\theta = (\alpha, \tau, k)^T$ para o modelo (1) tem a forma $l(\theta) = \sum_{i \in F} l_i(\theta) + \sum_{i \in C} l_i^{(c)}(\theta)$, em que $l_i(\theta) = \log[f(x_i)]$, $l_i^{(c)}(\theta) = \log[S(x_i)]$, $f(x_i)$ é a função densidade de probabilidade (1) e $S(x_i)$ é a função de sobrevivência (2). Dessa forma, o logaritmo da função de verossimilhança para θ é

$$l(\theta) = r \log(\tau) - r \log[\alpha \Gamma(k)] + (\tau k - 1) \sum_{i \in F} \log\left(\frac{t_i}{\alpha}\right) - \sum_{i \in F} \left(\frac{t_i}{\alpha}\right)^\tau + \sum_{i \in C} \log\left(1 - \left\{\gamma_1\left[k, \left(\frac{t_i}{\alpha}\right)^\tau\right]\right\}\right)$$

$$+ \sum_{i \in C} \log \left(1 - \left\{ \gamma_1 \left[k, \left(\frac{t_i}{\alpha} \right)^\tau \right] \right\} \right), \quad (1)$$

em que r é o número de falhas, F e C denotam o conjunto de observações não censuradas e censuradas, respectivamente.

As derivadas de primeira ordem de (1) em relação aos parâmetros em θ são dadas por

$$\begin{aligned} U_\alpha(\theta) &= -\frac{r\tau k}{\alpha} + \frac{\tau}{\alpha} \sum_{i=1}^r u_i + \sum_{i=r+1}^n \frac{-\tau(u_i)^k \exp(-u_i)}{\alpha \Gamma(k) \{1 - [\gamma_1(k, u_i)]\}}, \\ U_\tau(\theta) &= \frac{r}{\tau} + \frac{k}{\tau} \sum_{i=1}^r \log(u_i) - \frac{1}{\tau} \sum_{i=1}^r u_i \log(u_i) + \sum_{i=r+1}^n \frac{(u_i)^k \exp(-u_i) \log(u_i)}{\tau \Gamma(k) \{1 - [\gamma_1(k, u_i)]\}}, \\ U_k(\theta) &= -r\psi(k) + \sum_{i=1}^r \log(u_i) + \sum_{i=r+1}^n \left(\frac{-\psi(k) \gamma_1(k, u_i)}{\{1 - [\gamma_1(k, u_i)]\}} + \frac{[\dot{\gamma}(k, u_i)]_k}{\Gamma(k) \{1 - [\gamma_1(k, u_i)]\}} \right), \end{aligned}$$

em que, $\psi(\cdot)$ é a função digama, $u_i = \left(\frac{t_i}{\alpha}\right)^\tau$, $[\dot{\gamma}(k, u_i)]_k = \int_0^{u_i} w^{k-1} \exp(-w) \log(w) dw$ e n é o tamanho da amostra.

Consequentemente, o estimador de máxima verossimilhança (EMV) $\tilde{\theta}$ de θ é obtido numericamente a partir das equações não lineares

$$U_\alpha(\theta) = U_\tau(\theta) = U_k(\theta) = 0.$$

Para a estimação de intervalo e testes de hipóteses sobre os parâmetros do modelo é necessário obter a matriz 3×3 de informação observada

$$J = J(\theta) = \begin{pmatrix} j_{\alpha, \alpha} & j_{\alpha, \tau} & j_{\alpha, k} \\ j_{\tau, \alpha} & j_{\tau, \tau} & j_{\tau, k} \\ j_{k, \alpha} & j_{k, \tau} & j_{k, k} \end{pmatrix}$$

O ajuste da distribuição GG, foi comparado com o da distribuição Exponencial, por meio do teste da razão de verossimilhança (TRV). Pode-se usar a estatística TRV para checar se o ajuste usando a distribuição GG é estatisticamente “superior” ao ajuste usando a distribuição Exponencial para um determinado banco de dados. Em qualquer caso, o teste de hipóteses do tipo $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$, onde θ_0 é um vetor especificado, pode ser realizado usando a estatística TRV. O teste de $H_0: k = \tau = 1$ versus $H_1: H_0 \text{ não é verdade}$, é equivalente a comparar a distribuição GG com a distribuição Exponencial para o qual a estatística TRV reduz a

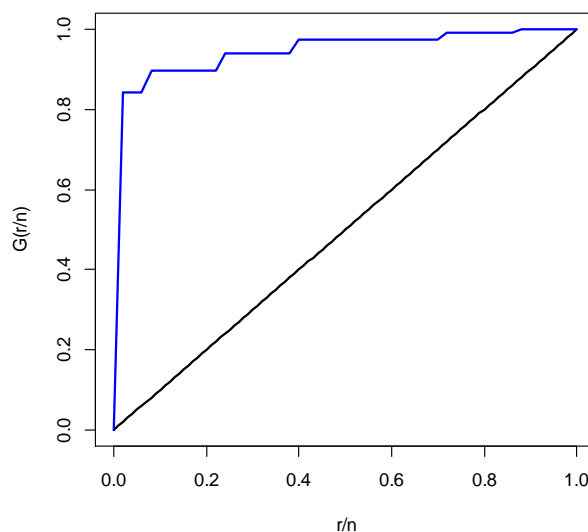
$$w = 2[\ell(\hat{\alpha}, \hat{\tau}, \hat{k}) - \ell(\tilde{\alpha}, 1, 1)],$$

em que $\hat{\alpha}$, $\hat{\tau}$ e \hat{k} são os estimadores de máxima verossimilhança sobre H_1 e $\tilde{\alpha}$ e $\tilde{\tau}$, são os estimadores sobre H_0 . Foram utilizadas as estatísticas AIC, BIC e CAIC para a seleção dos modelos. As análises foram implementadas no *software* estatístico R.

Resultados e discussões

A Curva TTT para o conjunto de dados, do período do nascimento à desmama, encontra-se na Fig. 1 e, indica uma função risco na forma crescente, razão pela qual a distribuição GG pode ser perfeitamente utilizada.

Figura 1: Curva TTT para dados de tempo até o desmame de suínos.



Na Tabela 1, podem ser vistos as estimativas de máxima verossimilhança (e os correspondentes erros-padrão que estão entre parênteses) dos parâmetros e os valores das estatísticas dos modelos, Exponencial e GG. Os resultados indicam que o modelo GG tem os menores valores das estatísticas AIC (Critério de Informação de Akaike), BIC (Critério de Informação Bayesiano) e CAIC (Critério de Informação Akaike Consistente) entre os modelos ajustados, portanto, o modelo GG é o mais adequado para os dados de tempo até o desmame de suínos. O teste da razão de verossimilhança é apresentado na Tabela 2. Os resultados nessa tabela sugerem que o modelo GG produz um ajuste mais adequado a esses dados quando comparado com a distribuição exponencial.

Tabela 1: Ajuste final dos modelos comparados, para os dados de tempo até o desmame de suínos

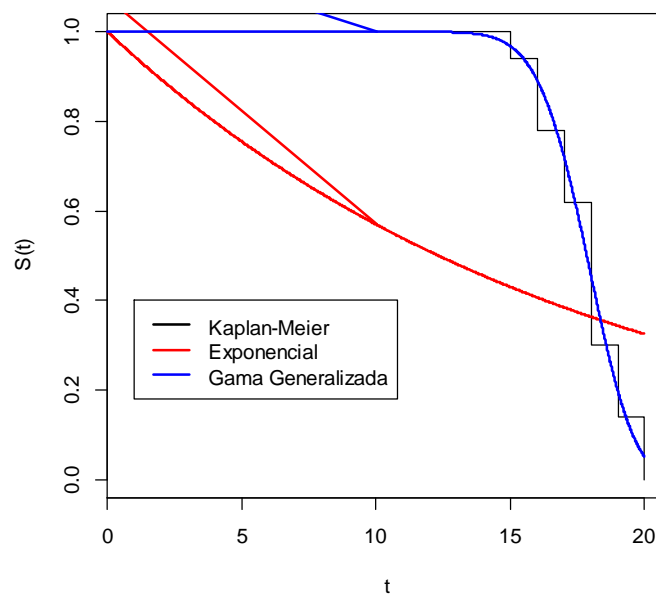
Modelo	α	τ	k	AIC	CAIC	BIC
GG	14.7092	6.6128	3.9174	182.4	183.8	188.2
	(7.1814)	(6.3939)	(7.0275)			
Exponencial	17.77448	1	1	389.8	390.3	391.7
	(2.5129)	(-)	(-)			

Tabela 2: Teste da razão de verossimilhança, para os dados de tempo até o desmame de suínos

Modelo	Estatística do Teste	Valor p
GG vs Exponencial	211.4	< 0,0001

A Figura 2 apresenta a comparação das estimativas da função de sobrevivência segundo Kaplan-Meier e segundo os modelos GG e exponencial, para os dados de tempo até o desmame de suínos. Observa-se pela figura que a distribuição GG nos fornece um ajuste satisfatório para os dados em estudo. Logo, conclui-se que em média o tem até o desmame dos suínos é de 17,8 dias. Com 19 dias 80% dos suínos haviam desmamado.

Figura 2: Estimativas da função de sobrevivência segundo Kaplan-Meier e segundo os modelos GG e Exponencial, para os dados de tempo até o desmame de suíno



Conclusão

A distribuição Gama Generalizada proposta por Stacy (1962), apresentou melhor ajuste para os dados em estudo, segundo o teste da razão de verossimilhança e as estatísticas AIC, BIC e CAIC, quando comparada com a distribuição exponencial, se mostrando mais flexível.

Referências bibliográficas

AARSET, M. V. How to identify bathtub hazard rate. **Transaction son Reliability**. v. 36, p. 106-108, 1987.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.

STACY, E.W. A generalization of the gamma distribution. The Annals of Mathematical Statistics, New York, v.33, p.409-419, 1962.