Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/090315

Model Suitability Analysis of Survival Time to Ovarian Cancer Patients Data

Manoj Kumar[1], *Sandeep K. Maurya*[2,*], *Sanjay K. Singh*[3], *Umesh Singh*[3] *and Anurag Pathak*[1]

[1]Department of Statistics, Central University of Haryana, Haryana, India
[2]Department of Statistics, Central University of South Bihar, Gaya, India
[3]Department of Statistics, Banaras Hindu University, Varanasi, India

Received: 27 Oct. 2018, Revised: 25 Nov. 2019, Accepted: 5 Dec. 2019
Published online: 1 Nov. 2020

Abstract: In this paper, we propose a suitable statistical model for survival time of the ovarian cancer patients data. The proposition followed by checking the suitability of twelve lifetime models through different statistical tools like the value of logarithmic of likelihood, Akaike Information Criterion, Kolmogorov-Smirnov distance and Bayesian Information Criterion. The maximum likelihood estimate of the parameters for the considered models has been obtained. Also, the non-parametric procedure has been used to show the validity of the conclusion.

Keywords: Lindley distribution, hazard rate, maximum likelihood estimation, Kaplan-Meier estimate, TTT plot

1 Introduction

The stage of cancer depends on the growth and spread of cancer cell which can be identified on the basis of various tests and scanning procedures. There are four stages in ovarian cancer as: stage 1, stage 2, stage 3 and stage 4. All the stages are again subdivided into three sub-stages. In stage 1, cancer is spread in the only ovary. When the cancer is spread inside one ovary, this is the condition of stage 1a. In stage 1b, cancer is spread inside both ovaries, and in stage 1c, cancer is spread inside both ovaries with some cancer cell on the surface of the ovary and the chance of bursting ovary before or during the surgery. In stage 2 ovarian cancer, the cancer spread outside the ovaries and extend inside the circular area by hip bones. In stage 2a, cancer cell spread into the womb. In stage 2b, the cancer spread to the other tissues in the pelvis, and in stage 2c, cancer grows inside the abdomen. In stage 3, cancer is spread outside the pelvis into the abdominal cavity and also spread to lymph nodes in the upper abdomen. In stage 3a, cancer cells are found in the lining of the abdomen. In 3b stage, cancer is 2cm or smaller on the lining of the abdomen, and in stage 3c, more than 2cm and cancer spread behind the womb. In stage 4, the cancer spread to other nearby ovaries organs. The 4a cancer stage is called pleural effusion, and in stage 4b, cancer spread inside other body part nearby ovaries like spleen or liver. Stage 4c is expansion of its earlier stage.

Surgery is the main treatment for most ovarian cancer and it depends on spread and health condition of the patients. The surgery is also done for the early stage and in advanced stage ovarian cancer. In early stage ovarian cancer, the surgeon removes ovaries, and womb and in the advanced stage, cancer is usually shrink and controlled for as long as possible. This might be for many months and sometimes years. Chemotherapy, radiotherapy, and surgery can all be used to treat advanced ovarian cancer. The aim of the use of chemotherapy is to shrink and, thus, make easier to remove the cancer. This depends on the condition that if cancer is removed after surgery then the aim of chemotherapy is to reduce the chance of occurrence of cancer again, and if cancer is not removed then the aim of chemotherapy is to shrink the cancer. There are also some side effects of chemotherapy depending on the drugs, amount of drugs, total treatment time and health of patients like nausea, vomiting, hair loss, mouth sores, increased infections, fatigue, bleeding after miner cut, etc. Chemotherapy may cause permanent infertility and early menopause. In most of the cases, both ovaries are removed (for more detailed about ovarian cancer see [1], [2], [3], [4] etc).
The aim of the paper is to propose a lifetime distribution for survival time of ovarian cancer patient data. The distribution

* Corresponding author e-mail: sandeepmaurya.maurya48@gmail.com

function plays an important role to solve uncertain real life problem and draw inference on the basis of deductive or inductive reasoning. Because a suitable model can cover the randomness and enables us to draw a valid conclusion about the population on the basis of the small sample in the face of uncertainty, which can be verified mathematically.
The rest of the paper is organized as follows: In Section 2, we have discussed the data set and its nature. Section 3, twelve lifetime models having same nature as the data show have been considered. Section 4, estimation procedure for parameters has been discussed. In Section 5, model section criterion is considered. In Section 6 and Section 7 provides the comparisons of the models and non-parametric procedure respectively. Finally, Section 8 concludes the paper.

2 Description of the data set

This dataset consists of survival time (in days) of 26 ovarian cancer patient having all tumour masses greater than 2cm (in diameter) after their surgical treatment of ovarian cancer. Dataset is propose by [5]. If the cancer comes back less than 6 months after having chemotherapy, a specialist may suggest one or more of the treatments like paclitaxel alone, liposomal doxorubicin, gemcitabine, etoposide and /or cyclophosphamide. [6] compares the anti-tumour effects of two different forms of chemotherapy treatment. [7] shows that Poisson - Exponential distribution (PED) fitted this data set and also studied Bayesian estimation under progressive type-II censored. The nature of the dataset has an increasing failure rate (IFR) and can be verified by using the concept of scaled total time to test (TTT) plot (follow [8] for the concept of scaled TTT plot). The scaled TTT plot is given in Figure 1. Hence, we need a probability model having the same nature of failure rate, i.e. IFR.

3 Some lifetime models

Let X be the random variable denoting the survival time of ovarian cancer patients having the surgery of ovarian cancer. And our aim is to develop a distribution function for the random variable which is denoted by F(x) and defined as:

Let $F(x) = P(X < x) = P$ (the survival time of ovarian cancer patient having surgical treatment before the survival time x), where $P(\cdot)$ denotes the probability.

In statistical literature, there is a number of lifetime distributions for various fields of life likewise in medical, engineering, finance, etc. to give a decision in the presence of uncertainty. Some frequently used IFR lifetime distributions are exponential, gamma, Weibull, Lindley, log-normal, etc. One model cannot be used in all real situation uniformly, because each has their utilities in different phenomenon of life depending on the problem and nature of hazard of the model. Lindley distribution is one of the famous lifetime distribution proposed by [9], and further statistical properties with real data application is studied by [10].

Many generalizations and transformations have been done by taking Lindley distribution as a baseline model by various authors with real data problems like in relief time of patients (see [11]), study of stress-strength reliability ([12]). [13] proposed Poisson-Lindley distribution, [14] proposed discrete Lindley distribution, [11] proposed two parameter generalized Lindley distribution and [15] proposed two parameter model having various failure rate shapes. But it is noticeable that nearly all (not necessarily for all see [16], [17] etc.) generalisations or transformations based on baseline distribution add some additional parameters to existing model which also create complexities in further inferential procedures. Perhaps taking this point in mind, [18] proposed exponential transformed Lindley (ETL) distribution and derived its various statistical properties with a real data set to show the model suitability for that dataset in comparison to eight other well-known distributions.

Here we consider twelve famous lifetime models having IFR and to find which distribution function support for the survival time of ovarian cancer patient data very well. The considered models have been discussed one by one as given below:

Model 1: The first model is exponential distribution having constant hazard rate, reason behind taking non-IFR model is that, we want to check whether the model is fit for ovarian cancer data or not. The reason for considering it here is that the exponential distribution is a very famous model and nearly all generalizations or transformations have been done on it. The CDF is

$$F(x) = 1 - e^{-\theta x} \qquad \theta > 0; x > 0.$$

Model 2: The second model which we consider is one parameter DUS exponential distribution proposed by [16]. They have shown the model applicability on Bladder cancer patients data. This is the reason for considering it here. The CDF is given as

$$F(x) = \frac{\exp\left[1 - e^{-\theta x}\right] - 1}{e - 1} \qquad \theta > 0; x > 0.$$

Model 3: The third model is GDUS exponential distribution proposed by [15]. They show the model applicability on four different real data sets over eight other competitive models. GDUS exponential is a very flexible model having different shapes of hazard rates. The CDF is given as

$$F(x) = \frac{\exp\left[1-e^{-\theta x}\right]^{\alpha}-1}{e-1} \qquad \alpha;\theta>0;x>0.$$

Model 4: The fourth model is Lindley distribution proposed by [9]. Lindley distribution is increasing popular lifetime model, and it is applicable in different fields of real life problems. The CDF is given as

$$F(x) = 1-e^{-\theta x}\left(1+\frac{\theta x}{1+\theta}\right) \qquad \theta>0;x>0.$$

Model 5: The fifth model is exponential transformed Lindley (ETL) distribution proposed by [18]. They show the model applicability on cycle of yarn failure data in comparison to seven other models. The CDF is given as

$$F(x) = \frac{\exp\left[1-e^{-\theta x}\left(1+\frac{\theta x}{1+\theta}\right)\right]-1}{e-1} \qquad \theta>0;x>0.$$

Model 6: The sixth model is generalized Lindley (GL) distribution proposed by [11]. They show the model applicability on relief time of patients data in comparison to three other models. The CDF is given as

$$F(x) = \left[1-e^{-\theta x}\left(1+\frac{\theta x}{1+\theta}\right)\right]^{\alpha} \qquad \alpha;\theta>0;x>0.$$

Model 7: The seventh model is gamma distribution. Gamma distribution is one of the famous model in various lifetime problems, and many research work have been done by taking the gamma distribution with the probability density function (PDF)

$$f(x) = \frac{\theta^{\alpha}e^{-\theta x}x^{\alpha-1}}{\Gamma(\alpha)} \qquad \alpha;\theta>0;x>0.$$

Model 8: The eighth model is Weibull distribution. Weibull distribution is also a famous model, and many more generalizations have been done by taking it as a baseline model with the CDF

$$F(x) = 1-e^{-(\frac{x}{\theta})^{\alpha}} \qquad \alpha;\theta>0;x>0.$$

Model 9: The ninth model is generalized exponential (GE) distribution proposed by [19]. GE distribution is a competitive model in comparison to gamma and Weibull distribution. The CDF of GE is

$$F(x) = (1-e^{-\theta x})^{\alpha} \qquad \alpha;\theta>0;x>0.$$

Model 10: The tenth model is Chen's distribution proposed by [20]. Chen's model is also a very flexible model having different shapes of hazard rates including bathtub shape. The CDF is

$$F(x) = 1-e^{\theta(1-e^{x^{\alpha}})} \qquad \alpha;\theta>0;x>0.$$

Model 11: The eleventh model is Poisson-Exponential (PE) distribution proposed by [21]. It shows the model applicability on ball bearing data set. The CDF of PE is

$$F(x) = \frac{e^{-\theta e^{-\alpha x}}-e^{-\theta}}{1-e^{-\theta}} \qquad \alpha;\theta>0;x>0.$$

Model 12: The twelve model is one parameter IFR logarithmic transformed exponential (LTE) distribution proposed by [17]. It shows the model applicability on item failure data with five other competitive models. The CDF is given as

$$F(x) = 1-\frac{\log(1+e^{-\theta x})}{\log 2} \qquad \theta>0;x>0.$$

All of the above models (except exponential) have IFR according to the nature of the survival time of ovarian cancer data.

4 Model selection criteria

In this section, we want to check which model is fitter to the dataset among the considered models. Here, we used Kolmogorov-Smirnov (KS) distance (D), p- value, minus of logarithmic of likelihood (-LogL) values, and model selection criteria, i.e. Akaike Information Criterion (AIC) ([22]) and Bayesian Information Criterion (BIC). First of all, we check whether at 5% level of significance the considered model is fit to the dataset or not. If some of the models are fit, then we check the model having the least value of these values. KS distance is the distance between empirical distribution function of sample data and considered distribution function. The hypothesis under the KS statistics is:

Null hypothesis H_0= Sample of survival time of ovarian cancer patients are coming from the considered model. Alternative hypothesis H_1=Sample of survival time of ovarian cancer patients are coming from any other model.

KS test is used for the goodness of fit. We have chosen the model that has maximum value of likelihood, so we choose the model having the least value of -LogL. The concept of AIC and BIC is the information loss due to assumed model. Along with this BIC, gives more penalties for increasing number of parameters in comparison to AIC, that's why the value of BIC is always more than AIC.

We prefer a model which has the minimum value of AIC and BIC as the less information loss for the considered model. The value of AIC and BIC is defined as

$$AIC = 2 * k - 2 * \log\hat{L}, \qquad BIC = k * \log(n) - 2 * \log\hat{L},$$

and KS statistics is defined as

$$KS = \underset{x}{Sup}\,|F_n(x) - F(x)| \qquad \text{where} \qquad F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{x_i \le x}$$

where $\hat{L}$ is the maximum likelihood for the considered distribution, a number of parameters are k with the sample size of n and $F_n(x)$ is the empirical distribution function.

5 Estimation procedure for the parameter of model

Once we have a model in hand, the next step is to find out its unknown value of the parameter. Here, we have considered the method of maximum likelihood estimation (MLE) for estimating the unknown parameter. This method of estimation is based on the concept of maximizing the likelihood function (L). Since the logarithmic function is a monotone function and maximising the logarithmic (Log) of the likelihood is same as maximising the likelihood function, we prefer to use Log likelihood function instead of likelihood function for easy calculation. Using the concept of maxima and minima, the MLE can be obtained by differentiating the log likelihood function (Log L) with respect to the parameter and put it to zero, i.e. $\frac{d\log L}{d\theta} = 0$ and find the value of the parameter. This equation is called likelihood equation. Then put the estimate of the parameter say, $\hat{\theta}$, in the second derivative and check it whether it is negative or not. If the second derivative got negative, then $\hat{\theta}$ is called the MLE of θ.

Sometimes the likelihood equation have a non-linear equation and the equation is not in closed form and cannot be solved analytically. So we have to use some numerical technique for the solution. Here, we propose to the use of Newton-Raphson method through [23]. In the case of choice of the initial guess, contour plot method is used. In the case of large samples, we can obtain the confidence intervals based on the diagonal elements of Fisher information matrix $I^{-1}(\hat{\theta})$ which provides the estimated asymptotic variance for the parameter θ. Thus, two-sided $100(1-\eta)\%$ confidence interval of θ can be defined as $\hat{\theta} \pm Z_{\eta/2}\sqrt{var(\hat{\theta})}$, where $Z_{\eta/2}$ stands for the upper $\eta/2\%$ points of standard normal distribution.

The Fisher Information matrix can be estimated by

$$I(\hat{\theta}) = \left[\frac{-d^2 \log L}{d\theta^2}\right]_{\hat{\theta}} \tag{1}$$

6 Comparison of the models

For all the above twelve considered lifetime models, we want to check which model is fitter to dataset. From Table 1, we can say that only exponential distribution is not fitted for survival time of ovarian cancer data set at 5% degree of significance, the reason is that exponential distribution has constant hazard rate while data shows IFR. In all of the accepted models, GDUSE has the least KS distance and when we see -LogL criterion; Weibull has the least value and

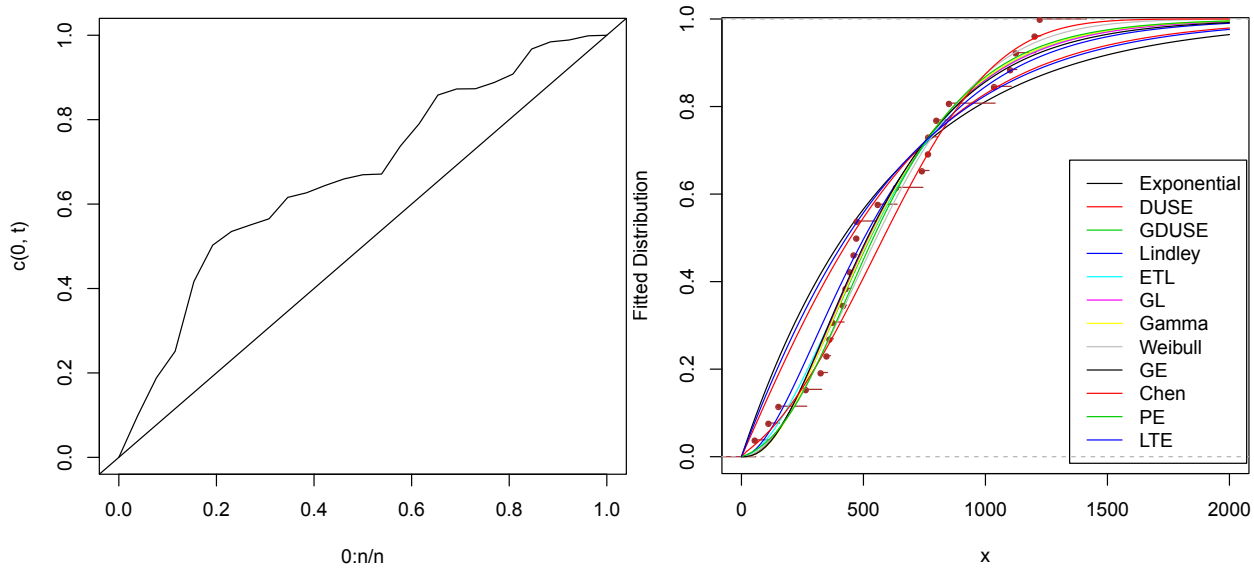

Fig. 1: TTT plot of the survival time (days) of ovarian cancer patient and empirical CDF with fitted CDF plot of considered data set.

ETL is at third place. But the table shows that all the values are the same in integer place and only changes in decimal place. So, we can consider ETL be comparative to Weibull, Chen, and PE distributions. Also, KS distance gives the maximum distance between empirical and fitted CDF, then it may be possible that a model fits to the dataset at all points except one or some points and results in the KS value becoming larger. Now, the model selection criterion AIC is least for ETL and is also least in BIC criterion in comparison to all others models. Hence, we conclude that ETL distribution is the most suitable model for survival time of ovation cancer patients having surgical treatment as compared to other eleven distributions. The empirical cumulative distribution function (ECDF) and fitted CDF plot, for the considered dataset, has been shown in Figure 1. This figure also justify our conclusion.

Also, the mean survival time of the 26 ovarian cancer patients are $599.5385 \approx 600$ days. The mean of ETL distribution for this data set is $597.4548 \approx 598$ days which is close to mean days of the data set and the standard deviation is 382 days.

Table 1: MLE, AIC, BIC and KS statistics with p-value and log likelihood value for fitted data sets.

Distribution	*Parameter*		*AIC*	*BIC*	*KS − Test*		*−LogL*
	α	θ			D	P-value	
Exponential	-	0.00167	386.6003	387.8584	0.2684	0.0382	192.3002
DUSE	-	0.00217	382.5291	383.7872	0.2333	0.0998	190.2646
GDUSE	2.43490	0.00326	378.1410	380.6572	0.0996	0.9361	187.0705
Lindley	-	0.00333	377.2004	378.4585	0.1467	0.5800	187.6002
ETL	-	0.00398	375.9227	377.1807	0.1138	0.8520	186.9163
GL	1.29765	0.00378	378.4365	380.9527	0.1060	0.9022	187.2183
Gamma	2.59293	231.22095	378.2609	380.7771	0.1003	0.9327	187.1304
Weibull	1.85005	674.12638	376.9536	379.4698	0.1286	0.7354	186.4768
GE	2.73319	0.00289	378.9196	381.4358	0.1106	0.8740	187.4598
Chen	0.29151	0.00116	377.0967	379.6129	0.1550	0.5106	186.5484
PE	4.39606	0.00345	377.8797	380.3959	0.1172	0.8273	186.9399
LTE	-	0.00205	384.0839	385.3420	0.2519	0.0610	191.0419

7 Non-parametric procedure

In this section, we have proposed to the use of non-parametric method for the dataset. Here, we use Kaplan-Meier estimate of survival. The procedure of Kaplan-Meier method is given below:
Let $S(t)$ be the response variable, denotes the survival time which represent the probability that a cancer patient survive at time t. The method of estimating this function is based on i.i.d. sample of the survival time of ovarian cancer patients. This method has been used in case of censored data. In ovarian cancer, the treatment of patient may undergo a course of chemotherapy treatment, the status of each patient that 0 represent censored and 1 represents uncensored, i.e. It means cancer does not come at the particular survival time then patients are censored or leave, otherwise, surviving till failure of chemotherapy treatment at a different stage. The Kaplan-Meier estimate can be obtained, a series of time in the interval is designed under one death in that interval, and this death occurs at the start of time interval. Here, the chemotherapy surgical treatment is conducted on n ovarian cancer patients with observed survival time $t_1, t_2, \cdots, t_n$. Some of ovarian cancer patients may be right censored, there may also be more than one patient censored with the same observed survival time due to unavailability of medical facilities, etc. There are r death time amongst n patients, where $r \leq n$. After arranging these death times in ascending order, the j^{th} is denoted $t_{(j)}$ for $j = 1,2,3,\cdots,r$, and so the r ordered death times are $t_{(1)} \leq t_{(2)}, \cdots, \leq t_{(r)}$.

The number of patients who are alive just before time $t_{(j)}$, including those who are about to die at the time, will be denoted n_j for $j = 1,2,3,\cdots,r$, and d_j will denote the number of those who die at the time $t_{(j)}$. The time interval from $\left(t_{(j)} - h\right)$ to $t_{(j)}$, where h is an infinitesimal time interval, which includes one death in that time interval. Since there are n_j patients who alive just before $t_{(j)}$ and d_j death at $t_{(j)}$, and the estimated probability of a patient dies in interval $\left(\left(t_{(j)} - h\right), t_j\right)$ is d_j/n_j, while $(n_j - d_j)/n_j$ is the estimated probability of survival for that interval. If no death occurs in interval from $t_{(j)}$ to $\left(t_{(j+1)} - h\right)$ then the correspondent of the probability of surviving is unity. The joint probability of surviving for $\left(\left(t_{(j)} - h\right), t_{(j)}\right)$ and $\left(t_{(j)}, \left(t_{(j+1)} + h\right)\right)$ can be estimated by $(n_j - d_j/n_j)$ (for more details see [24]). Moreover, if $h \to 0$, then $(n_j - d_j)/n_j$ became an estimate of the probability of surviving the interval from $t_{(j)}$ to $t_{(j+1)}$. In this experiment, the occurrence of death is independent to each interval. Therefore, the Kaplan-Meier estimate of the survivor function at time t is given as:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_{(1)}, \\ \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right) & \text{if } t_{(k)} \leq t < t_{(k+1)}; k = 1,2,3,...,r \\ \infty & \text{if } t_{(k+1)} \geq t \end{cases}$$

It is also known as *product − limit estimator* of the survival function.

Now, the standard error (SE) of the estimate $S(t)$ can be defined as the square root of the estimated variance of the estimate, and is used to the construction of an interval estimate for a quantity of interest. The Kaplan-Meier estimate of the survivor function for any value of t in the interval from $t_{(k)}$ to $t_{(k+1)}$ can be defined as follows:

$$\hat{S}(t) = \prod_{j=1}^{k} \hat{p}_j, \quad k = 1,2,3,\cdots,r. \tag{2}$$

where $\hat{p}_j = \frac{(n_j - d_j)}{n_j}$. It is estimated that probability of patient who survives through the time interval that begins at $t_{(j)}, j = 1,2,3,\cdots,r$. Taking both side logarithm of equation (2), then we have

$$\log \hat{S}(t) = \sum_{j=1}^{k} \log\left(\hat{p}_j\right), \quad k = 1,2,3,\cdots,r, \tag{3}$$

and the variance of $\log\left(\hat{S}(t)\right)$ is

$$var\left(\log \hat{S}(t)\right) = \sum_{j=1}^{k} var\left(\log\left(\hat{p}_j\right)\right), \quad k = 1,2,3,\cdots,r. \tag{4}$$

The number of individuals $(n_j - d_j)$ who survive through the interval beginning at $t_{(j)}$ can be assumed to have a $Binomial(n_j, p_j)$. Frequently, $\hat{p}_j = \frac{(n_j - d_j)}{n_j}$ is used for estimation. Since the variance of $\hat{p}_j$ is $var(\hat{p}_j) = \hat{p}_j(1 - \hat{p}_j)/n_j$. As we know that if X is a random variable with a function $g(x)$ then the variance $g(x)$ is

$$var\left(g(x)\right) = \left(\frac{dg(x)}{dx}\right)^2 var\left(x\right).$$

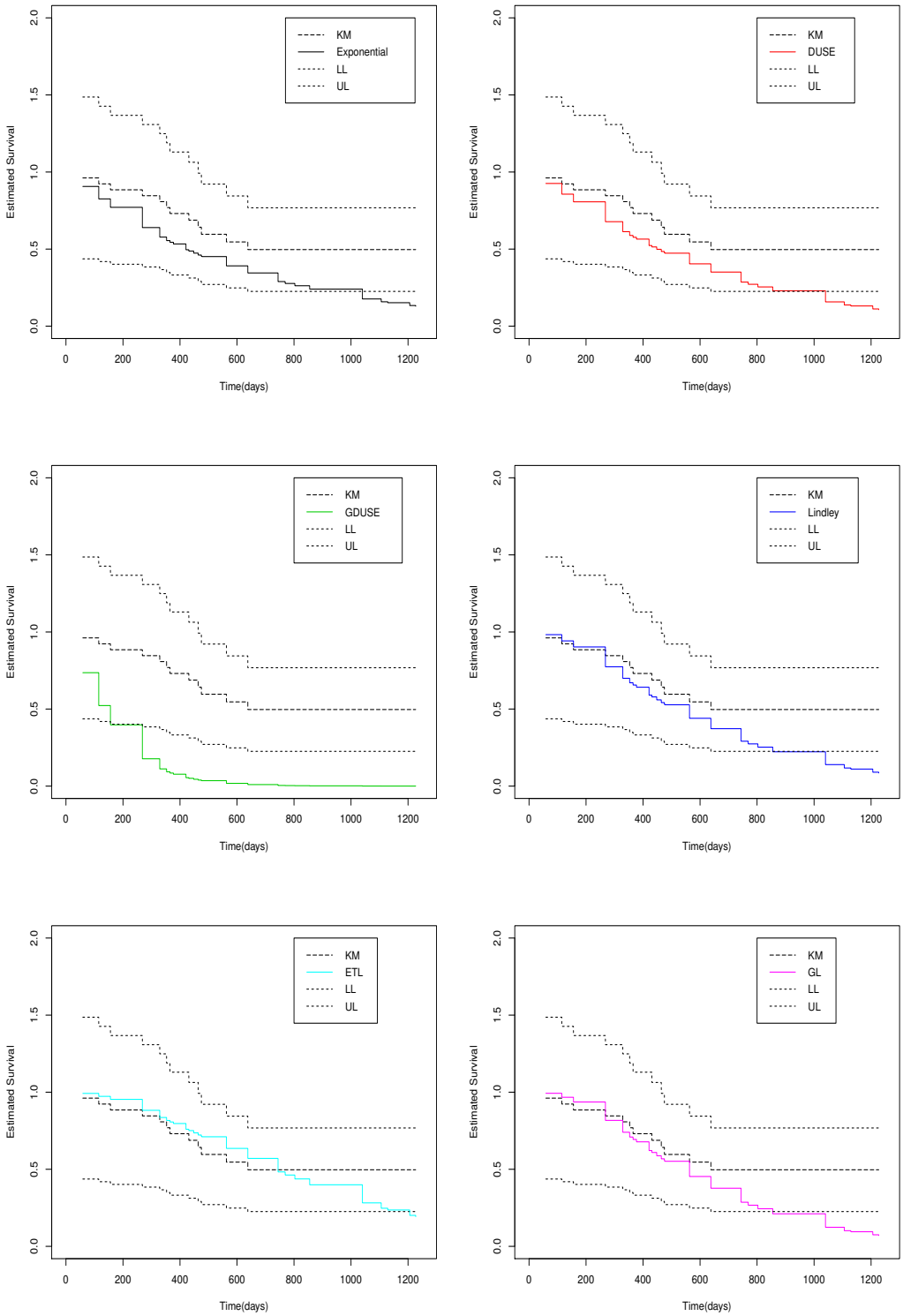

Fig. 2: Estimated survival function and 99% confidence limit for $S(t)$, top left panel: Exponential; top right panel: DUSE; middle left panel: GDUSE; middle right panel: Lindley; bottom left panel: ETL; bottom right panel: GL.

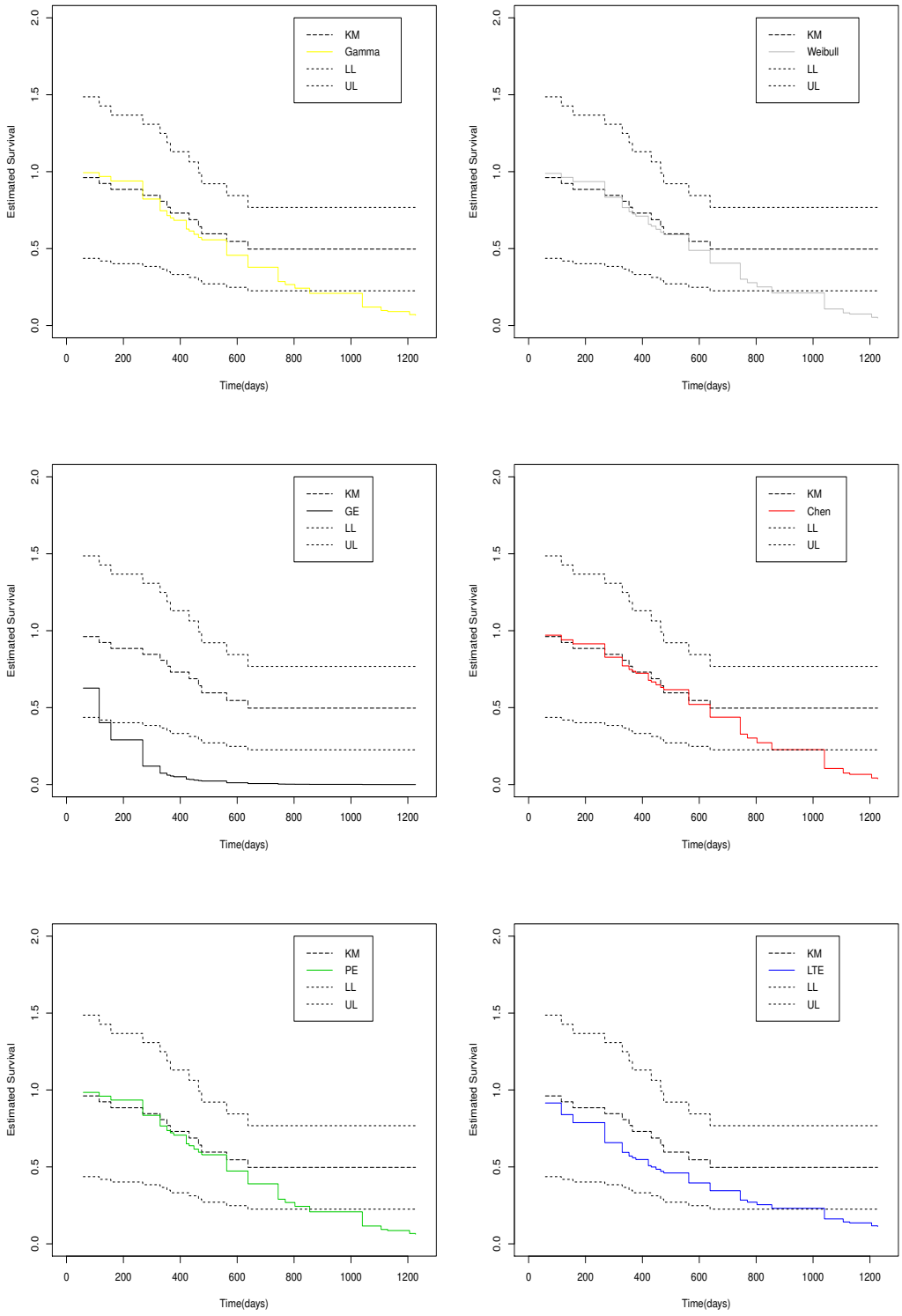

Fig. 3: Estimated survival function and 99% confidence limit for $S(t)$, top left panel: Gamma; top right panel: Weibull; middle left panel:GE; middle right panel: Chen; bottom left panel: PE; bottom right panel: LTE

Table 2: Kaplan-Meier estimate of the survival function, standard error of $\hat{S}(t)$ and confidence interval for $S(t)$

$Time(days)$	n	d	$(n-d)/n$	$\hat{S}(t)$	$SE(\hat{S}(t))$	99% confidence interval
59	26	1	0.96154	0.96154	0.20345	(0.43664, 1.48644)
115	25	1	0.96	0.92308	0.19531	(0.41917, 1.42698)
156	24	1	0.95833	0.88462	0.18717	(0.40171, 1.36753)
268	23	1	0.95652	0.84615	0.17904	(0.38424, 1.30807)
329	22	1	0.95455	0.80769	0.17090	(0.36678, 1.24861)
353	21	1	0.95238	0.76923	0.16276	(0.34931, 1.18915)
365	20	1	0.95	0.73077	0.15462	(0.33184, 1.12970)
377	19	0	1	0.73077	0.15462	(0.33184, 1.12970)
421	18	0	1	0.73077	0.15462	(0.33184, 1.12967
431	17	1	0.94118	0.68778	0.14553	(0.31232, 1.06324)
448	16	0	1	0.68778	0.14553	(0.31232, 1.06324)
464	15	1	0.93333	0.64193	0.13583	(0.29150, 0.99236)
475	14	1	0.92857	0.59608	0.12612	(0.27068, 0.92148)
477	13	0	1	0.59608	0.12612	(0.27068, 0.92148)
563	12	1	0.91667	0.54641	0.11561	(0.24812, 0.84469)
638	11	1	0.90909	0.49673	0.10510	(0.22557, 0.76790)
744	10	0	1	0.49673	0.10510	(0.22557, 0.76790)
769	9	0	1	0.49673	0.10510	(0.22557, 0.76790)
770	8	0	1	0.49673	0.10510	(0.22557, 0.76790)
803	7	0	1	0.49673	0.10510	(0.22557, 0.76790)
855	6	0	1	0.49673	0.10510	(0.22557, 0.76790)
1040	5	0	1	0.49673	0.10510	(0.22557, 0.76790)
1106	4	0	1	0.49673	0.10510	(0.22557, 0.76790)
1129	3	0	1	0.49673	0.10510	(0.22557, 0.76790)
1206	2	0	1	0.49673	0.10510	(0.22557, 0.76790)
1227	1	0	1	0.49673	0.10510	(0.22557, 0.76790)

then

$$var\left(\hat{S}(t)\right) = \left(\hat{S}(t)\right)^2 var\left(\log\left(\hat{S}(t)\right)\right) = \left(\hat{S}(t)\right)^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)},$$

where $var\left(\log\left(\hat{p}_j\right)\right) = \frac{d_j}{n_j(n_j-d_j)}$ and $var\left(\log\left(\hat{S}(t)\right)\right) = \sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)}$.
Therefore, the standard error of Kaplan-Meier estimate of survival function is the square root of the estimated variance of the estimate;

$$SE\left(\hat{S}(t)\right) = \hat{S}(t)\left\{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)}\right\}^{\frac{1}{2}}, \tag{5}$$

for $t_{(k)} \leq t_{(k+1)}$. This is known as Greenwood's formula. Further, we note that in the tails of the distribution of the survival times, that is, when $\hat{S}(t)$ is not close to zero or unity the variance of $\hat{S}(t)$ obtained by using Greenwood's formula can overestimate the actual variance. In such type of situation, Greenwood's formula will be used.

But if the tails of the distribution of the survival times, that is, when $\hat{S}(t)$ is close to zero or unity the variance of $\hat{S}(t)$ may be used. Then an alternative expression for the standard error of $\hat{S}(t)$ may be used. [25] propose that the standard error of $\hat{S}(t)$ should be obtained as

$$SE(\hat{S}(t)) = \frac{\hat{S}(t)\left(1-\hat{S}(t)\right)^{1/2}}{(n_k)^{1/2}}, \quad t_{(k)} \leq < t_{(k+1)}, k = 1,2,3,\cdots,r,$$

where $\hat{S}(t)$ is the Kaplan-Meier estimate of $S(t)$, and $n_{(k)}$ is the number of individual at risk at $t_{(k)}$, the start of the k^{th} constructed time interval. This expression for the standard error of $\hat{S}(t)$ is conservative in the sense that the standard error obtained will tend to be larger than it ought to be. Finally, for this considered problem, the Greenwood estimate is recommended to use.

8 Discussion of results

Figure 2 and Figure 3 show the Kaplan-Meier estimator of the survivor function $\hat{S}(t)$ and $1-F(t,\hat{\Theta})=S(t,\hat{\Theta})$, the survival function of considered lifetime models for 26 ovarian cancer patients with survival time of patients along with point wise 99% confidence interval (LL = Lower Limit, UL = Upper Limit) of $\hat{S}(t)$, where $\hat{\Theta}$ is the estimate of $\Theta=(\theta,\alpha)$. Plot of the Kaplan-Meier estimator $\hat{S}(t)$ and $S(t,\hat{\Theta})$ of considered lifetime models implicitly shows that the observed lifetime, drop at each distinct time. Further, we observed that a parametric model ETL in Figure 2 (bottom left panel) with survivor function $S(t,\hat{\theta})$ has not differ too much from non-parametric (Kaplan-Meier) estimates of $\hat{S}(t)$. Also, the sampling variability in Kaplan-Meier estimator $\hat{S}(t)$ at 99% confidence limit, contains survival time ($S(t,\hat{\theta})$) of ETL model.

From Table 2, we may say that, in general, the standard error $SE(\hat{S}(t))$ of $\hat{S}(t)$ is decreasing when increasing the time. It is important to observe that the length of confidence limit decreases as time increases. Hence, on the basis of Kaplan-Meier estimator plot Figure 2, 3 and Table 2, we have justified that the ETL satisfactorily is fit for this data.

9 Conclusion

In the present piece of work, we proposed a lifetime distribution for survival time of ovarian cancer patients data after their surgical treatment. Since the nature of the dataset is IFR so that, we have considered twelve famous IFR distributions. And we found that one parameter exponential transformed Lindley (ETL) distribution fit well in comparison to exponential, DUS exponential, GDUS exponential, Lindley, generalized Lindley, gamma, Weibull, generalized exponential, Chen's model, Poisson-exponential, and log transformed exponential (LTE) distributions. On the basis of different criterion, i.e. AIC, BIC, KS distance, p value and negative of logarithmic likelihood value, we can say that ETL distribution explains the dataset very well, which is also justified through the non-parametric procedure, i.e. the Kaplan-Meier estimator and ECDF plot. Hence, we recommended it for survival time of ovarian cancer patients after their chemotherapy surgical treatment.

Acknowledgement

The authors are grateful to the editor and the anonymous referees for the careful checking of the details and for the helpful comments that led to improvement of the paper.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

[1] Ledermann JA, Raja FA, Fotopoulou C, Gonzalez MA, Colombo N, Sessa C, and ESMO Guidelines Working Group. Newly diagnosed and relapsed epithelial ovarian carcinoma: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, **24(6)**, 24–32, (2013) .

[2] Cokkinides V, Albano J, Samuels A, Ward ME, and Thum JM. American cancer society: Cancer facts and figures. *Atlanta: American Cancer Society*, (2005).

[3] American Cancer Society. Cancer facts and figures 2013, (2013).

[4] American Joint Committee on Cancer. Ovary and primary peritoneal carcinoma. *AJCC cancer staging manual. 7th ed. New York: Springer*, 493–506, (2010).

[5] Collett D. *Modelling survival data in medical research*. CRC press, (2015).

[6] Edmonson JH, Fleming TR, Decker DG, Malkasian GD, Jorgensen EO, Jefferies JA, Webb MJ, and Kvols LK. Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. *Cancer Treatment Reports*, **63(2)**, 241–247, (1979).

[7] Singh SK, Singh U, and Kumar M. Bayesian estimation for Poisson-exponential model under progressive type-ii censoring data with binomial removal and its application to ovarian cancer data. *Communications in Statistics-Simulation and Computation*, **45(9)**, 3457–3475, (2016).

[8] Aarset MV. How to identify a bathtub hazard rate. *IEEE Transactions on Reliability*, **36(1)**, 106–108, (1987).

[9] Lindley DV. Fiducial distributions and Bayes theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, **20(1)**, 102–107, (1958).

[10] Ghitany ME, Atieh B, and Nadarajah S. Lindley distribution and its application. *Mathematics and computers in simulation*, **78(4)**, 493–506, (2008(a)).
[11] Nadarajah S, Bakouch HS, and Tahmasbi R. A generalized Lindley distribution. *Sankhya B*, **73(2)**, 331–359, (2011).
[12] Sharma VK, Singh SK, Singh U, and Agiwal V. The inverse Lindley distribution: a stress-strength reliability model with application to head and neck cancer data. *Journal of Industrial and Production Engineering*, **32(3)**, 162–173, (2015).
[13] Ghitany ME, Al-Mutairi DK, and Nadarajah S. Zero-truncated Poisson–Lindley distribution and its application. *Mathematics and Computers in Simulation*, **79(3)**, 279–287, (2008(b)).
[14] Deniz EG and Ojeda EC. The discrete Lindley distribution: properties and applications. *Journal of Statistical Computation and Simulation*, **81(11)**, 1405–1416, (2011).
[15] Maurya SK, Kaushik A, Singh SK, and Singh U. A new class of distribution having decreasing, increasing and bathtub-shaped failure rate. *Communications in Statistics-Theory and Methods*, **46(20)**, 10359-10372, (2017(a)).
[16] Kumar D, Singh U, and Singh SK. A method of proposing new distribution and its application to bladder cancer patient data. *Journal of Statistics Applications and Probability Letters*, **2(3)**, 235–245, (2015).
[17] Maurya SK, Kaushik A, Singh RK, Singh SK, and Singh U. A new method of proposing distribution and its application to real data. *Imperial Journal of Interdisciplinary Research*, **2(6)**, 1331–1338, (2016).
[18] Maurya SK, Kaushik A, Singh SK, and Singh U. A new class of exponential transformed Lindley distribution and its application to yarn data. *International Journal of Statistics & Economics*, **18(2)**, 135–151, (2017(b)).
[19] Gupta RC, Gupta PL, and Gupta RD. Modeling failure time data by Lehmann alternatives. *Communications in Statistics-Theory and Methods*, **27(4)**, 887–904, (1998).
[20] Chen Z. A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics and Probability Letters*, **49(2)**, 155–161, (2000).
[21] Cancho VG, Louzada NF, and Barriga GD. The Poisson-exponential lifetime distribution. *Computational Statistics Data Analysis*, **55**, 677–686, (2011).
[22] Akaike H. A new look at the statistical model identification. *IEEE transactions on automatic control*, **19(6)**, 716–723, (1974).
[23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL http://www.R-project.org/.
[24] Lawless JF. *Statistical models and methods for lifetime data.* John Wiley & Sons, (2011).
[25] Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, and Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. ii. analysis and examples. *British journal of cancer*, **35(1)**, 1–39, (1977).

Manoj Kumar is and Assistant Professor of Statistics at Central University of Haryana. He received the Ph. D. degree in Statistics from Banaras Hindu University. He is working on Bayesian Inferences for lifetime models. He published many research articles in various journals of repute.

Sandeep Kumar Maurya is and Assistant Professor of Statistics at Central University of South Bihar. Prior to this, he was working as an Assistant Professor at Banasthali Vidyapith. He received the Ph. D. degree in Statistics from Banaras Hindu University. He is working on Statistical Inferences for lifetime models. He is trying to develop some fruitful lifetime models that may be workable in various situations of real life.

Sanjay Kumar Singh is a Professor of Statistics at Banaras Hindu University. He received the Ph. D. degree in Statistics afrom Banaras Hindu University. His main area of interest is Statistical Inference. Presently, he is working on Bayesian principle in life testing and reliability estimation, analysing the demographic data and making projections based on the technique. He also acts as a reviewer in different international journals of repute. Under his supervision, more than six students were awarded with Ph.D. degree.

Umesh Singh is a Professor of Statistics, and the former Head of Department and coordinator of DST Centre for Interdisciplinary Mathematical Science at Banaras Hindu University. He received the PhD degree in Statistics from Rajasthan University. He is a referee and an Editor of several international journals in the frame of pure and applied Statistics. He is a founder Member of Indian Bayesian Group. He started research while dealing with the problem of incompletely specified models. A number of problems related to the design of experiment, life testing and reliability, etc. were dealt with. For some time, he worked on the admissibility of preliminary test procedures. After some time, he was attracted to the Bayesian paradigm. At present, his main field of interest is Bayesian estimation for lifetime models. Under his supervision, more than twelve students awarded with Ph.D. degree.

Anurag Pathak is a very young researcher. He recently completed his M. Sc. Degree in Statistics from Babasaheb Bhimrao Ambedkar University. Presently, he is enrolled as a research scholar in the Department of Statistics at Central Univeristy of Haryana.