INCOMPLETE GAMMA DISTRIBUTION: A NEW TWO PARAMETER LIFETIME DISTRIBUTION WITH SURVIVAL REGRESSION MODEL

Aliakbar Rasekhi[1], Mahdi Rasekhi[2], G.G. Hamedani[3]

We introduce a new two parameter lifetime distribution constructed via incomplete gamma function which includes exponential distribution as a limiting case. This distribution is more flexible than most of the two parameter extended exponential distributions. Various statistical properties such as moments, moment generating function and certain useful characterizations based on the ratio of two truncated moments are presented. Maximum likelihood estimation method is used for estimating parameters of this distribution and a survival regression model based on the proposed distribution is presented for fitting breast cancer data set.

Keywords: Incomplete gamma function, Accelerated failure time regression model, Characterizations, Limiting distribution.
MSC2010: 60E05, 62N01.

1. Introduction

In recent decades, several new distributions have been introduced based on the exponential distribution, which is a widely used distribution in many survival analysis problems. The main goal of this paper is to propose a two parameter lifetime distribution which includes exponential distribution and can accommodate practical applications where the underlying hazard functions have non-constant monotone shapes. In some real data applications, the exponential distribution does not provide a reasonable parametric fit, thus some researchers extend exponential distribution by adding one or more parameters e.g., [10], [12] and [3], among others.

In what follows, we use exponential integral, which is closely related to the incomplete gamma function, to introduce our distribution.

Generally,

$$\mathrm{E}_k(\lambda) = \int_1^\infty t^{-k} e^{-\lambda t} dt = \lambda^{k-1} \int_\lambda^\infty u^{-k} e^{-u} du = \lambda^{k-1} \Gamma(1-k, \lambda), \tag{1}$$

[1]Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran, e-mail: rasekhi@modares.ac.ir (Corresponding author)

[2]Department of Statistics, Faculty of Mathematical Sciences and Statistics, Malayer University, Malayer, Iran

[3]Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, WI, USA

where $k = 0, 1, 2, \ldots$ and $\lambda > 0$; see [1]. We need $\Gamma_0(\lambda)$ function for normalizing constant of the proposed distribution and it is obtained by setting $k = 1$ in (1), i.e.

$$\mathrm{E}_1(\lambda) = \int_\lambda^\infty u^{-1} e^{-u} du = \Gamma(0, \lambda) \equiv \Gamma_0(\lambda).$$

Another result of this paper is the introduction of a new survival regression model. In the last decades, many survival regression models were studied by many researchers, for example: [14], [5], [16], [2] etc. The main distributions of these regression models, however, have more than two parameters and only a few survival regression models exist based on two parameter lifetime distribution. The Weibull regression model is a well known and powerful regression model in survival analysis (see [11]) depending on two parameters distribution. In this paper, we present another survival regression model that is based on a new two parameters distribution. We show that our proposed regression model is a better fit than the Weibull regression model for some real data sets.

2. The incomplete gamma distribution

In this section, first we introduce incomplete gamma (ING) distribution and derive some of its properties.

Definition 2.1. A random variable X has a standard incomplete gamma distribution, if its pdf is given by

$$f(x) = \frac{1}{e\,\Gamma_0(1)} \log(x+1) e^{-x}, \quad x > 0. \tag{2}$$

By adding a shape (c) and scale (b) parameters to (2), the two parameters ING pdf is presented as

$$f(x|c,b) = \frac{1}{[c + e^{e^c}\,\Gamma_0(e^c)]b} \log\left(\frac{x}{b} + e^c\right) e^{-\frac{x}{b}}, \quad x > 0, c \geq 0, b > 0. \tag{3}$$

The random variable X with pdf (3) is denoted by $X \sim \mathrm{ING}(c, b)$.

Now, we prove that (2) and (3) are proper probability density functions. Using a change of variables, we have

$$\int_0^\infty \log(x+1) e^{-x} dx = e\,\Gamma_0(1),$$

thus the pdf of the standard version is (2). More generally, for $c \geq 0$,

$$\int_0^\infty \log(x + e^c) e^{-x} dx = c + e^{e^c}\,\Gamma_0(e^c),$$

and thus, we can write

$$f(x|c) = \frac{1}{c + e^{e^c}\,\Gamma_0(e^c)} \log\left(x + e^c\right) e^{-x}, \quad x > 0, c \geq 0. \tag{4}$$

From (4) and the scale family of densities $\frac{1}{b} f(\frac{x}{b})$ for $b > 0$, the pdf with shape and scale parameters is obtained (3).

Figure 1 shows the effect of both parameters on the shape of the pdf. This pdf can model unimodal and decreasing data sets. If $1 < e^c < \mathrm{LW}(1)^{-1} = 1.7632$, then we have a positive mode given by

$$x_m = b(\mathrm{LW}(1)^{-1} - e^c),$$

where LW(x) $= \sum_{n=1}^{\infty} \frac{(-n)^{n-1}}{n!} x^n$ is Lambert function; otherwise the mode is $x_m = 0$. An interesting property is that at $x = 0$,

$$f(0|c,b) = \left[b\left(1 + c^{-1} e \Gamma_0(e^c)\right)\right]^{-1}$$

and so this distribution moves continuously across the vertical axis and can take each value in $[0, b^{-1})$ (Figure 1), whereas at $x = 0$, the density of Gamma $(x^{a-1}e^{-x/b}/[b^a\Gamma(a)])$, takes zero (if $a > 1$), b^{-1} (if $a = 1$) or infinity (if $a < 1$).

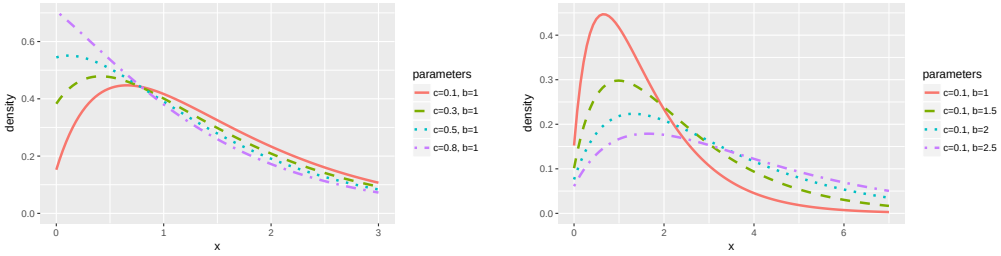

FIGURE 1. Density with different shape values and scale 1 (left), different scale values and shape 0.1 (right).

The cumulative distribution function (cdf) is

$$F(x) = 1 - \frac{1}{c + e^{e^c}\,\Gamma_0(e^c)} \left[e^{e^c}\,\Gamma_0(\frac{x}{b} + e^c) + \log(\frac{x}{b} + e^c)e^{-\frac{x}{b}}\right], \quad x \geq 0. \tag{5}$$

As $c \to \infty$, we have $X \xrightarrow{D} E(b)$, since $\lim_{c\to\infty} f(x|c,b) = \frac{1}{b}e^{-x/b}$. If $A(c) = [c + e^{e^c}\Gamma_0(e^c)]^{-1}$, then the first four moments are:

$$E(X) = bA(c)[c + (1 - e^c)e^{e^c}\Gamma_0(e^c) + 1],$$
$$E(X^2) = b^2A(c)[2c + (2 - 2e^c + e^{2c})e^{e^c}\Gamma_0(e^c) + 3 - e^c],$$
$$E(X^3) = b^3A(c)[6c + (6 - 6e^c + 3e^{2c} - e^{3c})e^{e^c}\Gamma_0(e^c) + 11 - 4e^c + e^{2c}],$$

and

$$E(X^4) = b^4A(c)[23c + (23 - 24e^c + 12e^{2c} - 4e^{3c} + e^{4c})e^{e^c}\Gamma_0(e^c) + 50 - 18e^c + 5e^{2c} - e^{3c}],$$

respectively. The mean and variance are increasing when c decreases and b increases and the skewness and kurtosis are increasing when c increases. After some calculations, the moment generating function is

$$M(t) = \frac{c + \Gamma_0[e^c(1 - bt)]e^{e^c(1-bt)}}{[c + e^{e^c}\Gamma_0(1)](1 - bt)}, \quad t < \frac{1}{b}.$$

3. Characterizations

This section deals with the characterizations of the ING distribution based on the ratio of two truncated moments. Note that our characterizations can be employed also when the cdf does not have a closed form. We would also like to mention that due to the nature of ING distribution, our characterizations may be the only possible ones. Our first characterization employs a theorem due to [7], see Theorem 1 of Appendix A. The result, however, holds also when the interval H is not closed, since the condition of the Theorem is on the interior of H.

Proposition 3.1 Let $X : \Omega \to (0, \infty)$ be a continuous random variable and let $q_1(x) = \left[\log\left(\frac{x}{b} + e^c\right)\right]^{-1}$ and $q_2(x) = q_1(x) e^{-\frac{x}{b}}$ for $x > 0$. The random variable X has pdf (3) if and only if the function η defined in Theorem 1 is of the form

$$\eta(x) = \tfrac{1}{2} e^{-\frac{x}{b}}, \quad x > 0.$$

Proof. Suppose the random variable X has pdf (3), then

$$(1 - F(x)) E[q_1(X) \mid X \geq x] = \frac{1}{c + e^{e^c}\Gamma_0(e^c)} e^{-\frac{x}{b}}, \quad x > 0,$$

and

$$(1 - F(x)) E[q_2(X) \mid X \geq x] = \frac{1}{2[c + e^{e^c}\Gamma_0(e^c)]} e^{-\frac{2x}{b}}, \quad x > 0.$$

Further,

$$\eta(x) q_1(x) - q_2(x) = -\frac{q_1(x)}{2} e^{-\frac{x}{b}} < 0\ , \quad for \quad x > 0.$$

Conversely, if η is of the above form, then

$$s'(x) = \frac{\eta'(x) q_1(x)}{\eta(x) q_1(x) - q_2(x)} = \frac{1}{b}, \quad x > 0,$$

and consequently $s(x) = x/b$ for $x > 0$.

Now, according to Theorem 1, X has density (3).

Corollary 3.1 Let $X : \Omega \to (0, \infty)$ be a continuous random variable and let $q_1(x)$ be as in Proposition 3.1 The random variable X has pdf (3) if and only if there exist functions q_2 and η defined in Theorem 1 satisfying the following differential equation

$$\frac{\eta'(x) q_1(x)}{\eta(x) q_1(x) - q_2(x)} = \frac{1}{b}, \quad x > 0.$$

Corollary 3.2 The general solution of the differential equation in Corollary 3.1 is

$$\eta(x) = e^{\frac{x}{b}} \left[- \int \frac{1}{b} e^{-\frac{x}{b}} (q_1(x))^{-1} q_2(x)\, dx + D \right],$$

where D is a constant. We like to point out that one set of functions satisfying the above differential equation is given in Proposition 3.1 with $D = 0$. Clearly, there are other triplets (q_1, q_2, η) which satisfy conditions of Theorem 1.

4. Estimation

Let $x_1, \ldots, x_n$ be n observed values of a random sample from the ING(c, b) distribution and $\theta = (c, b)^T$. The log-likelihood function is given by

$$\ell(\theta) = -n \log\left(\left[c + e^{e^c}\Gamma_0(e^c)\right] b\right) + \sum_{i=1}^{n} \log\left(\log\left(\frac{x_i}{b} + e^c\right)\right) - \sum_{i=1}^{n} \frac{x_i}{b}$$

and so the maximum likelihood estimations (MLEs) of the parameters, $\hat{\theta}$, are obtained by solving the the following nonlinear equations simultaneously,

$$\frac{d\ell(\theta)}{dc} = n \frac{e^{c+e^c}\Gamma_0(e^c)}{c + e^{e^c}\Gamma_0(e^c)} - \sum_{i=1}^{n} \frac{e^c}{\left(\frac{x_i}{b} + e^c\right) \log\left(\frac{x_i}{b} + e^c\right)} = 0,$$

$$\frac{d\ell(\theta)}{db} = n + \sum_{i=1}^{n} \frac{x_i}{b\left(\frac{x_i}{b} + e^c\right) \log\left(\frac{x_i}{b} + e^c\right)} - \sum_{i=1}^{n} \frac{x_i}{b} = 0.$$

This work can be performed by a numerical method such as the Newton-Raphson type procedure. Under standard regularity conditions when $n \to \infty$, the distribution of $\hat{\theta}$ can be approximated by a bivariate normal distribution, $N(\mathbf{0}, J(\hat{\theta})^{-1})$, where $J(\theta) = \{\frac{\partial^2 \ell}{\partial r\, \partial s}\}$ for $r, s = c, b$ and $J(\hat{\theta})$ is the observed information matrix evaluated at $\hat{\theta}$. In practice, e.g. for interval estimation of the parameters, the observed information matrix can be obtained by Hessian option in `optim` function of R statistical program. Our simulations show that this procedure works well and we applied it to the real data sets in [8] and the second real data set in [3]. The results show that ING is better than some other two parameter generalizations of the exponential distribution (exponential geometric, exponential Poisson and complementary exponential geometric distributions) based on the AIC criterion.

5. **Application in survival analysis**

As an application in survival analysis, we consider a data set analyzed in [15] which is $n = 686$ patients with primary node positive breast cancer. This data set is available in the package `flexsurv` of R software under the name `bc`; see [11]. The variables are time of death or censoring in years $(y_i, i = 1, \ldots, n)$, censoring ($\delta_i = 1$ if y_i is an observed death time, or $\delta_i = 0$ if this is censored), and prognostic group with three levels good, medium and poor. Let the first level of prognostic group (good) be reference level and z_{1i} and z_{2i} be indicators of medium and poor prognostic group respectively (that is, for $j = 1, 2$ let $z_{ji} = 1$ if the ith patient is in group j and 0 otherwise).

Suppose that survival times follow an ING distribution such that the scale parameter, but not the shape parameter, depends on the covariates. Then $S(y|c, b(z)) = S([b(z)]^{-1}y|c, 1)$ and we have an accelerated failure time (AFT) model. In these models, the effect of the covariates is to speed or slow the passage of time. The shape may depend on the covariates through log link, that is

$$\log b_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i}, \quad i = 1, \ldots, n. \tag{6}$$

We fitted this model to the `bc` data and obtained the MLEs by the function `flexsurvreg()` in the package `flexsurv`. For this purpose, we used the package `expint` ([9]) and defined appropriate functions and supplied the ING as a custom distribution. We compared the results with the Weibull AFT model presented in [11].

TABLE 1. Comparison of Weibull and ING models. The parentheses denote standard errors and brackets show p-values.

Model	log Shape	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	AIC
Weibull	0.3218 (0.04841)	2.4356 (0.1114)	−0.6136 (0.1269)	−1.2122 (0.1256)	1631.9
		[< .0001]	[< .0001]	[< .0001]	
ING	−8.3547 (8.5769)	1.7758 (0.0878)	−0.5746 (0.1098)	−1.1709 (0.1051)	1618.7
		[< .0001]	[< .0001]	[< .0001]	

The resutls are in Table 1 which compares these two models and based on the AIC criterion, the ING shows better fit than Weibull. Figure 2 shows the survival times of three prognostic group fitted by these two models and also Kaplan-Meier estimate of the survival function. Based on ING model, the scale parameter of reference group (good) is estimated as $e^{\hat{\beta}_0} = 5.91$, while for medium and poor groups this scale reduces to $5.91\, e^{\hat{\beta}_1} = 3.32$ and

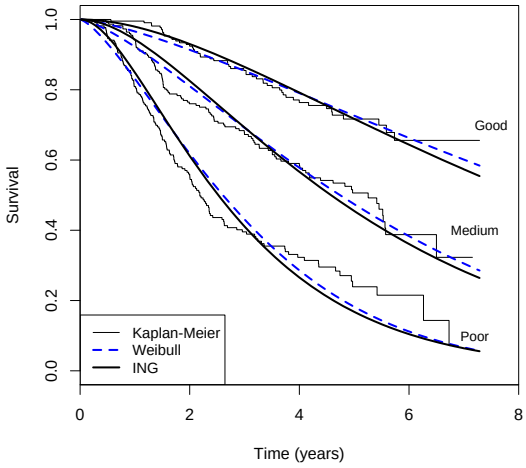

FIGURE 2. Survival by prognostic group from the data: Fitted from Weibull and ING models and Kaplan-Meier estimates.

$5.91\, e^{\hat{\beta}_2} = 1.83$, which means the expected survival time of these two groups reduces 44% and 69% with respect to good group, respectively.

REFERENCES

[1] *Abramowitz, M., Stegun, I. A.*, Handbook of Mathematical Functions, Dover, 1972.

[2] *Alizadeh, M., Altun, E., Cordeiro, G.M. and Rasekhi, M.*, The odd power cauchy family of distributions: properties, regression models and applications, Journal of statistical computation and simulation, **88** (2018), No. 4, 785-807.

[3] *Barreto-Souza, W. and Cribari-Neto, F.* , A Generalization of the Exponential-Poisson Distribution, (2008), (`https://arxiv.org/pdf/0809.1894`).

[4] *Burnham, K.P., Anderson, D.R.*, Model selection and multimodel inference, Springer, 2002.

[5] *Cruz, J.N.d., Ortega, E.M.M. and Cordeiro, G.M.* The log-odd log-logistic Weibull regression model: modelling, estimation, influence diagnostics and residual analysis, Journal of statistical computation and simulation. **86** (2016), No. 8, 1516-1538.

[6] *Glänzel, W.*, A characterization theorem based on truncated moments and its application to some distribution families, Mathematical Statistics and Probability Theory (Bad Tatzmannsdorf 1986) Reidel, Dordrecht, **B**, (1987), 75-84.

[7] *Glänzel, W.*, Some consequences of a characterization theorem based on truncated moments. Statistics: A Journal of Theoretical and Applied Statistics, **21**, (1990), No. 4, 613-618.

[8] *Gómez, Y.M., Bolfarine, H., Gómez, H.W.*, A new extension of the exponential distribution, Revista Colombiana de Estadística, **37**, (2014), N0. 1, 25-34.

[9] *Goulet, V.* (2016). expint: Exponential Integral and Incomplete Gamma Function. R package (`https://cran.r-project.org/package=expint`).

[10] *Gupta, R.D., Kundu, D.* Theory & methods: Generalized exponential distributions, Australian & New Zealand Journal of Statistics, **41**, (1999), No. 2, 173-88.

[11] *Jackson, C.* flexsurv: A Platform for Parametric Survival Modeling in R, Journal of Statistical Software, **70**,(2016), No. 8, 1-33, (`doi = 10.18637/jss.v070.i08`).

[12] *Kus, C.*, A new lifetime distribution. Computational Statistics and Data Analysis, **51**, (2007), 4497-4509.

[13] *Louzada, F., Roman, M. and Cancho, V.G.* The complementary exponential geomet-ric distribution: Model, properties, and a comparison with its counterpart. Computational Statistics and Data Analysis, **55**,(2011), 2516-2524.

[14] *Ortega, E.M.M., Cordeiro, G.M., Campelo, A.K., Kattan, M.W. and Cancho, V.G.* A power series beta Weibull regression model for predicting breast carcinoma, Statistics in medicine, **34**, (2015), No. 8, 1366-1388.

[15] *Royston, P. and Parmar, M.*, Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects, Statistics in Medicine, **21** (2002), No. 1, 2175-2197.

[16] *Yousof, H.M., Altun, E., Rasekhi, M., Alizadeh, M., Hamedani, G. G. and Ali, M.M.*, A new lifetime model with regression models, characterizations and applications, Communications in statistics-simulation and computation, (2017), 1-23.

Appendix A

Theorem 1. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a given probability space and let $H = [a, b]$ be an interval for some $d < b$ ($a = -\infty,\ b = \infty$ might as well be allowed). Let $X : \Omega \to H$ be a continuous random variable with the distribution function F and let q_1 and q_2 be two real functions defined on H such that

$$\mathbf{E}\left[q_2(X) \mid X \geq x\right] = \mathbf{E}\left[q_1(X) \mid X \geq x\right] \eta(x), \quad x \in H,$$

is defined with some real function η. Assume that $q_1, q_2 \in C^1(H)$, $\xi \in C^2(H)$ and F is twice continuously differentiable and strictly monotone function on the set H. Finally, assume that the equation $\eta q_1 = q_2$ has no real solution in the interior of H. Then F is uniquely determined by the functions q_1, q_2 and η , particularly

$$F(x) = \int_a^x C \left| \frac{\eta'(u)}{\eta(u) q_1(u) - q_2(u)} \right| \exp\left(-s(u)\right)\ du\ ,$$

where the function s is a solution of the differential equation $s' = \frac{\eta'\ q_1}{\eta q_1 - q_2}$ and C is the normalization constant, such that $\int_H dF = 1$.

We like to mention that this kind of characterization based on the ratio of truncated moments is stable in the sense of weak convergence (see, [7]), in particular, let us assume that there is a sequence $\{X_n\}$ of random variables with distribution functions $\{F_n\}$ such that the functions q_{1n}, q_{2n} and η_n $(n \in \mathbb{N})$ satisfy the conditions of Theorem 1 and let $q_{1n} \to q_1$, $q_{2n} \to q_2$ for some continuously differentiable real functions q_1 and q_2. Let, finally, X be a random variable with distribution F. Under the condition that $q_{1n}(X)$ and $q_{2n}(X)$ are uniformly integrable and the family $\{F_n\}$ is relatively compact, the sequence X_n converges to X in distribution if and only if η_n converges to η, where

$$\eta(x) = \frac{E\left[q_2(X) \,|X \geq x\right]}{E\left[q_1(X) \,|X \geq x\right]}.$$

This stability theorem makes sure that the convergence of distribution functions is reflected by corresponding convergence of the functions q_1, q_2 and η, respectively. It guarantees, for instance, the 'convergence' of characterization of the Wald distribution to that of the Lévy-Smirnov distribution if $\alpha \to \infty$.

A further consequence of the stability property of Theorem 1 is the application of this theorem to special tasks in statistical practice such as the estimation of the parameters of

discrete distributions. For such purpose, the functions q_1, q_2 and, specially, η should be as simple as possible. Since the function triplet is not uniquely determined it is often possible to choose η as a linear function. Therefore, it is worth analyzing some special cases which helps to find new characterizations reflecting the relationship between individual continuous univariate distributions and appropriate in other areas of statistics.