Decision Science Letters 10 (2021) 415–424

Contents lists available at GrowingScience

Decision Science Letters

homepage: www.GrowingScience.com/dsl

Simulation of generalized Gamma distribution with maximum likelihood estimation and expectation-maximization algorithm on right censored data type 1

Dian Kurniasari[a*], Warsono Warsono[a], Nourma Indryani[a], Mustofa Usman[a] and Sutopo Hadi[b]

[a]*Departement of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Lampung,, Indonesia*
[b]*Departement of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Indonesia*

C H R O N I C L E

Article history:
Received October 28, 2020
Received in revised format: December 29, 2020
Accepted January 4 2021
Available online January 4, 2021

Keywords:
Generalized Gamma Distribution
Right Censored Data type 1
Expectation Maximization (EM)
Maximum Likelihood Estimaton (MLE)
MSE

A B S T R A C T

The Generalized Gamma distribution is very suitable for modeling data with various forms of hazard (risk) functions, which makes the Generalized Gamma distribution useful in survival analysis. Survival analysis aims are to predict chances of survival, disease recurrence, death, and other events over a period of time. One characteristic of survival data is the possibility of sensors. Let X be the life span of the person being studied and the right censorship time of Cr, X is assumed to be independent with the probability density function $f(x)$, the survival function $S(x)$, and the hazard function $h(x)$. A person's X life span will be known if X is less than or equal to Cr. If X is greater than Cr, the individual X survives or is censored right now. Statistical inference, especially parameter estimation is needed in analyzing empirical data. Obviously the estimation results obtained are expected to be a good estimator, namely to meet the nature of unbiased and minimum variance. This paper will discuss the results of the estimation of Generalized Gamma distribution parameters with type 1 right censored data through simulations using the Expectation Maximization method and the Maximum Likelihood Estimation method. The simulation is conducted by generating data with the sample size: 25, 50, 100, 200, 500, 1000, 1500 and 2000 as well as determining censored data of 10%, 20% and 30% by first setting the parameters used which are obtained from the data of patients with gastric cancer namely $\alpha = 1.0649$, $\beta = 1{,}072$, $\theta = 59.766$. Based on the results obtained from the simulations on the two estimation methods that the parameter estimation using the Maximum Likelihood Estimation method is better than the Expectation Maximization method because it provides a smaller bias and MSE value where the larger the sample size used, the estimated parameter value will get closer to the parameter in fact. In addition, the Expectation Maximization method can also be used as an alternative estimation of generalized gamma distribution parameters with type 1 right censored data because it has a bias value and MSE approaching the MLE method.

1. Introduction

Survival time is a data to measure the time to certain events or events, such as failure, death, relapse, or the development of certain diseases. One of the characteristics of survival data is the possibility of sensors. Suppose X is the life span of a person being studied and the right censorship time of C_r, X is assumed to be independent identical distribution, $f(x)$, the survival function $S(x)$, and the hazard function $h(x)$. A person's X life span will be known if X is less than or equal to C_r. If X is greater than C_r, the individual X survives or is censored right (Klein & Moeschberger, 1997). The most popular generalized gamma distribution modeling is for analyzing skewed data. Generalized gamma was introduced by Stacy (1962), and was further discussed by Hoq *et al*. (1974), Lee and Gross (1991), Agarwal and Kalla (1996), Agarwal and Al-Saleh (2001), Kalla *et al*. (2001). The generalized distribution of gamma has been used in several research fields such as engineering, hydrology and survival analysis. Nadarajah and Gupta (2007) use this distribution with applications for drought data. Ortega

* Corresponding author.
E-mail address: dian.kurniasari@fmipa.unila.ac.id (D. Kurniasari)

doi: 10.5267/j.dsl.2021.1.003

et al. (2009) discuss the effects of diagnostics in generalized gamma regression models. Cox (2008) discusses and compares F-generalized families with generalized gamma models and Cox *et al*. (2007) present parametric survival analysis and taxonomies of generalized gamma hazard level functions.

Generalized gamma distribution is very suitable for modeling data with various forms of hazard (risk) function. This characteristic is useful for estimating a person's hazard function (risk) and relative risk and relative time, this is very much needed in survival analysis. To estimate the parameters of the generalized gamma distribution in survival analysis is using Maximum Likelihood Estimation (MLE) which is one of the methods of estimation that most widely used. According to Millar (2011) maximum likelihood as a tool for inference, including evaluation of statistical significance, calculation of confidence intervals, assessment of models, and estimation parameters. Maximum likelihood estimation has optimal properties for a large sample size. So that maximum likelihood is the most widely used methods of parametric inference. Some cases often found in the log likelihood function cannot be maximized analytically, so it is necessary to calculate the MLE by iteration such as by the Newton-Raphson Maximization Procedure, while the other alternative is the expectation-maximization algorithm (Mclachlan a& Krishnan, 2008; Wang & Cheng, 2009; Ruhi *et al*., 2015).

Expectation-maximization algorithm is an approach to calculate maximum likelihood estimation, it is useful in several of incomplete data problems. The procedure in the expectation-maximization algorithm, there are two steps namely, the expectation step (e-step) and the maximization step (m-step) with the basic idea of the expectation-maximization algorithm associating incomplete data problems with complete data problems to get the final result of maximization likelihood estimation. The e-step focuses on producing complete data, using a collection of observable data, so that the simple steps of m-step maximize the complete data (Mclachlan & Krishnan, 2008). This study will discuss the estimation of parameter of generalized gamma distribution by using the expectation-maximization algorithm on type 1 right censored data.

2. Statistical method

2.1 Survival analysis

Survival time is data to measure the time to certain events / events, such as failure, death, relapse, development of certain diseases, parole, or divorce. The survival time distribution is usually described by three functions, the S(x) survival function, the *f(x)* probability density function, and h(x) the hazard function (hazard). These three functions are used to describe various aspects of the data. The fundamental problem in survival data analysis is to estimate parameter from the sample one or more of these three functions and to draw conclusions about survival patterns in the population (Lee & Wang, 2003).

2.2 Survival function

According to Klein and Moeschberger (1997) the survival function is the basis used to describe the phenomenon of time-to-event, the chance of someone surviving beyond time x (experiencing events after time x). Survival function is defined as follows:

$$S(x) = \Pr(X > x) \tag{1}$$

In the context of the failure of the equipment or goods produced, S(x) is referred to as a reliability function. When X is a continuous random variable, the survival function is the complement of the cumulative distribution function, that is,

$$S(x) = 1 - F(x), \tag{2}$$

where $F(x) = \Pr(X < x)$, also the survival function is integral of the probability density function, $f(x)$ namely:

$$S(x) = \Pr(X > x) = \int_x^{\infty} f(t)\, dt \tag{3}$$

So that

$$f(x) = -\frac{dS(x)}{dx}. \tag{4}$$

2.3. Hazard Function (Risk)

According to Klein and Moeschberger (1997) hazard functions are known as conditional failure rates in reliability, mortality strength in demographics, function of intensity in stochastic processes, age-specific failure rates in epidemiology. The level of risk is determined by

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \le X < x + \Delta x \mid X \ge x]}{\Delta x}. \tag{5}$$

If X is continues random variable, so that

$$h(x) = \frac{f(x)}{S(x)} = \frac{d[ln\, S(x)]}{dx} \tag{6}$$

Cumulative hazard function (risk) H(x) is defined as follows:

$$H(x) = \int_0^x h(u)\, du = -\ln [S(x)] \tag{7}$$

So, for a sustainable life

$$S(x) = e^{-H(x)} = e^{-\int_0^x h(u)du}. \tag{8}$$

2.4. Generalized Gamma Distribution

According to Cordeiro *et al*. (2012) that the generalized gamma distribution is good for modeling data with various types of hazard rate functions (risk levels), adding, reducing, which makes it very useful for estimating each hazard function (risk). The generalized gamma distribution has been used in several fields of research such as engineering, hydrology, and survival analysis. The probability density function of the generalized gamma distribution is as follows:

$$f(x) = \frac{\beta\, x^{\beta\alpha-1}}{\Gamma(\alpha)\theta^{\beta\alpha}} e^{-(\frac{x}{\theta})^{\beta}} \tag{9}$$

Parameter α and β is known as shape and parameter θ is called as scale parameter.

2.5. Maximum Likelihood Estimation

According to Hogg et al. (2013), suppose random variables X1, X2, ..., Xn are mutually independent with the probability density function f(x; ψ), ψ ϵ Ω. The parameter ψ is unknown. The procedure for estimating the parameters are based on the likelihood function is as follows:

$$L(\psi;x) = \prod_{i=1}^{n} f(x_i|\psi),\ \theta \in \Omega \tag{10}$$

where $x = (x_1, x_2, \dots, x_n)$'. The log of the likelihood function is generally easier to use, i.e.

$$L(\psi) = \log L(\psi) = \prod_{i=1}^{n} \log f(x_i|\psi),\ \theta \in \Omega \tag{11}$$

The symbol $\widehat{\psi} = \widehat{\psi}(X)$ is a maximum likelihood estimator of θ if

$$\widehat{\psi} = Argmax\ L(\psi;X) \tag{12}$$

Argmax is a method such that L(ψ; X) attains its maximum values for $\widehat{\psi}$. To determine the maximum likelihood estimation, it often uses the log of the likelihood function and then determines the critical value (Hogg and Craig, 1995), because the logarithmic function is monotone increase at (0, ∞) which means it has the same critical value. For example L(ψ) =Log L(ψ), then solving the maximum likelihood estimation with the following equation,

$$\frac{\partial\, L(\psi)}{\partial \psi} = 0. \tag{13}$$

2.6. Expectation-MaximizationAlgorithm

According to Mclachlan and Krishnan (2008), the expectation-maximization (EM) algorithm is a method that is widely used to calculate the maximum likelihood iteration estimation that is useful in various incomplete data, which is if using an algorithm such as the Newton-Raphson method, it might be more complicated in overcoming the incompleteness data. At each EM iteration algorithm, there are two steps namely: the expectation step or e-step and the maximization step or m-step. The basic idea of the Expectation-Maximization Algorithm is to associate incomplete data problems with complete data problems to get the final result of maximum likelihood estimation. The e-step focuses on generating complete data, using a collection of observable data, so that the simple steps of m-step maximize the complete data. Let Y be a random vector that matches the data observed y and has p.d.f. g (y; ψ), where ψ = (ψ1, ..., ψd) T is an unknown parameter vector with parameter space Ω. The EM algorithm is a widely applied algorithm that provides iterative procedures for calculating maximization likelihood estimation. In this context, the observed vector data as incomplete. The idea of 'incomplete data' includes the notion of being lost in conventional data. Let x denote vectors containing complete data called augmented data,

and let z denote vectors containing additional data, referred to as data that cannot be observed or lost. Let $g_c(x; \psi)$ be p.d.f. from the random vector X that corresponds to the complete vector x data. Then the complete log likelihood data function that can be formed for ψ if x is fully observable is as follows:

$$\log L_c(\boldsymbol{\psi}) = \log g_c(x; \boldsymbol{\psi}). \tag{14}$$

Let's say there are two sample spaces x and y and a many one to one mapping from x to y. Vector data complete data x in x, while vector data incomplete vector y = y (x) in y. So that

$$g(x; \boldsymbol{\psi}) = \int g_c(x; \boldsymbol{\psi}) dx \tag{15}$$

where x (y) is the part of x that determines by the equation y = y (x). The EM algorithm is an approach to solving the likelihood equation for incomplete data indirectly by performing an iterative calculation for the complete log likelihood data function, $\log L_c(\boldsymbol{\psi})$. Because its cannot be observed, the incomplete data is replaced by the conditional expectation value given by y, using a match for $\boldsymbol{\psi}$. Let $\boldsymbol{\psi}^{(0)}$ be the initial value for $\boldsymbol{\psi}$. Then in the first iteration, e-step requires the following calculation

$$Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(0)}) = E_{\boldsymbol{\psi}^{(0)}}\{\log L_c(\boldsymbol{\psi}) \mid \mathbf{y}\}, \tag{16}$$

m-step, then maximization $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(0)})$ related to $\boldsymbol{\psi}^{(0)}$ on the parameter space Ω. Namely, we select $\boldsymbol{\psi}^{(1)}$ such that

$$Q(\boldsymbol{\psi}^{(1)}; \boldsymbol{\psi}^{(0)}) \geq Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(0)}), \tag{17}$$

for all $\boldsymbol{\psi} \in \Omega$. E-step and m-step process is rerepeat, but with $\boldsymbol{\psi}^{(0)}$ replaced by the fit of $\boldsymbol{\psi}^{(1)}$. At the-(k + 1) iterations, e-step and m-step is defined as follows, e-step calculate
$Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k)})$, where

$$Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k)}) = E_{\boldsymbol{\psi}^{(k)}}\{\log L_c(\boldsymbol{\psi}) \mid \mathbf{y}\} \tag{18}$$

m-step select $\boldsymbol{\psi}^{(k+1)}$ to find the value $\boldsymbol{\psi} \in \Omega$ that maximize $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k)})$ namely,

$$Q(\boldsymbol{\psi}^{(k+1)}; \boldsymbol{\psi}^{(k)}) \geq Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k)}) \tag{19}$$

for all $\boldsymbol{\psi} \in \Omega$.

2.7. Censored Data

Censorship is the loss of observation / information on the life span variables observed in the study. In survival data, censorship often occurs for various reasons. In clinical trials about the effectiveness of medical treatment for a diseases, for example, patients may leave the hospital due to migration or health problems. Sensors are generally divided into certain types. If someone has entered the study but cannot be followed up, the actual time of the event is placed somewhere to the right of the censored time along the time axis. This type of sensor is called the right sensor. Because right censors occur far more frequently than other types and information can be included in the estimation of survival models (Liu, 2012). According to Klein and Moeschberger (1997), there are three types of censorship as follows:

2.7.1. Right Sensor

Let X be the life span of a person and the time of censoring is C_r. A person's lifetime X will be known if $X \leq C_r$. If X is greater than C_r then the person is said to be surviving or the right sensor.

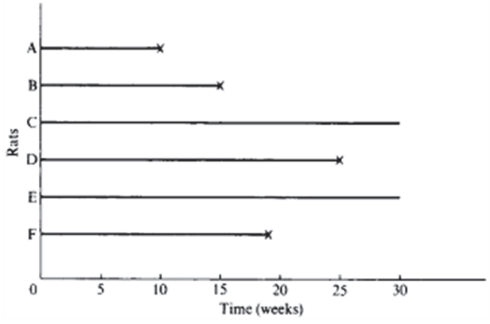


Fig. 1. The illustration of right sensor or type 1

For example, from Lee and Wang (2003), suppose that there are six mice that are exposed to carcinogens by injecting tumor cells into their foot pads. Time to develop tumors of a certain size is observed. Researchers decided to stop the experiment after 30 weeks. Figure 1 is a time graph of tumor development. Rats A, B, and D experienced tumors after 10, 15, and 25

weeks, respectively. C and E mice did not have a tumor at the end of the study, so the likelihood of getting a tumor is 30 weeks. F rat died without a tumor after 19 weeks of observation. So the survival time data are 10, 15, 30+, 25, 30+, and 19+ weeks. The plus sign indicates censored observations.

2.7.2. Left Sensor

Let X be the life span of a person in study. If there is a left C_l-sensor, it means that someone has experienced an event before it was observed in the study.

2.7.3. Interval sensor

A person's lifetime occurs at a time interval.

3. Characteristics of an estimator

3.1. Unbiased

One property that must be fulfiled by a parameter estimator of a distribution is the characteristic of the unbiased of the estimator. The estimator $U(\boldsymbol{X}) = U(X_1, X_2, \dots, X_n)$,which satisfied the following equation:

$E\big(U(\boldsymbol{X})\big) = g(\theta)$,
is called unbiased estimator of $g(\theta)$ (Warsono *et al*., 2019).

3.2. Minimum Variance

An estimator is said to be a good estimator if in addition to having an unbiased property, it also has a minimum variance property. Bain and Engelhardt (1992) states that for example, U(X) is an unbiased estimator for g(θ) with minimum variance, if for any other unbiased estimator U_1(X) of g(θ), $Var\ \big(U(\boldsymbol{X})\big) \leq Var\ (U_1(\boldsymbol{X}))$ for all $\boldsymbol{\theta} \in \Omega$, where

$$Var\big(U_1(\boldsymbol{X})\big) \geq \frac{\left(\frac{\partial}{\partial\theta} g(\boldsymbol{\theta})\right)^2}{n\,.\,E\left[\frac{\partial}{\partial\theta} ln\, f(\boldsymbol{X};\boldsymbol{\theta})\right]^2} \tag{20}$$

is an inequality for minimum variance (Hogg and Craig, 1995). According to Hogg and Craig (1995), $E\left[\frac{\partial}{\partial\theta} ln\, f(\boldsymbol{X};\boldsymbol{\theta})\right]^2$ is called Fisher Information and written as $I(\theta)$. Such that the Equation (20) above can be simplified as: $Var\big(U_1(\boldsymbol{X})\big) \geq \frac{[g'(\theta)]^2}{nI(\theta)}$,which is called as inequality of *Cramer-Rao Lower Bound*

4. Data analysis

The data used in this study are data from Monte Carlo simulation by using generalized gamma distribution. So that the choice of parameter values approaches the value of the reliability data, the selection of parameter values in the simulation is based on the reliability data available in libraries. This simulation is based on data from patients with gastric cancer who have been tested first using EASYFIT software that the data has generalized gamma distribution with parameters α = 1.0649, β = 1.072, θ = 59.766. Data generation is performed for several different size of data, namely for of n: 25, 50, 100, 200, 500, 1000, 1500 and 2000 with 100 replications and using R software and censoring data with percentages p = 10%, p = 20% and p = 30%.

4.1. Complete Data Likelihood Function from Generalized Gamma Distribution

Warsono (2009) states that the random variable X is said to have a generalized gamma distribution with parameters α, β, and θ if and only if the probability density function of X is,

$$f(x;\alpha,\beta,\theta) = \frac{\beta}{\Gamma(\alpha)}\theta^{-\beta\alpha}x^{\beta\alpha-1}\exp\left[-\left(\frac{x}{\theta}\right)^{\beta}\right]. \tag{21}$$

Estimation of the parameters of generalized gamma distribution by using the maximum likelihood estimation method begins by determining the likelihood function as follows:

$$L(\alpha,\beta,\theta) = \prod_{i=1}^{n} f(x_i;\alpha,\beta,\theta) = \prod_{i=1}^{n} \frac{\beta}{\Gamma(\alpha)}\theta^{-\beta\alpha}x^{\beta\alpha-1}\exp\left[-\left(\frac{x}{\theta}\right)^{\beta}\right] \tag{22}$$

$$L(\alpha,\beta,\theta) = \frac{\beta}{\Gamma(\alpha)}\theta^{-\beta\alpha} \dots \theta^{-\beta\alpha}\frac{\beta}{\Gamma(\alpha)}{x_1}^{\beta\alpha-1} \dots {x_n}^{\beta\alpha-1}e^{-\left(\frac{x_1}{\theta}\right)^{\beta}} \dots e^{-\left(\frac{x_n}{\theta}\right)^{\beta}} = \left(\frac{\beta}{\Gamma(\alpha)}\theta^{-\beta\alpha}\right)^n \left(\prod_{i=1}^{n} {x_i}^{\beta\alpha-1}\right) e^{-\frac{1}{\theta^{\beta}}\sum_{i=1}^{n} {x_i}^{\beta}} \tag{23}$$

To facilitate the derivative of the likelihood function of the Generalized Gamma distribution in the survival time data of patients suffering from gastric cancer, then equation (23) is changed to logarithmic form as follows:

$$\log L(\alpha,\beta,\theta) = n\log\beta - \beta\alpha n\log\theta - n\log\Gamma(\alpha) + (\beta\alpha-1)\log\sum_{i=1}^{n} x_i - \frac{1}{\theta^{\beta}}\sum_{i=1}^{n} {x_i}^{\beta} \quad . \tag{24}$$

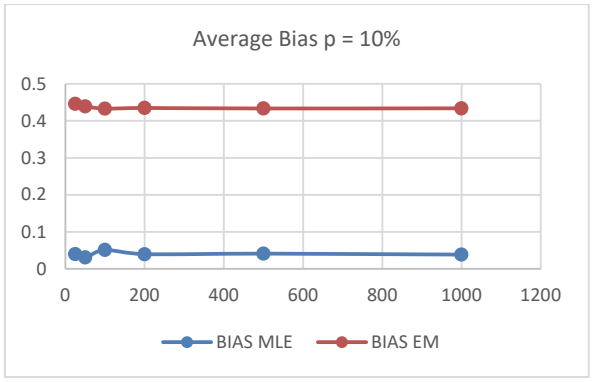
Table 1
The estimate of the Parameter α , β and θ by method MLE and EM

n	p	Estimate vale of Parameter α		Estimate value of Parameter β		Estimate value of Parameter θ	
		MLE	EM	MLE	EM	MLE	EM
25	10%	1.157	2.4024	1.098	1.072	59.767	59.766
	20%	1.257	2.7298	1.042	1.072	59.770	59.766
	30%	1.344	3.2955	0.985	1.072	59.772	59.766
50	10%	1.146	2.3825	1.062	1.072	59.768	59.766
	20%	1.245	2.6833	0.990	1.072	59.770	59.766
	30%	1.363	3.3266	0.949	1.072	59.773	59.766
100	10%	1.180	2.3643	1.036	1.072	59.768	59.766
	20%	1.238	2.6361	1.003	1.072	59.770	59.766
	30%	1.343	3.1050	0.956	1.072	59.771	59.766
200	10%	1.148	2.3695	1.037	1.072	59.767	59.766
	20%	1.236	2.6303	0.992	1.072	59.769	59.766
	30%	1.349	3.0690	0.938	1.072	59.771	59.766
500	10%	1.142	2.3656	1.028	1.072	59.767	59.766
	20%	1.227	2.6280	0.977	1.072	59.769	59.766
	30%	1.339	3.0477	0.920	1.072	59.772	59.766
1000	10%	1.135	2.3663	1.028	1.072	59.767	59.766
	20%	1.220	2.6269	0.978	1.072	59.769	59.766
	30%	1.339	3.0682	0.920	1.072	59.772	59.766
1500	10%	1.133	2.3621	1.026	1.072	59.767	59.766
	20%	1.219	2.6227	0.975	1.072	59.769	59.766
	30%	1.333	3.0582	0.919	1.072	59.771	59.766
2000	10%	1.135	2.3622	1.026	1.072	59.767	59.766
	20%	1.221	2.6223	0.976	1.072	59.769	59.766
	30%	1.336	3.0522	0.919	1.072	59.771	59.766

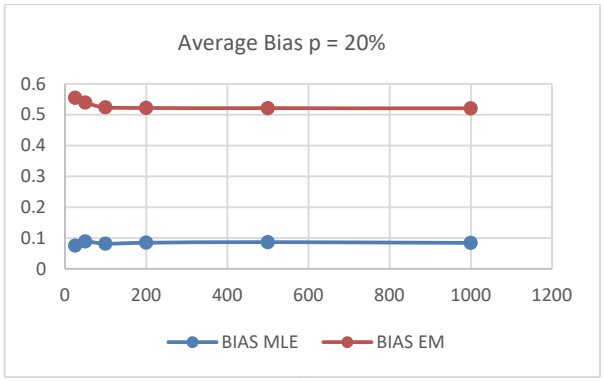
Based on the estimation results presented in Table 1, it can be seen that the two methods give different estimation results. Where the estimation of Parameter β and Parameter θ by using the EM method the estimate values are the same with the real parameter value. Next to see the accuracy of the estimation of the two methods, we compared the average values of Bias and MSE of each estimator.

Table 2
Estimate values of Bias and MSE by method MLE and EM

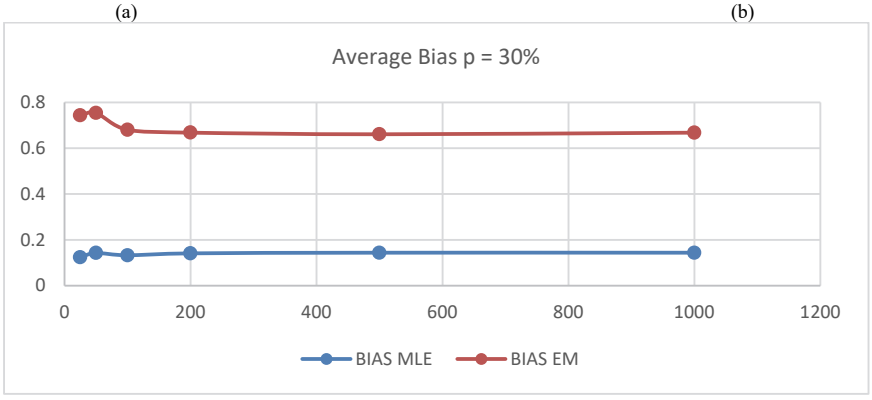
n	p	Average Bias		Average MSE	
		MLE	EM	MLE	EM
25	10%	0.0399	1.3375	0.0031	0.5963
	20%	0.0754	1.6649	0.0126	0.9240
	30%	0.1242	2.2306	0.0286	1.6585
50	10%	0.0310	1.3176	0.0022	0.5787
	20%	0.0887	1.6184	0.0131	0.8731
	30%	0.1428	2.2617	0.0348	1.7051
100	10%	0.0512	1.2994	0.0049	0.5628
	20%	0.0817	1.5712	0.0115	0.8229
	30%	0.1327	2.0401	0.0302	1.3873
200	10%	0.0396	1.3046	0.0027	0.5674
	20%	0.0849	1.5654	0.0119	0.8168
	30%	0.1413	2.0041	0.0330	1.3388
500	10%	0.0410	1.3007	0.0026	0.5639
	20%	0.0867	1.5631	0.0117	0.8145
	30%	0.1439	1.9828	0.0327	1.3104
1000	10%	0.0386	1.3014	0.0023	0.5645
	20%	0.0840	1.5620	0.0110	0.8132
	30%	0.1439	2.0033	0.0327	1.3377
1500	10%	0.0385	1.2972	0.0023	0.5609
	20%	0.0848	1.5507	0.0111	0.8016
	30%	0.1421	1.9933	0.0317	1.3244
2000	10%	0.0391	1.2973	0.0023	0.5610
	20%	0.0852	1.5574	0.0112	0.8085
	30%	0.1431	1.9873	0.0323	1.3165



(a)



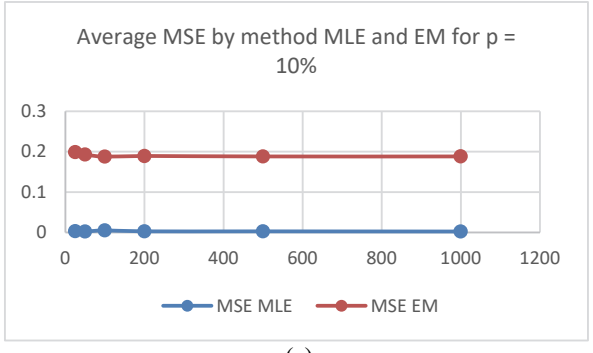
(b)



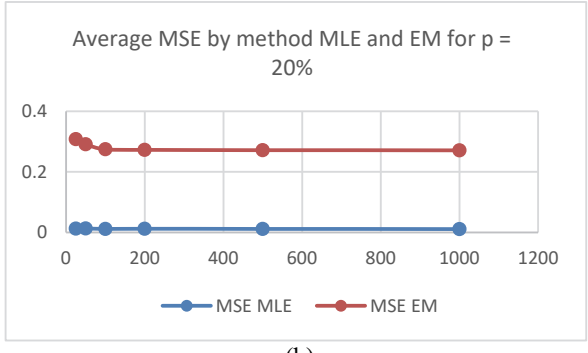
(c)

Fig. 2. Average Bias by method MLE and EM for the values of (a) p= 10%, (b) p=20%; and (c) p=30%

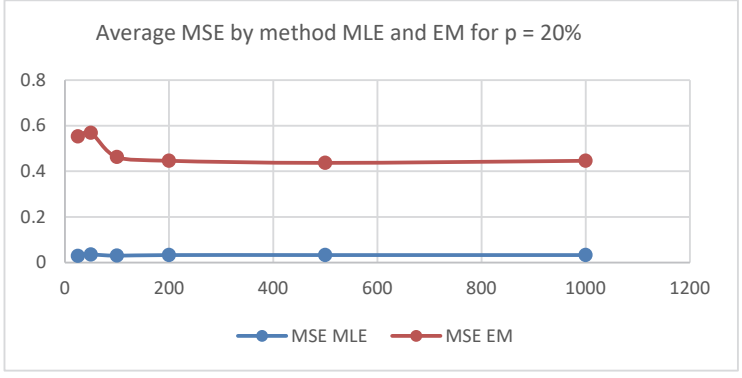
To make it easier to see the difference in average bias values given by the two methods, Table 2 is visualized in graphical form in Figure 2 and Figure 3. Based on the graph of the average bias values in Fig. 2. it can be seen that the bias value generated from the EM method is greater than the bias value produced by MLE. It can be seen that the greater the value of n for each sensor percentage p = 10%, p = 20% and p = 30% results in the smaller bias value produced by the EM and MLE methods. But the greater the percentage sensor value in the data, the greater the bias.



(a)



(b)



(c)

Fig. 3. Average MSE by method MLE and EM for the values of (a) p= 10%, (b) p=20% and(c) p=30%

Based on Fig. 3. the greater the n used, the smaller the average MSE of the EM and MLE methods. The average MSE value of the EM method is greater than the MLE method on the graph for each sensor percentage value p = 10%, p = 20% and p = 30%. So that, from the average bias and MSE values it is known that the MLE method gives a smaller average bias and MSE value than that of the EM method. So that, it can be said that the MLE method provides better estimator values regardless of the amount of data (n) and the percentage of censored data (p).

4.2. Establishing the Survival Function and Hazard Function from the First generating Data

After estimating parameter by using the Maximum Likelihood Estimation and Expectation-Maximization method assisted by the Newton Raphson method, the result given in Table 1, the survival and hazard functions of the data generated by the survival time parameters of patients with gastric cancer with parameter α = 1.0649 , β = 1,072, and θ = 59,766.The survival function of the Generalized Gamma distribution is as follows:

$$S(x) = \frac{\Gamma\left(\alpha, \left(\frac{x}{\theta}\right)^{\beta}\right)}{\Gamma(\alpha)}. \tag{25}$$

Hazard function of the Generalized Gamma distribution is as follows:

$$h(x) = \frac{\frac{\beta x^{\beta\alpha-1}}{\theta^{\beta\alpha}} e^{-(\frac{x}{\theta})^{\beta}}}{\Gamma\left(\alpha, \left(\frac{x}{\theta}\right)^{\beta}\right)} \tag{26}$$

From the survival function and hazard function the survival time of patients suffering from gastric cancer above can be found the value of survival probability and the probability of failure. Several setting of time will be tried when x = 50, x = 100, x = 150, x = 200, x = 250, and x = 1000. One of the estimation results of the survival and hazard functions by using the methods of MLE and EM are as follows:

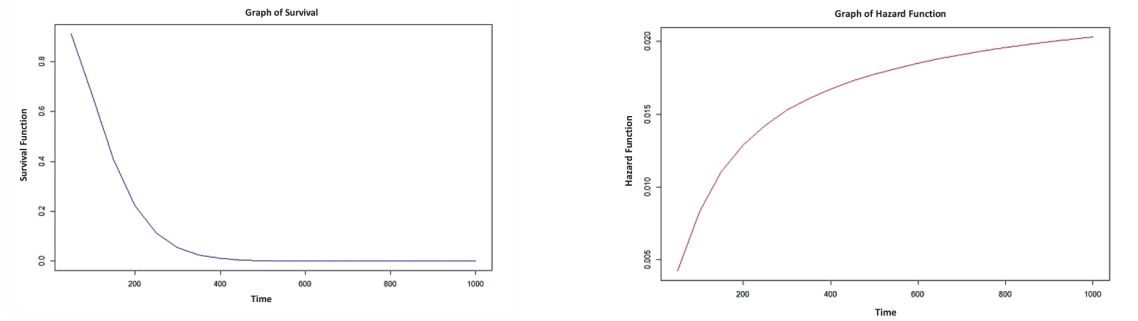


Fig. 4. Graph of Survival function and Hazard function by MLE method

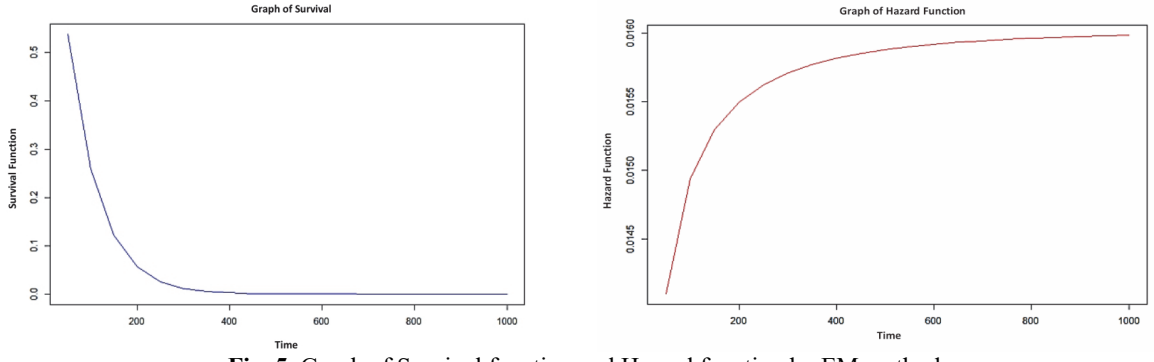


Fig. 5. Graph of Survival function and Hazard function by EM method

Fig. 4 and Fig. 5 show that the survival function curve for the MLE and EM methods, where the x-axis shows the value of the survival time of patients with gastric cancer and on the y-axis shows the estimation results of the survival function. On the survival function curve it is seen that for every increment of time, the value of the survival function continues to decrease approaches zero. So that, it can be concluded that each additional time, the chances of survival of patients suffering from gastric cancer will continue to decrease until the occurrence of death. In the hazard function curve by using the MLE and EM methods, where the x-axis is the value of the survival time of patients with gastric cancer and on the y-axis is the estimated results of the hazard function. The lines on the hazard function curve appear to go up along with the increasing survival time of patients with gastric cancer. So it can be concluded that the hazard function of the survival time data of

patients with gastric cancer has an increasing hazard rate that is the increasing survival time of an individual, the failure rate will increase.

5. Conclusion

Based on the results obtained by using data from patients with gastric cancer (in the Asaur package in Software R) has a generalized gamma of right-sensed type 1 distribution with parameters $\alpha = 1.0649$, $\beta = 1{,}072$, and $\theta = 59{,}766$. From the simulations that have been carried out it can be seen that the larger the sample size of the data used, the smaller the estimated value of the bias and MSE for both methods of EM and MLE. Estimation of Generalized Gamma distribution parameters with right censored data type 1 using Maximum Likelihood Estimation and Expectation Maximization produces estimators that cannot be solved analytically, so it needs to be solved numerically. From the two estimation methods used it can be concluded that the best method for estimating the Generalized Gamma distribution parameters with type 1 right censored data is the Maximum Likelihood Estimation method because it has the smallest average bias value and mean square error compared to EM method. Although the average bias and MSE values of the EM method are greater than the MLE method, the EM method still provides the results of the estimation of bias and MSE are close to the results of the MLE method. Thus, it can be said that the EM method can become an alternative estimation method for estimating parameters in the Generalized Gamma distribution with type 1 right censored data.

References

Agarwal, S.K. & Kalla. S.L. (1996). A generalized gamma distribution and its application in reliability. *Communication in Statistics Theory-Methods*, *25*, 201- 210.

Agarwal, S.K. & Al-Saleh, A.A. (2001). Generalized gamma type distribution and its hazard rate function. *Communication in Statistics Theory-Methods*, *30*, 309-318.

Bain, L.J. & Engelhardt, M. 1992. *Introduction to Probability and Mathematical Statistics.* Duxbury, United States.

Cordeiro, G.M., Castellares, F., Montenegro, L.C., & Casto, M. (2012). The Beta Generalizatiom of The Gamma Distribution. *Statistics, 47***,** 1-13.

Cox, C., Chu, H., Schneider, M.F., & Muñoz, A. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine*, *26*, 4352–4374.

Cox, C. (2008). The generalized F distribution: an umbrella for parametric survival analysis. *Statistics in Medicine*, *27***,** 4301–4312.

Hogg, R.V., & Craig, A.T. (1995). *Introduction of Mathematical Statistic. Fifth Edition*. John Wiley & Sons, The United States of America.

Hogg, V.H., Mckean, J.W., & Craig, A.T. (2013). *Introduction to Mathematical Statistics*. Pearson Education Inc., Boston.

Hoq, A., Ali, M. & Templetion. (1974). Estimation of parameters of a generalized life testing models. *Journal of Statistical Research*, *9*, 67-79.

Kalla, S.L. Al-Saqabi, B.N., & Khajah, H.G. 2001. A unified form of gamma type distributions. *Applied Mathematics and Computation*, *118*, 175-187.

Klein, J.P., & Moeschberger, M.L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data.* Springer-Verlag, New York.

Lee, E., & Wang, J.W. (2003). *Statistical Method for Survival Data Analysis.* John Wiley & Sons Inc., New York.

Lee, M., & Gross, A. (1991). Lifetime distributions under unknown environment. *Journal of Statistical Planning and Inference*, **29**, 137-143.

Liu, X. (2012). *Survival Analysis : Models and Applications*. John Wiley & Sons, USA.

Mclachlan, G.J., & Krishnan, T. (2008). *The EM Algorithm and Extensions.* John Wiley & Sons Inc., Canada.

Millar, R.B. (2011). *Maximum Likelihood Estimation and Inference: With Examples in R, SAS, and ADMB*. John Wiley & Sons, Ltd., United Kingdom.

Nadarajah, S., & Gupta, A.K. (2007). Moments and cumulants of the skew normal distribution, *Kobe Journal of Mathematics, 24*, 107-124

Ortega, E.M.M., Cancho, V.G., & Paula, G.A. (2009). Generalized log-gamma regression models with cure fraction. *Lifetime Data Analysis*, *15***,** 79–106.

Ruhi, S., Sarker, S., & Karim, M.R. (2015). Mixture Model for Analyzing Product Reliability Data: a Case Study. *SpringerPlus, 4***,** 634.

Stacy, E.W. (1962). A Generalization of The Gamma Distribution. *The Annals of Mathematical Statistics*, *33*, 1187-1192.

Wang, F.K. & Cheng, Y.F. (2009). EM Algorithm for Estimating The Burr XII Parameters with Multiple Censored Data. *John Wiley & Sons Inc.*, pp. 615-630.

Warsono. (2009). Moment Properties of the Generalized Gamma Distribution, *National Seminar on Sains, Matematika dan Aplikasinya IV,* pp. 157-162.

Warsono, Gustavia, E., Kurniasari, D., Amanto, & Antonio, Y. (2019). On The Comparison of The Methods of Parameter Estimation for Pareto Distribution. *Journal of Physics: Conf. Series*, *1338*, 012042.