Ibn Al Haitham Journal for Pure and Applied Science

Journal homepage: http://jih.uobaghdad.edu.iq/index.php/j/index

Parametric Models in Survival Analysis for Lung Cancer Patients

Layla A. Ahmed
layla.aziz@garmian.edu.krd
Department of Mathematics, College of Education, University of Garmian, Kurdistan Region, Iraq

Article history: Received, 7,January,2020, Accepted 20 ,Februry,2020, Published in April 2021

Doi: 10.30526/34.2.2617

Abstract

The aim of this study is to estimate the survival function for the data of lung cancer patients, using parametric methods (Weibull, Gumbel, exponential and log-logistic).

Comparisons between the proposed estimation method have been performed using statistical indicator Akaike information Criterion, Akaike information criterion corrected and Bayesian information Criterion, concluding that the survival function for the lung cancer by using Gumbel distribution model is the best. The expected values of the survival function of all estimation methods that are proposed in this study have been decreasing gradually with increasing failure times for lung cancer patients, which means that there is an opposite relationship failure times and survival function.

Keywords: Survival analysis, Weibull distribution, Gumbel distribution, Exponential distribution, Log-logistic distribution.

1.Introduction

Survival analysis is a branch of statistics which deals with the analysis of time to events, such as death in biological organisms and failure in mechanical systems. The topic of survival analysis is called reliability theory or reliability analysis in engineering, and duration analysis or duration modeling in economics or event history analysis in sociology [1]. In many applied sciences such as medicine, engineering and finance, amongst others, modeling and analyzing lifetime data are crucial.

Several lifetime distributions have been used to model such type of data. The quality of the executions and in a statistical analysis depends heavily on the presupposed probability distributions [2].

Parametric methods which involve the exponential, Weibull, lognormal, gamma and extreme value distribution have been widely used for fitting survival data [3].

Gumbel showed that the Weibull distribution and the type III smallest extreme value distribution are the same [4].

Log- logistic distribution is a very important reliability model as it fits well in many applied situations of reliability data analysis. Another advantage with the log- logistic distribution lies in its closed form expression for survival and failure rate functions that makes it important over log- normal distribution [5].

Salman and Farhan [6] estimated the survival function for the patients of lung cancer; they used several nonparametric estimation methods, and concluded that the shrinkage was the best method.

The main objective of this research is to estimate the survival function for the data of lung cancer patients, by using parametric methods and determine the best and most efficient distribution.

2. Theoretical Part

2.1 Survival Analysis

Survival function is the probability that a system or component will survive without failure during a specified time interval $[0, t]$ under given operating conditionals, denoted by S, which is defined as [7]:

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du \quad , t \geq 0 \tag{1}$$

Where, T is a random variable, t is the time of death.

The survival function $S(t)$ is the probability that the patient will survival till time t. Survival probability is usually assumed to approach zero as age increases, with:

$$S(0) = 1$$

$$\lim_{t \to \infty} S(t) = 0$$

$S(t)$ is decreasing and continuous from right side. It is linked with the failure distribution function $F(t)$ and in fact, it is the complement of it, i.e.:

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

$$\frac{d}{dt} S(t) = -f(t) \tag{3}$$

2.2 Life Time Distribution Function

The life time distribution function, is defined as the complement of the survival function [1],

$$F(t) = P(T \leq t) = 1 - S(t) \tag{4}$$

If $F(t)$ is differentiable then the derivative, which is the density function of the lifetime distribution is,

$$f(t) = \frac{d}{dt} F(t) \tag{5}$$

The function $f(t)$ is sometimes called the event density; it is the rate of death or failure events per unit time.

2.3 Hazard Function

Hazard function is also known as the immediate failure rate [8]. This is the limit of the conditional probability that an item will fail in the time interval $[t, t+\Delta t]$ when we know that the item is functioning at time t is

$$P(t < T \leq t + \Delta t / T > t) = \frac{P(t<T\leq t+\Delta t)}{P(T>t)}$$

$$= \frac{F(t+\Delta t) - F(t))}{S(t)} \quad (6)$$

By dividing this probability by the length of the time interval, Δt, and letting $\Delta t \to 0$, we get the rate function ($h(t)$) , and it is defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{S(t)}$$

$$h(t) = \frac{dF(t)}{dt} \frac{1}{S(t)}$$

$$h(t) = \frac{f(t)}{S(t)} \quad (7)$$

Where $f(t)$ is the failure density functions, and $S(t)$ is the survival function.

Then,

$$H(t) = \frac{-d ln S(t)}{dt} \quad (8)$$

$$S(t) = \exp(-\int_0^t h(s) ds)$$

So hazard is instantaneous mortality rate conditional on previous survival, and the integrated form of cumulative hazard

$$H(t) = \int_0^t h(s) ds = -\ln S(t) \quad (9)$$

3. Parametric Methods

3.1. Weibull Distribution

The Weibull distribution is continuous distribution. It is one of the most widely applied life distributions in reliability analysis [9] and [10].

The probability density function is:

$$f(t) = \begin{cases} \frac{\alpha}{\lambda} (\frac{t}{\lambda})^{\alpha-1} e^{-(\frac{t}{\lambda})^{\alpha}} & t > 0 \\ 0 & elsewhere \end{cases} \quad (10)$$

Where ($\alpha > 0$) is shape parameter and ($\lambda > 0$) is the scale parameter of the distribution. The mean and variance of Weibull distribution are respectively:

$$E(t) = \lambda\Gamma(1 + \frac{1}{\alpha}) \quad (11)$$

$$V(t) = \lambda^2[\Gamma\left(1 + \frac{1}{\alpha}\right) - \Gamma^2(1 + \frac{1}{\alpha}) \quad (12)$$

The cumulative distribution function is defined as:

$$F(t) = 1 - e^{-(\frac{t}{\lambda})^{\alpha}}, t > 0 \quad (13)$$

The survival function is defined as:

$$S(t) = e^{-(\frac{t}{\lambda})^{\alpha}} \quad (14)$$

The failure rate (or hazard rate) function is given by:

$$h(t) = \frac{\alpha}{\lambda}(\frac{t}{\lambda})^{\alpha-1} = \alpha\lambda^{-\alpha}t^{\alpha-1} \quad (15)$$

3.2. Gumbel Distribution

Gumbel (1958) denotes this distribution of the type I distribution of the smallest extreme, called the Gumbel distribution of the smallest extreme [8]. The Gumbel distribution is a very common distribution due to its global applicability in several fields and its wide applications [11] and [13]. The probability density function is:

$$f(t) = \begin{cases} \frac{1}{\beta}\exp[-(z + \exp(-z))] & -\infty < t < \infty \\ 0 & elsewhere \end{cases} \quad (16)$$

where $z = \frac{t-\mu}{\beta}$, μ is the location parameter, and β is the scale parameter of the distribution.

The mean and variance of $Weibull$ distribution are respectively:

$$E(t) = \mu + \beta\gamma , \quad (17)$$

γ is Euler's constant (0.577215)

$$V(t) = (\beta\pi)^2/6 \quad (18)$$

The cumulative distribution function is defined as:

$$F(t) = 1 - \exp[-\exp(-z)], -\infty < t < \infty \quad (19)$$

The survival function is defined as:

$$S(t) = \exp[-\exp(-z)] \quad (20)$$

The failure rate (or hazard rate) function is given by:

$$h(t) = \frac{\frac{1}{\beta}\exp[-(z+\exp(-z))]}{\exp[-\exp(-z)]} \quad (21)$$

3.3.Exponential Distribution

The exponential distribution is a special case of two- parameter $Weibull$ distribution (or gamma distribution) when the shape parameter is $(\alpha = 1)$ in equation 10, then the probability density function is [11]:

$$f(t) = \begin{cases} \frac{1}{\lambda}\, e^{-\frac{t}{\lambda}} & t > 0 \\ 0 & elsewhere \end{cases} \tag{22}$$

where $\lambda > 0$ is the scale parameter of the distribution, and the mean and variance of exponential distribution are respectively:

$$E(t) = \lambda \tag{23}$$

$$V(t) = \lambda^2 \tag{24}$$

The cumulative distribution function is defined as:

$$F(t) = 1 - e^{-\frac{t}{\lambda}} \ \ , t > 0 \tag{25}$$

The survival function is defined as:

$$S(t) = e^{-\frac{t}{\lambda}} \tag{26}$$

The failure rate (or hazard rate) is given by:

$$h(t) = \frac{1}{\lambda} \tag{27}$$

3.4. Log- Logistic Distribution

Log- logistic distribution is widely used in survival analysis when the failure rate function presents an unmoral shape [5]:

$$f(t) = \begin{cases} \frac{\beta}{\alpha}(\frac{t}{\alpha})^{\beta-1}[1 + (\frac{t}{\alpha})^{\beta}]^{-2} & t > 0 \\ 0 & elsewhere \end{cases} \tag{28}$$

where ($\alpha > 0$) is scale parameter and ($\beta > 0$) is the shape parameter of the distribution.

The mean and variance are respectively [12],

$$E(t) = \frac{\pi\alpha\beta^{-1}}{\sin(\pi\beta^{-1})} \quad , \beta > 0 \tag{29}$$

And

$$V(t) = \frac{2\pi\alpha^2\beta^{-1}}{\sin(2\pi\beta^{-1})} - [\frac{\pi\alpha\beta^{-1}}{\sin(\pi\beta^{-1})}]^2 \ , \beta > 2 \tag{30}$$

The cumulative distribution function is defined as:

$$F(t) = 1 - \ \ [1 + (\frac{t}{\alpha})^{\beta}]^{-1} \tag{31}$$

Also the survival function is defined as:

$$S(t) = [1 + (\frac{t}{\alpha})^{\beta}]^{-1} \quad (32)$$

The hazard function is given by:

$$h(t) = \frac{\beta}{\alpha}(\frac{t}{\alpha})^{\beta-1}[1 + (\frac{t}{\alpha})^{\beta}]^{-1} \quad (33)$$

4. Goodness of Fit Test

In order to compare the distributions, we consider some other criterion like Akaike information Criterion (AIC), Akaike information criterion corrected ($AICC$) and Bayesian information Criterion (BIC) for the real data set [2]. The best distribution corresponds to lower AIC, $AICC$, and BIC values [7] and [13]:

$$AIC = -2\log L + 2k \quad (34)$$

$$AICC = AIC + \frac{2k(k+1)}{n-k-1} \quad (35)$$

$$BIC = -2\log L + k\log n \quad (36)$$

Where k is the number of parameters in the statistical model, n the sample size and L is the maximized value of the likelihood function for the estimated model.

5. Data Analysis and Results

The dataset used in this study consists of a sample of (118) lung cancer patients obtained from Salman and Farhan [6] and given in Table 2.

We can find the estimated value of the parameters and its confidence intervals for the distributions by using maximum likelihood estimation method as follows:

Table1: Maximum likelihood estimates parameters of the distributions

$Model$	$Estimates$	$95\%\ C.I$
$Weibull$	$\alpha = 3.05$ $\lambda = 373.58$	$[2.62\ -\ 3.54]$ $[351.30\ -\ 397.28]$
$Gumbel$	$\mu = 394.02$ $\beta = 102.78$	$[374.42 -\ 413.61]$ $[89.521 - 118.01]$
$Exponential$	$\lambda = 337.15$	$[281.49\ -\ 403.82]$
$Log -\ logistic$	$\beta = 5.80$ $\alpha = 0.26$	$[5.72\ -\ 5.88]$ $[0.22 - 0.30]$

The survival function estimations of time are obtained by substituting these estimated values of the parameters in equations (14), (20), (26), and (32) as shown in table 2. And the survival plots of the W.D, G.D, E.D and L.L.D are shown in figures 1, 2, 3, and 4 respectively.

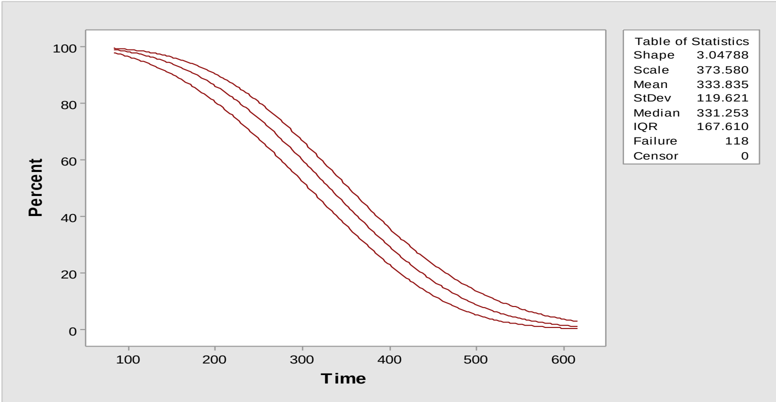

Figure 1: The curve of Weibull distribution for survival function

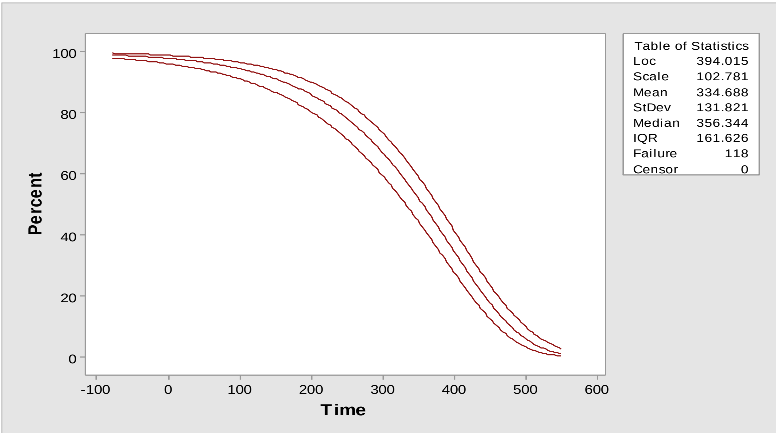

Figure 2: The curve of Gumbel distribution for survival function

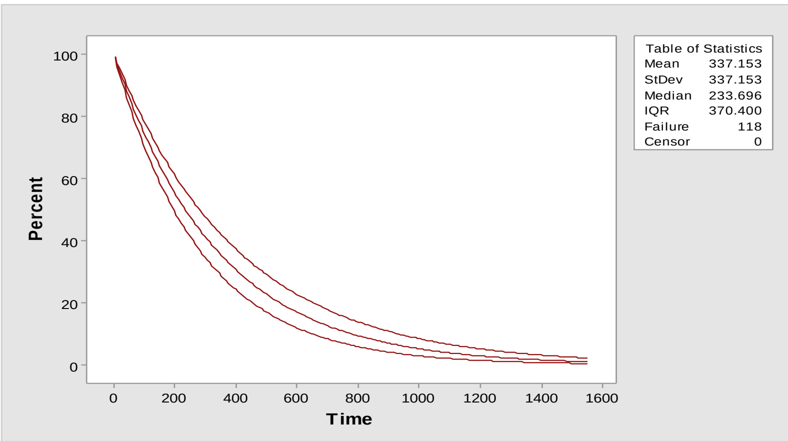

Figure 3: The curve of exponential distribution for survival function

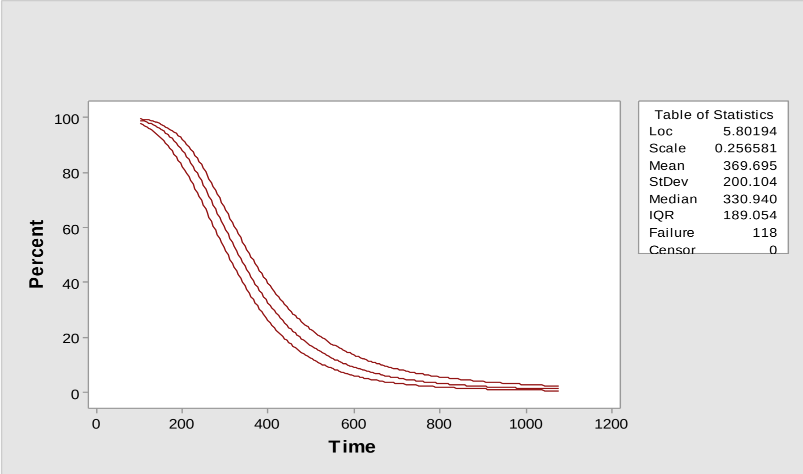

Figure 4: The curve of log logistic distribution for survival function

Table 2: Estimated values of the survival function [6]

No.	Time /Day	$\hat{S}(W.D)$	$\hat{S}(G.D)$	$\hat{S}(E.D)$	$\hat{S}(L.D)$	No.	Time /Day	$\hat{S}(W.D)$	$\hat{S}(G.D)$	$\hat{S}(E.D)$	$\hat{S}(L.D)$
1	3	1.000	0.978	0.991	1.000	60	341	0.469	0.550	0.364	0.471
2	37	0.999	0.969	0.896	0.999	61	342	0.466	0.547	0.363	0.468
3	72	0.993	0.957	0.808	0.997	62	345	0.456	0.538	0.359	0.460
4	75	0.993	0.956	0.801	0.997	63	349	0.444	0.524	0.355	0.448
5	91	0.987	0.949	0.763	0.994	64	354	0.428	0.508	0.349	0.435
6	100	0.982	0.944	0.743	0.991	65	357	0.419	0.498	0.347	0.427
7	103	0.981	0.943	0.737	0.989	66	363	0.400	0.477	0.341	0.411
8	121	0.968	0.932	0.699	0.981	67	364	0.397	0.474	0.339	0.408
9	127	0.964	0.928	0.686	0.977	68	364	0.397	0.474	0.339	0.408
10	140	0.951	0.919	0.660	0.966	69	366	0.391	0.467	0.338	0.403
11	154	0.935	0.908	0.633	0.952	70	367	0.388	0.464	0.337	0.401
12	156	0.933	0.906	0.629	0.949	71	368	0.385	0.460	0.336	0.398
13	164	0.922	0.899	0.615	0.939	72	371	0.376	0.449	0.333	0.391
14	186	0.888	0.876	0.576	0.904	73	373	0.369	0.443	0.331	0.386
15	211	0.839	0.845	0.535	0.853	74	380	0.349	0.419	0.324	0.369
16	212	0.837	0.844	0.533	0.850	75	387	0.328	0.393	0.317	0.352
17	213	0.835	0.842	0.532	0.848	76	387	0.328	0.393	0.317	0.352
18	217	0.826	0.836	0.525	0.838	77	392	0.314	0.375	0.313	0.341
19	218	0.824	0.835	0.524	0.836	78	393	0.311	0.372	0.312	0.339
20	221	0.817	0.831	0.519	0.828	79	397	0.300	0.357	0.308	0.330
21	221	0.817	0.831	0.519	0.828	80	399	0.295	0.350	0.306	0.325
22	233	0.789	0.812	0.501	0.797	81	400	0.292	0.346	0.305	0.323

23	240	0.772	0.789	0.491	0.778	82	400	0.292	0.346	0.305	0.323
24	241	0.769	0.798	0.489	0.775	83	401	0.289	0.343	0.304	0.321
25	243	0.764	0.794	0.486	0.769	84	402	0.286	0.339	0.304	0.319
26	249	0.748	0.784	0.478	0.752	85	407	0.273	0.322	0.299	0.309
27	254	0.735	0.774	0.471	0.737	86	409	0.268	0.314	0.297	0.305
28	266	0.701	0.749	0.454	0.701	87	419	0.242	0.279	0.289	0.285
29	273	0.681	0.735	0.445	0.679	88	421	0.237	0.272	0.287	0.281
30	276	0.672	0.728	0.441	0.669	89	421	0.237	0.272	0.287	0.281
31	277	0.669	0.726	0.440	0.667	90	422	0.235	0.269	0.286	0.279
32	278	0.666	0.724	0.438	0.664	91	422	0.235	0.269	0.286	0.279
33	281	0.657	0.717	0.435	0.654	92	423	0.232	0.266	0.285	0.278
34	290	0.630	0.695	0.423	0.626	93	427	0.222	0.252	0.282	0.270
35	301	0.596	0.667	0.410	0.591	94	428	0.220	0.247	0.281	0.269
36	301	0.596	0.667	0.410	0.591	95	430	0.215	0.242	0.279	0.265
37	301	0.596	0.667	0.410	0.591	96	446	0.179	0.191	0.266	0.238
38	302	0.593	0.665	0.408	0.588	97	450	0.171	0.178	0.263	0.232
39	304	0.587	0.659	0.406	0.582	98	454	0.163	0.167	0.260	0.226
40	304	0.587	0.659	0.406	0.582	99	416	0.249	0.289	0.291	0.291
41	306	0.581	0.654	0.404	0.576	100	463	0.146	0.141	0.253	0.213
42	307	0.577	0.651	0.402	0.573	101	470	0.133	0.123	0.248	0.203
43	307	0.577	0.651	0.402	0.573	102	477	0.122	0.106	0.243	0.194
44	308	0.574	0.649	0.401	0.569	103	481	0.115	0.097	0.240	0.189
45	313	0.558	0.635	0.395	0.554	104	481	0.115	0.097	0.240	0.189
46	313	0.558	0.635	0.395	0.554	105	483	0.112	0.093	0.239	0.186
47	314	0.555	0.632	0.394	0.551	106	483	0.112	0.093	0.239	0.186
48	318	0.542	0.621	0.389	0.539	107	497	0.092	0.066	0.229	0.170
49	330	0.504	0.585	0.376	0.503	108	511	0.074	0.044	0.220	0.155
50	331	0.501	0.582	0.375	0.499	109	512	0.073	0.043	0.219	0.154
51	332	0.498	0.579	0.374	0.497	110	512	0.073	0.043	0.219	0.154
52	332	0.498	0.579	0.374	0.497	111	516	0.069	0.038	0.216	0.150
53	334	0.491	0.573	0.371	0.491	112	517	0.068	0.037	0.216	0.149
54	334	0.491	0.573	0.371	0.491	113	519	0.066	0.034	0.215	0.148
55	335	0.488	0.569	0.370	0.488	114	533	0.052	0.021	0.206	0.135
56	335	0.488	0.569	0.370	0.488	115	534	0.051	0.021	0.205	0.134
57	335	0.488	0.569	0.370	0.488	116	535	0.050	0.19	0.205	0.133
58	335	0.488	0.569	0.370	0.488	117	540	0.046	0.016	0.202	0.129
59	338	0.479	0.560	0.367	0.479	118	550	0.039	0.010	0.196	0.121

Table 3. Criteria for comparison

$Model$	$-LL$	AIC	$AICC$	BIC
$Weibull$	737.89	1479.78	1479.88	1479.92
$Gumbel$	729.93	1463.86	1463.96	1464.00
$Exponential$	804.82	1611.65	1611.68	1611.71
$Log-logistic$	757.03	1518.06	1518.16	1518.20

The results in Table 3 indicate that the $Gumbel$ distribution has the lowest AIC, $AICC$ and BIC values than the Weibull, log- logistic, and exponential. Hence Gumbel distribution leads to a better fit than the other three distributions.

6. Conclusions

From the practical work, it is concluded that the Gumbel distribution has the lowest AIC, $AICC$ and BIC values than the Weibull, exponential, and log- logistic distributions. We conclude that the survival function for the lung cancer by using Gumbel distribution model is the best. And the expected values of the survival function of all estimation methods which are proposed in this article has been decreasing progressively with increasing failure times for lung cancer patients: this means that there is an opposite relationship failure times and survival function.

References

1. Mohammed, A. M., Shao T., Expected the Life Time for Heart Patients by Using Cox Regression Model, *International Journal of Applied Research*, **2015**, *1, 9*, 883-889.
2. Merovci, F., Transmuted Generalized Rayleigh Distribution, *Journal of Statistics Applications and Probability*, **2014**, 3, 1, 9-20, htt://dx.doi.org/10.12785/jsap/030102.
3. Zhu, H. P. et al., Application of Weibull Model for Survival of Patients with Gastric Cancer, *BMC Gastro enter bogy*, **2011**, 11,. 1, http://www.biomdcentral.
4. Abernethy, R. B., The New Weibull Hand Book, Ch.1, an over View of Weibull analysis, **2006**, http://www.barringer1.com/pdf/Chpt1-5th-edition.pdf
5. Akhtar, M. ; Khan, A., A. Log- Logistic Distribution as a Reliability Model: A Bayesian Analysis, *American Journal of Mathematics and Statistics*, **2014**, *4, 3*, 162-170, doi:10.5923/j.ajms20140403.05.
6. Salman, A. N. ; Farhan, I. H., Estimate Complete the Survival Function for Real Data of Lung Cancer Patients, *Ibn Al-Haitham Journal for Pure &Applied Sciences*, **2014**, *27, 3*, 531-541.
7. Yong, T., Extended Weibull Distributions in Reliability Engineering, *PhD. Theses*, **2004**, National University of Singapore.

8. Rausand, M. ; Hoyland, A., System Reliability Theory models, Statistical Methods, and Applications, 2nd Edition, **2004**, John Wiley &Sons, INC, Canada, ISBN 0-471-47133-X.
9. Chen, M.; Zhang, Z.; Cui, C., On the Bias of the Maximum Likelihood Estimators of Parameters of the Weibull Distribution, *Mathematical and Computational Application*, **2017**, *22*, *19*, 1-18, 2017, doi: 103390/mca2210019.
10. Varela, J. Gorgoso ; Rojo-Alboreca, A., Use of Gumbel and Weibull Functions to Model Extreme Values of Diameter Distribution in Forest Stands, *Annals of Forest Science*, **2014**, *71*, 741-750.
11. Deka, D., Bhanita Das, B. ; Baruah, B. K., Transmuted Exponential Gumbel Distribution and its Application to Water Quality DATA, *pak.j.stat.oper.res*, **2017**, *13, 1*, 115-126.
12. Santana, T. V.; Ortega, E. M. M. Cordeiro, G. M ; Silva, G. O., The Kumaraswamy-Log- Logistic Distribution, *Journal of Statistical Theory and Applications*, **2012**, *11, 3*, 265-291.
13. Vllah, E. ; Shahazad, M. N., Transmutation of the Two Parameters Rayleigh Distribution, *International Journal of Advanced Statistics and Probability*, **2016**, *4, 2*, 95-101, doi:10.1419/ijasp.v4i2.6100